فصلنامه علمی زبان‌پژوهی دانشگاه الزهرا(س)
نوع مقاله: پژوهشی

ساخت پیکره مقایسه‌ای تخصصی «پارسا»

الهام علایی ابوذر[۱]
علی‌اصغر حجت‌پناه[۲]

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

چکیده

پیکره‌ها بر اساس زبان به کار رفته در متون تشکیل‌دهندهٔ آن‌ها به پیکره‌های تک‌زبانه، دوزبانه و چندزبانه تقسیم می‌شوند. پیکرهٔ مقایسه‌ای، پیکره‌ای است دوزبانه یا چندزبانه که شامل متونی است مشابه در حوزه‌های موضوعی یکسان. علیرغم کاربرد فراوان این نوع پیکره‌ها در مطالعات مختلف از جمله 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 شده است. این پیکره از چکیده‌های فارسی و انگلیسی پایان‌نامه‌ها و رساله‌های ثبت‌شده در پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) ساخته شده است و شامل بیش از ۸۹ میلیون واژه فارسی و ۷۹ میلیون واژه انگلیسی است. محتوای این پیکره عمومی نیست و حاوی متون بسیار تخصصی در

[۱] استادیار پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک). alayi@irandoc.ac.ir
[۲] رئیس اداره سامانه‌های اطلاعاتی، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک). hojjatpanah@irandoc.ac.ir
تاریخ ارسال مقاله: September 17, 2023 و تاریخ پذیرش مقاله: October 31, 2023

حوزه‌های موضوعی کلان مانند علوم اجتماعی، علوم انسانی و هنر، فنی-ومهندسی و رشته‌های مربوط به این حوزه‌ها است و از این نظر برای پردازش‌های زبانی که مستلزم بهره‌گرفتن از متون 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 محاسبه شد. دقت، ۰.۵۶۱۴۰۳۵۰۸۸، فراخوان، ۰.۰۵۳۱۵۶۱۴۶۲ و در نهایت F1 ۰.۰۹۷۱۱۶۸۴۳۷۰۲۵۷۹۶۶ محاسبه شده است. سپس، داده‌ها برچسب-گذاری شدند (برچسب‌گذاری اجزای کلام) و برچسب‌های متون فارسی کنترل شدند. داده‌های انگلیسی نیز به صورت ماشینی برچسب‌گذاری شدند. تعداد واژه‌های محتوایی (فعل، اسم، صفت، قید) داده‌های فارسی این پیکره ۵۷۶۵۳۸۱۳ و تعداد واژه‌های دستوری به همراه اعداد و علائم سجاوندی ۳۱۳۵۰۱۲۵ است و بن‌واژه‌های فارسی استخراج‌شده نیز شامل ۴۱۰۶۴ بن‌واژه است. تعداد واژه‌های محتوایی متون انگلیسی ۴۵۶۰۶۶۸۶ و تعداد واژه‌های دستوری به همراه 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 استفاده 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 ظهور فناوری‌های رایانه‌ای و تولید حجم بسیار بزرگی از متون به زبان‌های گوناگون، منابع پیکره‌ای عظیمی برای پژوهشگران مشتاق به ساخت پیکره فراهم کرده

است. با افزایش قدرت و ظرفیت رایانه‌ها، پیکره‌ها نیز از نظر اندازه، تنوع و سهولت دسترسی، افزایش چشم‌گیری داشته‌اند و هم‌زمان با این تحولات، 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 گرفته است. گرچه در گذشته، بسیاری از زبان‌شناسان بر اهمیت پیکرهٔ زبانی در بیشتر بررسی‌هایشان تاکید کرده‌اند، اما در دوران جدید است که تکیه بر داده‌های واقعی زبان به صورت 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 استفاده از پیکره‌ها در پژوهش‌ها و کارهای گوناگون، سابقهٔ طولانی دارد. امکان سازمان‌دهی، تنظیم، تفکیک، جست‌وجو و دستیابی سریع داده‌های زبانی، افق‌های تازه‌ای در برابر پژوهشگران گشوده و باعث پیدایش شاخه‌ای تخصصی در حوزهٔ زبان‌شناسی گردیده است. این شاخه با نام زبان‌شناسی پیکره‌ای[1]، تنها در آخرین دهه‌های قرن بیستم ایجاد شده است و در همین زمان کوتاه تبدیل به یکی از فعال‌ترین و پرکاربردترین زمینه‌ها شده است. ویژگی این رشته این است که به همهٔ حوزه‌های زبان‌شناسی خدمات می‌دهد و در حقیقت، 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. در حقیقت، 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. در این میان، 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،

[1] corpus linguistics

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 اهداف 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 سایر 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[1]»، «نمونه‌ای[2]»، «نظارتی[3]»، «بسته[4]»، «باز[5]»، «هم‌زمانی[6]»، «درزمانی[7]»، «عمومی[8]»، «اصطلاح‌شناسی[9]»، «یک‌زبانه[10]»، «دوزبانه[11]»، «چندزبانه[12]»، «منفرد[13]»، «موازی[14]»، «مرکزی[15]»، «پوسته‌ای[16]»، «هسته‌ای[17]» و «پیرامونی[18]» طبقه‌بندی کرده‌اند.

پیکره‌ها بر اساس زبان بکاررفته در متون تشکیل‌دهندهٔ آن‌ها به پیکره‌های تک‌زبانه، دوزبانه و چندزبانه تقسیم می‌شوند. پیکره‌های دوزبانه یا چندزبانه با دو نام شناخته می‌شوند: پیکره‌های مقایسه‌ای/تطبیقی[19] و پیکره‌های موازی[20]. زمان زیادی طول کشید تا واژگان تخصصی حوزهٔ زبان‌شناسی پیکره‌ای جا بیفتد، در این میان، در ابتدا واژهٔ «موازی«

---

[1] whole text
[2] samples
[3] monitor
[4] open
[5] closed
[6] synchronic
[7] diachronic
[8] general
[9] thesaurus
[10] monolingual
[11] bilingual
[12] multilingual
[13] single
[14] parallel
[15] central
[16] cluster
[17] nuclear
[18] Perimeter
[19] comparable corpora
[20] parallel corpora

