گزارش نهایی طرح پژوهشی

ساخت انسانی پیکره موازی دو زبانه (فارسی-انگلیسی) عناوین مقالات مجلات رتبه‌دار نمایه‌شده در مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری

مجری:

دکتر محمد رضا فلاحتی قدیمی فومنی

گروه پژوهشی زبانشناسی رایانه‌ای

مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری

خرداد ۱۳۹۷

بسم الله الرحمن الرحيم

تشکر و قدردانی

در اینجا لازم می‌دانم از مجموعه‌ای از افراد که به نحوی در به ثمر رسیدن پژوهش حاضر سهیم بوده‌اند تشکر و قدردانی نمایم. در آغاز لازم می‌دانم از ریاست محترم سازمان جناب آقای دکتر محمد جواد دهقانی که با حمایت‌های سازنده خود زمینه‌ساز انجام این پژوهش بوده‌اند تشکر نمایم. سپس، ازجناب آقای دکتر محمد رضا صالحی معاون محترم پژوهش و فناوری سازمان که با استقبال از ایده اولیه این پژوهش، فرایند داوری پیشنهاده را پیگیری نمودند سپاس‌گزاری می‌نمایم. همچنین از سرکار دکتر صفاهیه به خاطر تعامل سازنده در طی فرایند انجام طرح قدردانی می‌نمایم. مدیر محترم گروه پژوهشی زبانشناسی رایانشی نیز که پیشنهاده پژوهش حاضر را در گروه مطرح و پس از تایید اولیه، آن را برای طی فرایندهای بعدی به سازمان ارسال داشتند نیز تشکر می‌نمایم. همچنین از داوران محترم که با صرف وقت فراوان نظرات سازنده خود را در مورد پیشنهاده اولیه این پژوهش و نیز گزارش نهایی ارسال داشتند و یقیناً باعث ارتقاء کیفیت آن شدند سپاس‌گزاری می‌نمایم. در پایان اذعان دارم موفقیت‌های پژوهش حاضر در گرو نظرات سازنده داوران محترم و راهنمایی همکاران بزرگوار بوده و اگر در این فرایند کاستی وجود دارد مسئولیت آن صرفاً برعهده اینجانب است.

دکتر محمد رضا فلاحتی قدیمی فومنی

خرداد ۹۷

فهرست مندرجات

فهرست علایم و نشانه‌ها

ردیف	علامت اختصاری	صورت کامل	معادل فارسی
۱	ACS	Advanced Conditional Sum	نرم‌افزار مشابه‌یاب
۲	ADJ	Adjective	صفت
۳	ADV	Adverb	قید
۴	ART	Article	حرف‌تعریف
۵	CON	Conjunction	حرف‌ربط
۶	DET	Determiner	حرف‌تعیین
۷	N	Noun	اسم
۸	NP	Noun Phrase	گروه اسمی
۹	PFX	Prefix	پیشوند
۱۰	POS	Part-of-Speech	مقوله نحوی
۱۱	PREP	Preposition	حرف‌اضافه
۱۲	PRO	Pronoun	ضمیر
۱۳	SFX	Suffix	پسوند
۱۴	SOV	Subject-Object-Verb	فاعل-مفعول-فعل
۱۵	V	Verb	فعل

چکیده فارسی

زبانشناسی پیکره‌ای یکی از حوزه‌های داغ تحقیقاتی است که توانسته محققان زیادی را به خود جلب کند. ساخت و تولید و همچنین ارزیابی انواع مختلف پیکره اساس این حوزه از علم را تشکیل می‌دهد. بر این اساس هدف از انجام پژوهش حاضر تولید انسانی پیکره موازی دو زبانه فارسی-انگلیسی عناوین مقالات مجلات رتبه‌دار نمایه‌شده در مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری بوده است. برای انجام این پژوهش زبانشناختی-رایانشی با استفاده از روش نمونه‌برداری چندمرحله‌ای تصادفی ۱۰۰۰۰ عنوان مقاله فارسی به همراه ۱۰۰۰۰ عنوان معادل انگلیسی از مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری اخذ گردید. سپس این عناوین در قالب یک فایل ماشین‌خوان اکسل تهیه و به عنوان داده اصلی پژوهش ثبت شد. در مرحله بعد، عناوین فارسی و انگلیسی مطابق ۵ حالت مختلف تقطیع و برابرسازی شد. فرایند الصاق برچسب نحوی به هر واژه پس از این مرحله صورت پذیرفت و سپس با استفاده از نرم‌افزار ACS مدخل های مشابه (در هر چهار متغیر) در هر یک از پنج دسته اطلاعات (حالت‌های اول تا پنجم) در هم ترکیب شده، فراوانی هر مدخل ثبت گردید. سپس اطلاعات پنج دسته مجدداً با ACS تحلیل و صورت‌های مشابه در هم مجدداً ترکیب شد. درپایان پیکره پژوهش حاضر در مجموع با ۲۴۹۰۹ مدخل و ۹۸۰۳۹ رخداد در قالب یک فایل اکسل و همچنین به فرمت پی دی اف و به عنوان ماحصل پژوهش حاضر ارائه شد. در این پیکره حرف‌ربط «و» با فراوانی ۵۰۱۶ مورد، فراوان‌ترین مدخل واژگانی بوده است. پس از آن مدخل «در» و مدخل «بر» به ترتیب با فراوانی ۴۸۸۷ و ۱۷۴۲ در رتبه‌های دوم و سوم قرار گرفتند. همین سه مدخل به تنهایی ۱۱/۸۸ درصد از کل رخدادهای موجود در این پیکره را شامل شد. ضمناً در رفتار نویسندگان در ترجمه واژه‌های فارسی به انگلیسی تنوع گسترده ای مشاهده شد تا آنجا که گاه تا ۲۰ معادل انگلیسی برای یک واژه فارسی به کار رفته است. از این پژوهش می‌توان برای انواع مختلفی از پژوهش‌های فرهنگ‌نگاری، تحلیل زبانی ترجمه و نیز آموزش زبان استفاده کرد.

واژگان کلیدی: زبانشناسی رایانشی، زبانشناسی پیکره‌ای، ساخت پیکره، پیکره‌های تخصصی، برچسب‌دهی نحوی، تنوع معادل‌گزینی.

۱- مقدمه

علم زبانشناسی به عنوان علمی که به تحلیل و بررسی زبان بشری می‌پردازد، از نقطهٔ آغاز در حوزه‌های مختلفی ایفای نقش نموده است. این علم دارای ده‌ها زیرحوزه است که از آن جمله می‌توان به معناشناسی، منظورشناسی، نحو، آواشناسی، واج‌شناسی، گفتمان، تجزیه و تحلیل کلام، جامعه‌شناسی زبان، روان‌شناسی زبان، عصب‌شناسی زبان، زبانشناسی سیاسی، زبانشناسی کلینیکی، زبانشناسی رایانه‌ای و نظیر آن اشاره نمود. از این میان، زبانشناسی رایانشی یکی از حوزه‌های تحقیقاتی علمی و مهندسی محسوب می‌شود که به درک زبان گفتاری و نوشتاری از منظری رایانشی می‌پردازد و می‌کوشد ابزاری را فراهم کند که در تولید و درک زبان موثر است (دایره‌المعارف فلسفه استنفورد، مدخل زبانشناسی رایانشی، پارا ۱، ۲۰۱۴). همچنین گولدواتر (۲۰۱۵) این حوزه از علم را به این صورت تعریف می‌کند: «کاربرد رایانه برای پاسخ دادن به پرسش‌های زبان از طریق تحلیل داده‌های زبان طبیعی». هدف عمده زبانشناسان رایانشی نیز آن است که «به نگارش برنامه‌هایی بپردازند که بتواند تا آنجا که مقدور باشد حجم بیشتری از داده‌های زبان طبیعی را درک یا تولید نماید» (اسپونیا پرات، ۱۹۹۴، ص. ۱۱). «حوزهٔ زبانشناسی رایانشی نیز خود زیرحوزه‌های وسیعی را پوشش می‌دهد که از آن جمله می‌توان به زبانشناسی پیکره‌ای اشاره نمود» (کِندی، ۲۰۱۴، ص. ۸۵). حوزه زبانشناسی رایانشی اگرچه در دهه ۱۹۵۰ و با مطالعات مرتبط با ماشین ترجمه آغاز شد اما توانسته در حوزه‌های مختلف به شکلی گسترده خود را مطرح نموده اثرگذاری مثبت و غیر قابل انکار خود را در زندگی انسان‌ها به اثبات برساند. جورافسکی (۲۰۰۶، ص. ۵۷۸) بر این باور است که امروزه هر آنچه پیرامون خود مشاهده می‌کنیم، نسخه‌ای رایانشی از آن نیز موجود و یا در دست تهیه است، مثل زیست‌شناسی رایانشی، موسیقی‌شناسی رایانشی، باستان‌شناسی رایانشی، زبانشناسی پیکره‌ای و صدها حوزه رایانشی دیگر.

در این میان، زبانشناسی پیکره‌ای شاخه‌ای از زبانشناسی است که در آن از طریق ایجاد پایگاه داده‌ای گسترده‌ای از متون و واژگان رایج یک زبان و بررسی این متون، به مطالعهٔ جنبه‌های گوناگون یک زبان می‌پردازد (کندی، ۲۰۱۴؛ مک‌انری و هاردی، ۲۰۱۲). این شاخه اگرچه در آخرین دهه‌های قرن بیستم ایجاد

شد، توانست در طی همین عمر کوتاه خود به یکی از فعال‌ترین و پرکاربردترین زمینه‌های تحقیقاتی تبدیل شود (عاصی، ۱۳۸۵).

پیکره که شالوده اصلی مطالعات در زبانشناسی پیکره‌ای را تشکیل می‌دهد، در تعریف عام به معنی حجم زیادی از داده‌های زبانی است که بر اساس معیارهای معین و برای هدفی مشخص (نظیر نمایاندن زبان یا گویشی خاص، ...) تهیه می‌شود (آتکینز، کلیِر و آستلر، ۱۹۹۲). بنا بر نظر عاصی (۱۳۸۵) پیکره مجموعه‌ای از متن‌های نوشتاری یا گفتاری است که می‌توان از آن برای توصیف و تحلیل زبان استفاده نمود. پیکره ممکن است به شکل برگه‌های یادداشت باشد یا به شکل پرونده‌های رایانه‌ای شامل متن‌های کامل یا گزیده‌هایی از آنها و حتی بخش‌های پیوسته‌ای از متون یا گزیده‌ای از نقل‌قول‌ها و نکات و فهرست‌های واژگانی ظاهر شود. البته باید توجه داشت که زبان گفتار با زبان نوشتار متفاوت است و طبیعتاً پیکره‌هایی که برای زبان گفتاری یا زبان نوشتاری تولید و تهیه می‌شوند باید به عنوان فعالیت‌هایی متمایز درنظر گرفته شوند و لذا عدم ایجاد تمایز بین زبان گفتاری و نوشتاری توسط برخی از متخصصان رایانه موجه نیست. در این خصوص عاصی (۱۳۸۹، ص. ۲۹) تصریح می‌کند: «هنگامی که کارشناسان رایانه از پردازش زبان طبیعی سخن می‌گویند اغلب مسائلی را عنوان می‌کنند که نشان می‌دهد تمایز روشنی میان خط و زبان قائل نیستند».

پیکره‌ها در زبانشناسی رایانشی و به ویژه در حیطه‌های خط و زبان نقش بسیار مهمی ایفا می‌نمایند (مک‌انری و ویلسون، ۲۰۰۱). پیکره‌ها انواع مختلفی دارند که از آن جمله می‌توان به پیکره‌های تک‌زبانه (برای مثال فارسی)، دوزبانه (برای مثال فارسی-انگلیسی) و چندزبانه (حاوی بیش از دو زبان) اشاره نمود. وقتی بیش از یک زبان داشته باشیم لازم است بین اطلاعات دو زبان در سطحی معین مثلاً «بند»، «جمله» یا «کلمه» نوعی توازی یا برابری ایجاد شود. به عنوان نمونه، در سطح واژه برای هر واژه در زبان «الف» یک واژه معادل و موازی در زبان «ب» در نظر گرفته می‌شود. به این نوع پیکره‌ها پیکره «موازی منطبق» اطلاق می‌گردد. در حقیقت، هدف از پژوهش حاضر نیز ساخت پیکره‌ای از این دست البته به صورت انسانی و در قالب فایل ماشین بوده است.

پیکره‌ها می‌توانند انواع مختلفی از اطلاعات را نیز در خود جای دهند مانند معادل کلمه زبان مبدأ در یک یا چند زبان مقصد، برچسب‌های معنایی، دستوری و نظیر آن. از این میان، برچسب‌های دستوری که غالباً مقوله‌های نحوی را تشکیل می‌دهد، رایج‌ترند (بیکِر، ۲۰۱۲). شایان ذکر است فرایند برچسب‌دهی فرایندی وقت‌گیر، پیچیده و پرهزینه است و معمولاً با کار تیمی بر روی پیکره‌های کوچک میسر می‌گردد. برای گویاتر شدن پیکره و کاربردهای خاص، کدهای متفاوتی نیز به آن اضافه می‌شود. این نوع نشانه‌گذاری از یک سو می‌تواند برای ارتباط دادن بخش‌های یک پیکره به ساختار کلی آن باشد، مانند شمارهٔ سطر، صفحه، فصل و مانند آن و یا بافت زبانی را مشخص کند مانند شرایط تولید زبانی، رسانه و مانند آن. از سوی دیگر، نشانه‌گذاری می‌تواند صرفاً زبانی باشد (عاصی، ۱۳۸۹). طبق نظر بِیکر (۲۰۱۲) انواع مختلفی از برچسب‌ها وجود دارد (مانند برچسب نحوی، معنائی، منظورشناختی، گفتمانی). یکی از شایع‌ترین آنها برچسب‌زنی به کلمات بر اساس مقولهٔ نحوی کلمه (اسم، فعل، صفت، قید و حرف اضافه، ...) است. این نوع تقسیم‌بندی اخیر در پژوهش حاضر مد نظر قرار گرفت. بنابراین موضوع پژوهش حاضر در چارچوب مباحث زبانشناسی رایانشی و در زیر حوزه زبانشناسی پیکره‌ای و در این حوزه در بخش تولید پیکره‌های موازی دستی و به فرمت ماشین‌خوان جای می‌گیرد.

۲- بیان مسئله

امروزه آنچه می‌تواند باعث افزایش و پیشرفت تحقیقات زبانشناسی رایانه‌ای شود ارایه مدل‌ها و نظریه‌های جدید نیست چرا که در این خصوص انواع و اقسام نظریه‌ها، مدل‌ها و روش‌ها وجود دارد (هر چند این گفته نباید به معنی عدم نیاز به پیشرفت در این حوزه‌های زیربنایی قلمداد شود). در حقیقت، مشکل تحقیقات در این حوزه نبود ابزار و پیکره‌های مرتبط و کافی است. پیش‌نیاز بسیاری از پژوهش‌ها در حوزهٔ ترجمه ماشینی نیز وجود همین ابزار و پیکره‌ها است. بسیاری از تحقیقات در حوزه زبانشناسی و برنامه‌ریزی‌های زبانی در صورت وجود پیکره‌های زبانی میسر است (مک‌انری و هاردی، ۲۰۱۲). پیکره‌های متنی خارج از حوزه پردازش زبان طبیعی نیز کاربرد دارند. مثلاً می‌توان کلمات با فراوانی بالا را از پیکره‌های متنی استخراج کرد و برای تسهیل فرایند آموزش زبان در کلاس درس به‌کار برد. در حال حاضر نرم‌افزارهای متعددی نیز وجود دارد که یا تنها و یا

به‌عنوان یکی از قابلیت‌های خود، فراوانی کلمات و انواع اطلاعات آماری واژگانی را استخراج و با اطلاعات دیگر ارایه می‌نمایند. با توجه به گرایش پژوهشگران به سمت حوزه‌هایی چون ماشین ترجمه، خلاصه‌سازی متن و ... و کاربردهای متعدد پیکره‌های زبانی در این خصوص، در پژوهش حاضر تهیه پیکره زبانی به‌عنوان یک منبع داده اولیه برای پژوهش‌های دیگر، مورد بررسی قرار گرفت.

