

## فهرست مطالب

| صفحه | عنوان                                         |
|------|-----------------------------------------------|
| ۱    | پیشگفتار                                      |
| ۹    | فصل اول: تاریخچه                              |
| ۹    | ۱-۱ مقدمه                                     |
| ۹    | ۱-۲ زبان‌های هند باستان                       |
| ۱۲   | ۱-۳ زبان‌های ایرانی باستان و میانه            |
| ۱۴   | ۱-۴ دوره زبان‌شناسی غرب                       |
| ۳۱   | ۱-۵ مطالعات اسلامی                            |
| ۳۸   | ۱-۶ گویشور بومی                               |
| ۴۲   | ۱-۷ زبان‌شناسی پیکره‌ای                       |
| ۵۳   | فصل دوم: دستگاه خط فارسی                      |
| ۵۳   | ۲-۱ مقدمه                                     |
| ۵۴   | ۲-۲ خط، نگاره‌شناسی و الفبا                   |
| ۶۰   | ۲-۳ نگاشت نگاره به واج                        |
| ۶۳   | ۲-۴ واج‌نگاری خط فارسی                        |
| ۶۵   | ۲-۵ هم‌نگاره و هم‌نگارگی                      |
| ۷۲   | ۲-۶ هم‌آوا و هم‌آوایی                         |
| ۷۳   | ۲-۷ عمق خط                                    |
| ۷۶   | ۲-۸ تنوع نگارشی                               |
| ۷۹   | ۲-۹ توزیع احتمال حروف                         |
| ۸۰   | ۲-۹-۱ توزیع احتمال حروف                       |
| ۸۱   | ۲-۹-۲ توزیع احتمال دوحرفی‌ها و سه‌حرفی‌ها     |
| ۸۴   | ۲-۹-۳ توزیع احتمال حروف آغازین و پایانی کلمات |
| ۸۵   | ۲-۹-۴ توزیع احتمال واج به شرط نگاره/حرف       |
| ۹۰   | ۲-۹-۵ الگوهای واژواجی عربی                    |
| ۹۴   | فصل سوم: متن و پیکره                          |
| ۹۴   | ۳-۱ مقدمه                                     |

| صفحه | عنوان                                         |
|------|-----------------------------------------------|
| ۹۴   | ۳-۲ تعریف پیکره                               |
| ۱۰۰  | ۳-۳ تعریف متن                                 |
| ۱۰۳  | ۳-۳-۱ متن تک وجهی و چندوجهی                   |
| ۱۰۸  | ۳-۳-۲ متن و ارتباط تک رسانه ای و چند رسانه ای |
| ۱۱۰  | ۳-۴ گونه های زبانی                            |
| ۱۱۶  | ۳-۵ انواع پیکره                               |
| ۱۲۶  | ۳-۶ پیکره گفتگوگر                             |
| ۱۲۷  | ۳-۷ تجزیه و تحلیل سیاق                        |
| ۱۳۷  | فصل چهارم: ساختن پیکره                        |
| ۱۳۷  | ۴-۱ مقدمه                                     |
| ۱۳۸  | ۴-۲ تعیین هدف، موضوع و عنوان                  |
| ۱۳۹  | ۴-۳ نمایندگی، توازن و مقایسه پذیری            |
| ۱۴۱  | ۴-۴ نمونه گیری                                |
| ۱۴۴  | ۴-۵ گردآوری داده ها                           |
| ۱۴۴  | ۴-۵-۱ پیکره گفتاری                            |
| ۱۴۷  | ۴-۵-۲ پیکره نوشتاری                           |
| ۱۴۸  | ۴-۶ بهنجار سازی                               |
| ۱۴۸  | ۴-۶-۱ پیکره گفتاری                            |
| ۱۴۹  | ۴-۶-۲ پیکره نوشتاری                           |
| ۱۵۳  | ۴-۷ تقطیع                                     |
| ۱۵۵  | ۴-۷-۱ پیکره گفتاری                            |
| ۱۵۷  | ۴-۷-۲ پیکره نوشتاری                           |
| ۱۵۹  | ۴-۸ غنی سازی                                  |
| ۱۶۰  | ۴-۸-۱ پیکره گفتاری (غنی سازی خاص)             |
| ۱۶۱  | ۴-۸-۲ برچسب زنی اجزای کلام                    |
| ۱۸۲  | ۴-۸-۳ بن واژه سازی                            |
| ۱۸۲  | ۴-۸-۴ تفسیر نحوی: درخت بانک                   |
| ۱۸۸  | ۴-۸-۵ تفسیر معنایی                            |
| ۱۹۱  | ۴-۸-۶ تفسیر گفتمانی                           |
| ۱۹۱  | ۴-۹ ساختارمندی                                |
| ۱۹۴  | ۴-۱۰ انتشار پیکره و مستندات آن                |
| ۱۹۵  | ۴-۱۱ ارزیابی پیکره                            |
| ۲۰۱  | ۴-۱۲ بیانیۀ داده ها                           |

