Ferdows-Lex: A Lexical Corpus of Persian Language Teaching Materials for Teaching Non-Persian Learners

Zahra Alizadeh Memar
PhD Candidate in General Linguistics; Linguistics Department; Faculty of Letters and Humanities; Ferdowsi University of Mashhad; Mashhad, Iran Email: memar.linguistics@gmail.com

Atiyeh Kamyabi Gol*
PhD in Applied Linguistics; Associate Professor; Department of Linguistics; Faculty of Letters and Humanities; Ferdowsi University of Mashhad; Mashhad, Iran Email: kamyabigol@um.ac.ir

Shahla Sharifi
PhD in General Linguistics; Associate Professor; Department of Linguistics; Faculty of Letters and Humanities; Ferdowsi University of Mashhad; Mashhad, Iran Email: sh-sharifi@um.ac.ir

Amirmasoud Iravani
PhD Candidate in General Linguistics; Department of Linguistics; Faculty of Letters and Humanities; Ferdowsi University of Mashhad; Mashhad, Iran Email: a.iravani@mail.um.ac.ir

Iranian Journal of Information Processing and Management

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 40 | No. 4 | pp. 1179-1218
Summer 2025
https://doi.org/10.22034/jipm.2025.2054865.1948

Received: 03, Mar. 2025 **Accepted: 03, May 2025**

Abstract: The purpose of this study was to develop a corpus according to the vocabulary overlaps in the materials for Teaching Persian to Non-Persian Speakers (TPNPS) in the elementary, intermediate, and advanced levels. Computer tools and Corpus-Informed approaches using a three-step protocol were applied in this study. First, the research data was prepared. The data was selected from among 26 TPNPS textbooks. These included Parsa, Mina, Shiraz, Parfa, Amozash e Novin e Zaban e Farsi at three language proficiency levels. The total number of tokens in the research dataset was 15,585. The data was typed out, and then computational pre-processing and parts of speech (POS) tagging were carried out. Normalization was mainly performed using Dadmatools Package and tokenization, lemmatization and POS tagging (POS) were carried out through the standard STANZA package. Then, the vocabulary overlap range in all textbooks at each level and among all levels were analyzed using Python programming. Finally, the corpus was designed in the mark-up language of XML. The corpus had three proficiency levels each including vocabulary information like lemma, overlap range,

* Corresponding Author

alphabet, token, POS and metadata. The results showed, that the vocabulary overlapping range followed a fixed rate at first, decreased as the proficiency level increased i.e., this rate stood at about 36 percent and 36.5 percent in the elementary and intermediate levels whereas it declined to 13 percent at advanced levels. Furthermore, with regards to the POS analysis, nouns, verbs and adjectives were the most repeated ones in all three levels. Comparing the overlap of vocabulary among different levels (elementary to intermediate, intermediate to advanced, elementary and advanced), nouns had the highest share. The findings emphasized systematic development of teaching materials to gradual improvements of language skills.

Keywords: Teaching Persian to Non-Persian Speakers, Vocabulary Range Overlap, Lexical Corpus Construction, Level-Structured Lexical Corpus, Computational Toolkits

فردوس-لکس: پیکره واژگانی از منابع آموزشی فارسی برای غیرفارسی‌زبانان

زهرا علیزاده معمار
دانشجوی دکتری زبان‌شناسی همگانی؛ دانشگاه فردوسی مشهد؛ مشهد، ایران؛
memar.linguistics@gmail.com

عطیه کامیابی‌گل
دکتری زبان‌شناسی کاربردی؛ دانشیار؛ دانشگاه فردوسی مشهد؛ مشهد، ایران؛
پدیدآور رابط kamyabigol@um.ac.ir

شهلا شریفی
دکتری زبان‌شناسی همگانی؛ دانشیار؛ دانشگاه فردوسی مشهد؛ مشهد، ایران؛
sh-sharifi@um.ac.ir

امیرمسعود ایروانی
دانشجوی دکتری زبان‌شناسی همگانی؛ دانشگاه فردوسی مشهد؛ مشهد، ایران؛
a.iravani@mail.um.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۴۰۳/۱۲/۱۳ | پذیرش: ۱۴۰۴/۰۲/۱۴ | مقاله برای اصلاح به مدت ۱۵ روز نزد پدیدآوران بوده است.

چکیده: پژوهش حاضر با هدف تدوین پیکره‌ای مطابق با همپوشانی واژگانی در سطوح مبتدی، میانه و پیشرفتهٔ منابع آموزش زبان فارسی به غیرفارسی‌زبانان در ایران، با رویکردی پیکره‌آگاه و روشی رایانشی انجام شد. این مطالعه در سه مرحله صورت گرفت. در مرحله اول، دادگان پژوهش شامل ۱۴۵۸۵ واحد واژگانی از ۲۶ منبع آموزشی در سه سطح مبتدی، میانه و پیشرفته انتخاب شدند. این منابع شامل مجموعه کتاب‌های «پرفا»، «مینا»، «شیراز»، «پارسا»، «رسا»، «نگارا»، «خوانا» و «آموزش نوین زبان فارسی» بودند. سپس، تمامی واژگان به‌صورت دستی تایپ شده، به‌صورت رایانشی پیش‌پردازش شده، و برچسب اجزای کلام دریافت کردند. هنجارسازی به‌طور عمده با ابزار «دادماتولز» انجام پذیرفت. واحدسازی، بن‌واژه‌سازی و برچسب‌دهی اجزای کلام با استفاده از «استنزا» صورت پذیرفت. در مرحله دوم، با بهره‌گیری از برنامه‌نویسی «پایتون»، با کمک اجتماع و اشتراک بین مجموعه‌های واژگان هر کتاب، میزان همپوشانی واژگان در هر سطح و بین سطوح مختلف تعیین شد. در مرحله سوم، ماشین‌خوان کردن صورت پذیرفت؛ بدین‌صورت که یک پیکره با استاندارد نشانه‌گذاری XML توسعه داده شد که برای هر واژه در سطوح

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۴۰ | شمارهٔ ۴ | ص‌ص ۱۱۷۹-۱۲۱۸
تابستان ۱۴۰۴
https://doi.org/10.22034/jipm.2025.2054865.1948

مختلف دارای حرف الفبا، واحد، بن‌واژه، دامنه همپوشانی، برچسب اجزای کلام و فراداده کتاب‌های منبع آن است. نتایج پژوهش نشان داد که با افزایش سطح زبان‌آموزی، میزان همپوشانی واژگان روندی ثابت و سپس کاهشی دارد؛ به‌طوری که در سطوح مبتدی و میانه به‌ترتیب، در حدود ۳۶ درصد و ۳۶/۵ درصد همپوشانی مشاهده شد، اما در سطح پیشرفته این میزان به ۱۳ درصد کاهش یافته است. واکاوی اجزای کلام در هر سطح نشان داد که اسم، فعل و صفت پرتکرارترین اجزای کلام در تمامی سطوح هستند. همچنین، در مقایسه همپوشانی واژگان بین سطوح مختلف (مبتدی و میانه، میانه و پیشرفته، مبتدی و پیشرفته) مقوله اسم بالاترین سهم را داشت. این یافته‌ها بر طراحی نظام‌مند منابع آموزشی جهت پیشرفت تدریجی مهارت‌های زبانی تأکید دارند

کلیدواژه‌ها: پیکره سطح‌بندی‌شده واژگان، آموزش فارسی به غیرفارسی‌زبانان، همپوشانی واژگانی، ساخت پیکره واژگان، ابزارهای رایانشی

۱. مقدمه

امروزه در جهان، افراد بسیاری بنا بر نیازهای گوناگون اقتصادی، شغلی، تحصیلی و غیره مهاجرت می‌کنند. این امر سبب می‌شود که آن‌ها به‌منظور کسب توانایی در رفع نیازهای خود، ایجاد ارتباط با محیط پیرامون، و تبادلات فرهنگی نسبت به یادگیری زبان مقصد اقدام کنند. در کشور ما نیز این مسئله باعث توجه ویژه به آموزش زبان فارسی به غیرفارسی‌زبانان شده است. در قلمرو آموزش زبان، اهمیت واژگان رکنی اساسی محسوب می‌شود؛ به‌طوری که محققان و مدرسان همواره به اهمیت آموزش واژه‌ها در زبان دوم/خارجی توجه داشته‌اند (Alenizi & Adawi 2024). دانش واژگانی یکی از راه‌هایی است که موجب افزایش مهارت زبان‌آموز در زبان دوم می‌شود (Youngblood & Folse 2017). در واقع واژه، سنگ بنای اصلی در تمام مهارت‌های زبانی در زبان اول و دوم به‌شمار می‌آید؛ به‌طوری که چگونگی آموزش و یادگیری آن در پیشبرد مهارت‌های ثمربخش[1] (نوشتاری و گفتاری) و دریافتی[2] (شنیداری و خوانداری[3]) و تسلط بر زبان دوم اثرگذار است (Tiansoodeenon et al. 2023; Nation 2001). بنابراین، مشخص است که مدرسان زبان و زبان‌شناسان کاربردی به‌طور عام، به اهمیت یادگیری واژگان پی برده و در جست‌وجوی روش‌هایی هستند که بدان‌وسیله سبب ارتقای کارآمد دامنه واژگانی زبان‌آموزان شوند. در

1. productive
2. receptive
3. reading

سال‌های اخیر، ترکیب زبان‌شناسی رایانشی و علم زبان‌شناسی پیکره‌ای سبب شده است تا پیکره‌های متنوعی در حوزه آموزش زبان طراحی شوند. در این راستا باید بیان نمود که زبان‌شناسی رایانشی شاخه‌ای میان‌رشته‌ای از زبان‌شناسی محسوب می‌شود که تحلیل، درک و تولید زبان گفتاری و نوشتاری به‌واسطه علوم رایانه‌ای انجام می‌گیرد. به‌گفته دقیق‌تر، در این شاخه علوم زبان‌شناسی، رایانه و هوش مصنوعی با یکدیگر تلفیق می‌شوند تا فهم یا تولید زبان از دیدگاهی رایانشی انجام شود (Ahmad et al. 2025).