فارسی یکی از زبان‌های هند و اروپایی در شاخه زبان‌های ایرانی جنوب غربی است که در کشورهای متعددی از جمله ایران، افغانستان، تاجیکستان و ازبکستان بدان تکلم می‌شود. شمس‌فرد (۲۰۱۱، ص. ۶۵) بر این باور است که «از نظر پردازش‌ها و تحلیل‌های رایانه‌ای، زبان فارسی چندان مورد توجه و تحقیق قرار نگرفته است». ضمن آنکه اذعان می‌دارد «پیکره‌های موازی یا دوزبانه از مهم‌ترین و ضروری‌ترین ملزومات ساخت ماشین‌های ترجمه‌اند» (ص. ۶۸). او همچنین بر روی فقدان یا کمبود ابزار پردازشی زیربنایی تأکید می‌کند، «ابزاری که بتوان از آنها برای تحلیل واژگانی، تقطیع، استانداردسازی، نرمال‌سازی و ... استفاده نمود» (صص. ۶۷-۶۶). او به مشکلات عمده‌ای نیز اشاره می‌کند که بر سر راه پردازش زبان فارسی در مقایسه با زبان‌های دیگری چون انگلیسی وجود دارد، مثل نوشته‌شدن فارسی از راست به چپ، وجود استثنائات فراوان برای الگوی ترتیب کلمات SOV، ماهیت افعال در فارسی، هسته‌آغازین بودن فارسی، وجود تنوع حروف برای آواهایی خاص، عدم وجود حروف تعریف معین در بیشتر اوقات و ... .

همانند افرادی چون شمس‌فرد (۲۰۱۱)، جباری، بخشایی، محمدزاده ضیابری و خدیری (۲۰۰۹) نیز اعتقاد دارند که برای زبان فارسی منابع و ابزار پردازش زبان طبیعی محدودی وجود دارد. آنها معتقدند برای ایجاد یک پیکره جدید دو رویکرد عمده موجود دارد: ۱) استفاده از ابزار خودکار برای هم‌تراز کردن اسناد و ۲) استفاده از انسان‌های مترجم. آنها در صفحات ۱۸ و ۱۹ پیکره‌های متعددی را با پوشش زبان‌های انگلیسی و فارسی معرفی کردند که از آن جمله می‌توان به Verb-Mobil، News، Ted، Central Asian و Misc اشاره نمود. آنها با درهم‌کرد این پیکره‌ها، پیکره خود را با نام امیرکبیر تولید کردند که حاوی ۷۰۰۹۱۶ خط انگلیسی و ۷۰۰۹۱۶ خط فارسی بود. دستاورد آنها به اذعان خود محققان نخستین پیکرهٔ انگلیسی-فارسی باز-حوزه بوده است.

با توجه به نکات ذکرشده در سطور بالا و اهمیت پیکره‌های زبانی برای انجام انواع پژوهش‌های زبانشناسی رایانه‌ای و تحرک گسترده فرهنگستان علوم، انجمن‌ها، گروه‌های زبانشناسی و رایانه دانشگاه‌ها در این زمینه و نیز اقدامات و کارگاه‌های متعدد مرتبط با این حوزه و نیز با توجه به اهداف ترسیم شده برای گروه پژوهشی زبانشناسی رایانه‌ای مرکز منطقه‌ای و نیز برنامه یک‌ساله گروه و اولویت قائل‌شدن برای پژوهش‌های پیکره‌ای، انجام پژوهش‌حاضر می‌تواند گامی در جهت رفع کمبودهای این حوزه باشد. در ضمن جهت پیکره موازی در این پژوهش از فارسی به انگلیسی است.

۳- اهداف پژوهش

در این پژوهش نیز مانند هر پژوهش دیگری اهدافی ترسیم و دنبال شد. هدف از پژوهش حاضر تولید انسانی یک پیکرهٔ (ماشین‌خوان) موازی دو زبانه فارسی-انگلیسی عناوین مقالات مجلات رتبه‌دار نمایه‌شده در مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری بود. مراد از لفظ انسانی آن بوده که پیرو نظرات داوران محترم، مقرر شد یک پیکرهٔ خام تولید و بدون افزودن امکانات نرم‌افزاری برای استخراج داده‌ها ارائه شود (یعنی پیکرهٔ خام به فرمت ماشین‌خوان). از این رو، محقق پس از تهیه فرمت ماشین‌خوان پیکره عناوین فارسی و انگلیسی، با استفاده از یک نرم‌افزار آماده (Advanced Conditional Sum (ACS، واژه‌های متمایز موجود در پیکره را به‌همراه فراوانی هر واژه مشخص و در قالب فایل اکسل و همچنین نسخه چاپی گزارش نمود. همچنین مقولهٔ نحوی هر واژه توسط محقق و با مشاوره یک زبانشناس دیگر با مدرک دکتری زبانشناسی و با حداقل ده سال سابقه کار مرتبط تعیین و ثبت شد. بنابراین برای رسیدن به هدف کلی پژوهش که «تهیه پیکرهٔ (ماشین‌خوان) موازی دو زبانه فارسی-انگلیسی عناوین مقالات مجلات رتبه‌دار نمایه‌شده در مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری» بوده است، اقدامات زیر مد نظر قرار گرفت:

الف- تهیهٔ فهرست ترازبندی شده جفت جملات (۱۰۰۰۰ جفت عنوان مقاله)

ب- تقطیع عناوین مقالات فارسی و معادل انگلیسی آنها

ج- تعیین واژه‌های متمایز (type) و فراوانی هر یک (token) و محاسبه نسبت این دو

د- تهیه پیکره حاوی واژه‌های فارسی، معادل انگلیسی آنها، مقوله نحوی و فراوانی رخداد هر واژه در قالب یک واژه‌نامهٔ دو زبانه با جهت فارسی به انگلیسی در قالب فایل اکسل و همچنین به فرمت پی دی اف به عنوان ماحصل پژوهش حاضر.

ه- معرفی فراوان‌ترین واژه‌های موجود در پیکره حاضر

۴- ضرورت انجام پژوهش

انجام پژوهش حاضر از جهات مختلف ضرورت پیدا می‌کند که به برخی از این موارد در این بخش اشاره می‌شود. امروزه در حوزه‌های مختلف به‌خصوص در حوزه ترجمه ماشینی کارهایی که توسعه بیشتری پیدا کرده‌اند عمدتاً بر مبنای پیکره‌ها کار می‌کنند. ماشین ترجمه گوگل‌ترانسلیت نمونه‌ای از این گونه فعالیت‌هاست. از این رو تولید پیکره‌ها می‌تواند به عنوان مقدمه‌ای بر انواع دیگر پژوهش‌ها در گروه زبانشناسی رایانشی مرکز منطقه‌ای و نیز در سایر سازمان‌های مرتبط به حساب آید. این نکته به‌خصوص از آن جهت حائز اهمیت است که دسترسی به منبع اصلی اطلاعات پیکره‌ها غالباً میسر نیست و از این‌رو با مهیا کردن دسترسی به اطلاعات این پژوهش برای علاقه‌مندان می‌توان زمینه انواع دیگری از پژوهش‌ها را نیز فراهم کرد.

انواع مختلفی از پیکره‌ها وجود دارند که برای زبان فارسی اغلب آنها از داده‌های عمومی نظیر محتوای وب و یا روزنامه‌ها و سایر متون عمومی گرفته شده‌اند که معمولاً با ضریب اصوات مزاحم (حجم حذف و درج در زمان ترجمه) بالایی همراه‌اند (لی، وو و ویجی-شانکر، ۲۰۱۷). کاهش اصوات مزاحم در این نوع پژوهش‌ها از اهمیت زیادی برخوردار است و از این جهت که در این پژوهش بر روی عناوین مقالات علمی کار شد که از ضریب اصوات مزاحم پایینی برخوردار است (و بنابراین واژه‌های بیشتری با هم تطبیق می شود)، فعالیت حاضر می‌تواند فعالیتی مورد نیاز محسوب گردد.

نکتهٔ دیگر به نوع محتوای پیکره حاضر مربوط است. با توجه به اینکه از عناوین فارسی و معادل انگلیسی مقالات رتبه‌دار از حوزه‌های مختلف موضوعی استفاده شد، می‌توانیم پیکره حاضر را نوعی پیکره تخصصی محسوب نماییم که در مقابل پیکره‌های عمومی‌ای قرار می‌گیرد که محتوای آنها از روزنامه‌ها و ... استخراج

می‌شود. در حقیقت، یکی از دلایل کمبود پیکرهای تخصصی عدم دسترسی به اطلاعات مربوطه یا دشوار و پرهزینه بودن تهیه منبع اصلی داده‌های آنها می‌باشد که خوشبختانه امکان تهیه این نوع اطلاعات از پایگاه داده مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری مهیا بود.

در مقایسه با زبان‌هائی چون انگلیسی، برای زبان فارسی مجموعه محدودی از پژوهش‌های پیکره‌ای صورت گرفته است. با توجه به این مهم و اینکه انجام پژوهش‌های پیکره‌ای مقدمه‌ای برای انواع پژوهش‌های دیگر محسوب می‌شود، انجام پژوهش حاضر ضروری می‌نماید.

مطلب دیگر به نوع زبان در پیکرهٔ حاضر مربوط می‌شود. غالباً در پیکره‌ها از گونهٔ رسمی و معیار زبان استفاده می‌شود که مخاطبان وسیع‌تری دارند. از این جهت که پیکرهٔ حاضر از واژگان مکتوب عناوین مقالات رتبه‌دار فارسی (و معادل انگلیسی آنها) گرفته شده مخاطبانی وسیع تر و واژگان علمی را پوشش می‌دهد.

در ایران فرهنگستان زبان و ادب فارسی بر روی زبان فارسی پژوهش‌های مختلفی را به انجام رسانده یا در دست انجام دارد. یکی از اولویت‌های پژوهشی در فرهنگستان تولید پیکره‌هاست که این تأکید به دلیل اهمیتی است که وجود پیکره‌ها می‌تواند در انجام پژوهش‌های دیگر داشته باشد. بنابراین، موضوع پژوهش حاضر را می‌توان در راستای اهداف کلان فرهنگستان زبان و ادب فارسی نیز ارزیابی کرد.

در گروه زبانشناسی رایانه‌ای مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری اهداف متعددی دنبال می‌شود که از جمله آنها می‌توان به تولید پیکره‌های مختلف اشاره کرد. پژوهش حاضر نخستین اقدام از سوی گروه در این خصوص است که از این جهت گامی در راستای تحقق اهداف گروه محسوب شده بر ضرورت انجام آن می‌افزاید.

۵- پیشینه پژوهش

زبانشناسی پیکره‌ای با وجود اینکه که حوزه‌ای جوان محسوب می‌شود (این شاخه در آخرین دهه‌های قرن بیستم ایجاد شد)، توانست در طی همین عمر کوتاه خود به یکی از فعال‌ترین و پرکاربردترین زمینه‌های تحقیقاتی تبدیل شود (عاصی، ۱۳۸۵). با توجه به کاربردی بودن این حوزه، محققان زیادی جذب این حوزه شده و پژوهش‌های متعددی را به‌انجام رسانده‌اند. با توجه به تعدد این نوع فعالیت‌ها به ویژه در دانشگاه‌های

بزرگی چون دانشگاه تهران، دانشگاه شهید بهشتی، دانشگاه فردوسی مشهد، دانشگاه شهید چمران اهواز، ... و فرهنگستان زبان و ادب فارسی، انجمن زبانشناسی ایران، پژوهشگاه علوم انسانی و مطالعات فرهنگی، دبیرخانهٔ شورای عالی انقلاب فرهنگی، وبگاه دادگان (۱۳۹۴) و ... در این قسمت به اختصار به نمونه‌ای از پژوهش‌های انجام شده در خصوص موضوع پژوهش حاضر پرداخته می‌شود.

پژوهش‌های این حوزه را می‌توان به چند حوزه کلی تقسیم کرد. برخی از محققان از اصول موجود در زبانشناسی پیکره‌ای برای انجام پژوهش هایی در حوزه ترجمه استفاده کرده‌اند. برای نمونه، عیار (۱۳۸۹) در پایان‌نامه کارشناسی ارشد خود از یک پیکره خودساخته حاوی ۲۱۵ رمان انگلیسی برای ارزیابی دو ترجمه انگلیسی از بوف کور استفاده کرد. در این پژوهش محقق نشان داد که با استفاده از پیکره و بررسی فراوانی رخداد واژه‌ها و نیز پراکنش و هم‌آیی آنها می‌توان به امکانات بهتری برای ارزیابی عینی ترجمه‌های صورت گرفته دست یافت. کشتکار (۱۳۹۱) نیز ضمن تولید یک پیکره دوزبانه موازی انگلیسی-فارسی کاربرد آن را در سامانه حافظه ترجمه بررسی کرد. او در این پژوهش بر این نکته تاکید می‌کند که غالباً در برونداد سیستم‌های ترجمه، خطاهای متعددی وجود دارد و در ترجمه‌های خودکار امکان دخالت انسان تا پیش از تولید برونداد وجود ندارد. او تصریح می‌کند که ابزاری چون سامانه‌های حافظه ترجمه با بهره‌گیری از پیکره‌های موازی هم‌ترازشده می‌توانند در رفع یا کاهش این خطاها پیش از مرحله تولید برونداد، نقشی بایسته را ایفا نمایند. او با افزودن این پیکره به سامانه حافظه ترجمه و ارزیابی این اقدام با استفاده از نمرات «فراخوانی» و «دقت» گزارش داد که استفاده از فنون خودکار همترازسازی، شانس جستجوی برابرنهاده برای زیرتوالی‌های جملات را تا حد زیادی افزایش می‌دهد و از این رو استفاده از این نوع پیکره‌ها را در سامانه‌های حافظه ترجمه توصیه کرد. همچنین، محمدی (۱۳۸۹) تاثیر استفاده از پیکره موازی را بر کیفیت ترجمه بررسی کرد. او ضمن انتخاب ۲۰ دانشجوی مترجمی زبان انگلیسی آنها را به صورت تصادفی به دو گروه تقسیم نمود. گروه اول به وسیله لغت‌نامه دوزبانه و گروه دوم به وسیله لغت‌نامه دوزبانه و یک پیکره موازی در یک آزمون ترجمه شرکت کردند. سپس کیفیت ترجمه بر اساس نمرات اخذ شده توسط دانشجویان مورد بررسی و مقایسه قرار گرفت. نتایج ارزیابی نمرات دانشجویان دو گروه نشان داد گروه دوم که علاوه بر واژه‌نامه دوزبانه از یک پیکره موازی هم استفاده

کرده بود در مقایسه با گروه اول به طرز معناداری عملکرد بهتری از خود نشان داد و پیشرفت این گروه ۵۵ درصد بوده است. او در پایان نتیجه‌گیری کرد که پیکره‌های موازی ابزاری بسیار مفید و کارا در ارتقاء کیفیت ترجمه محسوب می‌شوند اما سطح زبانی بر میزان اثرگذاری پیکره بر ترجمه دانشجویان هیچ گونه تاثیری ندارد.

