مجله علمی پژوهش در علوم انسانی و تحقیقات میان رشته ای ، شماره ۲۰ ، تابستان ۱۳۹۹جلد۱، ص ۱۲۰-۱۰۰
http://www.jase.ir

تهیه پیکره (فرهنگ) املائی واژه‌های هم‌آوا، هم‌نویسه، خطا در املاء و هم‌نام‌ها در زبان فارسی

مجله علمی پژوهش در علوم انسانی و تحقیقات میان رشته ای (سال ششم)
شماره ۲۰ / تابستان ۱۳۹۹ / ص ۱۲۰-۱۰۰

شاپوررضا برنجیان [۱]
[۱]عضو هیات علمی مرکز منطقه‌ای
نویسنده مسئول:
شاپوررضا برنجیان

چکیده

یکی از مهم‌ترین گام‌ها در طراحی و ساختِ سیستم‌های بازشناسی گفتار و تصحیح خطاهای آوایی و غلط یاب‌ها و استاندارد سازی متونِ فارسی نحوه‌ی تهیه‌ی یک پیکره و واژه‌نامه‌ی بی‌عیب و نقص در یک زبانِ خاص است. برای چنین عملی، داشتن مقادیر بسیار زیادی از داده‌های تصفیه شده‌ی استانداردی (که قبلاً بارها بررسی شده باشند) از آن زبان خاص ضروری است. هدف این پژوهش تهیه‌ی گنجینه‌ای از واژه‌های هم‌آوا و واژه‌های هم‌نویسه جهت جوابگویی به نیازهایی از این دست می‌باشد. جامعه‌ی پژوهش این پژوهش را تمامی واژه‌های بخش اول فرهنگ معین (۱۳۷۱) تشکیل می‌دهد. این بخش شامل پنج هزار و دویست و هفتاد و هفت (۵۲۷۷) صفحه و حدوداً بیش از سی و هفت هزار (۳۷۰۰۰) **مدخل** می‌باشد. روش پژوهش ما ترکیبی از روشهای زبانشناسی و رایانشی است. نتایج به دست آمده نشان داد که کل واژه‌های استخراج شده ۱۳۷۷۰ مدخل همراه با معانی واژه‌ها می باشند که از این تعداد ۲۸ ٪ را «هم-آواها» و ۳۸ ٪ را «هم‌نویسه‌ها» و ۱۵ ٪ را «امکان خطا در املاء» و ۱۹ ٪ را نیز «هم‌نام‌ها» تشکیل می‌دهند. در این تحقیق برای اولین بار واژه‌نامه‌ی «امکان خطا در املاء» که شامل هم‌آواها و واژه‌های دارای واج‌های مشابه بودند، استخراج گردید تا بتواند در اختیار آموزگاران و مدرسین دوره‌های ابتدایی برای تدریس واژه‌های دشوار قرارگیرد. این پیکره‌ها برای اشتراک گذاری و دسترسی آسانِ پژوهشگران، می‌توانند به صورت برخط در سامانهٔ مراکز اطلاع رسانی برای استفادهٔ کاربران نصب گردند.

کلید واژه‌ها: پیکره، هم‌آوا، هم‌نویسه، هم‌نام، خطا در املاء، زبان فارسی.

۱- مقدمه

پیکره در لغت به معنی "زمینه، شالوده، اساس" و در اصطلاح امروز "شکل، نقش، تصویر، صورت، تمثیل، چهره، عکس و صورت نگاشت" می‌باشد (دهخدا، ۱۳۷۱).

بنابر تعریف، پیکره زبانی عبارت است از مجموعه‌ای از متن‌های نوشتاری یا گفتاری آوانویسی شده که می‌توان در توصیف و تحلیل زبان از ان بهره گرفت (kennedy ،1998). به عبارت دیگر پیکرهٔ واژگانی (corpus) به مجموعه‌ای از داده‌های خام واژگانی زبانی گفته می‌شود که می‌توان از آن‌ها برای تحلیل زبان بهره گرفت. گر چه در گذشته، بسیاری از زبان‌شناسان بر اهمیت پیکره زبانی در بیشتر بررسی‌هایشان تاکید کرده‌اند، اما در دوران جدید است که تکیه بر داده‌های واقعی زبانی به صورت گسترده‌ای رواج یافته و شرط اساسی بسیاری از پژوهش-های نظری و کاربردی مانند: گویش‌شناسی، دستورنویسی، سازماندهی، تنظیم و تفکیک، جستجو و دستیابی سریع داده‌های زبانی شده است و افق‌های تازه‌ای در برابر پژوهندگان گشوده و باعث پیدایش شاخه‌ای تخصصی در حوزه زبان‌شناسی رایانه‌ای شده است. این شاخه، با نام زبان-شناسی پیکره‌ای تنها در آخرین دهه‌های قرن بیستم ایجاد شد و در همین زمان کوتاه تبدیل به یکی از فعال‌ترین و پرکاربردترین زمینه‌ها شده است. تالیف کتاب‌های متعدد و ایجاد مجامع علمی، مراکز پژوهشی، نشریه‌های تخصصی و بین المللی و سایت‌های اینترنت برای این شاخه، گویای این واقعیت است. پیکرهٔ هم‌آوا و هم‌نویسه، هم‌نام و امکان خطا در املاء حاضر، از فرهنگ معین (۱۳۷۱) استخراج گردید.

انسان‌ها غالباً به دلیل بی توجهی و یا عدم دقت یا عدم دانش کافی خطای تایپی انجام می‌دهند. خطاهای تایپی دستهٔ بزرگی از خطاهای معمول در متون الکترونیکی و تایپی هستند که توسط انسان و یا نرم‌افزارها ایجاد می‌شوند (دستغیب، ۱۳۹۵). در هنگام نوشتن نامه یا املاء یا هر سند دیگر ممکن است در اثر بی توجهی یا بی دقتی یا کم بود دانشِ نویسنده، واژه‌های مختلف به دلیل مشکلات موجود در خط فارسی به شکل‌های گوناگون نوشته شوند که فقط یکی از آن‌ها در جمله صحیح باشد و مابقی نادرست محسوب گردد.

به جهت کمک به حل مشکلات نوشتاری زبان فارسی، ما به انجام این پژوهش اقدام کردیم و چهار واژه‌نامه که بترتیب شامل هم‌آواها، هم‌نویسه‌ها، هم‌نام‌ها و واژه‌های امکان خطا در املاء، که مستخرج از فرهنگ شش‌جلدی معین می‌باشند را تهیه نمودیم، تا ضمن پاسخ‌گویی به مسئله‌ی فوق، در پردازش رایانه و تهیه‌ی غلط یاب مورد استفاده قرارگیرند و با کاربردی کردن این محصول در سایر زمینه‌ها بتوانیم خدمتی هر چند اندک در راستای آموزش صحیح خط و زبان فارسی به نو آموزان انجام داده باشیم. در ادامه به پژوهش‌های انجام‌گرفته در مورد اشکالات خط فارسی و چگونگی تهیه‌ی پیکره‌ها خواهیم پرداخت.

۱-۱- بیـــان مسئله

صفوی (۱۳۹۲) هم‌آوا و هم‌نویسه را در یکجا آورده و می‌نویسد:

معمولاً در کتاب‌های معنی شناسی به هنگام بحث در بارهٔ روابط مفهومی، بخشی نیز به معرفی هم‌آوا – هم‌نویسه یا تشابه اختصاص می-یابد که الفاظی متشابه‌اند و اتفاقاً هیچ رابطه‌ی مفهومی‌ای با یکدیگر ندارند. در سنتِ مطالعات ادبی، هم‌آوا – هم‌نویسه تحت عنوان جناس تام معرفی می‌شود که آمیزه‌ای از جناس خط و جناس لفظ است، جناس خط یا آنچه **هم‌نویسه** نامیده می‌شود، شرایطی است که دو یا چند واژه،

به یک شکل نوشته می‌شود، ولی تلفظ‌های مختلف داشته باشند، مانند «کرم» / karam /، «کرم» / kerem /، «کرم» / korom /، «کرم» / kerm / در این مورد می‌توان «کَرَم» / karam / به معنی «کر هستم» را نیز مطرح ساخت که با «کَرَم» به معنای «بخشش» هم‌نویسه به حساب می‌آیند و به دلیل اختلاف درجایگاه تکیه، به یک شکل تلفظ نمی‌شوند (صفوی، ۱۳۹۲: ۱۱۱). **هم‌آوایی** یا آنچه جناس لفظ نامیده می‌شود، شرایط چند واژهٔ مختلف است که به یک شکل تلفظ می‌شوند و به صورت‌های مختلف نوشته می‌شوند، برای نمونه «خوار» و «خار»، «قضا» و «غذا»، «خویش» و «خیش» و جز آن، هم نویسی ناشی از نظام نوشتاری است و هم‌آوایی از قرض‌گیری و تحول آوایی ناشی می‌شود.

در ادامه صفوی (۱۳۹۲) برای توضیح بیشتر در مورد هم‌آوا – هم‌نویسه یا چند معنایی (که ما آن‌ها را در اینجا هم‌نام می‌نامیم) چنین می‌نویسد:

هم‌آوا - هم‌نویسه (هم‌نام‌ها) یا جناس تام، شرایطی است که چند واژه‌ی متفاوت، بدون هیچ ارتباطی میان معانی‌شان، به یک شکل نیز نوشته شوند، به گونه‌ای که انگار یک واژه چند معنای متفاوت دارد. در این باره می‌توان سه واژه‌ی مختلف «شانه»{وسیله‌ی آرایش مو}، «شانه» {کتف}و «شانه»{جای تخم‌مرغ}یا «دوش»{کتف}، «دوش»{وسیله‌ی حمام}و «دوش» {دیشب}را نمونه آورد (صفوی، ۱۳۹۲). اهل زبان هیچ ارتباطی میان مفهوم «دوش» که واژه‌ای قرضی از فرانسه است و ابزاری برای حمام کردن به حساب می‌آید و «دوش» به معنی دیشب نمی‌یابند؛ به همین دلیل، اینها را واژه‌هایی متمایز از یکدیگر می‌دانند. در فرهنگ‌های مختلفی آورده می‌شوند تا مشخص شود که واژه-های متفاوتی به شمارمی‌روند، به عبارت دیگر چند معنایی که (نام دیگرِ هم‌نام‌ها و هم‌آوا- هم‌نویسه‌ها است) از جمله آشناترین روابط مفهومی-ای است که در سنت مطالعه‌ی معنی به آن پرداخته‌اند و آن را شرایطی معرفی می‌کنند که یک واحد زبانی از چند معنی برخوردار شود. چند معنایی را می‌توان در سطوح مختلفی چون تکواژ، واژه، گروه و جمله مورد بررسی قرار داد. هر چند در سنت، چند معنایی تنها در سطح واژه اهمیت یافته و تحلیل آن نیز به همین سطح محدود گشته است. به هنگام طرح چند معنایی در سطح واژه، به دونوع چند معنایی همزمانی و درزمانی برمی‌خوریم که یک صورت زبانی می‌تواند یا در یک مقطع زمانی از چند معنی برخوردار باشد و یا در طول زمان معنای متفاوتی بیابد، نمونه‌ای از چند معناییِ در زمانی را می‌توان در واژه‌هایی نظیر «گران» {سنگین / پرهزینه} «شوخ» {چرک / بزله گو} دانست (صفوی، ۱۳۹۲).

