مجله‌ی علمی پژوهش در علوم رایانه، شماره ۲۶، تابستان ۱۴۰۱، ص ۴۱- ۱۳

ISSN: 2745-3673
http://www.Rcsj.ir

مجله علمی پژوهش در علوم رایانه (سال هفتم)
شماره ۲۶ / تابستان ۱۴۰۱ / ص ۴۱-۱۳

معرفی واژه‌نامه پیکره‌محور فارسی به انگلیسی کلمات عناوین مقالات فارسی مجلات رتبه‌دار

محمد رضا فلاحتی قدیمی فومنی

دانشیار گروه پژوهشی زبانشناسی رایانه‌ای، مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، شیراز، ایران.

نام نویسنده مسئول:
محمد رضا فلاحتی قدیمی فومنی

تاریخ دریافت: ۱۴۰۱/۰۳/۰۱
تاریخ پذیرش: ۱۴۰۱/۰۵/۱۵

چکیده

هدف از انجام پژوهش حاضر آن بود تا: یک واژه‌نامه فارسی به انگلیسی پیکره‌محور بر اساس واژه‌های عناوین مقالات فارسی و معادل‌های انگلیسی آنها و همچنین یک فرهنگ اغلاط از اشکالات املایی، معنایی و مرتبط با مقوله نحوی واژگان (واژگان فارسی و معادل‌های انگلیسی آنها) هر دو در قالب ماشین‌خوان تهیه شود. با استفاده از روش نمونه‌برداری تصادفی سلسله‌مراتبی از واژه‌های فارسی موجود در ۱۰۰۰۰ عنوان مقاله مجلات معتبر فارسی و واژه‌های انگلیسی معادل آنها (برگرفته از ۱۰۰۰۰ عنوانِ معادل انگلیسی) استفاده به‌عمل آمد. از نرم‌افزار اکسل برای تولید واژه‌نامه (حاوی ۹ ستون) و فرهنگ اغلاط (حاوی ۸ ستون) استفاده شد. همچنین بر اساس مصوبات فرهنگستان و اصول فرهنگ‌نویسی شیوه‌نامه‌ای برای تدوین واژه‌نامه توسط محقق تهیه و پس از بازبینی و آزمایش بر روی نمونه کوچکِ داده‌ها و نیز اخذ نظر یک متخصص زبانشناس، مورد استفاده قرار گرفت. در نهایت، واژه‌نامه‌ای با ۱۱۹۴۹ مدخل فارسی و ۱۷۹۹۱ معادل انگلیسی تولید شد. در این واژه‌نامه در مجموع ۴۲ بار از علامت ر.ک. و ۱۰۹ بار از علامت مساوی استفاده شد. بیشترین تعداد کلمات عناوین مقالات، به مقوله نحوی اسم (۸۰۷۲ مورد، ۶۷/۵ درصد) مربوط بودند. مقوله نحوی صفت (ص.) با ۳۰۰۹ مورد (۲۵/۱۸ درصد) و گروه های اسمی با ۵۳۸ مورد (۴/۵ درصد) در جایگاه سوم قرار گرفت. سایر مقوله‌های نحوی در مجموع کمتر از ۳ درصد از کل مدخل‌ها را تشکیل داد. فرهنگ اغلاط پژوهش حاضر نیز در ۲۹۶ ردیف و ۲۶۸ واژه متمایز مدخل فارسی تولید شد. از این ۲۹۶ ردیف ۲۶۲ ردیف یعنی ۸۸/۵۲ درصد به خطای نوع سوم (خطای املایی) مربوط بود. پس از آن خطای معنایی و خطای نحوی هر یک با ۱۷ مورد (۵/۷۴ درصد) به صورت مشترک در جایگاه دوم قرار گرفتند. پژوهش حاضر با تولید فرهنگ اغلاط، نیاز به توجه بیشتر به رسم‌الخط فارسی و انگلیسی نشریات را به‌خصوص در بحث خطاهای املایی و تایپی منعکس کرد.

واژگان کلیدی: زبان‌شناسی پیکره‌ای، واژه‌نامه ماشین‌خوان، فرهنگ اغلاط، تحلیل خطا، خطاهای دستوری، خطاهای املائی.

ISSN: 2745-3673
http://www.Rcsj.ir

مقدمه

فرهنگ‌های لغت انواع و اقسام مختلف دارد. یکی از این نوع فرهنگ‌ها، فرهنگ‌های لغت پیکره‌محور است که تهیه‌کننده واژه‌نامه برای تهیه مدخل‌ها از یک پیکره‌ی واژگانی و محتوای لغات آن استفاده می‌کند. در این نوع واژه‌نامه‌ها برای چیدمان معادل‌های مختلف یک واژه نیز از معیارهای مختلفی استفاده می‌شود. معمولاً متداول‌ترین و مرسوم‌ترین معادل‌ها در ابتدا درج می‌شوند. برای تعیین متداول‌ترین معادل‌ها نیز معیارهای گوناگونی وجود دارد که یکی از آنها به‌خصوص در مباحث زبانشناسی پیکره‌ای به فراوانی رخداد معادل در پیکره برمی‌گردد. برای نمونه، اگر برای یک واژه دو معادل وجود داشته باشد و فراوانی رخداد یک معادل در پیکره عدد ۱۰۰۰ و معادل دیگر ۱۰ باشد غالباً (و نه همیشه) فرهنگ‌نویس معادل با فراوانی ۱۰۰۰ را با رعایت اصول نرمال‌سازی پیش از معادل با فراوانی ۱۰ ذکر می‌کند. بر این اساس هدف نخست از انجام پژوهش حاضر آن بود تا ضمن استفاده از داده‌های موجود در پیکره فلاحتی (۱۳۹۲) (حاوی ۱۰۰۰۰ عنوان مقاله فارسی و ۱۰۰۰۰ عنوان معادل انگلیسی مجلات رتبه‌دار وزارت عتف) واژه‌نامه‌ای پیکره‌محور تهیه شود که در آن هر یک از واژه‌های فارسی موجود در عنوان، یک مدخل فرض شده و معادل‌های انگلیسی هر واژه نیز ذیل آن مدخل ثبت گردد. هدف دیگر آن بود تا از اشکالات املایی، معنایی و مرتبط با مقوله نحوی واژگان (واژگان فارسی و معادل‌های انگلیسی آنها) یک فرهنگ اغلاط تهیه شود. بدیهی است تولید این واژه‌نامه می‌تواند برای نویسندگان مقالات و آن دسته از محققانی که در حال تألیف مقاله علمی به زبان انگلیسی هستند مفید واقع شود. همچنین فرهنگ اغلاط تولیدشده نیز می‌تواند در امر سیاست‌گذاری نشریات در وزارت عتف و بهداشت مفید واقع شود.

پیشینه پژوهش

هدف از فرهنگ‌نویسی تولید انواع واژه‌نامه‌ها است که هر کدام بسته به نوع و ماهیت خود به مجموعه‌ای از مخاطبان برای ابهام‌زدایی از مفهوم واژه یاری می‌رساند (دیما[1]، ۲۰۱۲). دیما (۲۰۰۸) واژه‌نامه را دارای منفعتی دوسویه می‌داند که هم در درک و هم در تولید متن به کمک کاربر می‌آید. همچنین مارتین[2] (۱۹۸۹) از میان انواع واژه‌نامه‌ها برای واژه‌نامه‌های تخصصی دو یا چندزبانه اهمیت خاصی قائل است. او بر این باور است که این نوع واژه‌نامه‌های تخصصی می‌توانند واژگان و چارچوب موضوعی رشته‌ها و حوزه‌های تخصصی را به خوبی ترسیم کنند.

برگنهولتس[3] و آگربو[4] (۲۰۱۵) در تمایز بین واژه‌نامه‌های چاپی و برخط بیان می‌دارند که بیشتر واژه‌نامه‌های چاپی دارای کارکردهای چندگانه‌اند و از این رو غالباً از آنها بر روی تلفن همراه و تبلت استفاده نمی‌شود چرا که فضای بسیار زیادی را می‌طلبد و لذا هزینه‌های جستجوی داده‌ها در آنها بیشتر است. در عوض واژه‌نامه‌های الکترونیک با تمرکز بر روی حوزه‌های خاص از حجم کمتر و سرعت بیشتر برخوردارند و به همین دلیل در ابزار الکترونیک بیشتر مورد استفاده قرار می‌گیرند. البته، فرت[5] و دولینگر[6] (۲۰۲۱) یکی از نقاط ضعف واژه‌نامه‌های برخط را درج آگهی‌های زیاد تبلیغاتی در آنها دانسته، به این نکته نیز اشاره می‌کنند که تغییر مستمر و حتی غیرمشخص و گاه غیرمستند محتوای واژه‌نامه‌های برخط می‌تواند سبب مخدوش شدن اعتبار واژه‌نامه شود. با این حال آنها کاربرپسند بودن و سهولت جستجو در واژه‌نامه‌های چاپی را از ویژگی‌های مثبت این نوع واژه‌نامه‌ها می‌دانند.

برخی محققان به کاربرد واژه‌نامه‌ها در حوزه‌های مختلف پرداختند. برای نمونه ژانگ[7]، زو[8] و ژانگ[9] (۲۰۲۱) اثر استفاده از واژه‌نامه‌ها را بر فراگیری واژگان زبان دوم بررسی کردند. آنها با کار بر روی نمونه‌ای حاوی ۳۴۷۵ آزمودنی متعلق به ۸۷ نمونه

---

[1] . Dima
[2] . Martin
[3] . Bergenholtz
[4] . Agerbo
[5] . Ferrett
[6] . Dollinger
[7] . Zhang
[8] . Xu

ISSN: 2745-3673
http://www.Rcsj.ir

مستقل دریافتند که در تمامی گروه‌ها، استفاده از واژه‌نامه در فراگیری واژگان زبان دوم تأثیر مثبت داشته است. آنها همچنین متغیرهای متعددی نظیر نوع واژه‌نامه، شکل واژه‌نامه، محیط یادگیری و ... را نیز بررسی کرده و به این نتیجه رسیدند که چنین عواملی نیز بر میزان یادگیری واژگان تأثیر مثبت دارند.

تهیه واژه‌نامه فرایند پیچیده و طولانی است و به طی مراحل مختلف نیاز دارد. از نظر چرماک[10] (۲۰۱۰) برای تهیه یک واژه‌نامه، فرهنگ‌نویس لازم است مجموعه‌ای از سیاست‌ها را اتخاذ نماید. او در صفحه ۵۶۰ این سیاست‌ها را در سه دسته کلی مطرح می‌کند. طبق نظر او گام نخست به تعیین منابعی مربوط است که فرهنگ‌نویس، واژگان را از آنها استخراج می‌کند. این منابع باید با دقت زیاد انتخاب شوند و نماینده خوبی از حوزه موضوعی واژه‌نامه مزبور باشند. دومین نکته‌ای که چرماک ذکر می‌کند تعیین نوع واژه‌نامه و مدخل‌های آن است. برای مثال، فرهنگ‌نویس باید تعیین کند که می‌خواهد فرهنگی یک‌زبانه بنویسید یا دوزبانه یا حتی چندزبانه. در صورتی که هدف نوشتن فرهنگ دوزبانه است آیا فرهنگ مورد نظر یک طرفه (برای نمونه فارسی به انگلیسی) است یا دو طرفه (فارسی به انگلیسی و انگلیسی به فارسی)؟ آیا هدف تولید یک فرهنگ جامع است یا یک فرهنگ فشرده؟ آیا واژگان زبان معاصر مدنظر است یا واژگان تاریخی یا ترکیبی از هر دو (هم‌زمانی و درزمانی)؟ ضمن آنکه اصولی کلی و مشترک نیز در نگارش واژه‌نامه باید پیوسته مدنظر فرهنگ‌نویس قرار گیرد نظیر اینکه داده‌ها باید نماینده حوزه موضوعی واژه‌نامه باشد و نباید تجویزی عمل شود بدین معنی که در ذکر و تعریف واژگان باید از برداشت شخصی اجتناب شده و نوع استفاده از لغات، توسط سخنوران زبان رصد و در واژه‌نامه ذکر گردد. در مقاله حاضر نوع فرهنگ لغت، یک واژه‌نامه ماشین‌خوان دوزبانه (فارسی به انگلیسی) یک طرفه (از فارسی به انگلیسی) و هم‌زمانی تعیین شد بدین معنی که از لغات موجود در مقالات اخیر وزارت عتف در این پژوهش استفاده به عمل آمد. همچنین هدف اگرچه در درازمدت تولید فرهنگ جامع است اما در این مرحله هدف تولید نمونه‌ای از این واژه‌نامه در مقیاس کوچکتر بوده است. با این همه سعی شد با استفاده از روش نمونه‌گیری تصادفی سلسله‌مراتبی برای تحقق اصل نماینده بودن[11] نمونه آماری تلاش شود. نکته دیگر اینکه در این فاز از تولید واژه‌نامه، هدف ارائه بافت و مثال برای کاربرد واژه‌ها نبوده و صرفاً به بیان مقوله نحوی، معادل(های) واژگانی انگلیسی و صورت(های) معادل (در صورت وجود) تمرکز شد.

سپس چرماک به معیار سوم (مخاطب واژه‌نامه) می‌پردازد. مخاطب می‌تواند خاص یا عام باشد. همچنان که برای گروه‌های سنی مختلف نیز نوع واژه‌نامه فرق می‌کند به طوری که برای سنین پایین‌تر واژه‌های عمومی‌تر و ساده‌تر غالب است حال آنکه در یک واژه‌نامه تخصصی مثلاً پزشکی با مجموعه‌ای از واژگان تخصصی سروکار خواهیم داشت. از این منظر، برای پژوهش حاضر نویسندگان مقالات علمی و محققان و دانشجویان تحصیلات تکمیلی، مخاطبان واقعی واژه‌نامه در نظر گرفته شدند. بنابراین، این واژه‌نامه را باید نوعی واژه‌نامه تخصصی با دامنه محدود به حساب آورد و هدف آن هرگز شامل کردن تمام واژه‌های زبان فارسی نبوده است. بلکه هدف صرفاً گردآوری واژگان موجود در عناوین ده هزار جفت مقاله علمی معتبر وزارت عتف بوده است. نکته دیگر آن است که با توجه به حجم متوسط داده ها، معادل‌های انگلیسی با فراوانی پایین نیز به سه دلیل از جریان تحقیق حذف نشدند: دلیل اول اینکه به هر حال این کلمات در عنوان مقالات ظاهر شده بودند که یکی از بخش‌های مهم مقاله است به گونه‌ای که نمایه‌سازی‌های متعددی صرفاً بر اساس واژگان عنوان تولید می‌شود مانند نمایه‌های کوئیک[12]، کوئک[13] و کواک[14]. دوم آنکه به دلیل اینکه حجم داده‌ها به خاطر محدودیت‌های پژوهش بزرگ نبوده است، امکان افزایش فراوانی این واژه‌ها در حجم بزرگتری از داده وجود دارد. سوم اینکه هدف تولید واژه‌نامه‌ای پیکره‌محور بوده و نه یک واژه‌نامه بسامدی و استفاده از لفظ فراوانی صرفاً برای تعیین ترتیب معادل‌های متعدد یک واژه فارسی بوده است.

---

9 . Zhang
10 . Cermak
11 . Representativeness
12 . KWIC
13 . KWAC
14 . KWOC

ISSN: 2745-3673
http://www.Rcsj.ir

واژه‌نامه‌های پیکره‌محور

واژه‌نامه پیکره‌محور در ساده‌ترین تعریف به آن‌دسته از واژه‌نامه‌هائی اطلاق می‌شود که در آنها از یک یا چند پیکره برای استخراج واژگان متمایز استفاده می‌شود. برای هر واژه متمایز نیز بسته به نوع پیکره (واژگانی است یا متنی)، انواع مختلفی از داده‌ها ثبت و درج می‌شود. برای نمونه، معادل واژگانی در یک یا چند زبان، جملات نمونه، تلفظ کلمات و ... . در واژه‌نامه‌های پیکره‌محور معادل‌های عنوان شده به نوع ساختار پیکره‌ای بستگی دارند که معادل‌ها از آن استخراج شده‌اند. برای نمونه اگر از یک پیکره متنی معاصر استفاده شود به ندرت واژه‌ها و معادل‌های منسوخ در آن ظاهر می‌شود. همچنین اگر کل واژه‌های زبان را یک ظرفیت کلی بدانیم واژه‌نامه پیکره‌محور آن بخش از واژگان زبان را پوشش می‌دهد که در متون گردآوری شده در پیکره متبلور است. البته، پیکره چه عمومی و چه تخصصی، چه درزمانی و چه همزمانی باید نماینده حوزه موضوعی ادعا شده باشد تا بتوان محتوای واژگانی آن را نماینده مناسبی از واژگان حوزه موضوعی مزبور قلمداد کرد.