برخی از محققان نیز از پیکره‌ها برای امر آموزش زبان شامل زبان مادری، دوم، خارجی و نظیر آن استفاده کردند. برای مثال، توکل (۱۳۹۶) پیکره‌ای از افعال خوانداری کتاب‌های پایه اول ابتدایی تا پایه ۱۲ دبیرستان را تولید نمود و در کنار آن فهرست افعال متمایز هر سال را مشخص کرد. برای انجام این کار ابتدا افعال، از کتاب فارسی خوانداری هر سال تحصیلی استخراج و در قالب یک فایل ذخیره شد (لازم به ذکر است که برای هر سال تحصیلی دو کتاب فارسی وجود دارد که یکی خوانداری است که افعال آن در پژوهش حاضر استخراج و تحلیل گردید). سپس، با استفاده از توابع اکسل افعال هر سال به همراه تحلیل بسامدی و نیز افعال متمایز هر سال شناسایی و گزارش شد. یا وحدت‌زاده (۱۳۹۱) فعل مرکب را در کتاب‌های درسی زبان فارسی در ایران با روشی پیکره‌بنیاد بررسی کرد. او ضمن بهره‌گیری از یک رویکرد آماری، ۱۴ نوع فعل مرکب را آنگونه که در کتاب‌های آموزش و پرورش معرفی شده بود مبنای استخراج داده‌ها قرار داد. او پس از تحلیل، افعال مرکب را به سه دسته تقسیم کرد که عبارت بود از: فعل گروهی گسترش‌ناپذیر (که در آنها فاصله جزء غیر فعلی از جزء فعلی صفر بود)، فعل گروهی گسترش‌پذیر (که در آنها فاصله جزء غیر فعلی و فعلی بین صفر تا یک محاسبه شد) و سرانجام توالی آزاد با فاصله جزء غیر فعلی و فعلی بزرگتر از یک. به همین ترتیب، صفری (۱۳۹۱) به تهیه پیکره زبان‌آموز فارسی پرداخت. او ضمن تهیه پیکره زبان‌آموز خود، تاکید کرد از این پیکره‌ها می‌توان برای تهیه برنامه درسی و نیز محتوای آموزشی مناسب استفاده کرد. او تولید این پیکره را یکی از نیازهای مهم در حوزه آموزش زبان فارسی عنوان می‌کند. او برای تولید این پیکره از انشاء فارسی فارسی‌آموزان خارجیِ مشغول به تحصیل در مدرسه المهدی قم استفاده کرد. ۱۷۰ انشاء در مرحله اول گردآوری و سپس ۱۵۰ مورد از آنها به عنوان داده نهایی تحقیق برگزیده شد. پس از تایپ این تعداد انشا کلمات، برچسب‌دهی شده و خطاهای موجود بر حسب نوع آن ثبت گردید. محقق در پایان ذکر می‌کند که این پیکره قادر است ۱۰ نوع تحلیل آماری و زبانی

را از انشاء دانش‌آموزان ارائه نماید که تعیین کمیت خطاها و نوع خطا یکی از اطلاعات بسیار مهم ارایه شده توسط این پیکره محسوب می‌شود

برخی از محققان بر روی تهیه پیکره از متون ادبی و به ویژه متون ادبی تاریخی کار کرده‌اند. برای مثال، قندی (۱۳۹۳) طرحی را برای تولید پیکره برای متون نثر و تاریخی زبان فارسی از قرن پنجم تا هفتم هجری ارایه کرد. محقق تلاش خود را اولین پیکره تاریخی برای نثر فارسی قرن پنجم تا هفتم هجری می‌نامد. برای انجام این کار، محقق ابتدا گزیده‌ای از متون این دوره را بر اساس معیارهای مشخص تعیین و تایپ کرد. سپس همه متون، وارد پایگاه داده زبان فارسی شده و نمایه گردید تا امکان استفاده از اطلاعات گردآوری شده برای مخاطبان وجود داشته باشد. این پیکره خام نیست بلکه برچسب‌های نحوی، شناسنامه‌ای و آوایی را نیز شامل می گردد.

مطالعات پیکره‌ای، با تهیه واژه‌نامه‌ها و فرهنگ‌نگاری آغاز شد و هنوز هم در این خصوص کارهای متعددی به انجام می‌رسد. برای مثال، جهانگردی (۱۳۸۹) طرحی را برای استفاده از پیکره‌های زبانی در فرهنگ‌نگاری معرفی کرد. هدف او از یک طرف اثبات ضرورت طراحی و ایجاد پیکره برای تدوین فرهنگ‌ها و لغت‌نامه بود و از طرف دیگر ارایه الگویی برای طراحی و ساخت این نوع پیکره ها. او سه رویکرد مختلف در فرهنگ‌نگاری را تشریح کرد و رویکرد پیکره‌بنیاد را به عنوان یکی از این سه رویکرد مورد تاکید قرار داد. وی در پایان نشان داد که نه تنها در فارسی بلکه در انگلیسی و زبان‌های دیگر نیز غالباً فرهنگ‌لغت‌های پیکره‌بنیاد در مقایسه با سایر منابع مورد اقبال بیشتری قرار می‌گیرند. در همین راستا، اسلامی‌زاده (۱۳۹۴) فهرست واژگان آکادمیک زبانشناسی را تهیه و تدوین کرد. او پس از تهیه فهرست لغات، آنها را با واژگان موجود در دو پیکره AWL (Academic Word List) و پیکره GSL (General Service List) مقایسه کرد تا میزان موفقیت خود در پوشش واژگان زبانشناسی را مشخص کند. پیکره تولیدی او حاوی ۷۰۰ مقاله زبانشناسی، حدود ۴ میلیون واژه از چهار زیرشاخه زبانشناسی (واج‌شناسی، واژه‌سازی، معناشناسی و نحو) جمع‌آوری شد. تحقیق او نشان داد که پوشش واژگان گردآوری شده توسط او با توجه به محتوای پیکره‌های فوق‌الذکر مناسب بوده است. در خارج از ایران نیز حِجا (۲۰۱۰) نقش پیکره‌های موازی را در فرهنگ‌نگاری دوزبانه بررسی کرد. او رویکردی را در

ترازبندی کلمات در پیکره‌های موازی ارائه کرد که هدف آن تسهیل کار فرهنگ‌نگاران در تهیه واژه‌نامه‌ها بوده است. او رویکرد خود را رویکردی جدید نمی‌داند اما استفاده از آن را برای ساخت واژه‌نامه‌های دوزبانه جدید معرفی می‌کند. او فواید این رویکرد در تولید واژه‌نامه‌ها را در قالب چند نکته کلی بیان می‌کند. مهم‌ترین نکته اینکه یک پیکره موازی با اندازه قابل توجه تضمین می‌کند که مرتبط‌ترین ترجمه‌های واژگان در واژه‌نامه وجود داشته باشد. نکته دیگر آنکه با توجه به وجود ترجمه‌های مختلف از یک واژه در پیکره موازی، امکان رتبه‌بندی معادل‌ها بر اساس فراوانی کاربرد وجود دارد و نکته آخر اینکه تمامی جملات حاوی واژگان خاص در دسترس‌اند و بنابراین کاربر می‌تواند با بررسی جملات اصلی در مورد استفاده از یک واژه یا عدم استفاده از آن با توجه به بافت پیشنهادی تصمیم بگیرد.

کاهانی و جکیان طوسی (۱۳۹۱) برای اولین بار مدلی ترکیبی را برای ترازبندی جملات ارائه دادند که از آن در ساخت پیکره‌های موازی انگلیسی-فارسی استفاده می‌شود. ویژگی خاص مدل پیشنهادی آنها غیروابسته بودن به زبان مبدأ و مقصد می‌باشد و از آن می‌توان برای تولید پیکره برای هر جفت زبان دیگر نیز استفاده کرد. البته، نویسندگان اذعان دارند که سیستم پیشنهادی آنها با چالشی نیز رو به رو است بدین معنی که در ترجمه متون سلیقه‌های مختلفی وجود دارد و از این رو در برخی موارد جملاتی تولید می‌شود که سیستم قادر به تشخیص آنها نیست. به بیان دیگر، صرفاً انسان (به صورت دستی) قدرت درک و تشخیص آن را دارد. در چنین وضعیتی کار استخراج جفت جمله‌های معادل بسیار دشوار می‌شود. ضمناً در مورد متونی که مقید به رعایت ساختار دستوری زبان فارسی و انگلیسی نمی‌باشند، این احتمال وجود دارد که تشخیص جملات هم‌تراز به خوبی انجام نشود.

پیکره میزان بیش از یک میلیون جمله را از متون مختلف به ویژه از حوزه ادبیات کلاسیک و نیز ترجمه این جملات را به فارسی در خود دارد. این پیکره توسط دبیرخانه شورای عالی اطلاع‌رسانی و در سال ۱۳۹۲ تهیه گردیده که از آن می‌توان برای انجام پژوهش در حوزه‌هایی چون پردازش زبان طبیعی، زبانشناسی رایانشی، زبانشناسی پیکره‌ای و به‌خصوص حوزه ماشین ترجمه استفاده نمود (خبرگزاری باشگاه خبرنگاران، ۱۶ اردیبهشت ۱۳۹۲). با توجه به اهمیت موضوع گسترش خط و زبان فارسی در محیط رایانه‌ای، تهیه پیکره موازی

انگلیسی-فارسی با یک میلیون جمله در دستور کار دبیرخانه شورای عالی اطلاع‌رسانی قرار گرفت. بر اساس همین پیکره، سامانه ترجمه آماری پایه با نام «مترجم برخط» به آدرس http://machinetranslation.ir نیز راه‌اندازی شد.

محمدی و قاسم‌آقایی (۱۳۸۸) اذعان داشتند پیکره‌های موازی یکی از منابع مهم برای بسیاری از تحقیقات زبانشناسی در حوزه چندزبانی مخصوصاً ترجمه ماشینی مبتنی بر پیکره‌های متنی است. در این مقاله، محققان یک پیکره متنی موازی و ترازشده را برای جفت زبان فارسی-انگلیسی با کاوش در محتویات ویکیپدیا ارایه دادند. آنها روشی را برای ترازبندی در سطح جمله ارایه کردند که از یک روش بهبودیافته لغت‌نامه دوزبانه استفاده می‌کند. نتایج نشان داد که دقت این روش نسبت به روش‌های مشابه با وجود کاهش ۵۰ درصدی تعداد کل جفت جملات کاندید تولیدشده، به دو برابر افزایش یافت.

«پیکره فارسی ۱۹۸۴» که نخستین پیکره موازی چندجانبه فارسی نیز محسوب می‌گردد، ۶۶۰۶ جمله، ۶۶۳۲ لما و ۱۳۵۹۷ کلمه دارد. در این پیکره که براساس متن اصلی رمان ۱۹۸۴ جورج اورول تهیه شده، ۴۴۸ برچسب نیز وجود دارد (قاسمی‌زاده، رحیمی و محمودی بختیاری، ۲۰۱۴). این پیکره در حوزهٔ متون ادبی برای زبان فارسی به همراه بیش از ۱۰ زبان دیگر اروپایی تولید شد.

پیله‌ور و همکاران (۲۰۱۱) پیکره موازی انگلیسی-فارسی تهران را با نام TEP طراحی کردند. آنها اذعان داشتند که با وجود اهمیت به کارگیری پیکره‌های موازی در حوزه‌های چندزبانه، تا زمان ساخت TEP پیکره بزرگ‌مقیاس انگلیسی-فارسی ساخته نشده بوده است. این پیکره از زیرنویس فیلم‌ها تهیه شد و حاوی چندین میلیون واژه است.

فرجیان (۲۰۱۱) نیز با استفاده از اطلاعات روزنامه‌ها، یک پیکره انگلیسی-فارسی موازی را با نام پِن (PEN) طراحی کرد. پیکره او یک پیکره تطبیق داده شده در سطح جمله بود که به شکلی نیمه-خودکار طراحی و ارایه گردید. برای تعیین شباهت بین جملات از دو معیار آماری سنجش شباهت استفاده شد. همچنین برای فرایند تأیید تطبیق خودکار جملات نیز از ماشین ترجمه گوگل استفاده شد. حجم داده‌های مورد استفاده نیز ۳۰۰۰۰ جفت جمله بود که از وبگاه‌های خبری گرفته شده بود.

موسوی میانگاه (۲۰۰۹) نیز یک پیکره موازی بزرگ مقیاس انگلیسی-فارسی را ارایه کرد. طبق اطلاعاتی که او ارایه می‌دهد این پیکره از متون الکترونیک و مدارک موجود در وب تهیه شده و از این رو حجم نویز در آن بسیار کم است. او علت انتخاب منابع وبی را سهولت دستیابی به معادل‌های متون به زبانهای دیگر و حجم زیاد اطلاعات در دسترس عنوان می‌کند. برای انجام کار پس از جستجوی متون معادل در انگلیسی و فارسی و ارزیابی کیفیت ترجمه آنها، نسبت به ذخیره این اطلاعات به فرمت ماشین‌خوان و تراز نمودن جملات اقدام شد. طبق گفته او، این پیکره یک پیکره باز است بدین معنی که می‌توان اطلاعاتی را به آن افزود. در پایان، او بیان می‌کند که در نظر دارد از این پیکره در سیستم ترجمه‌ای که به تازگی کار بر روی آن را شروع کرده، استفاده نماید.

محمدی (۱۳۹۱) نیز در پژوهش خود به ساخت پیکره تطبیقی فارسی-انگلیسی و استخراج جملات موازی از آن مبادرت ورزید. او برای تهیه پیکره تطبیقی خود از مقالات روزنامه همشهری و اخبار وبسایت شبکه بی بی سی انگلیسی استفاده کرد. معیارهای استخراج شده نیز عبارت بود از: تعداد کلمات کلیدی مشترک، اسامی خاص یکسان، عناوین مشابه و فاصله تاریخ انتشار دو خبر. سپس به هر یک از این چهار متغیر بر اساس اهمیت آن وزنی اختصاص یافت. محقق بیان می‌دارد که پیکره او از پیکره‌های تطبیقی دیگر تولید شده برای زبان فارسی عملکردی بهتر داشته است. محقق همچنین با استفاده از این متون، پیکره‌ای از جملات موازی را با استفاده از ویژگی‌های لغوی، طولی و همپوشانی لغات استخراج کرد.

موور (۲۰۰۲) روش جدیدی را برای منطبق کردن جملات با ترجمه‌شان در یک پیکره دوزبانه موازی ارایه کرد. رویکردهای قبلی عموماً بر طول جملات یا مطابقت کلمات استوار بود. روش‌های مبتنی بر طول جملات نسبتاً سریع‌اند و از دقت متوسط برخوردارند. در حالی که روش‌های مبتنی بر مطابقت واژه از نظر سرعت کندتر ولی از نظر دقت مناسب‌ترند. گروه اخیر غالباً بر روی واژگان دوزبانه متکی‌اند. موور این دو روش را درهم آمیخت و نشان داد که با این شیوه جدید با صرف هزینه متعارف می‌توان به دقت بسیار بالاتری دست یافت. مزیت این روش ترکیبی نیز در این است که به زبان مبدأ و مقصد وابسته نیست و تنها کاری که باید انجام شود این است که متن دو زبان به جملات و کلمات شکسته شود.

آنچه از این مرور مختصر بر می‌آید آن است که انواع متفاوتی از پیکره‌ها با اهداف متفاوت وجود دارد که توسط محققان مختلف به انجام رسیده و یا در حال انجام است. برخی از آنها مفاهیم عام را در بردارند و برخی نیز بر روی حوزه‌های خاص متمرکز شده‌اند. برخی به صورت دستی تهیه می‌شوند و برخی به صورت نیمه‌ماشینی یا تمام‌خودکار. با توجه به مأموریت مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری و وجود پایگاه مقالات تمام‌متن فارسی در آن (که در حال حاضر به عنوان بزرگ‌ترین پایگاه مقالات مجلات معتبر فارسی در سطح کشور شناخته می‌شود)، و با توجه به اینکه پیرو مصوبهٔ نهائی مقرر بوده پیکرهٔ تولیدی صرفاً خام و غیر نرم‌افزاری باشد، محقق حاضر کوشید ساخت انسانی پیکرهٔ موازی دوزبانهٔ (فارسی-انگلیسی) عناوین مقالات مجلات رتبه‌دار موجود در مرکز منطقه‌ای را به عنوان پژوهش موظف پیشنهاد دهد که از نظر ماهیت در چارچوب ترازبندی و تطبیق دستی (manual aligning) قرار می‌گیرد و از نظر محتوا نیز پیکره ای تخصصی و از نظر جهت از فارسی به انگلیسی است. پیکره‌هایی که از واژگان مقالات علمی استفاده کرده باشند تا آنجا که محقق اطلاع دارد بسیار محدود بوده است و به همین دلیل عملیاتی شدن آن در برنامه قرار گرفت.