به جای آنچه امروزه «مقایسه‌ای» خوانده می‌شود، به کار می‌رفته است. امروزه این دو نوع پیکره را به این صورت تعریف می‌کنند: پیکرهٔ مقایسه‌ای، پیکره‌ای است دوزبانه یا چندزبانه که شامل متونی است مشابه در حوزه‌های موضوعی یکسان؛ به عبارتی دیگر، پیکرهٔ مقایسه‌ای، مجموعه اسنادی در دو زبان متفاوت هستند که موضوعات مشابهی را پوشش می‌دهند. در حالیکه پیکرهٔ موازی شامل متونی است که برای هر جمله از آن در یک زبان، ترجمه معادل آن در زبان دیگر آورده شده است. در حقیقت، متون پیکره‌های موازی به منبع مشترکی مربوط می‌شوند، مانند ترجمه‌های فرانسه و آلمانی آثار چارلز دیکنز[1]. در پیکرهٔ مقایسه‌ای، برخلاف پیکرهٔ موازی، متون لزوماً ترجمهٔ یکدیگر نیستند، اما به یک حوزهٔ یکسان و فرادادهٔ یکسان مربوط می‌شوند. در حقیقت، آنچه مجموعه متون در پیکره‌های مقایسه‌ای را به هم ارتباط می‌دهد، معیارهای مشابهی همانند اندازهٔ متون، موضوع آن‌ها، تاریخ متون، ویژگی‌های تالیف (مانند ژانر، ملیت نویسنده و غیره) است. متون مطبوعات، سخنرانی‌های انتخاباتی، آگهی‌های استخدام و سایر موارد مشابه، موضوعاتی هستند که در همهٔ فرهنگ‌ها وجود دارند و برای نگارش آن‌ها معمولاً از قراردادهای مشابهی پیروی می‌شود؛ این منابع مورد علاقهٔ پژوهشگرانی است که اقدام به ساخت پیکره‌های مقایسه‌ای می‌کنند. نمونه‌ای از پیکرهٔ مقایسه‌ای، پیکره‌ای است که از ویکی‌پدیا[2] ساخته شده است 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 تحریف اجتناب‌ناپذیر که در ترجمهٔ متون در پیکره‌های موازی ایجاد می‌شود، پرهیز کرد. پیکره‌های مقایسه‌ای می‌تواند از متون عمومی ساخته شود که امکانات گوناگونی را برای تحلیل گفتمان، کاربردشناسی، تجزیه‌وتحلیل ژانرهای متون و زبان-شناسی اجتماعی فراهم می‌کند؛ نمونه‌هایی از چنین پیکره‌هایی می‌تواند شامل مجموعه مدخل‌های دایرةالمعارف‌ها یا متون ادبی یک دوره زمانی خاص باشد. اما متداول‌ترین نوع پیکره‌های مقایسه‌ای که مخاطبان بسیاری دارد، آن‌هایی هستند که به حوزه(های) تخصصی مربوط می‌شوند و دارای تراکم بالای واژگان و اصطلاحات تخصصی هستند.

---

[1] Dikens
[2] Wikipedia

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 متونی از ژانرهای گوناگون یا گونه‌های زبانی منطقه‌ای هستند یا جفت پیکره‌هایی که بر اساس 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، مراحل ساخت پیکره، شامل نمونه‌گیری، هنجارسازی و واحدسازی داده‌های فارسی و در نهایت برچسب‌گذاری داده‌های فارسی و انگلیسی توضیح داده می‌شود. شایان ذکر است فرایند هنجارسازی و واحدسازی تنها در مورد داده‌های فارسی انجام شده است و داده‌های انگلیسی فقط برچسب‌گذاری ماشینی روی آن‌ها انجام شده است.

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 توضیح می‌دهند. داده‌های موازی بخش مهمی از ترجمهٔ ماشینی را تشکیل می‌دهند؛ هر چه داده‌های بیشتری در دسترس 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 شده است. در این روش دو سیستم ترجمهٔ ماشینی برای ترجمه از فارسی به انگلیسی و برعکس، به کار گرفته شده است. پس از آن، از یک سامانه بازیابی اطلاعات [5] برای اندازه‌گیری شباهت جمله‌های ترجمه‌شده،

---

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

استفاده شده است. اضافه کردن جمله‌های استخراج‌شده به داده‌های آموزشی موجود در سیستم ترجمهٔ ماشینی، موجب بهبود کیفیت ترجمه شده است. علاوه بر آن، روش پیشنهادی کمی بهتر از رویکرد یک‌سویه عمل می‌کند. پیکرهٔ استخراج شده تقریباً شامل ۲۰۰ هزار جمله است که بر اساس درجهٔ شباهت مرتب شده‌اند که سیستم بازیابی اطلاعات این میزان شباهت را محاسبه کرده است. کولتانسکی (Koltunski, 2013) یک پیکرهٔ مقایسه‌ای ترجمه‌ای را معرفی می‌کند. هدف از ساخت چنین پیکره‌ای، بررسی تفاوت‌های ترجمه برحسب متغیرهای تفاوت در زبان، نوع متن و روش‌های ترجمه (ماشینی، به کمک ماشین در مقابل ترجمهٔ انسانی است). موارد ذکر شده در ویژگی‌های زبانی متن ترجمه شده منعکس می‌شوند. به منظور تجزیه‌وتحلیل ترجمه‌های انجام شده، تلفیقی از روش-هایی که در مطالعات ترجمه، تفاوت‌های گونه‌های زبانی و ترجمهٔ ماشینی، با تاکید بر تفاوت‌های متنی و دستوری-واژگانی به کار گرفته می‌شوند، مورد استفاده قرار گرفته است. در این پژوهش علاوه بر ساخت پیکرهٔ مقایسه‌ای، بررسی‌های انجام شده در زمینهٔ ترجمه در حوزه‌های گوناگون، از جمله مطالعات ترجمه، ترجمهٔ ماشینی و سایر حوزه‌های مشابه نیز به کار گرفته شده است. از جمله تلاش‌هایی که در زمینهٔ ساخت پیکره‌های دوزبانه فارسی- انگلیسی انجام گرفته است، 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 اصطلاحات 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 استفاده از معناشناسی قالبی برچسب‌گذاری شده است. برای این منظور یک هستان-شناخت[1] و یک شبکه قالبی برای این حوزه طراحی شده است. سپس یک رابط کاربری برای جستجو در این دادگان طراحی شده است که امکانات مختلفی شامل جستجوی سنتی الفبایی و جستجوی مبتنی بر معنا را فراهم می‌سازد. این کار به مترجمان، که در حقیقت