امکان خطا در املاء: لازم به ذکر است که با توجه به حروف زبان فارسی تعدادی واژه نیز در این زبان قابل توجه است و علت اصلی آن‌ها وجود حروف هم‌آوا (و نه هم‌نویسه) در نظام نوشتاری فارسی است که گاهی می‌توانند برای نوآموزان در نوشتار و املای خط فارسی تردید و مشکل ایجاد نمایند مانند: دو واژه‌ی «اَبَد و عبد» و «اِتر و عَطر» و ... که ما علاوه بر تهیه فهرست واژه‌های هم‌آوا و هم‌نویسه برای اولین بار به تهیه فهرست اینگونه واژه‌ها نیز پرداختیم همچنین برای اولین بار تعدادی واژه را نیز که نه هم‌آوا هستند و نه هم‌نویسه و نه هم‌نام اما در هنگام نوشتار، نوآموزان را دچار تردید می‌کنند (مانند: طِبّ و تَب) نیز استخراج کردیم و همه‌ی آن‌ها را «امکان خطا در املاء» نام نهادیم.

در زبان فارسی ۵ همخوان و یک واکه وجود دارد که در برابر آن‌ها، الفبای فارسی ۱۵ حرف به کار می‌برد که عبارتند از:

/s/ س – ث – ص /q/ غ – ق

/t/ ت – ط /h/ هـ - ح

/z/ ز – ذ – ض – ظ /ā/ و / a/ و / e/ اَ - ع

واج /s/ که به آوای حروف «س»، «ث» و «ص» تعلق دارد، همخوان ششی، برونسو، سخت، بیواک، سایشی، دهانی و لثوی است.

واج /t/ که به آوای حروف «ت» و «ط» تعلق دارد، همخوان ششی، برونسو، سخت، بیواک، دمیده، انفجاری، دهانی و دندانی می‌باشد.

واج /z/ که به آوای حروف «ز»، «ذ»، «ظ» و «ض» تعلق دارد، همخوان ششی، برونسو، نرم، واکدار، سایشی، دهانی و لثوی است.

واج /q/ که به آوای حروف «ق» و «غ» تعلق دارد، همخوان ششی، برونسو، نرم، واکدار، انفجاری، دهانی و ملازی است.

واج /h/ که به آوای حروف «ه» و «ح» تعلق دارد همخوان ششی، برونسو، سخت، بیواک، سایشی، دهانی و چاکنایی است.

واج /a/ که به آوای حروف «اَ» و «ع» تعلق دارد. واکه‌ی پیشین، باز، گسترده‌ی باز، کوتاه است.

واج /e/ که به آوای حرف «اِ» و «عِ» تعلق دارد. واکه‌ی پیشین، نیم‌بسته، گسترده، کوتاه است.

واج /ā/ کشیده، پسین، باز ، گرد.

بر همین اساس در خط فارسی اگرچه واج‌ها به یک حالت تلفظ می‌شوند اما به صورت‌های گوناگون تحریر می‌گردند و این عامل باعث تغییر معنایی (مانند حیاط و حیات) می‌شوند. همچنین بدلیل اینکه مصوت‌های کوتاه (َ ُ ِ) در خط و نوشتار فارسی نمایش داده نمی‌شوند، در هنگام خوانش نوآموز را دچار سردرگمی می‌کنند، لذا اینگونه واژه‌ها باعث مشکلاتی در املاء و آموزش نوشتاری فارسی بویژه برای نوآموزان می‌شود (مانند، طِب و تَب).

با کمی گشت و گذار در وب، نظرات تارنماهای گوناگون، فیس بوک و دیگر شبکه‌های اجتماعی به این نتیجه می‌رسیم که غلط‌های املایی فاحشی در نوشتارِ افرادی حتی با مدارک کارشناسی و کارشناسی‌ارشد به چشم می‌خورد و در نتیجه باید پذیرفت که امروزه نوشتن املاء مشکل بزرگ نسل جدید شده است، (اگرچه این نسل از نوشتن در فضای مجازی ترسی ندارند زیرا در این فضا گویا نوشتن املای غلط جزو خصوصیات منفی افراد محسوب نمی‌شود و در صورتی که شخص احساس کند ممکن است ارزش علمی وی پایین بیاید بلافاصله با تغییر صفحه کلید، مطالبش را به صورت فارسی انگلیسی "فنگلیش" می‌نویسد.) اما با این وجود هیچ کس دوست ندارد نوشته‌هایش غلط باشد. بی انصافی است اگر تمامی تقصیرهای این مسئله را فقط به گردن نویسندگان بیندازیم، زیرا خط فارسی سرشار از مشکلات است و کمتر کسی را می‌توان پیداکرد که هنگام نوشتن دچار تردید نشود. در زیر به تعدادی از اشکالات این خط اشاره می‌کنیم.

اشکالات خط فارسی

اولین مطلب در اغلب کتاب‌های دستور زبان مربوط به حروف الفبا است. شمار حروف الفبای فارسی را ۳۲،۳۳،۳۴،۳۵و گاهی تا ۳۷ حرف تخمین زده‌اند.

در لغت نامه دهخدا (۱۳۷۳) و فرهنگ فارسی معین (۱۳۷۱) در زیرِ کلمه الفبا نوشته شده که «الفبا، یا حروف فارسی ۳۳ حرف است» و در صفحه ۱۱۷ در زیر مدخل ابتث (abtas) ۲۸ حرف الفبای عربی آورده شده و در ادامه نوشته شده که «ایرانیان در این میان حروف ذیل

را افزوده‌اند «پ، چ، ژ، گ»». بدین ترتیب شمار حروف الفبای فارسی ۳۲ حرف فرض شده است. حسن عمید (۱۳۸۰) در فرهنگ عمید حروف زبان فارسی را ۳۲ عدد میداند. در صورتی که فرهنگستان زبان (۱۳۸۱) نشانه‌های خط فارسی را ۳۴ نشانه ذکر کرده و علاوه بر آن ۹ نشانه ثانوی را نیز تحت عنوان نشانه‌های خط فارسی ذکر نموده است. که کلاً ۴۳ نشانه محسوب می‌شود.

علاوه بر این، به مسائل و مشکلات اساسی خط فارسی می توان بطور اجمال چنین اشاره کرد.

- نبودن سه واکهٔ کوتاه (ـَ ـُ ـِ) در داخل خط فارسی، که در زمان خواندن مطلب می‌بایست سه واکه را حدس زد.
- مشخص نبودن تشدید.
- ناتوانی خط در نشان دادن واکه مرکب (ow) و همچنین عدم توانایی آن برای نشان دادن استرس‌های یک گویش خاص.
- وجود حرف‌هایی که عملاً یکی هستند و نقطه در آن‌ها ایجاد تفاوت کرده است، مانند: «ت– پ– ث».
- اشکال نقطه در خط این است که چون با اضافه و کم کردن نقطه حرف مبدل به حرف دیگری می‌شود لذا عملاً اگر بطور تصادفی در نوشتن، نقطه از قلم افتاده باشد، خواننده باید با حدس و گمان محل نقطه و بودن یا نبودن آنرا دریابد. مانند: «بر – بز – پُز – یز – تِز – یر – تر – پُر و ...» تازه این همه حدس در مورد دو حرف است، حال اگر این تعداد به سه یا چهار یا پنج حرف برسد بی‌شک مدت‌ها طول می‌کشد تا به نتیجه برسیم.
- خط فارسی از راست به چپ نوشته می‌شود، در حالی که فرمول‌ها و اعداد در علوم مختلف مثل ریاضی، فیزیک و شیمی از چپ به راست نوشته می‌شوند و مخلوط این دو، امکان پذیر نیست و در زمان نوشتن مشکلاتی را ایجاد می‌کنند.
- اکثر حروف در اوّل، وسط و آخر به یک شکل نوشته نمی‌شوند مانند: (عـ - ـعـ - ـع – ع) (یـ - ـیـ - ی) توجه داشته باشید که اگر حروف فارسی را با قرار گرفتن آن‌ها در جاهای مختلف کلمه حساب کنیم به بیش از ۱۲۵ شکل می‌رسد.
- ناتوانی خط فارسی در جداسازی واژه‌ها.
- کمبود منابع الکترونیکی در حوزهٔ NLP برای فارسی.

حال، وجود این همه مشکل در خط و زبان فارسی و سردرگمی در هنگام آموزش آن، ما را بر آن داشت، تا بتوانیم هر چند اندک در کمک به رفع بنیادی این مشکلات با تهیه‌ی پیکره‌ی فرهنگ املایی واژه‌های هم‌آوا و هم‌نویسه اقدام نماییم، و با این کار، راه را برای محققین و پژوهشگران جهت دستیابی به پیکره‌ها هموار نماییم.