۶- روش پژوهش

۶-۱ نوع مطالعه

مطالعهٔ حاضر مطالعه‌ای زبانشناختی-رایانشی است که در حوزهٔ زبانشناسی پیکره‌ای می‌گنجد و بنابراین از اصول و ضوابط موجود در این حوزه برای انجام پژوهش حاضر استفاده شد. البته با توجه به پیشنهاده طرح حاضر، نوع پیکره، یک پیکره تخصصی با برچسب‌های نحوی کلی (اسم، فعل، صفت، قید، حرف‌اضافه) بوده و جملات به صورت دستی تقطیع و برچسب‌دهی گردید. در این پژوهش از روش تحلیل داده‌ها نیز استفاده شد که در اینجا مراد از داده‌ها، ۱۰۰۰۰ جفت عنوان مقالات مجلات رتبه‌دار موجود در مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری می‌باشد. در این پژوهش همچنین از اصول موجود در حوزه تحقیقاتی واژه‌سازی (morphology) برای شناسائی و تقطیع واژگان در کنار روش‌های رایانشی (شکستن جملات از محل فاصله، استفاده از نرم‌افزار ACS و ...) استفاده شد. هر چند شیوه تقطیع کلمات بیشتر رایانشی و بر اساس اصل امکان

برابر سازی واحدها انجام شد، محقق از آثار پژوهشگران ایرانی در خصوص وندها و واژه‌سازی فارسی نیز بسیار بهره برد که برای نمونه می‌توان به کارهای شقاقی (۱۳۹۵)، صادقی (۱۳۹۴) و کلباسی (۱۳۸۸) اشاره کرد.

۲-۶ داده‌های پژوهش

با توجه به اهمیت و جایگاه ویژه مقالات مجلات رتبه‌دار و نیز جایگاه ویژه عنوان در ساختار مقالات علمی و با توجه به وجود اطلاعات خام عناوین مقالات رتبه‌دار وزارت علوم، تحقیقات و فناوری و وزارت بهداشت، درمان و آموزش پزشکی در مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، از طریق مراجعه حضوری، فهرست عناوین مقالات موجود اخذ گردید. (لازم به توضیح است که در این مرکز حدود یک میلیون مقاله تمام‌متن شامل مجلات رتبه‌دار و فاقد رتبه وجود دارد که محقق صرفاً به آن دسته از اطلاعات نیاز داشت که رتبه‌دار بوده و ضمناً برای آن عنوان فارسی کامل به همراه عنوان انگلیسی معادل موجود بوده و امکان در اختیار گرفتن آن توسط محقق مهیا بود. از این رو، عدد ۱۳۲۰۸۶ به عنوان جامعه آماری عناوین مربوطه تعیین و به محقق ارایه گردید.). ضمناً با توجه به اینکه این تعداد عنوان حوزه‌های موضوعی مختلفی را پوشش می‌داد (۸ حوزه با کدهای ۳ (فنی مهندسی)، ۴(علوم انسانی)، ۹(علوم پزشکی)، ۱۰(علوم کشاورزی)، ۱۱(هنر و معماری)، ۱۲(علوم پایه)، ۱۳(دامپزشکی) و ۱۶(منابع طبیعی) – شماره کدهای ذکر شده شماره کدهای موجود در نرم افزار رایسست بوده و صرفاً برچسبی برای هر حوزه موضوعی است و هیچگونه ارزش قراردادی یا خاص دیگری ندارد) برای پوشش منطقی و متناسب مقالات حوزه‌های مختلف در نمونه آماری، از روش نمونه‌برداری چندمرحله‌ای تصادفی استفاده شد. برای استفاده از این روش به اطلاعات زیر نیاز بود:

۱) تعداد کل عناوین مقالات (جامعه آماری کل): در این مورد در مجموع ۱۳۲۰۸۶ عنوان مقاله در قالب جامعه آماری کلی در اختیار محقق قرار گرفت.

۲) تعداد و حجم عناوین زیرحوزه‌ها: در داده‌های موجود ۸ زیرحوزه شناسایی شد. ضمناً تعداد عناوین (جفت‌عنوان‌ها) موجود در این زیرحوزه‌ها به شرح جدول ۱ بوده است:

جدول ۱. تعداد عناوین موجود در هر زیرحوزه

حوزه موضوعی	تعداد عناوین
۳	۱۰۷۲۲
۴	۵۵۱۴۶
۹	۳۴۷۱۰
۱۰	۱۶۵۱۶
۱۱	۲۶۹۱
۱۲	۸۲۶۰
۱۳	۲۲۱۱
۱۶	۱۸۳۰
مجموع	۱۳۲۰۸۶

۳) *حجم نهایی نمونه مورد نظر:* تعیین این حجم بر عهده محقق است. با توجه به اینکه در پیشنهاده، حجم نمونه ۱۰۰۰۰ عنوان فارسی (به همراه عناوین انگلیسی معادل) تعیین شده بود، عدد ۱۰۰۰۰ به عنوان حجم نهایی نمونه تعیین شد.

با استفاده از اطلاعات بالا و فرمول نمونه‌برداری چندمرحله‌ای تصادفی زیر، تعداد عناوین موجود در نمونه نهایی به شرح جدول ۲ تعیین شد:

*(حجم نهایی نمونه/حجم جامعه آماری کل) * تعداد کل عناوین در هر زیرحوزه = تعداد عناوین هر زیرحوزه در نمونه نهایی*

جدول ۲. نحوه محاسبه تعداد عناوین مورد نیاز از هر زیرحوزه در نمونه آماری نهایی

نام حوزه موضوعی	نحوه محاسبه*	حجم عناوین مورد نیاز از هر زیرحوزه در نمونه نهایی
۳	=۱۰۷۲۲*۱۳۲۰۸۶/۱۰۰۰۰	۸۱۲
۴	=۵۵۱۴۶*۱۳۲۰۸۶/۱۰۰۰۰	۴۱۷۵
۹	=۳۴۷۱۰*۱۳۲۰۸۶/۱۰۰۰۰	۲۶۲۸

۱۰	۱۰۰۰۰/۱۳۲۰۸۶*۱۶۵۱۶=	۱۲۵۰
۱۱	۱۰۰۰۰/۱۳۲۰۸۶*۲۶۹۱=	۲۰۴
۱۲	۱۰۰۰۰/۱۳۲۰۸۶*۸۲۶۰=	۶۲۵
۱۳	۱۰۰۰۰/۱۳۲۰۸۶*۲۲۱۱=	۱۶۷
۱۶	۱۰۰۰۰/۱۳۲۰۸۶*۱۸۳۰=	۱۳۹
مجموع		۱۰۰۰۰

*. فرمول از http://www.statisticshowto.com/stratified-random-sample گرفته شد.

در مرحله پایانی با استفاده از روش نمونه‌گیری تصادفی ساده، تعداد عناوین مورد نیاز از هر زیرحوزه استخراج و در قالب یک فایل اکسل گردآوری شد. هر یک از ۱۰۰۰۰ رکورد اطلاعاتی از این مجموعه کد حوزه موضوعی، کد مقاله، عنوان فارسی و عنوان انگلیسی معادل را شامل می‌شد هر چند در پیکره تولیدی نهایی کد حوزه موضوعی و کد مقاله لحاظ نگردید. نمونه زیر اطلاعات اولیه یک رکورد اطلاعاتی را نشان می‌دهد:

کد زیرحوزه	کد مقاله	عنوان فارسی	عنوان معادل انگلیسی
۱۱	۱۶۵۱۵۸۵	رویکردی به عنصر آب در اساطیر و فرهنگ اقوام مختلف	An approach to the element “water” in the myth and culture of various nations

۳-۶ نرم‌افزارها و توابع اکسل مورد استفاده

برای انجام پژوهش حاضر از نرم افزار اکسل به عنوان قالب ماشین‌خوان برای تولید پیکره حاضر استفاده شد. شایان ذکر است پیرو پیشنهاده، مقرر بوده فرایندهای موجود در پژوهش حاضر شامل تقطیع، برچسب‌دهی و ... به صورت دستی به انجام برسد که با توجه به حجم بالای داده‌ها (۲۰۰۰۰ عنوان شامل ۱۰۰۰۰ عنوان فارسی و

۱۰۰۰۰ عنوان معادل انگلیسی) عملاً این کار با اشکالات زیادی همراه بود. برای مثال، فرایند صرفاً دستی، عدم یکدستی در تقطیع کلمات و همچنین عدم یکدستی در برچسب‌زنی نحوی و خطاهای انسانی را سبب می‌شد. ضمناً، فرایند صرفاً دستی، ویرایش‌های هدفمند یا انجام تحلیل‌های دیگر نظیر تحلیل‌های بسامدی را دشوار می‌ساخت. از این رو، از توابع اکسل و نیز نرم‌افزار ACS به عنوان ابزار کمکی در کنار تحلیل‌ها و ویرایش‌های انسانی استفاده شد. استفاده از توابع اکسل و نرم‌افزار ACS نه تنها فرایند کار را تسهیل کرد بلکه سبب شد محقق تحلیل بسامدی مدخل‌های واژگانی را نیز به پیکره خود اضافه کند که این مورد فراتر از چارچوب پیشنهاده بوده اما سبب گردید تا پژوهش کامل‌تر و کاربردی‌تر شود. از آنجایی که مختصات نرم افزار ACS در بخش ۶-۵ توضیح داده خواهد شد، در ادامه صرفاً توابع اکسل مورد استفاده، معرفی می‌شود.

الف- Text to Columns: از این تابع برای تقطیع عناوین فارسی و انگلیسی بر اساس معیار فاصله کامل استفاده شد. برای مثال، ترکیب «مواد شیمیایی» به دو کلمه «مواد» و «شیمیایی» شکسته و در قالب ستون‌های جداگانه ثبت شد. سپس، واژه انگلیسی معادل هم به همین شیوه تقطیع و در مقابل آن درج شد.

ب- Trim: کاری که این تابع انجام می‌دهد این است که اگر در اول یا آخر یا حتی وسط واژه بیش از یک فاصله باشد فاصله اضافی را حذف می‌کند و تنها یک فاصله را نگه می‌دارد.

ج- Length: این تابع، طول هر مدخل واژگانی را بر حسب تعداد کاراکترهای موجود نشان می‌دهد. از این تابع به خصوص در فرایند اصلاح واژگان تقطیع شده و سنجش یکسانی یا تفاوت چند مدخلِ به ظاهر مشابه، استفاده شد. در داده‌ها مواردی وجود داشت که در آن به ظاهر چند مدخل از نظر هر چهار متغیر (واژه فارسی، واژه انگلیسی، برچسب نحوی فارسی و برچسب نحوی انگلیسی) یکسان بودند اما توسط ACS به عنوان مدخل‌های مجزا در نظر گرفته شده بودند. با استفاده از تابع «طول کلمه»، طول این نوع مدخل‌ها بازبینی شد و سمبل‌های نامتعارف (wildcards) شناسایی گردید. در این حالت سعی شد نگارش غالب حفظ و نگارش نادر به شکل نگارش غالب درآید تا نرم‌افزار ACS بتواند همه این مدخل‌ها را در قالب یک مدخل واحد تحلیل و گزارش کند.

د- **Right and Left:** تابع Right سمت راست یک واژه را با سمت راست یک واژه یا واژه‌های دیگر مقایسه کرده یکسان یا متفاوت بودن کاراکترها را گزارش می‌کند. به همین ترتیب تابع Left سمت چپ یک واژه را با سمت چپ یک واژه یا واژه‌های دیگر تطبیق داده یکسانی یا تفاوت بین آنها را گزارش می‌کند. به عبارت دیگر، در صورت وجود wild character در آغاز یا پایان واژه آن را مشخص کرده، محقق می‌تواند نسبت به حذف آن کاراکتر و اصلاح مدخل مربوطه اقدام نماید. شیوه کار نیز به این ترتیب بود که اگر دو کاراکتر اول در یک مدخل با کاراکتر اول در مدخل مشابه دیگر یکی باشد یا دو کاراکتر آخر در یک مدخل با کاراکتر آخر در یک مدخل مشابه دیگر یکی باشد پس یک wild character وجود داشته که باید حذف شود.

ه- **Proper:** این تابع اولین حرف کلمه را بزرگ می‌کند و باقی حروف را اگر بزرگ باشند به حرف کوچک تبدیل می‌کند. از این تابع در این پژوهش برای این هدف استفاده شد که مشخص شود آیا مدخل‌هایی وجود دارد که تنها تفاوت موجود بین آنها در بزرگ یا کوچک بودن حروف باشد؟ البته با توجه به اینکه در این پیکره اسامی خاص با حرف بزرگ ثبت شدند، به صورت خودکار این تابع عملیاتی نشد و تنها پس از تایید محقق و بسته به ضرورت، نسبت به اعمال آن اقدام گردید.

۴-۶ روش انجام کار

برای انجام پژوهش حاضر اقدامات زیر به ترتیب به انجام رسید. ابتدا داده‌های پژوهش حاضر (۱۰۰۰۰ عنوان مقاله فارسی و عناوین انگلیسی معادل آنها) آنگونه که در قسمت داده‌های پژوهش (۶-۲) توضیح داده شد، از مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری تهیه شد. سپس، اطلاعات مربوطه در قالب یک فایل اکسل و حاوی اطلاعات عنوان فارسی، عنوان معادل انگلیسی، کد مقاله و کد حوزهٔ موضوعی به شرح تصویر ۱ ذخیره و نگهداری شد. از این فایل به عنوان فایل داده خام و اولیه پژوهش حاضر استفاده شد. در حقیقت، تمامی فعالیت‌های بعدی در این پژوهش بر مبنای داده‌های موجود در این فایل به انجام رسید.

ID	Text	Language	doc_id	Subfield
1	معماري و شهرسازي ايران در دوران گذار	1	1418431	11
2	Iranian Urbanism and Architecture during Transmission Period	2	1418431	11
3	عبا بافي بهبهان	1	1416516	11
4	Aba-Weaving in Behbahan	2	1416516	11
5	جايگاه امام علي (ع) در نسخه خطي حمله حيدري	1	1001180	11
6	STATUS OF IMAM ALI «GOD BLESS HIM» IN HAMLEYE HEY DA RI MANUSCRIPT	2	1001180	11
7	تزيينات كتيبه بقعه شاه نعمت الله ولي در ماهان كرمان	1	1001230	11
8	INSCRIBED DECORATIONS OF SHAH-NEMAT-ALLAH VALI MAUSOLEUM	2	1001230	11
9	ملاحظات فرهنگي در شكل دهي به نماهاي شهري با تكيه بر ساختار نماهاي شهري ايراني در دوران اسلامي	1	1726950	11
10	Cultural Consideration in Urban Facade Formation (Emphasize on urban facade formation in Iranian Islamic architecture)	2	1726950	11
11	گردشگري روستايي و لزوم توجه به آن در برنامه هاي توسعه و آباداني روستاها	1	1857660	11
12	Rural tourism and the need of special attention to be paid in rural development programs	2	1857660	11
13	رويكردي به عنصر آب در اساطير و فرهنگ اقوام مختلف	1	1651585	11
14	An approach to the element "water" in the myth and culture of various nations	2	1651585	11

تصویر ۱: بخشی از فایل اکسل داده‌های ۱۰۰۰۰ جفت عنوان فارسی و انگلیسی به همراه کد مقاله و حوزه موضوعی آن

در مرحله بعد، ترازبندی جملات معادل فارسی و انگلیسی در قالب یک فایل اکسل به انجام رسید بدین ترتیب که هر یک از ۱۰۰۰۰ جفت عنوان با یک کد واحد ترازبندی شدند. برای مثال، جمله ۱، اولین عنوان فارسی و جمله ۲، معادل انگلیسی جمله ۱ فارسی بود که با کد ۱ ترازبندی شدند. به طریقه مشابه جمله ۳ دومین جمله فارسی و جمله ۴ معادل انگلیسی جمله ۲ فارسی بود که با کد ۲ ترازبندی شدند و به همین ترتیب تا آخر (تصویر ۲).