[1] ontology

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 (حوزهٔ فناوری ارتباطات و اطلاعات) پرداخته‌اند. این پیکره به صورت خودکار ساخته شده است و منابع آن، اسناد تخصصی حوزهٔ فاوا است. در این پژوهش، نرم‌افزاری برای ساخت پیکره طراحی شده است که هزینه و مدت زمان ساخت پیکره را کاهش می‌دهد. علاوه بر این، نرم‌افزار ارائه شده 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 شده است، حاشیه‌نویسی اسناد، جستجو در پیکره و تصحیح خطا را آسان می‌کند. قبل از شروع فرآیند پردازش اصلی، هر سند توسط نرم‌افزار پیش‌پردازش می‌شود؛ این کار به منظور انتخاب جمله‌های درست و معنی‌دار برای پردازش اصلی است؛ علاوه بر آن، در صورتیکه 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، اضافه کردن متن‌های جدید به پیکره، اندیس‌گذاری و حاشیه‌نویسی پیکره. بخش دوم، یک موتور جستجوی پیکره است که برای مدیریت مجموعه بزرگی از متون طراحی شده است. 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

---

[1] classifier
[2] parser
[3] stemmer

تخصصی حوزه فاوا و جملات انگلیسی پیکرهٔ تخصصی حوزه فاوا ارائه شده است. الگوریتم پیشنهادی آن‌ها از مدل ترجمهٔ کلمه‌به‌کلمه و تکنیک بلندترین زیردنبالهٔ مشترک[1] برای ترازبندی استفاده می‌کند و در نهایت امتیاز نشان دهندهٔ شباهت دو جمله، محاسبه 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 توضیح می‌دهد. برای ایجاد این پیکره از اسناد خبری روزنامه‌های همشهری و بی.بی.سی استفاده شده است و از اسناد بدست آمده، معیارهایی مانند تعداد کلمات کلیدی مشترک، اسامی خاص یکسان، عناوین مشابه و فاصلهٔ تاریخ انتشار دو خبر استخراج شده است. سپس، معیارهای بدست آمده از مرحلهٔ قبل، براساس میزان اهمیت آن‌ها در ترازبندی متون، با وزن‌های مختلف با یکدیگر ترکیب شده‌اند. در گام بعد، به استخراج جمله‌های موازی از پیکرهٔ مقایسه‌ای ساخته شده پرداخته شده است. بدین منظور، 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 بوده است، تصفیه شده است. پس از تصفیه، به استخراج ویژگی‌های واژگانی، طولی و هم‌پوشانی واژه‌ها از جمله‌های منتخب پرداخته شده است و در نهایت با استفاده از جمله‌های آموزشی پیکرهٔ موازی موجود و ویژگی‌های استخراج شده، با به کارگیری 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 شده است و از مقاله‌های ویکی‌پدیا استفاده کرده‌اند. جفت‌جمله‌ها در پیکرهٔ موازی امتیاز مشابه بین صفر تا یک دارند. در این پژوهش، اصغری و همکاران از

---

[1] longest common subsequence (LCS)-هدف این روش،مقایسهٔ دو رشته و پیدا کردن شباهت بین آن هاست

امتیازهای مشابه برای مشخص کردن درجهٔ ابهام به منظور ساخت موارد سرقت ادبی استفاده کرده‌اند.

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 حاشیه‌نویسی شده و مبتنی بر فراوانی واژگان در پیکره را گام‌به‌گام توضیح می‌دهد. وی از پروژهٔ «آموزش زبان انگلیسی برای اهداف دانشگاهی» که در دانشگاهی در خاورمیانه توسعه یافته است، نمونه‌هایی را ذکر می‌کند. مراحل ساخت پیکرهٔ تخصصی از نظر کلود توریدا شامل انتخاب متون آموزشی، حذف واژگانی که قاموسی نیستند (مانند حروف اضافه، حروف ربط و....)، تجزیه و تحلیل متن با استفاده از نرم‌افزار AntConc، ایجاد فهرست فراوانی واژه‌ها، توسعهٔ فهرست واژگان حاشیه‌نویسی شده که خود شامل: تعیین مقولهٔ نحوی (POS) واژگان موجود در فهرست، اضافه کردن تعریف واژگان، باهم‌آیی واژگان و نمونه‌ای از جمله‌ای که واژه در آن به کار رفته، است. بلوسو (Beloso, 2015) نیز پیکرهٔ تخصصی CADCE[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 شده است. این پیکره تک‌زبانه است، برچسب‌گذاری نشده است و شامل متون منتشر شده از گونه‌های زبانی انگلیسی آمریکای شمالی، بریتانیا، ایرلندی، کانادایی و استرالیایی است. از آنجائیکه شامل متون منتشر شده در سال‌های اخیر (۲۰۰۸-۲۰۰۷) است، پیکره هم‌زمانی است. این پیکرهٔ تخصصی دارای متونی در حوزه‌های مربوط به معماری، شامل ساخت‌وساز، شهرسازی، مواد ساخت‌وساز، معماری سبز، طراحی داخلی و سایر حوزه‌های این رشته است. این پیکره روی نمایندگی، معاصر بودن و قابلیت در دسترس بودن به عنوان اصول مهم ساخت پیکره تاکید دارد. همچنین در راستای ساخت پیکره‌های تخصصی، علایی ابوذر و همکاران (Alayiaboozar et al., 2021) و علایی و حجت‌پناه (*Alayiaboozar, Elham &*

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

(*Ali Asghar Hojjatpanah, 2022*) نیز دو پیکره تخصصی ساخته‌اند: پیکره پژوهش‌نامه، که از متون مقاله‌های «پژوهش‌نامه پردازش و مدیریت اطلاعات » ساخته شده است، و پکا، که از متون کتاب‌های دیجیتال ایرانداک ساخته شده است. هر کدام به ترتیب، شامل بیش از چهار میلیون و ۷۸۰ هزار واژه و سه میلیون و ۳۲۹ هزار واژه است. محتوای این پیکره‌ها، متون 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 به منظور 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 متن به گفتار، بازیابی اطلاعات، موتورهای جستجو و کمک به مدل‌های آماری است.

