دوفصلنامه زبان‌شناسی گویش‌های ایرانی – سال ۴، شماره ۲، (۱۳۹۸) ۲۶۷ – ۲۹۰

شاپا: ۳۵۷۴ – ۲۵۳۸

زبان‌شناسی گویش‌های ایرانی

نشانی اینترنتی مجله: http://jill.shirazu.ac.ir

شاخص‌ها و مراحل ساخت پیکرهٔ زبانی: گونهٔ نوشتاری و گفتاری

الهام علایی ابوذر *

استادیار پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، تهران، ایران

تاریخچه مقاله:

دریافت: ۲۴ دی ماه ۹۸

پذیرش: ۶ خرداد ماه ۹۹

واژه‌های کلیدی:

پیکره

شاخص‌های کلیدیِ ساخت پیکره

فرآیند ساخت پیکره

گونهٔ نوشتاری

گونهٔ گفتاری

چکیده

این پژوهش تلاش دارد با جمع‌آوری اطلاعات مربوط به شاخص‌ها و مراحل ساخت پیکرهٔ زبانی، به پژوهشگران در زمینهٔ ساخت انواع پیکره‌های زبانی کمک کند. در این راستا، در این مقاله، پس از بررسی نظرات پژوهشگرانی که اقدام به ساخت پیکره‌هایی در زبان‌های مختلف کرده‌اند، به شاخص‌های کلی ساخت پیکره‌های زبانی پرداخته می‌شود. این شاخص‌ها مربوط به ساخت گونه‌های متنی و گفتاری پیکره است که نمونه‌گیری، نمایندگی، توازن، اندازه، نوع پیکره و یک دستی را شامل می‌شوند. سپس، فرآیند ساخت پیکرهٔ متنی ارائه می‌شود که انتخاب متون، پیش‌پردازش متون و حاشیه‌نویسی را در بر می‌گیرد و در این راستا به تفصیل دربارهٔ هر یک از مراحل توضیح داده می‌شود. در پایان، فرآیند ساخت پیکرهٔ گفتاری بیان می‌شود که جمع‌آوری داده‌ها، آوانویسی، نمایش و حاشیه‌نویسی و دسترسی را در بر می‌گیرد. دربارهٔ هر یک از مراحل مذکور نیز به تفصیل توضیح داده می‌شود.

* نویسندهٔ مسئول

آدرس ایمیل: alayi@irandoc.ac.ir (الهام علایی ابوذر)

۱. مقدمه

برای تعریف پیکره[1]، و برای پیشگیری از هر گونه ابهام، ابتدا بهتر است بدانیم چه چیزی عمدتاً پیکره محسوب نمی‌شود. مجموعهٔ متون زبانی بسیاری وجود دارد که نمی‌توان آن‌ها را پیکره نامید. به عنوان مثال، وب جهانی پیکره نیست زیرا مدام در حال تغییر است و بر اساس رویکردهای زبان‌شناسی طراحی نشده است. آرشیو نیز پیکره محسوب نمی‌شود زیرا دلایل جمع‌آوری متون در پیکره و آرشیو متفاوت است و هر کدام اولویت‌های خاصی را در این راستا دنبال می‌کنند (سینکلر[2]: ۲۰۰۴). پیکره، مجموعه‌ای نظام‌مند، رایانه‌ای‌شده، معتبر و موثق از زبان است که برای مطالعات زبان‌شناسی مورد استفاده قرار می‌گیرد. مجموعهٔ داده‌های زبانی موجود در پیکره هم می‌تواند به صورت متن و هم برگردان مکتوب گفتار ضبط شده باشد. اصولاً هر مجموعه‌ای که بیش از یک متن داشته باشد را می‌توان پیکرهٔ متنی[3] خواند. اما، واژهٔ «پیکره» یا عبارت «پیکرهٔ متنی» نوعاً در زبان‌شناسی مدرن به‌کار می‌رود و به مجموعه‌ای از متون اطلاق می‌شود که برای تجزیه و تحلیل زبان‌شناسی یا پردازش زبان طبیعی مورد استفاده قرار می‌گیرد.

پیکره می‌تواند هم از متون کامل و هم از گزیده‌هایی بسیار طولانی از متون کامل تشکیل شود. معمولاً متون پیکره‌ها به نحوی انتخاب می‌شوند که نمایندهٔ نوع خاصی از زبان یا گونه‌ای از زبان باشند؛ مثلاً ممکن است پیکره به شکلی جمع‌آوری شود که زبان انگلیسی به‌کار رفته در کتاب‌های تاریخی یا گونهٔ کانادایی زبان فرانسوی را نمایندگی کند. در تهیهٔ پیکرهٔ متنی، ویژگی هایی مانند نمونه‌گیری و نمایندگی، اندازه، قابلیت خوانش توسط رایانه و ارجاع استاندارد و معتبر لحاظ می‌شود. از آن‌جا که تحلیل داده‌ها در پیکرهٔ مورد مطالعه، بر اساس شمارش دقیق جمله‌ها، عبارات، واژه‌ها و حتی تکواژها است، متون مورد مطالعه حتماً باید قابلیت خوانش توسط رایانه را داشته باشند. این متون معمولاً به شکل مجموعه داده‌های الکترونیکی ذخیره می‌شوند. در حقیقت، پیکره مجموعه‌ای نسبتاً بزرگ از متون الکترونیکی است که به صورت حساب‌شده‌ای

1. corpus
2. Sinclair
3. text corpus

حاشیه‌نویسی[1]، برچسب‌گذاری و دسته‌بندی شده‌اند و در نتیجه امکان بررسی‌های بسیار متنوعی را برای کاربر فراهم می‌آورند (مک‌انری[2] و ویلسون[3]: ۲۰۰۱).

اتکینز[4] و همکاران (۱۹۹۲) انواع پیکره را از چند دیدگاه مختلف بررسی کرده‌اند و تقسیم‌بندی نه‌گانه‌ای از آن‌ها ارائه کرده‌اند که عبارتند از: بر اساس اندازهٔ متون: متن کامل[5]، نمونه‌ای[6] و نظارتی[7]، بر اساس امکان افزودن متون به پیکره: بسته[8] و باز[9]، بر اساس تاریخ متون: هم‌زمانی[10] و درزمانی[11]، بر اساس نوع کاربرد: عمومی[12] و اصطلاح‌شناسی[13]، بر اساس تعداد زبان‌های به‌کار رفته: یک‌زبانه[14]، دوزبانه[15] و چندزبانه[16]، بر اساس زبان به‌کار رفته در پیکره (مثلاً فارسی، انگلیسی، روسی و ...)، بر اساس ارتباط زبان‌های استفاده‌شده در پیکره: منفرد[17]، موازی[18]، بر اساس نوع مدیریت متون: مرکزی[19] و پوسته‌ای[20] و بر اساس امکان اشتراک متون در پیکره‌ها: هسته‌ای[21] و پیرامونی[22].

البته تقسیم‌بندی‌های دیگری نیز بر اساس شکل داده‌ها انجام شده است که عبارتند از: (پرینت‌شده، متن الکترونیکی، گفتار دیجیتالی‌شده، ویدئویی یا تلفیقی از چند شکل)، روش طراحی پیکره (متوازن، هرمی و سایر)، اندازهٔ پیکره (اندازهٔ ثابت، اندازهٔ کنترل‌شده)، استفاده از

1. annotation
2. McEnery
3. Wilson
4. Atkins
5. whole text
6. samples
7. monitor
8. open
9. closed
10. synchronic
11. diachronic
12. general
13. thesaurus
14. monolingual
15. bilingual
16. multilingual
17. single
18. parallel
19. central
20. cluster
21. nuclear
22. Perimeter

متن اصلی یا ترجمه‌شده، نوع استفاده (عمومی یا تخصصی)، گفتاری[1] یا نوشتاری‌بودن پیکره، نوع حاشیه‌نویسی (ساده، نشانه‌گذاری‌شده). متون موجود در پیکره‌های متنی حاوی هزاران/ میلیون‌ها کلمه هستند که گویشوران در بافت طبیعی گفتار یا نوشتار استفاده می‌کنند. در تهیهٔ پیکره‌های متنی از منابع گوناگونی استفاده می‌شود که مهم‌ترین آن‌ها متون روزنامه‌ها و کتاب‌ها است. امروزه صفحات وب نیز به عنوان منبع غنی تهیهٔ پیکره‌های متنی مورد استفاده قرار می گیرند؛ زیرا صفحات وب امکان جمع‌آوری متون مختلف، به نحوی که نمایندهٔ زبان مورد مطالعه باشد، را فراهم می‌آورند و این امر از طریق جمع‌آوری متون گوناگون مربوط به ژانرها و نویسنده های مختلف انجام می‌پذیرد. در نتیجه، پیکرهٔ متنی می‌تواند تصویر معقولی از زبان مورد مطالعه ارائه دهد.