در کنار اطلاعات واژگانی، وجود پیکره ماشین‌خوان به فرهنگ‌نویس کمک می‌کند تا اطلاعات عددی نیز به واژه‌نامه اضافه کند. یکی از شایع‌ترین انواع اطلاعات عددی به فراوانی رخداد معادل‌ها در یک پیکره برمی‌گردد. میزان استفاده از اطلاعات آماری نیز در واژه‌نامه‌ها متفاوت است. به عبارت دیگر، محقق می‌تواند از این اطلاعات عددی در سطح محدود و صرفاً با هدف مرتب کردن معادل‌ها استفاده کند یا اینکه می‌تواند با تحلیل‌های آماری مختلف در انتخاب واژگان نیز از الگوهای بسامدی استفاده نماید. در پژوهش حاضر مبنا بر حالت اول قرار گرفت و از اطلاعات بسامدی واژگان صرفاً برای مرتب کردن معادل‌های انگلیسی کلمات استفاده شد و سپس معادل‌های با فراوانی یکسان بر اساس ترتیب حروف الفبا مرتب شدند. دلیل این انتخاب هم حجم متوسط داده‌های پژوهش بوده است که بر آن اساس نمی‌توان معادل دارای بسامد بیشتر را الزماً به معنی معادل مناسب‌تر تلقی کرد.

پژوهش‌های انجام‌شده

درخصوص واژه‌نامه‌های پیکره‌محور کارهای زیادی در دنیا و ایران انجام شده که در ادامه به نمونه‌هائی اشاره می‌شود. کورنای[15] و همکاران (۲۰۰۶) شیوه تهیه واژه‌نامه‌های پیکره‌محور بسامدی را تشریح کردند. این محققان اذعان داشتند این نوع واژه‌نامه‌ها در طرح‌ریزی آزمایش‌های روان‌شناسی زبان و نیز فناوری‌های زبانی نقش مهمی را ایفا می‌کنند. از این رو یک واژه‌نامه پیکره‌محور با دسترسی آزاد و رایگان را (که در هر دو حوزه فوق دارای کاربرد می‌باشد) برای زبان مجاری تهیه و معرفی کردند. کونرای و همکاران در اثر دیگر خود، نظامی کاملاً خودکار را برای دسته‌بندی موضوعی صفحات وب معرفی کردند (کونرای و همکاران، ۲۰۰۶).

برخی محققان نیز تاریخچه تهیه واژه‌نامه‌های پیکره‌محور و بسامدی را ارائه دادند. برای نمونه‌ای از این محققان می‌توان به پژوهش چِرماک و کرِن[16] (۲۰۰۵) و همچنین چِرماک (۲۰۱۱) اشاره کرد. نویسندگان این دو اثر ضمن اشاره به تاریخچه واژه‌نامه‌های پیکره‌محور بسامدی، بیان داشتند که نیاز خاصی به این نوع واژه‌نامه‌ها وجود دارد. آنها تأکید کردند که در دسترس بودن پیکره‌ها و نرم‌افزارهای مختلف، ساخت این نوع واژه‌نامه‌ها را بسیار تسهیل کرده است. آنها ضمن توصیف فرایند ساخت واژه‌نامه پیکره‌محور بسامدی خود، مشکلاتی را که در حین کار با آن مواجه بوده‌اند تشریح نموده و راه‌حل‌های خود را نیز ارائه دادند. آنها از فراوانی رخداد هر واژه در پیکره برای تعیین ترتیب معادل‌ها در هر مدخل استفاده کردند و معادل‌های دارای فراوانی یکسان را بر اساس ترتیب الفبایی ردیف کردند. این شیوه درست همان شیوه‌ای است که در پژوهش حاضر نیز مورد استفاده قرار گرفت. در پژوهشی دیگر دیویس[17] و گاردنر[18] (۲۰۱۰) واژه‌نامه پیکره‌محور بسامدی انگلیسی آمریکایی معاصر را تولید کردند. این واژه‌نامه بیش از ۵۰۰۰ واژه پربسامد موجود در انگلیسی آمریکایی معاصر را در خود دارد و هدف از نگارش آن این بوده تا واژه‌های پرکاربرد زبان در دسترس زبان‌آموزان قرار گیرد. اطلاعات این واژه‌نامه از پیکره‌ای حاوی ۳۸۵ میلیون واژه

15 . Kornai
16 . Kren
17 . Davies
18 . Gardner

ISSN: 2745-3673
http://www.Rcsj.ir

گرفته شده است. این پیکره حاوی حدود ۱۵۰۰۰۰ متن (گفتاری و نوشتاری) بوده که از حوزه‌هایی چون رادیو و تلوزیون، سینما، کتاب، روزنامه، مجلات معروف و ۱۰۰ نشریه علمی دانشگاهی گرفته‌شده‌اند. به عنوان چند کار کلاسیک، دروشِر[۱۹]، میرون[۲۰]، پَتن[۲۱] و پرَت[۲۲] (۱۹۷۳) ضمن برشمردن تفاوت میان واژه‌نامه‌های پیکره‌محور بسامدی و سایر انواع واژه‌نامه‌ها، روند تحول در تولید این نوع واژه‌نامه‌ها و شیوه شمارش واژه‌ها را تشریح کردند. برای نمونه آنها تحلیل‌های بسامدی را برای تعیین هم‌آیی واژگان بسیار مهم ارزیابی کرده، بیان داشتند در تعیین هم‌آیی واژه‌ها اطلاعات بسامدی به عنوان یک عامل مهم درنظر گرفته می‌شود. در همین خصوص، میرون و پرَت (۱۹۷۳) نیز در کتاب خود برای محاسبه فراوانی واژگان و نیز تولید فهرست‌های بسامدی دستورالعملی کامل را ارائه دادند.

کارهای انجام شده در ایران نیز تنوع زیادی دارد و به حوزه‌های مختلفی مربوط می‌شود که از آن‌جمله می‌توان به حوزه‌های علوم دینی، زبان، ادبیات و متون علمی اشاره کرد که در ادامه به نمونه‌هائی اشاره می‌شود.

به عنوان نمونه‌ای از کارهائی که در حوزه علوم دینی به انجام رسیده می‌توان به واژه‌نامه پیکره‌محور بسامدی رساله القدس نوشته سپنج (۱۳۵۵) اشاره کرد. سپنج (۱۳۵۵) برای تهیه این واژه‌نامه، ابتدا واژه‌ها را بر اساس حروف الفبا مرتب کرد. سپس با استفاده از دو خط، فراوانی رخداد هر واژه را جلوی آن نوشت. همچنین در کنار هر واژه از دو عدد دیگر نیز استفاده کرد که عدد اول شماره صفحه‌ای را که واژه در آن موجود بوده نشان می‌دهد و عدد دوم نشان‌دهنده شماره خطی است که آن واژه در آن صفحه ظاهر شده است. این دو عدد با دونقطه از هم جدا شده‌اند.

یا در حوزه ادبیات می‌توان به کارهائی چون تهیه فرهنگ بسامدی دیوان ناصرخسرو (دهقانی، ۱۳۵۴)، واژه‌نامه بسامدی، توصیفی و ریشه‌شناختی منتخبی از اندرزنامه‌های پهلوی (رحیمی، ۱۳۸۸)، بررسی بسامدی و معنائی واژگان و ترکیبات عربی در شعر رودکی (مارانی، ۱۳۹۲)، فرهنگ بسامدی صور خیال در دیوان فرخی سیستانی (ابراهیمی، ۱۳۹۳) و فرهنگ بسامدی صور خیال در غزلیات سنائی (ملکی، ۱۳۹۷) اشاره کرد که هر یک از این کارها بر اساس یک پیکره انجام شده است. دهقانی (۱۳۵۴) در پژوهش خود موضوعاتی از قبیل سبک زبانی، عادات زبانی شاعر و همچنین حجم واژگانی شاعر را مورد بررسی آماری قرار داد. در کنار مباحث دیگر، محقق کوشید تا فراوانی رخداد واژگان را در دیوان ناصرخسرو مشخص کند. برای انجام این کار ابتدا واژگان متمایز را فهرست کرد و سپس فراوانی هر واژه را محاسبه و جلوی هر واژه درج نمود. این همان شیوه‌ای بوده است که برای درج اطلاعات بسامدی لغات در پژوهش حاضر نیز توسط محقق مورد استفاده قرار گرفت که نظیر آن در کار مارانی (۱۳۹۲) نیز دیده می‌شود. به طریقه مشابه، رحیمی (۱۳۸۸) ضمن اشاره به گنجینه گسترده واژگان فارسی میانه و نیاز به فرهنگ جامع ریشه‌شناختی واژگانِ این دوره، پژوهش خود را به انجام رساند و در این راه بر روی واژگان منتخبی از اندرزنامه‌های پهلوی تمرکز کرد. واژه‌نامه او علاوه بر اطلاعات بسامدی واژگان، اطلاعات ریشه‌شناختی کلمات را نیز ثبت کرد و برای هر واژه فارسی میانه، معادل‌های آن در زبان‌های ایرانی باستان، فارسی باستان، اوستائی، فارسی میانه ترفانی، پارتی، پازند و فارسی نو ذکر شد. همچنین مارانی (۱۳۹۲) با استفاده از اشعار رودکی به عنوان داده اولیه، واژگان و ترکیبات عربی موجود در این اشعار را از نظر بسامدی و معنایی تحلیل کرد. نویسنده ضمن اشاره به وام‌گیری واژگانی فارسی از عربی به خصوص در قرن چهارم و پنجم، تأثیر این واژه‌ها را بر روی شعر رودکی با استفاده از روشی توصیفی-تحلیلی بررسی کرد. او ضمن استخراج واژه‌های فارسی و عربی موجود در اشعار، به بررسی تغییرات لفظی و معنائی واژگان پرداخت و از واژگان موجود، تحلیل بسامدی ارائه کرد. در تحقیقی دیگر ابراهیمی (۱۳۹۳) فرهنگ بسامدی صور خیال در دیوان فرخی سیستانی را تولید کرد. او در این پژوهش انواع صور خیال را مشخص کرد و سپس مصداق‌های این صور را در دیوان فرخی جستجو کرد. بر اساس گزارش او تشبیه فراوان‌ترین نمونه از صور خیال در دیوان فرخی سیستانی بوده است و کنایه، استعاره و تشخیص در رده‌های بعدی قرار گرفتند. به عبارت دیگر تحلیل بسامدی او به شناسایی صور خیال و درج تعداد رخداد هر صورت محدود شد. ملکی (۱۳۹۷) کاری مشابه ابراهیمی (۱۳۹۳) را به انجام رساند با این تفاوت که به جای استفاده از دیوان فرخی سیستانی از غزلیات سنائی به

[19] . De Rocher
[20] . Miron
[21] . Patten
[22] . Pratt

ISSN: 2745-3673
http://www.Rcsj.ir

عنوان داده‌های پژوهش خود استفاده کرد. او بر روی صور بیانی تشبیه، استعاره، مجاز، تمثیل، کنایه و صفت هنری تمرکز کرد و سعی کرد با استخراج داده‌ها از غزلیات سنائی، فراوان‌ترین صور خیال را شناسائی کند. نتایج پژوهش او نشان داد که تشبیه فراوان‌ترین صور خیال بوده و استعاره و کنایه در جایگاه بعدی قرار گرفتند.

در حوزه متون علمی نیز به عنوان یک نمونه می‌توان به پژوهش فلاحتی قدیمی فومنی (۱۳۹۲) اشاره کرد که واژه‌نامه برگردان نام و نام‌خانوادگی نویسندگان خارجی نوشته‌شده با حروف انگلیسی به فارسی با استفاده از تحلیل رخدادمحور را تولید کرد. در پژوهش او در مجموع، ۹۶۸ نام و نام خانوادگی از کتاب فهرست مستند اسامی مشاهیر و مؤلفان (ویراست سوم) استخراج گردید. سپس صورت‌های احتمالی هر نام توسط محقق فهرست و در موتور جستجوی گوگل بازیابی شد. در مرحله بعد، فراوانی رخداد هر صورت واژگانی ثبت و ضبط گردید. در پایان، ۹۶۸ واژه مزبور در قالب یک واژه‌نامه فهرست شد و ذیل هر مدخل تمامی صورت‌های واژگانی ممکن بر اساس فراوانی رخداد هر یک ثبت شد. برای نمونه، صورت‌های واژگانی مرتبط با نام «جان اشتاین بک» در زیر آورده شده است:

جان اشتاین بک (۴۴۴۰۰۰)
جان اشتاینبک (۷۵۱۰۰)
جان اشتین بک (۷۳۸۰)
جان استاین بک (۶۱۰۰)
جان استاینبک (۵۸۶۰)
جان استین بک (۱۳۹۰)

پژوهش حاضر قالبی مانند واژه‌نامه برگردان نام و نام‌خانوادگی نویسندگان خارجی نوشته‌شده با حروف انگلیسی به فارسی با استفاده از تحلیل رخدادمحور دارد با این تفاوت که در آن کار تنوع نگارشی اسامی نویسندگان خارجی به زبان فارسی تحلیل و ثبت شد درحالی که در واژه‌نامه پیکره‌محور حاضر، کلمات فارسی موجود در عناوین ۱۰۰۰۰ عنوان مقاله فارسی مجلات رتبه‌دار وزارت عتف به عنوان مدخل تعیین گردید و ذیل هر واژه فارسی تمام معادل‌های انگلیسی هر کلمه فارسی (برگرفته از ۱۰۰۰۰ عنوان معادل انگلیسی) درج شد. همچنین یکی دیگر از مواردی که پژوهش حاضر را از پژوهش‌های قبلی متمایز می‌کند تهیه فرهنگ اغلاط توسط محقق است که در آن صورت‌های خطای موجود در پیکره که در واژه‌نامه تصحیح شده بودند به همراه صورت درست این دسته از کلمات، فهرست شد.

روش پژوهش

این پژوهش در سطح کلان به حوزه زبانشناسی پیکره‌ای و در آن نیز به زیرحوزه فرهنگ‌نگاری مربوط می‌شود که البته در حوزه فرهنگ‌نگاری، محتوای پژوهش حاضر به نگارش فرهنگ تخصصی فارسی به انگلیسی غیرتوصیفی و پیکره‌محور معاصر مربوط می‌شود: تخصصی به این دلیل که از عناوین مقالات علمی مربوط به حوزه‌های (فنی مهندسی)، (علوم انسانی)، (علوم پزشکی)، (علوم کشاورزی)، (هنر و معماری)، (علوم پایه)، (دامپزشکی) و (منابع طبیعی) در نگارش واژه نامه استفاده شد؛ فارسی به انگلیسی از آن جهت که در هر مدخل، واژه رأس فارسی است و ذیل آن معادل(های) انگلیسی هر واژه فارسی ارائه می‌شود؛ غیرتوصیفی از آن جهت که برای هر واژه فارسی صرفاً معادل واژگانی انگلیسی و نه توصیف، توضیح و تشریح ارائه می‌شود؛ پیکره‌محور از آن جهت که محتوای خام و اولیه پژوهش حاضر از پیکره تولیدشده توسط فلاحتی (۱۳۹۲) گرفته شده است که پس از پالایش و تجمیع (مطابق دستورالعمل تهیه شده) در پژوهش حاضر مورد استفاده قرار گرفت. سرانجام معاصر به این دلیل که منبع استخراج عناوین مقالات، مجلات اخیر و معتبر وزارت عتف و موجود در وبگاه مرکز منطقه‌ای بوده و منابع متقدم بررسی نشده است.