۲- پیشینه تحقیق

اصطلاح پیکره را به ویژه زبان‌شناسان ساختگرا به کار می‌بردند و همواره تاکید می‌کردند که توصیف یک زبان یا گویش باید مبتنی بر داده‌های گردآوری شده و تحلیل این داده‌ها باشد و با آنکه فراگیری و بزرگی پیکره عامل تعیین کننده‌ای در افزایش دقت و اعتبار نتایج به شمار می‌آمد، محدودیت‌های نیروی انسانی و زمان در برابر این عامل قرار داشت و درنتیجه، بیشتر به نمونه‌های برگزیده‌ای اکتفا می‌شد. تنها از دهه ۱۹۶۰ بود که توانستند با کمک رایانه، مجموعه‌های بسیار بزرگی از متن‌های معتبر زبانی فراهم نمایند و آغازگر همه این‌ها، پیکره زبانی

براون بود که نمونه‌ای معیار از زبان انگلیسیِ امروز برای استفاده در رایانه‌های رقمی به شمار می‌آید (Svartvik؛ 7:1992). این پیکره را نلسون فرانسیس و هنری کوچرا (۱۹۶۴) در دانشگاه براون با گرداوری نزدیک به یک میلیون واژه از پانصد نمونه متنی پیوسته ایجاد کردند و کوشیدند انواع مقاله‌ها و گونه‌های زبانی را پوشش دهند. کار از ۱۹۶۱ اغازشد و در سال ۱۹۶۴ به پایان رسید. این مدت در مقایسه با کارهای گذشته و با توجه به حجم نسبتاً زیاد داده‌ها در آن زمان بسیار کم به نظر میرسید (francis and Kucera: 1964).

یکی از مهم‌ترین پیکره‌هایی که در همین دهه، اما بدون بهره‌گیری از رایانه فراهم آمد و از این نظر آن را در شمار پیکره‌های پیش از دوران الکترونیک قرار می‌دهند، پیکره بررسی کاربرد زبان انگلیسی با seu است که راندلف کرک از ۱۹۵۹ در دانشگاه لندن اغاز کرد و گرداوری، آوانگاری و اماده‌سازی آن تا سال ۱۹۸۹ طول کشید، چرا که برای یک میلیون واژه‌ی آن، یک میلیون برگه با هفده سطر از متن‌های مربوط به سال‌های ۱۹۵۳ تا ۱۹۸۷ با دست نوشته شد. البته در کنار این پیکره‌ی نوشتاری، بخش گفتاری آن با عنوان پیکره لندن - لوند (بررسی انگلیسی گفتاری ،...) به وسیله "یان سوارتویک"، در دانشگاه لوند سوئد از سال ۱۹۷۵ اغاز شد. از آنجا که این بخش با کمک رایانه فراهم شد، از همتای نوشتاری خود پیشی گرفت و با حدود نیم میلیون واژه در سال ۱۹۸۰ ارائه شد (kennedy, 1998: 32).

پیکره زبانی میتواند بسیار بزرگ، فراگیر و نماینده تمامی یک زبان و یا گونه‌ای از آن باشد، به شکل برگه‌های یادداشت یا پرونده‌های رایانه‌ای شامل متن‌های کامل یا گزیده‌هایی از آن‌ها، و یا بخش‌های پیوسته‌ای از متون یا گزیده‌ای از نقل قول‌ها و نکات و حتی فهرست‌های واژگانی پیکره می‌تواند، ویژه بررسی خاصی فراهم شود و یا در برگیرنده مجموعه‌ی عظیم و بی‌ساختاری از متون گوناگون باشد که برای منظورهای گوناگون به کار رود.

امروزه به دلیل نیازهای پژوهشی جامعه علمی در ایران، سازمان‌های مختلف نسبت به تهیه پیکره‌های مختلف اقدام نمودند که از آن جمله می‌توان به موارد زیر اشاره نمود:

پیکره‌ی به‌روزشونده‌ی پارسی ، این پیکره، یک پیکرهٔ به‌روزشوندهٔ (monitor) متون خبری در زبان فارسی است که هر روز متون جدید از خبرگزاری‌های معتبر به صورت خودکار در شش دستهٔ موضوعی «سیاسی»، «اقتصادی»، «اجتماعی»، «فرهنگی»، «علمی» و «ورزشی» به آن اضافه می‌شود (در آینده متون غیرخبری نیز به پیکره اضافه خواهند شد). پیکرهٔ حاضر از این حیث که تنوعات زبانی را با انتخاب داده از ژانرهای مختلف پوشش می‌دهد،

پیکره‌ی بی‌جن‌خان مجموعه‌ای است از متون فارسی شامل بیش از ۲ میلیون و ۶۰۰ هزار کلمه که با ۵۵۰ نوع برچسب POS برچسب‌گذاری شده‌اند. این پیکره که در پژوهشکدهٔ پردازش هوشمند علائم تهیه شده است همچنین شامل بیش از ۴۳۰۰ تگ موضوعی چون سیاسی، تاریخی و ... برای متون است. نسخه‌ی کامل پیکره بی‌جن‌خان با عنوان «پیکره‌ی متنی زبان فارسی».

فرهنگ بسامدی بر اساس پیکره متنی زبان فارسی امروز که توسط محمود بی جن خان و مهدی محسنی تهیه شده است، این فرهنگ بسامدی میزان کاربرد کلمات زبان را بر حسب فراوانی وقوع آن‌ها در یک پیکره زبانی که نماینده آماری زبان است نشان می‌دهد. کلمات پربسامد و کم‌بسامد می‌توانند شاخص ارزشمندی برای پی بردن به مولفه‌های فرهنگی یک جامعه زبانی باشند. میزان بسامد کلمات در یک

فرهنگ بسامدی تابع بازه زمانی است که پیکره فرهنگ در آن زمان گردآوری شده است. بنابراین اطلاعات موجود در فرهنگ بسامدی رابطه مستقیم با تحولات تاریخی یک جامعه زبانی دارد. امروزه فرهنگ بسامدی در تحقیقات بین‌رشته‌ای مانند آموزش زبان دوم به خارجیان، زبان‌شناسی نظری و کاربردی و فناوری‌های زبان در محیط رایانه از قبیل ترجمه ماشینی و طراحی مدل زبانی برای سامانه‌های پردازش گفتار، کاربرد موثری دارد.

پایگاه داده‌های زبان فارسی به منظور ایجاد مجموعه‌ای بزرگ از پیکره‌های گوناگون زبان فارسی امروز توسط پژوهشگاه علوم انسانی و مطالعات فرهنگی طراحی شده است و در آن متن‌های برگزیده ادبی، علمی، هنری، سیاسی و مانند این‌ها از گونه‌های نوشتاری و گفتاری فارسی وجود دارد. این سامانه امکان جستجوی واژه‌ها، ترکیب‌ها ، باهمایی‌ها و بررسی بسامد آن‌ها را به همراه گزارش‌های آماری متنوعی از متون فراهم می‌کند. افزون بر اصل متن‌ها و واژه‌های به کاررفته در آن‌ها، معنی، مقوله دستوری، آوانگاشت و ریشه یا (بن‌واژه) بسیاری از واژه‌ها نیز در پایگاه وجود دارد.

یکی از مهم‌ترین گام‌ها در طراحی و ساخت سیستم‌های بازشناسی گفتار پیوسته با دایرهٔ لغات بزرگ (LVCSR)، تعلیم مدل زبانی (Language Model)برای مدل‌کردن نحوهٔ توالی کلمات در یک زبان خاصّ است. برای انجام چنین کاری، داشتن مقادیر بسیار زیادی از دادگان متنی (Text) از آن زبان همراه با برچسب‌های لازم ضروری است. از طرفی در بسیاری از پروژه‌های تحقیقاتی در زمینهٔ پردازش زبان طبیعی(Natural Language Processing) ، چنین دادگانی مورد نیاز است. **پیکرهٔ متنی زبان فارسی**، دادگانی است که توسط پژوهشگاه توسعه فناوری‌های پیشرفته جواب‌گوی این‌گونه نیازها در شناسایی گفتار فارسی و پردازش زبان طبیعی است.

پایگاه انتشار و تولید داده‌های زبانی پیکره‌گان؛ این پیکره‌ی متنی زبان فارسی، که توسط پژوهشکده پردازش هوشمند علائم تهیه و منتشر شده است، مجموعه‌ای از متون نوشتاری و گفتاری رسمی زبان فارسی است که از منابع واقعی همچون روزنامه‌ها، سایت‌ها و مستنداتِ از قبل تایپ‌شده، جمع‌آوری شده، تصحیح گردیده و برچسب خورده است. حجم این دادگان حدوداً ۱۰۰ میلیون کلمه است و از منابع مختلف تهیه گردیده و دارای تنوع بسیار زیادی است. ۱۰ میلیون کلمه از این پیکره با استفاده از ۸۸۲ برچسب نحوی-معنایی به صورت دستی توسط دانشجویان رشته‌ی زبان‌شناسی برچسب‌دهی شده‌اند و هر پرونده بر حسب موضوع و منبع آن طبقه‌بندی شده است. این پیکره که توسط پژوهشکده پردازش هوشمند علائم تهیه شده است، برای استفاده در آموزش مدل زبانی و سایر پروژه‌های مربوط به پردازش زبان طبیعی مناسب است.

۲-۱- مبانی نظری

یکی از راه‌های توصیف و تحلیل معنا، بررسی روابط مفهومی واژه‌ها و مقایسه‌ی آن‌ها با یکدیگر است، در این میان «هم‌ها» که هم‌ردهٔ تضادمعنایی و شمول‌معنایی و چندمعنایی هستند، زیر مجموعهٔ روابط معنایی واژه‌ها محسوب می‌گردند و دستهٔ واژگانی مهمی را تشکیل می‌دهند.

هم‌ها

اصطلاح‌های «هم‌آوا»، «هم‌نویسه»، «هم‌نام» و «هم‌معنا» که به «هم‌ها» (Homos) معروف هستند، برای توصیف روابط بین کلمه‌های زبان به کار می‌روند، توضیح کامل هر کدام در زیر آمده است.

۱-هم‌آوا (Homophony) یعنی واژه‌هایی که شبیه هم تلفظ می‌شوند امّا بطور متفاوت نوشته می‌شوند و اغلب معانی متفاوت نسبت به هم دارند. مثلاً واژه‌های فارسی «خوار – خار؛ خواست – خاست؛ فترت – فطرت؛ ناضر – ناظر» دارای نوشتار و معنای متفاوتی هستند در حالیکه تلفظ یکسانی دارند. همچنین در فارسی کلمات «اثیر – اسیر – عصیر»، «صور (شیپور) - سور»، «قضا – غذا – غزا» و «عرض – ارز»، واژه‌های هم‌آوا محسوب می‌شوند. در دستور سنتی به اینگونه واژه‌ها جناس لفظ یا لفظی می‌گویند.