۲. پژوهش‌های انجام‌شده در زمینهٔ ساخت پیکرهٔ زبانی

نورلینگ[2] (۱۹۹۳) به بررسی برخی از ابعاد فنی ساخت و استفاده از پیکره‌های متنی به منظور ساخت فرهنگ لغت می‌پردازد. برخی از مواردی که او در این پژوهش به آن‌ها پرداخته است انواع پیکره، انواع ابزار، الگوی استاندارد، استفاده از پیکره، انواع تجزیه و تحلیل، ابزارهای محاسباتی و برچسب‌گذاری دستوری را شامل می‌شود. واینه[3] (۲۰۰۵) آرای نویسندگان و پژوهشگران سرشناس حوزهٔ ساخت پیکره را ارائه می‌کند که پیشنهادهای کاربردی برای ساخت پیکره ارائه می‌دهند. کسانی می‌توانند از راهنمایی‌های این پژوهشگران استفاده کنند که در مرحلهٔ ساخت پیکرهٔ زبانی هستند. از موضوعاتی که در این کتاب مورد بررسی قرار می‌گیرد می‌توان به تعاریف کلی مربوط به پیکره، اضافه‌کردن حاشیه‌نویسی زبانی، نوع داده‌های مورد استفاده و پیکرهٔ آوایی اشاره کرد.

بی‌جن‌خان و همکاران (۲۰۱۱) نیز به بررسی مسائلی می‌پردازند که طی ساخت یک منبع مکتوب زبانی، تحت عنوان «پیکره»، مورد توجه قرار گرفته‌اند. پیکرهٔ معرفی‌شده در این پژوهش حاوی ۳۵۰۵۸ پروندهٔ متنی است که هر کدام حاوی متن کامل یا نمونهٔ تصادفی از یک پروندهٔ متنی می‌باشد. اندازهٔ هر متن زنجیره‌ای با حداقل ۱۰۰۰ کلمه است. برای حاشیه‌نویسی فارسی

1. spoken corpus
2. Norling
3. Waynne

در پیکره، از دستورالعمل EAGLES استفاده شده است. در تهیهٔ این پیکره، به ساخت اضافه و هم‌نگاره‌ها که در تجزیه و تحلیل متن مشکلاتی ایجاد می‌کنند، توجه خاصی شده است. ری‌ریزو[1] (۲۰۱۰) نیز ابتدا رهنمودهایی برای طراحی و ساخت یک پیکرهٔ تخصصی بر اساس استانداردهای زبان‌شناسی پیکره‌ای ارائه می‌دهد؛ سپس، به توصیف مراحل ساخت پیکرهٔ مهندسی ارتباط از دور[2] و ویژگی‌های آن می‌پردازد. در نهایت، نتایج حاصل از تجزیه و تحلیل آماری مربوط به پیکره را ارائه می‌کند.

کلود توریدا[3] (۲۰۱۶) نیز به توضیح مرحله به مرحلهٔ ساخت پیکرهٔ تخصصی[4] و تهیهٔ فهرست واژگان حاشیه‌نویسی‌شده و مبتنی بر فراوانی واژگان در پیکره می‌پردازد. محمدی (۱۳۹۱) چگونگی ساخت پیکرهٔ تطبیقی فارسی- انگلیسی را توضیح می‌دهد. او برای ایجاد این پیکره از اسناد خبری روزنامه‌های همشهری و بی‌بی‌سی استفاده می‌کند و از اسناد به‌دست‌آمده، معیارهایی مانند تعداد کلمات کلیدی مشترک، اسم‌های خاص یکسان، عنوان‌های مشابه و فاصلهٔ تاریخ انتشار دو خبر را استخراج می‌کند. در این پژوهش معیارهای به‌دست‌آمده بر اساس میزان اهمیت آن‌ها در ترازبندی متون، با وزن‌های مختلف با یکدیگر ترکیب و سپس جمله‌های موازی از پیکرهٔ تطبیقی ساخته‌شده، استخراج می‌شوند.

پیکره‌ها می‌توانند آوایی باشند و برای اهداف خاصی تهیه شوند. دوراند[5] و همکاران (۲۰۱۴) در فصلی از کتاب خود به فرآیند ساخت پیکره‌های آوایی می‌پردازند. آن‌ها، در این راستا، پس از تعریف پیکرهٔ آوایی، به بحث پیرامون اجزای مهم ساخت چنین پیکره‌هایی می‌پردازند؛ این اجزا ذخیرهٔ پیکره، چگونگی به اشتراک گذاشتن و استفادهٔ مجدد از پیکره، سؤال‌های مربوط به نمایندگی و اندازهٔ پیکره، انتخاب دادهٔ خام و حاشیه‌نویسی پیکره را شامل می‌شوند. آن‌ها در نهایت، نظریه‌های مربوط به فرآیند ساخت پیکره را مورد بحث قرار می‌دهند. آیت (۱۳۸۹) نیز مراحل تهیه یک دادگان دایفون ویژهٔ زبان فارسی را این گونه بیان می‌کند:

> در این پژوهش، ابتدا پایگاه واژگانی که دایفون‌های زبان را شامل شوند، تهیه شد. سپس نرم‌افزاری طراحی و پیاده‌سازی شد که با گرفتن صورت‌های واجی واژه‌ها، دایفون‌هایی را مشخص می‌کند که قرار است از آن استخراج شوند. در مرحلهٔ بعد سیگنال‌های گفتاری

1. Rea Rizzo
2. telecommunication
3. Claude Toriida
4. specialized corpus
5. Durand

واژه‌ها ضبط و بررسی شد. در پایان نیز جداسازی دایفون‌ها و تهیهٔ دادگان مورد نظر صورت پذیرفت. برای افزایش دقت دادگان تهیه‌شده، مراحل جداسازی دایفون‌ها از سیگنال‌های گفتاری ضبط‌شده با استفاده از سه روش شنوایی، بررسی سیگنال زمانی و مطالعهٔ طیف‌نگاشت، ارزیابی و از ترکیب هر سه روش برای افزایش دقت دادگان استفاده شد.

بی‌جن‌خان (۱۳۸۳) به نقش پیکره‌های زبانی در نوشتن دستور زبان می‌پردازد. وی معتقد است که اگرچه با تأکید زبان‌شناسان زایشی بر داده‌های بالقوه و نه بالفعل زبانی برای کشف پیچیدگی‌های ذهن انسان، جایگاه پیکره‌ها در تجزیه و تحلیل دستوری تضعیف شده است، اما با رشد سریع فن‌شناسی اطلاعات، ضرورت تهیهٔ پیکره‌ها و دادگان‌های زبانی با حجم بسیار بالا از اولویت بالایی برخوردار شد. وی پس از تعریف پیکرهٔ زبانی که از کریستال[1] (۱۹۹۴) نقل قول می‌کند، به بررسی نقدهایی می‌پردازد که بر پیکره‌های زبانی وارد شده‌اند. سپس به رابطهٔ دستور زبان و پیکرهٔ زبانی می‌پردازد و در نهایت یک نرم‌افزار (نرم‌افزار پیکرهٔ زبانی) را معرفی و توصیف می‌کند. مراحل تولید پیکرهٔ زبانی عبارتند از جمع‌آوری داده‌ها، آماده‌سازی داده‌ها، برچسب‌دهی نحوی-معنایی و آمارگان. نحوهٔ جستجوی هوشمند در پیکرهٔ زبانی نیز در این اثر توضیح داده می‌شود.

