فصلنامه علمی زبان‌پژوهی دانشگاه الزهرا(س)
سال شانزدهم، شمارهٔ ۵۲، پاییز ۱۴۰۳
نوع مقاله: پژوهشی
صفحات ۲۴۶-۲۱۹

ساخت پیکره مقایسه‌ای تخصصی «پارسا»[۱]

الهام علایی ابوذر[۲]، علی‌اصغر حجت‌پناه[۳]

تاریخ دریافت: ۱۴۰۲/۰۶/۲۶
تاریخ پذیرش: ۱۴۰۲/۰۹/۱۱

چکیده

پیکره‌ها براساس زبان به‌کاررفته در متن‌های تشکیل‌دهندهٔ آن‌ها به پیکره‌های تک‌زبانه، دوزبانه و چندزبانه گروه‌بندی می‌شوند. پیکرهٔ مقایسه‌ای، پیکره‌ای است دوزبانه یا چندزبانه که شامل متن‌هایی است مشابه در حوزه‌های موضوعی یکسان. با وجود کاربرد فراوان این نوع پیکره‌ها در پژوهش‌های گوناگون همچون پژوهش‌های زبانی، ترجمهٔ ماشینی و سامانه‌های خودکار بازیابی اطلاعات بینازبانی، پژوهشگران همواره با کمبود پیکره‌های مقایسه‌ای مواجه بوده‌اند. در این مقاله، به معرفی مراحل ساخت یک پیکرهٔ مقایسه‌ای تخصصی به نام «پارسا» پراخته شده‌است. این پیکره از چکیده‌های فارسی و انگلیسی پایان‌نامه‌ها و رساله‌های ثبت‌شده در پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) ساخته شده‌است و شامل بیش از ۸۹ میلیون واژه فارسی و ۷۹ میلیون واژه انگلیسی است. محتوای این پیکره عمومی نیست و مشتمل بر متن‌های بسیار تخصصی در حوزه‌های موضوعی کلان مانند علوم اجتماعی، علوم انسانی و هنر، فنی‌ومهندسی و رشته‌های مربوط به این حوزه‌ها است و ازاین‌جنبه، برای پردازش‌های زبانی که نیازمند

---

[۱] شناسهٔ دیجیتال (DOI): 10.22051/jlr.2023.44928.2348

* این مقاله برگرفته از گزارش طرح پژوهشی «ساخت پیکرهٔ مقایسه‌ای تخصصی از چکیده‌های فارسی و انگلیسی پایان‌نامه‌ها و رساله‌های (پارساهای) ثبت شده در پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)» است که با حمایت مالی ایرانداک انجام شده‌است.

[۲] استادیار پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، تهران، ایران (نویسندهٔ مسئول)؛ alayi@irandoc.ac.ir

[۳] رئیس اداره سامانه‌های اطلاعاتی، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، تهران، ایران؛ hojjatpanah@irandoc.ac.ir

بهره‌گرفتن از متن‌های تخصصی است، بسیار ارزشمند است. برای ساخت این پیکره، پس از نمونه‌گیری، داده‌های فارسی وارد فرایند پیش‌پردازش (هنجارسازی و واحدسازی) شدند. برای ارزیابی این مرحله دقت (P)، فراخوان (R) و F1 سنجیده شد. دقت، ۰.۵۶۱۴۰۳۵۰۸۸، فراخوان، ۰.۵۳۱۵۶۱۴۶۲ و در پایان، F1 ۰.۹۷۱۱۶۸۴۳۷ محاسبه شده‌است. سپس، داده‌ها برچسب‌گذاری شدند (برچسب‌گذاری اجزای کلام) و برچسب‌های متون فارسی کنترل شدند. داده‌های انگلیسی نیز به‌صورت ماشینی برچسب‌گذاری شدند. شمار واژه‌های محتوایی (فعل، اسم، صفت، قید) داده‌های فارسی این پیکره ۵۷۶۵۳۸۱۳ و شمار واژه‌های دستوری به‌همراه اعداد و علائم سجاوندی ۳۱۳۵۰۱۲۵ است و بن‌واژه‌های فارسی استخراج‌شده نیز شامل ۴۱۰۶۴ بن‌واژه است. شمار واژه‌های محتوایی متون انگلیسی ۴۵۶۰۶۶۸۶ و شمار واژه‌های دستوری به‌همراه اعداد و علائم سجاوندی شامل ۳۳۶۶۲۳۰۴ و بن‌واژه‌های انگلیسی استخراج‌شده نیز شامل ۱۲۹۳۷ بن‌واژه است. پیکرهٔ ساخته‌شده قابلیت بسیار بالایی برای داده‌کاوی، پژوهش‌های مربوط به ترجمه ماشینی و به‌کارگیری در تمام پژوهش‌هایی که بر روی متون علمی انجام می‌شود را دارا است.

واژه‌های کلیدی: پیکره تخصصی، پیکره مقایسه‌ای، هنجارسازی، واحدسازی، برچسب‌گذاری

۱. مقدمه

امروزه پیدایش فناوری‌های رایانه‌ای و تولید حجم بسیار بزرگی از متن‌ها به زبان‌های گوناگون، منبع‌های پیکره‌ای عظیمی برای پژوهشگران مشتاق به ساخت پیکره فراهم کرده‌است. با افزایش قدرت و ظرفیت رایانه‌ها، پیکره‌ها نیز از نظر اندازه، گوناگونی و آسانی دسترسی، افزایش چشم‌گیری داشته‌اند و همزمان با این دگرگونی‌ها، نرم‌افزارهای بسیاری نیز برای پردازش پیکره‌ها و دسترسی به اطلاعات درون پیکره‌ها، توسعه داده شدند. طی سال‌های اخیر تهیهٔ پیکره‌های زبانی و تجزیه و تحلیل پیکره-بنیاد زبان بسیار مورد توجه پژوهشگران در حوزهٔ زبان‌شناسی، زبان‌شناسی رایانشی و هوش مصنوعی قرار گرفته‌است. گرچه در گذشته، بسیاری از زبان‌شناسان بر اهمیت پیکرهٔ زبانی در بیشتر بررسی‌هایشان تأکید کرده‌اند، ولی در دوران جدید است که تکیه بر داده‌های واقعی زبان به‌صورت گسترده‌ای رواج یافته و شرط اساسی بسیاری از پژوهش‌های نظری و کاربردی مانند نظریه‌پردازی و توصیف ساختمان زبان، گویش‌شناسی، دستورنویسی و فرهنگ‌نگاری به شمار می‌آید.

بحث پیکره‌ها و بهره‌گیری از پیکره‌ها در پژوهش‌ها و کارهای گوناگون، پیشینهٔ طولانی دارد.

امکان سازمان‌دهی، تنظیم، تفکیک، جست‌وجو و دست‌یابی سریع داده‌های زبانی، افق‌های تازه‌ای در برابر پژوهشگران گشوده و سبب پیدایش شاخه‌ای تخصصی در حوزهٔ زبان‌شناسی گردیده‌است. این شاخه با نام زبان‌شناسی پیکره‌ای[1]، تنها در آخرین دهه‌های قرن بیستم ایجاد شده‌است و در همین زمان کوتاه تبدیل به یکی از فعال‌ترین و پرکاربردترین زمینه‌ها شده‌است. ویژگی این رشته این است که به همهٔ حوزه‌های زبان‌شناسی خدمات می‌دهد و درواقع، زبان‌شناسی پیکره‌ای در خدمت همهٔ بررسی‌های زبانی است. همین مسئله دلیل پویایی و گسترش این رشته است. زبان‌شناسی پیکره‌ای به مطالعهٔ فقره‌های واژگانی، ساخت‌های دستوری، یا پیوند این دو با دیگر مشخصه‌های زبانی و غیرزبانی می‌پردازد. درواقع، پژوهش‌های پیکره‌ای بر الگوهای واقعی کاربرد زبان در پایگاه‌های حجیم دادگان یا همان پیکره‌ها تمرکز دارد و رویکردی مکمل برای رویکردهای سنتی‌تر به بررسی زبان در نظر گرفته می‌شود. دراین‌میان، هم از روش‌های کمی و هم از روش‌های کیفی برای تحلیل زبان بهره می‌برد و رایانه را برای انجام تحلیل‌های پیچیده به کار می‌گیرد. آن‌چه زبان‌شناسی پیکره‌ای را به کار زبان می‌آورد، توانمندی آن در بررسی کاربرد واقعی زبان است. همین توانمندی نیز واکاوی پایگاه‌های بزرگ داده‌های زبانی را به زبان‌شناسی پیکره‌ای می‌سپارد (Kouhestani, 2010).

پیکره‌ها برای هدف‌های گوناگونی ساخته می‌شوند که از میان آن‌ها می‌توان به این موارد اشاره کرد: دستورنویسی، فرهنگ‌نگاری، پردازش متن، ترجمهٔ ماشینی، تبدیل متن به گفتار و برعکس، تحلیل گفتمان و دیگر حوزه‌های زبان‌شناسی. پیکره انواع گوناگونی دارد؛ اتکینز و همکاران (Atkins et al., 1992) انواع پیکره را از چند دیدگاه مختلف بررسی کرده‌اند و بر آن اساس پیکره‌ها را به انواع «متن کامل[2]»، «نمونه‌ای[3]»، «نظارتی[4]»، «بسته[5]»، «باز[6]»، «هم‌زمانی[7]»، «درزمانی[8]»، «عمومی[9]»، «اصطلاح‌شناسی[10]»، «یک‌زبانه[11]»، «دوزبانه[12]»، «چندزبانه[13]»، «منفرد[14]»،

---

[1] corpus linguistics
[2] whole text
[3] samples
[4] monitor
[5] open
[6] closed
[7] synchronic
[8] diachronic
[9] general
[10] thesaurus
[11] monolingual
[12] bilingual
[13] multilingual
[14] single

«موازی[1]»، «مرکزی[2]»، «پوسته‌ای[3]»، «هسته‌ای[4]» و «پیرامونی[5]» طبقه‌بندی کرده‌اند.