۲-هم‌نام: هم‌نام‌ها (Homonymy) واژه‌هایی هستند که به یک صورت نوشته می‌شوند و به یک صورت نیز تلفظ می‌شوند امّا معناهای متفاوتی دارند (لانگمن و اشمیت، ۲۰۰۲: ۲۴۱) در سنت مطالعات ادبی، هم‌نامی تحت عنوان جناس تام معرفی می‌شود که ترکیبی از جناس خط و جناس لفظ است. هم‌نامی یا جناس تام، شرایطی است که چند واژه‌ی متفاوت، بدون هیچ ارتباطی میان معانی‌شان، به یک شکل تلفظ شده و به یک شکل نیز نوشته می‌شوند، گویا اینکه آن یک واژه، چند معنای کاملاً متفاوت دارد. در این مورد می‌توان سه واژه‌ی مختلفِ شانه (وسیلهٔ آرایش مو)، شانه (کتف) و شانه (جای تخم‌مرغ) و همچنین شیر (لبنی) و شیر (حیوان) و شیر (آب) یا دوش (وسیلهٔ حمام) و دوش (دیشب) را ذکر نمود. اهل زبان ارتباطی میان مفهوم دوش که واژه‌ای قرضی از زبان فرانسه است و ابزاری برای حمام کردن محسوب می‌شود و دوش به معنای (دیشب) نمی‌یابند؛ به همین دلیل، آن‌ها را واژه‌هایی متمایز از یکدیگر می‌دانیم. در فرهنگ لغت‌ها، این واژه‌ها تحت مدخل‌های مجزایی آورده می‌شوند تا مشخص شود که واژه‌های متفاوتی به شمار می‌روند (صفوی،۱۳۷۹: ۱۱۰). اغلب تشخیص بین کلمات هم‌نام و چند معنا (polysomy) واضح نیست زیرا واژه‌های **چندمعنایی**، واژه‌هایی هستند با یک صورت نوشتاری، چندین معنای مرتبط و نیز ریشه مشترک دارند. مثلاً واژهٔ head به بالایی بدن (سر) یا بالاترین مقام یک شرکت یا واحد اطلاق می‌شود که از این لحاظ به دلیل فرم نوشتاری یکسان با واژه‌های «هم‌نام» اشتباه می‌شوند. با این حال یکی از راه‌های تمیز بین آن‌ها، به اعتقاد یول (۲۰۰۶)، رجوع به مدخل خاص هر کلمه در فرهنگ لغت است به این معنا که اگر کلمه‌ای دارای چند معنای مرتبط به هم باشد فقط یک مدخل به آن تعلق می‌گیرد که زیر آن مدخل فهرستی از معانی مختلف کلمه آورده می‌شود ولی چنانچه دو کلمه هم‌نام باشند مدخل جداگانه‌ای به هر یک از آن‌ها اختصاص داده می‌شود. مثلاً واژه‌های فارسیِ شیر، شانه، دوش نمونه‌هایی از کلمات هم‌نام هستند که هر یک از معانی مختلف آن‌ها مدخلی جداگانه دارد. به عبارت دیگر کلمات هم‌معنا هریک با توجه به داشتن معانی مختلف و غیر مرتبط، مدخل‌های جداگانه‌ای دارند. در حالیکه در کلمات چندمعنا یک مدخل به هر یک اختصاص پیدا می‌کند و معنای مختلف مرتبط به هم در همان مدخل وارد می‌شوند. نتیجه آن که در مورد واژه‌های چندمعنا، سنّت غالب آن است که واژه‌ها، زیرِ یک مدخل درج و معانی مختلف، با شماره متمایز می‌شوند، زیرا بین معانی مختلف نوعی پیوند معنایی مشاهده می‌شود. تامس (۱۹۹۵) معتقد است که کلمات چندمعنا و هم‌نام برای گویشوران غیربومی ایجاد اشکال می‌کند زیرا آن‌ها به دانش زمینه‌ای فرهنگی که سخنوران بومی آن زبان بر آن تسلط دارند، آشنا نیستند. تا آنجا که، نمی‌دانند کدام معنا را در بافت خاص خود بکار

ببرند. حال آنکه سخنور بومی زبان حتی اگر معنای یک کلمه را هم نداند با مراجعه به فرهنگِ آن زبان به راحتی می‌تواند معنای صحیح بافت مورد نظر خود را بیابد (رحیمیان و رزمجو،۱۳۸۹: ۱۸۰). بنابراین همانگونه که قبلاً نیز ذکر شد، هم‌نام‌ها واژه‌هایی هستند که شکل آوایی و نوشتاری یکسان، اما معانی مختلف دارند، مانند واژهٔ «شیر» (= شیر آب، شیر جنگل، شیر لبنی). در هم‌نامی، یکسانی، هم در صورت آوایی و هم در صورت نوشتاری وجود دارد. باید توجه داشت که «هم‌آوایی» و «هم‌نویسی» مواردی از «هم‌نامی ناقص» هستند

۳-هم‌نویسه: واژه‌های هم‌نویسه (Homography) واژه‌هایی هستند که به صورت یکسان نوشته شده امّا بطور متفاوت تلفظ می‌شوند و معانی گوناگونی هم دارند (لانگمن و پلات ۱۹۹۲ و لانگمن و اشمیت، ۲۰۰۲: ۲۴۱). واژه‌های فارسی، مانند: شِکَر، شُکر، شُکِر؛ جَنگ، جُنگ؛ مِهر، مَهر، مُهر واژه‌هایی هم‌نویسه هستند. صفوی (۱۳۷۹: ۱۱۱) معتقد است که هم‌نویسه همان جناس خط است. برای نمونه واژه‌های کَرَم، کِرِم، کُرُم، کِرم. در این مورد می‌توان کَرَم به معنای کَرِ هستَم را نیز مطرح نمود که با کَرَم به معنای بخشش هم‌نویسه محسوب می‌شوند ولی به دلیل تفاوت در جایگاه تکیه، به یک شکل تلفظ نمی‌شوند. در سنت مطالعات ادبی، هم‌نویسه‌ای تحت عنوان جناس ناقص معرفی می‌شود که شامل واژه‌هایی است که فقط هم‌شکل می باشند امّا از نظر تلفظ متفاوت هستند. مانند: گَزیدن – گُزیدن، خَلق – خُلق و قَمَری – قُمری. در سنت ادبی گاهی جناس مرکب را نیز ذکر می‌کنند و آن بدین صورت است که کلمات مرکب شبیه کلمات دیگر می‌شوند، مانند: کَمَند و کَمَند (کم هستند) یا شهریار و شهرِ یار (همایی،۱۳۸۳).

۴- هم معناها (مترادف‌ها) واژه‌هایی هستند که هم‌نویسه و هم‌آوا نیستند و فقط در معنا مشترک می‌باشند.

۵-خطاهای املائی (هم واجی) گاهی ممکن است نوآموزان و کودکان در هنگام نوشتن املاء واج‌های هم‌آوا را بجای یکدیگر بکار ببرند. مانند: «تَب» بجای «طِبّ» یا «صُوَر» (ج صورت) بجای «سور» (جشن)، که اینگونه واژه‌ها نه جزو هم‌آواها هستند و نه جزو هم‌معناها و نه هم-نویسه‌ها، اینگونه خطاها فقط در واژه‌های نوآموزانِ زبانهایی امکان دارد مشاهده شود که دارای واج‌های هم آوا هستند. ما اینگونه واژه‌ها را «امکان خطا در املاء» یا «هم واجی» می‌نامیم.

۱-۱-۲- انواع پیکره

همان گونه که پیش‌تر اشاره شد، پیکره‌های زبانی می‌توانند عام و گسترده بوده، و یا نمونه تصادفی از یک زبان و یا تخصص و نماینده یک گونه خاص زبان باشند. پیکره‌ها را می‌توان برای منظورهای گوناگون به کار گرفت، از جمله برای فرهنگ‌نگاری، معناشناسی، بررسی‌های دستوری، اموزش زبان و مانند اینها، پیکره‌ها را می‌توان از نظر **اندازه و گستره** به دسته‌های محدود، متوسط و عظیم تقسیم کرد. بر اساس رویکردِ **«فرم و کارکرد»** نیز، می‌توان انواع زیر را برای پیکره‌ها در نظر گرفت:

-**پیکره اشباع شده** (**saturated corpus**) که با اشاره به عوامل و متغیرهای معیّن با اطلاعات اضافی پر شده است. اشباع، ویژگی آماری است که با نسبت غنای واژگانی (موارد کاربردی سرواژه یا...) ارتباط دارد. در این مورد، کل متن به پاره‌های مشابه بخش شده است.

-**پیکره موقعیتی، (situation corpus)**، که بنا بر نیازهای خاص یک بررسی، فراهم می‌شود.

-**پیکره پایشگر، (monitor corpus)**، که دارای ابعاد تاریخی است و برای ثبت تغییرات زبانی به کار می‌رود و از برهه‌های گوناگون زبانی، نمونه‌های مشابه را دارد.

-**پیکره تطبیقی، (comparative corpus)**، پیکره‌ای چند زبانه با نمونه‌های مشابه از هر زبان است.

-**پیکره موازی، (parallel corpus)**، پیکره‌ای چند زبانه با متنی اصلی از یک زبان و ترجمه آن در زبان‌های دیگر (یا زبان‌های هدف) است.

-**پیکره دوسویه، (two sided corpus)** ، پیکره‌ای چند زبانه با متن اصلی به همه زبان‌ها و ترجمه آن‌ها به همه زبان‌ها است (Teubert؛1999).

در اغاز دهه ۱۹۸۰ شاید می‌توانستیم پیکره‌های الکترونیکی را با انگشتان یک دست بشماریم. در اغاز دهه ۱۹۹۰ تعداد طرح‌هایی که برای ایجاد پیکره‌های زبانی فعال بودند به ده‌ها می‌رسید و چند سال بعد در کشورهای مختلف جهان، فعالیت‌های بسیاری برای زبان‌های گوناگون آغاز شده بود. اکنون پس از گذشت تنها دو دهه، با جستجویی ساده در شبکه جهانی رایانه‌ها (اینترنت) بیش از سیصد پیکره بزرگ و کوچک برای بسیاری از زبان‌های جهان فهرست میشوند.

علت این گسترش شتابنده در نخستین وهله، راه یافتن رایانه‌ها در این حوزه وتغییرات بنیادی در روش‌های پردازش متن و ذخیره سازی و دستیابی به آن است. ((بدین سان، زبان شناسی پیکره‌ای اکنون ناگزیر با رایانه پیوند یافته است و همین وسیله است که سرعت شگفت انگیز، شمار پذیری کامل، تکرار و روبرداری دقیق، صحت اماری و امکان به کارگیری حجم عظیم داده‌ها را به همراه آورده است)) (kennedy؛1998).