ISSN: 2745-3673
http://www.Rcsj.ir

جامعه و نمونه آماری

نمونه آماری در پژوهش حاضر داده‌های حاصل از پژوهش فلاحتی قدیمی فومنی (۱۳۹۲) در خصوص پیکره موازی بوده است. به عبارت دیگر، از واژه‌های فارسی موجود در ۱۰۰۰۰ عنوان مقاله فارسی و واژه‌های انگلیسی معادل آنها (برگرفته از ۱۰۰۰۰ عنوان معادل انگلیسی) استفاده به عمل آمد. این داده ها از کل عناوین مقالات موجود در وبگاه مرکز منطقه‌ای استخراج گردید. عدد ۱۳۲۰۸۶ به عنوان جامعه آماری عناوین مقالات تعیین شد. ضمناً با توجه به اینکه این تعداد عنوان حوزه‌های موضوعی مختلفی را چون فنی مهندسی، علوم انسانی، علوم پزشکی، علوم کشاورزی، هنر و معماری، علوم پایه، دامپزشکی و منابع طبیعی پوشش می‌داد، برای پوشش منطقی و متناسب مقالات حوزه‌های مختلف در نمونه آماری، از روش نمونه‌برداری تصادفی چندمرحله‌ای استفاده شد.

با توجه به اینکه جامعه آماری در این پژوهش ۱۳۲۰۸۶ جفت‌عنوان فارسی و انگلیسی بوده بر اساس جدول کرجسی و مورگان (۱۹۷۰) حجم نمونه برابر با ۳۸۳ جفت‌عنوان خواهد شد که این حجم نمونه برای کارهای واژه‌نامه‌ای بسیار محدود است. به همین دلیل عامدانه الگوی کرجسی و مورگان (۱۹۷۰) به کنار گذاشته شد. محقق ۱۰۰۰۰ جفت‌عنوان را به عنوان نمونه نهایی انتخاب شد که ۷/۶ درصد از کل جامعه آماری را تشکیل می‌دهد. سپس با استفاده از روش نمونه برداری تصادفی چندمرحله ای از هر حوزه تعداد عنوان متناسب با اندازه آن حوزه را به شرح جدول ۱ انتخاب کرد.

جدول ۱ . 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	=۱۰۰۰۰/۱۳۲۰۸۶*۱۰۷۲۲	۸۱۲
۴	=۱۰۰۰۰/۱۳۲۰۸۶*۵۵۱۴۶	۴۱۷۵
۹	=۱۰۰۰۰/۱۳۲۰۸۶*۳۴۷۱۰	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

*. *(حجم نهایی نمونه/حجم جامعه آماری کل) * تعداد کل عناوین در هر زیرحوزه = تعداد عناوین هر زیرحوزه در نمونه نهایی*. این فرمول از http://www.statisticshowto.com/stratified-random-sample گرفته شد.

ابزار گردآوری اطلاعات

برای گردآوری واژه‌های فارسی و معادل‌های انگلیسی کلمات فارسی از اطلاعات موجود در پیکره فلاحتی قدیمی فومنی (۱۳۹۲) استفاده شد و بنابراین نخستین ابزار مورد استفاده در این پژوهش، پیکره موازی تولیدشده توسط فلاحتی قدیمی فومنی (۱۳۹۲) بود. نرم‌افزار ACS تهیه شده توسط نعمتی (۱۳۹۸) ابزار دومی بود که برای محاسبه فراوانی ساده هر صورت واژگانی از آن استفاده شد. لازم به ذکر است در پژوهش حاضر به تولید فراوانی تجمعی صورت‌های انگلیسی واحد و نیز تجمیع صورت‌های تکراری نیاز بود که این کار به صورت دستی انجام شد. در پیکره پژوهش فلاحتی قدیمی فومنی (۱۳۹۲) در مجموع ۹۸۰۳۹ واژه فارسی (و به همین میزان معادل انگلیسی آنها) وجود داشت که از این تعداد ۲۴۹۰۹ واژه فارسی متمایز بوده‌اند. همین واژه‌های فارسی به همراه معادل انگلیسی آنها به عنوان داده خام مبنای تحلیل در پژوهش حاضر قرار گرفت. شیوه‌نامه تدوین مدخل‌ها نیز مورد بعدی بود که برای صرفه‌جویی در حجم مقاله صرفاً در بخش تدوین واژه‌نامه آورده شده است. همچنین از نرم‌افزار اکسل به عنوان زیرساخت ارائه واژه‌نامه و نیز فرهنگ اغلاط استفاده شد.

مجله‌ی علمی پژوهش در علوم رایانه، شماره ۲۶، تابستان ۱۴۰۱، ص ۴۱- ۱۳

ISSN: 2745-3673
http://www.Rcsj.ir

مراحل انجام کار

برای انجام پژوهش مراحل زیر به ترتیب به انجام رسید. ابتدا فایل اکسل پیکره تولیدشده در پژوهش فلاحتی قدیمی فومنی (۱۳۹۲) مورد استفاده قرار گرفت. سپس تمامی موارد ذکرشده در شیوه‌نامه مورد به مورد بر روی این داده‌های خام اعمال شد تا فایل اکسل واژه‌نامه تولید گردد. در نهایت خروجی اکسل و پی دی اف از واژه‌نامه گرفته شد. در پایان نیز اطلاعات آماری مختصر از محتوای واژه‌نامه ارائه گردید. در مورد فرهنگ اغلاط نیز پس از مشاهده و رصد خطاها در واژه‌نامه، مطابق شیوه‌نامه، موارد در واژه‌نامه اصلاح شد اما اطلاعات مربوط به خطاهای رؤیت‌شده به یک فایل اکسل جدید منتقل و فرهنگ اغلاط نام‌گذاری شد. در این فرهنگ با یک روش استقرایی سه نوع خطای مقوله نحوی، معنایی و املایی شناسایی و برای هر نوع خطا یک برچسب عددی (۱، ۲ و ۳) تعیین گردید. در پایان نیز انواع خطاهای مشاهده‌شده به صورت خلاصه، گزارش شد.

روش تحلیل

در حین تولید پیکره و فرهنگ اغلاط از روش‌های زیر استفاده شد. از نرم‌افزار ACS تهیه شده توسط نعمتی (۱۳۹۸) برای استخراج فراوانی معادل‌های انگلیسی کلمات فارسی استفاده شد و سپس محقق به صورت دستی موارد یکسان را در هم تلفیق کرد و فراوانی مربوط به کل رخداد را درج نمود. همچنین برای تلفیق واژه‌های فارسی یکسان از معیار یکسان بودن کامل استفاده شد بدین ترتیب که صورت‌های فارسی کاملاً یکسان و به تبع آن معادل‌های انگلیسی مشابه آنها در هم تلفیق شدند. از تحلیل بسامدی ساده نیز برای مرتب کردن معادل‌های انگلیسی کلمات فارسی به همراه ترتیب الفبایی استفاده شد. در نهایت از آمار توصیفی ساده (فراوانی رخداد) برای توصیف محتوای واژه‌نامه و نیز انواع سه‌گانه خطاهای موجود در فرهنگ اغلاط استفاده به عمل آمد.

تولید واژه‌نامه و فرهنگ اغلاط

برای آشنایی بیشتر با ساختار واژه‌نامه و فرهنگ اغلاط در ادامه ابتدا ستون‌های واژه‌نامه و فرهنگ اغلاط معرفی می‌شود.

معرفی ستون‌های واژه‌نامه

در این قسمت ستون‌های مختلف واژه‌نامه ماشین‌خوان معرفی می‌شود.

A	B	C	D	E	F	G	H	I
مدخل فارسی	معادل انگلیسی و فراوانی رخداد	واژه فارسی	معادل انگلیسی	مقوله نحوی و ر.گ.	فراوانی رخداد معادل انگلیسی	مقوله نحوی	ر.گ.	صورت‌های مساوی
آئوزینوزا (اِ)	aeruginosa 1	آئوزینوزا	aeruginosa	(اِ)	۱	(اِ)		
آئورت (اِ)	aorta 2	آئورت	aorta	(اِ)	۲	(اِ)		
آئورتی (ص.)	aortic 1	آئورتی	aortic	(ص.)	۱	(ص.)		
آئین (اِ) ر.گ. آیین		آئین		(اِ) ر.گ. آیین		(اِ)	آیین	
آئین نامه‌ای (گ.اِ) ر.گ. آیین نامه‌ای		آئین نامه‌ای		(گ.اِ) ر.گ. آیین نامه‌ای		(گ.اِ)	آیین نامه‌ای	
آئین ملی (گ.اِ) ر.گ. آیین ملی		آئین ملی		(گ.اِ) ر.گ. آیین ملی		(گ.اِ)	آیین ملی	

تصویر ۱. ستون‌های واژه‌نامه ماشین‌خوان.

همان‌گونه که در تصویر ۱ مشاهده می‌شود، واژه‌نامه ماشین‌خوان در مجموع دارای ۹ ستون است که با حروف انگلیسی A تا I مشخص شده‌اند.

ستون A مدخل اصلی واژه‌نامه است که در آن واژه فارسی، مقوله نحوی واژه فارسی (و در مدخل‌هایی که واژه فارسی واژه مرجح نباشد، صورت ارجح) ذکر می‌شود. برای نمونه، در مدخل «آئین» لفظ آئین در پیکره موجود بوده است. مقوله نحوی آن

ISSN: 2745-3673
http://www.Rcsj.ir

اسم است و با توجه به اینکه در فرهنگ معین لفظ «آیین» لفظ ارجح است با استفاده از ر.ک. مدخل آئین به مدخل آیین ارجاع داده شده است. در حقیقت، این ستون نشان‌دهنده مدخل‌های اصلی واژه‌نامه است.

ستون B حاوی معادل انگلیسی واژه فارسی موجود در ستون A به همراه فراوانی رخداد این معادل در کل پیکره می‌باشد. برای نمونه معادل انگلیسی لفظ آئورتی، واژه aortic 1 است.

در ستون C واژه فارسی آمده است. برای نمونه واژه آئورت برای مدخل آئورت (ا.) آورده شده است.

در ستون D معادل انگلیسی مدخل فارسی به تنهایی آمده است. برای مثال aorta برای آئورت. لازم به ذکر است که چنانچه واژه مدخل (در ستون A) واژه مرجع نباشد و در آن ر.ک. به کار رفته باشد ستون D و همچنین ستون‌های B و F خالی می‌ماند و در مدخل واژه مرجع، معادل انگلیسی و اطلاعات ستون‌های B و F درج می‌شود.

در ستون E مقوله نحوی واژه فارسی و همچنین در صورت وجود، اطلاعات واژه مرجع با کمک لفظ ر.ک. آورده می‌شود. برای نمونه در ستون E برای لفظ آئورت صرفاً مقوله نحوی (ا.) و برای لفظ آئین مقوله نحوی (ا.) به همراه ر.ک. آیین آورده شده است.

در ستون F فراوانی رخداد تجمعی هر یک از معادل‌های انگلیسی هر واژه مدخل فارسی آورده شده است. علت این امر آن بوده است که در پژوهش قبلی ردیف‌های پیکره بر اساس یکسان بودن چهار متغیر واژه فارسی، واژه انگلیسی، مقوله نحوی فارسی و مقوله نحوی انگلیسی تنظیم شده بود. با توجه به اینکه این واژه‌نامه صرفاً بر اساس یکسان بودن واژه فارسی و معادل انگلیسی آن تنظیم شد، بنابراین خطوط تکراری زیادی حاصل شد که به دلیل یکسان بودن لازم بود فراوانی رخداد آنها با هم جمع و عدد کلی به عنوان تعداد رخداد آن معادل انگلیسی برای یک واژه فارسی خاص ثبت شود. برای مثال، اگر برای آب معادل water در دو ردیف جدا و با فراوانی ۲ و ۳ رؤیت می‌شد در هم ترکیب و به صورت یک واژه فارسی با نام آب و معادل انگلیسی water با فراوانی ۵ ثبت شد. این فرایند یک فرایند زمان‌بر بود که محقق در کل داده‌ها آن را به انجام رساند.

در ستون G صرفاً مقوله نحوی درج شد. مقوله‌های نحوی مورد استفاده در این پژوهش به شرح جدول ۲ بود:

جدول ۲. مقوله‌های نحوی مورد استفاده در پژوهش.

علامت اختصاری	برای مقوله نحوی	علامت اختصاری	برای مقوله نحوی
(ا.)	اسم	(پی.)	پیشوند
(ص.)	صفت	(پس.)	پسوند
(ق.)	قید	(ح.ت.)	حرف تعریف
(ف.)	فعل	(گ.ا.)	گروه اسمی*
(ضم.)	ضمیر		
(ح.اض.)	حرف اضافه		

*. هدف اصلی در این پژوهش درج مقوله‌های نحوی ساده بود و نوع مقوله‌های نحوی نیز به شکل استقرایی و از کتاب گیوی و انوری (۱۳۷۱) استخراج شد بدین صورت که در صورت مشاهده یک مقوله نحوی، آن مقوله به فهرست حاضر اضافه شد. علت وجود مقوله نحوی (گ.ا.) نیز به این دلیل بوده است که گاه برای یک گروه اسمی یک معادل واژگانی انگلیسی واحد موجود بوده و به همین دلیل امکان تفکیک اجزای گروه اسمی وجود نداشته است.

در ستون H که با عنوان ر.ک. مشخص شده، واژه مرجع در صورت موجود بودن ذکر شد.

سرانجام در ستون I صورت‌های مساوی در صورت موجود بودن ذکر گردید. به عبارت دیگر، اگر در پیکره و در دو جایگاه مختلف دو صورت برای یک واژه موجود بود و هر دو درست بودند، هنگام ذکر هر صورت در مدخل A، صورت درست دیگر در مدخل I آورده شد.

ISSN: 2745-3673
http://www.Rcsj.ir

پیش از خاتمه این بخش ذکر چند نکته برای ایجاد شفافیت بیشتر ضروری است:

الف- در ستون A گاه جلوی کلمه از علامت (=واژه مساوی) استفاده می‌شود. این استفاده نشان می‌دهد که برای این کلمه صورت درست و مصطلح دیگری نیز وجود داشته که در پیکره نیامده است. برای مثال بوروکراسی (=بروکراسی) و پسیکوز (=سایکوسیس).

ب- اگر چند صورت درست برای یک مدخل وجود داشته باشد و همگی در پیکره موجود باشند جلوی هر صورت درج‌شده در ستون A صورت‌های درست دیگر در ستون I ذکر می‌شوند.

ج- حال اگر مثلاً دو صورت برای یک واژه وجود داشته باشد اما یکی ارجح و دیگری غیرمرجح باشد در این صورت هر دو واژه در جایگاه الفبایی خود ظاهر می‌شوند با این تفاوت که معادل‌ها و اطلاعات فراوانی در جلوی صورت مرجح ذکر می‌شود و صورت غیرمرجح با لفظ ر.ک. به صورت مرجح ارجاع داده می‌شود. در این حالت فراوانی صورت غیرمرجح نیز به صورت مرجح اضافه می‌شود. این حالت برای معادل‌های فراوان و نادر نیز صدق می‌کند.

د- حال اگر دو صورت برای یک واژه موجود باشد که یکی درست و دیگری از نظر نگارش یا معنی غلط باشد، صورت درست حفظ می‌شود و فراوانی صورت نادرست به فراوانی صورت درست اضافه می‌شود. ضمناً صورت نادرست در فرهنگ اغلاط درج می‌شود.

ه- همچنین محتوای ستون A مجموع اطلاعات خام ستون‌های C و E می‌باشد. به همین ترتیب، محتوای ستون B مجموع محتوای خام ستون‌های D و F می‌باشد.

معرفی ستون‌های فرهنگ اغلاط

در این قسمت ستون‌های مربوط به فرهنگ اغلاط توضیح داده می‌شود.