۳. شاخص‌های ساخت پیکرهٔ زبانی

خبرگانی که به تجزیه و تحلیل پیکرهٔ زبانی می‌پردازند لزوماً در ساخت پیکره/ پیکره‌هایی که از آن‌ها استفاده می‌کنند، تبحر ندارند؛ در حقیقت همیشه با این خطر مواجه هستند که ممکن است پیکره‌ای بسازند که حاوی اطلاعاتی باشد که یا از قبل می‌دانستند، و بنابراین اطلاعات جدیدی نیست، و یا می‌توانند جزئیات زبان‌شناسی آن را حدس بزنند. حال آن‌که شرایط مطلوب این است که پیکره باید طوری طراحی و ساخته شود که الگوهای ارتباطی جوامعی در پیکره منعکس شود که آن زبان در آن‌ها استفاده می‌شود. صرف نظر از محتوای اسناد مکتوب و گفتاری پیکره، اسناد باید حاصل جمع‌آوری مکالمات مردم و مطالبی باشد که مردم می‌نویسند یا می خوانند. بنابراین، محتویات پیکره باید بدون در نظر گرفتن زبان پیکره و بر اساس کارکرد کاربردی زبان در جامعهٔ مورد استفاده باشد. مهم‌ترین شاخص‌های اصلی ساخت پیکرهٔ زبانی عبارتند از:

1. Crystal

نمونه‌گیری[۱]، نمایندگی[۲]/ نماینده‌بودن اندازهٔ داده‌های زبانی، توازن[۳]، اندازه[۴]، نوع پیکره: عمومی[۵] یا تخصصی[۶] و یک‌دستی[۷] (سینکلر: ۲۰۰۴).

۱-۳. نمونه‌گیری

نمونه‌گیری در واقع عمل انتخاب متون مربوط به هر ژانر با توجه به هدف تهیهٔ پیکره است. سؤال اساسی این است که چگونه می‌توان از یک زبان نمونه‌گیری کرد. برای تصمیم‌گیری دربارهٔ نمونه-گیری دو مسئله باید در نظر گرفته شود: ۱. گرایش/ سوگیری[۸] ما نسبت به زبان یا گونه‌ای که قرار است از آن نمونه‌گیری کنیم. طراحان پیکره‌های اولیه، مانند «پیکرهٔ براون»[۹] و آن‌هایی که بر اساس مدل «براون» ساخته شده بودند، به دنبال ساخت پیکره‌ای نزدیک به زبان معیار[۱۰] بودند؛ بنابراین، تنها مستندات منتشرشده انتخاب می‌شدند و اکثر گونه‌های زبانی به طور خودکار حذف می‌شدند. بیشتر پیکره‌هایی که امروزه ساخته می‌شوند نیز از همان سیاست پیروی می‌کنند. برخی از پیکره‌ها بر اساس متغیری از پیش تعیین‌شده ساخته می‌شوند؛ به عنوان مثال، یک پیکرهٔ تاریخی (به لحاظ ساختار درونی)، عمداً مقابله‌ای طراحی شده است و هدف آن ارائهٔ تصویری واحد و منسجم از زبان طی زمان نیست.

نوع دیگر پیکره که بُعد زمان را نیز لحاظ کرده است، پیکرهٔ پایشگر/ نظارتی[۱۱] است که نمونه های زبانی را در بازهٔ منظم جمع‌آوری می‌کند و نرم‌افزار آن، تغییرات واژگان و عبارات را ثبت می‌کند. پیکره‌های موازی[۱۲] (هر پیکره‌ای که از بیش از یک زبان استفاده می‌کند)، حاوی اجزای درونی مقابله‌ای هستند که به تفاوت‌ها و شباهت‌های زبانی دو زبان یا بیشتر در سطوح مختلف

1. sampling
2. representativeness
3. balance
4. size
5. general corpus
6. specialized corpus
7. homogeneity
8. orientation
9. Brown corpus
10. standard language
11. monitor corpus
12. parallel corpora

می‌پردازند. این پیکره‌ها را می‌توان پیکره‌های مقابله‌ای[1] نامید، انگیزهٔ اصلی ساخت چنین پیکره-هایی، تقابل اجزای اصلی زبان‌های مورد مطالعه آن‌ها است. بنابراین، سوگیری طراحان پیکره و این‌که به دنبال ساخت چه نوع پیکره‌ای با تقابل قراردادن چه اجزایی باشند، محتوای پیکره‌ها و نمونه‌گیری را تعیین می‌کند. ۲. معیارهایی که بر اساس آن‌ها نمونه‌ها انتخاب می‌شوند. برخی از معیارهایی که بر اساس آن‌ها نمونه‌گیری صورت می‌پذیرد عبارتند از: شکل[2] متن (گفتاری/ نوشتاری/ الکترونیکی)، نوع متن (کتاب/ مجله/ حتی نامه)، حوزهٔ متن (آکادمیک/ عمومی)، زبان (زبان‌ها یا گونه‌های زبانی پیکره) و مکان متن (به عنوان مثال، انگلیسی بریتانیا باشد یا استرالیا).

نمونه‌های زبانی برای یک پیکره باید تا حد امکان شامل اکثر اسناد مکتوب یا رخدادهای گفتاری مکتوب‌شده باشند، یا حداقل به هدف مذکور نزدیک باشند. در نمونه‌گیری معمولاً مواردی از قبیل واحد نمونه‌گیری و محدودهٔ نمونه (کتاب/ روزنامه/ مقالات/ فصول کتاب/ مجله)، شبکهٔ نمونه (فهرستی از تمام واحدهای ممکن؛ به عنوان مثال، شبکهٔ نمونهٔ «پیکرهٔ براون» شامل فهرستی از کتاب‌ها، ماهنامه‌ها و هفته‌نامه‌های موجود در دانشگاه «براون» (سینکلر: ۲۰۰۴)، کاربران پیکره، روش نمونه‌گیری، اندازهٔ نمونه (متن کامل/ بریده‌هایی از متون) و ژانر متن در نظر گرفته می‌شوند. نمونه‌ها معمولاً بر اساس ژانرهایی مانند اثر منثور داستانی و غیرداستانی، اثر شعری از شاعران معاصر، مجله و نشریهٔ علمی و ادبی و تخصصی، نمایش‌نامه، فیلم‌نامه، ادبیات کودکان، روزنامه و نشریهٔ خبری همه‌پسند و متنوع، کتاب‌های درسی دبستانی و راهنمایی و دبیرستان، کتاب‌های تألیفی در زمینهٔ آموزش زبان به غیرفارسی‌زبانان، مجموعه‌ای از قوانین و مقررات، دفترچهٔ راهنما و بروشورها انتخاب می‌شوند. روش نمونه‌گیری معمولاً به دو صورت «نمونه‌گیری تصادفی ساده»[3] و «نمونه‌گیری لایه‌ای»[4] است. نمونه‌گیری تصادفی ساده، روشی است که در آن همهٔ واحدهای نمونه، داخل شبکهٔ نمونه قرار داده و شماره‌گذاری می‌شوند. در این رویکرد، اعضای جامعهٔ آماری به صورت تصادفی انتخاب می‌شوند و احتمال انتخاب هر عضو از جامعهٔ آماری از طریق نسبت اندازهٔ نمونه به اندازهٔ جامعهٔ آماری محاسبه می‌شود (واتام[5]: ۲۰۱۵).

1. Contrastive corpora
2. mode
3. Simple Random Sampling (SRS)
4. Stratified sampling
5. Wattam

در نمونه‌گیری تصادفی، همهٔ متون موجود در جامعهٔ آماری، از شانس یکسانی برای انتخاب شدن برخوردار هستند. این روش مرتکب خطاهایی نمی‌شود که ریشه در طبقه‌بندی دارند. در نمونه‌گیری لایه‌ای اندازهٔ متونی که به عنوان نمونه استفاده خواهند شد، از قبل تعیین می‌شود، داده‌ها به گروه‌های همگن تقسیم می‌شوند و نسبت استفاده از هر گروه از داده‌ها نیز مشخص می‌شود. در نهایت داده‌ها از هر لایه به طور تصادفی انتخاب می‌شوند. در نمونه‌گیری لایه‌ای، در واقع یک نمونهٔ تصادفی انتخاب‌شده به مجموعه شبکه‌هایی با اندازه‌های خاص، شکسته می‌شود. این روش مستلزم تقسیم‌کردن مجموعه به طبقات خاص متقابل و نمونه‌گیری تصادفی از این طبقات است (واتام: ۲۰۱۵).