پیکره‌ها بر اساس زبان به‌کاررفته در متن‌های تشکیل‌دهندهٔ آن‌ها به پیکره‌های تک‌زبانه، دوزبانه و چندزبانه گروه‌بندی می‌شوند. پیکره‌های دوزبانه یا چندزبانه با دو نام شناخته می‌شوند: پیکره‌های مقایسه‌ای/تطبیقی[6] و پیکره‌های موازی[7]. زمان بسیاری به طول انجامید تا واژه‌های تخصصی حوزهٔ زبان‌شناسی پیکره‌ای جا بیفتد، دراین‌میان، نخست، واژهٔ «موازی» به‌جای آن‌چه امروزه «مقایسه‌ای» خوانده می‌شود، به کار می‌رفته‌است. امروزه این دو نوع پیکره را به این صورت تعریف می‌کنند: پیکرهٔ مقایسه‌ای، پیکره‌ای است دوزبانه یا چندزبانه که شامل متن‌هایی است مشابه در حوزه‌های موضوعی یکسان؛ به‌بیان‌دیگر، پیکرهٔ مقایسه‌ای، مجموعه اسنادی در دو زبان متفاوت هستند که موضوع‌های مشابهی را پوشش می‌دهند. درحالی‌که، پیکره‌های موازی مجموعه‌ای از داده‌های برگرفته از دو زبان است که شامل متن‌های اصلی و ترجمهٔ آن‌ها است (Mohammadi, 2023; quoted in Zufferey, 2017). به‌بیان‌دیگر، پیکرهٔ موازی شامل متن‌هایی است که برای هر جمله از آن در یک زبان، ترجمه معادل آن در زبان دیگر آورده شده‌است. درواقع، متن‌های پیکره‌های موازی به منبع مشترکی مربوط می‌شوند، مانند ترجمه‌های فرانسه و آلمانی آثار چارلز دیکنز[8]. در پیکرهٔ مقایسه‌ای، برخلاف پیکرهٔ موازی، متن‌ها لزوماً ترجمهٔ یک‌دیگر نیستند، ولی به یک حوزهٔ یکسان و فرادادهٔ یکسان مربوط می‌شوند. درواقع، آن‌چه مجموعه متون در پیکره‌های مقایسه‌ای را به هم ارتباط می‌دهد، معیارهای مشابهی همانند اندازهٔ متون، موضوع آن‌ها، تاریخ متون، ویژگی‌های متن نگارش‌یافته (مانند ژانر، ملیت نویسنده و موارد مشابه) است. متن‌های مطبوعات، سخنرانی‌های انتخاباتی، آگهی‌های استخدام و دیگر موارد مشابه، موضوع‌هایی هستند که در همهٔ فرهنگ‌ها وجود دارند و برای نگارش آن‌ها معمولاً از قراردادهای مشابهی پیروی می‌شود؛ این منبع‌های مورد علاقهٔ پژوهشگرانی است که اقدام به ساخت پیکره‌های مقایسه‌ای می‌کنند. نمونه‌ای از پیکرهٔ مقایسه‌ای، پیکره‌ای است که از ویکی‌پدیا[9] ساخته شده‌است و به زبان‌های گوناگون است. چنین پیکره‌هایی این امکان را فراهم می‌آورند که بتوان زبان‌های گوناگون یا گونه‌های زبانی مختلف را در بافت مشابه مقایسه کرد و از

---

[1] parallel
[2] central
[3] cluster
[4] nuclear
[5] Perimeter
[6] comparable corpora
[7] parallel corpora
[8] Dikens
[9] Wikipedia

تحریف اجتناب‌ناپذیر که در ترجمهٔ متون در پیکره‌های موازی ایجاد می‌شود، پرهیز کرد.

پیکره‌های مقایسه‌ای می‌تواند از متن‌های عمومی ساخته شود که امکانات گوناگونی را برای تحلیل گفتمان، کاربردشناسی، تجزیه و تحلیل ژانرهای متون و زبان‌شناسی اجتماعی فراهم می‌کند؛ نمونه‌هایی از چنین پیکره‌هایی می‌تواند شامل مجموعه مدخل‌های دایرةالمعارف‌ها یا متون ادبی یک دوره زمانی ویژه باشد. ولی رایج‌ترین گونهٔ پیکره‌های مقایسه‌ای که مخاطبان بسیاری دارد، آن‌هایی هستند که به حوزه(های) تخصصی مربوط می‌شوند و دارای تراکم بالای واژگان و اصطلاح‌های تخصصی هستند. چنین پیکره‌هایی، پیکره مقایسه‌ای تخصصی[1] نامیده می‌شوند. برخی از کاربردهای چنین پیکره‌هایی شامل پژوهش‌های تطبیقی زبان‌ها، داده‌کاوی، موتورهای جستجوی دوزبانه، پژوهش‌های بین‌زبانی، ترجمهٔ ماشینی، فرهنگ‌نگاری محاسباتی[2] و بازیابی اطلاعات است. پیکره‌های بزرگی که شامل متن‌هایی از ژانرهای گوناگون یا گونه‌های زبانی منطقه‌ای هستند یا جفت‌پیکره‌هایی که براساس معیارهای مشابه گردآوری شده‌اند، مانند انگلیسی آمریکایی معیار پیکرهٔ براون[3] یا پیکرهٔ کولهاپور[4] (انگلیسی هندی) نیز می‌توانند پیکرهٔ مقایسه‌ای بسازند (Kenning, 2010). در این مقاله، پس از مرور پژوهش‌های پیشین ساخت پیکره‌های مقایسه‌ای و تخصصی، مرحله‌های ساخت پیکره، شامل نمونه‌گیری، هنجارسازی و واحدسازی داده‌های فارسی و درپایان، برچسب‌گذاری داده‌های فارسی و انگلیسی شرح داده می‌شود. گفتنی است فرایند هنجارسازی و واحدسازی تنها در مورد داده‌های فارسی انجام شده‌است و داده‌های انگلیسی فقط برچسب‌گذاری ماشینی روی آن‌ها انجام شده‌است.

۲. مرور پژوهش‌های پیشین ساخت پیکره‌های مقایسه‌ای و تخصصی

کریمی و همکاران (Karimi et al., 2017) چگونگی استخراج یک پیکرهٔ موازی از پیکرهٔ مقایسه‌ای را شرح می‌دهند. داده‌های موازی بخش مهمی از ترجمهٔ ماشینی را تشکیل می‌دهند؛ هرچه داده‌های بیشتری دردسترس باشد، کیفیت ترجمهٔ ماشینی بهتر خواهد بود. درمورد برخی جفت‌زبان‌ها مانند فارسی-انگلیسی، چنین منابع موازی نایاب است. در این پژوهش، یک روش دوسویه برای استخراج جمله‌های موازی از اسناد ترازبندی‌شدهٔ فارسی-انگلیسی ویکی‌پدیا پیشنهاد شده‌است. در این روش دو سیستم ترجمهٔ ماشینی برای ترجمه از فارسی به انگلیسی و برعکس، به کار گرفته شده‌است. پس از آن، از یک سامانه بازیابی اطلاعات[5] برای اندازه‌گیری شباهت

---

[1] specialized comparable corpus
[2] computational lexicography
[3] Brown copus of standard American English
[4] Kolhapur corpus
[5] information retrieval (IR)

جمله‌های ترجمه‌شده، بهره گرفته شده‌است. افزودن جمله‌های استخراج‌شده به داده‌های آموزشی موجود در سیستم ترجمهٔ ماشینی، سبب بهبود کیفیت ترجمه شده‌است. افزون‌برآن، روش پیشنهادی کمی بهتر از رویکرد یک‌سویه عمل می‌کند. پیکرهٔ استخراج شده تقریباً شامل ۲۰۰ هزار جمله است که براساس درجهٔ شباهت مرتب شده‌اند که سیستم بازیابی اطلاعات این میزان شباهت را اندازه‌گیری کرده‌است.

کولتانسکی (Koltunski, 2013) یک پیکرهٔ مقایسه‌ای ترجمه‌ای را معرفی می‌کند. هدف از ساخت چنین پیکره‌ای، بررسی تفاوت‌های ترجمه برمبنای متغیرهای تفاوت در زبان، نوع متن و روش‌های ترجمه (ماشینی، به کمک ماشین در مقابل ترجمهٔ انسانی است). موارد اشاره‌شده در ویژگی‌های زبانی متن ترجمه‌شده بازنمایی می‌شوند. برای تجزیه‌وتحلیل ترجمه‌های انجام‌شده، تلفیقی از روش‌هایی که در مطالعات ترجمه، تفاوت‌های گونه‌های زبانی و ترجمهٔ ماشینی، با تأکید بر تفاوت‌های متنی و دستوری-واژگانی به کار گرفته می‌شوند، به کار گرفته شده‌است. در این پژوهش، افزون‌بر ساخت پیکرهٔ مقایسه‌ای، بررسی‌های انجام‌شده در زمینهٔ ترجمه در حوزه‌های گوناگون، ازجمله مطالعات ترجمه، ترجمهٔ ماشینی و دیگر حوزه‌های مشابه نیز به کار گرفته شده‌است.

ازجمله تلاش‌هایی که در زمینهٔ ساخت پیکره‌های دوزبانه فارسی- انگلیسی انجام گرفته‌است، می‌توان به پیکره‌های امرایی و همکاران (Emrayi et al. 2019)، دشتبانی و همکاران (Dashtbani et al., 2014)، محمدی (Mohammadi, 2012) و اصغری و همکاران (Asghari et al., 2015) اشاره کرد. امرایی و همکاران (Emrayi et al. 2019) یک فرهنگ دوزبانه فارسی-انگلیسی در حوزه اصطلاح‌های راهنمایی و رانندگی و با تأکید بر نیازهای مترجمان تهیه کرده‌اند. ساختار این فرهنگ بر یک پیکرهٔ مقایسه‌ای دوزبانهٔ متون راهنمایی و رانندگی استوار است که با بهره‌گیری از معناشناسی قالبی برچسب‌گذاری شده‌است. به‌این‌منظور، یک هستان‌شناخت[1] و یک شبکه قالبی برای این حوزه طراحی شده‌است. سپس یک رابط کاربری برای جستجو در این دادگان طراحی شده‌است که امکانات مختلفی شامل جستجوی سنتی الفبایی و جستجوی مبتنی بر معنا را فراهم می‌سازد. این کار به مترجم‌ها، که درواقع گروه مخاطبان هدف این فرهنگ هستند، کمک می‌کند تا با دقت و کارآمدی بیشتری بتوانند به طبیعی‌ترین شیوهٔ بیان مفاهیم موردنظر خود در هر دو زبان دست یابند.

