KNO-1302-5204

تولید و کاربرد آزمایشی پیکره آموزشی ماشین‌خوان از الگوهای خطاهای دستوری رایج در انگلیسی: پیوند تحلیل خطا، زبان‌شناسی پیکره‌ای و تحلیل مقابله‌ای کاربردی

محمدرضا فلاحتی قدیمی فومنی، دانشیار گروه پژوهشی زبان‌شناسی رایانشی، مؤسسه استنادی و پایش علم و فناوری جهان اسلام (ISC)، شیراز، ایران، mrfalahati81@gmail.com

چکیده:

از داده‌های خطای ماشین‌خوان می‌توان به‌عنوان منبعی ارزشمند در آموزش زبان استفاده کرد. در این پژوهش دو پیکره ماشین‌خوان تولید شد: ۱) پیکره الگوهای خطاهای دستوری رایج انگلیسی و ۲) پیکره موازی جملات انگلیسی حاوی نکات دستوری به‌همراه ترجمه فارسی آن‌ها. داده‌ها از چاپ هفتم کتاب Common Mistakes in English استخراج و در نهایت ۶۵۸ جمله به‌دست آمد. برای هر رکورد، جمله صحیح، ترجمه فارسی، صورت سؤال، گزینه‌ها، کلید پاسخ و برچسب خطا ثبت شد. خطاها در ۵ بخش اصلی و ۲۰ زیربخش دسته‌بندی شدند. از این مجموعه، آزمون ۱۰۹ سؤالی متناسب با حجم زیربخش‌ها تهیه و در پاییز ۱۴۰۴ روی ۱۲۰ دانشجوی رشته زبان انگلیسی در سه دانشگاه شیراز اجرا شد. محصول نهایی، دو منبع ماشین‌خوان قابل استفاده در طراحی آزمون و تحلیل خطاست. محدودیت‌های پژوهش شامل نمونه‌گیری دردسترس، اتکا به یک منبع خاص و نبود داده‌های طبیعی تولیدشده توسط زبان‌آموزان است.

کلمات کلیدی: زبان‌شناسی پیکره‌ای، تحلیل خطا، پیکره آموزشی، پیکره موازی، سؤالات چندگزینه‌ای، خطاهای دستوری.

مقدمه

بروز خطا در تولیدات زبانی زبان‌آموزان جزئی لاینفک از فرایند یادگیری زبانی محسوب می‌شود. از این منظر، بروز خطا در تولیدات زبانی زبان‌آموزان یک عامل منفی تلقی نشده و برعکس می‌تواند اطلاعات متعددی را درخصوص مراحل یادگیری زبان و نیز دشواری‌هایی که زبان‌آموزان با آن مواجه هستند در اختیار محقق قرار دهد [1, 2]. از سویی دیگر، حوزه زبانشناسی مقابله‌ای و یا دقیق‌تر تحلیل کاربردی مقابله‌ای درصدد است تا از طریق بررسی شباهت‌ها و نیز تفاوت‌های موجود در بین زبان‌های مختلف، دشواری‌های احتمالی را که زبان‌آموزان در طی فرایند یادگیری زبان با آن مواجه می‌شوند، توضیح دهد[3, 4] .

با پیدایش زبان‌شناسی پیکره‌ای، مطالعه خطا به جای تمرکز بر داده‌های محدود، بر روی داده‌های بزرگ متمرکز شده است. البته، در این تحول صرفاً بر روی بزرگ‌تر کردن حجم داده تمرکز نشده، بلکه تلاش می‌شود تا به سمت شناسایی الگوهای جامعه‌تر و برچسب‌های منسجم‌تر و نیز کاربردی‌تر کردن پیکره‌های تولید شده حرکت شود [5, 6, 7]. از این منظر هدف پژوهش حاضر نیز صرفاً گردآوری و توصیف انواع خطاهای دستوری نبوده بلکه هدف اصلی آن بوده تا خطاهای دستوری شناسایی‌شده در قالب یک منبع و پیکره آموزشی کاربردی، ساختاریافته و ماشین‌خوان تنظیم و ارائه شود. در اینجا به جای پیکره زبان‌آموز از لفظ پیکره آموزشی استفاده به عمل آمد. دلیل این امر آن بوده که منبع داده‌های مورد استفاده در این پژوهش، داده‌های طبیعی زبان‌آموزان نبوده بلکه از جملات موجود در کتاب Common Mistakes in English نوشته ت. ج. فیتیکیدز استفاده شده است [8]. از این رو منظور از پیکره در این پژوهش، پیکره داده‌محور بوده و نه یک پیکره زبان‌آموز برگرفته از تولیدات زبانی واقعی زبان‌آموزان.

در توجیه استفاده از کتاب Common Mistakes in English نیز قابل ذکر است که این کتاب یکی از کتاب‌های کلاسیک در حوزه خطاهای رایج زبان‌آموزان زبان انگلیسی به عنوان یک زبان خارجی در دنیا محسوب می‌شود. در این کتاب تعداد زیادی از خطاهای دستوری در قالب صورت های درست (Say) و نادرست (Don’t Say) مورد بررسی قرار گرفته است. خطاها در این کتاب در قالب ۵ بخش کلی ارائه می‌شوند که عبارتند از: ۱) استفاده از صورت‌های نادرست، ۲) حذف نادرست، ۳) واژگان حشو، ۴) جایگاه دستوری نامناسب و ۵) به‌جای‌هم گرفتن واژه‌های مشابه. وجود این دسته‌بندی مشخص بین صورت‌های درست و نادرست و طیف گسترده نکات دستوری موجود در این کتاب آن را به منبع داده‌ای معتبر برای انجام پژوهش حاضر تبدیل می‌کند.

بر این اساس در پژوهش حاضر دو هدف کلی دنبال می‌شود: ۱) طراحی و مستندسازی یک پیکره آموزشی ماشین‌خوان از ۶۵۸ جمله حاوی نکات مختلف دستوری و سؤالات مستخرج از هر جمله و ۲) تولید یک پیکره موازی انگلیسی-فارسی از همان جمله‌ها. برای نشان دادن کاربردی بودن این پیکره، آزمونی حاوی ۱۰۹ سؤال (برگرفته از ۵ بخش کتاب و ۶۵۸ سؤال) بر روی ۱۲۰ دانشجوی سال سوم و چهارم رشته‌های زبان انگلیسی در سه دانشگاه شیراز، آزاد اسلامی واحد شیراز و مؤسسه آموزش عالی زند شیراز اجرا شد. در ادامه فرایند ساخت پیکره‌ها، ساختارداده، الگوی برچسب‌دهی، شیوه تولید گزینه‌های هر سؤال، قابلیت‌های استفاده و محدودیت‌های تعمیم توضیح داده خواهد شد.