زبان‌شناسی پیکره‌ای شاخه‌ای مهم از مطالعات پیکره‌ای است که می‌تواند به‌عنوان ابزاری کمک‌رسان در خدمت آموزش زبان دوم قرار گیرد و یکی از ابزارهای مهم در تدریس به‌شمار آید (Li et al. 2025). با توجه به نقش مهمی که زبان‌شناسی پیکره‌ای در سال‌های اخیر در زمینه‌های متعدد آموزشی ایفا کرده، در ارتباط با آموزش واژگان یک زبان، می‌توان از رویکرد پیکره‌ای بهره گرفت. نقطه شروع بسیاری از رویکردهای مبتنی بر پیکره در آموزش زبان، پژوهش‌های Sinclair (1987, 1991, 2004) بوده است. در دو دهه اخیر، مطالعات بسیاری بر تأثیر و نقش مهم پیکره‌های آموزشی به‌ویژه در ارتباط با آموزش واژگان، تأکید داشته‌اند (Chan & Liou 2005; Chen 2011; Cobb 1999; Daskalovska 2015; Li 2017; Varley 2009). زبان‌آموزان نیز به‌منظور یادگیری کلمات و نگارش متون دانشگاهی به استفاده از فناوری پیکره[1] علاقه‌مند هستند (Yan & Ma 2025). منظور از فناوری پیکره، اعمال فناوری مرتبط با زبان‌شناسی پیکره‌ای و به‌کارگیری پیکره به‌منظور یادگیری و آموزش است (Ma et al. 2024, 462). در ارتباط با یادگیری واژگان، استفاده از پیکره دستاوردهای قابل توجهی داشته است (Çalışkan & Kuru Gönen 2018). در این راستا، می‌توان بیان داشت که تاکنون فهرست‌های واژگانی متعددی نظیر AWL (Gardner & Davies 2014) و GSL (West 1953) در ارتباط با واژگان عمومی و تخصصی زبان انگلیسی طراحی شده‌اند. افزون بر این، در زبان فارسی نیز تلاش‌هایی به‌منظور تدوین پیکره‌های واژگانی انجام شده و پژوهشگرانی همچون «بی‌جن‌خان و محسنی» (۱۳۹۱)، «حسنی» (۱۳۸۴)، «عبادی، وکیلی‌فرد و بهراملو» (۱۳۹۳) به‌ترتیب، «فرهنگ بسامدی واژگان فارسی»، «واژه‌های پرکاربرد فارسی امروز»، و «فهرست واژگان پایه فارسی» را تدوین نموده‌اند. لازم به توضیح است که این پیکره‌های تولیدشده محدود هستند، و در

1. corpus technology (CT)

ارتباط با کل زبان فارسی بوده و برای گویشوران فارسی طراحی شده‌اند و در نتیجه، خاص آموزش زبان فارسی به غیرفارسی‌زبانان طراحی نشده‌اند. همچنین، توجه به تمایز واژه‌ها از نظر سطوح زبانی از اهمیت بسیار برخوردار است و ثبت اطلاعات مرتبط با سطوح زبانی در فرهنگ‌ها و واژه‌نامه‌ها ضروری به نظر می‌رسد که شوربختانه به این مهم، به‌علت مشخص نبودن معیارهای زبانی و فرازبانی توجه کافی نشده است (وکیلی‌فرد ۱۳۷۸)

با توجه به موارد اشاره‌شده، جای پیکره‌ای از واژگان که به‌صورت سطح‌بندی‌شده از منابع آموزشی غیرفارسی‌زبانان تدوین شده باشد، خالی است. این می‌تواند در زمینه‌های یادگیری، آموزش و ارزشیابی و طراحی آزمون نیز راهگشا باشد. بنابراین، قصد داریم در این پژوهش به تدوین پیکره‌ای سطح‌بندی‌شده (مبتدی، میانه، پیشرفته) از واژگان بر اساس منابع به‌روز آموزشی که در مراکز آموزش زبان فارسی به غیرفارسی‌زبانان کاربرد دارد، بپردازیم. در این راستا، به‌منظور پرهیز از اعمال سلیقه شخصی پژوهشگر و بر اساس همپوشانی که بین منابع مختلف در ارتباط با واژگان در سطوح مختلف آموزشی (پایه، مقدماتی، پیشرفته) وجود دارد، پیکره‌ای سطح‌بندی‌شده از واژگان[1] را مطابق با اصول و مبانی ساخت پیکره تهیه و تدوین می‌نماییم و سؤال‌های پژوهش را که در ادامه مطرح می‌شوند، پاسخ می‌دهیم

۱. میزان همپوشانی واژگان سطح مبتدی در منابع پرکاربرد مراکز «آزفا»[2] چقدر است؟
۲. میزان همپوشانی واژگان سطح میانی در منابع پرکاربرد مراکز «آزفا» چقدر است؟
۳. میزان همپوشانی واژگان سطح پیشرفته در منابع پرکاربرد مراکز «آزفا» چقدر است؟

به نظر ضروری می‌رسد که پیش از ورود به سایر بخش‌های پژوهش، منظور پژوهشگران از تدوین پیکره واژگانی سطح‌بندی‌شده شفاف‌سازی شود. ابتدا باید بیان داشت که پیکره‌ها منابعی زبانی هستند که حجم بسیاری از نمونه‌های زبانی نظیر واژه، جمله و متن به‌منظور تدوین آن‌ها گردآوری شده‌اند. همچنین، منابع واژگانی نیز منابع زبانی‌ای به‌شمار می‌آیند که دربردارنده مدخل‌های واژگانی بسیاری هستند و اطلاعاتی را در ارتباط با واژه‌ها و عبارات زبان در اختیار می‌گذارند. این منابع که واژگان نامیده

---

۱. این پیکره با عنوان «فردوس-لکس» نام‌گذاری شده است که بخش اول این عنوان «فردوس» به دانشگاه فردوسی مشهد اشاره دارد و بخش دوم که «لکس» است، از Lexical Corpus به معنای پیکره واژگانی گرفته شده است

۲. آموزش زبان فارسی به غیرفارسی‌زبانان

می‌شوند، ساختارهایی برای ذخیره‌سازی واژه‌ها و کلمات زبان هستند که می‌توانند شامل طیفی وسیع از فهرست کلمات تا واژگان معنایی شوند (شمس‌فرد ۱۴۰۱).

حال، در پژوهش حاضر، مفهوم پیکره را با منابع واژگانی ترکیب نمودیم و پیکره‌ای واژگانی در سطوح[1] مختلف زبان‌آموزی ساختیم که حاوی مؤلفه‌هایی نظیر حرف الفبا[2]، واحد[3]، بن‌واژه[4]، دامنه همپوشانی[5]، برچسب اجزای کلام[6] و فراداده کتاب‌های منبع[7] آن می‌شود. افزون بر این، لحاظ نمودن سطوح مبتدی، میانی و پیشرفتهٔ زبان‌آموزی به این پیکره، ویژگی سطح‌بندی‌شده به آن بخشیده است.

در این قسمت به توضیح پیرامون چند مفهوم پردازشی که در تدوین پیکره نقش دارند، می‌پردازیم. به‌منظور ایجاد یک پیکره بایستی ابتدا داده‌های گردآوری‌شده آماده‌سازی شوند. این مرحله را پیش‌پردازش می‌نامند. در این مرحله، مفاهیم هنجارسازی[8] و واحدسازی[9] مطرح می‌شوند. هنجارسازی به معنای کدگذاری یکسان حروف به‌کاررفته بر داده‌های ورودی است؛ به این منظور که به مجموعه‌ای واحد تبدیل شوند. در طی این مرحله، اعداد و شکل‌های گوناگون املایی نیز یکدست شده و جدول‌ها، شکل‌ها، حروف تکراری و سایر علامت‌های اضافی نظیر کشیدگی حروف حذف می‌شوند (قیومی ۱۴۰۱). گاهی اوقات چندواژگی در خط فارسی سبب می‌شود که واژه به‌درستی تشخیص داده نشده و به بروز مشکلاتی در تمام لایه‌های پردازشی منجر شود. از این‌رو، واحدسازی به فرایندی اطلاق می‌شود که طی آن یک واحد واژگانی، تشخیص داده می‌شود (همان). سپس، پردازش‌های پایه صورت می‌پذیرد که شامل مفاهیم بن‌واژه‌سازی[10] و برچسب‌دهی اجزای کلام[11] است. بن‌واژه‌سازی فرایندی ساخت‌واژی است که طی آن تمامی صورت واژه‌های یک واژه در زیر یک شکل پایه یا بن‌واژه قرار می‌گیرند تا در فرایند تحلیل، تنها

1. language proficiency level
2. alphabet
3. token
4. lemma
5. vocabulary overlap range
6. parts-of-speech tag (POS Tag)
7. metadata
8. normalization
9. tokenization
10. lemmatization
11. parts-of-speech tagging (POS Tagging)

بن‌واژه تحلیل شود (همان). در ارتباط با مفهوم برچسب‌گذاری در یک پیکره «می‌یر» بیان می‌دارد که نرم‌افزار برچسب‌دهنده، عمل برچسب‌گذاری را بر روی هر واژه انجام می‌دهد (Meyer 2004, 86). بدین‌ترتیب، با در نظر گرفتن پیکره‌ای مشخص به واژگان موجود در آن پیکره بر حسب مقوله دستوری آن‌ها در متن، برچسب (اسم، فعل، قید، صفت و غیره) تعلق می‌گیرد (علایی ابوذر ۱۳۹۷).

لازم به ذکر است که با توجه به ماهیت کیفی پرسش‌های پژوهش، فرضیه‌ای ذکر نشده است. همچنین، در این پژوهش برخی مواقع سرواژه[1] «آزفا» در متن مشاهده می‌شود که به‌جای آموزش زبان فارسی به غیرفارسی‌زبانان به‌کار رفته است

۲. مبانی نظری

مبانی نظری مرتبط با حوزه پژوهش که شامل مفهوم پیکره و زبان‌شناسی پیکره‌ای، مبانی ساخت پیکره و ارتباط زبان‌شناسی پیکره‌ای با حوزه آموزش، مفهوم واژگان چندجزئی و چارچوب مشترک اروپا[2] (۲۰۰۱) می‌شود، در ادامه مورد بررسی قرار خواهد گرفت.

۲-۱. پیکره و زبان‌شناسی پیکره‌ای

در زبان‌شناسی نوین، پیکره به مجموعه‌ای از داده‌های کلامی یا مکتوب اطلاق می‌شود که از چندین خصیصه برخوردار است. نخست اینکه این داده‌ها از متن‌های طبیعی نمونه‌برداری شده باشند، ماشین قابلیت خوانش آن‌ها را داشته باشد، و قابلیت برچسب‌گذاری داده‌های زبانی برای ماشین فراهم باشد (McEnery, Xiao & Tono 2006, 4). اصطلاح زبان‌شناسی پیکره‌ای نخستین‌بار در اوایل دهه ۱۹۸۰ ظاهر شد؛ گرچه مطالعات مبتنی بر پیکره قدمت قابل توجهی دارد (Leech 1992, 105). بدین‌صورت که روش‌شناسی پیکره به دوره قبل از «چامسکی»[3] بازمی‌گردد و پیشتر توسط زبان‌شناس‌هایی همچون «بواز»[4] و چندین زبان‌شناس ساختارگرا نظیر «سپیر»[5] و «بلومفیلد»[6] به‌کار گرفته شده بوده است (Biber &

1. acronym
2. The Common European Framework
3. Chomsky
4. Boas
5. Sapir
6. Bloomfield