| صفحه | عنوان                                         |
|------|-----------------------------------------------|
| ۲۰۷  | فصل پنجم: زبان‌شناسی پیکره‌ای                 |
| ۲۰۷  | ۵-۱ مقدمه                                     |
| ۲۰۸  | ۵-۲ زبان‌شناسی پیکره‌ای                       |
| ۲۱۴  | ۵-۲-۱ رویکرد پیکره‌بنیاد                      |
| ۲۴۸  | ۵-۲-۲ رویکرد پیکره‌محور                       |
| ۲۵۵  | ۵-۳ پیکره و وب                                |
| ۲۶۳  | ۵-۳-۱ وکی: مگاپیکره/مینی‌وب                   |
| ۲۶۵  | ۵-۳-۲ اسکچ انجین: پیکره‌های ده به ده          |
| ۲۶۹  | فصل ششم: سه نمونه پیکره                       |
| ۲۶۹  | ۶-۱ مقدمه                                     |
| ۲۶۹  | ۶-۲ پیکره گفتاری فارسی‌دات                    |
| ۲۷۰  | ۶-۲-۱ محتوای فارسی‌دات                        |
| ۲۷۳  | ۶-۲-۲ غنی‌سازی                                |
| ۲۷۶  | ۶-۲-۳ تقطیع آوایی                             |
| ۲۷۹  | ۶-۲-۴ تجزیه و تحلیل سیاق                      |
| ۲۸۰  | ۶-۳ پیکره متنی زبان فارسی معاصر               |
| ۲۸۴  | ۶-۳-۱ محتوای پیکره                            |
| ۲۸۵  | ۶-۳-۲ تجزیه و تحلیل سیاق                      |
| ۲۹۸  | ۶-۳-۳ زیرپیکره نشانه‌گذاری شده                |
| ۳۱۴  | ۶-۴ پیکره اچ‌ام‌بلاگز                         |
| ۳۱۸  | فصل هفتم: روش‌شناسی                           |
| ۳۱۸  | ۷-۱ مقدمه                                     |
| ۳۱۸  | ۷-۲ روش‌های تجزیه و تحلیل پیکره               |
| ۳۱۹  | ۷-۲-۱ توزیع فراوانی، تنوع واژگانی و قانون زیف |
| ۳۲۷  | ۷-۲-۲ واژه‌نما                                |
| ۳۳۲  | ۷-۲-۳ هم‌وقوعی                                |
| ۳۵۸  | فصل هشتم: آواشناسی                            |
| ۳۵۸  | ۸-۱ مقدمه                                     |
| ۳۵۹  | ۸-۲ آوا                                       |
| ۳۶۱  | ۸-۳ داده‌های یافتنی و ترازبندی                |
| ۳۶۳  | ۸-۴ دیرش و فضای واکه‌ای                       |

| صفحه | عنوان                                     |
|------|-------------------------------------------|
| ۳۶۸  | ۸۵ واج                                    |
| ۳۷۶  | ۸۶ فرایندهای واجی                         |
| ۳۸۲  | ۸۷ دیرش واج                               |
| ۳۸۴  | ۸۸ ساخت هجا و واج آرایی                   |
| ۳۹۶  | ۸۹ میزان کلمه بودن یک دنباله واجی         |
| ۴۰۲  | فصل نهم: صرف و واژگان                     |
| ۴۰۲  | ۹-۱ مقدمه                                 |
| ۴۰۲  | ۹-۲ قطعه، تکواژ، کلمه و واژه              |
| ۴۱۲  | ۹-۳ تقطیع و واحدهای چندقطعه‌ای            |
| ۴۱۳  | ۹-۳-۱ واحدهای چندقطعه‌ای ایستا            |
| ۴۱۵  | ۹-۳-۲ واحدهای چندقطعه‌ای پویا             |
| ۴۲۴  | ۹-۴ صرف پیکره‌بنیاد                       |
| ۴۲۵  | ۹-۵ پیچیدگی                               |
| ۴۲۶  | ۹-۵-۱ نسبت نوع به قطعه                    |
| ۴۲۸  | ۹-۵-۲ آنتروپی کلمه و بن‌واژه              |
| ۴۳۰  | ۹-۵-۳ اندازه میانگین صیغگان               |
| ۴۳۱  | ۹-۵-۴ ساخت تصریفی (فعل)                   |
| ۴۳۲  | ۹-۶ اشتقاق                                |
| ۴۳۷  | ۹-۶-۱ ساخت پیشوندی                        |
| ۴۴۶  | ۹-۶-۲ ساخت پسوندی                         |
| ۴۶۰  | ۹-۷ ترکیب                                 |
| ۴۶۸  | ۹-۸ زایایی                                |
| ۴۷۷  | ۹-۹ بن‌واژه‌سازی                          |
| ۴۸۳  | ۹-۱۰ واژگان و فرهنگ‌نگاری                 |
| ۴۸۹  | ۹-۱۰-۱ فرهنگ بسامدی                       |
| ۴۹۷  | پیوست: راهنمای نرم‌افزار ابزار زبان‌شناسی |
| ۴۹۷  | بخش ورودی متن برای پردازش                 |
| ۴۹۷  | بخش بهنجارسازی متون                       |
| ۴۹۹  | بخش چندتایی                               |
| ۵۰۱  | بخش واژه‌نما                              |
| ۵۰۳  | بخش گزارش                                 |