جفت جملات	عنوان فارسی	عنوان معادل انگلیسی
۱	۱	۲
۲	۳	۴
۳	۵	۶
۴	۷	۸
۵	۹	۱۰
۶	۱۱	۱۲
۷	۱۳	۱۴
۸	۱۵	۱۶
۹	۱۷	۱۸
۱۰	۱۹	۲۰
...	...	...
۱۰۰۰۰	۱۹۹۹۹	۲۰۰۰۰

تصویر ۲: فرایند ترازبندی هر یک از جفت جملات

در گام بعدی لازم بود تا واژگان موجود در عناوین فارسی (و معادل انگلیسی آنها) در مخزن عناوین شناسایی و تقطیع شود. اما پیش از تقطیع در فایل اصلی عناوین، به دلیل وجود کاراکترهای مختلف در فایل، و برای پرهیز از مشکل بازیابی کلمات و برای اطمینان از محاسبه دقیق رخداد واژگان و تحلیل آماری، تبدیل‌های زیر انجام شد: «ی» عربی به «ی» فارسی؛ «أ» به «ا» در کلماتی نظیر «تأمین/تامین»؛ «ی» به«ئ» در کلماتی نظیر «مسیله/مسئله»؛ تغییر نیم فاصله‌ها به فاصله کامل (به جز در استثنائات بالا برای ها، های، ای، ی). نکته شایان ذکر دیگر در خصوص ایجاد تمایز بین واژه‌ها آن است که در این پژوهش نشانه‌های زیر و زبر زنجیری مانند فتحه، ضمه، کسره، تشدید، تنوین که توسط نویسندگان لحاظ گردیده بود به صورت مجزا لحاظ شد بدین معنی که «حتماً» و «حتما» دو مدخل جدا فرض شد یا «حتی» و «حتّی». این نکته از آن جهت حایز اهمیت است که پیکره باید صورت طبیعی استعمال زبان را توسط کاربران به تصویر بکشد و چنانچه این ترکیب‌ها با نرمال‌سازی یکی فرض می‌شد تنوع نگارشی موجود در پیکره که صورت طبیعی کاربرد زبان بوده، ناخواسته از بین می‌رفت.

در این مرحله، محقق به بازبینی عناوین نمونه و معادل‌های آنها در عناوین پرداخت و پس از بررسی مشخص شد که می‌توان در حوزه تقطیع و معادل‌یابی واژگان فارسی و انگلیسی، پنج احتمال یا حالت را در نظر گرفت که در ادامه به تفکیک توضیح داده می‌شود.

۴-۶- ۱ حالت اول

دسته اول آن دسته از معادل‌های واژگانی را شامل می‌شد که در فارسی دارای چند کلمه بود و معادل آن در انگلیسی هم بیش از یک واژه را شامل می‌شد (فارسی چندکلمه‌ای و انگلیسی چندکلمه‌ای). این دسته از واژگان با لفظ «Req» شناخته شدند. جدول ۳ نمونه‌ای از این نوع مدخل‌ها را به تصویر می‌کشد.

جدول ۳: نمونه‌ای از مدخل‌های فارسی چندکلمه‌ای با معادل انگلیسی چندکلمه‌ای

فارسی چندکلمه ای	انگلیسی چندکلمه‌ای
دانشکده پرستاری آبادان	Abadan School of Nursing
درد شکم	abdominal pain
استیل ال کارنیتین	acetyl L carnitine
اردشیر چهارم هخامنشی	Achaemenid Artaxerxes IV
علم حصولی	acquired knowledge
آکریلیک اسید	acrylic acid
حضرت آدم	Adam prophet
ارزش افزوده	added value
وابسته به مواد	addict dependent
نظام اداری	administrative body
آلن هیورث	Alan Haworth

برای تقطیع این مدخل‌ها بدین ترتیب عمل شد که با استفاده از متغیر break در نرم‌افزار اکسل، عناوین فارسی و نیز انگلیسی معادل از محل فاصله شکسته و در قالب جدول ارایه شد. البته با توجه به متفاوت بودن جهت نگارش در فارسی و انگلیسی بسیاری از مدخل‌ها از نظر ترتیب اجزا درست نبود که توسط محقق و به صورت دستی یکایک تقطیع‌های انجام شده بازبینی و اصلاح شد (جدول ۴). همچنین برخی از واژگان انگلیسی در فارسی برابرنهاده نداشت که این گونه موارد حذف شد که برای مثال می‌توان به حذف حرف تعریف معین the و نیز حرف اضافهٔ of معادل کسرهٔ اضافه در فارسی اشاره کرد. در حقیقت کسره اضافه در این پژوهش به عنوان یک واحد مجزا تفکیک نشد و فقط در صورتی که در آخر واژه انگلیسی به صورت e– یا ye– موجود بود و در واژه فارسی هم با نشانه کسره یا «ی» همراه بود، حفظ شد. این فرایند در تمامی موارد مشابه دیگر نیز به انجام رسید.

جدول ۴: نمونه‌ای از تقطیع ترکیبات گروه اول

شکل تقطیع واژگان فارسی			شکل تقطیع واژگان انگلیسی معادل		
پرستاری	دانشکده	آبادان	Nursing	School	Abadan
	درد	شکم		pain	Abdominal
کارنیتین	ال	استیل	carnitine	L	acetyl
چهارم	اردشیر	هخامنشی	IV	Artaxerxes	Achaemenid
	علم	حصولی		knowledge	acquired
	اسید	آکریلیک		acid	acrylic
	حضرت	آدم		prophet	Adam
	ارزش	افزوده		value	added
	وابسته به	مواد		dependent	addict
	نظام	اداری		body	administrative
	هیورث	آلن		Haworth	Alan

در این دسته (حالت اول) شیوه شناسایی نیز بدین صورت بوده است که ترکیب‌هایی که در فارسی و انگلیسی در ساختار خود دارای فاصله بودند در این مقوله قرار گرفتند. پس از بررسی و حذف موارد نامرتبط در نهایت ۲۰۲۳ ترکیب متمایز به همین شیوه و در قالب فایل Req. Break Final ارایه شد.

۲-۴-۶ حالت دوم

دسته دوم ترکیباتی را شامل می‌شد که در فارسی دارای چند کلمه بودند (یعنی در بین اجزای آن فاصله وجود داشت) و معادل آن در انگلیسی دارای خط فاصله (-) بود. این دسته از کلمات و ترکیبات هم به عنوان

دسته دوم کلمات مستعد برای شکسته شدن تعیین و شناسایی شده و برای سهولت ارجاع، Req-1 Break Final نامیده شد. در جدول ۵ نمونه‌ای از این ترکیبات آورده شده است.

جدول ۵: نمونه‌ای از مدخل‌های فارسی چندکلمه‌ای با معادل انگلیسی دارای خط فاصله (-)

فارسی چندکلمه ای	انگلیسی دارای خط فاصله (-)
برگ پهن	broad-leaved
لبه پهن	broad-crested
سه بعدی	3-D
وابسته به سن	age-related
آلفا سیناکلین	alpha-synuclein
ضد قارچی	anti-fungal
کشت پائیزه	autumn-sown
پخت همزمان	co-cure
بافت آگاه	context-aware
یادگیری الکترونیکی	e-learning

برای تقطیع این نوع ترکیبات ابتدا از محل فاصله کلمات فارسی شکسته شد. سپس در کلمات انگلیسی خط فاصله حذف و از همان محل کلمات قبل و بعد از خط فاصله از هم تفکیک شد. مجدداً با توجه به تفاوت جهت نگارش در فارسی و انگلیسی ترکیبات شکسته شده توسط محقق بازبینی و خطاهای موجود که تعداد آنها نیز غالب بود مرتفع شد. در نهایت ۷۷۲ ترکیب متمایز در این گروه حاصل شد. جدول ۶ نمونه ای از تقطیع کلمات در این گروه را نشان می‌دهد:

جدول ۶: نمونه‌ای از تقطیع ترکیبات گروه دوم

شکل تقطیع واژگان فارسی	شکل تقطیع واژگان انگلیسی معادل

broad	leaved	پهن	برگ
broad	crested	پهن	لبه
3	D	سه	بعدی
age	related	سن	وابسته به
alpha	synuclein	آلفا	سیناکلین
anti	fungal	ضد	قارچی
autumn	sown	پائیزه	کشت
co	cure	همزمان	پخت
context	aware	بافت	آگاه
e	learning	الکترونیکی	یادگیری

۳-۴-۶ حالت سوم

این دسته مدخل هائی را شامل می‌شد که در فارسی دارای چند کلمه بودند اما ترکیب معادل آنها در انگلیسی نه چندکلمه‌ای بود و نه در ساختار آن خط فاصله وجود داشت. علت تعریف دسته اخیر آن بود که ممکن است در مقابل ترکیبات چندکلمه‌ای فارسی، کلمه‌ای به ظاهر منفرد در انگلیسی وجود داشته باشد اما با بررسی دقیق‌تر بتوان اجزای آن را از هم تفکیک کرد و برای هر جزء فارسی معادل و برابرنهاده‌ای در انگلیسی معرفی کرد. پس از بررسی، دلایل عمده که می‌تواند باعث منفرد دیده شدن ترکیب انگلیسی باشد عبارت بود از احتمال قرار گرفتن اجزا در کنار هم به صورت چسبیده و بدون فاصله. به ترکیبات موجود در این دسته لفظ Doubt-Final اطلاق گردید بدین معنی که به بررسی بیشتر توسط محقق نیاز داشت که آیا امکان شکسته شدن وجود دارد یا خیر؟ جدول ۷ نمونه‌ای از تجزیه ترکیبات در این گروه را نشان می‌دهد.

جدول ۷: نمونه مدخل‌های فارسی چندکلمه‌ای با معادل انگلیسی غیرچندکلمه‌ای فاقد خط فاصله

فارسی چندکلمه‌ای	انگلیسی غیرچندکلمه‌ای و فاقد خط فاصله (-)
ضد صرع	Anticonvulsant
ضد افسردگی	Antidepressant
دو جانبه	Bilateral

Biodiversity	تنوع زیستی
Geostatistics	زمین آمار
Groundwater	آب زیرزمینی
Hyperactivity	بیش فعالی
Hydro alcoholic	آبی الکلی
IbnHani	ابن هانی
Intracellular	درون سلولی
Teleworking	دور کاری
Nanotubes	نانو لوله ها

ترکیبات مندرج در جدول ۷ بالا بدین صورت تفکیک شد که ابتدا محل تقطیع کلمه فارسی از روی فاصله مشخص شد. سپس واژه معادل انگلیسی بررسی شد. در صورتی که امکان تعریف جزء معادل در واژه انگلیسی مطابق با اجزاء واژه فارسی وجود داشت واژه فارسی از محل فاصله تقطیع شد و در واژه انگلیسی هم بین پیشوند یا پسوند و باقی ترکیب یک فاصله ایجاد شد و سپس ترکیب انگلیسی از محل فاصله از هم جدا گردید. در غیر اینصورت ترکیب نهایی تلقی شد و فرایند تقطیع متوقف گردید. برای مثال، واژه «ضد صرع» قابل تفکیک به دو بخش «ضد» و «صرع» است. از طرفی در انگلیسی هم معادل مربوطه دارای پیشوند anti است که با «ضد» برابر است و واژه convulsant که با «صرع» برابری می‌کند. بنابراین، در این مورد ابتدا واژه فارسی به دو جزء «ضد» و «صرع» و واژه انگلیسی نیز به دو جزء anti و convulsant شکسته شد. جدول ۸ شیوه تقطیع ترکیبات موجود در جدول ۷ را نشان می‌دهد.

جدول ۸: نمونه‌ای از تقطیع ترکیبات گروه سوم

شکل تقطیع واژگان فارسی چندکلمه ای		شکل تقطیع واژگان انگلیسی معادل تک کلمه‌ای قابل تفکیک و تراز	
صرع	ضد	convulsant	anti
افسردگی	ضد	depressant	anti
انبه	دو	lateral	bi
تنوع	زیستی	diversity	bio
آمار	زمین	statistics	geo

ground	water	زیرزمینی	آب
hyper	activity	بیش	فعالی
Hydro	alcoholic	آبی	الکلی
Ibn	Hani	ابن	هانی
intra	cellular	درون	سلولی
tele	working	دور	کاری

در نهایت در این گروه ۴۷۴ ترکیب متمایز بدست آمد. در تمام فرایندها هرجا که واژه‌ای توسط محقق غیر قابل تجزیه تشخیص داده شد به کلاس آخر که حاوی واژگان غیر قابل تجزیه بود منتقل گردید.

۴-۴-۶ حالت چهارم

این دسته واژه‌هایی را شامل می‌شود که برخلاف سه گروه قبلی (که در فارسی دارای بیش از یک جزء بودند) در فارسی تنها دارای یک جزء است بدین معنی که بین اجزای آن در فارسی فاصله‌ای وجود ندارد. با این حال در انگلیسی بیش از یک جزء را دارا می‌باشد. این گروه Others-Separated Final نامیده شد. علت تعریف این گروه آن بوده که به چند دلیل ممکن است ترکیب فارسی یک واژه منفرد تلقی شود با این حال با دقت بیشتر بتوان آن را تقطیع کرد و پتانسیل تقطیع شدن را داشته باشد. مثلاً ممکن است واژه فارسی دارای دو جزء باشد و آخر جزء اول و اول جزء دوم حرف منفصل وجود داشته باشد و از این رو به هم چسبیده باشند. یا هنوز یک نیم‌فاصله عامل عدم تفکیک دو یا چند جزء بوده باشد. ترکیبات مندرج در حالت چهار بدین صورت تفکیک شد که ابتدا محل تقطیع کلمه فارسی مشخص شد. برای تشخیص محل تقطیع ضمن توجه به حروف منفصل و متصل، فهرستی از پیشوندها و پسوندهای فارسی و انگلیسی موجود در داده‌های تحقیق به شرح جدول‌های ۹ تا ۱۲ تهیه شد. استخراج و تعیین این وندها توسط محقق و با مشاوره یک زبانشناس و نیز با بهره گیری از کتاب دستور احمدی گیوی و انوری (۱۳۷۱) انجام شد و بنابراین وندها موارد موجود در داده ها بوده اند و مراد تهیه فهرست کاملی از وندهای زبان فارسی نبوده است. ابتدا وجود وند در واژه فارسی بررسی شد.

جدول ۹: پیشوندهای استخراج شده از داده‌های فارسی مورد بررسی

نیمه	مولتی	کم	ضد	دی	تکتونو	پسا	بینا	آنتی	ابر
نئو	میان	لا	عدم	ریز	تیو	پست	بیو	باز	اتو
هایپر	میکرو	ما	عکس	زیر	چند	پلی	پاد	بایو	از پیش
هگزا	مینی	مادون	غیر	ژمینو	خرد	پیش	پارا	بدون	الکترو
هم	ن	ماکرو	فرا	ژئو	خرده	پیشا	پالینو	برهم	اندر
هیپر	نا	ماوراء	فرآ	ساب	خلاف	تحت	پان	برون	اور
هیدرو	نانو	مایکرو	فوق	سایبر	خود	ترا	پتانسیو	بلا	اولترا
هیستو	نو	متا	فیتو	سوپر	درون	ترمو	پره	بی	ایمنو
وا	نورو	مگنتو	فیزیو	شبه	دگر	تری	پری	بیش	اینتر
یاب	نیم	مورفو	کربو	شناسی	دور	تک	پس	بین	آلترا

جدول ۱۰: پسوندهای استخراج شده از داده‌های فارسی مورد بررسی

نگری	مندی	لوژیک	گرا	کاوی	سنج	دهی	خوان	پایه	اندیشی
نویسی	ناپذیری	لوژیکی	گراف	کشی	سنجی	رسانی	خوانی	پذیر	ای
هراسی	نشده	مانند	گرایی	کلان	شکنی	روی	خواهی	پذیری	آزاری
واره	نشین	مآبی	گریزی	کوشی	شناختی	ریزی	خوری	پردازی	آفرینی
ورزی	نشینان	متر	گزینی	گانگی	شناسی	زارها	خیز	پروری	باز
یابی	نشینی	متری	گشتی	گانه	شویی	زایی	خیزی	پژوهی	برداری
	نگار	متریک	گونه	گاه	طلبی	زدایی	دار	پنداری	بری
	نگاری	محور	گویی	گاه ها	کار	ساز	داران	پوشانی	بندی
	نگاریها	محوری	گیری	گذاران	کاران	سازی	داری	تراپی	بینی
	نگر	مدار	لوژی	گذاری	کاری	ستیزی	درمانی	توانی	پاشی

جدول ۱۱: پیشوندهای استخراج شده از داده‌های انگلیسی مورد بررسی

anti	counter	geo	intera	morpho	palyno	pre	speedo	ultra
auto	cross	half	intra	multi	pan	pseudo	sub	un
back	cyber	hexa	Kam	nano	para	psycho	super	under

baro	de	histo	lexico	neo	peri	quasi	tectono	uni
bi	demo	hydro	litho	neuro	petro	re	tele	webo
biblio	di	hyper	macro	no	physio	sedi	thermo	
bio	eco	icono	magneto	non	phyto	seismo	thio	
carbo	electro	im	meta	nono	polluto	self	tomo	
chemo	epi	immuno	micro	ortho	poly	semi	trans	
chrono	extra	infra	mini	over	post	socio	tri	
co	gemino	inter	mono	palaeo	potentio	spectro	typo	

جدول ۱۲: پسوندهای استخراج شده از داده‌های انگلیسی مورد بررسی

fold	graphy	logic	Meter	metric	metry	scopy	Xani
grams	less	logical	Meteric	metrical	nity	therapy	
graph	like	logy	Metery	metrics	rizi	thermo	

سپس واژه معادل انگلیسی بررسی شد. در صورتی که در ترکیب فارسی وند وجود داشت و همچنین امکان تعریف جزء معادل در واژه انگلیسی مطابق با اجزاء واژه فارسی مهیا بود، بین پیشوند یا پسوند و باقی اجزا در ترکیب فارسی و انگلیسی یک فاصله ایجاد و سپس ترکیب از محل فاصله از هم جدا شد. در غیر اینصورت ترکیب، نهایی تلقی شد و فرایند تقطیع متوقف گردید.