۲-۳. نمایندگی

سؤال مطرح در طراحی هر پیکرهٔ زبانی این است که پیکرهٔ زبانی چه ویژگی‌هایی باید داشته باشد تا بتوان آن را نمایندهٔ زبان تلقی کرد. زمانی می‌توان گفت یک پیکره نمایندهٔ یک گونهٔ زبانی است که یافته‌های به‌دست‌آمده از تجزیه و تحلیل محتوای پیکره، قابل تعمیم به کل گونهٔ زبانی باشد. به عنوان مثال، فرض کنید بیش‌ترین سند مکتوبی که در ایران مورد استفاده قرار می‌گیرد، روزنامه باشد و از میان روزنامه‌ها، روزنامهٔ همشهری دارای بیش‌ترین خواننده باشد. در این صورت آیا در تهیهٔ پیکرهٔ فارسی باید بیش‌تر متون مستخرج از روزنامهٔ همشهری باشند؟ عمده‌ترین چالش چنین انتخاب‌هایی، مسئلهٔ زبان است؛ به این معنی که احتمالاً گونهٔ زبانی استفاده شده در روزنامه‌ها برای گزارش اخبار روز است و استفاده از این گونه به عنوان مدل کلی نوشتار توصیه نمی‌شود. این مثال اهمیت انتخاب نمونهٔ زبانی را نشان می‌دهد که گونهٔ زبان مورد استفاده در یک جامعه را نمایندگی می‌کند.

پیکره باید نمایندهٔ گونهٔ زبانی باشد تا به عنوان پایه و اساس تعمیم‌های زبانی مورد استفاده قرار گیرد. ارزیابی مسئلهٔ نمایندگی یک پیکره به چگونگی تعریف جامعهٔ آماری (زبان یا گونهٔ زبانی) مورد استفاده برای نمونه‌گیری بستگی دارد. در تعریف جامعهٔ آماری (زبان یا گونهٔ زبانی)، دو جنبه لحاظ می‌شوند: حد و اندازهٔ جامعه/ قالب نمونه‌گیری (چه متونی انتخاب و چه متونی حذف شوند؛ به عنوان مثال، در پیکرهٔ براون حد و اندازه به این صورت تعریف می‌شود: همهٔ متون انگلیسی منتشرشده در سال ۱۹۶۱) و سازمان‌دهی سلسله‌مراتبی درون جامعه (کدام طبقه‌بندی

های مربوط به متون در گونهٔ زبانی مورد مطالعه در نظر گرفته شوند؛ به عنوان مثال، در پیکرهٔ براون این سلسله‌مراتب به این صورت تعریف می‌شود: «۱۵ طبقهٔ اصلی از متون و تعداد بسیار زیادی زیرمجموعهٔ این طبقات به تفکیک ژانرهای مربوط به هر متن» (فرانسیس[1] و کوسرا[2]: ۱۹۶۴، نقل از بایبر[3]: ۱۹۹۳). به طور کلی، سؤال پژوهش، تعیین‌کنندهٔ میزان نماینده‌بودن پیکره است. به عنوان مثال، اگر کسی می‌خواهد پیکره‌ای بسازد که نمایندهٔ فارسی عمومی است، استفاده از متون روزنامه‌ها مناسب نیست؛ همین‌طور اگر کسی می‌خواهد پیکره‌ای بسازد که نمایندهٔ روزنامه‌های انگلیسی باشد، نمی‌تواند تنها از روزنامهٔ Times استفاده کند.

۳-۳. توازن

توازن پیش‌نیاز نمایندگی است. میزان استفاده از گونهٔ گفتاری زبان یکی از عواملی است که توازن یک پیکره را تعیین می‌کند؛ در حقیقت، پیکرهٔ متوازن حاوی گونهٔ نوشتاری و گفتاری زبان است. اکثر پیکره‌ها از این نظر، از توازن خوب و مناسبی برخوردار نیستند، زیرا به اندازهٔ کافی حاوی گونهٔ گفتاری زبان نیستند. این در حالی است که مردم از گونهٔ گفتاری زبان بیش‌تر از گونهٔ نوشتاری استفاده می‌کنند. بنابراین، در حالت بهینه میزان استفاده از گونهٔ گفتاری در یک پیکرهٔ عمومی باید ۵۰ تا ۹۰ درصد باشد. البته، باید این نکته را هم مدنظر قرار داد که اجزای نوشتاری یک پیکره زیرشاخه‌هایی مانند روزنامه، مجله، کتاب و غیره را دارد که برای هر کدام لزوماً معادل گونهٔ گفتاری زبان (مانند رسانه‌ها، جلسات، مکالمات) وجود ندارد، بنابراین حالت بهینه هیچ‌گاه فراهم نمی‌شود.

میزان تخصصی‌بودن متون پیکره عامل دیگر تأثیرگذار بر مسئلهٔ توازن پیکره است؛ استفاده از متون بسیار تخصصی در پیکرهٔ عمومی باعث بر هم خوردن توازن پیکره می‌شود. بنابراین، نوع متن از اهمیت بالایی برخوردار است. ساخت پیکرهٔ بسیار تخصصی مستلزم استفاده از متون بسیار تخصصی مربوط به هر حوزه است و این مسئله توازن را بر هم نمی‌زند و روش ساخت چنین پیکره‌هایی با ساخت پیکره‌های عمومی کمی تفاوت دارد. به طور کلی، دو مسئلهٔ «نمایندگی» و «توازن» تعریف خیلی دقیقی ندارند. با این وجود باید در طراحی پیکره و انتخاب اجزای پیکره لحاظ شوند (سینکلر: ۲۰۰۴). به عنوان مثال، اگر بخواهیم پیکره‌ای از روزنامه‌های ایرانی منتشرشده در

1. Francis
2. Kucera
3. Biber

یک بازهٔ زمانی خاص را تهیه کنیم، اولین مسئله‌ای که به نظر می‌رسد جمع‌آوری تمام مقالات روزنامه‌ها می‌باشد و سپس کار ساخت پیکره شروع می‌شود. چالش مربوط به توازن پیکره در این مورد به این صورت است که مقالات یک بخش از روزنامه ممکن است بسیار طولانی‌تر از بخش/ بخش‌های دیگر روزنامه باشند.

به اعتقاد برخی از پژوهشگران می‌توان مشکلاتی از این دست را به این صورت حل کرد که نمونه‌های حاوی ۲۰۰۰ تا ۵۰۰۰ کلمه را برای هر متن تهیه و در پیکره وارد کرد. دربارهٔ نوع متون هم این باور وجود دارد که زبان متون انتخاب‌شده نباید خیلی غیررسمی یا خیلی رسمی باشد. البته انتخاب نوع متن خود وابسته به هدف پژوهش است. اگر هدف، بررسی گونهٔ رسمی زبان در بافت اجتماعی باشد، قطعاً باید از پیکره‌ای استفاده شود که حاوی مطالبی به زبان رسمی باشد. بنابراین، توازن قابل قبول یک پیکره بر حسب استفاده‌های آن تعریف می‌شود. به عنوان مثال، یک پیکرهٔ عمومی (مانند «پیکرهٔ ملی بریتانیایی» یا پیکرهٔ نوشتاری «براون») که حاوی داده‌های نوشتاری و گفتاری باشد، متوازن در نظر گرفته می‌شود؛ چنین پیکره‌هایی حد وسیعی از انواع متون را پوشش می‌دهد که نمایندهٔ زبان یا گونهٔ زبانی مورد مطالعه هستند.

علی‌رغم اینکه در نظر گرفتن توازن، شرط اساسی طراحی پیکره است، هیچ مقیاس و معیار علمی قابل اعتمادی برای سنجش توازن یک پیکره وجود ندارد. این مسئله عمدتاً شمی و تخمینی است؛ با وجود این تنها اطمینانی که وجود دارد این است که طبقه‌بندی انواع متون در رسیدن به توازن، از اهمیت بالایی برخوردار است و باعث رده‌بندی متون می‌شود (اتکینز و همکاران: ۱۹۹۲). به عنوان مثال، «پیکرهٔ ملی بریتانیایی[1]» عموماً به عنوان پیکره‌ای متوازن شناخته می‌شود و مدل این پیکره در ساخت پیکره‌های بسیاری از جمله «پیکرهٔ ملی آمریکایی[2]»، «پیکرهٔ ملی کره‌ای[3]»، «پیکرهٔ ملی لهستانی[4]» و «پیکرهٔ مرجع روسی[5]» دنبال شده است (شاروف[6]: ۲۰۰۳، نقل از مک انری و همکاران: ۲۰۰۶). ساختار پیکرهٔ ملی بریتانیایی در جدول (۱) قابل مشاهده است.