Finegan 1991, 207). به بیانی ساده، زبان‌شناسی پیکره‌ای را می‌توان به‌عنوان مطالعه‌ای به‌منظور گردآوری و تحلیل مجموعه‌ای از متون (پیکره‌ها) تعریف کرد (Cheng 2012, 6). بنا بر نظر «هانستون»، به‌رغم اینکه زبان‌شناسی پیکره‌ای حوزه‌ای به‌نسبت جدید است، انقلابی در مطالعات زبان ایجاد کرده است؛ به این علت که روش‌های جدیدی برای تحلیل و توصیف کاربرد زبان با جست‌وجوی رایانه‌ای ارائه نموده است (Hunston 2002). در سال‌های اخیر، پیکره‌ها به منبعی ارزشمند در زمینه یادگیری و آموزش زبان تبدیل شده و از نقش به‌سزایی در تهیه و تدوین منابع آموزشی و آزمون‌سازی برخوردار هستند (Li et al. 2025). در ارتباط با ارزش آموزشی پیکره‌ها و ابزارهای پیکره‌ای Ma et al. (2024) آن‌ها را به‌عنوان فناوری‌ای ویژه در محیط یادگیری زبان به کمک ابزارهای رایانشی در نظر گرفت و با لقب فناوری پیکره به آن‌ها اشاره نمود. افزون بر این، در راستای ارزش پیکره در مطالعات زبان، «هانستون» و «مکارتی و اوکیف» بیان داشتند که رویکردهای پیکره‌ای می‌توانند در زمینه‌های مختلف مطالعات زبان‌شناسی از جمله آموزش و یادگیری زبان، تحلیل‌های مقابله‌ای، ترجمه، کاربردشناسی و کاربرد زبان با اهداف خاص به‌کار گرفته شوند (Hunston 2002; McCarthy & O' Keeffe 2010). از میان پیکره‌های زبان‌آموز متعدد می‌توان به پیکره‌های بین‌المللی زبان‌آموز انگلیسی[1] و پیکره بین‌المللی بینازبانی[2] اشاره کرد که پیکره‌هایی هستند که به‌صورتی مفید در پژوهش‌های مرتبط با واژه‌آموزی کاربرد دارند (Szudarski 2018, 4). رویکردهای مختلفی در زبان‌شناسی پیکره‌ای برای تفسیر نتایج وجود دارد که از میان آن‌ها «تاگنینی-بونلی» به تمایز میان دو رویکرد پیکره‌بنیاد[3] و پیکره‌محور[4] اشاره می‌کند (Tognini-Bonelli 2001). در رویکرد پیکره‌بنیاد، زبان‌شناسی پیکره‌ای به‌عنوان یک روش‌شناسی تلقی می‌شود که در آن از داده‌های پیکره به‌منظور استخراج شواهد در راستای بررسی نظریه‌های زبانی استفاده می‌شود. در مقابل، رویکرد پیکره‌محور می‌کوشد از پیکره و شواهد موجود در آن برای معناسازی و دستیابی به نظریه‌های زبانی استفاده کند. همچنین، «مونیر و رپن» رویکرد سومی را نیز تحت عنوان پیکره‌آگاه[5] معرفی می‌کنند که نه در آن به بررسی نظریه‌ها و فرضیه‌های زبانی در پیکره

1. international corpus of learner English (ICLE)
2. nternational corpus of crosslinguistic interlanguage
3. corpus-based
4. corpus-driven
5. corpus-informed

پرداخته می‌شود و نه قرار است با بررسی داده‌ها و شواهد موجود در پیکره به نظریه یا مشاهده دست یافت. در این رویکرد از نتایج و مشاهدات پژوهش‌های انجام‌شده بر پیکره استفاده می‌شود تا به تدوین منابع آموزشی از مشاهدات پیکره مانند واژگان، ساخت‌ها و الگوهای هم‌نشینی کمک شود (Meunier & Reppen 2015, 499).

۲-۲. مبانی ساخت پیکره

به‌طور کلی، پیکره یک مجموعه به‌نسبت بزرگ ماشین‌خوان[1] از نمونه‌های زبان طبیعی است که می‌تواند برچسب‌گذاری‌شده[2] باشد و در تحقیقات زبان‌شناسی کاربرد دارد (McEnery & Hardie 2011; Sinclair 2004). «رپن» و «بارث و شنل» در ساخت یک پیکره مواردی را مانند برچسب‌گذاری[3]، ماشین‌خوان بودن[4]، اندازه[5]، نمایندگی[6] و توازن[7] برشمرده‌اند (Reppen 2022; Barth & Schnell 2022).

یکی از جنبه‌های مهم در تدوین پیکره، برچسب‌گذاری است که به‌عنوان اصطلاحی کلی به فرایندهایی نظیر برچسب‌گذاری و افزودن اطلاعات زبان‌شناختی به پیکره اشاره دارد (Hunston 2002, 18). توجه به این نکته ضروری است که میزان برچسب‌گذاری به ماهیت داده‌های گردآوری‌شده و از همه مهم‌تر به اهداف تحقیقاتی بستگی دارد که پیکره در خدمت آن به‌کار گرفته می‌شود (Cheng 2012)؛ بدین‌معنا که برخی پیکره‌ها ممکن است در بعضی موارد برچسب‌گذاری داشته باشند یا به‌کل، هیچ‌گونه برچسبی دریافت نکنند. مفهوم برچسب‌گذاری بسته به هدف در سطوح متفاوتی مانند آوایی یا واجی، صرفی، نحوی، معنایی، و گفتمانی صورت می‌پذیرد (Kübler & Zinsmeister 2014). برچسب‌گذاری اجزای کلام بدان معناست که تمامی واژگان در یک پیکره از منظر مقوله‌های دستوری (اسم، فعل، صفت، قید، حرف اضافه) که به آن‌ها تعلق دارند، مشخص می‌شوند (Szudarski 2018, 10). ماشین‌خوان بودن عبارت است از ساختاردهی به دادگان پیکره به‌طوری که توسط ماشین قابل خوانش و جست‌وجو باشد و از یک رایانه

1. machine-readable
2. annotated
3. linguistic annotation
4. machine-readability
5. size
6. representativeness
7. balance

به رایانه دیگر تغییر نکند (Kübler & Zinsmeister 2014).

در زبان‌شناسی پیکره‌ای به‌طور معمول، از زبان‌های نشانه‌گذاری ساختاری مانند XML[1] و TEI[2] برای ماشین‌خوان بودن پیکره استفاده می‌شود. اندازه پیکره را به‌طور معمول، با تعداد واحدهای[3] آن ارزیابی می‌کنند. همچنین نمایندگی و توازن از موضوعات مهم در نمونه‌گیری[4] به‌شمار می‌روند تا از سوگیری یا گرایش خاص در جمع‌آوری دادگان جلوگیری شود (Reppen 2022; Barth & Schnell 2022). در نمایندگی بایستی به این مهم توجه شود که آیا نمونه‌های جمع‌آوری‌شده نماینده جمعیت هدف هستند؟ در مورد توازن که با نمایندگی نیز ارتباط تنگاتنگی دارد، بایستی به این مهم توجه شود که از انواع نمونه‌ها، دسته‌ها[5] و یا گونه‌های زبانی به میزان متعادل و متناسب نمونه‌گیری انجام شود. به‌طور کلی، در نمونه‌گیری مناسب برای ساخت پیکره باید نسبت به سبک یا سیاق زبان[6]، نوشتاری یا گفتاری بودن[7] و گونهٔ[8] زبان سوگیری دیده نشود

۲-۳. ارتباط زبان‌شناسی پیکره‌ای و آموزش زبان

اوایل دهه ۱۹۹۰، شاهد افزایش علاقه در به‌کارگیری یافته‌های پژوهش‌های مبتنی بر پیکره در آموزش زبان بود که خود، موضوعات مختلفی همچون توصیف زبان مبتنی بر پیکره، تحلیل‌های پیکره‌ای در کلاس درسی و پیکره‌های زبان‌آموز را شامل می‌شود (Keck 2004). افزون بر این، «لیچ» ضمن اشاره به مشهود بودن همسویی آموزش و پیکره زبانی بیان داشت که همگرایی این دو حوزه بر کاربرد مستقیم و غیرمستقیم پیکره (تدوین مطالب آموزشی و آزمون‌سازی) در آموزش زبان اول و دوم متمرکز است (Leech 1997). افزون بر این، کاربرد پیکره باید به دیدگاه‌های کاملاً متفاوتی در طراحی برنامه درسی منجر شود که این امر به‌طور اخص و گسترده در ارتباط با واژگان در برنامه درسی مورد بحث قرار می‌گیرد (Huston 2002, 189). درسنامه واژگان تمام جنبه‌های زبان را که شامل

1. extensible mark-up language (XML)
2. text encoding initiative (TEI)
3. token
4. sampling
5. genres
6. register
7. mode (written/spoken)
8. variety

صورت‌های واژگانی، الگوهای کاربردی اصلی و ترکیب‌های شکل‌گیری واژه‌ها می‌شود، مد نظر قرار می‌دهد و این نوع پیکره نباید با پیکره‌ای که صرفاً دربردارنده واژگان است، اشتباه گرفته شود (Sinclair & Renouf 1988, 148). افزون بر این، همواره این مسئله مورد توجه بوده است که چه واژگانی باید تدریس شوند و این موضوع نه‌تنها برای فراگیران، بلکه برای سیاست‌گذاران حوزه آموزش ضروری است تا از اصولی برای تشخیص انتخاب کلمات به‌منظور آموزش پیروی کنند (Nagy & Hiebert 2011, 388). در این راستا، در نظریه انتخاب واژگان به معرفی چندین مؤلفه (بسامد[1]، پراکندگی[2]، ارتباط ساختواژی[3]، ارتباط معنایی[4]، آشنایی[5]، دشواری مفهومی[6]، نقش کلمات در متون خاص[7]، نقش کلمات در برنامه‌های آموزشی گسترده[8]) پرداخته شده که لازم است در تدوین کتاب‌های آموزشی مورد توجه قرار گیرند (Nagy & Hiebert 2011). سپس، «سروتی» و همکاران با تکیه بر مؤلفه‌های هشت‌گانه نام‌برده‌شده در نظریه انتخاب واژگان، مؤلفه طول کلمه[9] را نیز به‌عنوان عاملی مؤثر در یادگیری واژگان برشمردند (Cervetti et al. 2015)

۲-٤. چارچوب مشترک اروپا

این چارچوب مبنایی مشترک برای تدوین برنامه‌های درسی زبان، راهنمایی به‌منظور تهیه و تدوین برنامه‌های آموزشی، طراحی آزمون‌ها و کتاب‌های درسی در سراسر اروپا فراهم می‌کند. چنین چارچوبی به‌طور جامع آنچه را که زبان‌آموزان لازم است یاد بگیرند تا قادر باشند از زبان برای تعامل استفاده نمایند، نشان می‌دهند (Council of Europe 2001, 1).