پیشینه پژوهش

در هر نوع بحث تحلیل خطا ابتدا لازم است تا بین دو واژه «خطا» و «اشتباه» تفاوت قایل شویم. خطا به طور کلی نوعی انحراف نظام‌مند است که به توانش زبانی مربوط می‌شود، درحالی که اشتباه به کنش زبانی مربوط می‌شود و غالباً بر اثر عواملی نظیر لغزش زبانی، بی دقتی، خستگی و نظیر آن اتفاق می‌افتد [2]. در [1, 9, 10] نیز تحلیل خطا به عنوان مجموعه‌ای از فرایندهای تشخیص، توصیف، دسته‌بندی صورت‌های نامتعارف زبانی تعریف شده است. همچنین [11] مشخص بودن واحد تحلیل و معیار تشخیص خطا را برای هر نوع تحلیل خطا ضروری می‌داند.

خطاها را می‌توان به شکل‌های مختلف دسته‌بندی کرد. برای نمونه، بر اساس شکل ظاهری خطاها به حذف، افزایش، جابجایی و ترتیب واژه تقسیم‌بندی می‌شوند. همچنین می آنها را بر اساس واحد زبانی به خطای واجی، واژه‌ای و نحوی تقسیم‌بندی کرد یا بر اساس منشأ اصلی خطا [2, 12]. در پژوهش حاضر، به جای تمرکز بر منشأ خطا بر روی نوع خطا تمرکز شده است، چرا که برای تعیین علت و منشأ خطا به وجود داده‌های زبانی مستقل و طبیعی، سوابق زبانی زبان‌آموز و موقعیت تولید زبان نیاز خواهیم داشت که در پژوهش حاضر مدنظر نبوده است.

طبق نظر [5] در زبان‌شناسی پیکره‌ای به داده‌های زبانی به‌عنوان داده‌هایی ساختاریافته و قابل پردازش با استفاده از رایانه نگریسته می‌شود. در این حوزه پژوهش‌های متعددی به انجام رسیده است که به اختصار به موارد زیر اشاره می‌شود. در [13] برای مطالعه قدرت تشخیص عبارت در دانشجویان زبان انگلیسی از چارچوبی پیکره‌محور استفاده شد. به همین ترتیب در پژوهش‌های [14, 15, 16] نیز خطاهای نوشتاری و دستوری با استفاده از رویکردی پیکره‌محور توصیف و بررسی شد. به عنوان یک نمونه دیگر، [17] برای ساخت داده‌های مرتبط با خطاهای دستوری، چارچوبی را پیشنهاد داد. پژوهش‌های بالا و نمونه‌های دیگر از این دست همگی مؤید این نکته‌اند که در پژوهش‌های پیکره‌محور تولید داده‌های ساخت‌یافته و قابل پردازش اولویتی اساسی به حساب می‌آید.

در ایران نیز تحقیقات پیکره‌محور بسیاری به انجام رسیده است. برای نمونه نویسندگان در [18] به تشریح ضرورت ایجاد یک پیکره زبان‌آموز فارسی و برچسب‌گذاری خطاها برای اهداف آموزشی پرداختند. همچنین [19, 20, 21, 22] پژوهش‌های خود را در حوزه پیکره‌های موازی انجام داده و آنها را در حوزه‌های مختلف نظیر بازیابی اطلاعات بین‌زبانی، ماشین ترجمه و ... دارای کاربرد دانستند.

دو حوزه تحقیقاتی تحلیل مقابله‌ای و تحلیل خطا به هم پیوسته‌اند. در حقیقت، در تحلیل مقابله‌ای نظام‌های زبانی مختلف با هم مقایسه می‌شوند، در حالی که در حوزه تحلیل خطا، هدف استخراج و توصیف خطاهایی است که در تولیدات زبانی بروز پیدا می‌کند. همچنین در حوزه زبان‌شناسی پیکره‌ای نیز امکان جستجو، برچسب‌دهی و بازیابی اطلاعات موجود در پیکره فراهم می‌گردد. در این مقاله، بررسی خطاها به بخش تحلیل خطا، برچسب‌زنی نوع خطا به بخش زبان‌شناسی پیکره‌ای و استفاده آموزشی از داده‌های این پژوهش به بخش تحلیل مقابله‌ای کاربردی مربوط می‌شود. نکته قابل ذکر در این قسمت آن است که هدف پژوهش حاض بررسی ریشه خطاها نبوده و صرفاً نوع خطای دستوری را پوشش می‌دهد. از این رو برای بررسی تأثیر زبان اول یا دوم به داده‌های جدید نیاز خواهد بود.

از آنجا که در این پژوهش بر اساس جملات کتاب و سؤال‌های مستخرج از آن یک آزمون ۱۰۹ سؤاله چندگزینه تولید گردید در این قسمت در خصوص نحوه طراحی آن توضیحاتی آورده می‌شود. یک سؤال چند گزینه‌ای بخش‌های مختلفی دارد که عبارتند از خود سؤال، گزینه‌ها، پاسخ درست، گزینه‌های انحرافی و ... . در [23] بیان می‌شود که در طراحی سؤال و آزمون چندگزینه‌ای ضرورت دارد تا به ارتباط بین سؤال با موضوع و هدف آزمون توجه شود. همچنین لازم است تا از راه‌کارهایی برای پنهان ماندن گزینه درست استفاده شود به گونه‌ای که زبان‌موز نتواند پاسخ درست را پیش‌بینی کند. مثلاً محقق می‌تواند پاسخ سؤالات را بین گزینه‌های مختلف توزیع کند،