دشتبانی و همکاران (Dashtbani et al., 2014) به معرفی پیکره‌ای دوزبانه در حوزهٔ فاوا

[1] ontology

(حوزهٔ فناوری ارتباطات و اطلاعات) پرداخته‌اند. این پیکره به‌صورت خودکار ساخته شده‌است و منبع‌های آن، اسناد تخصصی حوزهٔ فاوا است. در این پژوهش، نرم‌افزاری برای ساخت پیکره طراحی شده‌است که هزینه و مدت زمان ساخت پیکره را کاهش می‌دهد. افزون‌براین، نرم‌افزار ارائه‌شده قابلیت مدیریت پیکره را برای کاربران فراهم می‌کند. سیستم مدیریت پیکره دارای دو بخش اصلی است که بخش اول مربوط به ساخت پیکره است و بخش دوم مربوط به استخراج اطلاعات از پیکره است. نرم‌افزاری که برای مدیریت پیکره ایجاد شده‌است، حاشیه‌نویسی اسناد، جستجو در پیکره و درست‌نمودن خطا را آسان می‌کند. پیش از شروع فرآیند پردازش اصلی، هر سند توسط نرم‌افزار پیش‌پردازش می‌شود؛ این کار برای انتخاب جمله‌های درست و معنی‌دار برای پردازش اصلی است؛ افزون‌برآن، درصورتی‌که در سند نویسه‌های بی‌معنی وجود داشته باشد، در فرآیند پیش‌پردازش از سند حذف می‌شوند. از جمله کارهایی که این سیستم انجام می‌دهد می‌توان به این موارد اشاره کرد: ویرایش پیکره، افزودن متن‌های جدید به پیکره، اندیس‌گذاری و حاشیه‌نویسی پیکره. بخش دوم، یک موتور جستجوی پیکره است که برای مدیریت مجموعه بزرگی از متون طراحی شده‌است. پردازش متون به کمک سیستمی انجام شد که دارای این بخش‌ها است: طبقه‌بند[1] برای پذیرش اسناد حوزهٔ فاوا، برچسب‌گذار نقش دستوری واژگان (برچسب‌گذاری اجزای کلام) و تجزیه‌کننده[2] برای اسناد فارسی و یک تجزیه‌کننده، برچسب‌گذار نقش دستوری واژگان و ریشه‌یاب[3] برای اسناد انگلیسی است. اسناد حوزهٔ فاوا به کمک این سیستم حاشیه‌نویسی می‌شوند و اطلاعات پردازش‌شدهٔ اسناد در پایگاه داده پیکره ذخیره می‌شوند. مهم‌ترین مرحلهٔ ساخت پیکره‌های چندزبانی، ترازبندی داده‌های پیکره است. در این پروژه روشی برای ترازبندی جمله‌های پیکرهٔ فارسی تخصصی حوزه فاوا و جمله‌های انگلیسی پیکرهٔ تخصصی حوزه فاوا ارائه شده‌است. الگوریتم پیشنهادی آن‌ها از مدل ترجمهٔ واژه‌به‌واژه و تکنیک بلندترین زیردنبالهٔ مشترک[4] برای ترازبندی بهره می‌گیرد و درپایان امتیاز نشان دهندهٔ شباهت دو جمله، سنجیده می‌شود و اطلاعات مربوط به نگاشت جمله‌های دو مجموعهٔ انگلیسی و فارسی در پایگاه دادهٔ پیکره، ذخیره می‌گردد.

محمدی (Mohammadi, 2012) ابتدا ساخت پیکرهٔ مقایسه‌ای فارسی–انگلیسی را شرح می‌دهد. برای ایجاد این پیکره از اسناد خبری روزنامه‌های همشهری و بی.بی.سی بهره گرفته

---

[1] classifier
[2] parser
[3] stemmer
[4] longest common subsequence (LCS)
هدف این روش، مقایسهٔ دو رشته و یافتن شباهت میان آن‌هاست.

شده‌است و از اسناد به‌دست‌آمده، معیارهایی مانند تعداد واژه‌های کلیدی مشترک، اسم‌های خاص یکسان، عنوان‌های مشابه و فاصلهٔ تاریخ انتشار دو خبر استخراج شده‌است. سپس، معیارهای به‌دست‌آمده از مرحلهٔ پیشین، براساس میزان اهمیت آن‌ها در ترازبندی متون، با وزن‌های مختلف با یک‌دیگر ترکیب شده‌اند. در گام پسین، به استخراج جمله‌های موازی از پیکرهٔ مقایسه‌ای ساخته‌شده پرداخته شده‌است. به‌این‌منظور، پس از استخراج متن‌های منطبق با یکدیگر، مجموعه‌ای از جمله‌ها را ایجاد کرده و با استفاده از معیارهای طول و تعداد هم‌پوشانی واژه‌ها، جمله‌هایی را که احتمال موازی بودن آن‌ها بسیار کم بوده‌است، تصفیه شده‌است. پس از تصفیه، به استخراج ویژگی‌های واژگانی، طولی و هم‌پوشانی واژه‌ها از جمله‌های منتخب پرداخته شده‌است و درپایان، با استفاده از جمله‌های آموزشی پیکرهٔ موازی موجود و ویژگی‌های استخراج‌شده، با به‌کارگیری یک طبقه‌بند، جمله‌های منتخب در دو دستهٔ موازی و غیرموازی دسته‌بندی شده‌اند.

اصغری و همکاران (Asghari et al., 2015) یک پیکرهٔ دوزبانهٔ فارسی–انگلیسی تشخیص سرقت ادبی را که ساخته‌اند، معرفی می‌کنند. پیکرهٔ تشخیص سرقت ادبی برای ارزیابی سیستم‌های تشخیص سرقت ادبی به کار گرفته می‌شود. در راستای ساخت پیکره، آن‌ها از یک پیکرهٔ موازی دوزبانهٔ فارسی–انگلیسی که جمله‌های آن ترازبندی شده‌است و از مقاله‌های ویکی‌پدیا استفاده کرده‌اند. جفت‌جمله‌ها در پیکرهٔ موازی امتیاز مشابه بین صفر تا یک دارند. در این پژوهش، اصغری و همکاران از امتیازهای مشابه برای مشخص‌کردن درجهٔ ابهام به منظور ساخت موارد سرقت ادبی بهره گرفته‌اند.

با توجه به کارایی بالای پیکره‌های تخصصی، بسیاری از پژوهشگران اقدام به ساخت چنین پیکره‌هایی کرده‌اند که از میان آن‌ها می‌توان به کلاد توریدا (Claude Toriida, 2016) اشاره کرد. وی ساخت پیکرهٔ تخصصی و تهیهٔ فهرست واژگان حاشیه‌نویسی‌شده و مبتنی بر فراوانی واژگان در پیکره را گام‌به‌گام شرح می‌دهد. وی از پروژهٔ «آموزش زبان انگلیسی برای اهداف دانشگاهی» که در دانشگاهی در خاورمیانه توسعه یافته‌است، نمونه‌هایی را بیان می‌کند. مراحل ساخت پیکرهٔ تخصصی از دید کلود توریدا (همان) شامل انتخاب متون آموزشی، حذف واژگانی که قاموسی نیستند (مانند حروف اضافه، حروف ربط و مانند آن)، تجزیه و تحلیل متن با استفاده از نرم‌افزار انت‌کانک[1]، ایجاد فهرست فراوانی واژه‌ها، توسعهٔ فهرست واژگان حاشیه‌نویسی شده که خود شامل تعیین مقولهٔ نحوی[2] واژگان موجود در فهرست، افزودن تعریف واژگان، باهم‌آیی واژگان و نمونه‌ای از جمله‌ای که واژه در آن به‌کاررفته، است. بلوسو (Beloso, 2015) نیز پیکرهٔ

[1] AntConc
[2] POS

تخصصی سی.ای.دی.سی.ای[1] را معرفی می‌کند. این پیکره شامل مجموعه‌ای از ۵۰۰ هزار واژه زبان نوشتاری از منابع گوناگون است که نمایندهٔ زبان معماری در انگلیسی معاصر است و برای بررسی واژگان این حوزه ساخته شده‌است. این پیکره تک‌زبانه است، برچسب‌گذاری نشده‌است و شامل متن‌های منتشرشده از گونه‌های زبانی انگلیسی آمریکای شمالی، بریتانیا، ایرلندی، کانادایی و استرالیایی است. از آن‌جایی‌که شامل متون منتشرشده در سال‌های اخیر (۲۰۰۸-۲۰۰۷) است، پیکره هم‌زمانی است. این پیکرهٔ تخصصی دارای متن‌هایی در حوزه‌های مربوط به معماری، شامل ساخت‌وساز، شهرسازی، مواد ساخت‌وساز، معماری سبز، طراحی داخلی و دیگر حوزه‌های این رشته است. این پیکره روی نمایندگی، معاصربودن و قابلیت دردسترس‌بودن به‌عنوان اصول مهم ساخت پیکره تأکید دارد.

همچنین درراستای ساخت پیکره‌های تخصصی، علایی ابوذر و همکاران (Alayiaboozar et al., 2021) و علایی و حجت‌پناه (Alayiaboozar, Elham & Ali Asghar Hojjatpanah, 2022) نیز دو پیکره تخصصی ساخته‌اند: پیکره پژوهش‌نامه، که از متن‌های مقاله‌های «پژوهش‌نامه پردازش و مدیریت اطلاعات» ساخته شده‌است، و پکا، که از متن‌های کتاب‌های دیجیتال ایرانداک پدید آمده‌است. هرکدام به‌ترتیب، شامل بیش از چهار میلیون و ۷۸۰ هزار واژه و سه میلیون و ۳۲۹ هزار واژه است. محتوای این پیکره‌ها، متن‌های عمومی نیست، بلکه دارای نوشته‌های بسیار تخصصی و میان‌رشته‌ای مانند علم اطلاعات و دانش‌شناسی، فناوری اطلاعات، مدیریت دانش، زبان‌شناسی رایانشی، مدیریت اطلاعات و مانند آن‌هاست. بنابراین، برای پردازش‌هایی که نیازمند بهره‌گیری از نوشته‌های تخصصی باشند، ارزشمند هستند. همچنین به‌منظور افزایش کارایی، این دو پیکره برچسب‌گذاری شده‌اند (برچسب‌گذاری اجزای کلام)؛ این نوع برچسب‌دهی، عملی کاربردی در بسیاری از حوزه‌های پیشرفته‌تر پردازش زبان طبیعی از جمله ترجمه ماشینی، خطایاب، تبدیل متن‌به‌گفتار، بازیابی اطلاعات، موتورهای جستجو و کمک به مدل‌های آماری است.