برای مثال واژه «اسیدپاشی» به صورت یک واژه نوشته شده است. با مراجعه به انگلیسی آن به صورت acid throwing برمی‌خوریم. مقایسه‌ای ساده بین اجزا ترکیب فارسی و انگلیسی نشان می‌دهد که اولاً در واژه فارسی یک پیشوند قابل تشخیص است. ثانیاً امکان تفکیک واژه فارسی به دو جزء آن هم بر اساس اجزاء واژه انگلیسی وجود دارد و بنابراین واژه فارسی به بخش‌های «اسید» و «پاشی» و واژه معادل انگلیسی هم به تبع آن به دو جزء acid و throwing تقطیع شد. جدول ۱۳ نمونه‌های دیگری از ترکیبات این دسته را نشان می‌دهد.

جدول ۱۳: نمونه ای از ترکیبات حالت چهارم (کلمات فارسی به ظاهر تک‌کلمه‌ای اما قابل تفکیک و برابرسازی با اجزاء ترکیب انگلیسی)

فارسی ظاهراً تک‌کلمه‌ای اما قابل تفکیک	انگلیسی چندکلمه ای
*عبدالحمید	Abd al-Hamid
اسیدپاشی	acid throwing
*اسرارالتوحید	Asrar Al-Tawhid

Ahmad Aghaei	❋احمدآقائی
case study	موردکاوی
balance sheet	❋ترازنامه
mismatch compliance	ناهمسانی
blood pressure	❋فشارخون
before constitution	پیشامشروطه
dual space	دوفضائی
grammar writing	دستورنویسی
regeneration	باززائی

❋. کلمات ستاره‌دار به خاطر حروف منفصل ترکیب شده بوده و از اینرو از هم تفکیک شدند. باقی ترکیبات به خاطر وند تفکیک شدند

جدول ۱۴: شیوه تقطیع کلمات حالت چهارم

شکل تقطیع واژگان فارسی ظاهراً تک‌کلمه‌ای		شکل تقطیع واژگان انگلیسی معادل چندکلمه‌ای	
الحمید	عبد	al-Hamid	Abd
پاشی	اسید	Throwing	acid
التوحید	اسرار	Al-Tawhid	Asrar
آقائی	احمد	Aghaei	Ahmad
کاوی	مورد	Study	case
نامه	تراز	Sheet	balance
همسانی	نا	Compliance	mismatch
فشار	خون	Pressure	blood
مشروطه	پیشا	constitution	before
فضائی	دو	Space	dual
نویسی	دستور	Writing	grammar
زائی	باز	Generation	re

جدول ۱۴ شیوه تقطیع ترکیبات گروه چهارم را به تصویر می‌کشد. همانگونه که ملاحظه می‌شود از حروف منفصل و متصل و نیز وندها برای تفکیک ترکیبات استفاده شده است. در این گروه در مجموع ۸۷۴ ترکیب متمایز و تقطیع شده حاصل شد.

۵-۴-۶ حالت پنجم

بعد از تکمیل مرحله چهار و جداسازی پیشوندها و پسوندها آنچه باقی ماند در این دسته قرار گرفت. به این گروه لفظ Others-Sum Final اطلاق گردید چرا که در هیچ یک از چهار گروه قبلی قرار نمی‌گرفت. در این گروه بیشترین تعداد واژگان متمایز وجود داشت که عبارت بود از ۲۳۸۱۴ واژه فارسی و ۲۳۸۱۴ واژه معادل انگلیسی آنها. این دسته دیگر قابل تجزیه نبود و به عنوان واحدهای تحلیلی و تقطیعی نهایی دست نخورده باقی ماند. جمع عددی مدخل‌های متمایز تولید شده در هر یک از این پنج گروه عدد ۲۷۹۵۷ بوده است که پس از ترکیب این پنج فایل و تحلیل با نرم‌افزار ACS به ۲۴۹۰۹ مدخل با فراوانی کل ۹۸۰۳۹ رسید. جدول زیر نمونه‌هایی از واژگان این دسته را نشان می‌دهد.

جدول ۱۵: حالت پنجم، نمونه‌ای از ترکیبات غیر قابل تفکیک

فارسی غیر قابل تفکیک	انگلیسی غیر قابل تفکیک
معماری	architecture
و	and
شهرسازی	urbanism
دوران	period
ایران	Iran
امام	Imam
علی	Ali
جایگاه	status
کرمان	Kerman
در	in
شکل دهی	formation
روستائی	rural

همانگونه که در جدول بالا مشاهده می‌شود عملاً واژگان موجود در این گروه کوچک‌ترین واحدها به شمار می‌روند و امکان شکستن آنها وجود ندارد. گاه، البته واژه فارسی ممکن است دارای جزئی جداشدنی باشد اما از این جهت در این گروه قرار گرفته شده باشد که صورت انگلیسی آن تک‌واژه‌ای و غیر قابل تقطیع بوده است مانند ترکیب «شکل دهی» که معادل آن formation یعنی یک واژه منفرد است.

۵-۶ تعیین واژه‌های متمایز در هر پنج گروه (word types)، برچسب نحوی و بسامد هر واژه

پس از شناسائی پنج گروه و تقطیع واژگان موجود در آنها مشاهده شد که در فهرست مدخل‌های هر گروه، واژه‌های تکراری زیادی وجود دارد (tokens). در این مرحله لازم بود صورت‌های مشابه در هر یک از پنج گروه در هم ترکیب شود تا فهرستی از واژه‌های متمایز (types) به دست آید. برای انجام این مهم از نرم‌افزار ACS (نعمتی، ۱۳۹۷) استفاده شد.

نرم‌افزار ACS با دریافت تا سقف چهار متغیر یا شرط برای هر مدخل، آن را با مدخل‌های دیگر مقایسه می‌کند و در صورتی که مدخل‌هایی در تمام متغیرها با هم یکی باشند آنها را یک مدخل فرض کرده به تعداد تکرار، جلوی آن مدخل، عددی را درج می‌کند که نشان‌دهنده فراوانی رخداد آن صورت است. جدول ۱۶ شکل ظاهری و نحوه عملکرد این نرم‌افزار را نشان می‌دهد.

Start

جدول ۱۶: شکل ظاهری و نحوه عملکرد نرم‌افزار ACS

Advanced Conditional Sum					
NO.	**Cond.1**	**Cond.2**	**Cond.3**	**Cond.4**	**Qty. 1**
*	Abad	آباد	N	N	۲
	Abd	عبد	N	N	۱
	absorbent	جاذب	ADJ	N	۲
	abuse	آزاری	N	SFX	۲
	academic	دانشگاهی	ADJ	ADJ	۱

	acceleration	شتابان	N	ADJ	۱
	acceptance	پذیری	N	SFX	۱
*	Abad	آباد	N	N	۲
***	admitted	مقیم	ADJ	ADJ	۱
**	action	کنش	N	N	۲
	addict	معتاد	N	ADJ	۱
	addiction	اعتیاد	N	N	۱
***	admitted	مقیم	ADJ	ADJ	۱
	advantage	بهره	N	N	۱
*	Abad	آباد	N	N	۲
	admitted	مقیم	ADJ	ADJ	۱
**	action	کنش	N	N	۲

برای مثال در جدول نمونه بالا، ۱۷ مدخل وارد نرم‌افزار ACS شده است و هر مدخل حاوی واژه فارسی، واژه معادل انگلیسی، مقوله نحوی واژه فارسی و مقوله نحوی واژه انگلیسی (چهار شرط مورد نظر) است. پس از فعال‌سازی نرم‌افزار و کلید زدن بر روی دکمه Start، واژه‌های مشابه در هر چهار متغیر شناسایی و به شکل جدول ۱۷ درهم‌کرد می‌شوند.

Start

جدول ۱۷: نتیجه اعمال نرم‌افزار ACS بر روی فایل نمونه بالا

Advanced Conditional Sum					
NO.	**Cond.1**	**Cond.2**	**Cond.3**	**Cond.4**	**Qty. 1**
*	Abad	آباد	N	N	۶
	Abd	عبد	N	N	۱

۲	N	ADJ	جاذب	absorbent	
۲	SFX	N	آزاری	abuse	
۱	ADJ	ADJ	دانشگاهی	academic	
۱	ADJ	N	شتابان	acceleration	
۱	SFX	N	پذیری	acceptance	
۴	N	N	کنش	action	
۱	ADJ	N	معتاد	addict	
۱	N	N	اعتیاد	addiction	
۳	ADJ	ADJ	مقیم	admitted	
۱	N	N	بهره	advantage	

مشاهده می‌شود که پس از عملکرد نرم‌افزار، ۱۷ مدخل به ۱۲ مدخل کاهش یافته که نشان‌دهنده حذف تکرار است. مشاهده می‌شود که در جدول بالا ۳ رخداد واژه «آباد» در هم ادغام گردیده و فراوانی آنها (۲،۲،۲) با هم ترکیب شده و به صورت عدد ۶ نشان داده شده است. یا واژه «کنش» به همین شیوه دو رخداد آن هر کدام با فراوانی ۲ به یک رخداد با فراوانی ۴ تغییر یافته است. به طریقه مشابه، دو رخداد واژه «مقیم» هر کدام با فراوانی ۱ در هم ترکیب شده و با فراوانی ۲ نشان داده شده است.

با توجه به اینکه در پنج گروهِ مدخل‌های ذکر شده، واژه فارسی و معادل انگلیسی آن (متغیر ۱ و ۲) وجود داشت، و با توجه به پتانسیل نرم‌افزار در اخذ چهار متغیر، محقق تصمیم گرفت مقوله نحوی هر واژه فارسی (متغیر ۳) و مقوله نحوی هر واژه معادل انگلیسی (متغیر ۴) را نیز تعیین و جلوی هر مدخل درج نماید. برای درج مقوله نحوی مطابق پیشنهاده مقرر بود پنج مقوله نحوی عمده شامل «اسم»، «صفت»، «قید»، «فعل» و «حرف اضافه» درج شود با این حال در فرایند کار مدخل‌هایی مشاهده شد که در هیچ یک از مقوله‌های ذکر شده قرار نمی‌گرفت. بر این اساس و متناسب با مدخل‌های موجود، فهرست مقوله‌های نحوی برای جلوگیری از حذف مدخل‌ها به شرح زیر گسترش پیدا کرد. شایان ذکر است برای تشخیص نوع مقوله نحوی از مختصات کتاب دستور زبان فارسی تالیف احمدی گیوی و انوری (۱۳۷۱) استفاده شد. البته، لفظ گروه اسمی (NP) بالاجبار و به خاطر وجود گروه اسمی در فارسی با معادل غیر قابل برابرسازی در انگلیسی به این مجموعه اضافه شد. لفظ DET هم به خاطر مقتضیات مدخل‌های انگلیسی لحاظ شد. تعریف هر مقوله نحوی هم همان

تعریفی لحاظ شد که توسط احمدی گیوی و انوری در کتاب دستور زبان فارسی‌شان ارایه شده بود. ضمناً در تعیین مقوله واژه ها صرفاً معیار صوری و مقوله واژه نامه ای کلمات مد نظر بوده و از پرداختن به معیارهای معنایی در این پژوهش به دلیل پیچیدگی آن ها، خودداری شد.

جدول ۱۸: مقوله‌های نحوی مورد استفاده در فرایند برچسب‌زنی

مقوله نحوی	صورت انگلیسی	کد مورد استفاده
اسم	Noun	N
فعل	Verb	V
صفت	Adjective	ADJ
قید	Adverb	ADV
حرف‌اضافه	Preposition	PREP
حرف‌ربط	Conjunction	CON
ضمیر	Pronoun	PRO
حرف‌تعیین	Determiner	DET
حرف تعریف	Article	ART
پیشوند	Prefix	PFX
پسوند	Suffix	SFX
گروه اسمی	Noun Phrase	NP

پس از تکمیل فرایند درج مقوله نحوی برای واژه فارسی و واژه انگلیسی معادل در هر مدخل، برای افزایش دقت کار یک متخصص زبانشناس دیگر با مدرک دکتری زبانشناسی نیز برچسب‌های نحوی را بررسی و تایید و یا اصلاح کرد. همچنین با استفاده از نرم‌افزار ACS مدخل‌هایی که در متغیرهای اول و دوم (واژه فارسی و انگلیسی) مشابه بودند اما مقوله نحوی متفاوتی داشتند با کمک فیلتر در برنامه اکسل فهرست شد و مجدداً مورد وارسی قرار گرفت تا در صورت وجود تنوع غیرموجه در کدها، اصلاح و یکدست‌سازی شود. قابل ذکر است که در فرایند برچسب‌دهی، از مقوله نحوی واژه‌نامه‌ای و خارج از بافت کلمات استفاده شد چرا که کلمات به واحدهای خود تقطیع شده و عملاً ارتباط بین اجزا قطع شده بود.

پس از اصلاح و تکمیل برچسب‌های نحوی اصلی، در هر یک از پنج گروه (حالت‌های اول تا پنجم)، اطلاعات هر چهار متغیر (واژه فارسی، واژه انگلیسی، برچسب نحوی واژه فارسی و برچسب نحوی واژه انگلیسی) برای هر مدخل در نرم‌افزار ACS وارد گردید و بر روی دکمه Start کلید زده شد. بدین ترتیب تمامی مدخل‌هایی که در هر چهار متغیر یکسان بودند درهم‌کرد شدند و فراوانی رخداد هر مورد نیز در جلوی آن و سمت راست ثبت گردید. جدول ۱۹ نمونه‌ای از تحلیل بخشی از فایل نمونه Req (حالت اول) را نشان می‌دهد.