همچنانکه می‌تواند طول گزینه‌ها را ثابت یا نزدیک به هم نگه دارد و نظیر آن. برای رعایت این نکات محقق از یک دستورالعمل به شرح زیر استفاده کرد: در گام نخست، محقق از محتوای ارائه شده در هر مدخل و نیز تجربه مدرس (بیش از ۳۰ سال تجربه تدریس دستور زبان انگلیسی) برای تولید گزینه‌ها استفاده کرد. سپس این شش مرحله انجام شد: ۱) ویژگی دستوری هدف در سؤال به همراه کد خطا تعیین شد؛ ۲) از صورت صحیح "Say" به عنوان کلید سؤال استفاده شد؛ ۳) از آن صورت دستوری که در قسمت "Don't Say" آمده بود، به عنوان یکی از گزینه‌های انحرافی استفاده شد؛ ۴) گزینه‌های انحرافی دیگر با رعایت موضوع سؤال و ارتباط منطقی گزینه انحرافی با سؤال تولید شد؛ ۵) گزینه‌های هر سؤال برای صحت ارتباط و نیز منحصر به فرد بودن گزینه درست (وجود تنها یک پاسخ صحیح) بررسی شد و سرانجام ۶) گزینه پاسخ صحیح در سؤال‌های مختلف جابجا شد تا گزینه درست قابل پیش‌بینی نباشد. در پایان در آن دسته از سؤالات که هدف صرفاً قدرت تمیز دو واژه نزدیک به هم مطرح بود (مانند lay/lie یا raise/rise) قالب دوگزینه‌ای حفظ شد و گزینه‌های سوم و چهارم اضافه نشد تا تمرکز بر قدرت تشخیص واژه‌های نزدیک به هم حفظ شود. همچنین لازم به ذکر است که تمامی گزینه‌های هر سؤال توسط خود محقق ساخته و بازبینی شد و چون برای این آزمون از داوری مستقل چند خبره استفاده نشده، این آزمون را باید صرفاً یک منبع آموزشی ساخت‌یافته تلقی کرد که در صورت استفاده در آزمون‌های دیگر نیاز به اعتبارسنجی دارد.

مرور پیشینه حاضر نشان می‌دهد کهبا وجود مطالعات بسیار در حوزه تحلیل خطا، پژوهش‌هایی که به ایجاد منابع آموزشی خطامحور و در قالبی ماشین‌خوان انگلیسی-فارسی پرداخته باشند محدود بوده و محقق از آنها بی‌اطلاع است. در بخش پیکره‌ای این پژوهش هم علاوه بر ساخت پیکره، محقق کوشیده تا در خصوص منطق برچسب‌دهی و طراحی گزینه‌های انحرافی نیز توضیحاتی را ارائه کند تا از این طریق بتواند یک منبع حاوی سؤال و خطامحور را به یک پیکره آموزشی ماشین‌خوان و پیکره موازی تبدیل کند.

اهداف و سؤالات پژوهش

هدف اصلی در پژوهش حاضر آن بوده تا فرایند ساخت و مستندسازی یک منبع ماشین‌خوان از خطاهای دستوری را برای کاربردهای آموزشی تشریح کند. بر این اساس، سؤالات زیر مطرح شد:

1) **فرایند تبدیل مدخل‌های یک منبع آموزشی خطامحور به رکوردهای ماشین‌خوان دارای جملات صحیح دستوری، سؤال، گزینه‌ها، کلید و برچسب‌های خطا به چه صورت است؟**
2) **فرایند مستندسازی گزینه‌های انحرافی در سؤالات آزمون به چه صورت است؟**
3) **ساختار پیکره دستوری تولید شده (حاوی پنج بخش و بیست زیربخش) به چه صورت است؟**
4) **فرایند تولید یک پیکره موازی از جملات انگلیسی منبع آموزشی مورد استفاده و ترجمه این جملات به فارسی به چه صورت است؟**
5) **پیاده‌سازی آزمایشی آزمون ۱۰۹ سؤاله در این پژوهش در میاندانشجویان دانشگاه شیراز، آزاد اسلامی واحد شیراز و مؤسسه آموزش عالی زند شیراز از نظر امکان اجرا و زمان پاسخگویی چه ویژگی‌هایی داشته است؟**

روش پژوهش

این پژوهش را می‌توان در دسته پژوهش‌های مرتبط با توسعه یک منبع آموزشی دسته‌بندی کرد. محصول این پژوهش یک پیکره آموزشی ماشین خوان از خطاهای دستوری دانشجویان زبان انگلیسی سه دانشگاه در شهر شیراز و نیز یک پیکره موازی انگلیسی-فارسی از جملات صحیح انگلیسی و ترجمه آنها به فارسی بوده است. ضمناً اطلاعات خام این پیکره از کتاب خطاهای رایج در زبان انگلیسی گرفته شده است و بنابراین از داده‌های طبیعی استفاده نشده است. از این رو نباید اطلاعات آماری و بسامدی گزارش شده در این پژوهش را منعکس کننده حقایق طبیعی زبانی ارزیابی کنیم. در حقیقت، اطلاعات و آمار بسامدی صرفاً نشان‌دهنده محتوای یک کتاب است که

می تواند با اطلاعات کتاب‌های دیگر در این زمینه متفاوت باشد. همچنین به دلیل استفاده از آزمودنی‌های در دسترس، نتایج گزارش شده در این پژوهش الزاماً قابل تعمیم به کل جامعه آماری دانشجویان زبان انگلیسی در کشور نیست.

منبع استخراج داده‌ها

داده‌ها از چاپ هفتم کتاب Common Mistakes in English تألیف فیتیکیدز استخراج شد. این کتاب حاوی ۵۸۴ مدخل بود که با احتساب زیرمدخل‌ها و مدخل‌های دارای بیش از یک نکته، در نهایت ۶۵۸ جمله به‌دست آمد. نکته‌های متعدد در برخی مدخل‌ها با پسوند a، b، c و نظیر آن از هم متمایز شدند. برای هر رکورد، اطلاعاتی نظیر شناسه، جمله صحیح انگلیسی، ترجمه هر جمله انگلیسی به فارسی، سؤال مستخرج از هر جمله (صورت سؤال)، تعداد گزینه‌ها، کلید پاسخ، گزینه‌ها و برچسب کلی و جزئی خطا ثبت شد. خطاها در ۵ بخش اصلی و ۲۰ زیربخش تنظیم شدند.