۳- مراحل ساخت پیکره پارسا

برای ساخت پیکره مقایسه‌ای تخصصی پارسا مراحل زیر طی شده است که در شکل ۳- ۱ نیز نمایش داده شده است:

- نمونه‌گیری
- پیش‌پردازش (هنجارسازی و واحدسازی) داده‌های فارسی
- برچسب‌گذاری اجزای کلام (POS tagging) داده‌های فارسی
- برچسب‌گذاری اجزای کلام (POS tagging) داده‌های انگلیسی
- کنترل صحت برچسب‌های داده‌های فارسی

شکل ۳-۱: مراحل ساخت پیکره مقایسه‌ای تخصصی «پارسا»

۳-۱- نمونه‌گیری

نمونه‌گیری در واقع عمل انتخاب متون مربوط به هر ژانر با توجه به هدف تهیهٔ پیکره است. برخی از معیارهایی که بر اساس آن‌ها نمونه‌گیری صورت می‌پذیرد شامل این موارد است: شکل متن[1] (گفتاری/ نوشتاری/ الکترونیکی)، نوع متن (کتاب/ مجله/ نامه)، حوزهٔ متن (آکادمیک (علمی)/ عمومی)، زبان متن (زبان‌ها یا گونه‌های زبانی پیکره) و مکان متن (به عنوان مثال، انگلیسی بریتانیا باشد یا استرالیا) است (سینکلر: ۲۰۰۴). داده‌های پیکره پیش از انجام پژوهش انتخاب شده بودند؛ به دلیل دسترسی به چکیده‌های فارسی و انگلیسی پایان‌نامه‌ها و رساله‌های (پارساهای) ثبت‌شده در ایرانداک، این متون برای ساخت پیکره استفاده شدند. متون این چکیده‌ها شامل متون تخصصی یا میان‌رشته‌ای است، بنابراین، حوزهٔ متن آکادمیک (علمی) است. شکل متون، نوشتاری و به صورت الکترونیکی بوده است و از طریق خروجی گرفتن از پایگاه مربوطه در ایرانداک تهیه شده است. نوع متون، چکیده‌های فارسی و انگلیسی پایان‌نامه‌ها و رساله‌ها است. حوزهٔ متون، آکادمیک (علمی) است، چون شامل پایان‌نامه‌ها و رساله‌هایی است که در دانشگاه‌ها و موسسات پژوهشی توسط دانشجویان رشته‌های گوناگون نوشته شده است و گونهٔ زبانی متون، گونهٔ نوشتاری و رسمی است.

---

[1] mode

۳-۱-۱-جامعهٔ آماری

جامعه آماری استفاده شده در ساخت این پیکره، چکیده پایان‌نامه‌ها و رساله‌های ثبت‌شده در ایرانداک است که از سامانه‌های *ثبت و ویرایش* گرفته شده است. ایرانداک کار مدیریت اطلاعات علم و فناوری را از سال ۱۳۴۷ شروع کرده است و افزون بر اطلاعات جاری، اطلاعات پیش از آن را نیز سازمان داده است. 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 شده است، به این ترتیب، این مرجع منبع غنی 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 شده است. 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 ۱-۳ ارائه شده است.

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

جدول ۳-۱: 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

۳-۱-۲- نمونه‌گیری از جامعه آماری

برای نمونه‌گیری از سامانه ثبت پایان‌نامه‌ها و رساله‌ها، ابتدا باید پرونده‌هایی انتخاب می‌شدند که علاوه بر چکیده فارسی، چکیده انگلیسی نیز داشته باشند. شایان ذکر است که پرونده چکیده‌ها به صورت مجزا در سامانه بارگزاری می‌شود و نیازی به جداکردن بخش چکیده از کل پایان‌نامه یا رساله نیست. نمونه‌گیری از سامانه ثبت به صورت ماشینی و خودکار و از طریق اتصال به منبع داده، کپی کردن داده‌ها و بارگزاری 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 قطعاً در این پژوهش نمی‌تواند مورد استفاده قرار گیرد. بر اساس این بررسی، تصمیم گرفته شد فیلتر تعداد واژه موجود در هر پرونده برای نمونه‌گیری قرار داده شود، به این صورت که پرونده‌هایی به صورت ماشینی انتخاب شوند که حداقل 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 حاوی 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 مورد استفاده قرار می‌گیرد. اطلاعات پایان‌نامه‌ها و رساله‌ها

(پارساها) ثبت‌شده دانشجویان پس از ذخیره‌سازی در سامانه ثبت، به صورت خودکار و سیستمی وارد سامانه ویرایش می‌شود. اطلاعات این سامانه توسط کارشناسان مدیریت سازماندهی و تحویل اطلاعات، فهرست‌نویسی و نمایه‌سازی و ویرایش می‌شود و در نهایت از طریق سامانه ویرایش، اطلاعات وارد پایگاه اطلاعات علمی ایران (گنج)[1] 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 بر اساس وب‌گاه وب‌آوساینس[2] 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 ذکرشده مدنظر قرار گرفت: پرونده‌ها حاوی معادل چکیده انگلیسی باشند و حداقل تعداد واژه ۲۰۰ و حداکثر ۲۰۰۰ برای خروجی گرفتن مدنظر قرار گیرد. پس از اعمال فیلتر حدوداً ۲۹۵ هزار پرونده چکیده انتخاب شدند که معادل انگلیسی نیز 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[3] و واحدسازی[4]. «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 یکدست کردن انواع «ی» و «ک» و «کسره اضافه روی «ه» در حالت اضافی («ۀ»)). همچنین اصلاحات دیگری نیز به منظور پردازش دقیق‌تر