| صفحه | عنوان                       |
|------|-----------------------------|
| ۵۰۴  | منابع                       |
| ۵۲۴  | واژه‌نامهٔ فارسی به انگلیسی |
| ۵۳۴  | واژه‌نامهٔ انگلیسی به فارسی |
| ۵۴۴  | نمایه                       |

## پیشگفتار

دانش بشر در حوزه‌های مختلف علوم ناشی از بررسی نظامی است که در تاروپود موجودات و جهان هستی حاکم است. با وجود شناسایی چند هزار زبان مختلف و حدود ۴۰۰ خانوادهٔ زبانی تاکنون (کمبل<sup>۱</sup>، ۲۰۲۴)، نظم حاکم بر شیوهٔ تلفظ آواها، چینش آن‌ها در هجا، چینش تکواژها در کلمه، چینش کلمات در جمله، چینش جملات در گفتمان و معنای حاصل از آن‌ها توجه زبان‌شناسان را به شباهت‌ها و تفاوت‌های میان زبان‌ها جلب کرده است. نظم در مطالعات زبانی با عنوان «ساخت»<sup>۲</sup> مطالعه می‌شود. زبان‌شناسان ساخت‌های زبانی را در پنج سطح بررسی می‌کنند: آوا و واج، کلمه، جمله، گفتمان و معنی. هر کدام از این سطوح را می‌توان با استفاده از عناصر زبانی توصیف کرد و ساخت‌های حاکم بر عناصر زبانی در هر سطح را با استفاده از مقوله‌ها، قواعد، اصول یا محدودیت‌ها صورت‌بندی کرد و دستور زبان را برای هر سطح به دست آورد. از طریق تحلیل دستور زبان در سطح آوا و واج می‌توان به نظام واج‌شناسی زبان پی برد که تلفظ خوش ساخت کلمات و پاره‌گفتارها را پیش‌بینی می‌کند. بررسی دستور زبان در سطح تکواژ، ساخت‌واژه (یا صرف) زبان را آشکار می‌سازد که فرایند کلمه‌سازی و خوش ساختی کلمات را پیش‌بینی می‌کند. تحلیل دستور زبان در سطح عبارت و جمله به شناسایی ساختار نحوی زبان می‌انجامد؛ ساختاری که امکان تبیین جملات خوش ساخت را فراهم می‌سازد. با استفاده از دستور زبان در سطح معنای کلمه، عبارت و جمله می‌توان معنی‌شناسی زبان را به دست آورد که خوش ساختی معنای کلمات، عبارات و جملات را پیش‌بینی می‌کند، و از طریق بررسی دستور زبان در سطح جملات یا پاره‌گفتارهای متن، می‌توان به گفتمان‌شناسی زبان دست یافت که ساخت‌های دستوری را در سطح کلام تبیین می‌کند.

شاید مهم‌ترین مسئله‌ای که زبان‌شناسان با آن مواجه‌اند، میزان خوش ساختی ساخت‌های دستوری باشد که در چارچوب نظریه‌های زبانی برای هر سطح پیش‌بینی می‌شود. واقعیت این است که در موقعیت‌های مختلف ارتباطی گفته‌ها و نوشته‌های بسیاری وجود دارند که نمی‌توان براساس شم زبانی خبرگان یا اهل زبان ساخت دستوری آن‌ها را با معیار دوارزشی

---

1. Campbell

2. structure

خوش ساخت و بدساخت پیش‌بینی کرد. دست کم در حوزه‌ی نحو اساساً غیرممکن است یک توصیف دقیق و کامل از پاره‌گفتارهای خوش ساخت ارائه کرد که آن‌ها را به طور شفاف از سایر دنباله‌های کلمات، که پاره‌گفتارهای بدساخت محسوب می‌شوند، متمایز کرد؛ زیرا اهل زبان برای برآورده کردن نیازهای ارتباطی خود همواره در حال قبض و بسط قواعد دستوری هستند (منینگ<sup>۱</sup> و شوتزه<sup>۲</sup>، ۲۰۰۰: ۳). مسلم است که یافته‌های زبان‌شناسان در چارچوب نظریه‌های زبانی حجم عظیمی از الگوها و ساخت‌های زبانی را به خوبی توصیف و تبیین می‌کند، اما باید بین امکان وجود ساخت‌های دستوری و احتمال وقوع ساخت‌های زبانی که در زبان گفتاری و نوشتاری به کار می‌روند، فرق گذاشت.