در این کتاب هر مدخل بخش معرفی نکته دستوری، نمونه نادرست (که با عبارت انگلیسی "Don't Say" مشخص می‌شود)، نمونه صحیح (که با عبارت انگلیسی "Say" مشخص می‌شود) و در برخی مدخل‌ها توضیح تکمیلی "Notes" را شامل می‌شود. از جملات صحیح انگلیسی به عنوان داده برای تکمیل ستون جمله انگلیسی در پیکره و نیز استخراج کلید (گزینه صحیح) استفاده شد. برای مثال مدخل مربوط به حرف اضافه، تقابل "Don't Say: She accused the man for stealing." و "She accused the man of stealing." موضوع دستوری هدف (تقابل for با of) را مشخص می‌سازد.

ساختار داده و فرایند برچسب‌دهی

ساختار فیلدهای پیکره آموزشی ماشین‌خوان در جدول زیر آمده است.

جدول ۱. توضیح ساختار فیلدهای موجود در پیکره آموزشی ماشین‌خوان

ستون	نام فیلد	محتوا
A	شناسه	شماره یکتا برای هر جمله یا سؤال. ضمناً شماره یکتا در مدخل‌های حاوی نکات متعدد با پسون حروف انگلیسی همراه می‌شود مانند 2a, 2b, …
B	ترجمه فارسی	ترجمه فارسی ابتدا توسط ماشین ترجمه گوگل تولید می‌شود و سپس محقق به عنوان یک متخصص زبان برای رفع خطاهای دستوری، املایی و ... ترجمه هر جمله را بازبینی می‌کند.
C	جمله انگلیسی	صورت صحیح استخراج شده از بخش "Say" کتاب
D	صورت سؤال	جمله حاوی سؤال یا جای خالی مانند "Which sentence is correct?"
E	تعداد گزینه‌ها	۲ یا ۴ گزینه متناسب با موضوع هدف در سؤال‌های مختلف
F	کلید پاسخ	شماره گزینه صحیح
G-J	گزینه‌ها	گزینه‌های ۱ تا ۴ و در سؤال‌های دوگزینه‌ای، گزینه‌های ۱ و ۲
K	برچسب کلی خطا	کد یکی از پنج بخش اصلی خطا
L	برچسب جزئی خطا	کد یکی از زیربخش‌های خطا

در این پژوهش فرایند برچسب‌دهی بر اساس پنج بخش اصلی کتاب Common Mistakes in English انجام شد. پنج بخش موضوعی این کتاب عبارت بودند از: ۱) استفاده از صورت‌های نادرست، ۲) حذف نادرست، ۳) واژگان حشو، ۴) به کار گیری واژه در موقعیت دستوری نامناسب و ۵) خطا در استفاده از واژه‌های مشابه. برای این پنج بخش اصلی بیست زیربخش نیز تعریف شد. به بخش‌های

اصلی کدهای ۱۰، ۲۰، ۳۰، ۴۰ و ۵۰ تعلق گرفت و برای زیربخش‌ها اعدادی مانند ۱۱، ۲۵، ۳۴ و ... درنظر گرفته شد. برای نمونه عدد ۲۵ نشان می‌دهد که خطا مربوط به بخش اصلی ۲ و زیربخش ۵ بوده است. تنها در یک مورد (در زیربخش مربوط به استفاده نادرست از مصدر با to پنج بخش فرعی ذیل این زیربخش وجود داشت که با اعداد ۱۲۱ تا ۱۲۵ نام‌گذاری شدند. برای نمونه عدد ۱۲۱ نشان می دهد که خطا مربوط به بخش اصلی ۱ و زیربخش ۲ و در آنجا زیربخش ۵ از زیربخش ۲ بوده است.

پروتکل تولید سؤال و گزینه‌های انحرافی

برای تولید سؤال و گزینه‌های انحرافی مراحل زیر طی شد: ۱) تعیین یک نکته دستوری و کد خطای متناظر آن به عنوان یک سازه هدف، ۲) تعیین کلید (پاسخ صحیح). این مورد از قسمت "Say" در کتاب گرفته شد، ۳) تولید صورت سؤال. برای تولید سؤال در بیشتر موارد موضوع سؤال از جمله حذف و به جای آن نقطه‌چین گذاشته شد. همچنین در برخی موارد صورت سؤال به شکل "Which sentence is correct?" تولید گردید که دلیل این امر ماهیت و ساختار متفاوت این نوع از سؤال‌ها بود، ۴) تولید گزینه انحرافی اصلی. این گزینه از بخش "Don't Say" کتاب گرفته شد. لازم به ذکر است این گزینه ترکیبی است که زبان‌آموزان به احتمال زیاد آن را به‌جای گزینه صحیح برمی‌گزینند، ۵) تولید گزینه انحرافی فرعی. علاوه بر گزینه درست و گزینه انحرافی اصلی دو گزینه دیگر نیز برای سؤال چهارگزینه‌ای مورد نیاز است. این موارد بر اساس تجربه ۳۰ ساله محقق در تدریس دستور زبان انگلیسی و با استفاده از گزینه‌های خطای محتمل تولید شد. به عبارت شفاف‌تر این گزینه‌ها در حوزه پاسخ صحیح و مرتبط با آن تولید شدند، ۶) وارسی مجدد سؤالات برای اطمینان از تک‌کلیدی بودن سؤالات و یکسان نبودن گزینه صحیح در سؤالات مختلف. در این قسمت محقق به بررسی مجدد تمام گزینه‌ها پرداخت تا اولاً اطمینان یابد که برای هر سؤال بیش از یک پاسخ صحیح وجود ندارد و ثانیاً کلید همه سؤالات یکسان نباشد و اگر چنین بود محقق جای گزینه‌ها را به شکلی تغییر داد تا کلید سؤالات مختلف متفاوت باشد و الگو یا نظمی که زبان‌آموز را به سمت پاسخ هدایت کند، وجود نداشته باشد. البته لازم به ذکر است که چون گزینه‌ها توسط محقق انتخاب شده است، روایی محتوایی برای آزمون محاسبه نشده است که این نکته به‌عنوان یکی از محدودیت‌های پژوهش گزارش شده است.