برای زبان شناسان قدیمی‌تر، اصطلاح «زبان شناسی پیکره ای» اغلب یادآور دوران رونق پیکره زبانی یعنی دهه ۱۹۵۰ است، دوران هریس، فریس، هیل و دیگر ساختگرایان آمریکایی. اما از آن دوران به بعد است که این اصطلاح، کاربرد بیشتری یافته و به ویژه در اوایل دهه ۱۹۹۰ است که به عنوان شاخه‌ای مشخص از زبان‌شناسی، مانند زبان‌شناسی اجتماعی، روانشناسی‌زبان و زبان‌شناسیِ‌متنی که در دهه-های۱۹۶۰و۱۹۷۰ شکل گرفته‌اند، شناخته شده‌است (leech؛1992). اما زبان‌شناسی پیکره‌ای با رشته‌های یاد شده تفاوت دارد، زیرا نماینده یک حوزه مشخص از مطالعات زبانی نیست بلکه بنیادی روش شناختی برای پژوهش‌های زبانی به شمار می‌آید. در اصل و عملاً زبان شناسی-پیکره‌ای، به آسانی با شاخه‌های دیگر زبان‌شناسی می‌آمیزد. می‌توان با کمک پیکره، به بررسی‌های آوایی، نحوی و اجتماعی یا دیگر زمینه‌های زبان پرداخت و در این صورت می‌گوییم که روش‌ها و فنون زبان‌شناسی پیکره‌ای را با موضوعات آوایی، نحوی و اجتماعی زبان و مانند ان آمیخته‌ایم (leech؛1992).

از آنجا که حجم مواد زبانی (در اغلب زبان‌های امروز جهان) برای گنجاندن در پیکره بسیار زیاد است. در برخی موارد می‌توان پیکره را به صورت نمونه ایجاد کرد. تعیین حدود و گستره پیکره، وظیفه زبان‌شناسان است. برخی از متغیرهایی که در تعیین حدود دخالت دارند، عبارتند از: زبان، بُرِش یا محدوده‌ای از زمان، منطقه، شرایط و موقعیت، ویژگی‌های بیرونی و درونی متنی و .....

از این‌رو ما به تهیه پیکره املایی واژه‌های «هم‌آوا» و «هم‌نویسه» در زبان فارسی پرداختیم.

صادقی و زندی مقدم (۱۳۸۵) نسخهٔ رقومی فرهنگ‌املایی خط‌فارسی براساس دستور خط‌فارسی، مصّوب فرهنگستان زبان و ادب فارسی را تهیه نمودند. این پژوهشگران در این مورد می‌نویسند:

از آنجایی که به خاطر سپردنِ تمام قواعدی که در دستور خط فارسی مصوب فرهنگستان زبان و ادب فارسی آمده، برای همگان امکان پذیر نیست، و از سوی دیگر، عموم نویسندگان، اعم از محققان، نثر نویسان، شاعران، ویراستاران، روزنامه نگاران، منشیان و غیره، هنگام نوشتن کلمات مرکب و مشتق و بعضی کلمات دو املائی یا سه املائیِ دیگر دچار تردید می‌شوند، فرهنگستان راه چاره را دراین دید که دست به انتشار یک فرهنگ املائی مبتنی بر قواعد دستور خط فارسی بزند و املای کلماتی را که در این دستور به عهده سلیقه و ذوق نویسندگان گذاشته بود. تا آنجا که امکان دارد، بیشتر قاعده‌مند سازد، یا شیوهٔ مرجع خود را (در مواردی که کلمه دو املائی است) در آن بگنجاند. از این رو، مقرر شد که گروه دستور زبان فارسی مجموعه‌ای از کلمات رایج فارسی امروز را با قواعد دستور خط فارسی منطبق و منتشر کند (صادقی و زندی مقدم؛ ۱۳۸۵).

این اثر با وجود اینکه بسیار سودمند است، اما چون به صورت کتاب و PDF است، لذا امکان استفاده از آن محدود می‌باشد.

سمیعی (۱۳۶۷) معتقد است؛ برخی از دشواری‌ها نوعاً مربوط می‌شوند به عناصری از زبان که صورت لفظی آن‌ها به عینه در زبان زندهٔ ادبی باقی مانده، لیکن معنایی که در متون از آن‌ها مراد گرفته‌شده، منسوخ یا مهجور گردیده یا در عرف زبان کاربرد کمتری دارد.

این تشابهاتِ صوری را می‌توان به انواع زیر تقسیم کرد:

الف -تشابه در کتابت و تلفظ (مانند «افسوس» که در متون کهن به معنی استهزاء نیز به کار رفته است). ممکن است دو متشابه؛ دو یا چند لغت جداگانه باشند (مانند «بید» به معنی «باشید» و به معنی «درخت بید» و به معنی «نوعی حشره که به فرش آسیب می‌رساند»).

ب - تشابه در کتابت و فرق در تلفظ به جهت فرق در حرکات یا وجود «واومعدوله» و «واو» و «یای‌مجهول» (مانند: «شسته: شِسته، شسته»؛ «پول: به «واو معلوم» و به «واو مجهول» که لغتی است در معنی پُل»؛ «گو: فعل امر از گفتن به ضّم کاف فارسی»، «گُو به معنی گودال»).

ج -تشابه در کتابت و فرق در تلفظ به جهت فرق در تکیه (مانند، انگشتری: انگشتر + یای نکره، با تکیه در هجای اول؛ انگشتری با یای جزء کلمه و با تکیه در هجای آخر).

د - تشابه در کتابت و فرق در تلفظ به جهت فرق در مکث بالقوه و تقطیع تکواژی (اوزار: او/ زار؛ اوزار به معنی افزار).

ه - تشابه در کتابت و فرق در تلفظ به جهت فرق در دو یا چند عامل سابق الذکر. فرق معنایی این کلمات و عبارات متشابه نیز علل گوناگون دارد، از جمله:

۱- منسوخ یا مهجور شدن معنای مراد («خانه» به معنی «اتاق»).

۲- مهجوری معنای اصطلاحی («رخصت» در مقابل «عزیمت»، در اصطلاح فقهی؛ «حفاظ» در اصطلاح اهل فتوت).

– فرق در تقطیع تکواژی و ساخت واژه‌ای («دستوری» به معنیِ «اجاره» و «خدمتی» به معنی «تحفه» و «بربند» به معنی «سینه بند» در مقابل «دستوری» و «خدمتی» با یای نکره و «بربند» فعل امر از فعل مرکبِ «بربستن»).

-مهجوریِ قاعدهٔ دستوری به کاررفته (حذف شناسهٔ فعلی به قرینه).

- اختلاف در ریشهٔ («اهداء»: از «هدی» به معنی هدیه فرستادن، از «هدء» به معنی آرام دادن).

- اختلاف در تعلق زبانی (دمدمه: بر وزن فعلله در عربی به معنی خشم گرفتن، هلاک کردن؛ در فارسی به معنی مکر و فریب و افسون).

در تشابه صوری و فرق معنایی عناصرِ گویشی و لهجه‌ای و تحول آوایی نیز، همواره دخیل است. سیاههٔ این تشابهات برای استفادهٔ دانشجویان و علاقه‌مندان به متون فارسی مفید خواهد بود (سمیعی، ۱۳۶۷).

اگرچه تحقیقات زیادی در این حوزه انجام پذیرفته است، اما هیچکدام از آن‌ها به صورت عمومی منتشر نشده‌اند، و تنها آگاهی ما از این محصولات همانا مقالات این پژوهشگران ارجمند می‌باشد.

هدف این پژوهش تهیهٔ فرهنگ املائی واژه‌های هم‌آوا و واژه‌های هم‌نویسه برای زبان فارسی است. طبق پژوهش‌های انجام شدهٔ اخیر، پژوهشگران (کاشفی و همکاران (۲۰۱۴)، پیله ور و همکاران (۲۰۱۲)، دستغیب و همکاران (۲۰۱۶)) به این نتیجه رسیده‌اند که یکی از بزرگترین دستهٔ خطاها، آن دسته از خطاهائی هستند که کاربر، دانش تصحیح آنرا ندارد، این دسته همانا دستهٔ «خطاهای هم آوا» هستند.

دستغیب (۱۳۹۵) می‌نویسد: غالب روش‌های تصحیح خطا به روش تصحیح واژه‌ها به صورت منفرد تاکید دارد و تصحیح واژه را بدون در نظر گرفتن اطلاعات زبانشناسی و همچنین متنی که واژه در آن قرارگرفته است انجام می‌دهد. این نوع تصحیحِ خطا نمی‌تواند کسری از خطاها را (که شامل خطاهای آوایی (Phonetic)، شناختی (cognitive)، گرامری و چاپی (typographic) هستند) شناسایی و تصحیح نماید. خطاهای آوایی معمولاً در استفاده‌ی نادرست از حروف هم‌آوا پدید می‌آید و خطای شناختی نیز به دو دلیل پدید می‌آید، نخست در اثر عدم تسلط کاربر به زبان پدید می‌آید که معمولاً کاربر معنا و املای صحیح واژه را نمی‌داند، دوم بدلیل بیماری یا اختلالات شناختی است. خطای تایپی نیز زمانی رخ می‌دهد که کاربر املای صحیح لغت را می‌داند، ولی در زمانِ تایپِ واژه دچار مشکل شده و حرف کنارِ حرف مورد نظر را فشار می‌دهد و در نتیجه واژه را غلط تایپ می‌کند. خطای شناختی به دلیل عدم آگاهیِ کاربر از هجی و یا املای صحیح واژه نشات می‌گیرد (مانند خواندن که می‌شود خاندن)، یا شخص هنگام نوشتن به موضوع دیگری می‌اندیشد و در نوشته‌ی خود کلمه‌ای از آن موضوع را به متن می‌افزاید. خطاهای هم‌آوایی که حدوداً ۴۴ ٪ از خطاها را تشکیل می‌دهند زیر شاخه‌ای از خطاهای شناختی محسوب می‌شوند و بدین صورت است که یک حرف با حرف دیگرِ هم‌صدا یا هم‌آوا جایگزین می‌شود (مانند غلطک که می‌شود غلتک). در بسیاری از موارد تشخیص و تصحیح خطا بدون در نظر گرفتن متن امکان پذیر نمی‌باشد. استفاده از متن واژه‌ها در شناسایی و تصحیح خطا علاوه بر امکانِ شناسایی و تصحیح خطاهای معنایی (Real word error) در تصحیح واژه‌های مرکب و ... نیز کاربرد دارد.