1. British National Corpus (BNC)
2. American National Corpus
3. Korean National Corpus
4. Polish National Corpus
5. Russian Reference Corpus
6. Sharoff

جدول (۱) ساختار پیکرهٔ ملی بریتانیایی : نوع متنی این پیکره

(برگرفته از شاروف: ۲۰۰۳، نقل از مک‌انری و همکاران: ۲۰۰۶)

منابع (درصد)		تاریخ (درصد)		حوزه/موضوع (درصد)	
کتاب	٪۵۸,۵۸				
نشریه	٪۳۱,۰۸	۱۹۶۰-۱۹۷۴	٪۲,۲۶	تخیلی	٪۲۱,۹۱
متفرقه. منتشرشده	٪۴,۳۸	۱۹۷۵-۱۹۹۳	٪۸۹,۲۳	هنر	٪۸,۰۸
متفرقه. منتشرنشده	٪۴,۰۰	طبقه‌بندی‌نشده	٪۸,۴۹	عقاید و افکار	٪۳,۴۰
To-be-spoken	٪۱,۵۲			تجارت/ سرمایه‌گذاری	٪۷,۹۳
طبقه‌بندی‌نشده	٪۰,۴۰			فراغت	٪۱۱,۱۳
				علوم طبیعی/محض	٪۴,۱۸
				علوم کاربردی	٪۸,۲۱
				علوم اجتماعی	٪۱۴,۸۰
				امور جهان	٪۱۸,۳۹
				طبقه‌بندی‌نشده	٪۱,۹۳

این پیکره تقریباً شامل ۱۰۰ میلیون کلمه است. ۹۰٪ آن‌ها متون نوشتاری است و ۱۰٪ داده‌های مکتوب‌شدهٔ گفتاری است.

۴-۳. اندازه

منظور از اندازه در ساخت پیکره، تعداد کلمات پیکره است. فرض کلی در رویکرد آماری این است که پیکره هر چه بزرگ‌تر باشد، بهتر است. اما در واقع، هیچ رویکرد نظام‌مندی دربارهٔ تأثیر اندازهٔ پیکره و ساختار آن بر کیفیت پیکره و کاربرد آن وجود ندارد. اندازهٔ اولین نسل پیکره‌ها به یک میلیون کلمه می‌رسید که در نوع خود در آن دوره با امکانات محدود، پیشرفتی عظیم در حوزهٔ ساخت پیکره محسوب می‌شد. نسل دوم پیکره‌ها به بزرگی چندصد میلیون کلمه بودند. همین

باعث می‌شود تصور کنیم هر چه پیکره بزرگ‌تر باشد، آسان‌تر می‌توان به بررسی موارد زبانی نادر در آن‌ها پرداخت (مک‌انری و ویلسون: ۲۰۰۱، نقل از بیانچی[1]: ۲۰۱۲).

اندازهٔ پیکره تا حدی به هدف مطالعه مربوط می‌شود. به عنوان مثال، بارش[2] و اورت[3] (۲۰۱۳) معتقدند پیکره‌ای کوچک‌تر اما تمیز و متوازن، برای مطالعاتی مانند بررسی هم‌آیندها در بافت، بهتر از پیکره‌ای بزرگ برگرفته از صفحات وب یا متون روزنامه است. پیکره‌های تخصصی، به طور کلی، کوچک‌تر از پیکره‌های عمومی هستند. به عنوان مثال، اشکال گوناگون کلمه (تصریفی یا اشتقاقی) در پیکره‌های تخصصی بسیار کم‌تر از پیکره‌های عمومی است. از طرفی، پیکره‌های بزرگ امکان مطالعهٔ علل اصلی وقوع ساخت‌های واژگانی یا دستوری خاص با فراوانی بالا در زبان را فراهم می‌آورند. به عنوان مثال، مطالعهٔ واژهٔ get در ساخت مجهول انگلیسی، مستلزم بهره‌گیری از پیکره‌ای بزرگ است که بتوان این واژه را در بافت بررسی کرد.

برخی از عواملی که در تعیین اندازهٔ پیکره لحاظ می‌شوند، سرعت و کارایی نرم‌افزار و توانایی انسان در کارکردن با مقادیر زیاد داده است؛ هم رایانه و هم مغز انسان ممکن است در پردازش مقادیر زیاد داده ناتوان باشد؛ لحاظ‌کردن این مسئله ممکن است منجر به ساخت پیکره‌های کوچک یا زیرمجموعه‌سازی از پیکره‌های دیگر شود. مسئلهٔ اندازهٔ پیکره از چند منظر حائز اهمیت است: برای مقایسهٔ پیکره با پیکره‌های دیگر، انجام تجزیه و تحلیل‌های کمّی و محاسبات آماری و مسئلهٔ نمایندگی پیکره (بیانچی: ۲۰۱۲).

۵-۳. نوع پیکره: عمومی یا تخصصی

هدف تهیهٔ پیکره‌های عمومی، نمایش گونه‌های مختلف و ابعاد گوناگون زبان است. «پیکرهٔ ملی بریتانیایی» نمونه‌ای از این نوع پیکره‌ها، است (استون[4] و برنارد[5]: ۱۹۹۷). پیکره‌های عمومی معمولاً بزرگ‌تر از پیکره‌های تخصصی هستند؛ هر چه پیکره بزرگ‌تر باشد، بهتر است. پیکره‌های بزرگ، غالباً پیکره‌های مرجع خوانده می‌شوند؛ زیرا عمدتاً به عنوان پایه و اساس انواع مقایسه و قضاوت دربارهٔ گونه‌های زبانی محسوب می‌شوند و امکان مطالعهٔ مقابله‌ای زبان‌ها یا گونه‌های

1. Bianchi
2. Bartsch
3. Evert
4. Aston
5. Burnard

زبانی را فراهم می‌آورند. پیکره‌های تخصصی حاوی متونی از ژانرهای خاص یا دورهٔ زمانی یا بافت خاص هستند. آن‌ها حاوی نمونه‌هایی از ژانر مورد مطالعه هستند. به عنوان مثال، یک پیکرهٔ تخصصی می‌تواند حاوی همهٔ نمایش‌نامه‌های شکسپیر باشد. این پیکره‌ها، به عنوان منابع ارزشمند بررسی و مطالعهٔ حوزه‌ای خاص یا ژانری خاص در نظر گرفته می‌شوند. «پیکرهٔ گفتاری میشیگان»[1] و «پیکرهٔ بین‌المللی آموزشی انگلیسی»[2] نمونه‌هایی از این نوع پیکره هستند. پیکره های تخصصی حاوی متون خاصی هستند و هدف آن‌ها نمایندگی زبان آن متون است. این پیکره‌ها عمدتاً برای پاسخ به سؤالات بسیار خاصی ساخته می‌شوند. به عنوان مثال، یک پیکرهٔ پزشکی از زبانی استفاده می‌کند که پرستاران و کارکنان بیمارستان استفاده می‌کنند و حاوی کلمات تخصصی رشتهٔ پزشکی و پیراپزشکی است.