تمرکز اصلی در این کتاب بر روی ساخت‌های پربسامد است که در به کارگیری زبان فراوانی وقوع بیشتری دارند و به توصیف و تبیین قضاوت دوارزشی زبان‌شناسان درباره ساخت‌های ممکن دستوری نمی‌پردازد. بنابراین فراوانی مطلق و احتمال وقوع ساخت‌هایی که در زبان گفتاری و نوشتاری به کار می‌روند، کانون توجه محتوای این کتاب است. در نتیجه، مرجع قضاوت درباره ساخت‌های زبانی به جای شم زبانی اهل فن، پیکره زبانی است که از متون معتبر به دست آمده باشد.

رویکرد پیکره‌ای به مطالعات زبانی دست کم چهار مزیت دارد: ۱) داده‌شناسی: استفاده از فناوری‌های نوین رقمی در جمع‌آوری گفته‌ها و نوشته‌های اهل زبان که باعث می‌شود تا ماهیت آوایی و زبانی داده‌های گفتاری و نوشتاری بیش از پیش شناخته شود و نظم غالب و حاکم در الگوها و ساخت‌های زبانی با دقت بیشتری بررسی شود؛ ۲) گونه‌شناسی: گونه‌های زبانی را که ناشی از عوامل مختلفی چون موقعیت ارتباطی، سبک، لهجه، اعتبار اجتماعی و مناطق گویشی هستند، می‌توان در داده‌های پیکره یافت و ساخت‌های آن‌ها را در همه سطوح مطالعه کرد و ساخت‌های پیکره‌ای را با دو رویکرد پیکره‌بنیاد و پیکره‌محور، که به تفصیل توضیح داده می‌شوند، به دست آورد؛ ۳) آموزش زبان: با آگاهی از ساخت‌های پربسامد و مشخصه‌های دستوری در گونه‌های زبانی می‌توان محتوای آموزشی معتبری برای آموزش و آزمون‌سازی زبان طراحی کرد؛ ۴) هوش مصنوعی: امروزه ساختن پیکره‌های بسیار بزرگ برای آموزش و ارزیابی یا محک سامانه‌های پردازش زبان طبیعی یکی از نیازمندی‌های اولیه در فناوری‌های هوش مصنوعی است. اینکه سامانه‌های هوشمند چه نوع داده‌های زبانی را باید

1. Manning

2. Schutze

بیاموزند و مهارت‌های گفتاری و نوشتاری اهل زبان چگونه باید به سامانه‌های رایانشی منتقل شوند، اهمیت چشمگیری یافته است. ارتباط گفتاری و نوشتاری امروزه در زمره توانمندی‌های ابتدایی سامانه‌های هوشمند محسوب می‌شود. بنابراین، اگر یک سامانه هوش مصنوعی نتواند مانند انسان‌ها، در زبان گفتاری و نوشتاری رسمی و محاوره‌ای، به راحتی با کاربران خود ارتباط برقرار کند، نمی‌توان آن را سامانه‌ای موفق دانست.

با توجه به اینکه آمار و احتمال و روش‌های کمی، ابزار اصلی رویکرد کتاب در مطالعه ساخت‌های زبانی است، خواننده ناگزیر به آشنایی با ریاضیات جدید است. اما خواننده نباید اعداد و ارقام موجود در متن کتاب و جدول‌ها را یافته‌هایی در نظر بگیرد که ماهیت ریاضی و عددی دارند، بلکه باید آن‌ها را به صورت کیفی و در راستای آشنایی با ساخت‌های پربسامد زبانی تعبیر و تفسیر کند. فرض بر این است که ساخت‌هایی با فراوانی آماری معنی‌دار نسبت به سایر ساخت‌ها در پیکره زبانی باید از دیدگاه زبان‌شناسی نظری و کاربردی مورد توجه ویژه قرار گیرند. با وجود این، سعی شده است برای وفاداری به ماهیت زبان‌شناختی کتاب از بحث‌های ریاضی و آماری صرف خودداری شود. به همین دلیل، در تجزیه و تحلیل کمی داده‌ها به توصیف بعضی از شاخص‌های مرکزی بسنده شده است.