A	B	C	D	E	F	G	H
فرهنگ اغلاط							
مدخل	معادل انگلیسی و فراوانی	معادل فارسی	معادل انگلیسی	مقوله نحوی	فراوانی	صورت درست	نوع خطا
آبزی (ص.)	acquatic 1	آبزی	acquatic	(ص.)	۱	aquatic	۳
آبی (ص.)	acquatic 1	آبی	acquatic	(ص.)	۱	aquatic	۳
آبی (ص.)	irritated 67	آبی	irritated	(ص.)	۶۷	irrigated	۳
آپگار (ا.)	apgar 1	آپگار	apgar	(ا.)	۱	آپگار	۳
آتروسکلروتیک (ص.)	atherosclorosis 1	آتروسکلروتیک	atherosclorosis	(ص.)	۱	atherosclerosis	۳
آتش‌فشان (ا.)	volcano 1	آتش‌فشان	volcano	(ا.)	۱	آتشفشان	۳
آتش‌فشانی (ص.)	volcanic 2	آتش‌فشانی	volcanic	(ص.)	۲	آتشفشانی	۳
آدرنرژیک (ص.)	adrenergic 1	آدرنرژیک	adrenergic	(ص.)	۱	آدرنرژیک	۳

تصویر ۲. ستون‌های فرهنگ اغلاط.

در فرهنگ اغلاط، ستون‌های A، B، C و D به ترتیب مدخل واژه فارسی، معادل انگلیسی به همراه فراوانی، معادل فارسی و معادل انگلیسی را شامل می‌شود. ستون‌های E و F نیز به ترتیب مقوله نحوی واژه فارسی و فراوانی رخداد معادل انگلیسی یک واژه فارسی را در بر می‌گیرد. در ستون G صورت درست خطای مشاهده‌شده درج می‌شود. ریشه خطا یا در واژه فارسی است و یا در واژه معادل انگلیسی که در هر صورت در مدخل G مشخص می‌شود. سرانجام در ستون H نوع خطا درج می‌گردد که در ادامه به اختصار توضیح داده می‌شود:

خطای نوع ۱ خطای مقوله گفتار است و مواردی را شامل می‌شود که معادل انگلیسی از نظر ریشه درست اما از نظر اشتقاق درست نیست مثل اینکه برای واژه امکان‌پذیر معادل انگلیسی possibility به جای معادل درست possible به کار رفته باشد. یا برای واژه تلاطمی معادل turbulence به جای turbulent استفاده شده باشد.

خطای نوع ۲ خطای معنایی است بدین صورت که گاه مشاهده شد که برای یک واژه مدخل فارسی، معادل انگلیسی به کار رفته از نظر معنایی کاملاً اشتباه است. ریشه این خطا در معادل‌گزینی نویسنده و در مواردی اشکال در تقطیع عناوین در فرایند پژوهش بوده که در هر صورت در فرهنگ اغلاط ذکر و در واژه‌نامه اصلی اصلاح شد. برای مثال، برای کلمه فارسی تأمین به جای

ISSN: 2745-3673
http://www.Rcsj.ir

fulfilling از معادل security استفاده شده است که غلط است. یا برای لفظ پایان‌نامه معادل dissertation آمده که صورت درست آن thesis است.

خطای نوع ۳ شایع‌ترین خطا در داده‌های مورد بررسی بود. این نوع خطا، خطاهای املایی را شامل می‌شد. به عبارت دیگر، گاه املای واژه فارسی و گاه املای معادل انگلیسی واژه‌های فارسی درست نبود. بنابراین، این موارد به عنوان خطای نوع ۳ شناخته شدند و صورت درست آنها در ستون G و همچنین در ستون‌های مربوطه در واژه‌نامه اصلی نوشته شد. برای مثال، معادل انگلیسی واژه آترواسکلروتیک به‌جای atherosclerosis به صورت atherosclerosis نوشته شده بود. یا معادل فارسی واژه apgar به صورت آپکار نوشته شده بود حال آنکه صورت درست آن آپگار بود. به همین ترتیب، معادل انگلیسی واژه آبزی acquatic نوشته شده بود در صورتی که صورت درست آن aquatic بود. همچنین خطاهای جدا و سرهم‌نویسی هم در این دسته قرار گرفت. برای نمونه، واژه آتش‌فشان در پیکره آمده بود در صورتی که در واژه‌نامه‌ها از جمله واژه‌نامه معین، صورت آتشفشان صورت درست محسوب می‌شود و از این رو صورت درست آتشفشان به عنوان واژه درست درج شد.

مراحل تهیه واژه‌نامه و فرهنگ اغلاط

مراحل تهیه واژه‌نامه

برای تهیه واژه‌نامه ماشین‌خوان مراحل مختلفی طی شد که در ادامه به اختصار و به ترتیب اجرا، توضیح داده می‌شود.

گام اول: استفاده از داده‌های پیکره پژوهش فلاحتی قدیمی فومنی (۱۳۹۲) به عنوان مبنای اولیه کار

همان‌گونه که پیشتر نیز ذکر گردید از داده‌های حاصله از پژوهش فلاحتی قدیمی فومنی (۱۳۹۲) به عنوان داده‌های اولیه در پژوهش حاضر استفاده شد.

گام دوم: اعمال شیوه‌نامه تهیه واژه‌نامه

در این مرحله لازم بود تا با استفاده از شیوه‌نامه تهیه‌شده اطلاعات پیکره بازآرایی شود. علت نیاز به این بازآرایی نیز تفاوت ساختار واژه‌نامه حاضر با پیکره قبلی، نیاز به تجمیع ردیف‌ها و تصحیح مدخل‌ها بود. برای تحقق این مهم از شیوه‌نامه تدوین‌شده و به شرح زیر استفاده شد.

-**ترتیب مدخل‌ها:** با توجه به تجمیع ردیف‌های مختلف و رفع خطاهای املایی و ... ترتیب مدخل‌ها به هم ریخته بود و نیاز به بازآرایی از نظر ترتیب الفبایی داشت که این کار وفق شیوه‌نامه و با ترتیب: آ، ا (به ترتیب با فتحه، ضمه و کسره)، همزه با فتحه، ضمه و کسره، ب، پ، ت، ث، ج، چ، ح، خ، د، ذ، ر، ز، س، ش، ص، ض، ط، ظ، ع، غ، ف، ق، ک، گ، ل، م، ن، و، ه، ی انجام شد.

-**معیار درست یا نادرست بودن رسم‌الخط یک واژه:** در این مرحله مطابق شیوه‌نامه، نگارش نادرست کلمات اصلاح شد که این امر خود باعث تغییر در ترتیب الفبایی واژه‌های فارسی شد که محقق مجدد آنها را نیز اصلاح کرد. همان‌گونه که در فصل ۳ ذکر شد معیار درست یا نادرست بودن رسم‌الخط واژه‌های فارسی، شیوه نگارش فرهنگ معین، واژه‌نامه‌های تخصصی (در صورت لزوم) و همچنین دستور خط فرهنگستان زبان و ادب فارسی بود. موارد اصلاحی نیز در فرهنگ اغلاط درج شد.

-**استفاده از علامت (=) در مدخل اصلی:** در این مرحله تمامی واژه‌های فارسی موجود پیکره که دارای صورت دیگر درست اما غیرموجود در پیکره بود، شناسایی شد و جلوی واژه فارسی داخل پرانتز و با علامت مساوی ذخیره شد. برای نمونه، کراکاور (=کراکائر) (ا.)، ایده‌آل (=ایدئال) (ص.)، ... .

-**فاصله و نیم‌فاصله:** در این قسمت طبق شیوه‌نامه فاصله کامل بین ریشه و وند در واژه‌های فارسی شناسایی و به نیم‌فاصله تبدیل شد. برای نمونه بهره مند به بهره‌مند، آموزش پذیر به آموزش‌پذیر، امکان سنجی به امکان‌سنجی و آزاد شده به آزادشده (در این مورد صرفاً با حذف فاصله) تبدیل شد. اگر ریشه کلمه به حرف غیرچسبان ختم می شد صرفاً فاصله بین ریشه و وند حذف شد.

ISSN: 2745-3673
http://www.Rcsj.ir

-**جدا و سرهم‌نویسی:** در این قسمت وفق شیوه‌نامه اشکالات مربوط به جدا و سرهم‌نویسی مرتفع شد و موارد خطا در فرهنگ اغلاط درج شد. برای نمونه کلمات آبپاش و آبپایه به صورت آب‌پاش و آب‌پایه، آب‌زی به آبزی، آب‌گیر به آبگیر، آب‌شستگی به آبشستگی، کانیسازی به کانی‌سازی، کانیزایی به کانی‌زایی، ... تبدیل شد.

-**حروف کوچک و بزرگ:** همان گونه که در شیوه‌نامه در فصل ۳ ذکر شد، به خاطر تنوع شیوه نگارش در نشریات، بزرگ و کوچک نویسی کلمات انگلیسی به طور نامنظم مشاهده شد. این بزرگ و کوچک نویسی متمایز از قواعد زبان انگلیسی مربوط به بزرگ و کوچک‌نویسی (برای مثال بزرگ‌نویسی حرف اول اسامی خاص و ...) بود. این گونه موارد در پیکره شناسایی و با حرف کوچک نوشته شدند برای نمونه isfahan به Isfahan، ISFAHAN به Isfahan، Acrylonitrite به acrylonitrite و Fissure به fissure تبدیل شد.

-**اسامی مفرد و جمع:** در این مورد صورت‌های جمع وفق شیوه‌نامه به صورت مفرد تبدیل شدند مگر در مواردی که تبدیل به صورت مفرد باعث تغییر معنا یا تولید صورت نامتعارف می‌شد. برای نمونه کتاب ها و کتابها هر دو به کتاب، آزمایشات و آزمایش ها هر دو به آزمایش، عناصر به عنصر، شهدا به شهید، علائم به علامت، عقاید به عقیده، عقود به عقد، وعاظ به واعظ، قرون به قرن تبدیل شد اما حبوبات و خاندان چون به نوع اشاره داشتند و ابزار، عوارض (toll) اوراق بهادار، پرسنل، خوارج، ضایعات، ... به دلیل اینکه تبدیل آنها به صورت مفرد باعث تغییر معنی و نامعمول شدن واژه می‌شد، به همان صورت حفظ شدند.

-**صورت‌های فینگلیش:** وفق شیوه‌نامه تمام واژه‌های فینگلیش و انگلیسی‌نویسی کلمات فارسی مصطلح (موجود در واژه‌نامه‌های تخصصی) به همان صورت حفظ شد. برای نمونه، معادل Adam برای آدام، Agha در کنار Mr برای آقا، دلیشز برای واژه انگلیسی delicious و ... .

-**توضیح نام و نام خانوادگی افراد:** در این قسمت نام و نام خانوادگی موجود در عنوان (غالباً وقتی نام روش یا مدل باشند) حفظ شدند. مانند اسم Alvin که نام مدل است یا آقایی که نام نوعی برنج است.

-**القاب:** مطابق دستورالعمل، القابی نظیر دکتر، شیخ و ... موجود در عناوین فارسی و انگلیسی به همان صورت حفظ شدند. مانندحفظ صورت‌های Agha و Mr برای لفظ آقا.

- **بازگرداندن خطاهای سیستمی حاصل از پیش‌پردازش:** در این قسمت خطاهای سیستمی و حاصل از پیش‌پردازش با هدف انطباق صورت واژگان با فرهنگ‌های لغت و مصوبات فرهنگستان مجدد به حالت اول بازگردانده شد. برای نمونه بوپیواکایین به بوپیواکائین، بویین‌زهرا به بوئین‌زهرا، پالئوزوییک به پالئوزوئیک، یاس به یأس، آدنویید به آدنوئید، آمیلوویید به آمیلوئید، وب کویست به وب کوئست (webquest)) تبدیل شد. همچنین عدم درج علائم فتحه، ضمه و کسره در کلمات با پیچیدگی تلفظ اصلاح شد. برای نمونه فتحه در باشکل به صورت باشْکَل اصلاح گردید، یا هرثمه (Harsameh) به هَرثَمه یا هجویری (Hojviri) به هُجویری تبدیل شد. یا در مورد تنوین کاملا به کاملاً، احتمالا به احتمالاً تغییر کرد.

-**گونه انگلیسی (انگلیسی آمریکایی/بریتانیایی):** وفق شیوه‌نامه صورت‌های مربوط به انگلیسی آمریکایی و بریتانیایی در صورت وجود در پیکره به همان صورت حفظ شدند مانند behavior و behavior یا labor و labour.

-**مقوله‌های مختلف نحوی برای یک واژه:** طبق شیوه‌نامه شیوه درج صورت‌های فارسی متعلق به مقوله‌های نحوی مختلف به صورت زیر بود: اسم، صفت، فعل، قید، حرف اضافه، حرف ربط، حرف تعریف، ضمیر و گروه حرف اضافه‌ای.

-**حذف تکرار:** چون معیار چیدمان ردیف‌ها در پیکره قبلی بر اساس یکسان بودن صورت فارسی، معادل انگلیسی و مقوله نحوی فارسی و انگلیسی تعیین شده بود و در مقابل در پژوهش حاضر مقوله نحوی کلمه انگلیسی ملاک نبود، ناگزیر ردیف‌های تکراری در پیکره ایجاد شد که این موارد طبق شیوه‌نامه تجمیع شد. برای نمونه واژه فرزندپروری با معادل parenting در دو ردیف تشکیل شد که با هم تجمیع شدند. یا لفظ سازمان‌یافته با معادل organized، وسیله نقلیه با معادل vehicle و واژه‌های دیگر نظیر آن نیز به همین ترتیب عمل شد. همچنین در صورت وجود معادل‌های انگلیسی متفاوت طبیعتاً واژه مدخل فارسی در ردیف‌های مختلف تکرار می‌شد که در این مرحله واژه فارسی در جلوی معادل انگلیسی اول حفظ و باقی از ستون A پاک شد. مانندحذف تکرار واژه فارسی آب در نمونه زیر:

ISSN: 2745-3673
http://www.Rcsj.ir

water 42	آب (ا.)
juice 6	
moisture 6	
drinking water 5	
hydro 3	
fluid 1	

تصویر ۳. حذف تکرار واژه فارسی در ستون A.

-**حذف یا تغییر برخی اسامی مغایر با ضوابط دولتی یا اداری:** مطابق شیوه‌نامه در این قسمت کلماتی چون اسراییل حذف شد و لفظی نظیر افغانی به افغان تبدیل شد.

-**کلمات هم آوا-هم نویسه:** بر اساس شیوه‌نامه کلمات هم آوا-هم نویسه نیز شناسایی و ضمن تفکیک با درج عدد از هم متمایز شدند. برای نمونه باز۱، باز۲ و باز۳ به ترتیب برای اشاره به re، open و eagle. یا بازی۱ و بازی۲ برای اشاره به game و basicity.

-**کلمات هم‌نویسه:** وفق شیوه‌نامه کلمات دارای املای یکسان اما تلفظ و معنی متفاوت شناسایی و با درج اِعراب، شیوه خوانش آنها شفاف شد. برای مثال بَردگی و بُردگی، کِشتی و کُشتی یا سَبْک و سَبُک.

-**علامت تشدید:** بر اساس شیوه‌نامه تشدید از واژه‌ها حذف شد مگر در مواردی که حذف آن باعث دشواری در تلفظ کلمه می‌شد. برای نمونه، تشدید از لفظ علّامه حذف شد اما در مورد عیار (grade) و عیار (Ayyar) صرفاً برای صورت دوم تشدید لحاظ شد.

-**چیدمان معادل‌ها در هر مدخل:** وفق شیوه‌نامه در هر مدخل معادل‌ها به ترتیب فراوانی در ردیف‌ها درج شدند و در صورت برابر بودن فراوانی، ترتیب الفبایی ملاک تعیین ترتیب معادل‌ها قرار گرفت. برای نمونه، چهار معادل انگلیسی کلمه فارسی مجزا (ص)، به ترتیب isolated 15، decoupled 1، distinct 1 و separate 1 ثبت شد.

-**کلمات عربی موجود در پیکره:** مطابق شیوه‌نامه کلمات عربی موجود در عناوین نظیر المناط، المنتظر، المؤمن و نظیر آن حفظ شد.