---

[1] Ganj.irandoc.ac.ir
[2] Web Of Science
[3] normalization
[4] tokenization

متون در این مرحله صورت می‌پذیرد. برای رسیدن به این هدف، قبل از مقایسهٔ متون، پیش‌پردازش‌هایی روی آن‌ها انجام می‌شود. هنجارسازی در متون فارسی می‌تواند شامل این موارد باشد: یک‌دست کردن فاصله‌ها و نشانه‌گذاری‌های درون متن، یکسان کردن یونیکد نویسه‌های استفاده شده در متون، یکسان کردن روش اتصال وندهای گوناگون به ستاک، اصلاح غلط‌های املایی، ارتباط دادن کلمات چندواژه‌ای و یکسان در نظر گرفتن آن‌ها و غیره. در فرایند پردازش داده‌های زبانی، معمولاً فاصله کامل به عنوان مرزنمای تشخیص واژه شناخته می‌شود. ولی تشخیص صحیح واژه در خط فارسی و عربی با چالش پراهمیت چندواژگی مواجه است. عدم تشخیص واحدهای واژگانی به تاثیرگذاری بر تمام لایه‌های پردازشی می‌انجامد. به فرایند تشخیص یک واحد واژگانی، واحدسازی می‌گویند (Ghayoomi, 2022). به طور کلی، در پردازش رسم‌الخط زبان فارسی، با توجه به قرابتی که با رسم‌الخط عربی دارد، همواره در پردازش تعدادی از نگاره‌ها مشکلاتی وجود دارد که در اولین گام باید مشکلات مربوط به این نگاره‌ها را برطرف ساخت. علاوه بر این، اصلاح و یکسان‌سازی نویسه‌ی نیم‌فاصله و فاصله در کاربردهای مختلف آن و همچنین حذف نویسه‌ی «ـ» که برای کشش نویسه‌های چسبان مورد استفاده قرار می‌گیرد و مواردی مشابه برای یکسان‌سازی متون، از اقدامات لازم قبل از شروع مراحل مختلف می‌باشد. از دیگر تفاوت‌های نظام نوشتاری فارسی در مقایسه با انگلیسی، نحوه اتصال وندها به ستاک است. در زبان فارسی بسیاری از وندها به ستاک متصل می‌شوند و چگونگی اتصال وندها به ستاک، می‌تواند به صورت با فاصله، نیم‌فاصله یا حتی بدون فاصله باشد. به عنوان نمونه، هر سه حالت «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 نتایج درستی نخواهیم رسید. قیومی و همکاران (Ghayoomi et al., 2010) معتقدند برای صرفه‌جویی در وقت و انرژی، داده‌های خام هم به صورت خودکار و هم به صورت دستی پیش‌پردازش شوند. انجام بسیاری از عملیات خودکار بر روی زبان مانند ترجمه،

خلاصه‌سازی، تصحیح املا و غیره، 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 صورت می‌گیرد. البته در برخی موارد ترکیبی از هر دو روش مورد استفاده قرار می‌گیرد. از این جهت طراحی و پیاده‌سازی این ابزارها برای زبان‌های مختلف به طرق مختلف و مخصوص زبان مربوطه صورت می‌گیرد. در پژوهش حاضر از میان ابزارهای موجود برای زبان فارسی (مانند «پارسی‌پرداز»، «هضم»، «پرژن–پی[1]»، «پارسی‌وار»، « ویراستیار»، « نگار»، « وارسیگر وفا»، « ویراسباز»، « به‌نویس» « پرشین یوتیلز[2]») از مجموعه ابزارهای موجود در کتابخانه «هضم» استفاده شد. هضم، یک کتابخانه است که به عنوان یک مجموعه ابزار پردازشی پایه به زبان‌های پایتون، سی‌شارپ و جاوا نوشته شده است. فعالیت‌های تعریف شده در این کتابخانه عبارت است از هنجارسازی داده‌ها، واحدسازی در سطح واژه و جمله، بن‌واژه‌سازی، برچسب‌دهی مقولات دستوری، تجزیه سطحی نحوی و تجزیه نحوی وابستگی. از این ابزار پیش از این در پژوهش‌های دیگر استفاده شده بود و در دسترس بود. از کدهای ابزار «هضم»، کد واحدسازی در پیکره استفاده شده است. 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. در حقیقت، واحدسازی، تکه‌تکه‌کردن متن به قسمت‌های کوچکی به نام واحد است. واحدسازی در سطح واژه رخ می‌دهد و واحدهای استخراج شده به عنوان ورودی پیمانه‌های دیگر مانند ریشه‌یاب، برچسب‌گذاری و غیره استفاده می‌شوند.

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

---

[1] Persianp
[2] Persianutils

مختلف بودند و حرف «ه» بصورت کدهای افزوده در مرحله هنجارسازی به‌کار گرفته شدند. برای یکدست کردن انواع «ی» و «ک» و «کسره اضافه روی «ه»/ «ه» در حالت اضافی («ۀ»)، از دستورهای TSQL در برنامه پایگاه داده SQL Server نیز استفاده شده است.

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

روش انتخاب نمونه بصورت تصادفی انتخاب گردید وسپس از بین دانش آموزان این دبیرستان انتخاب شدند و به دو گروه آزمایش کنترل تقسیم شدند

در این مرحله سعی شد تا حد امکان اصلاحات نگارشی به صورت دستی یا به صورت 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. به عنوان نمونه، واژه‌ای مانند «اطلاعات» بدون فاصله در ابتدا و انتهای واژه در پرونده موجود بود، دستور داده‌شده این بود که این صورت نوشتاری را با واژۀ «اطلاعات« به علاوه 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 رفع این مشکل از دستور دیگری استفاده شد، مانند اینکه اگر « اطلاعات ی» در متن وجود دارد، آن را تبدیل به «اطلاعاتی« کند. در برخی موارد این دستورها مشکلاتی دیگر را ایجاد کرده بود که اصلاح ماشینی آن امکان‌پذیر نبود و در صورت برخورد با مشکل نگارشی حاصل از اعمال دستور اصلاحی در متن، باید واژه‌ها تک‌تک یا گروهی اصلاح می‌شدند. استفاده از دستورهای داده‌شده در بسیاری موارد موفقیت‌آمیز بوده است، مانند نمونه‌های جدول ۳-۲ که ابتدا صورت نوشتاری موجود در پرونده آورده شده است و پس از آن صورت‌های اصلاح‌شده آورده شده است:

جدول ۳-۲: نمونه‌ای از اصلاح ماشینی عبارات به هم چسبیده

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

علاوه بر اصلاح ماشینی و پس از آن، کنترل دستی به صورتی که توضیح داده شد، از فهرست وندها و واژه‌بست‌های فارسی نیز استفاده شد. این فهرست برگرفته از مجموعه ۱۲ مقالهٔ دکتر علی‌اشرف صادقی تحت عنوان «*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 (Gahtre, 2007) (در بحث تصریف) است. پیش از استفاده از اطلاعات ساختواژی موجود در این منابع، مجدداً میزان استفاده 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 بر اساس آن بخواهیم یکدست‌سازی را انجام دهیم (به عنوان مثال 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 ۳-۳ از فهرست وندها در منابع ذکرشده استخراج شد و اطلاعات موجود در این جدول برای پیش‌پردازش متون پیکره مورد استفاده قرار گرفت.