در فصل اول، تاریخچه مطالعات زبانی از منظر توجه دست‌نویسان با تأکید بر ماهیت داده‌های زبانی بررسی شده است و تا حد امکان بر فراوانی وقوع ساخت‌ها در داده‌های زبانی توجه شده است. بررسی محتوای منابع دستوری در زبان‌های هند باستان، زبان‌های ایرانی باستان و میانه، زبان‌های باستانی و میانه تمدن غرب، مانند یونانی و لاتین در اروپا و همچنین زبان عربی قدیم نشان می‌دهد که احتمالاً اولین بار در دوره مطالعات اسلامی، از قرن هشتم میلادی به بعد، سیبویه به اهمیت داده‌های «زبان بومی» و «گویشور بومی» به عنوان منبعی موثق برای استخراج دستور زبان توجه نشان داد؛ زیرا وی برای نوشتن دستور زبان عربی از گفتار خود به خودی چادر نشینان عرب به عنوان داده‌های تحقیق استفاده کرد. در حالی که قبل از سیبویه، در سرزمین هند، و در غرب از یونان باستان تا قرن سیزدهم برای استخراج دستور زبان از متون ادبیات قدیم و کتب مقدس استفاده شده است. همچنین در این فصل مبانی نظری رویکرد پیکره‌ای در زبان‌شناسی، از نقش‌گرایی فرث در اروپا و تجربه‌گرایی بلومفیلد و نیز روال‌های کشف هریس در آمریکا در نیمه اول قرن بیستم، افول آن در چهار دهه پایانی قرن، و سپس احیای آن در قرن بیست و یکم هم‌زمان با گسترش هوش مصنوعی و کاربرد آن در ساخت پیکره‌های زبانی بزرگ، بحث شده است.

در فصل دوم، دستگاه خط فارسی بر پایه مبانی زبان‌شناسی و نگاره‌شناسی توضیح داده شده است. آگاهی ساختاری از دستگاه خط در رویکرد پیکره‌ای به زبان فارسی بسیار اهمیت دارد؛ زیرا خط فارسی در نوشتن و تایپ متون الکترونیکی از دستور خاصی پیروی می‌کند. همچنین روش رایج در غنی‌سازی پیکره‌های گفتاری، افزودن صورت نوشتاری گفته‌ها به محتوای صوتی گفتار است. در این فصل به این موضوعات هم پرداخته شده است: تمایز حرف و نگاره، بازنمایی یونیکدی برای پردازش متن، نگاشت نگاره به واج، واج‌نگاری خط فارسی، هم‌نگاره‌ها و هم‌آواها، عمق خط، عوامل ایجاد تنوع نگارشی در متون فارسی مانند همزه و <ح>، اعراب، تنوین، تشدید، کسره اضافه، ادغام، گونه‌های آزاد، تکرار و نشانه‌های سجاوندی، توزیع احتمال تک‌حرفی‌ها، دوحرفی‌ها و سه‌حرفی‌ها، توزیع احتمال حروف آغازین و پایانی کلمات، توزیع احتمال واج به شرط نگاره/حرف و الگوهای واژوایی عربی و ارتباط بعضی از آن‌ها با ساخت‌واژه فارسی.

در فصل سوم، «متن» فراتر از مفهوم عامیانه‌اش که صرفاً ماهیت نوشتاری دارد، معرفی شده است. متن و پیکره موجودیت الکترونیکی دارند و در فضای رایانشی باید توصیف شوند. متن و پیکره در چارچوب نظریه ارتباط و انواع متن برحسب وجه و رسانه تبیین می‌شوند. پیکره مجموعه‌ای از متون الکترونیکی است که براساس معیارهای غیرزبان‌شناختی جمع‌آوری می‌شود. انواع پیکره برحسب متغیرهای دوشقی تعریف شده و برای هر کدام دست‌کم یک پیکره فارسی معرفی شده است. ویژگی‌های دستوری هر گونه زبانی در متون پیکره را می‌توان با استفاده از تجزیه و تحلیل سیاق، یعنی رابطه بین موقعیت‌های ارتباطی و مقوله‌های دستوری، به دست آورد.

در فصل چهارم، به مراحل ساخت و ارزیابی پیکره‌های زبانی نوشتاری و گفتاری پرداخته شده است. فعالیت‌های اصلی در طراحی پیکره تا جایی که به خط و زبان فارسی مربوط است، بررسی شده‌اند که عبارت‌اند از: تعیین هدف، موضوع و عنوان؛ نمایندگی، توازن و مقایسه‌پذیری؛ نمونه‌گیری؛ گردآوری داده‌ها؛ بهنجارسازی؛ تقطیع؛ غنی‌سازی؛ ساختارمندی؛ انتشار پیکره و مستندات آن و ارزیابی پیکره. برخی از این فعالیت‌ها مانند گردآوری داده‌ها، بهنجارسازی، تقطیع و غنی‌سازی به تفکیک برای پیکره‌های گفتاری و نوشتاری مورد بحث قرار گرفته‌اند.