یکی از اهداف چارچوب مشترک اروپا کمک به توصیف سطوح مهارتی مورد نیاز، مطابق با استانداردها و آزمون‌های موجود است تا از این طریق امکان مقایسه بین سیستم‌های تحصیلی مختلف فراهم شود. این چارچوب از شش سطح گسترده که شامل

1. frequency
2. dispersion
3. morphological relatedness
4. semantic relatedness
5. familiarity
6. conceptual difficulty
7. role in the particular texts
8. role in the larger Curriculum
9. word length

نوآموز[1]، پایدار[2]، آستانه[3]، پیشرفته[4]، مستقل[5]، و تسلط کامل[6] می‌شود، تشکیل شده است که پوشش کافی از فضای یادگیری مرتبط با زبان‌آموزان اروپایی را فراهم می‌آوَرَد. به‌طور کلی، با توجه به مهارت‌های چهارگانه خواندن، شنیدن، صحبت کردن و نوشتن، این شش سطح قابل دسته‌بندی به سه سطح مبتدی[7]، میانه[8] و پیشرفته[9] هستند (Council of Europe 2001, 23 & 24). لازم به ذکر است که در پژوهش حاضر نیز سه سطح کلی مبتدی، میانه و پیشرفته در نظر گرفته شده است

۲-۵. واژگان چند-جزیی[10]

در این بخش ابتدا به تمایز دو مفهوم «واحدهای چند-قطعه‌ای» و «قطعه‌های چند-واحدی» پرداخته می‌شود که ناشی از مسئله چند-واژگی در زبان فارسی است. در این راستا، «شریفی‌آتشگاه و بی‌جن‌خان» بیان داشته‌اند که واحدهای چند-قطعه‌ای به زنجیره‌ای از حروف اشاره دارند که به‌دلیل چسبندگی دو یا چند بخش بدون درج فاصله کامل ایجاد می‌شوند. این امر باعث می‌شود که رایانه، این زنجیره را به‌عنوان یک واحد واژگانی تشخیص دهد؛ در حالی که از منظر زبان‌شناسی شامل چند واژه است. برای مثال، «آنهادرعذاب‌اند» به‌جای «آنها در عذاب‌اند» به‌صورت یک واحد واژگانی شمرده می‌شود، زیرا نبودِ فاصله کامل، پردازش رایانه‌ای را مختل می‌کند. این مشکل به‌دلیل ویژگی برخی حروف فارسی مانند «آ-ا»، «د»، «ذ»، «ر»، «ز»، «ژ» و «و» است که به حرف بعدی نمی‌چسبند و در صورت عدم درج فاصله، خوانش مشکلی ندارد، اما پردازش داده‌ها را دشوار می‌کند (Sharifi Atashgah & Bijankhan 2009). «قیومی» نیز به چالش دیگری از واحدهای چند-قطعه‌ای اشاره کرده است که ناشی از کوتاه‌سازی و ادغام چند واژه در یک واحد است؛ مانند «آن‌هاست» که از ترکیب «آن‌ها» و «است» به‌وجود آمده است

1. breakthrough (A1)
2. waystage (A2)
3. threshold (B1)
4. vantage (B2)
5. effective operational proficiency (C1)
6. mastery (C2)
7. basic user
8. independent user
9. proficient user
10. multi-word tokens (MWTs)

(۱۳۹۶). این ادغام، اغلب به دلایل زیبایی‌شناختی در متون ادبی و شعر، با حذف حروف مشترک انتهای واژه قبلی و ابتدای واژه بعدی رخ می‌دهد و تجزیهٔ نحوی جمله را پیچیده می‌کند. «شریفی آتشگاه و بی‌جن‌خان»، در ارتباط با قطعه‌های چند-واحدی بیان نمودند که این قطعه‌ها به واژه‌هایی گفته می‌شود که به‌اشتباه به‌جای نیم‌فاصله، فاصله کامل بین اجزای تشکیل‌دهنده یک واحد واژگانی درج شده و این امر باعث جدایی اجزای آن واحد می‌شود؛ مانند نوشتن «بین الملل» به‌جای «بین‌الملل». این خطا ساخت‌واژه را تغییر داده و گاهی واژه‌های بی‌معنایی مانند «الملل» تولید می‌کند (ibid). در این پژوهش بنا به هدف که ساخت پیکره واژگانی است، تنها واژگان بسیط و «قطعه‌های چند-واحدی» توسط پژوهشگران در فایل‌ها تایپ شدند. به‌گفته دیگر، «واحدهای چند-قطعه‌ای» در دادگان پژوهش اخیر لحاظ نگردیده‌اند. از این‌رو، در پژوهش کنونی تعریف عملیاتی واژگان چند-جزئی، همان واژگان مرکب و «قطعه‌های چند-واحدی» است.

این واحدهای واژگانی چند-جزئی نباید به واحدهای کوچک‌تر تجزیه شوند. در زبان فارسی، افعال مرکب مانند «استفاده کردن» و «انجام دادن»، افعال پیشوندی مانند «برچیدن» و «برگرداندن» و افعال چند-جزئی مانند «به کار بردن» و «از دست دادن» از این قبیل هستند که ساخت‌های فعلی سبک[1] را تشکیل می‌دهند. همچنین، اسامی مرکب از این قبیل، واحدهای واژگانی چند-جزئی به‌شمار می‌روند که فاصله‌گذاری بین اجزای آنان دارای اهمیت است؛ مانند «آب‌وهوا» و «گفت‌وگو». در نگارش این واحدها به‌طور معمول، از نیم‌فاصله استفاده می‌شود و طبق قواعد نگارش در شیوه‌نامه خط فارسی، این واژگان نباید به‌صورت متصل نوشته شوند (فرهنگستان زبان و ادب فارسی ۱۴۰۱، ۲۳)[2]. شوربختانه، شیوه‌نامه نگارش خط فارسی همواره به‌صورت یکنواخت و یک‌دست رعایت نمی‌شود و این مسئله به تنوع شیوه‌های نگارشی در متون فارسی دامن می‌زند؛ مانند (گفت و گو، گفت‌وگو، گفتگو). همان‌طور که در قسمت روش‌شناسی شرح داده می‌شود، تمام این شکل‌های گوناگون در نگارش واحدهای چند-جزئی هنگام تایپ کردن و هنگام پیش‌پردازش، یک‌دست و یکسان‌سازی شدند تا پردازش رایانه‌ای زبان فارسی را با مشکل مواجه نسازند. لازم به ذکر است که برای جست‌وجوی ساخت‌های سبک فعلی در زبان فارسی، یک درگاه

1. light verb constructions (LVCs)

2. https://apll.ir/wp-content/uploads/2023/07/Dastour-e-Khat-17.04.1402.pdf

آنلاین دادگان[1] مبتنی بر فرهنگ ظرفیت نحوی طراحی شده تا زبان‌شناسان رایانه‌ای در این زمینه به مشکل برنخورند.

یک نکته مهم در پژوهش حاضر این است که برای یافتن واژگان همپوشان، واحدهای واژگانی چند-جزئی به‌عنوان معیار مقایسه و شمارش استفاده شده است تا نتایج، با دقت گزارش شوند. به‌عنوان مثال، اگر فعل «زمین خوردن» به دو واحد تجزیه شود، ممکن است با «خوردن» یا «زمین» در منابع مختلف همپوشانی داشته باشد؛ در صورتی که «زمین خوردن» به‌خودی خود، یک واحد مستقل است و برای بررسی همپوشانی آن نباید به زیرواحدهایش تجزیه شود. در قسمت روش پژوهش به شرح پیش‌پردازش این واحدهای چندجزئی اشاره شده است. برچسب‌گذاری پیکره در پژوهش حاضر شامل برچسب اجزای کلام بوده و ساختاردهی پیکره برای ماشین‌خوان بودن نیز با زبان نشانه‌گذاری XML صورت پذیرفته است. همچنین، از آنجا که در پژوهش حاضر کل واژگان منابع آموزش فارسی به غیرفارسی‌زبانان همراه با اطلاعات همپوشانی آن‌ها در ساخت یک پیکره سطح‌بندی‌شده مد نظر هستند، اندازه پیکره شامل تمامی واژگان در هر سطح است تا نمایندگی و توازن نیز رعایت شود.

۳. پیشینه پژوهش

با جست‌وجوی منابع متعدد مشخص شد که در ارتباط با آموزش زبان فارسی به غیرفارسی‌زبانان پژوهش‌های پیکره-بنیاد محدودی انجام شده است. در ادامه، به معرفی مطالعات خارجی و داخلی پرداخته می‌شود.

«النیزی و اداوی»، به «بررسی اثربخشی استفاده از روش توسعه‌یافته مبتنی بر پیکره در یادگیری واژگان برای دانشجویان EFL عربستان» با طرح آزمایشی پنج هفته‌ای پرداختند. شرکت‌کنندگان شامل ۲۴ دانشجوی دختر رشته زبان انگلیسی از عربستان می‌شدند که به‌طور مساوی به دو گروه کنترل و آزمایش تقسیم شدند. در این پژوهش از پیش‌آزمون، پس‌آزمون، پرسشنامه، و یادداشت‌های روزانه استفاده شد. داده‌ها با استفاده از نرم‌افزار «اس‌پی‌اس‌اس»[2] و آزمون «تی»[3] بررسی شدند. تجزیه‌وتحلیل یادداشت‌های روزانهٔ فراگیران

1. http://search.dadegan.ir/
2. SPSS
3. T-test

نشـان داد کـه هـر دو گـروه بـر اسـاس نتایـج پس‌آزمـون، واژگان بیشـتری در طـول دوره مطالعـه کسـب نمودنـد. نتایـج پرسشـنامه در ارتبـاط بـا نگـرش زبان‌آمـوزان نشـان داد کـه بیشـتر زبان‌آمـوزان به‌رغـم مواجهـه بـا بـرخی مشـکلات، نگـرش مثبـتی نسـبت بـه اسـتفاده از پیکـره بـرای یادگیـری واژگان داشـتند (Alenizi & Adawi 2024).

«یـو لییـو»، اهمیـت فهرسـت واژگان دانشـگاهی را بـرای درک زبان‌آمـوزان زبـان دوم برجسـته کـرد. ایـن مطالعـه پیکـره‌ای ۱۰/۲ میلیـون-واژه‌ای از دوره‌هـای آزاد آنلایـن[1] در چهـار حـوزه مختلـف (مهنـدسی، علـوم انسـانی و هنـر، علـوم و ریاضیـات، و علـوم اجتمـاعی) ایجـاد نمـود. ایـن پیکـره نیازهـای واژگانی دوره‌هـای آزاد آنلایـن را مـورد بـررسی قـرار داد و پوشـش‌دهی فهرسـت واژگان عمـومی و تخصـصی دانشـگاهی را در پیکـره MOOCs بـررسی نمـود. نتایـج نشـان داد کـه بـا در نظـر گرفتـن اسـامی خـاص، واژگان حاشیـه‌ای، اسـامی مرکـب و مخففـ‌هـا، متداول‌تریـن خانواده‌هـای واژگانی از انگلیـسی عمـومی بیـن ۹۰ تـا ۹۵ درصـد بـا پیکـره مـورد نظـر پوشـش داشـتند (Yu Liu 2023). ایـن امـر بیانگـر ایـن مطلـب اسـت کـه به‌دلیـل قـرار گرفتـن واژگان عمـومی و تخصـصی در پیکـره MOOCs، یادگیـری واژگان بـرای زبان‌آمـوزان تسـهیل شـده اسـت

«برزینـا و گابلاسـوا» فهرسـتی از واژگان جامـع جدیـد[2] را ارائـه نمودنـد. ایـن فهرسـت حاصـل مقایسـه‌ای بیـن چهـار پیکـره زبـانی بـود کـه شـامل بیـش از ۱۲ میلیـارد واژه متـداول می‌شـد. در ایـن پژوهـش، همپوشـانی واژگانی میـان پیکره‌هـا در ۳۰۰۰ واژه پایـه بـر اسـاس بسـامد متوسـط کاهش‌یافتـه[3] انجـام شـد. نتایـج ایـن پژوهـش حـاکی از وجـود ۲۱۲۲ واژه اصـلی در بیـن چهـار پیکـره بـود. محصـول نهـایی ایـن بـررسی پیکـره‌ای شـامل ۲۴۹۴ بـن‌واژه (مدخـل) شـد کـه ۸۱ درصد از متن در پیکره اصلی را پوشش می‌داد (Brezina & Gablasova 2015).

«صحـرائی و میـرزایی» در پژوهـشی، کاربردهـای زبان‌شـناسی پیکـره‌ای در آمـوزش فـارسی بـه غیرفارسی‌زبانـان را بـررسی نمودنـد. در ایـن مقالـه پیکره‌هـای مرتبـط و مفیـد در ارتبـاط بـا آمـوزش زبـان نظیـر پیکـره آمـوزشی زبـان و پیکـره زبان‌آمـوز معـرفی شـدند. همچنیـن، پژوهشـگران بـه فهرسـت بـرخی داده‌هـای زبـانی برچسـب‌خورده زبـان فـارسی پرداختنـد و نشـان دادنـد کـه چگونـه می‌تـوان از آن‌هـا در آمـوزش زبـان فـارسی اسـتفاده نمـود. نتایـج ایـن پژوهـش

1. massive open online courses (MOOCs)
2. NEW-GSL
3. average reduced frequency (ARF)

حاکی از کاربردی بودن پیکره‌های نامبرده در زمینه آموزش زبان بود. افزون بر این، مشخص شد که داده‌های زبانی اثر مثبتی بر تولید محتوای آموزشی مناسب می‌گذارند (Sahraei & Mirzaei 2023).