جدول ۳-۳: فهرست پیشوندها و پسوندها

پیشوندهایی که باید با نیم‌فاصله به واژهٔ بعد از خود متصل شوند	مثال
می- / نمی-/ بی-/ فرا-/ فرو-	می‌گفت/ می‌توان/ نمی‌گفت/نمی‌توان/ 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/ سگ-دانی/ریاضیدان/ کشتزار / سنگ‌لاخ/ نگهبان/ آهنگر/ وحشی‌گری/ 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/ گوش-نواز/ 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/ کینه-توز/ خوش‌گوار/ خون‌آشام/ دیوانه‌وار/ سازمان‌دهی

برنامه‌ای در پایتون نوشته شد که بر اساس آن اگر وندها جدا از ستاک باشند، با نیم‌فاصله به ستاک متصل شوند. به عنوان نمونه، هر جا تکواژ «‌–ها»، «‌–های» و «‌–هایی» جدا از تکواژ قبل باشد و یک واژهٔ مجزا 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 قبل باشد و یک واژهٔ مجزا 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 متون تخصصی استخراج‌شده مورد توجه قرار گرفته‌اند و برای آن برنامه مجزا 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 ۳-۳ 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؛ در حقیقت، ۳۲ مورد به صورت ماشینی اصلاح شد. بنابراین، ۵۶٪ به صورت ماشینی و بقیه موارد به صورت دستی اصلاح شد. برای محاسبه F1 مراحل زیر دنبال شد. دقت[1] ۰.۵۶۱۴۰۳۵۰۸۸، فراخوان[2] ۰.۰۵۳۱۵۶۱۴۶۲ و F1 ۰.۰۹۷۱۱۶۸۴۳۷۰۲۵۷۹۶۶ محاسبه شده است.

```
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.
```

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

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

۳-۳- برچسب‌گذاری اجزای کلام (POS tagging)

با توجه به کاربرد برچسب اجزای کلام که در حقیقت، خوراک بسیاری از فرایندهای نشانه‌گذاری مانند بن‌واژه‌سازی، تقطیع نحوی[1]، نشانه‌گذاری معنایی و غیره است، در پردازش متن، تصمیم گرفته شد این نوع برچسب نیز به پیکره اضافه شود. در این راستا، 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 به صورت دستی کنترل شوند. برای برچسب‌گذاری متون انگلیسی (چکیده‌های انگلیسی) نیز از برنامهٔ NLTK2 استفاده شد. کتابخانه NLTK مجموعه‌ای است که شامل

[1] syntactic parsing
[2] Natural Language Toolkit

کتابخانه‌ها و برنامه‌هایی برای پردازش زبان‌های آماری است. در حقیقت، یک کتابخانه به زبان پایتون است که اولین‌بار در سال ۲۰۰۱ منتشر شد و کارهای بسیاری مانند دسته‌بندی متون، واحدسازی/جداسازی، ریشه‌یابی، برچسب‌گذاری، تجزیه نحوی و سایر وظایف مربوط به تحلیل معنایی را پوشش می‌دهد . در فرایند ساخت پیکرهٔ مقایسه‌ای از چکیده-های فارسی و انگلیسی پایان‌نامه‌ها و رساله‌های ثبت‌شده در ایرانداک، ابتدا قرار بود از سیستم برچسب‌گذاری لنکس‌باکس[1] استفاده شود که در دانشگاه لنکستر توسعه داده شده است. اما از آنجائیکه امکان واردکردن همهٔ رکوردها (پرونده‌ها) به‌دنباله هم و به طور همزمان، تفکیک آن‌ها به منظور 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 شده است. این فهرست با مثال در جدول ۳-۴ آورده شده است:

جدول ۳-٤: فهرست برچسب‌های داده‌های فارسی

POS tag	Tag Name	مثال	آمار برچسب در داده‌های فارسی پیکره پارسا
N	Noun	کشاورز، خانه، کتاب	17383371
PREP (P)	Preposition	از، در، برای	10276735
PUNC	Punctuation	نقطه، ویرگول، علامت سوال	4363936
AJ	Adjective	زیبا، اجتماعی، بزرگ	7617693
V	Verb	می‌تواند، خورد، شمرد	6283020
CON	Conjunction	که	6731744
NUM	Number	۱۲، ۵۰،۹۰	3380242
PRO	Pronoun	من، تو، ایشان	824305
DET	Determiner	این، آن، هر	1733948
ADV	Adverb	یقیناً، خوب، بسیار	682723
POSTP	Postposition	را	620247

[1] LancsBox

RES	Residual	هر برچسبی غیر از برچسب‌های اصلی	1563942
CL	Classifier	نوع، دست، چنین	119125
INT	Interjection	ای، یا	685

نکته: در فهرستی که در جدول ۳-۴ آورده شده است، کسره اضافه نمایش داده نشده است؛ «هضم» به منظور نمایش کسره اضافه، هرگاه پس از واژه‌ای، کسرهٔ اضافه وجود داشته، کنار برچسب، e را اضافه می‌کند. به عنوان نمونه، برچسب عبارت «شناسایی ساختار محتوایی مطالعات» به صورت: شناسایی (Ne) ساختار (Ne) محتوایی (ADJe) مطالعات (N) نشانه‌گذاری می‌شود. آمار برچسب‌های همراه با کسره اضافه به صورت زیر است:

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

کنترل دستی برچسب‌ها به این صورت انجام شد که فهرستی از واژه‌های پیکره به صورت 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 صحت خوب و قابل‌قبولی انجام شده است. البته علیرغم کنترل دستی برچسب‌ها، کماکان خطاهایی ممکن است در برچسب-گذاری مشاهده شود. اما این خطاها بسیار کم و قابل چشم‌پوشی است. به عنوان نمونه، در بریده‌ای از متن برچسب‌گذاری‌شده، تنها چهار خطای برچسب‌گذاری مشاهده می‌شود.

شکل ۳-۲: نمونه‌ای از متن فارسی برچسب‌گذاری‌شده

همانگونه که قبلاً ذکر شد، برای برچسب‌گذاری داده‌های چکیده‌های انگلیسی از ابزار NLTK استفاده شد. فهرست برچسب‌ها که در سامانه مربوط به این ابزار آمده است در جدول ۳-۵ ارائه شده است.