تولید آزمون ۱۰۹ سؤالی

در این پژوهش با بهره‌گیری از ۶۵۸ سؤال، و با استفاده از روش نمونه‌گیری تصادفی چندمرحله‌ای یک آزمون حاوی ۱۰۹ سؤال ساخته شد. علت استفاده از این روش نمونه‌گیری آن بوده است که در کتاب پنج بخش اصلی و بیست زیربخش وجود داشت و به دلیل تفاوت این بخش‌ها و همچنین تفاوت در حجم سؤالات هر بخش، استفاده از روش نمونه‌گیری تصادفی چندمرحله‌ای روشی مناسب محسوب می‌شد. در این آزمون از نظر گزینه‌ها دو دسته سؤال وجود داشت. دسته اول سؤالات چهارگزینه‌ای بودند که اتفاقاً بیشترین تعداد سؤالات آزمون را تشکیل می‌دادند. این سؤالات عبارت بودند از: سؤالات ۱ تا ۶۸، ۸۳ تا ۹۱ و ۱۰۴ تا ۱۰۹. دسته دوم سؤالات دوگزینه‌ای بودند. تعداد این دسته از سؤالات کمتر بود و سؤال‌های ۶۹ تا ۸۲ و ۹۲ تا ۱۰۳ را تشکیل می‌دادند.ضمناً نحوه نمره‌دهی به برگه‌های زبان‌آموزان بدیت صورت بود که به ازای هر پاسخ صحیح نمره یک درنظر گرفته شد. به هر پاسخ اشتباه هم نمره صفر تعلق گرفت. با توجه به وجود ۱۰۹ سؤال، حداقل و حداکثر نمره قابل اخذ صفر تا ۱۰۹ بود. در جدول شماره ۲ تعداد سؤالات موجود در هر زیربخش از سؤالات به همراه کد شناسایی هر زیربخش آورده شده است.

جدول ۲. توزیع سؤالات آزمون ۱۰۹ سؤالی در ۲۰ زیربخش

کد سؤال	عنوان زیربخش	تعداد سؤال
۱-۱	استفاده از حرف اضافه نادرست	۱۳
۱-۲	استفاده نادرست از مصدر با to	۵
۱-۳	استفاده نادرست از زمان فعل	۵
۱-۴	مثال‌های متفرقه بخش ۱	۹
۱-۵	عبارات نامعتبر در انگلیسی	۶

۱-۲	حذف حرف اضافه	۳
۲-۲	مثال‌های متفرقه بخش ۲	۷
۱-۳	حرف اضافه زائد	۳
۲-۳	حرف تعریف زائد	۳
۳-۳	مصدر با to غیرضروری	۲
۴-۳	مثال‌های متفرقه بخش ۳	۴
۱-۴	موقعیت نامناسب قید	۱
۲-۴	مثال‌های متفرقه بخش ۴	۳
۱-۵	به‌جای‌هم‌گرفتن حروف اضافه	۴
۲-۵	به‌جای‌هم‌گرفتن افعال	۱۵
۳-۵	به‌جای‌هم‌گرفتن قیود	۲
۴-۵	به‌جای‌هم‌گرفتن صفات	۶
۵-۵	به‌جای‌هم‌گرفتن اسامی	۵
۶-۵	عدم رعایت مطابقت عددی	۷
۷-۵	استفاده از مقوله نحوی/گفتاری نادرست	۶

روش انتخاب زبان‌آموزان

برای اجرای آزمون ابتدا نیاز بود تا زبان‌آموزان انتخاب شوند. برای این کار از روش نمونه‌گیری در دسترس استفاده شده و در مجموع ۱۲۰ دانشجوی سال سوم و چهارم رشته‌های زبان انگلیسی از دانشگاه شیراز، دانشگاه آزاد اسلامی واحد شیراز و مؤسسه آموزش عالی غیرانتفاعی غیردولتی زند شیراز انتخاب شدند. از سه دانشگاه تعداد برابر از زبان‌آموزان انتخاب شدند (۴۰ زبان‌آموز از هر دانشگاه). علت انتخاب دانشجویان سال سوم و چهارم این بوده که تصور می‌رفت این دانشجویان تمامی درس‌های پایه دستوری را گذرانده باشند و در مرحله مشابهی از برنامه دوره کارشناسی قرار داشته باشند. این نکته ارزیابی بهتری از عملکرد دستوری زبان‌آموزان فراهم می‌کرد. همچنین ذکر این نکته ضروری است که افرادی که مایل به شرکت در آزمون نبودند از جریان پژوهش حذف شدند. محدوده سنی زبان‌آموزان بین ۲۱ تا ۲۴ سال بود و معدل دانشجویان نیز بین ۱۷ تا ۱۸,۷۸ متغیر بود.

با این حال باید توجه داشت که به دلیل نوع روش نمونه‌گیری (نمونه‌های در دسترس) یافته‌های پژوهش حاضر الزاماً قابل تعمیم به کل زبان‌آموزان رشته‌های زبان انگلیسی در کشور نیست و از این رو محقق ادعایی درباره تعمیم یافته‌ها به کل جامعه آماری پژوهش در کشور ندارد. بلکه برعکس نتایج بدست آمده عملکرد نمونه‌های مورد بررسی را منعکس می کند.

نحوه اجرای آزمون

برای این آزمون زبان‌آموزان هر دانشگاه در یک زمان معین (هفته سوم آبان ۱۴۰۴) در یک کلاس جمع شدند و آزمون ۱۰۹ سؤالی بین آنها توزیع شد. با توجه به تعداد سؤالات، زمان آزمون برای زبان‌آموزان ۸۵ دقیقه تعیین شد. با این حال چنانچه زبان‌آموزی به زمان بیشتری نیاز داشت، زمان در اختیار او قرار گرفت تا بتواند به همه سؤالات پاسخ بدهد و به عبارتی عدم ارائه پاسخ به دلیل کمبود زمان نبوده باشد. تنها یکی از زبان‌آموزان تقاضای زمان بیشتری داشت که به او این زمان داده شد و باقی زبان‌آموزان بین ۵۰ تا ۷۰ دقیقه آزمون را کامل کرده و برگه‌های خود را تحویل دادند. در فرایند تصحیح برگه‌ها نیز به هر پاسخ درست نمره ۱ و به هر پاسخ خطا نمره صفر تعلق گرفت.