«جهانگردی»، در پژوهشی سه کتاب آموزشی از منابع آموزشی زبان فارسی به غیرفارسی‌زبانان را که در سه مرکز آموزشی متفاوت تدریس می‌شدند، با هدف بررسی وضعیت آموزش واژگان انتخاب نمود. رویکرد این پژوهش پیکره‌ای-شناختی بود. وی ابتدا از هر یک از کتاب‌های منتخب یک پیکره زبانی ایجاد نمود و پربسامدترین واژه‌ها از این پیکره‌های آموزشی را با پیکره‌ای مبنا که نماینده زبان طبیعی فارسی بود و شامل برگزیده‌ای از متون گفتاری و نوشتاری می‌شد، مقایسه کرد. از جمله دستاوردهای این پژوهش، ارائه پنج فهرست واژگانی-بسامدی بود که سه فهرست برای هر یک از کتاب‌ها، یک فهرست مربوط به پیکره زبان‌آموزی و یک فهرست برای پیکره مبنا طراحی شد. یافته‌های این پژوهش نشان داد که بدون در نظر گرفتن سطوح واژه‌آموزی، کتاب‌های آموزشی در تأمین نیازهای واژگانی زبان‌آموزان ناکارآمد خواهند بود. بدین‌ترتیب، وی با دستیابی به فهرست واژگانی ۵۰۰۰-واژه‌ای سعی داشت مؤلفان کتاب‌های آموزشی را در تنظیم متن‌های لازم برای مهارت خواندن و واژه‌آموزی یاری نماید (Jahangardi 2016). همان‌طور که مشخص است در این پژوهش تنها سه کتاب از مجموعه منابع «آزفا» مورد بررسی قرار گرفته و به سایر منابع پُرکاربرد توجهی نشده، و توجه پژوهشِ صورت‌گرفته بر مقایسه واژه‌های پربسامد منابع «آزفا» با پیکره مبنای واژگان زبان فارسی و سنجش همپوشانی این واژه‌ها بوده است.

«عبادی، وکیلی‌فرد و بهراملو» در پژوهشی، تدوین فهرست واژگان پایه برای زبان فارسی را با رویکردی تلفیقی انجام دادند. در این پژوهش با توجه به ۳ معیار بسامد، پوشش متنی و گستره واژگان پایه در یک پیکره ۳۴۰۰۰-کلمه‌ای، به اصلاح فرهنگ «بی‌جن‌خان و محسنی» (۱۳۹۱) پرداخته شده است. در انتها، مشخص شد که واژگان این پیکره، گستره مناسبی دارد و با اصلاح آن می‌توان به فهرستی از واژگان پایه فارسی دست یافت (Ebadi, Vakilifard & Bahramlu 2014).

«ترابی»، به این موضوع اشاره کرده است که پیکره‌ها در زمینه آموزش زبان، به‌ویژه زبان فارسی نقش‌آفرین هستند. وی به تدوین پیکره‌ای آموزشی شامل ۶۰۰۰۰ واژه از دسته‌های مختلف مبادرت ورزید که با توجه به ۶۰ متن از متن‌های نوشتاری معاصر تولید شده است.

واژه‌های این پیکره از گونه نوشتاری بودند و بر اساس اجزای کلام برچسب‌گذاری شدند (۱۳۸۹). این پیکره به نشانی www.corpus.ir در دسترس عموم قرار دارد

پژوهش‌هایی که در این بخش معرفی شدند، به اهمیت پیکره‌ها در حوزه آموزش، نقش مؤثر پیکره در آموزش واژگان و اثربخش بودن آن در رشته‌های تحصیلی مختلف به‌منظور رفع نیاز واژگانی فراگیران اشاره داشتند. همچنین، در سایر پژوهش‌ها نظیر Jahangardi (2016) ملاحظه شد که آنچه که مد نظر بوده، مقایسه واژه‌های پربسامد از منابع محدود در حوزه «آزفا» با پیکره‌ای از زبان فارسی بوده است. افزون بر این، محصول پژوهش Ebadi, Vakilifard & Bahramlu (2014) تهیه فهرستی از واژگان پایه بوده است. پیکره آموزشی Torabi (2010) به آموزش کاربرد واژه‌ها در بافت توجه داشته و فاقد فهرستی از واژگان برای غیرفارسی‌زبانان است و منبع گردآوری داده‌های آن متن‌های چاپی و غیرچاپی نظیر کتاب‌ها، مجله‌ها، وبگاه‌ها و نشریات است. همان‌طور که مشخص است، نیاز به منابع آموزشی مؤثر در ارتباط با یادگیری واژگان، به‌عنوان یک چالش بسیار مهم برای زبان‌آموزان مورد توجه پژوهشگران مختلف قرار گرفته است. با وجود این، واضح است که در هیچ‌یک از بررسی‌های صورت‌گرفته به سطوح مهارت‌های زبانی زبان‌آموزان و تدوین پیکره‌ای مطابق با منابع آموزشی در حوزه «آزفا» توجهی نشده است. از این‌رو، پژوهش حاضر و نتایج حاصل از آن قدمی نو در مسیر یادگیری واژگان زبان فارسی توسط غیرفارسی‌زبانان به‌شمار می‌آید

٤. روش پژوهش

پژوهش حاضر که نوعی تحلیل محتوای کیفی با رویکرد پیکره‌آگاه برای مقایسه واژگان همپوشان در سه سطح مبتدی، میانه و پیشرفته محسوب می‌شود، با کمک ابزارهای رایانشی و برنامه‌نویسی «پایتون»[1] در سه گام صورت پذیرفته است. در گام اول، آماده‌سازی دادگان صورت پذیرفته که شامل تایپ کردن واژگان، پیش‌پردازش[2] آن‌ها و افزودن برچسب‌زنی اجزای کلام است. پیش‌پردازش در پژوهش حاضر شامل هنجارسازی، واحدسازی، بن‌واژه‌سازی است و پس از آن برای تکمیل اطلاعات پیکره، برچسب‌زنی اجزای کلام پیاده شده است. در گام دوم، همپوشانی واژگان در هر یک از سه سطح زبان‌آموزی به

1. Python
2. preprocessing

کمک برنامه‌نویسی «پایتون» محاسبه شده تا به سؤالات پژوهش پاسخ داده شود. در گام سوم نیز، پیکره پژوهش در قالب استاندارد نشانه‌گذاری XML با کمک برنامه‌نویسی «پایتون» توسعه داده شده است. در ادامه، ابتدا دادگان پژوهش معرفی شده، آنگاه گام‌های پژوهش به‌ترتیب بیان می‌شوند

٤-١. دادگان پژوهش

دادگان پژوهش از ۲۶ منبع آموزشی که شامل مجموعه کتاب‌های «پرفا»، «مینا»، «شیراز»، «پارسا»، «رسا»، «نگارا»، «خوانا» و «آموزش نوین زبان فارسی» بودند، در سه سطح انتخاب و به‌صورت دستی تایپ شدند. این مهم دارای اهمیت است که به دو علت نیاز به استخراج و تایپ‌کردن دستی واژه‌ها بود. نخست، عدم دسترسی عمومی به فایل بیشتر منابع به‌رغم درخواست پژوهشی از مؤلفان و انتشارات، و دوم، فایل «پی‌دی‌اف» برخی از منابعی هم که در دسترس عموم بود، بعد از تبدیل نویسه نوری به متن[1]، دچار ایرادات نگارشی مثل به‌هم‌خوردگی کلمات به‌صورت غیرقابل فهم (درهم‌ریختگی حروف) می‌شدند. بنابراین، تایپ‌کردن دستی توسط پژوهشگران غیرقابل احتناب بود. برخی منابع واژگان بیشتری دارند؛ مانند «مینا ۳» با ۱۷۸۳ واحد، و برخی دیگر واژگان محدودتری دارند؛ مانند مقدمه با تنها ۲۷۹ واحد. همچنین، تعداد کل کلمات دادگان پژوهش ۱۴۵۸۵ واحد بوده است. مشخصات دادگان پژوهش در جدول ۱، آمده است:

جدول ۱. مشخصات دادگان پژوهش

نام کتاب	مؤلفان	سال انتشار	سطح کتاب	تعداد واژه
پرفا ۱	میردهقان، حاجی‌باقری، عبداللهی پارسا، باقری، آقایی، منتظری	۱۳۹۸	پایه	۱۱۱۳
پرفا ۲	میردهقان، حاجی‌باقری، عبداللهی پارسا، باقری، آقایی، منتظری	۱۳۹۸	میانی	۷۰۵
پرفا ۳	میردهقان، حاجی‌باقری، عبداللهی پارسا، باقری، آقایی، منتظری	۱۳۹۸	پیشرفته	۶۹۵
مینا ۱	صحرائی، غریبی، ملک‌لو، صادقی، شهباز، سلطانی	۱۳۹۹	پایه	۱۰۹۸
مینا ۲	صحرائی، غریبی، ملک‌لو، صادقی، شهباز، سلطانی	۱۴۰۱	میانی	۱۴۲۴
مینا ۳	صحرائی، غریبی، شهباز، سلطانی، طالبی	۱۳۹۹	میانی	۱۷۸۳

1. optical character recognition (OCR)

نام کتاب	مؤلفان	سال انتشار	سطح کتاب	تعداد واژه
مینا ۴	صحرائی، غریبی، سلطانی، شهباز، طبسی	۱۴۰۰	پیشرفته	۱۰۳۹
شیراز ۱	صحرائی، آقایی، طبسی، رستم‌زاده	۱۴۰۱	مقدماتی	۴۲۶
شیراز ۲	صحرائی، آقایی، طبسی، سلطانی، رضاپور	۱۴۰۰	میانی	۶۶۷
شیراز ۳	صحرائی، آقایی، رضاپور، سلطانی، طبسی	۱۴۰۱	میانی	۶۰۲
پارسا ۱	کولی‌وندی، فاطمی‌منش، پورمند	۱۳۹۷	مقدماتی	۱۱۲۲
پارسا ۲	کولی‌وندی، فاطمی‌منش	۱۳۹۷	میانی	۱۷۶۲
رسا ۱	کولی‌وندی، پورمند	۱۳۹۷	مقدماتی	۴۶۷
رسا ۲	کولی‌وندی، فاطمی‌منش	۱۳۹۷	میانی	۵۳۳
نگارا ۱	فاطمی‌منش، سعیدنیا	۱۳۹۷	مقدماتی	۴۶۳
نگارا ۲	فاطمی‌منش، سعیدنیا	۱۳۹۸	میانی	۹۲۸
خوانا ۱	کولی‌وند، فاطمی‌منش، مختاری، کرمانی	۱۳۹۷	مبتدی	۸۶۵
خوانا ۲	کرمانی، فاطمی‌منش، مختاری، خاکی‌صحنه	۱۳۹۷	میانه	۱۱۰۲
خوانا ۳	مختاری، خاکی‌صحنه، کرمانی، فاطمی‌منش	۱۳۹۸	پیشرفته	۱۲۶۸
مقدمه	پورمند، فاطمی‌منش، مصطفی‌پور	۱۳۹۷	مقدماتی	۲۷۹
نوشتن ۲	سعیدنیا، فاطمی‌منش	۱۳۹۸	میانی	۷۸۱
آموزش نوین زبان فارسی ۱	احسان‌قبول	۱۳۸۸	مبتدی	۱۳۴۶
آموزش نوین زبان فارسی ۲	احسان‌قبول	۱۳۸۸	میانی	۷۳۵
آموزش نوین زبان فارسی ۳	احسان‌قبول	۱۳۸۸	میانی	۹۷۰
آموزش نوین زبان فارسی ۴	احسان‌قبول	۱۳۸۸	میانی	۸۴۶
آموزش نوین زبان فارسی ۵	احسان‌قبول	۱۳۹۴	پیشرفته	۱۰۹۰