جدول ۳-۵: فهرست برچسب‌های متون 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 ظهور فناوری‌های رایانه‌ای و تولید حجم بسیار بزرگی از متون به زبان‌های گوناگون، منابع پیکره‌ای عظیمی برای پژوهشگران مشتاق به ساخت پیکره فراهم کرده است. تعداد پیکره‌های تخصصی که برای اهداف خاص و پردازش‌های خاص زبانی استفاده می‌شود، به اندازه 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 ثبت شده در ایرانداک است را توضیح می‌دهد. در این پژوهش، چکیده‌ای از تجربیات پژوهشگران این حوزه دربارهٔ ساخت پیکره‌های تخصصی و مقایسه‌ای و چالش-هایی که هنگام ساخت پیکره‌های متنی/ نوشتاری با آن‌ها مواجه شده‌اند، ارائه شده است. برای ساخت این پیکره مقایسه‌ای تخصصی، ابتدا فرایند نمونه‌گیری، سپس، هنجارسازی و واحدسازی متون فارسی پیکره انجام شده است و در نهایت برچسب‌گذاری متون فارسی و انگلیسی (POS) انجام شده و برچسب‌های فارسی کنترل شده‌اند.

پیکره ساخته شده 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 به همراه 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 به همراه 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 استفاده در پردازش‌های زبانی است. در این راستا می‌توان گفت پیکرهٔ پارسا غنی است، کمتر پیکره-ای را می‌توان یافت که شامل این تعداد واژگان باشد که حوزه‌های گوناگون تخصصی را پوشش دهد. این پیکره شامل متون 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. علاوه بر تعداد بالای واژگان در این پیکره، برچسب اجزای کلام نیز به داده‌های این پیکره اضافه شده است که یکی از پرکاربردترین نوع برچسب‌گذاری محسوب می‌شود. این نوع برچسب‌دهی، عملی کاربردی در بسیاری از حوزه‌های پیشرفته‌تر پردازش زبان طبیعی از جمله ترجمهٔ ماشینی،

خطایاب، تبدیل متن به گفتار، بازیابی اطلاعات، موتورهای جستجو و کمک به مدل‌های آماری است. این پیکره می‌تواند به عنوان یک مرجع تخصصی برای اهداف پردازش زبان طبیعی، به ویژه در ترجمه ماشینی مورد استفاده قرار گیرد. مهم‌ترین چالشی که در ساخت این پیکره وجود داشت، هنجارسازی و واحدسازی داده‌های فارسی پیکره بود که سعی شد با ابزارهای مربوطه، برنامه نویسی ماشینی و اصلاح نگارشی به صورت دستی، تا حد امکان این مرحله با صحت خوبی انجام شود و سپس، داده‌ها برچسب‌گذاری شوند. در مورد داده‌های انگلیسی این مشکلات وجود نداشت؛ چون در انگلیسی مشکلات مربوط فاصله و نیم‌فاصله، نحوه اتصال وندها به ستاک و غیره وجود ندارد. بنابراین، نیازی نبود متون انگلیسی از نظر هنجارسازی و واحدسازی چک 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.

سپاسگزاری: 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 شده است.

فهرست منابع فارسی

امرایی، علیرضا.، حسابی، اکبر. 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: ۹۷-۶۵.

دشتبانی، شکوفه. منصوری‌زاده، محرم.، و نصیری، محمد. (۱۳۹۳). پیکرهٔ متنی تطبیقی فارسی-انگلیسی حوزهٔ تخصصی فاوا. نشریهٔ پژوهش‌های زبان‌شناسی تطبیقی. سال چهارم. شماره ۸. پاییز و زمستان ۱۳۹۳. ۱۴۱-۱۲۱.

صادقی، علی‌اشرف. (۱۳۷۰-۱۳۷۲). «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱۲-۱)». تهران: نشر دانش. شمارهٔ ۶۴-۸۰.

علایی، الهام.، پاک‌نیت، نصراله.، حجت‌پناه، علی‌اصغر.، زالی، مجتبی.، و آقالویی آغمیونی، محمدهادی.(۱۴۰۰). *ساخت پیکرهٔ متنی از مقاله‌های پژوهش‌نامهٔ پردازش و مدیریت اطلاعات*. پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک).

قطره، فریبا. (۱۳۸۶). «مشخصه‌های تصریفی در زبان فارسی امروز». *دستور*. شماره ۳. ۵۲-۸۱.

قیومی، مسعود. (۱۴۰۱). «پیش‌پردازش و ابزارهای پایه». در *پردازش و متن گفتار فارسی: مروری بر مبانی نظری و آخرین یافته‌های پژوهشی*. زیر نظر مهرنوش شمس‌فرد و محمود بی‌جن‌خان. ۱۱۳-۸۶. سازمان مطالعه و تدوین کتب دانشگاهی در علوم اسلامی و انسانی (سمت). پژوهشکده تحقیق و توسعه علوم انسانی.

کشانی، خسرو. (۱۳۷۱). *اشتقاق پسوندی در زبان فارسی امروز*. تهران: مرکز نشر دانشگاهی.

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

لازار، ژیلبر. (۱۳۸۹). *دستور زبان فارسی معاصر*. ترجمهٔ مهستی بحرینی و توضیحات و حواشی هرمز میلانیان. تهران: انتشارات هرمس. چاپ دوم.

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

فهرست منابع لاتین

Alayiaboozar, E., & Hojjatpanah, A (2022). Steps for creating two Persian specialized corpora. International Journal of Information Science and Management (IJISM). Volume 20, Issue 4, Pages 231-243.

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). [in Persian].

Asghari, H., Khoshnava, Kh., Fatemi, O., & Faili, H. (2015). Developing bilingual plagiarism detection corpus using sentence aligned parallel corpus. *Conference: Notebook for PAN at CLEF 2015. CLEF (working Notes).*

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

Beloso, B. S. (2015). Designing, describing and compiling a corpus of English for architecture. *7th International conference on corpus linguistics: current work in corpus linguistics: working with traditionally-conceived corpora and beyong (CILC)*. Procedia-social and behavioral sciences 198. 459-464. Elseveir.

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. Springer.

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.