۳. مراحل ساخت پیکره پارسا

برای ساخت پیکره مقایسه‌ای تخصصی پارسا مراحل زیر پیموده شده‌است که در شکل (۱) نیز نمایش داده شده‌است:

- نمونه‌گیری
- پیش‌پردازش (هنجارسازی و واحدسازی) داده‌های فارسی
- برچسب‌گذاری اجزای کلام (POS tagging) داده‌های فارسی

[1] Corpus of Architecture Discourse in Contemporary English (CADCE)

- برچسب‌گذاری اجزای کلام (POS tagging) داده‌های انگلیسی
- کنترل درستی برچسب‌های داده‌های فارسی

شکل ۱: مراحل ساخت پیکره مقایسه‌ای تخصصی «پارسا»

۳-۱. نمونه‌گیری

نمونه‌گیری درواقع عمل انتخاب متن‌های مربوط به هر ژانر با توجه به هدف تهیهٔ پیکره است. برخی از معیارهایی که براساس آن‌ها نمونه‌گیری صورت می‌پذیرد شامل این موارد است: شکل متن[1] (گفتاری/ نوشتاری/ الکترونیکی)، نوع متن (کتاب/ مجله/ نامه)، حوزهٔ متن (دانشگاهی (علمی)/ عمومی)، زبان متن (زبان‌ها یا گونه‌های زبانی پیکره) و مکان متن (برای‌نمونه، انگلیسی بریتانیا باشد یا استرالیا) است (Sinclair, 2004). داده‌های پیکره پیش از انجام پژوهش انتخاب شده بودند؛ به‌سبب دسترسی به چکیده‌های فارسی و انگلیسی پایان‌نامه‌ها و رساله‌های (پارساهای) ثبت‌شده در ایرانداک، این متن‌ها برای ساخت پیکره به کار گرفته شدند. متن‌های این چکیده‌ها شامل متن‌های تخصصی یا میان‌رشته‌ای است، بنابراین، حوزهٔ متن دانشگاهی (علمی) است. شکل متون، نوشتاری و به‌صورت الکترونیکی بوده‌است و به‌وسیلهٔ خروجی‌گرفتن از پایگاه مربوطه در ایرانداک تهیه شده‌است. نوع متن‌ها، چکیده‌های فارسی و انگلیسی پایان‌نامه‌ها و رساله‌ها است. حوزهٔ متن‌ها، دانشگاهی (علمی) است، چون شامل پایان‌نامه‌ها و رساله‌هایی است که در دانشگاه‌ها و موسسه‌های پژوهشی به‌وسیلهٔ دانشجویان رشته‌های گوناگون نوشته شده‌است و گونهٔ زبانی متون، گونهٔ نوشتاری و رسمی است.

---

[1] mode

۳-۱-۱. جامعهٔ آماری

جامعه آماری به‌کاررفته در ساخت این پیکره، چکیده پایان‌نامه‌ها و رساله‌های ثبت‌شده در ایرانداک است که از سامانه‌های *ثبت و ویرایش* گرفته شده‌است. ایرانداک کار مدیریت اطلاعات علم و فناوری را از سال ۱۳۴۷ آغاز کرده‌است و افزون‌بر اطلاعات جاری، اطلاعات پیش‌ازآن را نیز سازمان داده‌است. دستاوردهای این کار در سامانه‌ها و پایگاه‌های اطلاعات و منابع مرجع ارائه شده‌اند. همچنین ایرانداک مرکز ثبت و تنها بایگانی ملی اطلاعات پایان‌نامه‌ها، رساله‌ها، و پیشنهاده آن‌هاست و بایگانی پایان‌نامه‌ها و رساله‌های دانش‌آموختگان ایرانی خارج از کشور نیز در ایرانداک ثبت شده‌است، به‌این‌ترتیب، این مرجع منبع سرشاری برای تولید پیکره‌های تخصصی است. برای ساخت این پیکره ابتدا از سامانه ثبت[1] خروجی گرفته شد. دانشجویان پس از تصویب پیشنهاده، به سامانه ملی ثبت پایان‌نامه، رساله، و پیشنهاده (سامانه ثبت) مراجعه و اطلاعات پیشنهاده خود را ثبت و شناسه ره‌گیری دریافت می‌کنند. پایان‌نامه‌ها و رساله‌ها نیز پس از ثبت و بارگذاری فایل تمام‌متن و تأیید ایرانداک و دانشگاه مربوطه، ثبت نهایی می‌شوند. در سامانه ثبت حوزه‌های موضوعی کلانی وجود دارد که رشته‌های گوناگون در آن‌ها گنجانده شده‌است. این حوزه‌ها شامل هفت حوزه موضوعی است: علوم انسانی، فنی‌ومهندسی، علوم پایه، کشاورزی، هنر، علوم پزشکی و دامپزشکی. هنگام نگارش گزارش ساخت این پیکره، آماری از تعداد پایان‌نامه‌ها و رساله‌های ثبت‌شده در سامانهٔ ثبت از معاونت مربوطه در ایرانداک که کار ثبت پایان‌نامه‌ها و رساله‌ها را انجام می‌دهند، دریافت شد که در جدول (۱) ارائه شده‌است.

جدول ۱: گزارش سامانه ثبت پایان‌نامه‌ها و رساله‌ها از تاریخ ۱۳۸۷/۱۰/۱ تا ۱۴۰۱/۱۲/۲۹

حوزه‌های موضوعی	پارسای (پایان‌نامه و رساله)داخل کشور
علوم انسانی	۳۱۷۰۵۰
فنی و مهندسی	۱۵۱۰۴۳
علوم پایه	۱۰۱۰۸۱
کشاورزی	۵۶۵۸۴
هنر	۳۲۹۴۲
علوم پزشکی	۱۳۴۵۹
دامپزشکی	۳۴۰۴
جمع	۶۷۵۵۶۳

[1] https://sabt.irandoc.ac.ir/Home/AboutUs

۳-۱-۲. نمونه‌گیری از جامعه آماری

برای نمونه‌گیری از سامانه ثبت پایان‌نامه‌ها و رساله‌ها، ابتدا باید پرونده‌هایی انتخاب می‌شدند که افزون‌بر چکیده فارسی، چکیده انگلیسی نیز داشته باشند. گفتنی است که پرونده چکیده‌ها به‌صورت جداگانه در سامانه بارگذاری می‌شود و نیازی به جداکردن بخش چکیده از کل پایان‌نامه یا رساله نیست. نمونه‌گیری از سامانه ثبت به‌صورت ماشینی و خودکار و از طریق اتصال به منبع داده، تهیه رونوشت[1] از داده‌ها و بارگذاری در فایل اکسل انجام شد. بررسی اولیه خروجی نشان داد در برخی موارد دانشجویان هنگام ثبت حتی در چکیده نیز کل پایان‌نامه یا رساله را بارگذاری کرده‌اند که بی‌گمان در این پژوهش نمی‌تواند به کار گرفته شود. براساس این بررسی، تصمیم گرفته شد فیلتر تعداد واژه موجود در هر پرونده برای نمونه‌گیری قرار داده شود، به‌این‌صورت که پرونده‌هایی به‌صورت ماشینی انتخاب شوند که دست‌کم تعداد واژه‌های آن ۲۰۰ و حداکثر ۲۰۰۰ باشد. پس از دریافت خروجی، واژه‌های فارسی در پرونده اکسل قرار داده شد تا بتوان وضعیت املایی/نگارشی واژه‌های موجود در خروجی را بررسی کرد. این پرونده دارای مشکلات نگارشی بسیاری بود که به نظر می‌رسید ویرایش آن امکان‌پذیر نباشد. در مرحله بعد تصمیم بر آن شد تا خروجی از سامانه ویرایش دریافت شود. سامانه ویرایش، سامانه‌ای است که در داخل ایرانداک به کار می‌رود. اطلاعات پایان‌نامه‌ها و رساله‌ها (پارساها) ثبت‌شده دانشجویان پس از ذخیره‌سازی در سامانه ثبت، به‌صورت خودکار و سیستمی وارد سامانه ویرایش می‌شود. اطلاعات این سامانه توسط کارشناسان مدیریت سازماندهی و تحویل اطلاعات، فهرست‌نویسی و نمایه‌سازی و ویرایش می‌شود و درپایان ازطریق سامانه ویرایش، اطلاعات وارد پایگاه اطلاعات علمی ایران (گنج)[2] می‌شود و در اختیار کاربران قرار می‌گیرد.

حوزه‌های موضوعی پایان‌نامه‌ها و رساله‌ها در این سامانه کمی متفاوت از سامانه *ثبت* است. در این پایگاه حوزه‌های موضوعی براساس وب‌گاه وب‌آوساینس[3] مشخص شده‌اند. در این وب‌گاه سه حوزه موضوعی کلان وجود دارد: علوم اجتماعی، علوم انسانی و هنر، فنی‌ومهندسی. همه رشته‌های تحصیلی (حدود ۲۸۰ رشته) در این سه حوزه کلان قرار داده شده‌اند. برای نمونه‌گیری از سامانه ویرایش نیز همان دو فیلتر مورداشاره موردنظر قرار گرفت: پرونده‌ها دارای معادل چکیده انگلیسی باشند و دست‌کم تعداد واژه ۲۰۰ و حداکثر ۲۰۰۰ برای خروجی گرفتن موردنظر قرار گیرد. پس از کاربرد فیلتر حدوداً ۲۹۵ هزار پرونده چکیده انتخاب شدند که معادل انگلیسی نیز

---

[1] copying
[2] https://www.Ganj.irandoc.ac.ir
[3] web of science

داشتند (در مجموع حدوداً ۸۹ میلیون واژه فارسی و ۷۹ میلیون واژه انگلیسی).

۳-۲. هنجارسازی و واحدسازی داده‌ها

پیش از واردکردن داده‌ها در پیکره لازم است پیش‌پردازش‌هایی روی متون انجام پذیرد. اصولاً پیش‌پردازش دو مرحله دارد: هنجارسازی[1] و واحدسازی[2]. «هنجارسازی» خود شامل چندین مرحله است: یکدست‌سازی رمزگذاری حروف، یکدست‌سازی تنوع نگارشی، حذف شکل‌ها و جدول‌ها، حذف کدها و علامت‌های اضافی (Ghayoomi, 2022). برای هنجارسازی یک متن ابتدا باید همهٔ نویسه‌های متن با جایگزینی با معادل استاندارد آن، یکسان‌سازی گردند (مانند یکدست‌کردن انواع «ی» و «ک» و «کسره اضافه روی «ه» در حالت اضافی («ۀ»)). همچنین اصلاح‌های دیگری نیز برای پردازش دقیق‌تر متون در این مرحله انجام می‌پذیرد. برای رسیدن به این هدف، پیش از مقایسهٔ متون، پیش‌پردازش‌هایی روی آن‌ها انجام می‌شود. هنجارسازی در متون فارسی می‌تواند شامل این موارد باشد: یک‌دست‌کردن فاصله‌ها و نشانه‌گذاری‌های درون متن، یکسان‌کردن یونیکد نویسه‌های استفاده‌شده در متن‌ها، یکسان‌کردن روش اتصال وندهای گوناگون به ستاک، اصلاح غلط‌های املایی، ارتباط‌دادن واژه‌های چندواژه‌ای و یکسان در نظر گرفتن آن‌ها و مانند آن. در فرایند پردازش داده‌های زبانی، معمولاً فاصله کامل به‌عنوان مرزنمای تشخیص واژه شناخته می‌شود. ولی تشخیص درست واژه در خط فارسی و عربی با چالش پراهمیت چندواژگی روبه‌رو است. تشخیص‌ندادن واحدهای واژگانی به تأثیرگذاری بر همهٔ لایه‌های پردازشی می‌انجامد. به فرایند تشخیص یک واحد واژگانی، واحدسازی می‌گویند (Ghayoomi, 2022). روی‌هم‌رفته، در پردازش شیوهٔ نگارش زبان فارسی، با توجه به شباهتی که با خط عربی دارد، همواره در پردازش برخی از نگاره‌ها مشکلاتی وجود دارد که در نخستین گام باید مشکلات مربوط به این نگاره‌ها را از بین برد. افزون‌براین، اصلاح و یکسان‌سازی نویسه‌ی نیم‌فاصله و فاصله در کاربردهای گوناگون آن و همچنین حذف نویسه‌ی «ـ» که برای کشش نویسه‌های چسبان به کار گرفته می‌شود و مواردی مشابه برای یکسان‌سازی متون، از اقدام‌های لازم پیش از شروع مرحله‌های مختلف است.