بنا براین در این پژوهش گنجینه‌ای از واژه‌های هم آوا و واژه‌های هم‌نویسه گردآوری می‌شوند، تا بتوانند زمینه‌ای برای سایر پژوهش‌های حوزهٔ پردازش زبان طبیعی باشند. در ضمن بسیاری از کاربردهای حوزه‌ی NLP مانند غلط‌یاب‌ها به چنین منابعی نیاز دارند بنابراین، این

گنجینه لغت املایی می‌تواند در راستای تولید منابع برای زبان فارسی مفید باشد.

۲-۲- ضرورت انجام تحقیق

ضرورت اجرای این تحقیق بیشتر و در درجه‌ی اول در بازیابی اطلاعات در مراکز اطلاع رسانی الکترونیکی و رایانه‌ها در اسناد و مدارک به زبان فارسی و قابل رویت است، زیرا اگر کاربر در هنگام جستجو نتواند املای درست و صحیح واژه‌ای را تایپ نماید، بی شک اسناد بازیابی شده بسیار متفاوت‌تر از نتایج مورد انتظار خواهد بود و نتایج نهایی این پژوهش می‌تواند کمک شایانی برای رفع اینگونه نواقص در بر داشته باشد. همچنین از نتایج این پژوهش می‌توان در آموزش زبان فارسی به غیر فارسی زبانان و نوآموزان زبان فارسی و ساخت و برنامه نویسی غلط‌یاب‌های املایی و تصحیح متون و تهیه‌ی ویراستارهای الکترونیکی سود برد. همچنین با توجه به کمبود منابع زبانی برای زبان فارسی، این پیکره‌ها می‌توانند در مواردی مانند: غلط‌یاب‌ها، شناسائی فوری نویسه‌ها، تبدیل گفتار به متن و .... مورد استفاده قرارگیرند.

باید توجه داشته باشیم که فعالیت‌های دیگر این حوزه مانند "فرهنگ املائی خط فارسی" (صادقی و دیگران؛ ۱۳۸۵)، علیرغم ارزش والائی که دارند، به صورت کتاب ویا فایل pdf تهیه شده‌اند و مستقیما قابل استفاده به صورت الکترونیکی نیستند.

باید دقت داشت که خطا یا غلط نویسی در هم‌آواها معمولاً زمانی اتفاق می‌افتد که کاربر یا نویسنده معنای واژه‌ای را می‌داند ولی شکل لغوی واژه و املای درست آنرا نمی‌داند. بنابراین، فرهنگ حاضر می‌تواند در این راستا به کاربران کمک کند تا نرم افزارهای کاربردی، از این پیکره به عنوان یک منبع برای تهیهٔ کاربرد غلط‌یابی استفاده نمایند.

اکثر سامانه‌هایی که محتوای متنی تولید می‌کنند می‌بایست دارای اجزای ویراستاری و بررسی لغوی متنی باشند، به همین علت نتیجهٔ این پژوهش می‌تواند برای سامانه‌ها مورد استفاده قرار گیرد.

خط و زبان فارسی علاوه بر مشکلات تایپی و نگارشیِ دستی، مشکلات متعددی از نظر نگارش در رایانه نیز دارند. برای تهیهٔ سامانه‌های محتوایی و واژه‌نامه‌های مورد نیازِ اینگونه سامانه‌ها در کنارِ سایر منابع، نیاز است ابزارهای پردازش نیز برای زبان فارسی آماده گردد تا بتوان از آن به نحو مناسب در پژوهش‌های گوناگون استفاده نمود.

بدیهی است بکارگیری ابزارهای استاندارد موجب افزایش کیفیت و بازدهی پژوهش‌های مختلف خواهد بود. در حال حاضر عدم دسترسی به منابع استاندارد زبانی برای پژوهش‌های مختلف در حوزهٔ زبان فارسی، موجب شده تا پژوهش‌های این حوزه هم با مشکلات متعدد روبرو شود و هم از کیفیت مناسب برخوردار نباشد. لذا به نظر می‌رسد ساخت و تولید منابع و ابزارهای مناسب و با کیفیت، برای زبان فارسی بسیار مهم و از جمله پیش‌نیازها و فعالیت‌های بنیادی این حوزه باشد. از مهمترین دلایل استخراج این پیکره‌ها می‌توان به نیاز مبرم نرم‌افزارهای فارسی از جمله غلط یاب‌ها به یک پیکرهٔ کامل از یک منبع موثق و قابل استناد اشاره نمود.

۲-۳- جامعه پژوهش

جامعه‌ی پژوهش طرح حاضر را تمامی واژه‌های هم‌آوا و هم‌نویسه‌ی موجود در فرهنگ ۶ جلدی معین تشکیل می‌دهند که این واژه‌ها به دو دسته‌ی مجزا تقسیم می‌گردند:

دسته نخست عبارتند از واژه‌هایی که دارای یکی از ۱۵ حروف زیر هستند:

t: ت – ط

s: ث – س – ص

h: ح - ه –

z: ذ – ز - ض – ظ –

q: غ – ق

/ā/ و / a/ و / e/ : أ – ع

این واژه‌ها به دو دلیل به وجود آمده‌اند نخست به دلیل استفاده از رسم الخط عربی دوم به دلیل نفوذ آن زبان در زبان فارسی (مانند: ثواب و صواب، حیاط و حیات، زهر و ظَهر ، غریب و قریب).

دسته دوم واژه‌هایی که فارسی می‌باشند و به دلایل تاریخی بوجود آمده‌اند (مانند: خوار و خار).

ما از فرهنگ معین (۱۳۷۱)، «واژه‌های هم‌آوا» را که می‌توانند شامل «هم‌نام‌ها و تعدادی از هم نویسه‌ها» باشند را استخراج کردیم و همچنین «هم‌نویسه‌ها» را که می‌توانند «هم‌نام‌ها و واژه‌های غیر هم‌آواها» را نیز شامل شوند، استخراج نمودیم. همچنین واژه‌های «هم نام (چند معنایی)» که شرح آن گذشت و واژه‌های «امکان خطا در املاء» که شامل «هم‌آواها مانند: هما و حمی» و «واژه‌هایی که هنگام نوشتن ایجاد تردید می‌کنند» نیز می‌باشند را استخراج نمودیم.

فرهنگ معین شامل یک مقدمه و سه بخش است:

بخش ۱- لغات.

بخش ۲- ترکیبات خارجی.

بخش ۳- اعلام.

واژه‌های مورد نظر ما مستخرج از بخش اول فرهنگ معین (۱۳۷۱) می‌باشند. این بخش شامل پنج هزار و دویست و هفتاد و هفت (۵۲۷۷) صفحه و حدوداً بیش از سی و هفت هزار (۳۷۰۰۰) **مدخل** می‌باشد، از این تعداد، چهار هزار و صد و نود (۴۱۹۰) مدخل **هم‌آوا** و پنج هزار و شش صد و پنجاه و چهار (۵۶۵۴) مدخل **هم‌نویسه**، که شامل هم‌نام و غیر هم‌آوا و.... نیز می‌باشد. همچنین دو هزار و نه صد و دوازده (۲۹۱۲) مدخل واژه‌های **هم‌نام** و دو هزار و صد و هشتاد و هشت (۲۱۸۸) مدخل نیز شامل **امکان خطا در املاء** می‌باشد.

۳- روش پژوهش

روش مورد استفاده در تحقیق حاضر ترکیبی است از روش‌های زبانشناسی و رایانه، بدین معنی که ابتدا برای تشخیص و چگونگی تفاوت بین واژه‌های هم آوا و هم‌نویسه و .... از دانش زبانشناسی و برای سهولت و دسته بندی و الفبایی کردن و ... از رایانه، بویژه از نرم افزار اکسل استفاده کردیم؛ در این تحقیق از روش توصیفی – تحلیلیِ زمینه یابِ پیکره‌ای استفاده شده است. پیکره‌ی مورد استفاده در این پژوهش

فرهنگ شش جلدی معین بوده است.

ابتدا اشکالات مختلف املائی در زبان فارسی با استفاده از کتب گوناگون دستور زبان و زبانشناسی مورد بررسی قرارگرفت، سپس واژه-های دارای هم‌آوایی و هم‌نویسه استخراج شده و جهت استفاده‌های الکترونیکی وارد نرم افزارِ صفحه‌ی گسترده‌ی excel گردید و سپس مورد تجزیه و تحلیل قرارگرفت و در نهایت پیکره‌ای واژه‌نامه‌ای هم‌آوا و هم نویسه استخراج گردید. لذا این پژوهش شامل مراحل زیر می‌باشد:

۱-ابتدا پیشینه و شکل‌گیری واژه‌های هم‌آوا و واژه‌های هم‌نویسه در زبان فارسی مورد تحقیق و بررسی قرار گرفت.

۲-سپس تمامی واژه‌هایی که دارای حروف «ت – ط - ث – س – ص - ح - ه – ذ – ز - ض – ظ – غ – ق - أ - ع و همزه (ء) و همچنین واژه‌هایی که دارای گروهِ همخوانِ (خو)» بوده‌اند و همچنین واژه‌های هم شکل و هم‌نویسه‌ی غیر هم‌معنا و واژه‌های هم‌آوا از کتاب فرهنگ معین (۱۳۷۱) استخراج گردید.

۳-در ادامه تمامی واژه‌های فارسی هم‌آوا و هم‌نویسه، مشخص شد.

۴-برای تأیید متداول بودن واژه‌های استخراج شده، فهرست این واژه‌ها در اختیار تعدادی گویشور باسواد فارسی زبان قرارگرفت تا آن‌ها واژه-های متداول را از ناواژه‌ها (منظور از ناواژه (non sense word) شِبهِ واژه‌هایی هستند که در زبان وجود ندارد و احتمالاً به دلیل تایپ اشتباه تایپیست‌ها در لیست ما قرار گرفته‌اند مانند: «یاهاوه‌ژاو» و ...) متمایز سازند.

۵-سپس ورود واژه‌های متداول به رایانه (نرم افزار اکسل) انجام پذیرفت.

۶-در ادامه در اکسل برای واژه‌های هم‌آوا و هم‌نویسه داده پردازی (مجموعه‌سازی و مرتب‌سازی) انجام پذیرفت، پس از آن، لیست نهایی از واژه‌های هم‌آوا و هم‌نویسه در نرم افزار اکسل تهیه گردید.