جدول ۱۹: نمونه‌ای از تحلیل فایل Req توسط نرم‌افزار ACS

فراوانی	برچسب فارسی	برچسب انگلیسی	واژه فارسی	واژه معادل انگلیسی
۱	N	N	عباس	Abbas
۱	N	N	عبدالله	Abdollāh
۱	N	ADJ	شکم	abdominal
۱	SFX	N	پذیری	ability
۲	ADJ	ADJ	آبسیزیک	abscisic
۱	NP	N	سوءمصرف	abuse
۱	ADJ	N	کاربره	access
۱	N	N	حادثه	accident
۱	N	N	اکرتا	accreta
۱	ADJ	N	استاتی	acetate
۱	ADJ	ADJ	استیک	acetic
۱	N	N	استوبوتیلیکوم	acetobutylicum
۱	N	N	استیل	acetyl
۱	ADJ	ADJ	هخامنشی	Achaemenid
۳۵	N	N	اسید	acid

پس از اتمام این فرایند، برونداد ACS برای هر پنج گروه با هم ترکیب و مجدداً در نرم‌افزار ACS درونداد و تحلیل شد تا مدخل‌هایی که در این فایل‌های پنج‌گانه با هم مشابهت دارند نیز شناسایی و درهم‌کرد شود و فراوانی رخداد آنها نیز مجدداً اصلاح گردد. نکته قابل ذکر آن است که علت عدم درونداد تمامی اطلاعات به

یکباره (به جای کار بر روی پنج فایل به صورت جدا و سپس درهم‌کرد آنها) آن بوده که تعداد رکوردها بسیار زیاد بود و از این رو برای سهولت تحلیل، ابتدا محتوای هر گروه تحلیل شد و سپس نتایج هر گروه با هم ترکیب و در نرم‌افزار مجدداً اجرا شد.

۷- تولید پیکره نهایی بسامدی (فارسی-انگلیسی)

پس از انجام مراحل بالا یک فایل اکسل بزرگ حاصل شد حاوی ۲۴۹۰۹ مدخل فارسی به همراه معادل انگلیسی، برچسب نحوی اصلی هر واژه و فراوانی رخداد هر مدخل در پیکره. ضمناً مجموع فراوانی این تعداد مدخل در کل پیکره، عدد ۹۸۰۳۹ بوده است. این پیکره که در ۶۶۷ صفحه تهیه و در پایان همین گزارش ارایه گردیده حاصل نهایی کار در پژوهش حاضر است. جدول ۲۰ بخش کوچکی از پیکره تولید شده را به عنوان نمونه نشان می‌دهد. بدین ترتیب نسبت مدخل متمایز به تعداد کل تکرار (type تقسیم بر token) معادل ۰/۲۵۴ بدست آمد. شایان ذکر است این عدد هر چه به ۱ نزدیک‌تر باشد نشان دهنده واژه‌های متمایز بیشتر و فراوانی کم‌تر است و هر چه به صفر نزدیک‌تر شود نشان دهنده آن است که واژه‌های متمایز کم‌تر و فراوانی آنها بیشتر بوده است. البته لازم به ذکر است که عدد ۱ و ۰ هر دو اعدادی آرمانی‌اند که غالباً هیچگاه در عمل محقق نمی‌شوند، یعنی تصور متنی که در آن هر کلمه فقط یکبار به کار رفته باشد یا متنی داشته باشیم که در آن تنها از یک کلمه با صدها بار تکرار استفاده شده باشد، ممکن نیست. بنابراین عددِ نسبت، همواره بین این دو حد نهایت قرار می‌گیرد.

جدول ۲۰: چند مدخل نمونه از پیکره تولید شده (شامل ۲۴۹۰۹ مدخل با فراوانی کل ۹۸۰۳۹)

واژه فارسی	واژه انگلیسی	مقوله نحوی فارسی	مقوله نحوی انگلیسی	تعداد رخداد
ام اس	MS	N	N	۱
امپرازول	omeprazole	N	N	۱
امتحان	test	N	N	۱
امتحان‌های کوچک	quizzes	NP	N	۱
امتحانی	exam	ADJ	N	۵
امسل	amsel	N	N	۱
امنتین	serum omentin	N	NP	۲
امنیت	security	N	N	۱
امواج	waves	N	N	۱
امواج مسدودکننده	jammer	NP	N	۱
امید	hope	N	N	۱
امیدواری	hope	N	N	۱
اندازه	size	N	N	۱۳
اندازه گیری	measurement	N	N	۱
...	...	...	...	...

۸- کاربردهای عملی پژوهش

پیکره‌ها به طور کلی دارای کاربردهای متعددی هستند و پیکره حاضر نیز از این قاعده مستثنی نیست. پیکره‌ها نشان‌دهنده کاربرد عینی زبان از سوی کاربران هستند (مک‌کارتی و اُکیفِ، ۲۰۱۰) و همین نکته اولین کاربرد و دلیل اهمیت تولید پیکره‌ها را نشان می‌دهد. در پژوهش حاضر محقق عناوین ۱۰۰۰۰ عنوان مقاله رتبه‌دار وزارتین و ۱۰۰۰۰ عنوان معادل انگلیسی آنها را بررسی کرد و نشان داد که نحوه تولید عنوان فارسی و انگلیسی توسط نویسندگان مقالات رتبه‌دار به چه صورتی بوده است.

همچنین گروه‌ها و حوزه‌های تحقیقاتی مختلفی می‌توانند از پیکره‌هایی از این دست بهره‌برداری نمایند. یکی از این حوزه‌های تحقیقاتی، حوزه وسیع آموزش زبان است. در حقیقت، مدرسان زبان همواره به بررسی دانش املایی، واژگانی، نحوی و ... کاربران زبان می‌پردازند. نگاهی مختصر به پیکره تولیدشده نشان داد که کاربران (نویسندگان مقالات) در حین ترجمه واژه‌های فارسی، از واژگان انگلیسی متفاوت و متعددی استفاده کرده‌اند. هرچند گاه این تنوع می‌تواند به خاطر ماهیت رشته و تفاوت‌های بین‌رشته‌ای باشد اما موارد متعددی از اشکالات املایی و حتی دستوری در واژگان مشاهده شد. بنابراین اشکالات املایی و نیز دستوری از این نوع می‌تواند منبع خوبی برای پژوهش باشد.

زبانشناسان نیز می‌توانند از یافته‌های پژوهش حاضر به شکل‌های مختلف بهره‌برداری نمایند. برای مثال آنها می‌توانند واژگان به کار رفته در عنوان را از نظر واژه‌سازی بررسی نمایند یا واژگان مربوط به مقوله‌های نحوی مختلف را از نظر بسامدی تحلیل کنند که در صورت موافقت سازمان انجام این موارد با توجه به داده‌های فراهم‌شده در آینده میسر خواهد بود.

دانشجویان و محققان حوزه‌های زبانشناسی، زبانشناسی رایانشی و رایانه نیز می‌توانند از داده‌های موجود در این پژوهش شامل پیکره تولید شده و نیز فهرست پیشوندها و پسوندهای فارسی و انگلیسی تولید شده به عنوان مبنایی برای پژوهش‌های خود استفاده نمایند. البته، این نوع وندها از داده‌های مورد بررسی به دست آمد و بنابراین نیاز به گسترش آن در پژوهش‌های دیگر محتمل است با این حال در پژوهش حاضر به گسترش آن نیازی نبود.

از یافته‌های این پژوهش در حوزه فرهنگ‌نگاری نیز می‌توان استفاده کرد. در حقیقت، شروع پژوهش‌های پیکره‌ای با مطالعات فرهنگ‌نگاری بوده است و می‌توان واژگان موجود در پیکره ارایه شده را در قالب یک فرهنگ واژگانی دو زبانه ارایه کرد که خود می‌تواند برای نویسندگان مقالات بسیار مفید باشد. در چنین لغت‌نامه‌ای واژه فارسی با معادل‌های متنوع آن در انگلیسی گردآوری شده و واژه‌نامه‌ای دو زبانه از فارسی به انگلیسی تولید می‌شود. یا می‌توان این کار را از جهت دیگر (از زبان فارسی به زبان انگلیسی) انجام داد که در آن جلوی هر واژه انگلیسی معادل‌های متنوع آن به فارسی درج می‌گردد.

۹- نمونه‌ای از اطلاعات قابل استخراج از پیکره تولید شده

از پیکره تولید شده انواع متنوعی از اطلاعات قابل استخراج است. با توجه به اینکه استخراج این نوع اطلاعات در چارچوب پژوهش حاضر نبوده و خود به تحلیل‌های مستمر و گسترده نیاز دارد در این قسمت صرفاً برای نشان دادن پتانسیل اطلاعاتی پیکره حاضر، به چند نمونه مختصر از اطلاعات قابل استخراج از این پیکره اشاره می‌شود.

۹-۱ استخراج واژه‌های پربسامد (Highly Frequent Words)

تحلیل واژه‌های موجود در یک پیکره بر اساس تحلیل فراوانی، یکی از شایع‌ترین و جذاب‌ترین تحلیل‌هایی است که عموماً از پیکره‌ها صورت می‌گیرد. برای نشان دادن این قابلیت در پیکره حاضر، فایل پیکره تولید شده حاوی ۲۴۹۰۹ مدخل با فراوانی کل ۹۸۰۳۹ مورد بر اساس فراوانی رخداد واژه‌ها فیلتر و از فراوان‌ترین به نادرترین واژه مرتب شد. جدول ۲۱ فراوان‌ترین واژه‌های موجود در این پیکره را به تصویر می‌کشد.

جدول ۲۱: فراوان‌ترین واژه‌های موجود در پیکره تولید شده

و	and	CON	CON	۵۰۱۶
در	in	PREP	PREP	۴۸۸۷
بر	on	PREP	PREP	۱۷۴۲

همانگونه که در جدول بالا مشاهده می‌شود در پیکره تولید شده مدخل «و» با فراوانی ۵۰۱۶ مورد، فراوان‌ترین مدخل واژگانی بوده است. پس از آن مدخل «در» و مدخل «بر» به ترتیب با فراوانی ۴۸۸۷ و ۱۷۴۲ در رتبه‌های دوم و سوم قرار گرفتند. همین سه مدخل به تنهایی ۱۱/۸۸ درصد از کل رخدادهای موجود در این پیکره را شامل شد. این نوع تحلیل به شکل‌های مختلف قابل گسترش است. برای مثال، می‌توان فراوان‌ترین واژه‌ها در هر مقوله نحوی را نیز رصد کرد یا واژه‌ها را از نظر طول آنها بر اساس تعداد کاراکترها مورد بررسی قرار داد.

۹-۲ تعیین خطاهای نگارشی

بررسی صحت نگارش کلمات یکی دیگر از زمینه‌های تحقیقاتی در خصوص پیکره‌هاست. در این پژوهش نیز در املای کلمات عناوین فارسی و انگلیسی خطاهای نگارشی وجود داشت که کم و کیف این نوع خطاها خود می‌تواند نوع دیگری از پژوهش را سبب گردد. برای مثال، برای لفظ «ابعاد» dimensions از لفظ dimentions هم یک بار استفاده شده که مشخصاً خطای املائی است. یا برای «اثربخشی» در کنار معادل‌های انگلیسی درست از نگارش غلط efficacy هم استفاده شده که صورت درست آن efficacy بوده است. یا برای لفظ «اداره» از املای نادرست manegement هم در کنار املای درست management استفاده شده است. یا برای لفظ «ارتباط» در کنار معادل‌های درست، صورت نادرست انگلیسی relationship و نیز خطای تایپی elationship هر کدام با فراوانی ۱ مشاهده شد.

۹-۳ تعیین تنوع معادل گزینی از فارسی به انگلیسی و برعکس

یکی از جذاب‌ترین کاربردهای پیکره‌ها آن است که از طریق آنها می‌توان معادل‌گزینی واژه‌ها و تنوع معادل‌ها را از زبانی به زبان دیگر بررسی کرد. این امکان از طریق بررسی این پیکره نیز وجود دارد. به عنوان نمونه‌ای مختصر و در ادامه معادل‌های انگلیسی منتخب نویسندگان برای سه واژه «اجرا»، «احکام» و «ارتباط» آمده است (جدول‌های ۲۲-۲۴) .

جدول ۲۲: معادل‌های انتخاب شده توسط نویسندگان ایرانی برای واژه فارسی «اجرا»

واژه فارسی	واژه انگلیسی	مقوله نحوی فارسی	مقوله نحوی انگلیسی	تعداد رخداد
اجرا	application	N	N	۱
اجرا	enforcement	N	N	۶۲
اجرا	executability	N	N	۱
اجرا	execution	N	N	۱
اجرا	implementation	N	N	۲
اجرا	perform	N	N	۱
اجرا	performance	N	N	۲

اجرا	performing	N	N	۱

همانگونه که در جدول ۲۲ آمده است، نویسندگان ایرانی صورت واژگانی «اجرا» را با ۸ واژه مختلف در زبان انگلیسی به تصویر کشیده‌اند که در این میان لفظ enforcement با فراوانی ۶۲، فراوان‌ترین معادل انگلیسی برگزیده از سوی این نویسندگان بوده است.

جدول ۲۳: معادل‌های انتخاب شده توسط نویسندگان ایرانی برای واژه فارسی «احکام»

واژه فارسی	واژه انگلیسی	مقوله نحوی فارسی	مقوله نحوی انگلیسی	تعداد رخداد
احکام	commands	N	N	۷
احکام	decrees	N	N	۶
احکام	injunctions	N	N	۱
احکام	judgments	N	N	۱
احکام	laws	N	N	۱۵
احکام	precepts	N	N	۱
احکام	provisions	N	N	۱
احکام	verdicts	N	N	۱
احکام	principles	N	N	۱۱

مطابق جدول ۲۳ نویسندگان ایرانی صورت واژگانی «احکام» را با ۹ واژه مختلف در زبان انگلیسی نشان داده‌اند. در این میان، واژه‌های انگلیسی laws، principles و commands به ترتیب با ۱۵، ۱۱ و ۷ مورد، فراوان‌ترین معادل‌های انگلیسی صورت واژگانی «احکام» بودند.

جدول ۲۴: معادل‌های انتخاب شده توسط نویسندگان ایرانی برای واژه فارسی «ارتباط»

واژه فارسی	واژه انگلیسی	مقوله نحوی فارسی	مقوله نحوی انگلیسی	تعداد رخداد

ارتباط	analysis	N	N	۱
ارتباط	association	N	N	۲۱
ارتباط	communication	N	N	۱
ارتباط	connection	N	N	۲
ارتباط	correlation	N	N	۱
ارتباط	elationship*	N	N	۱
ارتباط	friendship	N	N	۱
ارتباط	highway	N	N	۱
ارتباط	interrelationship	N	N	۱
ارتباط	relationship*	N	N	۱
ارتباط	relation	N	N	۲
ارتباط	relations	N	N	۶۳
ارتباط	relationships	N	N	۲

سرانجام همانگونه که در جدول ۲۴ آمده است، برای واژه فارسی «ارتباط» از ۱۳ معادل انگلیسی مختلف استفاده شده است که در این میان، relations و association به ترتیب با ۶۳ و ۲۱ مورد فراوان‌ترین معادل‌های انگلیسی برای واژه «ارتباط» بودند.

۱۰-محدودیت‌های پژوهش

مانند هر پژوهش دیگری، پژوهش حاضر نیز با محدودیت‌هایی مواجه بوده است که در این قسمت به اختصار به برخی از آنها اشاره می‌شود. اولین محدودیت پژوهش حاضر، حجم داده‌های مورد بررسی بوده است. فرایند تحلیل داده‌ها و نیز فرایند برچسب‌زنی واژه‌ها فرایندی وقت‌گیر و پیچیده است و با توجه به محدودیت زمانیِ انجام پژوهش و نیز این حقیقت که کلیه فرایندها جز درونداد بخشی از اطلاعات صرفاً توسط محقق به انجام رسیده است، محقق ۱۰۰۰۰ عنوان مقاله فارسی و ۱۰۰۰۰ عنوان انگلیسی معادل عناوین فارسی را بررسی کرد. طبیعتاً در صورت وجود یک تیم تحقیقاتی امکان کار بر روی عناوین بیشتری وجود داشت.