۶-۳. یک‌دستی

هیچ مقیاس دقیقی برای اندازه‌گیری میزان یک‌دستی یک پیکره تعریف نشده است. یک پیکره زمانی یک‌دست در نظر گرفته می‌شود که حاوی تفاوت‌های بارز در مشخصه‌های درون مستندات خود نباشد (کاواگلیا[3]: ۲۰۰۲). در هر گونهٔ زبانی که در پیکره استفاده می‌شود، متونی وجود دارند که با متون دیگر متفاوت هستند و در یک طبقه قرار نمی‌گیرند و در حقیقت نمی‌توانند نمایندهٔ گونهٔ زبانی مورد مطالعه باشند. سینکلر (۲۰۰۴) بر این باور است که واردکردن این متون در پیکره، تنها اندازهٔ پیکره را افزایش می‌دهد، بدون آن‌که کارکرد خاصی داشته باشد یا وجودشان در تحلیل‌های زبانی مفید واقع شود. اگر محدودیت اندازهٔ داده‌های پیکره مسئله‌ای مهم در تهیهٔ پیکره باشد، بهتر است این متون از پیکره حذف شوند یا هنگام ساخت پیکره، وارد نشوند. حال آن‌که برخی دیگر از متخصصان حوزهٔ پیکره معتقدند اعمال نظر در واقعیت داده‌های پیکره درست نیست. باید هر آن‌چه وجود دارد با توجه به توزیع فراوانی آن، نمونه‌گیری شود و به همان میزان در پیکره سهم داشته باشد. اگر یک ساخت حاشیه‌ای هم باشد، در توزیع فراوانی منعکس می‌شود.

۴. فرآیند ساخت پیکرهٔ متنی

1. Michigan Corpus of Academic Spoken English
2. international corpus of learner English
3. Cavaglia

در ساخت پیکرهٔ متنی، پس از انتخاب متونی که در پیکرهٔ متنی مورد استفاده قرار خواهند گرفت، باید پیش‌پردازش‌هایی روی متن انجام شود که به اصطلاح «بهنجارسازی متن»[1] نامیده می‌شود. این پیش‌پردازش‌ها در متون فارسی می‌تواند شامل، یک‌دست‌کردن فاصله‌ها، نشانه‌گذاری‌های درون متن، یکسان‌کردن نویسه‌های[2] استفاده‌شده در متون (مانند انواع «ی»، «ک»، همزه و غیره)، یکسان‌کردن روش اتصال وندهای گوناگون به ستاک، اصلاح غلط‌های املایی، ارتباط‌دادن کلمات چندواژه‌ای و یکسان در نظر گرفتن آن‌ها باشد. ابتدا باید همهٔ نویسه‌های متن با جایگزینی با معادل استاندارد آن، یکسان‌سازی شوند. ممکن است متون مختلف بسیار به هم شبیه باشند اما به دلیل تفاوت‌های سادهٔ ظاهری از نظر ماشین متفاوت باشند؛ به همین دلیل ابتدا باید این تفاوت‌های سادهٔ ظاهری برطرف شوند. همچنین به منظور پردازش دقیق‌تر متون، اصلاحات دیگری نیز انجام می‌شود. به طور کلی، در جهت تسهیل فرآیند ساخت پیکره و صرفه‌جویی در استفاده از نیروی انسانی و تسریع فرآیند ساخت یا تجزیه و تحلیل داده‌های پیکره، ابزارهایی تعریف شده است که می‌توان از آن‌ها استفاده کرد، یکی از این ابزارها، ابزاری برای پیش‌پردازش متن است.

پس از پیش‌پردازش متون، حاشیه‌نویسی انجام می‌شود. حاشیه‌نویسی پیکره به عمل اضافه کردن اطلاعات زبانی توضیحی-تفسیری به پیکره اطلاق می‌شود. به عنوان مثال، یکی از رایج ترین نوع حاشیه‌نویسی پیکره، اضافه‌کردن برچسب‌ها به کلمات پیکره است. برخی از پژوهشگران مانند سینکلر (۲۰۰۴، نقل از لیچ[3]: ۲۰۰۴) ترجیح می‌دهند وارد مقولهٔ حاشیه‌نویسی پیکره نشوند؛ زیرا معتقدند پیکرهٔ حاشیه‌نویسی‌نشده، پیکره‌ای خالص است و برای مطالعات زبان‌شناسی چنین پیکره‌هایی را ترجیح می‌دهند. این پژوهشگران به حاشیه‌نویسی اعتماد ندارند و با شک به آن می‌نگرند و باور دارند که حاشیه‌نویسی نمی‌تواند بدون خطا باشد. اما اکثر پژوهشگران این حوزه معتقدند که حاشیه‌نویسی، پیکره را بسیار مفیدتر و غنی‌تر از پیکرهٔ خام می‌کند؛ از نظر آن‌ها حاشیه‌نویسی، ارزش مضاعفی به پیکره می‌دهد. برخی از رایج‌ترین حاشیه‌نویسی‌ها شامل موارد زیر است (لیچ: ۲۰۰۴):

1. normalization
2. characters
3. Leech

- حاشیه‌نویسی آوایی[۱]: اضافه‌کردن اطلاعاتی دربارهٔ نحوهٔ تلفظ یک کلمه در پیکرهٔ گفتاری یا اضافه‌کردن مشخصه‌های زبرزنجیری[۲] (مانند تکیه[۳]، آهنگ کلام[۴]، وقفه[۵]) به کلمات تشکیل‌دهندهٔ یک پیکره است.
- حاشیه‌نویسی نحوی[۶]: اضافه‌کردن اطلاعاتی دربارهٔ اجزای کلام[۷] (نحوهٔ تجزیهٔ یک جمله به عبارت‌ها و اجزای تشکیل‌دهندهٔ آن) است.
- حاشیه‌نویسی معنایی[۸]: اضافه‌کردن اطلاعاتی دربارهٔ مقوله/ طبقه‌بندی معنایی کلمات پیکره است. به عنوان مثال، صورت نوشتاری <cricket> در انگلیسی، فارغ از صورت نوشتاری یا تلفظی یکسان، هم در طبقه‌بندی انواع ورزش (نوعی ورزش) قرار می‌گیرد و هم در طبقه‌بندی دیگر و متفاوتی تحت عنوان حشرات (نوعی حشره).
- حاشیه‌نویسی کاربردی[۹]: اضافه‌کردن اطلاعاتی دربارهٔ کنش گفتار[۱۰]/ نشانگرهای گفتمانی است. به عنوان مثال، پاره‌گفتار <okay> در انگلیسی، در موقعیت‌های مختلف می‌تواند به منزلهٔ اقرار و تصدیق، درخواست بازخورد، پذیرفتن یا نشانهٔ شروع مرحلهٔ جدیدی از بحث باشد.
- حاشیه‌نویسی گفتمانی[۱۱]: اضافه‌کردن اطلاعاتی دربارهٔ ارتباطات ارجاعی[۱۲] در متن. به عنوان مثال، ارتباط‌دادن ضمیر <them> و مرجع آن <the horses> در جملهٔ زیر:

I'll saddle the horses and bring them round. o

1. phonetic annotation
2. prosodic
3. stress
4. intonation
5. pause
6. syntactic annotation
7. Part Of Speech (POS)
8. semantic annotation
9. pragmatic annotation
10. speech act
11. discourse annotation
12. anaphoric

- حاشیه‌نویسی سبکی[1]: اضافه‌کردن اطلاعاتی دربارهٔ نمود گفتار (گفتار مستقیم[2]، گفتار غیرمستقیم[3]).
- حاشیه‌نویسی واژگانی[4]: اضافه‌کردن اطلاعاتی مانند ریشهٔ کلمه.