از دیگر تفاوت‌های نظام نوشتاری فارسی در مقایسه با انگلیسی، چگونگی پیوستن وندها به ستاک است. در زبان فارسی بسیاری از وندها به ستاک پیوسته می‌شوند و چگونگی پیوستن وندها به ستاک، می‌تواند به‌صورت بافاصله، نیم‌فاصله یا حتی بدون فاصله باشد. برای‌نمونه، هر سه حالت

---

[1] normalization
[2] tokenization

«کتاب‌ها»، «کتاب ها» و «کتابها»، صورت نوشتاری درستی در نظر گرفته می‌شوند. در پردازش متون، چنان‌چه حروف، نشانه‌های نگارشی و واژه‌ها به شکل یکسانی نوشته نشده باشند، پردازش متون به‌درستی انجام نخواهد شد و در بازیابی اطلاعات به یافته‌های درستی نخواهیم رسید.

قیومی و همکاران (Ghayoomi et al., 2010) بر این باورند برای صرفه‌جویی در وقت و انرژی، داده‌های خام هم به‌صورت خودکار و هم به‌صورت دستی پیش‌پردازش شوند. انجام بسیاری از عملیات خودکار برروی زبان مانند ترجمه، خلاصه‌سازی، تصحیح املا و مانند آن، مستلزم استفاده از مجموعه‌ای از ابزارها برای پیش‌پردازش و آماده‌سازی متون است. تهیه این ابزارها به دو صورت انجام می‌شود: دسته اول روش‌های وابسته به زبان هستند که براساس برخی قواعد نحوی و ساختاری زبان انجام می‌شوند. روش‌های دیگر مستقل از زبان هستند و بیشتر براساس پیکره‌های زبانی و با استفاده روش‌های یادگیری ماشینی انجام می‌گیرد. البته در برخی موارد ترکیبی از هر دو روش مورد استفاده قرار می‌گیرد. ازاین‌جهت، طراحی و پیاده‌سازی این ابزارها برای زبان‌های مختلف به گونه‌های مختلف و ویژهٔ زبان مربوطه انجام می‌گیرد.

در پژوهش حاضر از میان ابزارهای موجود برای زبان فارسی (مانند «پارسی‌پرداز»، «هضم»، «پرژن‌پی[1]»، «پارسی‌وار»، « ویراستیار»، « نگار»، « وارسیگر وفا»، «ویراسباز»، « به‌نویس» « پرشین یوتیلز[2]») از مجموعه ابزارهای موجود در کتابخانه «هضم» بهره گرفته شد (Kokabi et al., 2023). هضم، یک کتابخانه است که به‌عنوان یک مجموعه ابزار پردازشی پایه به زبان‌های پایتون، سی‌شارپ و جاوا نوشته شده‌است. فعالیت‌های تعریف شده در این کتابخانه عبارت است از هنجارسازی داده‌ها، واحدسازی در سطح واژه و جمله، بن‌واژه‌سازی، برچسب‌دهی مقوله‌های دستوری، تجزیه سطحی نحوی و تجزیه نحوی وابستگی. از این ابزار پیش از این در پژوهش‌های دیگر استفاده شده بود و دردسترس بود. از کدهای ابزار «هضم»، کد واحدسازی در پیکره استفاده شده‌است. ابزار واحدسازی، مرز واژه‌ها را در متون تشخیص می‌دهد و متن را به دنباله‌ای از واژه‌ها تبدیل می‌کند و آن را برای تحلیل‌های بعدی آماده می‌کند. درواقع، واحدسازی، تکه‌تکه‌کردن متن به قسمت‌های کوچکی به نام واحد است. واحدسازی در سطح واژه رخ می‌دهد و واحدهای استخراج شده به‌عنوان ورودی پیمانه‌های دیگر مانند ریشه‌یاب، برچسب‌گذاری و مانند آن به کار گرفته می‌شوند.

برای هنجارسازی، برخی موارد به‌صورت خودکار توسط ابزار «هضم» انجام شد. این موارد عبارتند از حروف عربی «ی» و «ک» و برخی موارد مانند نیم‌فاصله که با دو یونیکد مختلف بودند

---

[1] Persianp
[2] Persianutils

و حرف «ه» به‌صورت کدهای افزوده در مرحله هنجارسازی به‌کار گرفته شدند. برای یکدست‌کردن انواع «ی» و «ک» و «کسره اضافه روی «ه»/ «ه» در حالت اضافی («هٔ»)، از دستورهای TSQL در برنامه پایگاه داده SQL Server نیز استفاده شده‌است.

خروجی از داده‌های فارسی (چکیده‌های فارسی) گرفته شد که از هر واژه یک نمونه در پرونده اکسل قرار داده شده بود. به‌این‌معناکه برای‌نمونه صورت نوشتاری «اطلاع‌رسانی» که به‌عنوان یک نمونه در پرونده اکسل آورده شده‌است، ممکن است خود ۱۰۰ هزار بار در داده‌های فارسی تکرار شده باشد و اگر در این پرونده اصلاح شود (به‌صورت «اطلاع‌رسانی»)، در همه آن ۱۰۰ هزار بار که رخ داده‌است اصلاح خواهد شد و صورت اصلاح‌شده جایگزین صورت نادرست خواهد شد. تعداد واژه‌های این پرونده ۸۷۹۳۳۸ واژه بودند. بررسی داده‌ها نشان داد که باوجود اصلاح‌هایی که در مورد نیم‌فاصله و جایگزینی یونکدهای عربی با یونیکدهای فارسی انجام شده بود، کماکان متن‌های چکیده‌ها نیازمند اصلاحات نگارشی اساسی بود. به‌این‌معنا که در بسیاری موارد واژه‌های یک جمله کاملاً به هم چسبیده بودند. متأسفانه امکان جداکردن واژه‌های به‌هم‌چسبیده متون به‌صورت ماشینی وجود نداشت؛ درواقع، ابزاری وجود ندارد که بتوان واژه‌های به هم چسبیده یک متن را از هم جدا کرد. موارد بسیاری مانند سطر زیر در خروجی گرفته‌شده از سامانه ویرایش وجود داشت که بیشتر واژه‌ها ناخوانا بودند و روشن نبود مرز واژه‌ها کجاست:

۱. روشانتخابنمونهبصورتتصادفیانتخابگردیدوسپسازبیندانشآموزانایندبیرستانانتخابشدند دوبهدوگروهآزمایشکنترلتقسیمشدند

در این مرحله کوشش شد تا اندازهٔ ممکن اصلاح‌های نگارشی به‌صورت دستی یا به‌صورت دستورهای جایگزینی یک صورت نگارشی با صورتی دیگر و استفاده از گزینه «replace» انجام شود تا برچسب‌گذاری داده‌ها در مرحله بعد با صحت بالاتری انجام شود. برای‌نمونه، واژه‌ای مانند «اطلاعات» بدون فاصله در ابتدا و انتهای واژه در پرونده موجود بود، دستور داده‌شده این بود که این صورت نوشتاری را با واژهٔ «اطلاعات» همراه با یک فاصله در ابتدا و یک فاصله در انتهای واژه در تمام پرونده جایگزین کند. با انجام این دستور مشکلاتی نیز ایجاد می‌شود؛ اگر واژه دارای وندهای تصریفی مانند «ی»، «کسره اضافه به شکل «ی»، «ای»، «ها»، «های»، «تر/ترین» و مانند آن باشد، آن وند نیز از واژه جدا می‌شود. برای ازبین‌بردن این مشکل از دستور دیگری استفاده شد، مانند اینکه اگر « اطلاعات ی» در متن وجود دارد، آن را تبدیل به «اطلاعاتی» کند. دربرخی‌موارد، این دستورها مشکلاتی دیگر را ایجاد کرده بود که اصلاح ماشینی آن امکان‌پذیر نبود و در صورت برخورد با مشکل نگارشی برآمده از به‌کارگیری دستور اصلاحی در متن، باید واژه‌ها

تک‌تک یا گروهی اصلاح می‌شدند. بهره‌گیری از دستورهای داده‌شده در بسیاری موارد موفقیت‌آمیز بوده است، مانند نمونه‌های جدول (۲) که ابتدا صورت نوشتاری موجود در پرونده آورده شده‌است و پس از آن صورت‌های اصلاح‌شده آورده شده‌است:

جدول ۲: نمونه‌ای از اصلاح ماشینی عبارت‌های به‌هم‌چسبیده

نمونه‌ای از صورت‌های نوشتاری موجود در پرونده که نیاز به اصلاح نگارشی داشتند	صورت‌های اصلاح‌شده
توقفاقدامات	توقف اقدامات
توقف تمرینات	توقف تمرینات
ساختارورفتارهای	ساختار و رفتارهای
مساحتپهنه‌هایمتفاوتخطراتموردمطالعهمشخصگردید	مساحت پهنه‌های متفاوت خطرات مورد مطالعه مشخص گردید
برایسنجش	برای سنجش
برایسنجشارتباطبیندومتغیرازضریبهمبستگیپیرسونورگرسیونگامبهگام استفاده می‌شود	برای سنجش ارتباط بین دو متغیراز ضریب همبستگی پیرسون و رگرسیون گام‌به‌گام استفاده می‌شود

افزون‌بر اصلاح ماشینی و پس از آن، کنترل دستی به‌صورتی که شرح داده شد، از فهرست وندها و واژه‌بست‌های فارسی نیز استفاده شد. این فهرست برگرفته از مجموعه ۱۲ مقالهٔ دکتر علی‌اشرف صادقی با نام «*شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر* ۱ تا ۱۲ (Sadeghi, 1991-1993)» و خسرو کشانی (Keshani,1992) (در بحث اشتقاق) و ژیلبر لازار (Lazard, 2010)، و فریبا قطره (Ghatre, 2007) (در بحث تصریف) است. پیش از بهره‌گیری از اطلاعات ساخت‌واژی موجود در این منبع‌ها، دوباره میزان بهره‌گیری از اطلاعات موجود در این منابع بررسی شد. این بررسی نشان داد که همهٔ اطلاعات ساخت‌واژی موجود در این منابع مورد نیاز این مرحله نیست، چون در بسیاری موارد نگارش واژه تنها به یک صورت در پرونده خروجی دریافت شده وجود داشت و تنوع نگارشی وجود نداشت که براساس آن بخواهیم یکدست‌سازی را انجام دهیم (برای‌نمونه، واژهٔ «دانشگاه» تنها یک صورت نگارشی دارد). بنابراین، برخی واژه‌ها که کاربرد نیم‌فاصله در نگارش آن‌ها رعایت نشده بود باید اصلاح می‌شدند. در نتیجه جدول (۳) از فهرست وندها در منابع اشاره‌شده استخراج شد و اطلاعات موجود در این جدول برای پیش‌پردازش متون پیکره به کار گرفته شد.