Kokabi, A., et al. (2023, Oct-5).-Persian NLP Toolkit.- github.- available on Oct-16 at https://github.com/roshan-research/hazm

Dashtbani, Sh., Mansoorizade, M., & Nasiri, M. (2014). English-Persian comparable textual corpus in FAVA domain. Comparative linguistic research. 4 (8). Pp 121-141. [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). Pp 65-97. [in Persian]

Ghatre, F. (2007). Inflectional features in modern Persian. Dastoor. No.3. pp 52-81. [in Persian]

Ghayoomi, M., Momtazi, S., & Bijankhan, M. (2010). A Study of Corpus Development for Persian. *International Journal of Asian Language Processing.* Vol 20. No 1. 17-34.

Ghayoomi, M. (2022). Preprocessing and basic tools. *Text and speech processing for the Persian language: the state of art and a brief review of the theoretical foundations.* SAMT. 86-113.[in Persian]

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

Kenning, M. M. (2010). *The Routledge Handbook of Corpus Linguistics: What are parallel and comparable corpora and how can we use them*. Edited by: Anne O'Keeffe , Michael McCarthy.

Keshani, Kh. (1992). Derivation suffix in modern Persian. Markaz Nashr Daneshgahi. [in Persian].

Koltunski, E. L. (2013). VARTRA: A comparable corpus for analysis of translation variation. In *Proceedings of the 6th workshop on building and using comparable corpora*. Pp. 77-86. Association for computational linguistics.

Kouhestani, M. (2010). *Studying written errors in Persian weblogs and their linguistic nature*. M.A. thesis. University of Tehran. [in Persian]
Lazard, G. (2010). Persian Grammar. Hermes. [in Persian]

Mohammadi, R. (2012). Building Persian-English comparable corpus and extracting parallel sentences. M.A thesis. University of Alzahra. [in Persian]

Sadeghi, A. A. (1991-1993). Word formation methods in Persian. Danesh publication. [in Persian]

Sinclair, J. (2004). *Developing Linguistic Corpora: a Guide to Good Practice. Chapter 1: Corpus and Text — Basic Principles. Edited by Martin Wynne .ahds.literature, languages and linguistics.* The Oxford Text Archive.

Building a specialized comparable corpus: PARSA

Elham Alayiaboozar[1]
Aliasghar Hojjatpanah[2]

Received: 2023/09/17
Accepted: 2023/12/02

Introduction

Based on the language used in their constituent texts, the corpora are divided into monolingual, bilingual and multilingual. A comparable corpus is a bilingual or multilingual corpus that includes similar texts in the same subject areas; In other words, comparable corpus is a collection of documents in two different languages that cover similar topics. Comparable corpus can be made from general texts, which provide various possibilities for discourse analysis, pragmatics, analysis of text genres, and sociolinguistics; Examples of such corpora could include a collection of encyclopedia entries or literary texts of a certain time period. But the most common types of comparative corpora that have many audiences are those that are related to specialized fields and have a high density of vocabulary and technical terms. Such corpus is called specialized comparable corpus. In this study a specialized comparable corpus was built from the Persian and English abstracts of theses and dissertations registered in IranDoc. The corpus is named PARSA.

Method

The process of building PARSA is as follows: sampling, preprocessing (normalization and tokenization), POS tagging and checking the Persian tags. In sampling process 295000 Persian abstract with their English version were chosen, which include texts in different fields of science (humanities, social sciences, art, engineering and the like). Preprocessing may include two main

[1] Assistant Professor, Iranian Research Institute for Information Science and Technology (IranDoc); alayi@irandoc.ac.ir
[2] Iranian Research Institute for Information Science and Technology (IranDoc); hojjatpanah@irandoc.ac.ir

processes: *Normalization* and *Tokenization*. Before annotation, it is essential to check the homogeneity of the text units; this process is called *Normalization.* Normalization process turns the text into a machine-readable one. In normalization process, some parts were automatically normalized, using "HAZM" tool. For Unicode homogeneity of some characters whose Unicode is different in Persian and Arabic (<ی>, <ک> and *Ezafe construction*1 over <ۀ>/je/), TSQL was followed in the SQL Server database, and Arabic Unicode of mentioned letters were replaced by the Persian ones. *Tokenization* is a way of separating a piece of text into smaller units called tokens. Here, tokens can be either word, characters, or subwords. For space homogeneity in the *Tokenization* process, the morphological information (regarding Persian inflectional and derivational affixes) in Sadeghi (1991-1993), Keshani (1992), Lazar (2010) and Ghatre (2007) was used. The output of using "HAZM" tool and Persian morphological information was checked; lots of words were not separated and couldn`t be read. So, the output file was checked manually to correct the wrong spell and separate words manually or using search and replace tools.

An example:

۱) مساحت پهنه‌های متفاوت خطرات مورد مطالعه مشخص گردید

After checking:

۲) مساحت پهنه‌های متفاوت خطرات مورد مطالعه مشخص گردید

For POS tagging of Persian texts, HAZM tool was used and the tags were controlled. For English texts, NLTK was used.

Result

In this study a specialized comparable corpus was built (PARSA), which include Persian and English abstracts of theses and dissertations registered in IranDoc. To build such corpus the following process was followed:

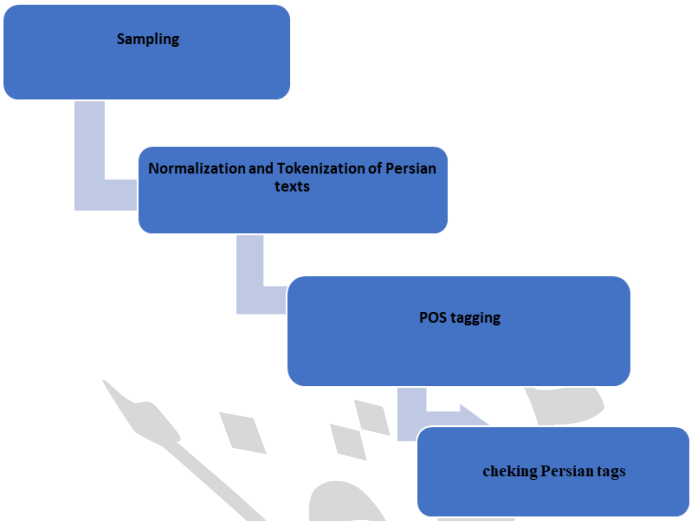

Fig 1: The process of building PARSA corpus

Key words: specialized corpus; comparable corpus; normalization; tokenization; tagging