۴-۲. آماده‌سازی دادگان

در گام نخست، ابتدا واژگان منابع «آزفا» هر یک به‌صورت دستی در ردیف‌های مجزا از فایل «اکسل» (صفحه گسترده) تایپ شدند. همان‌طور که در قسمت مبانی نظری به واژگان

چند-جزئی اشاره گردید، برخی واژگان در پژوهش اخیر چند-جزئی بودند. بنابراین، این نکته هنگام تایپ کردن آنان لحاظ گردید؛ بدین‌معنا که واژگان چند-جزئی هر کدام به‌عنوان یک واحد در یک ردیف از فایل «اکسل» تایپ شدند. جمع‌آوری دادگان زبانی در بستر رایانه مانند تایپ‌کردن یا کپی‌کردن، بسته به استفاده از صفحه‌کلید متفاوت توسط کاربران، منجربه پیدایش تفاوت‌هایی در کدگذاری حروف می‌شود. بنابراین، در پژوهش حاضر که تایپ کردن واژگان توسط چندین نفر صورت پذیرفته، هنجارسازی متون یک گام اساسی از پیش‌پردازش‌ها به‌شمار می‌رود. پیش‌پردازش‌های رایانشی مهم طبق دیدگاه Indurkhya & Damerau (2010) و Jurafsky & Martin (2024) شامل هنجارسازی، واحدسازی، بن‌واژه‌سازی هستند. برچسب‌دهی اجزای کلام نیز می‌تواند به‌عنوان اطلاعات تکمیلی پس از پیش‌پردازش یا ساخت پیکره اضافه شود. در پژوهش حاضر برچسب‌دهی اجزای کلام پس از پیش‌پردازش و در گام نخست صورت پذیرفته است.

پیش‌پردازش و برچسب‌دهی اجزای کلام

در سال‌های اخیر، ابزارهای رایانشی متعددی مانند «پارسی‌ور»[1]، «استپ‌وان»[2] (Shamsfard, Jafari & Ilbeygi 2010) و ابزار ارائه‌شده توسط (Seraji, Megyesi & Nivre 2012)[3] برای پیش‌پردازش زبان فارسی توسعه یافته‌اند، اما با یکدیگر قابل مقایسه نیستند؛ زیرا بر پیکره‌های متفاوتی ارزیابی شده‌اند. بر خلاف ابزارهای پیشین، عملکرد «استنزا»[4] (Qi et al. 2020) و «دادماتولز»[5] (Etezadi et al. 2022) بر پیکره‌های مشابهی ارزیابی شده و بنابراین، قابل سنجش هستند. در واحدسازی واژگان، ابزار «استنزا» توانسته دقت ۹۹/۹۶ درصد را در پیکره Rasooli, Kouhestani & Moloodi (2013) و دقت ۱۰۰ درصد را در پیکره Seraji (2015) کسب کند. همچنین این ابزار در برچسب‌زنی اجزای کلام روی همین پیکره‌ها به دقت‌هایی معادل ۹۷/۳۵ درصد و ۹۷/۴۳ درصد دست یافته است. طبق بررسی Etezadi et al. (2022)، ابزار «دادماتولز» در واحدسازی واژگان و برچسب‌زنی اجزای کلام در پیکره‌های فوق به عملکرد بالاتری نسبت به ابزار «استنزا» دست یافته است. همچنین، «دادماتولز»

1. ParsiVar (F-Measure = 95%)
2. step-1 (Accuracy = 90.9%)
3. Seraji (Accuracy = 96.9%)
4. Stanford STANZA Package
5. Dadmatools Package

دارای هنجارسازی متن است که در «استنزا» به‌طور مستقل دیده نمی‌شود. در همین راستا، با توجه به عملکرد برتر این دو ابزار، «استنزا» و «دادماتولز» به‌عنوان ابزارهای اصلی پیش‌پردازش در پژوهش حاضر انتخاب گردیدند.

از آنجا که تایپ کردن واژگان توسط چندین نفر صورت گرفته، این است که هنجارسازی در پژوهش حاضر الزامی است. این مهم شامل حذف علائم نگارشی مانند «!» و «؟»، یکسان‌سازی فاصله در واحدهای چند-جزئی مانند «آب‌وهوا» و «آب و هوا» و یکسان‌سازی کدگذاری حروف مانند «ی» و «ک» در صفحه‌کلید فارسی[1] است. همان‌طور که ذکر شد، ابزار «استنزا» قادر به هنجارسازی نبود، و بنابراین، از ابزار «دادماتولز» برای حذف علئم نگارشی، یکسان‌سازی کدگذاری و حذف فاصله‌های غیرضروری استفاده گردید. برخی موارد نگارشی مانند فتحه (ـَ)، ضمه (ـُ) و کسره (ـِ) نیز توسط «دادماتولز» قابل تشخیص نبودند که به‌کمک برنامه‌نویسی «پایتون» توسط پژوهشگران حذف گردیدند. اصلاح خطاهای املایی در تایپ مانند «گفتم» به‌جای «کفتم» و یکسان‌سازی نگارش واژگان چند-املایی مانند مسئول» به‌جای «مسوول» نیز به‌صورت دستی انجام گردید.

در مورد واژگان چند-جزئی نیز که بایستی به‌عنوان یک واحد مستقل واحدسازی شوند، یعنی به اجزای کوچک‌تر شکسته نشوند، دو کار صورت پذیرفت: یکی قبل از پیش‌پردازش و هنگام تایپ‌کردن دستی که هر یک از واحدهای واژگانی چند-جزئی در یک ردیف از فایل‌های «اکسل» تایپ شدند، و یکی هم در مرحله پیش‌پردازش توسط ماژول واحدساز چند-جزئی در «استنزا» و «دادماتولز». «استنزا» و «دادماتولز» هر دو دارای یک بخش واحدساز چند-جزئی[2] برای بررسی واحدهای چند-جزئی زبان فارسی طبق آمار و احتمال همنشینی واژگان هستند تا در هنگام واحدسازی واژگان، آن‌ها را یک واحد در نظر بگیرند. بدین‌صورت در پژوهش حاضر واژگان چند-جزئی به واحدهای تک‌جزئی و منفرد دچار تجزیه نخواهند شد.

همچنین، برای اطمینان از مقایسه‌پذیر بودن صورت‌های واژگانی، نیاز به بن‌واژه‌سازی بعد از واحدسازی احساس می‌گردد. به‌عنوان مثال، برای مقایسه‌پذیر بودن واحد «به خطر انداخت» یا «به خطر می‌اندازد» بایستی هر دو به یک مدخل «به خطر انداختن»

1. UNICODE (u064a and u06a9 versus u0649 and u0643)
2. multi-word tokenizer (MWT)

بن‌واژه‌سازی شوند. در این راستا، به چند دلیل، از ابزار «استنزا» برای بن‌واژه‌سازی استفاده گردید که در ادامه شرح داده می‌شود. در مرحله برچسب‌زنی اجزای کلام نیز با کمک ابزار «استنزا» به هر واحد یا بن‌واژه یک برچسب نحوی اعطا گردید.

گرچه طبق گفته «اعتزادی» و همکاران، «دادماتولز» در برچسب‌زنی اجزای کلام، تجزیه وابستگی و بن‌واژه‌سازی عملکرد بهتری نسبت به «استنزا» نشان داده است، اما در واحدسازی واژگان چند-جزئی در پژوهش حاضر و بن‌واژه‌سازی افعال خطاهای فاحشی نسبت به «استنزا» نشان می‌دهد (Etezadi et al. 2022). به‌عنوان مثال نتیجه واحدسازی کلمه «گفت و گو» توسط «دادماتولز»، دو واحد جدا از هم «#گفت» و «#گو» است، در صورتی که «گفت و گو» یک واحد چند-جزئی است و نباید به اجزای سازنده‌اش تجزیه شود. همچنین، ماژول واحدسازی، بن‌واژه‌سازی و برچسب‌دهی اجزای کلام در «استنزا» به‌هم‌پیوسته هستند. در همین راستا، در پژوهش حاضر بعد از تایپ کردن هر واژه چند-جزئی در یک ردیف، هنجارسازی دستی و هنجارسازی توسط «دادماتولز»، از ابزار «استنزا» برای واحدسازی، بن‌واژه‌سازی و سرانجام، برچسب‌زنی اجزای کلام استفاده گردید. همچنین، ذکر این نکته دارای اهمیت است که شیوه‌نامه برچسب‌زنی اجزای کلام در «استنزا» از نوع وابستگی جهانی[1] است.

لازم به ذکر است که برای اطمینان از دقت برچسب‌زنی ابزار «استنزا»، پنج درصد از واژگان نیمه اول هر سطح انتخاب شده و به‌صورت دستی بازنگری شدند. در این بازنگری، میانگین صحت برچسب‌ها در سه سطح ۸۸ درصد به‌دست آمد که نسبت به دقت گزارش‌شده ابزار بر پیکره‌های نحوی (۹۷ درصد) تفاوت چشمگیری دارد. با بررسی خطاهای رایج برچسب‌زنی این نتیجه حاصل آمد که این ابزار، مصدر افعال چند-جزئی را اسم محسوب می‌کند، و بنابراین، برچسب تمامی مصدرهای افعال به‌صورت دستی اصلاح شدند و برچسب‌زنی پنج درصد از نیمه دوم واژگان در سه سطح بازنگری شد. در بازنگری دوم، میانگین صحت برچسب‌ها در سه سطح ۹۴ درصد به‌دست آمد که حاکی از دقتی نزدیک به دقت ابزار است. در جدول ۲، آمار دقت ابزار قبل و بعد از بازنگری آمده است.