جدول ۳: فهرست پیشوندها و پسوندها

پیشوندهایی که باید با نیم‌فاصله به واژهٔ بعد از خود متصل شوند	مثال
می- / نمی-/ بی-/ فرا-/ فرو-	می گفت/ می توان/ نمی گفت/نمی توان/ بی‌بنیه/ بی‌ریشه/ فراگرفت/ فرورفت
پسوندهایی که باید با نیم‌فاصله به واژهٔ قبل از خود متصل شوند	**مثال**
-ها/ -های/ -هایی/ -تر/ -ترین/ -ام/ -ای/ -ایم/ -اید/ -اند/ -مان/ -تان/ -شان/ -ستان/ -کده/ -دان/ -دانی/ -زار/ -زاری/ -لاخ/ -لاخی/ -بان/ -گر/ -گری/ سازی/ -باف/ -بافی/ -پز/ -پزی/ -گان/ -واره/ -بندی/ -مند/ -مندی/ -ناک/ -دار/ -داری/ -نده/ -گار/ -ند/ -گانه/ -وند/ -گین/ -گون/ -فام/ -آسا/ -آسایی/ -سان/ -سانان/ -وار/ -وش/ -انه/ -باره/ -آگین/ -گرا/ -گرایی/ -پرست/ -پرستی/ -سنج/ -سنجی/ -پذیر/ -پذیری/ -شناس/ -شناسی/ -نشین/ -گیر/ -گیری/ -خوار/ -خواری/ -پوش/ -خوان/ -خوانی/ -آور/ -آوری/ -جو/-جویی/ -انداز/ -کن/ -آمیز/ -آمیزی/ -کوب -کوبی/ -انگیز/ -انگیزی/ -نویس/ -نویسی/ -پرور/ -پروری/ -شکن/-شکنی/ -افکن/ -افکنی/ -افزار/ -افزاری/ -گداز/ -گدازی/ -نواز/ -نوازی/ -افشان/ -کش/ گذار/ -گذاری/ -یاب/ -یابی/ -اندیش/ -زا/ -گشا/ -گشایی/ -رس/ -رسی/ -گستر/ -گستری/ -نورد/ -نوردی/ -پسند/ -پسندی/ -افزا/ -افزایی/ -آموز/ -آموزی/ -پاش/ -خند/ -تاز/ -رسان/ -رسانی/ -فرسا/- فرسایی/ -ورز/ -اندوز/ -اندوزی/ -زدا/ -زدایی/ -گسار/ -آزما/ -آزمایی/ -توز/ -توزی/ -گوار/ -گواری/ -آشام/ -وار/ -واری/ -دهی	کتاب‌ها/ کتاب‌های/ کتاب‌هایی/ مومن‌تر/ خوب‌ترین/ زنده‌ام/ خورده‌ام/ زنده‌ای/ خواننده‌ای/ زنده‌ایم/ خسته‌ایم/ زنده‌اید/ رفته‌اید/ رفته‌اند/ خوانده‌اند/ طرحمان/ پایگاهمان/ پایگاه‌مان/ لباستان/ روسری‌تان/ دانشگاه‌تان/ لباسشان/ دانشگاه‌شان/ کردستان/ دانشکده/ نمکدان/ سگ‌دانی/ریاضیدان/ کشتزار / سنگ‌لاخ/ نگهبان/ آهنگر/ وحشی گری/ آهنگ‌سازی/ فرش‌باف/ شیرینی‌پزی/ ناوگان/ جشنواره/ طبقه‌بندی/ ثروتمند/ اندوهناک/ کتابدار/ کتابداری/ فرساینده/ خواستگار/ خوشایند/ پنج‌گانه/ شهروند/ غمگین/ گندم‌گون/ گل‌فام/ رعدآسا/ گربه‌سانان/ دیوانه‌وار/ مهوش/ دلیرانه/ شکمباره/ زهرآگین/ سنت‌گرا/ خداپرست/ فشارسنج/ امکان‌پذیر/ هواشناس/ کارگرنشین/ دندانگیر/ گوشتخوار/ گیاه‌خواری/ ژنده‌پوش/ آوازه‌خوان/ سودآور/ جمع‌آوری/ حقیقت‌جو/ چشم‌انداز/ خردکن/ تمسخرآمیز/ دیوارکوب/ رقت‌انگیز/ خوشنویس/ شهیدپرور/ دندان‌شکن/ مردافکن/ نرم‌افزار/ نرم‌افزاری/ جانگداز/ گوش‌نواز/ مهمان‌نوازی/ زرافشان/ دخترکش/ خدمتگذار/ برچسب‌گذاری/ طلایاب/ دستیابی/ دوراندیش/ بیماری‌زا/ مشکل‌گشا/ فریادرس/ دادرسی/ مهرگستر/ دادگستری/ دریانورد/ مشکل‌پسند/ روح‌افزا/ دست‌آموز/ نمک‌پاش/ پوزخند/ یکه‌تاز/ پیام‌رسان/ اطلاع‌رسانی/ طاقت‌فرسا/ دادورز/ مال‌اندوز/ ثروت‌اندوزی/ گندزدا/ غم‌گسار/ بخت‌آزما/ کینه‌توز/ خوش‌گوار/ خون‌آشام/ دیوانه‌وار/ سازمان‌دهی

برنامه‌ای در پایتون نوشته شد که براساس آن اگر وندها جدا از ستاک باشند، با نیم‌فاصله به ستاک متصل شوند. برای‌نمونه، هر جا تکواژ «–ها»، «–های» و «–هایی» جدا از تکواژ پیشین باشد و یک واژهٔ جدا در نظر گرفته شده باشد، با نیم‌فاصله به واژه/تکواژ قبل از آن متصل گردد، یا هر جا تکواژ «–ی» و «–ای» جدا از تکواژ پیشین باشد و یک واژهٔ جدا در نظر گرفته شده باشد، با نیم‌فاصله به واژه/تکواژ قبل از آن متصل گردد. این روش ماشینی متفاوت از روش قبلی بود که دستورها را بر اساس خود واژه‌ها داده می‌شد. در این روش وندها و تکواژهای پرکاربرد در متن‌های تخصصی استخراج‌شده مورد توجه قرار گرفته‌اند و برای آن برنامه جداگانه‌ای نوشته شد.

برای ارزیابی متن یکی از پرونده‌های انتخاب‌شده در مرحله نمونه‌گیری مورد استفاده قرار گرفت و سپس هنجارسازی و واحدسازی و استفاده از اطلاعات جدول (۳) روی متن اعمال شد. تعداد واژه‌های متن ورودی ۶۰۲ واژه، تعداد مواردی که نیاز به اصلاح نگارشی داشتند، ۵۷ مورد بود که ۳۲ مورد توسط ابزار هضم و بررسی ساخت‌واژی اصلاح شد؛ درواقع، ۳۲ مورد به‌صورت ماشینی اصلاح شد. بنابراین، ۵۶ درصد به‌صورت ماشینی و بقیه موارد به‌صورت دستی اصلاح شد. برای محاسبه F1 مراحل زیر دنبال شد. دقت[1] ۰٫۵۶۱۴۰۳۵۰۸۸، فراخوان[2] ۰٫۰۵۳۱۵۶۱۴۶۲ و F1 ۰٫۰۹۷۱۱۶۸۴۳۷۰۲۵۷۹۶۶ اندازه‌گیری شده‌است.

```
Def Calculate_F1_Score (all_words, need_to_correct, program_detect_true):
"""Calculates the F1 score.
Args:
All_Words: The total number of words.
Need_To_Correct: The number of words that need to be corrected.
Program_Detect_True: The number of words that the program correctly detected as needing correction.
Returns:
The F1 score.
"""
Precision = program_detect_true / need_to_correct
Recall = program_detect_true / all_words
F1_Score = 2 * (precision * recall) / (precision + recall)
Return F1_score
# Calculate the F1 score.
F1_Score = Calculate_F1_Score (all_words=602, need_to_correct=57, program_detect_true=32)
# Print the F1 score.
Print (F1_score)
# 0.09711684370257966
```

[1] Precision (P)
[2] recall (R)

۳-۳. برچسب‌گذاری اجزای کلام (POS tagging)