مرحله تولید پیکره موازی انگلیسی-فارسی

همانگونه که در قسمت‌های قبل توضیح داده شد، از کتاب Commom Mistakes in English در مجموع ۶۵۸ جمله حاوی نکات دستوری مختلف استخراج شده بود. برای تولید پیکره موازی به صورت زیر عمل شد. ابتدا هر یک از این جملات انگلیسی به ماشین ترجمه گوگل داده شد تا به فارسی ترجمه شود. سپس، محقق ترجمه‌ها را از نظر دستوری، املایی، واژگانی و ... بازبینی کرد و اشکالات احتمالی را مرتفع نمود. بنابراین داده‌های فارسی (۶۵۸ جمله فارسی) را نباید به عنوان داده‌ای خام و صرفاً ماشینی تلقی کرد بلکه باید آن را ترجمه ماشینیِ بازبینی شده توسط محقق قلمداد نمود. در مرحله بعد، به هر جمله انگلیسی و ترجمه فارسی یک شناسه واحد تخصیص داده شد تا هم‌ترازی آنها در سطح جمله حفظ شده و بتوان از این ساختار هم‌تراز ماشین‌خوان در حوزه‌های مختلفی نظیر بازیابی اطلاعات بین‌زبانی، ترجمه ماشینی و نظیر آن استفاده کرد.

نتایج پژوهش

در این قسمت یافته‌های پژوهش تشریح می‌شود:

محصول اول پژوهش حاضر یک پیکره آموزشی با ۶۵۸ رکورد است. از این تعداد رکورد اطلاعاتی، ۵۰۳ مورد (۷۶,۴ درصد) چهار گزینه‌ای و ۱۵۵ مورد (۲۳,۶ درصد) دو گزینه‌ای بودند. ضمناً هر رکورد اطلاعاتی موارد زیر را شامل می‌شد: جمله صحیح به انگلیسی، سوال انگلیسی مستخرج از جمله انگلیسی، کلید سؤال، گزینه‌های انحرافی و برچسب‌های کلی و جزئی خطا. از این منظر، داده‌های پیکره بر اساس نوع خطا، شناسه و نوع سؤال (دو یا چهار گزینه‌ای) قابل بازیابی خواهد بود. جدول ۳ توزیع رکوردهای پیکره آموزشی تولید شده را بر اساس پنج بخش اصلی کتاب نشان می‌دهد.

جدول ۳. توزیع رکوردهای پیکره آموزشی تولید شده بر اساس پنج بخش اصلی کتاب

بخش	عنوان بخش	تعداد و درصد
۱	استفاده از صورت‌های نادرست	۲۳۳ (۳۵,۴ درصد)
۲	حذف نادرست	۶۱ (۹,۲ درصد)
۳	واژگان حشو	۶۹ (۱۰,۵ درصد)
۴	موقعیت دستوری نامناسب	۲۶ (۴ درصد)
۵	به‌جای‌هم‌گرفتن واژه‌های مشابه	۲۶۹ (۴۰,۹ درصد)

همانگونه که در جدول بالا نشان داده شده است، در این پیکره بیشترین سهم به جملات بخش ۵ (به‌جای‌هم‌گرفتن واژه‌های مشابه) داده شده است. پس از آن بخش‌های ۱ (استفاده از صورت‌های نادرست) و ۳ (واژگان حشو) در جایگاه دوم و سوم قرار می‌گیرند. البته باید توجه داشته باشیم که این نوع توزیع کاملاً به نظر نویسنده کتاب وابسته بوده و نمی‌توان بر اساس آن تعداد و فراوانی واقعی خطاها را در میان زبان‌آموزان در جامعه آماری مشخص کرد. در جدول ۴ توزیع رکوردهای پیکره آموزشی تولید شده بر اساس ۲۰ زیربخش کتاب نشان داده شده است.

جدول ۴. توزیع رکوردهای پیکره آموزشی تولید شده بر اساس بیست زیربخش کتاب

کد سؤال	عنوان زیربخش	تعداد/درصد سؤال
۱-۱	استفاده از حرف اضافه نادرست	۷۶ (۱۱,۶ درصد)
۱-۲	استفاده نادرست از مصدر با to	۳۱ (۴,۷ درصد)
۱-۳	استفاده نادرست از زمان فعل	۳۴ (۵,۲ درصد)
۱-۴	مثال‌های متفرقه بخش ۱	۵۵ (۸,۴ درصد)
۱-۵	عبارات نامعتبر در انگلیسی	۳۷ (۵,۶ درصد)

۱-۲	حذف حرف اضافه	۱۹ (۲,۹ درصد)
۲-۲	مثال‌های متفرقه بخش ۲	۴۲ (۶,۴ درصد)
۱-۳	حرف اضافه زائد	۱۷ (۲,۶ درصد)
۲-۳	حرف تعریف زائد	۲۰ (۳ درصد)
۳-۳	مصدر با to غیرضروری	۱۱ (۱,۷ درصد)
۴-۳	مثال‌های متفرقه بخش ۳	۲۱ (۳,۲ درصد)
۱-۴	موقعیت نامناسب قید	۷ (۱,۱ درصد)
۲-۴	مثال‌های متفرقه بخش ۴	۱۹ (۲,۹ درصد)
۱-۵	به‌جای‌هم‌گرفتن حروف اضافه	۲۵ (۳,۸ درصد)
۲-۵	به‌جای‌هم‌گرفتن افعال	۸۷ (۱۳,۲ درصد)
۳-۵	به‌جای‌هم‌گرفتن قیود	۱۴ (۲,۱ درصد)
۴-۵	به‌جای‌هم‌گرفتن صفات	۳۵ (۵,۳ درصد)
۵-۵	به‌جای‌هم‌گرفتن اسامی	۳۲ (۴,۹ درصد)
۶-۵	عدم رعایت مطابقت عددی	۴۳ (۶,۵ درصد)
۷-۵	استفاده از مقوله نحوی/گفتاری نادرست	۳۳ (۵ درصد)

همانگونه که در جدول بالا مشاهده می‌شود، از میان کل زیربخش‌های کتاب، زیربخش‌های « به‌جای‌هم‌گرفتن افعال»، «استفاده از حرف اضافه نادرست» و «مثال‌های متفرقه بخش ۱» به ترتیب با فراوانی ۸۷ (۱۳,۲ درصد)، ۷۶ (۱۱,۶ درصد) و ۵۵ (۸,۴ درصد) بیشترین تعداد از رکوردهای اطلاعاتی پیکره آموزشی تولید شده را به خود اختصاص داده‌اند.