1. universal dependency POS (UPOS)

جدول ۲. دقت برچسب‌زنی اجزای کلام

سطح	دقت ٥ درصد از واژگان پیکره قبل از بازنگری			دقت ٥ درصد دیگر از واژگان بعد از بازنگری		
	برچسب صحیح	کل برچسب‌ها	درصد صحت	برچسب صحیح	کل برچسب‌ها	درصد صحت
مبتدی	۱۶۶	۱۸۶	۸۹/۲۴	۱۷۷	۱۸۶	۹۵/۱۶
میانه	۲۹۳	۳۳۵	۸۷/۴۶	۳۱۵	۳۳۵	۹۴/۰۲
پیشرفته	۱۸۵	۲۱۰	۸۸/۰۹	۱۹۷	۲۱۰	۹۳/۸۰
میانگین	–	–	۸۸/۲۶	–	–	۹۴/۳۲

در تحلیل خطاهای برچسب‌زنی هر جزء کلام، بالاترین میزان خطا مربوط به برچسب‌دهی جزء کلام فعل بوده (تقریباً ۷۲ درصد)، و باقی خطاها مربوط به تشخیص صحیح برچسب اسم خاص، حرف اضافه، اعداد و قیدها بوده است. در مورد خطای افعال، همگی در افعال مرکب (چند-جزئی) مشاهده شدند که تعدادی از آن‌ها به‌اشتباه اسم تلقی شده بودند و تعدادی دیگر به‌اشتباه برچسب حرف اضافه دریافت کرده بودند. به‌احتمال، علت این خطا به برچسب جزء ابتدایی فعل برمی‌گردد که در مثال‌هایی مانند «غذا خوردن» برچسب اسمی و «به کار گرفتن» برچسب حرف اضافه دریافت کرده است. در این راستا، تمامی افعال پیکره به‌صورت دستی بازنگری شدند و اصلاح گردیدند. در مورد اسم خاص نیز مواردی به‌اشتباه اسم تلقی شدند که چون هر دو نوعی از اسم بودند، از اصلاح آن‌ها در کل پیکره چشم‌پوشی شد. در مورد خطای اعداد نیز تعداد اندکی مشاهده گردید که به‌اشتباه اسم تشخیص داده شده بودند که به‌علت میزان کم بودن آن‌ها از بین سایر خطاها، اصلاح آن‌ها در کل پیکره نادیده گرفته شد. همچنین، خطای حروف اضافه در موارد انگشت‌شماری مشاهده گردید که به‌اشتباه برچسب حرف ربط و همچنین حرف تعریف دریافت کرده بودند و تعدادی قید نیز به‌اشتباه برچسب حرف اضافه گرفته بودند. از اصلاح دستی این‌ها نیز در کل پیکره چشم‌پوشی شد. لازم به ذکر است که در بازنگری دوم از پنج درصد نیمه دوم پیکره یعنی بعد از اصلاح دستی نیز خطای جدیدی در برچسب‌زنی مشاهده نگردید و خطاها در اسامی خاص، حروف اضافه، اعداد و قیدها مشاهده شدند

٤-۳. مقایسه همپوشانی واژگان

در این گام، با توجه به اینکه تمامی واژگان یک‌دست و یکسان‌سازی شدند، منابع مرتبط

با هر سطح خاص از چارچوب مشترک اروپا مانند مبتدی، میانه یا پیشرفته به‌طور مستقل برای مقایسه و تحلیل انتخاب شدند. به بیان شفاف‌تر، در مقایسه سطح مبتدی، واژگان ۹ کتاب «شیراز ۱»، «رسا ۱»، «نگارا ۱»، «پارسا ۱»، «خوانا ۱»، «مقدمه»، «مینا ۱»، «پرفا ۱»، «و آموزش نوین فارسی ۱» استخراج شده و با هم مقایسه شدند تا واژگان همپوشان مشخص گردند. در سطح میانه واژگان ۱۲ کتاب «شیراز ۲ و ۳»، «رسا ۲»، «نگارا ۲»، «پارسا ۲»، «خوانا ۲»، «مینا ۲ و ۳»، «پرفا ۲»، و «آموزش نوین فارسی ۲ و ۳ و ۴» استخراج شده و با هم مقایسه شدند. در سطح پیشرفته نیز واژگان ۵ کتاب «نگارا ۳»، «خوانا ۳»، «مینا ۴»، «پرفا ۳»، و «آموزش نوین فارسی ۵» استخراج شده و با هم مقایسه شدند. برای مقایسه واژگان همپوشان از ساختار داده مجموعه[1] در زبان «پایتون» استفاده گردید. مزیت استفاده از این ساختار داده در پژوهش حاضر آن است که محتویات مجموعه در «پایتون» غیرقابل تغییر در نظر گرفته شده و در نتیجه، هر آنچه بین مجموعه‌ها کاملاً یکسان بوده، همپوشان محسوب شده است. برای این منظور، یک کُد به زبان «پایتون» نوشته شد که برای ۹ کتاب سطح مبتدی، ۹ مجموعه از بن‌واژه (با حفظ واژگان چند-جزئی به‌عنوان یک واحد) تشکیل می‌دهد. سپس، همپوشانی بین مجموعه‌ها را محاسبه نموده و دامنه اشتراک یا همپوشانی[2] را گزارش می‌دهد. همین کُد برای ۱۲ کتاب میانه و ۵ کتاب پیشرفته نیز پیاده شده تا مقایسه و تحلیل انجام پذیرد. همچنین، تفاضل اشتراک و اجتماع مجموعه بن‌واژه‌ها در هر سطح محاسبه گردید تا بن‌واژه‌های یکتا نیز مشخص گردند. خروجی این کُد همچنین مشخص می‌کند که بن‌واژه‌های همپوشان در چه تعداد کتاب رخداد داشته‌اند.

٤-٤. ماشین‌خوان[3] کردن پیکره

طبق استانداردهای ساخت پیکره که در قسمت مبانی پژوهش نیز ذکر گردید، در ساخت پیکره ملزوماتی از جمله ماشین‌خوان بودن، اندازه، نمایندگی و توازن دارای اهمیت هستند و به برچسب‌گذاری نیز بنا بر هدف پژوهش از حیث ارائه اطلاعات تکمیلی توجه می‌شود. از آنجا که کلیه واژگان در نظر گرفته شده بودند، پیکرهٔ پژوهش حاضر معیارهای اندازه، نمایندگی و توازن را داراست. همچنین، اطلاعات دستوری در قالب برچسب اجزای

1. set data structure
2. vocabulary overlap range
3. machine-readable

کلام به هر کدام از واژگان پیکره اضافه گردید. تنها معیار لازم، ماشین‌خوان بودن پیکره است که بتواند واژگان پیکره پژوهش حاضر را همراه با اطلاعات برچسب‌های اجزای کلام در قالب یک استاندارد ماشین‌خوان به پژوهشگران ارائه کند. اکنون دادگان پژوهش که واژگان همپوشان در هر سطح بودند، همراه با برچسب اجزای کلام آنان در دست است که برای تبدیل آنان به قالب ماشین‌خوان باید از یک زبان نشانه‌گذاری استفاده نمود. برای این منظور از زبان XML استفاده گردید، زیرا XML یک زبان استاندارد در زبان‌شناسی پیکره‌ای و ساخت پیکره‌های متنی به‌شمار می‌رود (McEnery & Brookes 2022)، و ساختاری سلسله‌مراتبی برای نمایش اطلاعات به‌دست می‌دهد که برای خوانش توسط انسان نیز پیچیدگی بالایی ندارد (Ide & Sperberg-McQueen, 2023). این زبان داده‌ها را در ساختار براکت‌مانند[1] خاصی ذخیره می‌کند تا اطلاعات بدون تغییر از یک رایانه به رایانه دیگر منتقل شوند و امکان جست‌وجو در داده‌ها توسط رایانه یا ماشین فراهم گردد. در این راستا، به کمک برنامه‌نویسی «پایتون» و «ابزار درخت ایکس‌ام‌ال»[2] پیکره نهایی ماشین‌خوان شد. این پیکره سه سطح مبتدی، میانه و پیشرفته دارد و در هر سطح مطابق با شکل ۲، شامل اطلاعات واژگان است. این اطلاعات حرف الفبا، واحد، بن‌واژه، دامنه همپوشانی، برچسب جزء کلام و فراداده کتاب‌های منبع هستند که به فرمت استاندارد ماشین‌خوان XML در یک ساختار سلسله‌مراتبی نشانه‌گذاری شده‌اند

۵. ارائه یافته‌ها

در این قسمت ابتدا ابر بسامدی واژگان و نسبت نوع‌واژه به واحد[3] در هر سطح مقدماتی، میانه و پیشرفته با توجه به واژگان تایپ‌شده توسط نرم‌افزار «انت‌کانک»[4] ارائه شده است. سپس همپوشانی واژگان در هر سطح (مبتدی، میانه و پیشرفته) محاسبه و گزارش شده است

۵-۱. ابر بسامدی واژگان

همان‌طور که مشاهده می‌شود، کلیه واژگان تایپ‌شده در سطح مبتدی شامل ۳۱۸۹ نوع‌واژه و ۵۵۴۲ واحد هستند که نسبت نوع‌واژه به واحد آنان ۵۷/۵ درصد است. واژگان سطح میانه

1. bracket-shaped
2. XLM element tree
3. type / token ratio (TTR)
4. AntConc 4.3.1

دارای ۵۳۷۶ نوع‌واژه و ۱۱۲۰۵ واحد است که نسبت نوع‌واژه به واحد آنان ۴۷/۹ درصد است. در سطح پیشرفته نیز ۳۹۶۲ نوع‌واژه و ۶۶۵۶ واحد محاسبه شده که نسبت نوع‌واژه به واحد آنان ۵۹/۵ درصد را به‌دست می‌دهد. با محاسبه نسبت نوع‌واژه به واحد واژگان تایپ‌شده سطح پیشرفته بالاترین و سطح میانه کمترین غنای واژگانی را بین سه سطح نشان داده‌اند. لازم به ذکر است که این نسبت فقط از روی واژگان تایپ‌شده محاسبه گردیده و نمی‌تواند معیار دقیقی از غنای واژگانی منابع هر سطح باشد، زیرا برای این منظور به متن کامل کتاب‌ها نیاز است تا معیار بسامد و تکرار واژگان به‌دقت گزارش شود. همچنین ابرهای بسامدی صرفاً از واژگان تک-جزئی (بدون پردازش واحدهای چند-جزئی) ترسیم گردیده‌اند تا یک شمای کیفی از واژگان پربسامد و در نتیجه، مفاهیم پربسامد در هر سطح ارائه دهند که در شکل ۱، قابل مشاهده است

شکل ۱. ابر بسامدی ۱۵ واژه پربسامد در هر سطح (از راست به چپ: مقدماتی، میانه و پیشرفته)

۵-۲. همپوشانی واژگان

۵-۲-۱. سطح مبتدی

در محاسبه همپوشانی واژگان با احتساب واحدهای چند-جزئی پردازش شده و در سطح مبتدی، نتایج زیر مطابق جدول ۳، به‌دست آمد. در حدود ۶۴ درصد از واژگان سطح مبتدی همپوشانی بین منابع نداشتند و فقط دارای دامنه وقوع ۱ بودند؛ یعنی فقط در یک کتاب وقوع داشتند. در حدود ۳۶ درصد از واژگان این سطح دارای همپوشانی بودند که البته تعداد واژگان همپوشان بین کلیه ۹ کتاب این سطح ۷ عدد بوده که نسبت به کلیه واژگان یعنی ۳۷۱۰ نسبت ناچیزی در حدود ۰/۲ درصد را دارا هستند.