محتوای فصل پنجم به چارچوب نظری کتاب، یعنی زبان‌شناسی پیکره‌ای، اختصاص دارد. در این شاخه از دانش زبان‌شناسی، ساخت‌های دستوری گونه‌های زبانی با اتکا به متون

نوشتاری و گفتاری پیکره و با توجه به الگوی آماری‌شان توصیف و تبیین می‌شوند. تحقیقات پیکره‌ای بر روی متون زبانی معمولاً به یکی از دو رویکرد پیکره‌بنیاد و پیکره‌محور تعلق دارد. رویکرد پیکره‌بنیاد به مجموعه‌ای از روش‌ها دلالت دارد که از پیکره عمدتاً به منظور تبیین، آزمایش و مثال برای نظریه‌ها و توصیف‌های دستوری استفاده می‌کند. سه روش پیکره‌بنیاد برای لحاظ کردن ساخت داده‌های پیکره زبانی در یافته‌های دستوری توضیح داده می‌شود که عبارت‌اند از: عایق‌بندی، هنجارسازی و نمونه‌سازی. رویکرد پیکره‌محور به مجموعه‌ای از روش‌ها دلالت دارد که ساخت‌های زبانی را با استفاده از الگوهای مکرر و توزیع فراوانی از پیکره استخراج می‌کند. واژه‌نما در زبان‌شناسی پیکره‌محور یکی از مهم‌ترین ابزارها برای مطالعه الگوهای مکرر است. ساخت‌های زبانی مورد توجه در این رویکرد عبارت‌اند از: باهمایی، هم‌ساختی و باهماساختی.

در فصل ششم، جریان ساخت سه پیکره زبانی، شامل پیکره گفتاری فارسی‌دات، پیکره نوشتاری زبان فارسی معاصر و پیکره نوشتاری اچ‌ام‌بلاگز، به عنوان یک پیکره بزرگ، براساس مراحل ساخت پیکره و تجزیه و تحلیل سیاق توضیح داده شده‌اند.

در فصل هفتم کتاب به رایج‌ترین و پرکاربردترین روش‌های کمی برای تجزیه و تحلیل پیکره پرداخته شده است؛ زیرا استخراج اطلاعات زبانی از یک پیکره با اندازه بسیار زیاد، محقق را ناگزیر به استفاده از روش‌های کمی برای رسیدن به تعمیم‌های زبانی و آماری می‌کند. روش‌های پرکاربرد کمی برای تجزیه و تحلیل متون پیکره شامل توزیع فراوانی، تنوع واژگانی، قانون زیف، واژه‌نما و هم‌وقوعی است. روش هم‌وقوعی برای توصیف و تبیین شش ساخت بنیادین در زبان انگلیسی به صورت گسترده در حوزه نحو، معنی‌شناسی و گفتمان‌شناسی استفاده شده است. این ساخت‌ها عبارت‌اند از: چندتایی، باهمایی، هم‌ساختی، ترجیح معنایی، نوای معنایی و باهماساختی. در عین حال، تأکید شده است همراه شدن ساخت‌های زبانی با مفاهیم ریاضی و آمار از قبیل توزیع فراوانی، احتمال مشروط و امثال آن نباید محقق را از تجزیه و تحلیل یافته‌های زبانی دور کند، بلکه باید زمینه بهره‌گیری از آن‌ها برای کشف شیوه به کارگیری ساخت‌های زبانی در پیکره را فراهم آورد؛ ساخت‌هایی که در مقایسه با ساخت‌های دیگر در پیکره زبانی از نظر آماری به طرز معنی‌داری فراوانی بیشتری دارند، از دیدگاه زبان‌شناسی نظری و کاربردی باید توجه ویژه‌ای به آن‌ها بشود. همچنین راهنمای نرم‌افزار «ابزار زبان‌شناسی» برای استفاده عملی از برخی روش‌های کاربردی‌تر در پیوست کتاب معرفی شده است.