محصول دوم پژوهش حاضر پیکره موازی تولید شده است. این پیکره موازی از ۶۵۸ جمله انگلیسی و برابرنهاده آن (بر اساس ترجمه ماشینی گوگل و بازبینی محقق) به زبان فارسی می‌باشد. هر جفت جمله (انگلیسی و معادل آن به فارسی) با یک شناسه واحد ثبت شده‌اند که می‌توان آن را به رکورد آموزشی اصلی شامل سؤالات و گزینه‌های مربوط به آن ربط داد. مزیت این کار آن است که چنانچه محقق یا کاربر نیاز داشته باشد می‌تواند یا صرفاً از اطلاعات پیکره موازی (جمله انگلیسی و معادل فارسی آن) استفاده کند یا از این اطلاعات در پیوند با اطلاعات دیگر موجود در پیکره نظیر نوع خطا و ... نیز بهره‌مند شود.

اجرای آزمایشی آزمون ۱۰۹ سؤالی

در مرحله اجرا، هر ۱۲۰ نفر به تمامی ۱۰۹ سؤال پاسخ گفتند. از نظر زمانی بالاترین زمان مصرف شده توسط زبان‌آموزان برای تمام کردن آزمون ۸۵ دقیقه بود (فقط یک نفر) و باقی برگه‌ها در زمانی بین ۵۰ تا ۷۰ دقیقه توسط زبان‌آموزان تکمیل شده و به محقق تحویل داده شدند. در این پژوهش تعیین دشواری سؤال از اهداف پژوهش نبوده و از این رو آزمون مورد استفاده به عنوان یک مجموعه داده برای مدرسان قابل استفاده است که می‌توانند با توزیع آن در میان زبان‌آموزان خود نسبت به بررسی معیارهایی چون دشواری سؤالات مطالعه نمایند.

بحث و نتیجه‌گیری

در این پژوهش، پیوند میان سه حوزه تحلیل خطا، زبان‌شناسی پیکره‌ای و تحلیل مقابله‌ای کاربردی ترسیم شد. در حقیقت عنوان شد که مبحث تحلیل خطا انحرافات تولیدی زبانی زبان‌آموزان را به تصویر می‌کشد. از طرفی زبان‌شناسی پیکره‌ای با ملزوماتی نظیر ثبت استاندارد رکوردهای اطلاعاتی و نیز برچسب‌دهی به عناصر زبانی، امکان بازیابی داده‌ها را به شکلی آسان فراهم آورد. در نهایت تحلیل

مقابله‌ای کاربردی می‌کوشد تا از طریق مقایسه و مقابله داده‌های زبانی در دو یا چند زبان، زمینه استفاده آنها را در محیط‌های آموزشی فراهم آورد.

نکته دیگر به ماهیت داده‌های موجود در پیکره آموزشی ماشین خوان تولید شده از خطاهای دستوری دانشجویان زبان انگلیسی سه دانشگاه در شهر شیراز مربوط می‌شود که مهم‌ترین دستاورد پژوهش حاضر نیز بوده است. در حقیقت، داده‌های دستوری مورد استفاده در این پژوهش از داده‌های طبیعی تولید شده توسط زبان‌آموزان استخراج نگردیده و از این رو پیکره تولید شده را باید یک پیکره آموزشیِ ساخته شده از الگوهای خطا به حساب آوریم آن هم الگوهایی که توسط نویسنده کتاب به مخاطب منعکس گردیده است [6, 7]. از این رو پیکره تولید شده برای طراحی آزمون توسط آموزگاران، اجرای تمرین‌های زبانی و نیز مطالعات روش‌شناختی مناسب است و نباید آن را منعکس کننده داده‌های واقعی زبان طبیعی به حساب آوریم.

نکته بعدی به پروتکل تولید سؤال و گزینه‌های انحرافی مربوط می‌شود. در حقیقت از طریق این پروتکل ۶ مرحله‌ای فرایند تولید سؤال و گزینه‌ها به انجام رسید. البته با توجه به اینکه درجه سختی و نیز روایی سؤالات محاسبه نگردیده نباید آن را دارای روایی مفهومی قلمداد کنیم بلکه باید صرفاً به آن به عنوان مخزنی از سؤالات دستوری نگریسته شود که بر اساس چارچوب فکری نویسنده کتاب منبع ساختاربندی شده است. در گرینه‌های سؤالات نیز گزینه انحرافی گرفته شده از بخش "Don't Say" از اهمیت بسیاری برخوردار است چرا که نویسنده کتاب آن را بر اساس خطاهای رایجی تنظیم کرده که در جامعه زبان‌آموزان خود با آن مواجه بوده است.

نکته دیگر به پیکره موازی تولید شده در این پژوهش باز می‌گردد که یکی از دستاوردهای اصلی این پژوهش بوده است. این نوع از پیکره‌ها در حوزه‌های مختلفی از جمله بازیابی اطلاعات بین زبانی کاربد بسیاری دارند. در این پژوهش، همترازی در سطح جمله تنظیم شد و به هر جفت جمله شناسه‌ای واحد تعلق گرفت تا در صورت لزوم بتوان این لایه را به بخش‌های پیکره اول تولید شده از ۶۵۸ جمله وصل کرد یا صرفاً و بسته به نیاز از هم‌ترازی جملات فارسی و انگلیسی استفاده کرد.

در نهایت اجرای آزمایشی آزمون بر روی دانشجویان سه دانشگاه در شیراز نشان داد که این آزمون از نظر اجرایی دارای قابلیت است. با این همه نوع نمونه‌گیری از زبان‌آموزان، عدم نمونه‌گیری از نقاط مختلف کشور و همچنین عدم همگون‌سازی کامل زبان‌آموزان سبب می‌گردد تا نتوانیم نتایج به دست آمده در این پژوهش را لزوماً به جامعه آماری زبان‌آموزان کشور ایران تعمیم دهیم.