برای مدیریت و کنترل واژیِ این حجم از داده‌ها می‌بایست از راهکارهای خودکار کمک گرفت، ما نیز برای این کار از نرم‌افزار اکسل استفاده نمودیم.

۴- روش و ابزار گردآوری داده‌ها

واژه‌های این فرهنگ‌ها از بخش اول (لغات) فرهنگ شش جلدی معین که شامل ۵۲۷۷ صفحه می‌باشد استخراج شده است.

مرحله اول روش استخراج واژه‌ها توسط ما، در بندهای ۲ و ۳ و ۵ مراحل پژوهش شرح داده شد، تعداد واژه‌های استخراج شده ۱۳۷۷۰ واژه را شامل گردید و همراه با معانی در نرم‌افزار اکسل وارد شد، از این تعداد پس از بررسی، تعداد ۴۱۹۰ مدخل تحت عنوانِ «پیکره‌ی هم‌آوا» مشخص و در اکسل ثبت گردید. از این پیکره با بررسی‌های به عمل آمده، ۵۶۵۴ ردیف (مدخل همراه با معانی) «هم‌نویسه» و ۲۹۱۲ ردیف (مدخل همراه با معانی) «هم‌نام» و ۲۱۸۸ ردیف نیز شامل «امکان خطا در املاء» می‌باشند. چنانچه مشاهده می‌شود مجموع واژه‌های این چهار واژه‌نامه به ۱۴۹۴۴ واژه می‌رسد، این تفاوت از آنجا ناشی می‌شود که تعدادی از واژه‌ها در بعضی از وژه‌نامه‌ها تکرار شده‌اند مثلاً در واژه‌نامه‌ی «غلط‌های املایی» برخی از واژه‌های هم‌آوا نیز آورده شده‌اند.

این واژه‌نامه‌ها به صورت جدول در اکسل برای استفاده در سایر نرم افزارها، استخراج گردید. که همراه با این گزارش در اختیار دفتر

معاونت محترم پژوهشی مرکز منطقه‌ای اطلاع رسانی علوم و فناوری قرار گرفت.

۵- روابط واژگانی

برای درک بهتر، از روابط واژگانی و بین کلمات هم‌آوا، هم‌نام و هم‌نویسه، ویژگی‌های هر کدام در جدول ذیل آمده است.

جدول۱: ویژگی‌های هم‌آوا، هم‌نام، هم‌نویسه و هم‌معنا

ویژگی‌ها و نمونه‌ها	تلفظ	معنا	نوشتار(نویسه)
هم‌آوا (غیرهم معنا، غیرهم‌نویسه، هم‌آوا) سَفَر؛ صَفَر هُما (پرنده)؛ حُمی (تب) خوار؛ خار	+	–	–
هم‌نام (غیرهم معنا، هم نویسه، هم‌آوا) شیر؛ شیر	+	–	+
هم‌نویسه(غیرهم معنا ، غیر هم‌آوا) صَفَر؛ صِفر کَرَم؛ کِرِم	–	–	+
هم معنا (غیرهم نویسه، غیر هم‌آوا) زِهی؛ آفرین	–	+	–
هم معنا، هم‌نویسه، غیر هم‌آوا سَفَر؛ سِفر (تفاوت گویشی و لهجه)	–	+	+
هم‌معنا، هم‌آوا، غیر هم‌نویسه طهران؛ تهران	+	+	–
خطاهای املایی (امکان خطا در املاء) غیر هم‌آوا، غیر هم‌نویسه، غیر هم-معنا مزارع؛ مضارع / ادات؛ عُدات (دشمنان)	–	–	–

۵-۱- شرح روابط

۱-هم‌آوا (غیرهم معنا، غیرهم‌نویسه، هم‌آوا): واژه‌های هم‌آوا واژه‌هایی هستند که، فقط از نظر آوایی شبیه هم تلفظ می‌شوند ولی از نظر معنایی و شکل نوشتاری با هم متفاوت هستند. مانند: سَفَر؛ صَفَر هُما (پرنده)؛ حُمی (تب) خوار؛ خار

۲-هم‌نام (غیرهم معنا، هم نویسه، هم‌آوا): هم‌نام‌ها واژه‌هایی هستند که، هم شبیه هم نوشته می‌شوند و هم شبیه هم تلفظ می‌گردند اما از نظر معنا با هم کاملاً متفاوت هستند. مانند: شیر؛ شیر شانه؛ شانه و ...

۳-هم‌نویسه(غیرهم معنا ، غیر هم‌آوا، هم نوشتار): هم‌نویسه‌ها واژه‌هایی هستند که، شبیهِ هم نوشته می‌شوند، اما از نظر معنا و شکل تلفظ باهم کاملاً متفاوت هستند. مانند: صَفَر؛ صِفر کَرَم؛ کِرِم

۴-هم معنا (غیرهم نویسه، غیر هم‌آوا، هم‌معنا): هم‌معناها واژه‌هایی هستند که فقط هم معنا هستند ولی از نظر تلفظ و نوشتار کاملاً متفاوت می باشند. مانند: زِهی؛ آفرین

۵-هم معنا، هم‌نویسه، غیر هم‌آوا نیز واژه‌هایی هستند که از نظر آوایی فقط باهم تفاوت دارند ولی از نظر معنا و نوشتار کاملاً شبیه همدیگر می‌باشند. مانند: سَفَر؛ سِفر (معمولاً اینگونه واژه‌ها بیشتر در تفاوت‌های گویشی و لهجه‌ها مشاهده می شوند)

۶-هم‌معنا، هم‌آوا، غیر هم‌نویسه نیز واژه‌هایی هستند که مانند هم تلفظ می‌شوند و هم‌معنا هستند ولی با شکل نوشتاری متفاوت. مانند: طهران؛ تهران تشت؛ طشت

۷- غیر هم‌آوا، غیر هم‌نویسه، غیر هم‌معنا که ما در اینجا آن‌ها را خطاهای املایی (امکان خطا در املاء) نامیدیم، و هدف ما از مطرح کردن اینگونه واژه‌ها، نشان دادن واژه‌هایی است که در هنگام نوشتار ممکن است نویسنده بخاطر شباهت تلفظ مشترک بعضی از حروفِ فارسی مرتکب خطا گردد. مانند: مزارع؛ مضارع / ادات؛ عُدات (دشمنان).

اگر چه برای هم‌معنایی در این پژوهش پیکره‌ای تهیه نشده اما بنا به ضرورت اجمالاً مورد بررسی قرار می‌دهیم.

۲-۵- هم‌معنایی

دو صورت زبانی متفاوت با معنای یکسان هم‌معنا تلقی می‌شوند. عاصی (۱۳۸۳) معتقد است که هم‌معنایی (Synonymy) مطلق کمتر وجود دارد زیرا کلمات دارای کاربرد یکسان نیستند. نمونه‌ی بارزِ هم‌معنایی مطلق دو کلمه کامپیوتر و رایانه در زبان فارسی است. به اعتقاد پالمر (۱۹۷۶) با مطالعه‌ی واژه‌های هم‌معنا می‌توان پنج امکان تمایز میان آن‌ها در نظر گرفت. **نخست** اینکه گروهی واژه‌های هم‌معنا، متعلق به گونه‌های خاص زبانی هستند مانند: «شدن، در گویش گیلکی و رفتن در گویش مشهدی» و «پاییز و خزان». **دوم** اینکه چند واژه ممکن است در سبک‌ها یا کاربردهای مختلف هم‌معنا باشند مانند: آقا، مرد، جناب. **سوم** بعضی از واژه‌ها تنها از نظر معنایی عاطفی یا سنجش معنایی از یکدیگر متمایز هستند و جنبه‌های دیگر معنایی یا معناشناختی آن‌ها یکسان است. برای مثال دولتمرد / سیاستمدار، نهان / مخفی یا صورت‌های سه‌گانه‌ی صرفه‌جو / مقتصد / خسیس را می‌توان نام برد. **چهارم**، بعضی از واژه‌ها از نظر همنشینی در محدودیت هستند، یعنی ارائهٔ آن‌ها تنها در ارتباط با واژه‌هایی خاص میّسر است. مثل آشپزخانه /* آشپزمنزل. **پنجم**، اینطور به نظر می‌رسد که معنای بعضی از واژه‌ها نزدیک به هم یا منطبق بر هم هستند. این هم‌معنایی از آن نوعی است که مورد استفادهٔ فرهنگ نویسان قرار می‌گیرد. برای نمونه هم‌معناهای ممکن برای صفت بالغ عبارتند از: به حد رشد رسیده، جاافتاده، کامل و شایسته.

مدرسی (۱۳۸۷: ۴۶۵) می‌نویسد: واژه‌ها، یا عباراتی که دارای معنای یکسانی هستند، هم‌معنا یا مترادف خوانده می‌شوند. هم‌معنایی وقتی ممکن است که بتوان به جای جزء یا اجزایی از یک زنجیرهٔ زبانی، جزء یا اجزای دیگر نهاد، بدون اینکه معنی زنجیره تغییر یابد، باید توجه داشت که دو واژه هنگامی هم‌معنا تلقی می‌شوند که مجموعهٔ مشخصه‌های آن‌ها یکسان باشد. بنظر می‌رسد یافتن دو هم‌معنای کامل دشوار است. مثلاً واژهٔ «ثابت» را همه کس به کار می‌برد، اما واژهٔ «پایا» یا «برجای» کاربرد خاص و ادبی دارد، بنابراین، با مطالعه در محتوا و مشخصه‌های واژه، معلوم می‌شود که دو واژه هیچ‌گاه به‌طور دقیق هم‌معنا نیستند، زیرا ممکن است بار محتوایی و معنایی آن‌ها به مقدار کمی با هم تفاوت داشته باشد.

به نظر می‌رسد که امکانات زبانی، اجازه نمی‌دهد که دو واژه، معنای واحدی را بطور کامل داشته باشند. فقط گاهی واژه‌ها در برخی همنشینی‌ها روی زنجیرهٔ گفتار، می‌توانند معنای واحدی بیابند. مثلاً «تر» و «خیس» در دوجملهٔ: الف) این لباس خیس است. ب) این لباس تر

است. می‌توانند در این زنجیره جانشین یکدیگر شوند، اما در جملهٔ «او تر دامن است» نمی‌توان واژهٔ «خیس» را جانشین «تر» کرد، زیرا این بافت تنها با واژهٔ «تر» می‌تواند معنای «آلوده» را برساند و قابل جانشینی با «خیس» نیست.