در فصل هشتم کتاب، برخی مباحث بنیادین آواشناسی و واج‌شناسی پیکره‌ای، با عنوان «آواج‌شناسی» تشریح شده است. در بخش آواشناسی پیکره‌ای مفاهیم داده‌های یافتنی و ترازبندی در چارچوب نظریه لیبرمن (۲۰۱۹) توضیح داده شده و از دو پیکره فارسی‌دات با سیاق‌های گفتار خواندنی تک‌گویی و مکالمه تلفنی دونفره برای بررسی ساخت‌های دیرش و فضای واکه‌ای در زبان فارسی استفاده شده و نتایج جالبی در مقایسه با تحقیقات رایج آزمایشگاهی به دست آمده است. در بخش واج‌شناسی پیکره‌ای جفت‌های کمینه برحسب تفاوت واجی در جایگاه ابتدا، وسط و پایان کلمات با استفاده از واژگان زبانی اسلامی و همکاران (۱۳۸۳) آورده شده‌اند. توزیع فراوانی مطلق و نسبی واج‌های زبان فارسی با استفاده از صورت واجی مدخل‌های واژگان زبانی به دست آمده است. توزیع فراوانی فرایندهای واجی زبان فارسی با استفاده از پیکره فارسی‌دات با سیاق گفتار خواندنی تک‌گویی گزارش شده است. همچنین دیرش واج با تعریف عملیاتی «متوسط مدت زمان تلفظ واج گونه‌های هر واج در پیکره آوایی» برای واکه‌ها و همخوان‌های فارسی محاسبه شده که با گزاره‌های واجی موجود در واج‌شناسی شهودی و آزمایشگاهی فارسی قابل مقایسه است. در پایان این فصل، از توزیع فراوانی واج‌ها و مشخصه‌های تمایزدهنده آوایی شواهدی برای یافته‌های واج‌شناسی نظری درباره ساخت هجای فارسی و میزان کلمه بودن یک دنباله واجی به دست آمده است.

در فصل نهم، به برخی مباحث صرف و واژگان زبان فارسی در چارچوب زبان‌شناسی پیکره‌ای پرداخته شده است. ابتدا ساخت کلمات نوشتاری و گفتاری زبان فارسی در چارچوب مفاهیم قطعه، تکواژ، کلمه، واژه، واحد چندقطعه‌ای و قطعه چندواحدی توصیف شده و شواهد متعددی از پیکره عرضه شده است. ساخت درونی کلمات اشتقاقی و مرکب بر پایه توزیع فراوانی بررسی شده و همچنین میزان پیچیدگی و زیایابی آن‌ها با استفاده از پیکره زبانی که در سطح اجزای کلام برچسب‌زنی شده، تجزیه و تحلیل گردیده است. میزان پیچیدگی ساخت‌واژی زبان فارسی، براساس متغیرهای نسبت نوع به قطعه، آنتروپی کلمه و بن‌واژه، اندازه میانگین صیغگان و ساخت تصریفی به دست آمده است. در بخش اشتقاق کلمات، ساخت پیشوندی و پسوندی کلمات فارسی و بازنمایی سلسله‌مراتبی آن‌ها با استفاده از واژگان زبانی و پیکره بن‌واژه‌سازی شده به تفصیل تبیین شده است. همچنین فرایند بن‌واژه‌سازی کلمات نوشتاری و گفتاری به منظور فرهنگ‌نگاری یا استخراج فرهنگ لغت (لغت‌نامه) و فرهنگ بسامدی از پیکره بررسی شده و به مراحل نه‌گانه فرهنگ‌نگاری الکترونیکی اشاره شده است که با مجهز شدن به مدل زبانی بزرگ مناسب برای استخراج معنی مدخل‌ها مبنای مناسبی برای

افزایش دقت و کاهش زمان در دستیابی به یک لغت‌نامه به‌روز خواهد بود.

مخاطبان این کتاب دانشجویان زبان‌شناسی محض و رایانشی، آموزش زبان فارسی به غیرفارسی‌زبانان، زبان و ادبیات فارسی، متخصصان پردازش زبان فارسی و به‌طور کلی همه علاقه‌مندان به خط و زبان فارسی هستند. چون ساخت‌های پیکره‌ای در خط و زبان فارسی موضوع اصلی کتاب است، استفاده از مبانی آمار و احتمال برای استخراج ساخت‌های پربسامد روش اصلی این کتاب به شمار می‌آید. باز هم تأکید می‌شود که مقدار مطلق فراوانی‌ها که در جدول‌های کتاب آمده است نباید معیار قضاوت درخصوص معنی‌دار بودن آن‌ها باشد، کمالینکه اگر اندازه پیکره تحقیق بیشتر شود مقادیر مطلق فراوانی نیز تغییر خواهند کرد. اما می‌توان این فرض را پذیرفت که اگر یک مقوله یا ساخت زبانی در مقایسه با ساخت‌های دیگر در یک پیکره استاندارد فراوانی بیشتری داشته باشد، باید به لحاظ نظری و کاربردی توجه ویژه‌ای به آن کرد. می‌توان این ادعا را نیز مطرح کرد که ساخت‌های پربسامد زبان در زمره عناصر توانشی‌اھل زبان هستند نه صرفاً کنشی. همچنین ساخت‌های پربسامد معمولاً در آموزش و آزمون‌سازی زبان نقش مهمی دارند. در سامانه‌های فناوری زبان نیز هرچه فراوانی یک ساخت در پیکره آموزش بیشتر باشد، یادگیری آن بهتر انجام می‌شود.