کاربردها، محدودیت‌ها و جمع‌بندی

از یافته‌های پژوهش حاضر می‌توان در حوزه‌های مختلفی استفاده کرد که در ادامه به چند مورد از آنها اشاره می‌شود و سپس به بیان محدودیت‌های پژوهش پرداخته خواهد شد.

آموزگاران و مدرسان زبان می‌توانند از پیکره تولید شده برای تولید تمرین برای زبان‌آموزان استفاده نمایند. همچنین از این پیکره می‌توان برای طراحی و تولید آزمون استفاده کرد. البته برای ارزیابی‌های غیر تمرینی و جدی لازم خواهد بود که اعتبارسنجی محتوایی آزمون نیز به انجام برسد. کاربرد دیگر پیکره تولید شده آن است که می‌توان از آن برای بررسی و تحلیل خطاها استفاده کرد، هرچند منظور اینجا نوع خطا است و نه ریشه آن چرا که مورد اخیر جزء اهداف محقق حاضر نبوده است. در پایان می‌توان از پیکره موازی تولید شده نیز در آموزش ترجمه و مباحث بازیابی اطلاعات بین‌زبانی بهره‌برداری کرد.

با وجود دستاوردهای پژوهش حاضر، محدودیت‌هایی نیز وجود داشته است که باید مورد توجه قرار بگیرد. برای نمونه، نوع خطاها و فراوانی‌های گزارش شده تا حد زیادی تحت تأثیر محتوای منبع اصلی نکات دستوری بوده و نه موارد برگرفته از زبان طبیعی و بنابراین نباید این پیکره را یک پیکره زبان‌آموز شمرد، بلکه باید آن را صرفاً یک پیکره آموزشی به حساب آورد. روش نمونه گیری زبان‌آموزان نیز روش نمونه‌های در دسترس بوده که تعمیم نتایج به کل جامعه زبان‌آموزان کشور را ممکن نمی‌سازد.

فهرست منابع

[1] Corder, S. P. (1967). The significance of learners' errors. International Review of Applied Linguistics in Language Teaching, 5(4), 161–170.

[2] James, C. (1998). *Errors in language learning and use: Exploring error analysis*. New York: Longman.

[3] Keshavarz, M. H. (2010). *Contrastive analysis and error analysis*. Tehran: Rahnama.

[4] یارمحمدی، ل. ا. و رشیدی، ن. (۱۳۹۴). *زبان‌شناسی مقابله‌ای کاربردی در زبان انگلیسی و فارسی با تأکید بر دستور*. تهران: انتشارات راهنما.

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

[6] Callies, M. (2015). Learner corpus methodology. In S. Granger, G. Gilquin, & F. Meunier (Eds.), The Cambridge handbook of learner corpus research (pp. 35–56). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.003

[7] Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). The Cambridge handbook of learner corpus research. Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

[8] فیتیکیدز، ت. ج. (۱۳۷۰). *خطاهای رایج در انگلیسی*. تهران: انتشارات راهنما.

[9] Lennon, P. (1991). Error: Some problems of definition, identification and distinction. *Applied Linguistics*, *12*(2), 180-196.

[10] Botley, S. P. (2015). Errors versus mistakes: A false dichotomy? *Malaysian Journal of ELT Research*, *11*(1), 81-94.

[11] Ellis, R. & Burkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.

[12] Bussmann, H., Kazzazi, K., & Trauth, G. (1996). *Routledge dictionary of language and linguistics*. London: Routledge.

[13] Mala, M. (2020). Phraseology in learner academic English corpus-driven approaches. *Discourse and Interaction*, *13*(2), 75-88.

[14] Chen, X. (2024). Corpus-based study on language errors in English Writing. *International Journal of Education and Humanities*, *15*(2), 144-153. DOI: 10.54097/f2cmct16

[15] Gazioglu, M., & Salami, A. (2024). Identifying grammatical errors and mistakes via a written learner corpus in a foreign language context. *Journal of Language Research*, *8*(2), 91-106. https://doi.org/10.51726/jlr.1553484

[16] Amirsheibani, M. (2025). A corpus-based description of grammatical errors in Uzbek EFL learner writing. European *International Journal of Philological Sciences*, *5*(10), 65-70. https://doi.org/10.55640/eijps-05-10-14

[17] Lin, M., Zeng, M., Huang, W., Jiang, Sh., Xiao, L., & Yang, A. (2025). A simple yet effective corpus construction framework for Indonesian grammatical error correction. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *24*(1), 1-15. https://doi.org/10.1145/3704264

[18] صفری، س.، عاصی، م. و صحرایی، ر. (۱۳۹۱). طراحی و ایجاد پیکره تولیدی زبان‌آموز فارسی (پایان‌نامه کارشناسی ارشد، دانشگاه علامه طباطبائی).

[19] براتعلی‌پور، خ.، فیلی، ه. و شاکری، آ. (۱۳۹۱). تولید یک پیکره موازی فارسی-انگلیسی با استفاده از دادگان استخراج شده از وب (پایان‌نامه کارشناسی‌ارشد، دانشگاه تهران، پردیس دانشکده‌های فنی، دانشکده علوم مهندسی).

[20] کشتکار، ح.، موسوی میانگاه، ط. و غیاثیان، م. (۱۳۹۱). ساخت پیکره دوزبانه موازی انگلیسی-فارسی و کاربرد آن در سامانه حافظه ترجمه (مبحثی در زبانشناسی پیکره‌ای) (پایان‌نامه کارشناسی‌ارشد، دانشگاه پیام نور استان تهران، مرکز پیام نور تهران).

[21] آذربنیاد، ح.، شاکری، آ. و فیلی، ه. (۱۳۹۲). بازیابی اطلاعات بین‌زبانی فارسی-انگلیسی با استفاده از پیکره‌های موازی (پایان‌نامه کارشناسی‌ارشد، دانشگاه تهران، پردیس دانشکده‌های فنی، دانشکده علوم مهندسی).

[22] هاشمی، آ.، فیلی، ه. و شاکری، آ. (۱۳۹۴). ابهام‌زدایی معنایی کلمات با استفاده از پیکره‌های موازی (پایان‌نامه کارشناسی‌ارشد، دانشگاه تهران، پردیس دانشکده‌های فنی، دانشکده علوم مهندسی).

[23] Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education, 15(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5