باتوجه‌به کاربرد برچسب اجزای کلام که درواقع، خوراک بسیاری از فرایندهای نشانه‌گذاری مانند بن‌واژه‌سازی، تقطیع نحوی[1]، نشانه‌گذاری معنایی و مانند آن است، در پردازش متن، تصمیم گرفته شد این نوع برچسب نیز به پیکره افزوده شود. دراین‌راستا، برای متون فارسی (چکیده‌های فارسی) تصمیم گرفته شد از یکی از ابزارهای آماده برای برچسب‌گذاری ماشینی اجزای کلام در فارسی (ابزار هضم) استفاده شود و سپس، برچسب‌ها به‌صورت دستی کنترل شوند. برای برچسب‌گذاری متن‌های انگلیسی (چکیده‌های انگلیسی) نیز از برنامهٔ «NLTK2» بهره گرفته شد. کتابخانه «NLTK» مجموعه‌ای است که شامل کتابخانه‌ها و برنامه‌هایی برای پردازش زبان‌های آماری است. درواقع، یک کتابخانه به زبان پایتون است که نخستین بار در سال ۲۰۰۱ منتشر شد و کارهای بسیاری مانند دسته‌بندی متن‌ها، واحدسازی/جداسازی، ریشه‌یابی، برچسب‌گذاری، تجزیه نحوی و دیگر وظایف مربوط به تحلیل معنایی را پوشش می‌دهد. در فرایند ساخت پیکرهٔ مقایسه‌ای از چکیده‌های فارسی و انگلیسی پایان‌نامه‌ها و رساله‌های ثبت‌شده در ایرانداک، ابتدا قرار بود از سیستم برچسب‌گذاری لنکس‌باکس[3] بهره گرفته شود که در دانشگاه لنکستر توسعه داده شده‌است. ولی از آن‌جایی که امکان واردکردن همهٔ رکوردها (پرونده‌ها) به‌دنباله هم و به‌طور همزمان، جداسازی آن‌ها به‌منظور امکان تطبیق با پرونده‌های فارسی وجود نداشت، تصمیم گرفته شد از ابزار NLTK که در آن امکان استفاده به شیوه‌ای که بیان شد وجود دارد، به کار گرفته شود. فهرست برچسب‌های فارسی در ابزار هضم شامل برچسب‌هایی است که در مقالهٔ بی‌جن‌خان و همکاران (Bijankhan et al., 2011) معرفی شده‌است. این فهرست با آوردن نمونه‌هایی در جدول (۴) ارائه شده‌است:

جدول ٤: فهرست برچسب‌های داده‌های فارسی

POS tag	Tag Name	مثال	آمار برچسب در داده‌های فارسی پیکره پارسا
N	Noun	کشاورز، خانه، کتاب	۱۷۳۸۳۳۷۱
PREP (P)	Preposition	از، در، برای	۱۰۲۷۶۷۳۵
PUNC	Punctuation	نقطه، ویرگول، علامت سوال	۴۳۶۳۹۳۶

[1] syntactic parsing
[2] Natural Language Toolkit
[3] LancsBox

AJ	Adjective	زیبا، اجتماعی، بزرگ	۷۶۱۷۶۹۳
V	Verb	می‌تواند، خورد، شمرد	۶۲۸۳۰۲۰
CON	Conjunction	که	۶۷۳۱۷۴۴
NUM	Number	۱۲، ۵۰،۹۰	۳۳۸۰۲۴۲
PRO	Pronoun	من، تو، ایشان	۸۲۴۳۰۵
DET	Determiner	این، آن، هر	۱۷۳۳۹۴۸
ADV	Adverb	یقیناً، خوب، بسیار	۶۸۲۷۲۳
POSTP	Postposition	را	۶۲۰۲۴۷
RES	Residual	هر برچسبی غیر از برچسب‌های اصلی	۱۵۶۳۹۴۲
CL	Classifier	نوع، دست، چنین	۱۱۹۱۲۵
INT	Interjection	ای، یا	۶۸۵

نکته مهم اینکه، در فهرستی که در جدول (۳) آورده شده‌است، کسره اضافه نمایش داده نشده‌است؛ «هضم» برای نمایش کسره اضافه، هرگاه پس از واژه‌ای، کسرهٔ اضافه وجود داشته، کنار برچسب، e را می‌افزاید. برای‌نمونه، برچسب عبارت «شناسایی ساختار محتوایی مطالعات» به‌صورت شناسایی (Ne) ساختار (Ne) محتوایی (ADJe) مطالعات (N) نشانه‌گذاری می‌شود. آمار برچسب‌های همراه با کسره اضافه به‌صورت زیر است:

ADVe	96534
AJe	3106973
CONJe	6797
DETe	174172
Ne	22483499
NUMe	94391
Pe	1416302
PROe	8912
RESe	34642

کنترل دستی برچسب‌ها به این‌صورت انجام شد که فهرستی از واژه‌های پیکره به‌صورت پرونده اکسل تهیه شد تا برچسب‌های واژه‌های فارسی پیکره بررسی شوند. با توجه به

پیش‌پردازش‌های انجام‌شده و هنجارسازی داده‌های فارسی (پیش‌پردازش‌های ماشینی و استفاده از اطلاعات ساخت‌واژی فارسی و اصلاح دستی املای واژه‌های فارسی)، می‌توان گفت فرایند برچسب‌گذاری با درستی خوب و قابل‌قبولی انجام شده‌است. البته باوجود کنترل دستی برچسب‌ها، کماکان خطاهایی ممکن است در برچسب‌گذاری مشاهده شود. ولی این خطاها بسیار کم و قابل چشم‌پوشی است. برای‌نمونه، در بریده‌ای از متن برچسب‌گذاری‌شده، تنها چهار خطای برچسب‌گذاری دیده می‌شود.

شکل ۲: نمونه‌ای از متن فارسی برچسب‌گذاری‌شده

همان‌گونه که پیش‌تر اشاره شد، برای برچسب‌گذاری داده‌های چکیده‌های انگلیسی از ابزار «NLTK» بهره گرفته شد. فهرست برچسب‌ها که در سامانه مربوط به این ابزار آمده‌است در جدول (۵) ارائه شده‌است.

جدول ۵: فهرست برچسب‌های متن‌های انگلیسی

POS tag	Tag Name	آمار برچسب در داده‌های انگلیسی پیکره پارسا
CC	It is the conjunction of coordinating	3504811
CD	It is a digit of cardinal	1707777
DT	It is the determiner	6916770
EX	Existential	95839
FW	It is a foreign word	53067
IN	Preposition and conjunction	10179166

JJ	Adjective	7551971
JJR and JJS	Adjective and superlative	179109, 135049
LS	List marker	6
MD	Modal	320426
NN	Singular noun	15833944
NNS, NNP, NNPS	Proper and plural noun	6025180, 4678139, 27827
PDT	Predeterminer	21819
WRB	Adverb of wh	57361
WP$	Possessive wh	4472
WP	Pronoun of wh	46615
WDT	Determiner of wp	337079
VBZ	Verb	1646600
VBP, VBN, VBG, VBD, VB	Forms of verbs	1040006, 2533053, 1448143, 1647568, 1295483
UH	Interjection	3412
TO	To go	1349688
RP	Particle	38301
RBS, RB, RBR	Adverb	93102, 1403802, 67710
PRP, PRP$	Pronoun personal and professional	527942, 324839

٤. نتیجه‌گیری

امروزه پیدایش فناوری‌های رایانه‌ای و تولید حجم بسیار بزرگی از متن‌ها به زبان‌های گوناگون، منبع‌های پیکره‌ای عظیمی برای پژوهشگران مشتاق به ساخت پیکره فراهم کرده‌است. تعداد پیکره‌های تخصصی که برای هدف‌های ویژه و پردازش‌های خاص زبانی استفاده می‌شود، به‌اندازه پیکره‌های عمومی نیست. بنابراین، ساخت پیکره‌هایی که شامل متون تخصصی و میان‌رشته‌ای است، بسیار ارزشمند است. این طرح پژوهشی فرایند ساخت یک پیکرهٔ تخصصی دوزبانه (انگلیسی–فارسی) که شامل متون چکیده‌های پایان‌نامه‌ها و رساله‌های ثبت‌شده در ایرانداک است را شرح می‌دهد. در این پژوهش، چکیده‌ای از تجربیات پژوهشگران این حوزه دربارهٔ ساخت پیکره‌های تخصصی و مقایسه‌ای و چالش‌هایی که هنگام ساخت پیکره‌های متنی/نوشتاری با آن‌ها روبه‌رو شده‌اند، ارائه شده‌است. برای ساخت این پیکره مقایسه‌ای تخصصی، ابتدا فرایند نمونه‌گیری، سپس، هنجارسازی و واحدسازی متون فارسی پیکره انجام شده‌است. درپایان، برچسب‌گذاری متن‌های فارسی و انگلیسی (POS) انجام شده و برچسب‌های فارسی کنترل شده‌اند.

پیکره ساخته‌شده شامل بیش از ۸۹ میلیون واژه فارسی و ۷۹ میلیون واژه انگلیسی است. تعداد

واژه‌های محتوایی (فعل، اسم، صفت، قید)، ۵۷۶۵۳۸۱۳ است و تعداد واژه‌های دستوری به‌همراه اعداد و علائم سجاوندی شامل ۳۱۳۵۰۱۲۵ است. بن‌واژه‌های فارسی نیز استخراج شد و تعداد آن‌ها ۴۱۰۶۴ است. تعداد واژه‌های محتوایی متون انگلیسی (فعل، اسم، صفت، قید)، ۴۵۶۰۶۶۸۶ است و تعداد واژه‌های دستوری به‌همراه اعداد و علائم سجاوندی شامل ۳۳۶۶۲۳۰۴ است. بن‌واژه‌های انگلیسی نیز استخراج شد و تعداد آن‌ها ۱۲۹۳۷ است.

از ویژگی‌های مهم هر پیکره که در معرفی و گزارش‌های مربوط به هر پیکره وجود دارد، تعداد واژگان، تنوع حوزه‌های موضوعی و قابلیت‌های پیکره در بهره‌گیری از پردازش‌های زبانی است. دراین‌راستا، می‌توان گفت پیکرهٔ پارسا غنی است، کمتر پیکره‌ای را می‌توان یافت که شامل این تعداد واژگان باشد که حوزه‌های گوناگون تخصصی را پوشش دهد. این پیکره شامل متن‌های چکیده‌های پایان‌نامه‌ها و رساله‌های سه حوزه موضوعی کلان علوم اجتماعی، علوم انسانی و هنر، فنی‌ومهندسی و حدود ۲۸۰ رشته مربوط به این سه حوزه است. افزون‌بر شمار بالای واژگان در این پیکره، برچسب اجزای کلام نیز به داده‌های این پیکره افزوده شده‌است که یکی از پرکاربردترین نوع برچسب‌گذاری به شمار می‌آید. این نوع برچسب‌دهی، عملی کاربردی در بسیاری از حوزه‌های پیشرفته‌تر پردازش زبان طبیعی از جمله ترجمهٔ ماشینی، خطایاب، تبدیل متن‌به‌گفتار، بازیابی اطلاعات، موتورهای جستجو و کمک به مدل‌های آماری است. این پیکره می‌تواند به‌عنوان یک مرجع تخصصی برای هدف‌های پردازش زبان طبیعی، به‌ویژه در ترجمه ماشینی به کار گرفته شود. مهم‌ترین چالشی که در ساخت این پیکره وجود داشت، هنجارسازی و واحدسازی داده‌های فارسی پیکره بود که تلاش شد با ابزارهای مربوطه، برنامه‌نویسی ماشینی و اصلاح نگارشی به‌صورت دستی، تا اندازهٔ ممکن این مرحله با درستی خوبی انجام شود و سپس، داده‌ها برچسب‌گذاری شوند. در مورد داده‌های انگلیسی این مشکلات وجود نداشت؛ چون در انگلیسی مشکلات مربوط فاصله و نیم‌فاصله، چگونگی اتصال وندها به ستاک و مانند آن وجود ندارد. بنابراین، نیازی نبود متون انگلیسی از نظر هنجارسازی و واحدسازی بررسی شوند و تنها کاری که روی داده‌های انگلیسی انجام شد برچسب‌گذاری ماشینی بود. پیش از استفاده از این پیکره در ترجمه ماشینی لازم است ابتدا پیکرهٔ پارسا، که پیکره‌ای است مقایسه‌ای، به پیکرهٔ موازی تبدیل شود تا برای هر جمله از آن در فارسی، ترجمه معادل آن در انگلیسی آورده شود.