-**عدم حذف معادل‌های با فراوانی پائین:** وفق شیوه‌نامه تمام معادل‌های انگلیسی موجود در پیکره و نیز واژه‌های فارسی دارای فراوانی کم به دلایلی که در فصل ۳ عنوان شد، حفظ شد.

-**نحوه ثبت مقوله نحوی:** مقوله نحوی کلمات فارسی بر اساس مقوله واژه‌نامه‌ای کلمات ثبت شد و ملاحظات بافتی مورد توجه قرار نگرفت. ضمناً، به مقوله نحوی کلمات انگلیسی توجه نشد.

-**علائم اختصاری:** علائم اختصاری موجود به همان صورت حفظ شد. برای نمونه، اختلال بیش‌فعالی کم‌توجهی، اختلال نقص توجه بیش‌فعالی و اختلال وسواس به ترتیب با معادل‌های ADHD، ADHD و OCD و موجود در پیکره ثبت شدند. ضمناً در صورتی که در پیکره علاوه بر سرنام صورت کامل هم در همان مدخل برای چنین واژه‌هایی موجود بود هر دو نوع معادل به ترتیب فراوانی و سپس بر اساس ترتیب الفبایی ردیف شدند.

-**علامت تنوین:** وفق شیوه‌نامه هر جا که وجود علامت تنوین ضروری بود، محقق آن را به واژه فارسی اضافه کرد. بر این اساس کلمه کاملا به کاملاً و احتمالا به احتمالاً تغییر کرد.

-**توجیه وجود گروه اسمی:** هر جا در عناوین مقالات برای چند واژه فارسی (یک گروه واژه) صرفاً یک معادل انگلیسی واحد وجود داشت، برای تعیین مقوله نحوی از لفظ (گ.ا.) برای نشان دادن گروه اسمی استفاده شد. برای نمونه برای لفظ قشر آدرنال و قشر مغز به ترتیب کلمات adrenocotice و neocortex به کار رفتند و دو جزء کلمات فارسی از هم جدا نشدند. یا فیبروآنوم پستان به عنوان یک واحد شناخته شد چون برای کل این عبارت معادل fibroadenoma وجود داشت.

ISSN: 2745-3673
http://www.Rcsj.ir

-**حذف علامت‌های نامانوس:** برخی علائم نامأنوس مانند ā برای نشان دادن حرف آ کشیده از جریان تحقیق کنار گذاشته شدند و به جای آنها علائم مأنوس بکار رفت. برای مثال، ā به a ساده تغییر کرد و بنابراین معادل انگلیسی علّامه از Allāmeh به Allameh تغییر کرد، یا رامایانا از Rāmāyānā به Ramayana تبدیل شد.

-**اشکال ناشی از تقطیع نادرست کلمات:** هر جا به خاطر اشکال در فرایند تقطیع عناوین فارسی و انگلیسی، بین واژه فارسی و معادل انگلیسی تناسب معنایی وجود نداشت، واژه فارسی اصل فرض شد و معادل درست آن واژه از عناوین خام استخراج و برای آن واژه ثبت گردید. مثلاً معادل واژه فارسی هجا از satirical poetry به syllable تغییر کرد یا معادل فسفوکیناز از creatine به phosphokinase تبدیل شد.

-**کلمات نوشته‌شده با و بدون همزه:** طبق شیوه‌نامه چنانچه در یک واژه فارسی همزه آخر حذف شده بود همزه اضافه شد و صورت بدون همزه به صورت دارای همزه ارجاع شد، مانند اسما و اسماء، اشیا و اشیاء، اعضا و اعضاء، افشا و افشاء، انحنا و انحناء، اولیا و اولیاء، اهدا و اهداء، ابتلا و ابتلاء، ابتدا و ابتداء.

گام سوم: تولید واژه‌نامه

با کمک شیوه‌نامه تهیه‌شده و پس از اعمال این شیوه‌نامه بر روی پیکره، واژه‌نامه پژوهش حاضر با طی مراحل زیر تولید شد:

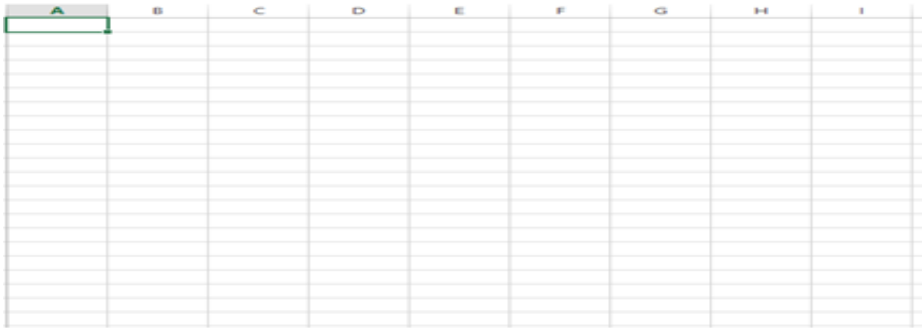
- **استفاده از فایل اکسل پیکره فلاحتی قدیمی فومنی (۱۳۹۲) به عنوان داده خام:** تصویر ۴ فایل نهایی پیکره مزبور را نشان می‌دهد. همان‌گونه که مشاهده می‌شود در این فایل معیار چیدمان ردیف‌ها بر اساس یکسان بودن چهار متغیر واژه فارسی، واژه انگلیسی، مقوله نحوی فارسی و سرانجام مقوله نحوی انگلیسی بوده است که پس از این چهار ستون تعداد رخداد توأمان این چهار متغیر نیز درج شده است.

واژه فارسی	واژه انگلیسی	مقوله نحوی فارسی	مقوله نحوی انگلیسی	تعداد رخداد
اتاقک	chamber	N	N	۱
اتانازی	euthanasia	N	N	۵
اتانل	ethanol	N	N	۱
اتانول	ethanol	N	N	۱
اتانولی	ethanolic	ADJ	ADJ	۱
اتحاد	union	N	N	۲
اتحاد	unity	N	N	۱
اتحادیه	union	N	N	۱
اتحادیه اروپا	EU	N	N	۲
اتر	ether	N	N	۳
اترک	Atrak	N	N	۱
اترها	ethers	N	N	۲
اتساع مری	esopfiageafdistention	NP	N	۱
اتستیک	autistic	ADJ	ADJ	۳
اتصاف	endowed	N	ADJ	۱
اتصال	bond	N	N	۱
اتصال	bonding	N	N	۷
اتصال	connection	N	N	۱
اتصال	coupling	N	N	۱۰
اتصال	joint	N	N	۲
اتصال	mount	N	N	۴
اتصال ها	connections	N	N	۱

تصویر ۴. نمونه ای از ردیف‌های پیکره را در پژوهش قبل نشان می‌دهد.

ISSN: 2745-3673
http://www.Rcsj.ir

- **باز کردن یک فایل اکسل جدید:** سپس یک فایل اکسل جدید باز شد و ۹ ستون آن با توجه به نیاز پژوهش حاضر انتخاب گردید (تصویر ۵).



تصویر ۵. بازکردن یک فایل اکسل جدید حاوی ۹ ستون.

در این مرحله ستون‌های A تا I به ترتیب با الفاظ مدخل فارسی؛ معادل انگلیسی و فراوانی رخداد؛ واژه فارسی؛ معادل انگلیسی؛ مقوله نحوی و ر.ک.؛ فراوانی رخداد هر معادل انگلیسی؛ مقوله نحوی؛ ر.ک. و صورت‌های مساوی نام‌گذاری شدند (تصویر ۶):

مدخل فارسی	معادل انگلیسی و فراوانی	واژه فارسی	معادل انگلیسی	مقوله نحوی و ر.ک.	رخداد هر معادل انگ	مقوله نحوی	ر.ک.	صورت‌های مساوی
	aeruginosa 1		aeruginosa					
	aorta 2		aorta					
	aortic 1		aortic					

تصویر ۶. نام‌گذاری ستون‌های واژه‌نامه.

- **استخراج اطلاعات از پیکره فلاحتی قدیمی فومنی (۱۳۹۲) و پرکردن ستون‌های واژه‌نامه:** برای استخراج اطلاعات از پیکره فلاحتی قدیمی فومنی (۱۳۹۲) به شرح زیر عمل شد:

ابتدا ستون واژه‌های فارسی موجود در پیکره (ستون سمت راست تصویر ۴) در ستون C فایل اکسل جدید کپی شد. سپس وفق دستورالعمل این واژه‌ها از نظر ترتیب و همچنین املا یا تکرار ویرایش شدند. برای نمونه، چنانچه یک مدخل در جایگاه الفبایی درست بر اساس دستورالعمل قرار نداشت به جایگاه درست منتقل شد. یا همان‌گونه که در دستورالعمل ذکر شد چنانچه کلماتی از نظر قرائت ثقیل بودند با علامت مصوت کوتاه کامل شدند؛ صورت‌های جمع به صورت مفرد تبدیل شد و در صورت وجود صورت مفرد، فراوانی صورت‌های جمع به فراوانی صورت‌های مفرد اضافه گردید؛ صورت‌های تکراری در هم تجمیع شد؛ چنانچه کلمه‌ای دارای غلط املایی یا معنایی بود طبق دستورالعمل به صورت درست تبدیل شد و خطای مشاهده‌شده به فرهنگ اغلاط منتقل گردید. پس از اعمال کامل دستورالعمل بر روی واژه‌های فارسی، بین هر دو واژه متفاوت (دو مدخل) یک ردیف خالی ایجاد شد تا تفکیک مدخل‌ها برای کاربر راحت‌تر باشد.

سپس بر روی ستون دوم در تصویر ۴ (واژه انگلیسی) کار شد. واژه‌های انگلیسی موجود در این ستون معادل‌های واژه‌های فارسی بودند. اطلاعات این ستون در فایل اکسل جدید در ستون D قرار گرفت. با توجه به اینکه در بسیاری از موارد برای یک واژه فارسی بیش از یک معادل انگلیسی وجود داشت برای تعیین ترتیب معادل‌های انگلیسی ذیل هر مدخل ابتدا از معیار فراوانی رخداد و در صورت برابر بودن فراوانی رخداد، بر اساس ترتیب حروف الفبا اقدام شد. برای نمونه، به مدخل مربوط به واژه بخش در تصویر ۷ مستخرج از پیکره اولیه توجه کنید.

ISSN: 2745-3673
http://www.Rcsj.ir

بخش	unit	N	N	۱
بخش	department	N	N	۱
بخش	district	N	N	۱
بخش	making	ADJ	ADJ	۱
بخش	part	N	N	۳
بخش	sect	N	N	۱
بخش	section	N	N	۱
بخش	sector	N	N	۱
بخش	segment	N	N	۱
بخش	unit	N	N	۲
بخش	ward	N	N	۱
بخش بندی	segmentation	N	N	۳
بخش بندی	segmenting	N	N	۱
بخش قشری	cortex	N	NP	۲
بخش مراقبت ویژه	ICU	N	NP	۴
بخش میانی	section	N	NP	۱
بخش ها	fractions	N	N	۱
بخش ها	parts	N	N	۱
بخش ها	sectors	N	N	۴
بخش ها	wards	N	N	۱
بخش های	settings	N	N	۱
بخش	department	N	N	۱
بخش مراقبت های ویژه	ICU	NP	N	۱
بخش های	units	N	N	۲
بخشایش	forgiveness	N	N	۳
بخشها	sectors	N	N	۲

تصویر ۷. داده اولیه مربوط به واژه بخش.

در این داده، کلمه بخش به چند صورت مختلف ظاهرشده که عبارتند از بخش، بخش ها، بخشها و بخش های. برای تجمیع این موارد به این صورت عمل شد که ابتدا صورت مفرد اصل فرض شد و صورت‌های جمع بخش ها و بخشها با حذف علامت جمع و صورت بخش‌های با حذف ی پایانی و سپس حذف علامت جمع، به صورت مفرد خود یعنی بخش تبدیل شدند. سپس، چنانچه در ردیف‌های مختلف آن واژه فارسی یک معادل انگلیسی واحد وجود داشت فراوانی این معادل‌ها با هم جمع و عدد بدست آمده برای آن معادل انگلیسی ثبت و تکرار معادل انگلیسی حذف شد. البته، با توجه به اینکه در نقاط دیگر پیکره هم گاه لفظ بخش ممکن بود ظاهر شده باشد تمام آن موارد نیز به جایگاه اول (مدخل ب و جایگاه مدخل بخش) منتقل شد. بدین ترتیب، ۱۹ ردیف مربوط به واژه بخش و صورت‌های جمع آن بر اساس فراوانی صورت انگلیسی و نیز ترتیب الفبایی به ۱۱ ردیف تقلیل یافت. (تصویر ۸). در تصویر ۸ معادل‌های sector تا ward بر اساس فراوانی و معادل‌های district تا setting بر اساس ترتیب حروف الفبا ردیف شده‌اند. فراوانی معادل‌های انگلیسی نیز در ستون F درج شد. همین فرایند در کل طرح دنبال گردید.

ISSN: 2745-3673
http://www.Rcsj.ir

F	C	D
۷	بخش	sector
۵	بخش	unit
۴	بخش	part
۳	بخش	department
۲	بخش	ward
۱	بخش	district
۱	بخش	fraction
۱	بخش	sect
۱	بخش	section
۱	بخش	segment
۱	بخش	setting

تصویر ۸. مدخل واژه‌نامه برای واژه بخش پس از اعمال تغییرات در پیکره اولیه.

در مرحله بعد مقوله نحوی فارسی (تصویر ۴) از پیکره قبلی به ستون G در فایل اکسل جدید منتقل شد. البته این اطلاعات در پیکره قبلی به انگلیسی بود و چون در طرح جدید هدف، تهیه واژه‌نامه ماشین‌خوان فارسی بود مقوله‌های نحوی از انگلیسی به فارسی تبدیل شدند. برای نمونه ADJ به (ص.) به معنی صفت تبدیل شد یا اسم از N به (ا.) تغییر یافت. لازم به ذکر است این مقوله، مقوله نحوی کلمه فارسی را نشان می‌دهد و ذکر یا درج مقوله نحوی معادل‌های انگلیسی در این پژوهش مد نظر نبوده است. بنابراین در مواردی که برای یک واژه فارسی معادل‌های متفاوت انگلیسی وجود دارد، مقوله نحوی در ستون G صرفاً به مقوله نحوی واژه فارسی سازنده مدخل اشاره دارد و نه معادل انگلیسی مندرج در یک ردیف (تصویر ۹). نکته دیگر اینکه مقوله‌های نحوی بر اساس مقوله واژه‌نامه‌ای[۲۳] و نه بافتی[۲۴] کلمات فارسی تعیین شدند.

C	D	F	G
تکمیل	completion	۱	(ا.)
تکمیل	supplementation	۱	(ا.)
تکمیلی	post	۸	(ص.)
تکمیلی	supplemental	۷	(ص.)
تکمیلی	complementary	۱	(ص.)
تکمیلی	supplementary	۱	(ص.)
تکنار	Taknar	۱	(ا.)
تکنولوژی	technology	۱	(ا.)
تکنولوژیک	technological	۱	(ص.)

تصویر ۹. تولید ستون G در واژه‌نامه ماشین‌خوان.

سپس ستون H برای درج واژه مرجع (در صورت موجود بودن) تشکیل شد. برای نمونه، صورت مرجع واژه افشا لفظ افشاء بوده است بنابراین وقتی لفظ غیرمرجع افشا در ستون C ظاهر می‌شود صورت مرجع آن در ستون H به صورت افشاء درج می‌گردد. لازم به ذکر است که صورت درج شده در ستون H خود نیز به عنوان واژه مدخل در ردیف الفبایی مربوطه در واژه‌نامه ظاهر می‌شود. یا به نمونه‌های دیگر در تصویر ۱۰ توجه کنید:

[23] . Denotative
[24] . Connotative

ISSN: 2745-3673
http://www.Rcsj.ir

C	D	F	G	H
آئین			(ا.)	آیین
آئین دادرسی			(گ.ا.)	آیین دادرسی
آئین مانی			(گ.ا.)	آیین مانی
آئین‌نامه			(ا.)	آیین‌نامه
آئینه			(ا.)	آیینه
آئین هندو			(گ.ا.)	آیین هندو

تصویر ۱۰. تکمیل ستون H در واژه‌نامه ماشین‌خوان.