۳-۵- هم معنا هم شکل، غیر هم‌آوا

غالباً واژه‌های هم‌معنا هم‌شکل و غیرهم‌آوا در تلفظ واژگان در گویش‌های متفاوت مشاهده می‌شود، مانند: اَلوار – اِلوار (گلپایگانی)، الکل – اَلکل. مدرسی این هم‌معنایی را هم‌نویسه می‌نامد و می‌نویسد: هم‌نویسه (هم‌نامی ناقص) دو واژه، زمانی هم‌نویسه‌اند که صورت نوشتاری یکسان، اما صورت‌های آوایی متفاوت داشته باشند. مانند شور (Šour / Šur) و دور (Dour / Dur). صورت‌های نوشتاری و معنایی این واژه‌ها همسان، اما صورت‌های تلفظی آن‌ها جداگانه است (مدرسی؛ ۱۳۸۷: ۴۶۸).

۴-۵- چند معنایی

یک واژه که دارای چندین معنای به هم وابسته و مرتبط با یکدیگر هست چند معنایی (ploysemy) نام دارد. مانند دل به معنای قلب، مرکز، میان، جرأت، و شکم (عاصی، ۱۳۸۳). اغلب در انتخاب میانِ واژه‌های چندمعنا و واژه‌های هم‌معنا با مشکل مواجه می‌شویم ولی به عقیدهٔ پالمر (۱۹۷۶) عاملی که موجب انتخاب مقوله‌ی چندمعنایی به جای هم‌نامی می‌شود، در نظر گرفتن معنای مرکزی یا یک مرکز ثقل معنایی است. (رحیمیان و رزمجو،۱۳۸۹: ۱۸۴).

۵-۵- هم‌آوا هم‌معنا، غیر هم‌شکل

نوشتار برخی از واژه‌ها در طول زمان به دلایل مختلف از جمله سلیقه‌ها، متفاوت نوشته می‌شوند، به عبارت دیگر نگارشآن‌ها به دو شکل می‌باشد که هر دو شکل نیز صحیح محسوب می‌شوند، مانند: طهران – تهران، اطاق – اتاق، طشت - تشت و ... اینگونه واژه‌ها در لیستِ واژه-های هم‌آوا قرار می‌گیرند.

۲-۵- نتیجه گیری

با بکارگیریِ تصمیم‌گیری و منطق انسانی در مراحل مختلفِ تهیهٔ واژه‌نامه‌ها و پیکره‌های هم‌آوا و هم‌نویسه، توانستیم دو واژه‌نامه‌ی دیگری نیز تهیه نماییم، چهار واژه‌نامه‌ی زیر را بر اساس فرهنگ شش جلدی معین تهیه و همراه با این گزارش به دفتر معاونت محترم پژوهشی مرکز منطقه‌ای اطلاع رسانی علوم و فناوری تحویل نماییم که عبارتند از:

۱-واژه‌نامه‌ی (پیکره‌ی) هم‌آوای زبان فارسی براساس فرهنگ معین.

۲-واژه‌نامه‌ی (پیکره‌ی) هم‌نویسه‌ی زبان فارسی براساس فرهنگ معین.

۳- واژه‌نامه‌ی (پیکره‌ی) هم‌نام‌ها براساس فرهنگ معین.

۴- واژه‌نامه‌ی (پیکره‌ی) امکان خطا در املاء براساس فرهنگ معین.

این پیکره‌های زبانی به منظور بکارگیری در آموزش زبان فارسی، خطایاب‌های رایانه‌ای، تصحیح متون خودکار و ..... تهیه گردیده‌اند. و می‌توانند مورد استفاده‌ی پژوهشگرانِ سامانه‌های غلط یاب، تشخیص متن و .....قرارگیرند.

نتایج این پژوهش نشان داد که کل واژه‌های استخراج شده ۱۴۹۴۴ مدخل همراه با معانی واژه‌ها می باشند که از این تعداد ۲۸ ٪ را «هم‌آواها» و ۳۸ ٪ را «هم‌نویسه‌ها» و ۱۵ ٪ را «امکان خطا در املاء» و ۱۹ ٪ را نیز «هم‌نام‌ها تشکیل می‌دهند. همچنین معادل ادبیات سنتی هم‌نویسه‌ها «جناس خطی» هم‌نویسه‌های غیر هم‌آوا «جناس ناقص»، هم‌نام‌ها «جناس تام»، هم‌آواها «جناس لفظی»، و برخی واژه‌ها که با پیشوندها یا پسوندها با واژه‌های دیگر هم‌نویسه می‌شوند «جناس مرکب» (مانند کمند و کمند) می‌باشند.

در این تحقیق برای اولین بار واژه‌نامه‌ی «امکان خطا در املاء» که شامل هم‌آواها و واژه‌های دارای واج‌های مشابه بودند، استخراج گردید تا بتوانید در اختیار آموزگاران و مدرسین دوره‌های ابتدایی برای تدریس واژه‌های دشوار قرارگیرد.

در انتها پیشنهاد می‌شود که این تحقیق در مورد سایر واژه‌نامه‌ها از جمله: لغت نامه‌ی دهخدا، برهان قاطع و ... نیز انجام پذیرد و در نهایت با یکدیگر مقایسه گردند.

منابع

- برنجیان، شاپوررضا؛ (۱۳۸۹)؛ "نظری بر نواقص خط فارسی " ،اباختر ، شماره ۱۹ و ۲۰ ، اسفند ۱۳۸۹

-دستغیب ، محمد باقر (۱۳۹۵). تولید مجموعه‌ای برای مقایسهٔ اثر طول واژه و محل رخداد خطا در تصحیح خطاهای لغوی برای واژگان زبان فارسی، طرح تحقیقاتی، مرکزمنطقه‌ای اطلاع رسانی علوم و فناوری.

- دهخدا، علامه،" لغت نامه دهخدا"(۱۳۷۳)، انتشارات موسسه لغت نامه دهخدا ،دانشگاه تهران.

- رحیمیان، جلال و رزمجو، سید آیت اله (۱۳۸۹). مبانی زبانشناسی ۱. مرکز نشر دانشگاه شیراز، شیراز.

- سمیعی، احمد (۱۳۶۷). واژه‌های فریبکار، ناشناسهای آشنا. نشر دانش، بهمن و اسفند ۱۳۶۷، شمارهٔ دوم، سال نهم، شمارهٔ مسلسل ۵۰.

- صادقی، علی اشرف ؛ زندی مقدم، زهرا؛ (۱۳۸۵). فرهنگ املایی خط فارسی (براساس دستور خط فارسی مصوب فرهنگستان زبان و ادب فارسی) نسخهٔ رقومی. فرهنگستان زبان و ادب فارسی، تهران.

- صفوی، کورش؛ (۱۳۹۲) " در آمدی بر معناشناسی" انتشارات پژوهشکدهٔ فرهنگ و هنر اسلامی، تهران.

- عاصی ، م (۱۳۸۳). پردازش دستوری زبان فارسی با رایانه، دستور، ویژه نامهٔ فرهنگستان، شمارهٔ ۱. تهران.

- عمید، حسن، (۱۳۸۰)" فرهنگ فارسی عمید"انتشارات امیر کبیر ، تهران ۱۳۸۰

- فرهنگستان زبان و ادب فارسی، (۱۳۸۱)" دستور خط فارسی"، خرداد ۱۳۸۱

- مدرسی، فاطمه؛ (۱۳۸۷) ؛ "از واج تا جمله" انتشارات نشر چاپار چاپ دوم؛ تهران

- معین، محمد؛ (۱۳۷۱) ؛ فرهنگ فارسی معین (فرهنگ لغت ۶ جلدی)؛ انتشارات امیر کبیر. تهران.

- همایی، جلال الدین (۱۳۸۳). فنون بلاغت و صناعات ادبی. انتشارات نشر هما. تهران ۱۳۸۳.

Damerau. F. J and. Mays. E. (1989). “An examination of undetected. Typing errors.” Inf process. Manag, vol 25 , no 6.

Dastgheib, Mohammad Bagher . Fakhrahmad, seyed Mostafa. Zolghadri jahromi, Mansoor (2016) Perspell: A New Persian Semantic Based Spelling Correction System . Digital Scholarship in the HUMANITIES Advance Access Published March27,2016.

Faili, H., Ehsan, N ,. Montazery,.M,. Pilehvarm. M. T (2014).Vafa Spell – checker for detecting spelling, grammatical, and real – word errors of Persian language . literary and linguistic computing, fgu 043.

Francis, W N and H. Kucera (1964) Manual of Information to Accompany
Standard Sample of Present-Day Edited American English, for Use with Digital Computers" (revised 1979), Providence, RI: Department of Linguistics, BrownUniversity.

Kashefi, o, Sharifi, M., Minaie, B. (2012). A nove 1 string distance metric for ranking Persian respelling suggestions . natural language Engineering, 19(02)

Kennedy, Graeme (1998) **An Introduction to Corpus Linguistics**, London: Longman.

Leech, Geoffry (1992) "Corpora and theories of linguistic performance" in: Svartvik (1992): 105-122.

Mitton, R (1987). “Spelling chechers, Spelling correctors and the misspellings of poor Spellers” inf. Process. Manag, vol. 23, no 5.

Palmer, F.R. (1976). Semantics: a new outline, Cambridge: Cambridge University press. Pollock. j. j and A.Zamora, (1984). Automatic. Spelling correction in Scientific and scholarly. Text” commun. ACM. Vol, 27, no 4.

Pollock. j. j and A.Zamora, (1984). Automatic. Spelling correction in Scientific and scholarly. Text” commun. ACM. Vol, 27, no 4.

Richards. J. C. , platt, J and platt.H (1992). Longman dictionary of Language teaching & applied Linguistics . Essex: Longman.

Richards. J. C. and Schmidt, R. (2002). Longman dictionary of Language teaching & applied linguistics: Essex: Longman.

Svartvik, Jan (ed.) (1992) Directions in Corpus Linguistics, Berlin: Mouton de Gruyter.

Teubert, Wolfgang (1999) "Corpus linguistics: A partisan view" in: International Journal of Corpus Linguistics, Vol.4, No.1.

Thomas, J. (1995). Meaning in interaction: an introduction to pragmatics. New York: longman Group.

Yule, G. (2006). The study of language. Cambridge: Cambridge University Press.