فهرست منابع

امرایی، علیرضا، اکبر حسابی و عباس اسلامی راسخ (۱۳۹۸). «طراحی پیکره و فرهنگ دوزبانهٔ اصطلاحات راهنمایی و رانندگی بر پایهٔ معناشناسی قالبی». *مطالعات زبان و ترجمه*. دوره ۵۲. شماره ۲. صص ۹۷-۶۵. https://doi.org/10.22067/lts.v52i2.80823/

دشتبانی، شکوفه، محرم منصوری‌زاده و محمد نصیری (۱۳۹۳). «پیکرهٔ متنی تطبیقی فارسی-انگلیسی حوزهٔ تخصصی فاوا». *پژوهش‌های زبان‌شناسی تطبیقی*. سال ۴. شماره ۸. صص ۱۴۱-۱۲۱.
Retrieved from <https://rjhll.basu.ac.ir/article_972.html>

صادقی، علی‌اشرف (۱۳۷۰-۱۳۷۲). *شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱-۱۲)*. تهران: نشر دانش. شمارهٔ ۶۴-۸۰. Retrieved from >https://ensani.ir/fa/article/293365/<

علایی، الهام، نصراله پاک‌نیت، علی‌اصغر حجت‌پناه، مجتبی زالی و محمدهادی آقالویی آغمیونی (۱۴۰۰). *ساخت پیکرهٔ متنی از مقاله‌های پژوهش‌نامهٔ پردازش و مدیریت اطلاعات*. تهران: پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک).
Retrieved from <https://irandoc.ac.ir/sites/fa/files/attach/research/559pf.pdf>

قطره، فریبا (۱۳۸۶). «مشخصه‌های تصریفی در زبان فارسی امروز». *دستور*. شماره ۳. صص ۵۲-۸۱.
Retrieved from <https://ensani.ir/fa/article/99232>

قیومی، مسعود (۱۴۰۱). «پیش‌پردازش و ابزارهای پایه». در *پردازش و متن گفتار فارسی: مروری بر مبانی نظری و آخرین یافته‌های پژوهشی*». به کوشش مهرنوش شمس‌فرد و محمود بی‌جن‌خان. تهران: سازمان مطالعه و تدوین کتب دانشگاهی در علوم اسلامی و انسانی (سمت). پژوهشکده تحقیق و توسعه علوم انسانی. صص ۸۶-۱۱۳.
Retrieved from <https://samt.ac.ir/fa/book/6143>

کشانی، خسرو (۱۳۷۱). *اشتقاق پسوندی در زبان فارسی امروز*. تهران: مرکز نشر دانشگاهی.
Retrieved from <https://daneshnegar.com/fa/product/39614>

کوهستانی، منوچهر (۱۳۸۹). *بررسی خطاهای املایی و نگارشی در وبلاگ‌های فارسی و ماهیت زبان‌شناختی آن‌ها*. پایان‌نامهٔ کارشناسی ارشد. دانشگاه تهران.

لازار، ژیلبر (۱۳۸۹). *دستور زبان فارسی معاصر*. ترجمهٔ مهستی بحرینی. چ ۲. تهران: انتشارات هرمس.
Retrieved from <https://www.hermespub.ir/product/>

محمدی، علی محمد (۱۴۰۲). «رابطهٔ بین عناصر گفتمانی در پیکره های موازی: موردپژوهیِ ترجمهٔ شفاهی همزمان». *زبان‌پژوهی. دورهٔ ۱۵. شمارهٔ ۴۷. صص ۲۶۲-۲۹۳.*
https://doi.org/10.22051/jlr.2021.36750.2056

محمدی، رویا (۱۳۹۱). ساخت پیکرهٔ تطبیقی فارسی-انگلیسی و استخراج جملات موازی از آن. پایان‌نامهٔ کارشناسی ارشد. دانشگاه الزهرا (س).
Retrieved from <https://elmnet.ir/doc/10526832-12611>

References

Alayiaboozar, E., & Hojjatpanah, A (2022). Steps for creating two Persian specialized corpora. *International Journal of Information Science and Management (IJISM)*, *20*(4), 231-243. https://ijism.isc.ac/article_698428.html

Alayiaboozar, E., Pakniat, N., Zali, M., & Aghalooyi Aghmiyooni,.M.H. (2021). *Building a corpus from the published articles of Iranian Journal of Information Management and Processing*. Iranian Research Institute for Information Science and Technology (Irandoc). https://irandoc.ac.ir/sites/fa/files/attach/research/559pf.pdf [In Persian]

Asghari, H., Khoshnava, Kh., Fatemi, O., & Faili, H. (2015, September 8-11). *Developing bilingual plagiarism detection corpus using sentence aligned parallel corpus* [Conference presentation]. Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France. https://ceur-ws.org/Vol-1391/148-CR.pdf

Atkins, S. J. Clear., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*, 7(1), 1-16. https://doi.org/10.1093/llc/7.1.1

Beloso, B. S. (2015). Designing, describing and compiling a corpus of English for architecture. *Procedia-social and behavioral sciences*, *198*, 459-464. https://doi.org/10.1016/j.sbspro.2015.07.466

Bijankhan, M., Sheykhzadegan, J., Bahrani, M., & Ghayoomi, M. (2011). Lesson from building a Persian written corpus: Peykare. *Language resources and evolution*, *45*(2), 143-164. https://doi.org/10.1007/s10579-010-9132-x

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. https://doi.org/10.18806/tesl.v34i1.1257

Dashtbani, Sh., Mansoorizade, M., & Nasiri, M. (2014). English-Persian comparable textual corpus in FAVA domain. *Comparative linguistic research*, *4*(8), 121-141. https://rjhll.basu.ac.ir/article_972.html [In Persian]

Emrayi, A., Hesabi, A., & Eslami Rasekh, A. (2019). Designing corpus and bilingual traffic terms based on frame semantics. *Language and translation studies*, **52**(2), 65-97. https://doi.org/10.22067/lts.v52i2.80823 [In Persian]

Ghatre, F. (2007). Inflectional features in modern Persian. *Dastoor*, *3*, 52-81. https://ensani.ir/fa/article/99232 [In Persian]

Ghayoomi, M. (2022). Preprocessing and basic tools. In Shams Fard, M. & Bijan Khan, M. (Eds.), *Text and speech processing for the Persian language: the state of art and a brief review of the theoretical foundations* (pp. 86-113). SAMT. https://samt.ac.ir/fa/book/6143 [In Persian]

Ghayoomi, M., Momtazi, S., & Bijankhan, M. (2010). A Study of Corpus Development for Persian. *International Journal of Asian Language Processing*, *20*(1), 17-34. https://www.colips.org/journals/volume20/20.1.02-Masood-Ghayoomi.pdf

Karimi, A., Ansari, E., & Sadeghi Bigham, B. (2017). Extracting an English-Persian parallel corpus from comparable corpora. (Project: Machin translation. Parallel sentence extraction from comparable corpora using statistical machine translation). Arxiv: 1711.00681v3 [cs.CL]. https://doi.org/10.48550/arXiv.1711.00681

Kenning, M. M. (2010). What are parallel and comparable corpora and how can we use them. In O'Keeffe, A., McCarthy, M. (Eds.), *The Routledge Handbook of*

Corpus Linguistics (pp. 487–500). Routledge. https://www.routledge.com/The-Routledge-Handbook-of-Corpus-Linguistics/OKeeffe-McCarthy/p/book/ 9780367076382? srsltid= AfmBOorRl_ 9evhDUirMihhq4DAgTjE8fSY1aCgOygHEO9igc2RRFNTf7

Keshani, Kh. (1992). *Derivation suffix in modern Persian*. Markaz Nashr Daneshgahi. https://daneshnegar.com/fa/product/39614 [In Persian]

Kokabi, A., Nourian, A., Ghafourzadeh, E., Imani, M., Fallah, M., Mahdavi Mortazavi, M., Ghorbani, M., Ruhollah, R., Ebrahimi, M., Riasati, R., Khallash, M., Khosrotabar, M., Bashari, H., Mahdizade, M., Souri, Y., Kharazi, V… Qayyoomi, A. (2023, October 5). Persian NLP Toolkit. github. https://github.com/roshan-research/hazm

Koltunski, E. L. (2013). VARTRA: A comparable corpus for analysis of translation variation. In Sharoff, S., Zweigenbaum, P., & Rapp, R. (Eds.), *Proceedings of the 6th workshop on building and using comparable corpora*. (pp. 77-86). Association for computational linguistics. https://www.researchgate.net/publication/263352667_VARTRA_A_Comparable_Corpus_for_Analysis_of_Translation_Variation

Kouhestani, M. (2010). *Studying written errors In Persian weblogs and their linguistic nature* [Unpublished master's thesis]. University of Tehran. [In Persian]

Lazard, G. (2010). *Persian Grammar*. Hermes. https://www.hermespub.ir/product/ [In Persian]

Mohammadi, A. M. (2023). A study of the relationship between discoursal elements in parallel corpora: a case study of simultaneous interpretation. *ZABANPAZHUHI (journal of language research)*, *15*(47), 236-262. https://doi.org/10.22051/jlr.2021.36750.2056 [In Persian]

Mohammadi, R. (2012). Building Persian-English comparable corpus and extracting parallel sentences [Unpublished master's thesis]. Alzahra University. https://elmnet.ir/doc/10526832-12611 [In Persian]

Sadeghi, A. A. (1991-1993). *Word formation methods In Persian*. Danesh publication. https://ensani.ir/fa/article/293365/ [In Persian]

Sinclair, J. (2004). Corpus and Text-Basic Principles. In *Wynne, M.* (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 5-25). The Oxford Text Archive. https://users.ox.ac.uk/~martinw/dlc/chapter1.htm