در مرحله بعد لازم بود با استفاده از این ستون‌ها ظاهر واژه‌نامه‌ای نیز به داده‌ها داده شود. برای این کار فرایند زیر طی شد:

ابتدا ستون E تدوین شد بدین صورت که محتوای داده‌های مربوط به ستون‌های G (مقوله نحوی) و H (واژه مرجع) درهم‌کرد شد و لفظ ر.ک. بین محتوای دو ستون قرار گرفت. برای نمونه، برای واژه آرا، در ستون G (ا.) و در ستون H (آراء) وجود داشت که این دو در ستون E به صورت (ا.) ر.ک. آراء تجمیع شد (تصویر ۱۱).

C	D	E	F	G	H
آرا		(ا.) ر.ک. آراء		(ا.)	آراء
آراء	viewpoints	(ا.)	۱۰۶	(ا.)	
آرا	ideas	(ا.)	۴	(ا.)	
آرا	perspectives	(ا.)	۳	(ا.)	
آرا	opinions	(ا.)	۲	(ا.)	
آرا	thoughts	(ا.)	۲	(ا.)	
آرا	views	(ا.)	۲	(ا.)	
آرابیدوپسیس	arabidopsis	(ا.)	۱	(ا.)	
آراسته	bedecked	(ا.)	۱	(ا.)	

تصویر ۱۱. تکمیل ستون E در واژه‌نامه ماشین‌خوان.

لازم به ذکر است که محتوای ستون H و همچنین ترکیب بعد از مقوله نحوی در ستون E در صورتی پر می‌شود که واژه فارسی موجود در ستون A واژه مرجع نبوده باشد و بنابراین بخواهیم آن را به واژه مرجع موجود در پیکره ارجاع دهیم. برای نمونه این دو ستون در مدخل آراسته در تصویر ۱۱ خالی است اما برای مدخل آرا پر شده است.

در مرحله بعد با توجه به اینکه واژه فارسی مدخل ممکن بود در داده‌های پژوهش دارای صورت معادل و مساوی دیگری هم باشد، ستون I ایجاد شد و واژه معادل به شرطی که در پیکره وجود داشته باشد در آن درج شد. فرق این ستون با ستون H در این است که ستون H برای مدخلی پر می‌شود که واژه فارسی موجود در مدخل، غیرمرجع بوده باشد و ناگزیر باید به واژه مرجع موجود در پیکره ارجاع داده می‌شد که در این حالت ستون H پر می‌شود. اما ستون I مربوط به مواردی است که یک واژه فارسی خود صحیح است و ضمناً دارای صورت یا صورت‌های صحیح دیگری نیز در همان پیکره می‌باشد. در چنین حالتی هر صورت از آن واژه در جایگاه الفبایی خود آورده می‌شود و در ستون I صورت یا صورت‌های معادل و درست دیگر درج می‌شوند. مانند آربسکولار و آربوسکولار در تصویر ۱۲:

ISSN: 2745-3673
http://www.Rcsj.ir

C	D	E	F	G	H	I
آربسکولار	arbuscular	(ص.)	۲	(ص.)		آربوسکولار
آربوسکولار	arbuscular	(ص.)	۶	(ص.)		آربسکولار
آرتروپلاستی	arthroplasty	(ا.)	۱	(ا.)		

تصویر ۱۲. تکمیل ستون I در واژه‌نامه ماشین‌خوان.

برای تکمیل ستون‌های فایل اکسل، دو کار دیگر باقی ماند که عبارت بودند از پر کردن ستون‌های A و B. در حقیقت دو ستون A و B به همراه ستون I بخش نهائی واژه‌نامه را تشکیل می‌دهند. ستون A حاوی واژه مدخل است و ستون B هم حاوی معادل(های) انگلیسی به ترتیب و با ذکر فراوانی رخداد هر معادل است. برای ساختن ستون B از ترکیب ستون‌های D و F با استفاده از فانکشن "=dx&Fx&" استفاده شد. در این فرمول x شماره ردیف را در فایل اکسل نشان می‌دهد بنابراین d2 به این معنی است که واژه مندرج در ردیف یا خط دوم از ستون D آمده است. به همین ترتیب C22000 به واژه مندرج در خط یا ردیف ۲۲۰۰۰ از ستون C اشاره می‌کند. برای نمونه این فانکشن در ردیف ۱، معادل انگلیسی aeruginosa در ستون D را با فراوانی رخداد ۱ مندرج در ستون F با هم ترکیب کرد و صورت aeruginosa 1 تولید شد. به مثال‌های دیگری در تصویر ۱۳ توجه کنید:

B	C	D	E	F
Abadan 11	ابادان	Abadan	(ا.)	۱۱
planning 1	ابادانی	planning	(ا.)	۱
Abadeh 1	ابادہ	Abadeh	(ا.)	۱
cistern 1	آب‌انبار	cistern	(ا.)	۱
rainfall 3	آب باران	rainfall	(گ.ا.)	۳
cut off 2	آب‌بند	cut off	(ا.)	۲
grout 1	آب‌بند	grout	(ا.)	۱
sprinkler 1	آب‌پاش	sprinkler	(ا.)	۱

تصویر ۱۳. شیوه تکمیل ستون A واژه‌نامه ماشین‌خوان.

سرانجام در گام نهایی، واژه اصلی مدخل (ستون A) تولید شد. برای انجام این کار از داده‌های دو ستون C و E استفاده شد. فرمول "=Cx&Ex" ترکیبات موجود در دو ستون را در هم تجمیع کرد و بدین ترتیب مدخل اصلی هر واژه تولید شد. برای نمونه، واژه آئروژینوزا در ستون C با (ا.) در ستون E تجمیع شد (تصویر ۱۴).

A	B	C	D	E
آئین بهوه (گ.ا.) رک. آیین بهوه		آئین بهوه		(گ.ا.) رک. آیین بهوه
آب (ا.)	water 42	آب	water	(ا.)
	juice 6	آب	juice	(ا.)
	moisture 6	آب	moisture	(ا.)
	drinking water 5	آب	drinking water	(ا.)
	hydro 3	آب	hydro	(ا.)
	fluid 1	آب	fluid	(ا.)
ابادان (ا.)	Abadan 11	ابادان	Abadan	(ا.)
ابادانی (ا.)	planning 1	ابادانی	planning	(ا.)
ابادہ (ا.)	Abadeh 1	ابادہ	Abadeh	(ا.)

ISSN: 2745-3673
http://www.Rcsj.ir

تصویر ۱۴. تکمیل ستون A در واژه‌نامه ماشین‌خوان.

توصیف مختصر واژه‌نامه

برای اینکه تمایز بین واژه‌ها و مدخل‌ها در این واژه‌نامه آسان‌تر باشد محقق بین هر دو مدخل یک خط خالی قرار داد مانند تصویر ۱۵.

A	B	C	D	E	F	G	H	I
مدخل فارسی	معادل انگلیسی و فراوانی	واژه فارسی	معادل انگلیسی	مقوله نحوی و ر.ک.	خدادهر معادل	مقوله نحوی	ر.ک.	صورت‌های مساوی
آئوژنوزا (ا.)	aeruginosa 1	آئوژنوزا	aeruginosa	(ا.)	۲	(ا.)		
آئورت (ا.)	aorta 2	آئورت	aorta	(ا.)	۲	(ا.)		
آئورتی (ص.)	aortic 1	آئورتی	aortic	(ص.)	۲	(ص.)		
آئین (ا.) ر.ک. آیین		آئین		(ا.) ر.ک. آیین		(ا.)	آیین	
آئین نادرستی (اگ.ا.) ر.ک. آیین نادرستی		آئین نادرستی		(اگ.ا.) ر.ک. آیین نادرستی		(اگ.ا.)	آیین نادرستی	
آئین مانی (اگ.ا.) ر.ک. آیین مانی		آئین مانی		(اگ.ا.) ر.ک. آیین مانی		(اگ.ا.)	آیین مانی	

تصویر ۱۵. ایجاد یک خط خالی بین هر دو مدخل برای خوانش آسان‌تر.

برای توصیف مختصر مختصات واژگانی این واژه‌نامه محقق داده‌ها را فیلترگذاری کرد که نتایج زیر بدست آمد:

- در این واژه‌نامه در مجموع ۱۱۹۴۹ مدخل فارسی بدست آمد.
- تعداد کل معادل‌های انگلیسی ثبت‌شده برای ۱۱۹۴۹ واژه مدخل فارسی، ۱۷۹۹۱ مورد بوده است. به عبارت دیگر به طور متوسط برای هر دو واژه فارسی سه معادل انگلیسی ثبت شده است.
- در این واژه‌نامه در مجموع ۴۲ بار از علامت ر.ک. استفاده شده است (مربوط به ستون H).
- تعداد صورت‌های مساوی مشاهده‌شده در این واژه‌نامه ۱۰۹ مورد بوده است (مربوط به ستون I)
- از نظر مقوله‌های نحوی در مجموع و پس از انجام تمام تغییرات و اصلاحات مشخص شد که بیشترین تعداد کلمات عناوین مقالات، به مقوله نحوی اسم مربوط بودند. در مجموع ۸۰۷۲ مدخل فارسی به مقوله نحوی اسم (ا.) مربوط تعلق داشت که معادل ۶۷/۵ درصد مدخل‌ها بود.
- مقوله نحوی صفت (ص.) با ۳۰۰۹ مورد و ۲۵/۱۸ درصد دومین مقوله نحوی رایج در کلمات عناوین مقالات بود.
- گروه اسمی (گ.ا.) با ۵۳۸ مورد و ۴/۵ درصد در جایگاه سوم قرار گرفت.
- سایر مقوله‌های نحوی در مجموع کمتر از ۳ درصد از کل مدخل‌ها را تشکیل دادند. این مقوله‌ها به ترتیب به شرح زیر بود: حرف اضافه (۱۵۴ مورد)، پیشوند (۱۰۰ مورد)، پسوند (۱۶ مورد)، قید (۱۹ مورد) و ضمیر (۴ مورد).

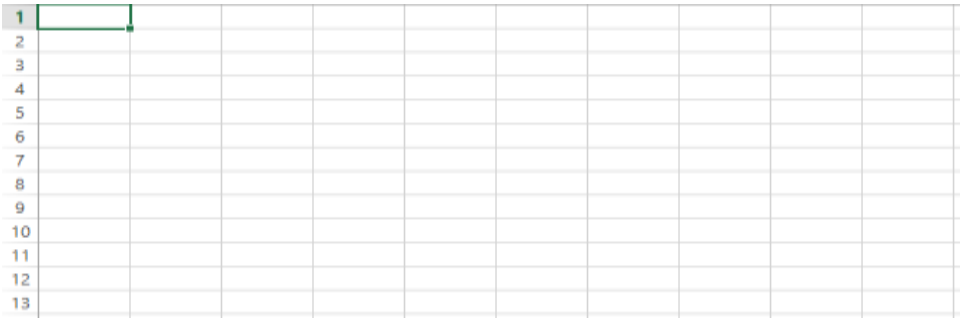
مراحل تهیه فرهنگ اغلاط

برای تهیه فرهنگ اغلاط مراحل زیر طی شد:

گام اول: بازکردن یک فایل اکسل

برای تهیه فرهنگ اغلاط به عنوان محصول دوم پژوهش حاضر، ابتدا یک فایل اکسل باز شد (تصویر ۱۶).

ISSN: 2745-3673
http://www.Rcsj.ir



تصویر ۱۶. باز کردن فایل اکسل.

گام دوم: شناسایی خطاها و انتقال ردیف‌های دارای خطا از واژه‌نامه به فایل اکسل جدید

محقق در حین تولید واژه‌نامه از پیکره متوجه شد که مجموعه‌ای از خطاها در پیکره و به تبع آن در واژه‌نامه وجود دارد. بنابراین در گام دوم تصمیم گرفت به محض مواجهه با هر خطا، ردیف مربوطه را در این فایل اکسل جدید بارگذاری کند.

گام سوم: برچسب‌گذاری خطاها

برای برچسب‌گذاری خطاها نیز از روش استقرایی استفاده شد بدین ترتیب که ابتدا هر نوع خطایی که در واژه‌نامه مشاهده شد، استخراج گردید و در مرحله بعد سعی شد این خطاها به انواع تقسیم شود.

پس از درج ردیف‌های دارای خطا و بررسی آنها، خطاهای زیر در داده‌های پژوهش مشاهده شد: بزرگ و کوچک نویسی اسامی خاص و ...؛ اشکالات مربوط به درج یا عدم درج علائم زبرزنجیری نظیر مصوت‌های کوتاه، تشدید، تنوین و ...؛ خطای مربوط به مقوله نحوی، خطای املایی و خطای معنایی. از این میان با توجه به اینکه امکان تعیین ریشه خطا برای دو مورد اول وجود نداشت این موارد از فهرست خطاها خارج شدند. ضمناً با توجه به اینکه در واژه‌نامه مدخل‌ها عموماً مفرد در نظر گرفته می‌شوند رؤیت صورت‌های جمع هم خطا محسوب نشد و محقق به صورت دستی آنها را به صورت مفرد تبدیل کرد. بنابراین سه نوع خطا باقی ماند که عبارت بودند از: خطای مقوله نحوی، خطای معنایی و خطای املایی که به ترتیب با اعداد ۱، ۲ و ۳ برچسب خوردند.

خطای مقوله نحوی به مواردی مربوط می‌شد که برای یک کلمه فارسی مشخص معادل انگلیسی از نظر ریشه درست اما از نظر مقوله نامناسب بود. به عبارت دیگر صورت متناسب با واژه فارسی از زبان انگلیسی اخذ نشده بود. این موارد با عدد ۱ برچسب خوردند و صورت درست آنها نیز در فرهنگ اغلاط ذکر شد. البته همان‌گونه که پیشتر در این فصل ذکر شد صورت درست کلمات در واژه‌نامه ثبت گردید تا واژه‌نامه عاری از هر نوع خطایی باشد. خطای معنایی نیز که با عدد ۲ برچسب خورد مواردی را توصیف می‌کرد که در آن واژه انگلیسی از نظر معنایی ارتباطی با واژه فارسی نداشت. سرانجام خطای نوع سوم بررسی شد که خطای املایی بود. این نوع خطا شایع‌ترین خطای موجود در داده‌ها محسوب می‌شد و هر نوع خطای املایی مانند حذف، اضافه یا جابجایی حروف مشکل فاصله و جدانویسی را شامل می‌شد. عمده خطاهای املایی، خطاهای نگارشی سهوی بودند و موارد متعدد نیز خطاهای تایپی بود که به دلیل بی‌دقتی حروف نزدیک به هم بر روی صفحه کلید به جای هم استفاده شده است. در پایان لازم به ذکر است که درج یا حذف بی مورد علائم زبرزنجیری، تشدید، تنوین و ... به دلیل عدم امکان رصد ریشه خطا مد نظر قرار نگرفت.