۵. فرآیند ساخت پیکرهٔ گفتاری

کارکردن روی داده‌های گفتاری زبان کاری بسیار دشوار در مقایسه با داده‌های نوشتاری است. زبان نوشتاری به راحتی قابل ذخیره در پرونده‌های متنی الکترونیکی است. مشکل اساسی در مورد داده‌های گفتاری زبان، نمود نوشتاری متنی است که خوانده‌شده یا گفتاری است که در گذشته ضبط شده است. در حالی که در مورد پیکره‌های متنی این مشکل وجود ندارد. زیرا کلمات موجود در یک پیکرهٔ متنی، قبل از ساخت پیکره وجود داشته‌اند. به طور کلی، این نکته را باید در نظر داشت که ساخت پیکرهٔ گفتاری باید تحت شرایط خاص و با توجه به هدف ساخت پیکره صورت پذیرد. به عنوان مثال، چنان‌چه زبان‌شناسی قصد بررسی فراوانی واژگانی کلمات در حجم عظیمی از داده‌های زبانی را داشته باشد، نیازی به آوانویسی دقیق و موشکافانهٔ گفتار نیست و مسئلهٔ اصلی کمیت و سرعت آوانویسی است. بنابراین آوانویسی ساده کفایت می‌کند. این در حالی است که یک آواشناس برای بررسی آوایی داده‌های گفتاری، نیازمند داده‌های کم‌تری است، اما این داده‌ها باید با دقت بسیار بالا همراه با جزئیات آوایی دقیق، آوانویسی شده باشند و امکان دسترسی به صدای ضبط‌شدهٔ مربوط به هر آوانویسی هم وجود داشته باشد. یک تحلیل‌گر گفتمان برای تجزیه و تحلیل داده‌های گفتاری زبان، نیازمند اطلاعات دقیق دربارهٔ ویژگی‌های بافتی است که گفتار/ مکالمه در آن صورت گرفته است. بنابراین، هدف مطالعه بر چگونگی ساخت پیکرهٔ گفتاری تأثیرگذار است. لیچ و همکاران (۱۹۹۵، به نقل از تامسون[5]، ۲۰۰۴) پنج مرحله را برای ساخت پیکرهٔ گفتاری در نظر می‌گیرند: ضبط، آوانویسی، نمایش/ نمود[6]، کدگذاری (یا حاشیه نویسی)، به‌کارگیری.

1. stylistic annotation
2. direct speech
3. indirect speech
4. lexical annotation
5. Thompson
6. representation

دربارهٔ مرحلهٔ اول باید چند مسئله را در نظر گرفت: ویژگی‌های فنی ضبط صوتی/ تصویری، جمع آوری اطلاعات بافتی و رضایت شرکت‌کنندگان؛ این مرحله را می‌توان «جمع‌آوری داده‌ها» نامید. پس از جمع‌آوری داده‌های گفتاری زبان، فرآیند آوانویسی آغاز می‌شود. مرحلهٔ سوم، تبدیل آوانویسی به شکلی است که برای رایانه قابل خواندن باشد. پس از این مرحله، تحلیل‌گر می‌تواند، به تناسب هدف ساخت پیکره، اطلاعات اضافی از قبیل طبقه‌بندی کنش‌های گفتاری داده‌ها یا مشخص‌کردن طبقهٔ دستوری کلمات را اضافه کند؛ این مرحله «حاشیه‌نویسی» خوانده می‌شود. تامسون (۲۰۰۴) دو مرحلهٔ اخیر را یک مرحله در نظر می‌گیرد و عنوان «نمایش و حاشیه‌نویسی» را برای آن انتخاب می‌کند. وی مرحلهٔ نهایی را «دسترسی»[1] می‌نامد؛ زیرا تأکید روی دسترسی به پیکره است و نه لزوماً به کار بردن آن. نمی‌توان دقیقاً رویکردهای مربوط به کاربرد چنین پیکره‌هایی را مورد بررسی قرار داد. اما می‌توان بررسی کرد که آیا پیکره در دسترس دیگر پژوهشگران قرار می‌گیرد؟ و به چه شکلی در دسترس است؟ بنابراین، تامسون (۲۰۰۴) با در نظر گرفتن تغییرات مذکور، چهار مرحله را برای ساخت پیکرهٔ گفتاری تعریف می‌کند: جمع‌آوری داده‌ها، آوانویسی، نمایش و حاشیه‌نویسی، دسترسی.

مرحلهٔ اول داده‌ها ضبط و جمع‌آوری می‌شوند. ادواردز[2] (۱۹۹۳، نقل از تامسون :۲۰۰۴) سه اصل را برای مرحلهٔ آوانویسی دستی و کامپیوتری در نظر می‌گیرد: طبقه‌بندی‌ها باید مجزا از هم، قابل تشخیص و دقیق باشند، برای پژوهشگران قابل خواندن باشند و برای کامپیوتر باید نظام‌مند و قابل پیش‌بینی باشد. در آوانویسی نکتهٔ اساسی این است که باید نوع آوانویسی مشخص باشد: وابسته به املا[3] (در این حالت یک فرهنگ لغت مرجع در نظر گرفته می‌شود و از استانداردهای آوانویسی آن پیروی می‌شود و هر جا لازم باشد اطلاعات اضافی گنجانده می‌شود)، زبرزنجیری (تکیه، آهنگ کلام و مانند آن نمایش داده شود) یا آوانگاری (استفاده از الفبای جهانی آوانگاری که زبان‌شناسان استفاده می‌کنند؛ همچنین یکی از علائم نشانه‌گذاری آوایی SAMPA است که بیش‌تر مهندسان گفتار استفاده می‌کنند).

می‌توان حتی بیش از یک نوع آوانویسی را در پیکره گنجاند، در این صورت باید انواع آوانویسی مربوط به هر کلمه در ستون‌ها یا ردیف‌های مجزا از هم قرار گیرند. در آوانویسی باید دربارهٔ چگونگی نمایش داده‌های غیرکلامی مانند اطلاعات بافتی، وقفه‌ها و مشخصه‌های فرازبانی نیز

1. access
2. Edwards
3. orthographic

تصمیم‌گیری کرد. به عنوان مثال، وقفه می‌تواند کوتاه، کمی طولانی یا زمان‌بر باشد. هر کدام را می‌توان به شکل‌های گوناگون در آوانویسی لحاظ کرد. باید روشی برای نشانه‌گذاری و حاشیه نویسی اتخاذ کرد که مستقل از هر سیستم خاص کاربری و در هر رایانه‌ای قابل خواندن باشد. قبل از حاشیه‌نویسی داده‌های گفتاری زبان (چه دستی یا ماشینی)، در مقایسه با داده‌های نوشتاری، باید پیش‌پردازش‌های بیشتری روی داده‌ها صورت پذیرد. به عنوان مثال، یکی از پیش پردازش‌های خاص داده‌های گفتاری، تصمیم‌گیری دربارهٔ حذف یا تصحیح تکرار (مانند تکرار یک هجا یا یک کلمه در گفتار) یا عبارت‌های ناتمام است. نکتهٔ مهم این است که آوانویسی باید درست و طبق استانداردهای تعریف‌شده در زبان‌شناسی باشد تا به عنوان درونداد درست مورد استفاده قرار گیرد.

آخرین مرحله از ساخت پیکره، فراهم‌کردن امکان دسترسی دیگران به پیکره است. می‌توان نسخهٔ چاپی پیکره را در اختیار دیگران قرار داد، اما این امر تجزیه و تحلیل‌ها مبتنی بر پیکره را محدود می‌کند. بهتر است نسخهٔ الکترونیکی منتشر شود تا به راحتی بتوان جست‌وجو یا هر نوع تجزیه و تحلیل پیکره‌بنیاد را انجام داد. در نسخهٔ منتشرشده بهتر است غیر از آوانویسی گفتار، مستنداتی مانند گفتار ضبط‌شده و اطلاعات مربوطه نیز وارد شود تا برای کاربر امکان مقایسهٔ آوانویسی با پرونده‌های صوتی/ تصویری فراهم شود؛ به این منظور، برنامه‌های رایانه‌ای/ نرم افزارهایی مانند Transana و Anvil وجود دارد که به تحلیل‌گر این امکان را می‌دهد تا پروندهٔ صوتی/ تصویری را به آوانویسی ارتباط دهد. ساخت پیکرهٔ مبتنی بر داده‌های گفتاری، کاری بسیار پیچیده و مستلزم برنامه‌ریزی دقیق است. باید این اطمینان حاصل شود که همهٔ اطلاعات مربوطه جمع‌آوری شده است و از نظر کیفیت داده‌ها، فرآیند آوانویسی، پیروی از قراردادها و استانداردهای مربوطه، نمایش و حاشیه‌نویسی، یک‌دست هستند. باید دربارهٔ میزان داده‌های گفتاری جمع‌آوری شده و جزئیات حاشیه‌نویسی و میزان واردکردن جزئیات، تصمیم‌گیری شود. در صورت وجود منابع کافی، بهتر است تا حد امکان داده‌های بیش‌تری در ساخت پیکره وارد شوند، امکان نمایش آن‌ها به شکل‌های گوناگون فراهم شود و حتی‌الامکان سعی شود بین پرونده‌های صوتی/ تصویری و آوانویسی مربوطه ارتباط ایجاد شود.