محدودیت دیگر در این پژوهش به برچسب‌ها مربوط می‌شود. در حقیقت هدف اولیه درج پنج مقوله نحوی از میان مقوله‌های نحوی اصلی کلمات (اسم، فعل، قید، صفت و حرف‌اضافه) بوده است که البته با توجه به

ضرورت کار، این مقوله‌ها در جریان پژوهش گسترش پیدا کرد و اسم، فعل، صفت، قید، حرف‌اضافه، حرف‌ربط، حرف‌تعریف، حرف‌تعیین، ضمیر، گروه اسمی و مقوله‌هایی نظیر پیشوند و پسوند را نیز شامل شد. علت درج گروه اسمی در این فهرست آن بوده که با توجه به اینکه در برخی موارد در مقابل یک گروه اسمی یک معادل انگلیسی غیر قابل تجزیه وجود داشت بالاجبار می‌بایست شکستن آن ترکیب فارسی خاص به واحدهای کوچکتر متوقف می‌شد و از این رو در کنار مقوله‌های دیگر به ناچار از NP هم برای نشان دادن گروه اسمی استفاده شد. علت افزایش پیشوند و پسوند هم این بوده که در طی فرایند برابرگزینی واژه‌های فارسی و انگلیسی، گاه در انگلیسی واژه مشخصاً قابل تقطیع بود و در معادل فارسی نیز می‌توانستیم پیشوند یا پسوندی را تقطیع کرده و به عنوان برابرنهاده یکی از اجزاء واژه انگلیسی قرار دهیم. لازم به ذکر است که در فرایند برچسب‌زنی امکان درج انواع و سطوح مختلفی از برچسب‌ها وجود دارد. برای مثال، در وهله نخست چنانچه واژه «کرمان» به عنوان یک اسم معرفی شود می‌توان در مرحله بعد نوع اسم را که در این مورد اسم خاص است مشخص کرد و ... . با این حال، با توجه به چارچوب پیشنهاده پژوهش حاضر و محدودیت‌های زمانی و نبود همکاران پژوهش، اضافه کردن سطوح دیگری از برچسب میسر نشد.

به عنوان یک نمونه دیگر از محدودیت‌های موجود می‌توان به عمومیت موضوعی پیکره تولید شده اشاره کرد. در حقیقت در این پژوهش هدف تولید یک پیکره از عناوین مقالات مجلات رتبه‌دار بوده است و از این رو از روش نمونه‌برداری سلسله‌مراتبی برای انتخاب عناوین از زیرحوزه‌های موضوعی استفاده شد تا عناوین از تمام حوزه‌های موضوعی موجود متناسب با اندازه هر زیرحوزه استخراج و در پیکره درج گردد. با این حال در این پژوهش لفظ «زیرحوزه» یکی از متغیرهای پژوهش محسوب نشد همچنانکه برچسبی برای تفکیک زیرحوزه‌ها تعریف نگردید. در حقیقت از زیرحوزه‌ها صرفاً در مرحله استخراج نمونه آماری از جامعه آماری استفاده به عمل آمد.

سرانجام هر چند هدف اولیه و مصوب پژوهش حاضر ارایه اطلاعات بسامدی از کاربرد کلمات توسط نویسندگان نبوده است، با این همه محقق ضمن انجام این نوع تحلیل بسامدی، فراوانی رخداد کلمات را در قالب یک ستون در پیکره درج کرد. البته باور بر این است که این اطلاعات بسامدی اطلاعاتی اولیه است که

می‌تواند در قالب پژوهش‌های دیگر دنبال شده و انواع تحلیل‌های بسامدی را در اختیار مخاطب قرار دهد. با این حال در این پژوهش به نمونه‌هایی مختصر از تحلیل‌های بسامدی اشاره شد (ر.ک. بخش‌های ۷ و ۹-۱).

۱۱- زمینه‌هایی برای مطالعه بیشتر

همواره انجام یک پژوهش افق‌های جدیدی را برای پژوهش بر روی محقق می‌گشاید. در حقیقت محقق حاضر در صدد خواهد بود در راستای پژوهش حاضر و برای تکمیل آن پژوهش‌های دیگری را نیز در آینده به انجام برساند که در این قسمت به برخی از آنها اشاره می‌شود.

افزایش عناوین جدید به ۱۰۰۰۰ جفت عنوان موجود، یکی از فعالیت‌هایی است که می‌تواند به عنوان گام بعدی پژوهش حاضر به انجام برسد. افزایش این عناوین به ۲۰۰۰۰ جفت یا حتی ۵۰۰۰۰ جفت می‌تواند الگوهای واژگانی و معادل‌گزینی جامع‌تری را پیش روی خوانندگان و محققان قرار دهد.

گسترش برچسب‌زنی انجام شده در پژوهش حاضر به تعداد بیشتری از برچسب‌ها و نیز لحاظ کردن زیر برچسب‌ها برای هر یک از برچسب‌های نحوی کلی یا استفاده از برچسب‌های جزئی‌نگر موجود برای فارسی نیز می‌تواند پژوهشی مکمل پژوهش حاضر محسوب گردد. برای مثال برای اسم انواع اسامی مانند اسم عام و خاص، ذات و معنی و ... و به همچنین برای فعل برچسب‌های جزئی‌تر نظیر فعل لازم و متعدی، ساده یا استمراری و .... همچنین می‌توان برچسب‌های دیگری نیز تعریف و اضافه کرد که از آن جمله می‌توان به برچسب حوزه موضوعی اشاره کرد. برای مثال، با این برچسب مشخص می‌شود که این واژه از مقالات حوزه موضوعی علوم انسانی گرفته شده یا از حوزه فنی و مهندسی، کشاورزی یا ... .

بررسی فراوانی‌محور واژه‌ها و نیز بررسی تنوع معادل‌گزینی نویسندگان دو موضوعی است که اگر چه محتوای اولیه آن در پیکره تولید شده وجود دارد اما به دلیل آنکه خود، تحلیل‌های فراوان، متعدد و مستقلی را می‌طلبد در این پژوهش صرفاً به گزارش آن (با تحلیل مختصر) بسنده شد. بدیهی است یکی از اهداف عمده محقق حاضر انجام پژوهش دیگر برای تحلیل بسامدی واژه‌ها و به خصوص ارایه تحلیلی جامع از رفتار معادل‌گزینی نویسندگان از فارسی به انگلیسی خواهد بود.

همچنین در این پژوهش جهت پیکره از فارسی به انگلیسی بوده است و از این رو پیکره تولید شده و مندرج در انتهای این گزارش، از فارسی به انگلیسی تنظیم شده است. طبیعی است امکان ارایه پیکره از جهت دیگر (از انگلیسی به فارسی) و نیز زبان‌های دیگر در قالب پژوهش‌های بعدی وجود خواهد داشت.

فهرست منابع فارسی

احمدی گیوی، ح.، و انوری، ح. (۱۳۷۱). *دستور زبان فارسی (چاپ دهم)*. تهران: انتشارات فاطمی.

اسلامی زاده، ز. (۱۳۹۴). تهیه و تدوین لیست واژگان آکادمیک زبانشناسی (پایان‌نامه کارشناسی ارشد، دانشگاه کاشان، دانشکده ادبیات و علوم انسانی، ۱۳۹۴).

پیکره موازی انگلیسی-فارسی با نام میزان ارائه شد. (۱۶ اردیبهشت ۱۳۹۲). خبرگزاری باشگاه خبرنگاران. بازیابی شده در ۲۳ خرداد ۱۳۹۶ از www.yjc.ir.

توکل، م. (۱۳۹۶). تولید پیکره افعال فارسی بر اساس کتابهای خوانداری سالهای اول تا دوازدهم و تهیه فهرست افعال متمایز هر سال با رویکردی بسامدی-مقایسه‌ای (پایان‌نامه کارشناسی ارشد، گروه آموزشی زبانشناسی رایانشی، مرکز منطقه‌ای اطلاع رسانی علوم و فناوری، ۱۳۹۶).

جهانگردی، ک. (۱۳۸۹). طرحی برای استفاده از پیکره‌های زبانی در فرهنگ نگاری (پایان‌نامه کارشناسی ارشد، پژوهشگاه علوم انسانی و مطالعات فرهنگی، پژوهشکده زبانشناسی، ۱۳۸۹).

دبیرخانه شورای عالی انقلاب فرهنگی. (۱۳۸۸). طرح جامعه پیکره زبان فارسی. بازیابی شده در ۱۲ مهر ۱۳۹۲ از www.prosody.ir/attachments/058_17-Tag.pdf.

شقاقی، و. (۱۳۹۵). *مبانی صرف*. تهران: سازمان مطالعه و تدوین کتب علوم انسانی دانشگاه‌ها (سمت).

صادقی، ع. ا. (۱۳۹۴). *فرهنگ املایی خط فارسی*. تهران: فرهنگستان زبان و ادب فارسی.

صفری، س. (۱۳۹۱). طراحی و ایجاد پیکره تولیدی زبان آموز فارسی (پایان‌نامه کارشناسی ارشد، دانشگاه علامه طباطبایی، دانشکده ادبیات و علوم انسانی، ۱۳۹۱).

عاصی، م. (۱۳۸۵). از پیکره زبانی تا زبانشناسی پیکره ای. *پژوهشگران*، شماره‌های ۸ و ۹ مرداد-آبان.

عاصی، م. (۱۳۸۹). پردازش دستوری زبان فارسی با رایانه. بازیابی شده در ۱۲ مهر ۱۳۹۴ از http://persianacademy.ir/UserFiles/File/D/01/D-01-03.pdf

عیار، م. (۱۳۸۹). کاربرد عبارات هم آیی در ترجمه انگلیسی بوف کور نوشته صادق هدایت: بررسی بر اساس زبانشناسی پیکره‌ای (پایان‌نامه کارشناسی ارشد، دانشکده زبانهای خارجی، دانشگاه اصفهان، ۱۳۸۹).

قندی، س. (۱۳۹۳). طرحی برای تهیه پیکره تاریخی متون نثر زبان فارسی از قرن پنجم تا هفتم هجری (پایان‌نامه کارشناسی ارشد، پژوهشگاه علوم انسانی و مطالعات فرهنگی، پژوهشکده زبانشناسی، ۱۳۹۳).

کاهانی، م.، و جکیان طوسی، ا. (۱۳۹۱). ارائه رهیافتی جدید برای تولید پیکره موازی انگلیسی-فارسی (پایان‌نامه کارشناسی ارشد، دانشکده مهندسی، دانشگاه فردوسی مشهد، ۱۳۹۱).

کشتکار، ح. (۱۳۹۱). ساخت پیکره دوزبانه موازی انگلیسی-فارسی و کاربرد آن در سامانه حافظه ترجمه (مبحثی در زبانشناسی پیکره ای) (پایان‌نامه کارشناسی ارشد، دانشگاه پیام نور استان تهران، مرکز پیام نور تهران، ۱۳۹۱).

کلباسی، ا. (۱۳۸۸). *ساخت اشتقاقی واژه در فارسی امروز (چاپ چهارم)*. تهران: پژوهشگاه علوم انسانی و مطالعات فرهنگی.

محمدی، س. ر. (۱۳۹۱). ساخت پیکره تطبیقی فارسی-انگلیسی و استخراج جملات موازی از آن (پایان‌نامه کارشناسی ارشد، دانشگاه الزهرا (س)، دانشکده فنی و مهندسی، ۱۳۹۱).

محمدی، م. (۱۳۸۹). تاثیر استفاده از پیکره موازی بر کیفیت ترجمه: یک مطالعه موردی (پایان‌نامه کارشناسی ارشد، دانشگاه علامه طباطبایی، دانشکده ادبیات فارسی و زبانهای خارجی، ۱۳۸۹).

محمدی، م.، و قاسم آقایی، ن. (۱۳۸۸). ارائه پیکره متنی موازی فارسی-انگلیسی با کاوش در ویکیپدیا. پانزدهمین کنفرانس بین المللی سالنه انجمن کامپیوتر ایران. تهران، انجمن کامپیوتر، مرکز توسعه فناوری نیرو. قابل بازیابی از طریق http://www.civilica.com/Paper-CSICC15-CSICC15_008.html.

نعمتی، ح. (۱۳۹۷). نرم‌افزار Advanced Conditional Sum (ارائه شده به شرکت همپا شیراز).

وبگاه دادگان. (۱۳۹۴). http://www.dadegan.ir/catalog.

وحدت زاده، س. (۱۳۹۱). بررسی پیکره بنیاد فعل مرکب بر مبنای کتابهای دستور زبان فارسی دبیرستانهای ایران. (پایان‌نامه کارشناسی ارشد، دانشگاه پیام نور استان تهران، مرکز پیام نور تهران، ۱۳۹۱).

English references

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

Baker, P. (2012). *Contemporary corpus linguistics*. New York: Continuum International Publishing Group.

Bijankhan, M., Sheykhzadegan, K., Bahrani, M., Ghayoomi, M. (2011). Lessons from building a Persian written corpus: Peykare. *Language Resources and Evaluation*, *45*(2), 143-164.

Espunia I Prat, A. (1994). Computational linguistics: A brief introduction. Retrieved June 6, 2018 from www.raco.cat/index.php/LinksLetters/article/download/49811/87789.

Farajian, M. A. (2011). PEN: Parallel English-Persian News Corpus. Retrieved Jan. 1st, 2018, from http://www.lidi.info.unlp.edu.ar/WorldComp2011-Mirror/ICA4953.pdf.

Goldwater, Sh. (2015). Introduction to computational linguistics: Introductory information. (PPT file). Retrieved May 21, 2018 from http://homepages.inf.ed.ac.uk/sgwater/teaching/lsa2015/lectures/day1-nup.pdf.

Heja, E. (2010). The role of parallel corpora in bilingual lexicography. In *Proceedings of the IREC Conference* (pp. 2798-2805). Retrieved May 23, 2018 from http://www.lrec-conf.org/proceedings/lrec2010/pdf/559_Paper.pdf.

Jabbari, F., Bakhshaei, S., Mohammadizadeh Ziabary, S. M., & Khadivi, Sh. (2009). Developing an open-domain English-Farsi translation system using AFEC: Amirkabir bilingual Farsi-English corpus. *10th Biennial Conference of the Association for Machine Translation in the Americas*. Retrieved Feb. 24,

2018, from http://www.researchgate.net/publication/234025183_Developing_an_Open-domain_English-farsi_Translation_System_Using_AFEC_Amirkabir_Bilingual_farsi-English_Corpus.

Kennedy, G. (2014). *An introduction to corpus linguistics*. New York: Routledge.

Li, G., Wu, C. H., & Vijay-Shanker, K. (2017). Noise reduction methods for distantly supervised biomedical relation extraction. *Proceedings of the BioNLP 2017 Workshop* (pp. 184-193), August 4, Vancouver, Canada.

McCarthy, M. & O'Keefe, A. (2010). What are corpora and how have they evolved? *The Routledge handbook of corpus linguistics* (Anne O'Keeffe and Michael McCarthy, eds.). New York: Routledge.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics, method, theory and practice*. UK: Cambridge University Press.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In S. D. Richardson, editor, *Machine translation: From research to real users*, 5th Conference of the Association of Machine Translation in the Americas (AMTA-02), volume 2499 of Lecture Notes in Computer Science, pp. 135-144, Berlin, Springer-Verlag, 2002.

Mosavi Miangah, T. (2009). Constructing a large-scale English-Persian parallel corpus. Meta: Translators' Journal, *54*(1), 181-188.

Pilevar, M. T., Faili, H., & Pilevar, A. H. (2011). TEP: Tehran English-Persian Parallel Corpus. Retrieved Jan. 3rd, 2016, from http://www.pilevar.com/taher/pubs/CICLING_2011_Pilehvars_faili.pdf.

Qasemizadeh, B., Rahimi, S., & Mohammadi Bakhtiari, B. (2014). The first parallel multilingual corpus of Persian: Toward a Persian BLARK. Retrieved on Sep. 22, 2015, from http://arxiv.org/ftp/arxiv/papers/1404/1404.4572.pdf.

Shamsfard, M. (2011). Challenges and open problems in Persian text processing. Retrieved June 26, 2015 from http://hnk.ffzg.hr/bibl/ltc2011/book/papers/MPLRL-6.pdf.

Stanford encyclopedia of philosophy (Computational Linguistics entry, 2014). Retrieved June 5, 2018 from https://plato.stanford.edu/entries/computational-linguistics/

The Handbook of pragmatics (2006). (Laurence R. Horn and Gregory Ward, eds.). Pragmatics and computational linguistics (Dan Jurafsky). Oxford: Blackwell Publishing.