-تحلیل آماری مختصر فرهنگ اغلاط

در مجموع پس از تکمیل گام سوم، فرهنگ اغلاط پژوهش حاضر با ۲۹۶ ردیف و ۲۶۸ واژه متمایز مدخل فارسی تولید شد. تصویرهای ۱۷ تا ۲۰ نمونه هایی از فرهنگ اغلاط ماشین‌خوان را نشان می‌دهد:

ISSN: 2745-3673
http://www.Rcsj.ir

نوع خطا	صورت درست	فراوانی	مقوله نحوی	معادل انگلیسی	معادل فارسی	معادل انگلیسی و فراوانی	مدخل
۳	aquatic	۱	(ص.)	acquatic	آبزی	acquatic 1	آبزی (ص.)
۳	aquatic	۱	(ص.)	acquatic	آبی	acquatic 1	آبی (ص.)
۳	irrigated	۶۷	(ص.)	irritated	آبی	irritated 67	
۳	آپگار	۱	(ا.)	apgar	آپگار	apgar 1	آپگار (ا.)
۳	atherosclerosis	۱	(ص.)	atherosclorosis	آترواسکلروتیک	atherosclorosis 1	آترواسکلروتیک (ص.)
۳	آتشفشان	۱	(ا.)	volcano	آتش‌فشان	volcano 1	آتش‌فشان (ا.)
۳	آتشفشانی	۲	(ص.)	volcanic	آتش‌فشانی	volcanic 2	آتش‌فشانی (ص.)
۳	آدرنرژیک	۱	(ص.)	adrenergic	آدرینرژیک	adrenergic 1	آدرینرژیک (ص.)
۳	آدرنرژیک	۱	(ص.)	adrenergic	آدرنرژیک	adrenergic 1	آدرنرژیک (ص.)
۳	آذرگشسب	۱	(ا.)	Azar Goshnasb	آذرگشسب	Azar Goshnasb 1	آذرگشسب (ا.)
۳	arbuscular	۱	(ص.)	Arbuskular	آربوسکولار	Arbuskular 1	آربوسکولار (ص.)

تصویر ۱۷. نمونه اول از فرهنگ اغلاط ماشین‌خوان.

H	G	F	E	D	C	B	A
۳	enzyme	۱	(ا.)	enzime	انزیم	enzime 1	انزیم (ا.)
۳	enzymes	۴	(ا.)	enzyms	انزیم‌ها	enzyms 4	
۳	aeruginosa	۲	(ا.)	aeroginosa	آیروژینوزا	aeroginosa 2	آیروژینوزا (ا.)
۳	mirror	۱	(ا.)	miror	آیینه	miror 1	آیینه (ا.)
۳	initiation	۴	(ا.)	bias	ابتکار	bias 2	ابتکار (ا.)
۳	infection	۱	(ا.)	affection	ابتلاء	affection 1	ابتلاء (ا.)
۳	dimensions	۱	(ا.)	dimentions	ابعاد	dimentions 1	ابعاد (ا.)
۳	ایدیوپاتیک	۱	(ص.)	idiopathic	ایدیوپاتیک	idiopathic 1	ایدیوپاتیک (ص.)
۳	epigallocatechin	۱	(ا.)	epigallcoatechin	اپیگالوکتکین	epigallcoatechin 1	اپیگالوکتکین (ا.)

تصویر ۱۸. نمونه دوم از فرهنگ اغلاط ماشین‌خوان.

H	G	F	E	D	C	B	A
۱	effect	۳	(ا.)	effectiveness	اثر	effectiveness 3	اثر (ا.)
۱	effect	۱	(ا.)	efficacy	اثر	efficacy 1	
۱	effectiveness	۶	(ا.)	effect	اثربخشی	effect 6	اثربخشی (ا.)
۳	efficacy	۱	(ا.)	effecacy	اثربخشی	effecacy 1	
۳	Asir	۱	(ا.)	Atir	اثیر	Atir 1	اثیر (ا.)
۳	socio	۱	(ص.)	economic	اجتماعی	economic 1	اجتماعی (ص.)
۳	implementation	۲	(ا.)	implemenatation	اجراء	implemenatation 2	اجراء (ا.)
۱	performance	۱	(ا.)	perform	اجراء	perform 1	
۳	Ahmad	۱	(ا.)	Ahamd	احمد	Ahamd 1	احمد (ا.)
۳	resuscitation	۱	(ا.)	ressuscitation	احیاء	ressuscitation 1	احیاء (ا.)
۳	free will	۱	(ا.)	freewill	اختیار	freewill 1	اختیار (ا.)
۳	free will	۱	(ا.)	free-will	اختیار	free-will 1	

تصویر ۱۹. نمونه سوم از فرهنگ اغلاط ماشین‌خوان.

H	G	F	E	D	C	B	A
۳	اسکولیوز	۱	(ا.)	scoliosis	اسکولیوز	scoliosis 1	اسکولیوز (ا.)
۳	document	۱	(ا.)	ducument	سند	ducument 1	سند (ا.)
۳	diarrhea	۴	(ا.)	diarrheia	اسهال	diarrheia 4	اسهال (ا.)
۳	Isfahan	۱	(ا.)	Isfafan	اصفهان	Isfafan 1	اصفهان (ا.)
۳	principle	۴	(ا.)	principal	اصل	principal 4	اصل (ا.)
۳	modification	۱	(ا.)	modiffication	اصلاح	modiffication 1	اصلاح (ا.)
۳	principle	۲	(ص.)	principal	اصلی	principal 2	اصلی (ص.)
۳	depression	۱	(ا.)	depession	افسردگی	depession 1	افسردگی (ا.)
۳	depression	۴	(ا.)	deperssion	افسردگی	deperssion 4	
۳	robinia pseudoacacia	۲	(ا.)	rubinia pseudoacacia	اقاقیا	rubinia pseudoacacia 2	اقاقیا (ا.)
۳	economic	۱	(ص.)	economic	اقتصادی	economic 1	اقتصادی (ص.)
۳	اقلیم	۱۸	(ا.)	climate	اقلیم	climate 18	اقلیم (ا.)

تصویر ۲۰. نمونه چهارم از فرهنگ اغلاط ماشین‌خوان.

از این ۲۹۶ ردیف ۲۶۲ ردیف یعنی ۸۸/۵۲ درصد به خطای نوع سوم (خطای املایی) مربوط بود. پس از آن خطای معنایی و خطای نحوی هر یک با ۱۷ (۵/۷۴ درصد)مورد به صورت مشترک در جایگاه دوم قرار گرفتند. در ادامه نمونه‌ای از خطاهای هر دسته توضیح داده می‌شود:

از میان رایج‌ترین خطای مشاهده شده در داده‌ها یعنی خطای املایی می‌توان به نمونه‌های زیر اشاره کرد:

ISSN: 2745-3673
http://www.Rcsj.ir

جدول ۳. نمونه‌ای از خطاهای املایی (با برچسب ۳).

ردیف	واژه فارسی	واژه انگلیسی	صورت خطای واژه فارسی	صورت خطای واژه انگلیسی
۱	اصفهان	Isfahan	----	Isfafan
۲	افسردگی	depression	----	deperssion/Depression
۳	الکل	alcohol	----	alchohol
۴	اقلیم	climate	اقیلم	----
۵	الکترونیک	electronic	التکرونیک	----
۶	اولانزاپین	olanzapine	اولانزاپین	----

همان‌گونه که در جدول ۳ و همچنین تصاویر ۱۷ تا ۲۰ مشاهده می‌شود خطاهای املایی غالباً به دلیل عدم دقت در تایپ کلمات ایجاد می‌شود. برای نمونه گاه حرفی حذف یا اضافه می‌شود (مانند alchohol به جای alcohol) یا جای برخی حروف تغییر می‌کند (مانند deperssion به جای depression یا التکرونیک به جای الکترونیک) و گاه با فشردن کلید خطا و همجوار با یک کلید، حرف خطا تایپ می‌شود. همچنین گاه کلمات بر اساس اصول نگارش فرهنگستان باید جدا یا پیوسته باشند و این جدا یا پیوسته‌نویسی خطا نیز در این دسته قرار گرفت. برای نمونه مشاهده کانیزایی به جای کانی‌زایی یا کانیسازی به جای کانی‌سازی.

خطاهای معنایی و مقوله نحوی هر دو با فراوانی ۱۷ به صورت مشترک در جایگاه بعدی قرار گرفتند. در ادامه نمونه‌هایی از هر دسته ذکر می‌شود:

خطاهای معنایی: در جدول ۴ به چند نمونه از خطاهای معنایی اشاره شده است:

جدول ۴. نمونه‌ای از خطاهای معنایی (با برچسب ۲).

ردیف	واژه فارسی	واژه انگلیسی	صورت خطای واژه فارسی	صورت خطای واژه انگلیسی
۱	قرمز	red	----	black
۲	دریا	sea	----	ocean
۳	خلاصه	summary	----	introduction
۴	پایان نامه	thesis	----	dissertation
۵	بیده	bidet	----	fleece
۶	ارتقاء	enhancement	----	intervention

همان‌گونه که در جدول ۴ مشاهده می‌شود، خطاهای معنایی به مواردی اشاره داشت که برای واژه فارسی مندرج در مدخل، معادل انگلیسی نادرست استفاده شده بود. برای نمونه، واضح است که معادل انگلیسی کلمه قرمز در انگلیسی واژه red است که به خطا black درج شده است. یا معادل انگلیسی لفظ پایان نامه همان thesis است که به خطا dissertation درج شده که به معنی رساله دکتری است و ... .

جدول ۵. نمونه‌ای از خطاهای مقوله نحوی (با برچسب ۱).

ردیف	واژه فارسی	واژه انگلیسی	صورت خطای واژه فارسی	صورت خطای واژه انگلیسی
۱	آکریل	acryl	----	acrylic
۲	اجراء	performance	----	perform
۳	اختیار	option	----	optional

ISSN: 2745-3673
http://www.Rcsj.ir

۴	ایمن	safe	----	safety
۵	تولیدمثلی	reproductive	----	reproduction
۶	یونان	Greece	----	Greek

همانطور که در در جدول ۵ آمده است خطاهای مقوله نحوی خطاهایی را شامل می‌شوند که برای یک واژه فارسی، معادل انگلیسی از نظر ریشه درست اما از نظر مقوله نحوی نادرست است و بدین ترتیب معنی واژه فارسی هم به درستی منتقل نمی‌شود. برای نمونه ایمن safe و ایمنی safety است حال آنکه در داده‌ها برای لفظ ایمن از safety استفاده شده است. به عنوان یک نمونه دیگر معادل انگلیسی واژه فارسی آکریل لفظ acryl است که به خطا لفظ acrylic استفاده شده است و ... .

بحث و نتیجه‌گیری

در این پژوهش هدف آن بود تا ضمن استفاده از اطلاعات پیکره پژوهش فلاحتی قدیمی فومنی (۱۳۹۲) (به عنوان داده خام و اولیه) و انجام اصلاحات و تغییرات (وفق شیوه‌نامه تدوین‌شده توسط محقق) یک واژه‌نامه پیکره‌محور فارسی به انگلیسی تولید شود. با توجه به اینکه پیکره‌ها داده‌های طبیعی را نشان می‌دهند منطقاً می‌بایست خطاهای رؤیت‌شده در پیکره به همان صورت حفظ می‌شد. از طرفی طبق اصول و قواعد فرهنگ‌نگاری، واژه‌نامه باید عاری از هر نوع خطایی باشد. برای حل این مشکل خطاهای رؤیت‌شده در پیکره در واژه‌نامه اصلاح شد و صورت خطا به همراه سایر اطلاعات مربوط به آن ردیف از فایل اکسل به فرهنگ اغلاط منتقل شد. مختصات محصول اول (واژه‌نامه پیکره‌محور) بدین شرح بود:

- در این واژه‌نامه در مجموع ۱۱۹۴۹ مدخل فارسی بدست آمد. تعداد کل معادل‌های انگلیسی ثبت‌شده برای ۱۱۹۴۹ واژه مدخل فارسی، ۱۷۹۹۱ مورد بوده است. به عبارت دیگر به طور متوسط برای هر دو واژه فارسی سه معادل انگلیسی ثبت شده است. در این واژه‌نامه در مجموع ۴۲ بار از علامت ر.ک. استفاده شده است (مربوط به ستون H). تعداد صورت‌های مساوی مشاهده‌شده در این واژه‌نامه ۱۰۹ مورد بوده است (مربوط به ستون I). از نظر مقوله‌های نحوی در مجموع و پس از انجام تمام تغییرات و اصلاحات مشخص شد که بیشترین تعداد کلمات عناوین مقالات، به مقوله نحوی اسم مربوط بودند. در مجموع ۸۰۷۲ مدخل فارسی به مقوله نحوی اسم (ا.) مربوط تعلق داشت که معادل ۶۷/۵ درصد مدخل‌ها بود. مقوله نحوی صفت (ص.) با ۳۰۰۹ مورد و ۲۵/۱۸ درصد دومین مقوله نحوی رایج در کلمات عناوین مقالات بود. گروه اسمی (گ.ا.) با ۵۳۸ مورد و ۴/۵ درصد در جایگاه سوم قرار گرفت. سایر مقوله‌های نحوی در مجموع کمتر از ۳ درصد از کل مدخل‌ها را تشکیل دادند. این مقوله‌ها به ترتیب به شرح زیر بود: حرف اضافه (۱۵۴ مورد)، پیشوند (۱۰۰ مورد)، پسوند (۱۶ مورد)، قید (۱۹ مورد) و ضمیر (۴ مورد).

محصول دوم در پژوهش حاضر فرهنگ اغلاط بود. این فرهنگ حاوی ۲۹۶ ردیف و ۲۶۸ واژه متمایز فارسی بود. از این ۲۹۶ ردیف ۲۶۲ ردیف به خطای نوع سوم (خطای املایی) مربوط بود. پس از آن خطای معنایی و خطای نحوی هر دو با ۱۷ مورد به صورت مشترک در جایگاه دوم قرار گرفتند. خطاهای املایی غالباً به دلیل عدم دقت در تایپ کلمات ایجاد می‌شود. برای نمونه گاه حرفی حذف یا اضافه می‌شود یا جای برخی حروف تغییر می‌کند و گاه با فشردن کلید خطا و هم‌جوار با یک کلید، حرف خطا تایپ می‌شود. همچنین گاه کلمات بر اساس اصول نگارش فرهنگستان باید جدا یا پیوسته باشند و این جدا یا پیوسته‌نویسی خطا نیز در این دسته قرار گرفت. برای نمونه، مشاهده کانیزایی به جای کانی‌زایی یا کانیسازی به جای کانی‌سازی.

خطاهای معنایی و مقوله نحوی هر دو با فراوانی ۱۷ در جایگاه بعدی قرار گرفتند. خطاهای معنایی به مواردی اشاره داشت که برای واژه فارسی مندرج در مدخل، معادل انگلیسی نادرست استفاده شده بود. برای نمونه، معادل انگلیسی کلمه قرمز در انگلیسی واژه red است که به خطا black درج شده بود. یا معادل انگلیسی لفظ پایان‌نامه thesis است که به خطا dissertation درج شده بود که به معنی رساله دکتری است و ... . خطاهای مقوله نحوی نیز خطاهایی را شامل می‌شوند که معادل انگلیسی یک واژه فارسی از نظر ریشه درست اما از نظر مقوله نحوی نادرست است و بدین ترتیب معنی واژه فارسی هم به درستی منتقل

ISSN: 2745-3673
http://www.Rcsj.ir

نمی‌شود. برای نمونه، ایمن safe و ایمنی safety است حال آنکه در داده‌ها برای لفظ ایمن از safety استفاده شده است. به عنوان یک نمونه دیگر معادل انگلیسی واژه فارسی آکریل لفظ acryl است که به خطا لفظ acrylic استفاده شده است.

آنچه از کل فرایند پژوهش حاضر می‌توان نتیجه‌گیری کرد آن است که:

الف- با وجود انجام پژوهش بر روی نشریات معتبر وزارت عتف، در مقالات نشریات کم و بیش خطای املایی و انواع دیگر خطاها وجود دارد و این بدین معنی است که لازم است در این خصوص دقت‌نظر بیشتری به عمل آید. برای نمونه، فلاحتی قدیمی فومنی (۱۳۹۲) با بررسی چکیده‌های انگلیسی مجلات رتبه‌دار وزارت عتف دریافت که در چکیده‌های ۲۴ مجله، ۲۷ نوع خطا و در مجموع ۱۴۳۹ خطا وجود داشته است که عمده این خطاها دستوری، واژگانی و املائی بوده‌اند.

ب- جمع بندی دیگر از این پژوهش آن بود که با توجه به مقتضیات محیط رایانه و الزام انجام پیش‌پردازش در داده‌ها، رسم‌الخط پایگاه‌های اطلاعاتی گاه از شیوه‌های استاندارد فاصله می‌گیرد که این تنوع گاه باعث می‌شود اطلاعات با وجود حضور در منابع پایگاه، به درستی بازیابی نشوند.