از میان پیکره‌های گفتاری فارسی می‌توان به فارس‌دات[1] و پیکرهٔ گفتار محاوره‌ای زبان فارسی اشاره کرد. دادگان فارس‌دات مجموعه‌ای از عبارات و جملات است که توسط گویندگان فارسی‌زبان از مناطق مختلف کشور بیان شده است. این دادگان در سطح واج (آوا) با دقت میلی‌ثانیه تقطیع و برچسب‌دهی شده و به صورت فایل‌های مجزا ذخیره شده است. این دادگان، به عنوان دادگان استاندارد گفتاری زبان فارسی در داخل و خارج کشور شناخته شده و برای آموزش سیستم‌های هوشمند تشخیص گفتار استفاده می‌شود[2]. پیکرهٔ گفتار محاوره‌ای زبان فارسی ۳۵۰ ساعت دادهٔ گفتاری دارد که به صورت سیگنال آکوستیکی از فارسی‌زبانان داوطلب در موقعیت‌های مختلف ارتباطی با هدف تحقیقات کاربردی و آموزش و آزمون سامانه‌های محاوره‌ای ضبط شده است و در سطح نوبت و پاره‌گفتار نشانه‌گذاری می‌شود. خروجی‌های پیکره از جمله پروندهٔ شرکت‌کنندگان برحسب سن، جنسیت، لهجه، میزان تحصیلات، نوع سیاق، مدت‌زمان گفتمان، پرونده‌های صوتی و شبکهٔ متنی متناظر، پروندهٔ متنی نشانه‌گذاری نوشتاری، پروندهٔ واژگان پیکره و مستندات پیکره را شامل می‌شوند[3].

۶. جمع‌بندی و نتیجه‌گیری

در این پژوهش، ابتدا به شاخص‌های کلی ساخت پیکره‌های زبانی پرداخته شد. این شاخص‌ها ساخت انواع پیکرهٔ متنی و گفتاری را در بر می‌گیرند که نمونه‌گیری، نمایندگی، توازن، اندازه، نوع پیکره و دسترسی را شامل می‌شوند. سپس، فرآیند ساخت پیکرهٔ متنی ارائه شد که شامل انتخاب متون، پیش‌پردازش متون و حاشیه‌نویسی است و در این راستا به تفصیل دربارهٔ هر یک از مراحل توضیح داده شد. در پایان، فرآیند ساخت پیکرهٔ گفتاری بیان شد که جمع‌آوری داده ها، آوانویسی، نمایش و حاشیه‌نویسی و دسترسی را شامل می‌شود. دربارهٔ هر یک از مراحل مذکور نیز به تفصیل توضیح داده شد.

1. FarsDat
2. peykaregan.ir
3. peykaregan.ir

فهرست منابع

آیت، سید سعید. (۱۳۸۹). طراحی و پیاده‌سازی دادگان دایفون زبان فارسی برای کاربرد زبانشناسی رایانه‌ای. *پژوهش‌های زبان‌شناسی*، ۲ (۳)، صص. ۱۱-۱.

بی‌جن‌خان، محمود. (۱۳۸۳). نقش پیکره‌های زبانی در نوشتن دستور زبان: معرفی یک نرم‌افزار رایانه‌ای. *مجلهٔ زبان‌شناسی*. سال نوزدهم، ۲، صص. ۴۸-۶۷.

محمدی، رؤیا. (۱۳۹۱). *ساخت پیکرهٔ تطبیقی فارسی-انگلیسی و استخراج جملات موازی از آن*. پایان نامهٔ کارشناسی ارشد. دانشگاه الزهرا (س). دانشکدهٔ فنی و مهندسی.

Aston, G. & Burnard, L. (1997). *The NBC handbook exploring the British National Corpus with SARA*. Edinburgh University Press.

Atkins, S. Clear, J. Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*. 7 (1). 1-16

Bartsch, S. and Evert, S. (2013). *Corpus linguistics. Exploring the Firthian notion of collocation*. Lancaster. UCREL.

Bianchi, F. (2012). *Culture corpora and semantics*: methodological issues in using elicited and corpus data for cultural comparison. Chapter 3: corpora and corpus linguistics. University of Salento.

Biber, D. (1993). Representativeness in corpus design. *Literaray and linguistic computing*. 8 (4). Oxford University Press.

Bijankhan, M. Sheykhzadegan, J. Bahrani, M. and Ghayoomi, M. (2011). Lesson from building a Persian written corpus: Peykare. *Language resources and evolution. Springer*. Netherland. 45 (2). 143-164.

Cavaglia, G. (2002). Measuring corpus homogeneity using a range of measures for inter-document distance. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. European Language Resources Association (ELRA).

Claude Toriida, M. (2016). Steps for creating specialized corpus and developing an annotated frequence-based vocabulary list. *TESL Canada journal/ revue TESL du Canada*. 34 (11). 87-105.

Durand, J. Gut, U. and Kristoffersen, G. (2014). *The handbook of corpus phonology*. Oxford.

Edwards, J. (1993). *Principles and contrasting systems of discourse transcription. In Talking Data: Transcription and coding in discourse research*. eds. J. Edwards and M. Lampert, 3-32. Hillsdale, NJ: Lawrence Erlbaum Associates

Francis, W. N. and Kucera, H. (1964/1979). *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Department of Linguistics, Brown University.

Leech, G. (2004). *Developing Linguistic Corpora: a Guide to Good Practice*. adding linguistic annotaion. Edited by Martin Wynne .ahds.literature, languages and linguistics. The Oxford Text Archive.

Leech, G., Myers, G., and Thomas, J. eds. (1995). *Spoken English on computer*. Harlow: Longman.

McEnery, T. & Wilson, A. (2001). *Corpus Linguistics: An Introduction*: Edinburgh University Press.

McEnery, T. Xiao, R. and Tono, Y. (2006). *Corpus-based language studies: and advanced resource book*. Routledge. London and New York.

Norling- Christensen, O. (1993). Methods and tools for corpus lexicography. *Proceedings of the 9th Nordic Conference of Computational Linguistics (NODALIDA)*. Stockholm university. Sweden. pp. 187-196

Rea Rizzo, C. (2010). Getting on with corpus compilation: from theory to practice. *ESP World*, 1 (27), 9, pp. 1-23. Spain.

Sharoff, S. (2003). Methods and tools for development of the Russian Reference Corpus. in D. Archer, P. Rayson, A. Wilson and A. McEnery (eds.) *Corpus Linguistics Around the World*. Amsterdam: Rodopi.

Sinclair, J. (2004). Developing Linguistic Corpora: A Guide To Good Practice Corpus and Text–Basic Principles. *Tuscan Word Centre, Available online from http://www. ahds. ac. uk/creating/guides/linguistic-corpora/chapter1. htm.*

Thompson, P. (2004). *Developing Linguistic Corpora: a Guide to Good Practice*. Spoken language corpora. Edited by Martin Wynne .ahds.literature, languages and linguistics. The Oxford Text Archive

Wattam, S. M. (2015). *Technological Advances in Corpus Sampling ethodology*. Mathematics and Statistics School of Computing and Communications Lancaster University.

Waynne, M. (2005). *Developing linguistic corpora: a guide to good practice.* Oxbow books. Literary and linguistic computing. 22 (1).

https://www.peykaregan.ir/dataset

Steps to be followed in corpus construction: written and spoken language corpora

Elham Alayiaboozar

Assistant professor; Information Science Research Department, Tehran, Iran

Abstract

The aim of this paper is to take readers through the basic steps involved in building a corpus of language data for different purposes. This is done via gathering information about corpus construction from related sources. After a review of literature (regarding corpus construction and the use of corpus in different fields) , this article offers advice in a non-technical style to help the researchers to make sure that their corpus is well-designed and fit for the intended purpose. Key points to be considered in constructing any corpus (written or spoken language) include: Sampling, Size, Representativeness, Balance, General vs. Specialized corpus and Homogeneity. The steps involved in constructing a text corpus are: text selection, text normalization and different kinds of annotation. The steps to be followed in constructing a spoken language/speech-based corpus are: data gathering, transcription, representation, annotation and access. In this paper all the afore-mentioned steps have been explained with related details.

Keywords: Corpus, Key Points in Corpus Construction, Corpus Construction Process, Text Corpus, Spoken Language Corpus