امروزه صدها پیکره زبانی با اهداف مختلف برای شناخت و کسب دانش بیشتر از ساخت‌های زبانی و همچنین گسترش و توسعه فناوری‌های زبانی در فضای رایانشی به صورت رایگان یا برای کسب درآمد توزیع می‌شود. پیکره‌های تأثیرگذار در رویکرد پیکره‌ای که عمدتاً برای زبان انگلیسی طراحی و ساخته شده‌اند، در پانویس معرفی شده‌اند و سعی شده است در متن کتاب صرفاً بعضی از پیکره‌ها و منابع زبان فارسی، که برای طرح مباحث در این کتاب اهمیت داشته‌اند، معرفی یا توصیف شوند. در پایان کتاب، واژه‌نامه فارسی به انگلیسی و انگلیسی به فارسی برای بیشتر کلمات و عبارت‌هایی که پانویس شده‌اند، پیوست شده است.

در این کتاب، برای نمایش توزیع فراوانی واحدهای زبانی استفاده از جدول بر هیستوگرام ترجیح داده شده است. در بعضی موارد برای صرفه‌جویی در اضافه کردن جدول مقدار فراوانی در داخل پرانتز آورده شده است. نویسندگان در استنتاج‌های خود از فراوانی مقوله‌ها و ساخت‌های زبانی به‌ندرت از آمار استنباطی استفاده کرده‌اند؛ زیرا اولاً در بسیاری از موارد معنی‌داری در اختلاف نسبی مقادیر فراوانی مطلق از یکدیگر کاملاً مشهود است و ثانیاً استفاده از آمار استنباطی، که خود در موضوع این کتاب بحث مفصلی دارد، به عهده خواننده محقق گذاشته شده است. همچنین به منظور حفظ ابتکار عمل نویسندگان در محاسبات،

به جای بهره‌گیری از نرم‌افزارها یا ابزارهای آماده و رایگان، برنامه‌های رایانه‌ای مربوط به هر فعالیت توسط همکاران نویسنده تهیه شده‌اند تا با توجه به نیاز و تجزیه و تحلیل نتایج در صورت لزوم بتوان برنامه‌ها را اصلاح کرد. تعدادی از این برنامه‌ها با عنوان «ابزارهای زبان‌شناسی»<sup>۱</sup> در سکوی گیت‌هاب<sup>۲</sup> ایجاد شده‌اند تا علاقه‌مندان بتوانند به صورت رایگان از آن استفاده کنند.

در تدوین این کتاب عزیزانی یاریگر نویسنده بودند که شایسته است از آنان تشکر و قدردانی کنم: دکتر مهدی احمدی همچون همیشه، یار و یاور نویسنده در یافتن منابع بودند و در بحث‌های رودررو و تلفنی از استدلال‌ها و نکات ارزشمند ایشان بهره‌مند شدم. آقای عرفان بنیادی کارشناس ارشد زبان‌شناسی رایانشی در دانشگاه تهران زحمت نوشتن بیش از ۹۰ درصد برنامه‌های رایانه‌ای را کشیدند و هر تغییری را که در جریان برنامه‌نویسی درخواست کردم با شکیبایی انجام دادند. آقای داوود حیدرپور کارشناس ارشد زبان‌شناسی رایانشی در دانشگاه تهران نیز در نوشتن برخی از برنامه‌های رایانه‌ای نویسنده را یاری کردند، به‌خصوص در نگاشت بین صورت نوشتاری و واجی مدخل‌های واژگان زیبا برای برنامه توزیع احتمال واج به شرط نگاره/حرف. آقای سینا فرزام کارشناس ارشد زبان‌شناسی رایانشی در دانشگاه صنعتی شریف نیز در استخراج ترتیب کلمات فاعل، مفعول و فعل در محتوای شبکه‌های اجتماعی تلگرام، تویتر، اینستاگرام و بعضی از شبکه‌های خبری نویسنده را یاری دادند؛ اگرچه نتایج آن در این کتاب گزارش نشده است. همچنین قدردان زحمات سرکار خانم مژگان ضرغامی در امر ویرایش محتوای کتاب هستم.

محمود بی‌جن‌خان

استاد دانشگاه تهران

فروردین‌ماه ۱۴۰۴

1. Linguistics Tools

2. <https://github.com/>