ج- همچنین مشخص شد که جدای از بحث تألیف مقاله، مسئله ویراستاری پیش از انتشار نیز از اهمیت خاصی برخوردار است و غفلت از این مهم مسئول بخشی از خطاهای موجود بوده است. این خطاها هم به واژه‌های فارسی و هم به واژه‌های انگلیسی مربوط بوده که در فرهنگ اغلاط ثبت و ضبط شد.

د-بحث دیگر و مهم در خصوص این تحقیق نیاز به استفاده از متخصصان زبان در نشریات علمی است. منظور از متخصص زبان، زبان فارسی برای مقالات فارسی و همچنین زبان خارجی متناسب با زبان خارجی به کار رفته در نشریات است. به عبارتی، برای نگارش چکیده و عنوان انگلیسی حضور متخصص زبان می‌تواند مؤثر باشد. البته لازم است متخصص زبان تسلط کافی بر اصول نگارش را داشته باشد و به دلیل عدم تسلط موضوعی لازم است با نویسندگان مقالات در ارتباط باشد و این در حالی است که نشریات حتی در صورت استفاده از ویراستار نیز غالباً بین این متخصصان و نویسندگان مقالات ارتباطی برقرار نمی‌کنند.

ه- در مجموع می‌توان نتیجه گرفت که پژوهش‌های پیکره‌محوری نظیر پژوهش‌های حاضر به‌خصوص در ابعادی بزرگتر در حوزه فرهنگ‌نگاری می‌تواند نحوه استفاده متخصصان از معادل‌های انگلیسی کلمات تخصصی فارسی را به گونه‌ای مناسب به تصویر بکشد.

کاربردهای عملی پژوهش

برونداد پژوهش حاضر در حوزه‌های مختلف دارای کاربرد است که در این بخش به اختصار توضیح داده می‌شود. تولید این واژه‌نامه می‌تواند به محققان ایرانی که در حال نگارش یا ترجمه مقاله به انگلیسی هستند، یاری برساند. انتخاب معادل از میان معادل‌های مختلف برای نویسندگان و مترجمان دغدغه‌ای مهم محسوب می‌شود که در این خصوص نیز پژوهش حاضر می‌تواند مفید باشد. همچنین روش استفاده‌شده در این پژوهش (مطالعه پیکره‌محور) بوده که می‌تواند رویکردی عینی در فرهنگ‌نگاری و چیدمان معادل‌های هر واژه را پیش روی فرهنگ‌نویسان قرار دهد.

در این پژوهش سه نوع خطای عمده در معادل‌گزینی (مقوله نحوی، معنایی و املایی) استخراج و ثبت شد. این اطلاعات می‌تواند چراغ راه متولیان و سیاست‌گذاران نشریات کشور باشد تا در این خصوص دقت‌نظر بیشتری داشته باشند. از این رو پژوهش حاضر نه تنها یک واژه‌نامه و منبع استخراج واژه که منبعی برای مطالعات زبانشناسی دیگر نیز محسوب می‌شود بدین صورت که محققان می‌توانند از طریق رفتار معادل‌گزینی نویسندگان و حتی خطاهای مشاهده‌شده، تحلیل‌های مختلفی را به انجام برسانند. یافته‌ها و به ویژه خطاها می‌تواند عاملی برای یافتن منشأ این مشکلات باشد که گاه اشکال توسط نویسنده و گاه اشکال تایپی و مربوط به اپراتور می‌باشد و بررسی این موارد و اقدام برای رفع آنها می‌تواند باعث ارتقاء کیفیت انتشار مقالات باشد.

درونداد اطلاعات خطا در نظام‌های بازیابی اطلاعات هم خود تبعاتی را به همراه دارد که از آن جمله می‌توان به عدم بازیابی اطلاعات با وجود موجود بودن آن اطلاعات در پایگاه اشاره کرد. معطوف کردن توجه مخاطب و جامعه علمی به امکان بروز

ISSN: 2745-3673
http://www.Rcsj.ir

این‌گونه خطاها می‌تواند ایجاد نظام‌های کنترلی بیشتر بر نحوه درونداد اطلاعات توسط اپراتورها را سبب گردد که این مهم می‌تواند در بازیابی درست اطلاعات تأثیرگذار باشد.

دانشجویان، اساتید و متخصصان رشته‌های زبانشناسی، زبانشناسی‌رایانشی، زبان و ادبیات فارسی و زبان‌های دیگر می‌توانند به حوزه ویراستاری و ترجمه مقالات علمی به عنوان یک حوزه کاری مؤثر نگاه کنند. نویسندگان مقالات حوزه‌های موضوعی مختلف در حین تبدیل مقاله خود به انگلیسی به واژگان انگلیسی نیاز پیدا می‌کنند، به ویژه در ترجمه عنوان و چکیده. این محققان می‌توانند معادل‌های انگلیسی بخشی از واژه‌های مورد نیاز خود را از این واژه‌نامه تهیه کنند. اطلاعات بسامدی معادل‌ها نیز می‌توانند در مواردی و نه همیشه یک عامل کمکی برای انتخاب معادل‌های مناسب‌تر توسط کاربر باشد.

این واژه‌نامه ماشین‌خوان می‌تواند به عنوان یک منبع مرجع از طریق کتابخانه‌ها در اختیار محققان و اعضای هیئت‌علمی یا دانشجویان تحصیلات‌تکمیلی قرار گرفته و تمایل نویسندگان فارسی مقالات علمی را به استفاده از واژه‌های انگلیسی در حین برگردان مقالات خود از فارسی به انگلیسی به تصویر بکشد، یعنی مشخص کند چه واژه‌های انگلیسی بیشتر از سوی نویسنگان برای یک واژه فارسی خاص مورد استفاده قرار می‌گیرد.

محدودیت‌های پژوهش

پژوهش حاضر مانند هر پژوهش دیگری دارای محدودیت‌هایی است که به برخی از آنها اشاره می‌شود. نخستین محدودیت حجم داده‌های مورد استفاده بوده است به طوری که در مجموع از واژگان موجود در ده هزار جفت عنوان مقاله (فارسی و انگلیسی) استفاده شد. طبیعی است در صورت استفاده از حجم بزرگتری از داده‌ها امکان افزایش تعداد لغت و نیز بسامد رخداد معادل‌ها و درنتیجه تغییر ترتیب معادل‌ها در هر مدخل وجود می‌داشت. محدودیت دوم به نوع داده مورد استفاده باز می‌گردد که صرفاً عنوان مقالات فارسی و عناوین معادل انگلیسی آنها بوده است. بنابراین، در این واژه‌نامه بافت یا نحوه استفاده از واژه ها در قالب عبارات و جملات لحاظ نشده که چنانچه در پژوهش‌های آتی لحاظ شود می‌تواند کمک‌کننده باشد. محدودیت دیگر به واژه‌های معادل موجود در پیکره بر می‌گردد که از معیار یکسانی کامل استفاده شد و مفاهیم هم‌معنا و دارای صورت‌های مختلف به عنوان واژه‌های یکسان یا مترادف شناسایی نشدند. از درج تلفظ واژه فارسی هم به خاطر مخاطبان فارسی زبان این پژوهش پرهیز شد و صرفاً در مواردی که تلفظ واژه دشوار بود از اِعراب برای تسهیل قرائت واژه استفاده شد.

زمینه‌هایی برای مطالعه بیشتر

هیچ پژوهشی کامل نیست و بنابراین همواره برای توسعه هر پژوهشی راه‌های مختلفی وجود دارد. از طرفی متغیرهایی چون زمان، بودجه و اهداف، چارچوبی مشخص را برای هر پژوهش مشخص می‌کنند تا پژوهش در چارچوبی معین و مشخص به انجام برسد. با این مقدمه مشخص است که پژوهش حاضر نیز نمی‌تواند پژوهشی کامل محسوب شود و بنابراین در این بخش چند نمونه از موضوعات پژوهشی ذکر می‌شود که در امتداد پژوهش حاضر در آینده قابل انجام است.

در این پژوهش بر روی ۱۰۰۰۰ جفت عنوان مقاله کار شد. محققان دیگر می‌توانند نمونه آماری بزرگتری را برگزیده و بر روی تعداد بیشتری از عناوین این کار را انجام دهند. در این پژوهش صرفاً بر روی واژه‌های موجود در عنوان کار شد. محققان علاقه‌مند می‌توانند واحدهای بزرگتری از متن مانند چکیده یا متن کامل مقالات را برگزینند یا دست‌کم در کنار واژه‌های عنوان، واژگان کلیدی را نیز مد نظر قرار دهند. محقق حاضر با استفاده از روش نمونه‌برداری تصادفی سلسله‌مراتبی نمونه‌های واژگانی را از حوزه‌های موضوعی مختلف استخراج کرد. سایر پژوهشگران می‌توانند بر روی یک حوزه تخصصی خاص این کار را تکرار کنند. برای نمونه آنها می‌توانند صرفاً عناوین مربوط به رشته هنر، فنی مهندسی، علوم انسانی یا هر رشته و حوزه دیگری را انتخاب کنند. همان‌گونه که پیشتر ذکر شد دسته‌بندی داده‌ها در وبگاه مرکز منطقه‌ای در قالب دسته‌بندی‌های کلان بوده است. برای نمونه حوزه علوم انسانی، فنی مهندسی، ... . این در حالی است که در هر حوزه نظیر حوزه علوم انسانی نیز زیرحوزه‌های مختلفی وجود دارد. بنابراین، محققان دیگر در صورت علاقه‌مندی می‌توانند نه تنها بر روی یک حوزه نظیر علوم انسانی که بر روی یکی از زیرحوزه‌های آن مانند روانشناسی، جامعه‌شناسی، زبان‌شناسی و ... تمرکز کنند.

مجله‌ی علمی پژوهش در علوم رایانه، شماره ۲۶، تابستان ۱۴۰۱، ص ۴۱- ۱۳

ISSN: 2745-3673
http://www.Rcsj.ir

در این پژوهش واژه‌نامه و فرهنگ اغلاط به صورت ماشین‌خوان و در قالب فایل اکسل خام تولید شد. محققان دیگر می‌توانند از طریق کدنویسی برنامه‌ای تهیه کنند تا جستجوی واژه و معادل‌ها برای کاربر آسان‌تر شود. همچنین در این پژوهش در بخش فرهنگ اغلاط چند نوع خطای غالب که تشخیص مرجع و ریشه آن مشخص بود، بررسی شد. محققان دیگر می‌توانند سایر خطاها را نیز بررسی کرده و به این مجموعه اضافه نمایند. برای نمونه علائم زبرزنجیری و اِعراب.

ISSN: 2745-3673
http://www.Rcsj.ir

منابع و مراجع

[۱] ابراهیمی، ف. (۱۳۹۳). فرهنگ بسامدی صور خیال در دیوان فرخی سیستانی (پایان‌نامه کارشناسی ارشد، دانشکده ادبیات و علوم‌انسانی، دانشگاه ایلام).

[۲] دهقانی، ع. ا. (۱۳۵۴). تهیه فرهنگ بسامدی دیوان ناصرخسرو (پایان‌نامه کارشناسی ارشد، دانشکده ادبیات و علوم‌انسانی، دانشگاه شیراز).

[۳] رحیمی، س. (۱۳۸۸). واژه‌نامه بسامدی، توصیفی و ریشه‌شناختی منتخبی از اندرزنامه‌های پهلوی (پایان‌نامه کارشناسی ارشد، دانشکده ادبیات و علوم‌انسانی، دانشگاه شیراز).

[۴] سپنج، د. (۱۳۵۵). واژه‌نامه بسامدی رساله القدس (پایان‌نامه کارشناسی ارشد، دانشکده ادبیات و علوم‌انسانی، دانشگاه تهران).

[۵] فلاحتی قدیمی فومنی، م. ر. (۱۳۹۲). بررسی وضعیت چکیده‌نویسی انگلیسی در مجلات فارسی علمی-پژوهشی وزارت علوم، تحقیقات و فناوری (حوزه علوم پایه) سال ۱۳۹۰ و ارایه راهکاررای بهبود آن. شیراز: مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری و نامه پارسی.

[۶] فلاحتی قدیمی فومنی، م. ر. (۱۳۹۲). واژه‌نامه برگردان نام و نام‌خانوادگی نویسندگان خارجی نوشته‌شده با حروف انگلیسی به فارسی با استفاده از تحلیل رخدادمحور. شیراز: تخت جمشید.

[۷] فلاحتی قدیمی فومنی، م. ر. (۱۳۹۲). ساخت انسانی پیکره موازی دوزبانه (انگلیسی-فارسی) عناوین مقالات مجلات رتبه‌دار وزارت عتف. گزارش نهایی طرح پژوهشی، شیراز: مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری.

[۸] مارانی، ف. (۱۳۹۲). بررسی بسامدی و معنائی واژگان و ترکیبات عربی در شعر رودکی (پایان‌نامه کارشناسی ارشد، دانشکده زبان‌های خارجی، دانشگاه اصفهان).

[۹] ملکی، ا. (۱۳۹۷). فرهنگ بسامدی صور خیال در غزلیات سنائی (پایان‌نامه کارشناسی ارشد، دانشکده ادبیات و علوم‌انسانی، دانشگاه ایلام).

[۱۰] نعمتی، ح. (۱۳۹۸). نرم افزار *ACS*. دریافت شده در تاریخ ۱۵ تیر ۱۳۹۹ از طریق ایمیل.

[11] Bergenholtz, H., & Agerbo, H. (2015). Lexicographical structuring: The number and types of fields, data distrtibution, searching and data presentation. Retrieved March 12, 2021, from https://pure.au.dk/portal/files/104533592/_Lexicographica_Lexicographical_structuring_the_number_and_types_of_fields_data_distribution_searching_and_data_presentation.pdf.

[12] Čermák (2010). Notes on compiling a corpus-based dictionary. *Lexikos*, 20, 559-579.

[13] Čermák, F. (2011). Notes on compiling a frequency-based dictionary. Retrieved February 12, 2020, from https://www.researchgate.net/publication/307661883_Notes_on_Compiling_a_Corpus-_Based_Dictionary.

[14] Čermák, F., & Kren, M. (2005). New generation corpus-based frequency dictionaries: The case of Czech. *International Journal of Corpus Linguistics*, *10*(4), 453-467.

[15] Cornai, A., Halacsy, P., Nagy, V., Oraveez, C., Tron, V., & Varga, D. (2006). Retrieved January 12, 2021 from https://www.aclweb.org/anthology/W06-1701/

[16] Davies, M., & Gardner, D. (2010). *A frequency dictionary of contemporary American English* (1st ed.). The USA: Routledge.

[17] De Rocher, J. E., Miron, M. S., Patten, S., & Pratt, C. C. (1973). Retrieved November 22, 2020 from https://files.eric.ed.gov/fulltext/ED098814.pdf.

[18] Dima, G. (2008). *Lexicography, translation and dictionary use*. Proceedings of the *International Conference on Translation Studies: Retrospective and Prospective Views* (2nd ed.), Nov. 1st-2nd, 2007, (pp.68-75) Galati, RSEAS.

[19] Dimaa, G. (2012). A terminological approach to dictionary entries: A case study. *Procedia - Social and Behavioral Sciences*, *63*, 93-98.

[20] Ferrett, Emma, & Dollinger, Stefan (2021). Is digital always better? Comparing two English print dictionaries with their digital counterparts. International Journal of Lexicography, *34*(1), 66-91.

ISSN: 2745-3673
http://www.Rcsj.ir

[21] Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. Retrieved on Jan. 12, 2020 from http://www.statisticshowto.com/stratified-random-sample

[22] Martin, J. R. (1989). Factual writing: Exploring and challenging social reality (language education), (2nd ed.). England: Oxford University Press.

[23] Miron, M., & Pratt, C. (1973). Manual for the development of language frequency counts. Retrieved Dec. 15, 2020 from https://files.eric.ed.gov/fulltext/ED098815.pdf.

[24] Zhang, Songshan, Xu, Hai, & Zhang, Xian (2021). International Journal of Lexicography, *34*(1), 1-38.