جدول ۳. نتیجه همپوشانی واژگان سطح مبتدی

	دامنه ۱	دامنه ۲	دامنه ۳	دامنه ٤	دامنه ٥	دامنه ٦	دامنه ۷	دامنه ۸	دامنه ۹
واژگان	۲۳۷۱	۵۷۴	۳۰۹	۱۶۰	۱۲۷	۸۸	۵۱	۲۳	۷

۵-۲-۲. سطح میانه

در محاسبه همپوشانی واژگان در این سطح نیز واحدهای چند-جزئی لحاظ شده است که در حدود ۶۳/۵ درصد از واژگان سطح میانه همپوشانی بین منابع نداشته‌اند و فقط دارای دامنه وقوع ۱ بودند؛ یعنی فقط در یک کتاب وقوع داشتند. در حدود ۳۶/۵ درصد از واژگان این سطح دارای همپوشانی بودند که البته واژگان همپوشان در ۱۲ کتاب رؤیت نشد و تنها بین ۹ کتاب و کمتر همپوشانی مشاهده گردید که در جدول ۴، آمده است

جدول ٤. نتیجه همپوشانی واژگان سطح میانه

	دامنه ۱	دامنه ۲	دامنه ۳	دامنه ٤	دامنه ٥	دامنه ٦	دامنه ۷	دامنه ۸	دامنه ۹
واژگان	۴۲۴۵	۱۲۳۵	۵۸۶	۳۴۶	۱۷۴	۸۴	۲۲	۳	۱

۵-۲-۳. سطح پیشرفته

در محاسبه همپوشانی واژگان سطح پیشرفته نیز واحدهای چند-جزئی پردازش شده و نتایج مطابق با جدول ۵، به‌دست آمد. در حدود ۸۷ درصد از واژگان سطح پیشرفته، همپوشانی بین منابع نداشتند و فقط دارای دامنه وقوع ۱ بودند؛ یعنی فقط در ۱ کتاب وقوع داشتند. در حدود ۱۳ درصد از واژگان این سطح دارای همپوشانی بودند که البته واژگان همپوشان در ۵ کتاب رؤیت نشد و تنها بین ۴ کتاب و کمتر همپوشانی مشاهده گردید.

جدول ۵. نتیجه همپوشانی واژگان سطح پیشرفته

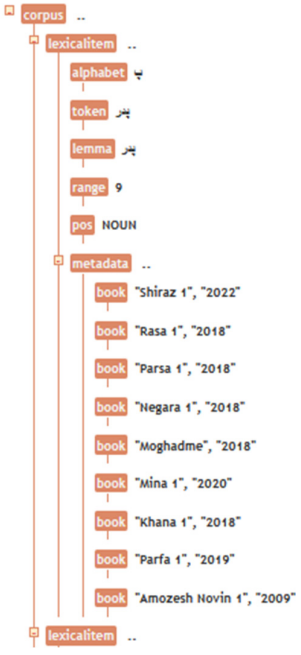
	دامنه ۱	دامنه ۲	دامنه ۳	دامنه ٤	دامنه ٥	دامنه ٦	دامنه ۷	دامنه ۸	دامنه ۹
واژگان	۳۶۴۸	۴۴۶	۷۴	۸	–	–	–	–	–

۵-۳. پیکره پژوهش

با توجه به اطلاعات موجود، پیکره پژوهش در قالب XML مطابق با ساختار شکل ۲، طراحی شد و توسعه یافت. در این شکل در سمت چپ، شمای منطقی پیکره و در سمت راست، شمای کد یا نشانه‌گذاری XML دیده می‌شود. پیکره شامل سطح زبان‌آموزی،

واحد واژگانی، بن‌واژه، دامنه همپوشانی، حرف الفبا، صورت واژه، برچسب جزء کلام، و فراداده کتاب‌های منبع (نام کتاب و سال چاپ) است

```xml
<?xml version='1.0' encoding='UTF-8'?>
<Corpus>
  <LexicalItem>
    <Alphabet>پ</Alphabet>
    <Token>پدر</Token>
    <Lemma>پدر</Lemma>
    <Range>9</Range>
    <POS>NOUN</POS>
    <Metadata>
      <Book>"Shiraz 1", "2022"</Book>
      <Book>"Rasa 1", "2018"</Book>
      <Book>"Parsa 1", "2018"</Book>
      <Book>"Negara 1", "2018"</Book>
      <Book>"Moghadme", "2018"</Book>
      <Book>"Mina 1", "2020"</Book>
      <Book>"Khana 1", "2018"</Book>
      <Book>"Parfa 1", "2019"</Book>
      <Book>"Amozesh Novin 1", "2009"</Book>
    </Metadata>
  </LexicalItem>
```



شکل ۲. شمای پیکره توسعه‌یافته در پژوهش

همچنین، برای خوانش پیکره و جست‌وجو[1] در آن، یک کُد «پایتون» نوشته شد که اجازه می‌دهد واژگان در سطوح خاص، واژگان در دامنه همپوشانی خاص و همچنین برچسب‌های دستوری خاص جست‌وجو شوند. بدین روش، اطلاعات پنج جزء کلام با بیشترین تعداد همپوشانی در پیکره پژوهش در هر سطح استخراج گردید و در ادامه ارائه می‌شود.

٥-٤. پنج جزء کلام با بیشترین تعداد همپوشانی در هر سطح

پنج نقش دستوری با بیشترین تعداد همپوشانی در سطح مبتدی شامل این واژگان در هر دامنه هستند

با دامنه همپوشانی ۱: ۱۸۲۲ اسم، ۴۱۸ فعل، ۳۲۸ صفت، ۱۴۹ اسم خاص، ۷۵ عدد

با دامنه همپوشانی ۲: ۴۰۷ اسم، ۹۴ فعل، ۷۴ صفت، ۲۵ اسم خاص، ۱۶ عدد

با دامنه همپوشانی ۳: ۲۲۲ اسم، ۴۵ فعل، ۲۸ صفت، ۱۲ اسم خاص، ۱۰ حرف اضافه

با دامنه همپوشانی ۴: ۱۱۲ اسم، ۲۶ فعل، ۱۴ صفت، ۵ حرف اضافه، ۴ اسم خاص، ۴ عدد

با دامنه همپوشانی ۵: ۸۴ اسم، ۱۲ فعل، ۱۰ صفت، ۱۰ اسم خاص، ۴ عدد، ۴ قید

با دامنه همپوشانی ۶: ۵۶ اسم، ۱۳ صفت، ۵ عدد، ۵ اسم خاص، ۳ فعل

با دامنه همپوشانی ۷: ۳۹ اسم، ۵ صفت، ۲ ضمیر، ۱ عدد، ۱ اسم خاص

با دامنه همپوشانی ۸: ۲۰ اسم، ۲ فعل، ۱ صفت

با دامنه همپوشانی ۹: ۶ اسم، ۱ حرف اضافه

پنج نقش دستوری با بیشترین تعداد همپوشانی در سطح میانه شامل این واژگان در هر دامنه هستند:

با دامنه همپوشانی ۱: ۳۰۰۱ اسم، ۱۱۶۸ فعل، ۶۳۲ صفت، ۳۱۸ اسم خاص، ۷۶ قید

با دامنه همپوشانی ۲: ۸۷۰ اسم، ۲۵۸ فعل، ۱۴۱ صفت، ۲۴ اسم خاص، ۱۷ قید

با دامنه همپوشانی ۳: ۴۱۸ اسم، ۹۲ فعل، ۵۶ صفت، ۱۵ قید، ۱۰ حرف اضافه

با دامنه همپوشانی ۴: ۲۲۸ اسم، ۴۴ فعل، ۴۲ صفت، ۱۸ قید، ۵ حرف اضافه

با دامنه همپوشانی ۵: ۱۱۹ اسم، ۲۱ فعل، ۱۹ صفت، ۸ قید، ۳ حرف اضافه

با دامنه همپوشانی ۶: ۵۰ اسم، ۲۰ صفت، ۵ قید، ۷ فعل، ۱ حرف ربط

با دامنه همپوشانی ۷: ۱۵ اسم، ۴ صفت، ۱ فعل، ۱ حرف اضافه، ۱ قید

1. query

با دامنه همپوشانی ۸: ۳ اسم
با دامنه همپوشانی ۹: ۱ اسم

پنج نقش دستوری با بیشترین تعداد همپوشانی در سطح پیشرفته شامل این واژگان در هر دامنه هستند:
با دامنه همپوشانی ۱: ۲۷۰۳ اسم، ۷۷۱ فعل، ۴۸۹ صفت، ۱۸۸ اسم خاص، ۴۵ قید
با دامنه همپوشانی ۲: ۳۲۵ اسم، ۶۸ فعل، ۵۵ صفت، ۱۰ حرف اضافه، ۶ قید
با دامنه همپوشانی ۳: ۶۷ اسم، ۴ صفت، ۶ فعل، ۲ اسم خاص، ۱ ضمیر
با دامنه همپوشانی ۴: ۵ اسم، ۱ صفت، ۱ فعل، ۱ حرف ربط

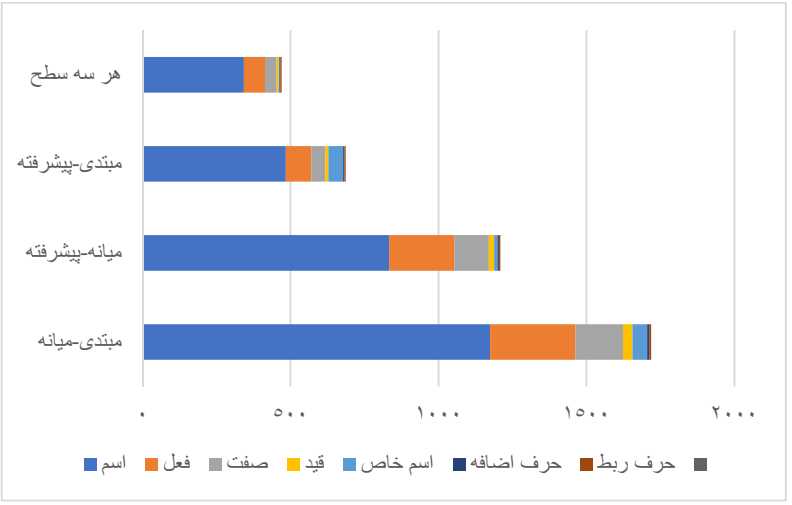
به‌طور کلی، از بین واژگان همپوشان در سطح مبتدی ۲۳۴۲ اسم، ۵۵۶ فعل و ۲۹۲ صفت مشاهده شد. در سطح میانه ۳۸۶۷ اسم، ۱۴۳۷ فعل و ۵۶۲ صفت مشاهده شد. در سطح پیشرفته نیز ۲۶۵۷ اسم، ۷۸۵ فعل و ۳۳۴ صفت به‌عنوان پرتکرارترین اجزای کلام مشاهده شد. در چارک پایین سطح مبتدی (دامنه یک و دو) بیشترین تعداد اجزای کلام متعلق به اسم، فعل، صفت و عدد بود. در چارک بالای سطح مبتدی (دامنه هشت و نه)، بیشترین تعداد اجزای کلام متعلق به اسم، فعل، صفت و حرف اضافه بود. در چارک پایین سطح میانه (دامنه یک و دو) بیشترین تعداد اجزای کلام متعلق به اسم، فعل، صفت، اسم خاص و قید بود. در چارک بالای سطح میانه (دامنه هشت و نه)، بیشترین تعداد اجزای کلام متعلق به اسم بود. در چارک پایین سطح پیشرفته (دامنه یک) بیشترین تعداد اجزای کلام متعلق به اسم، فعل، صفت، اسم خاص و قید بود. در چارک بالای سطح پیشرفته (دامنه چهار)، بیشترین تعداد اجزای کلام متعلق به اسم، فعل، صفت و حرف ربط بود

۵-۵. همپوشانی اجزای کلام بین سطوح مختلف

گرچه هدف این پژوهش بررسی همپوشانی واژگان در راستای طراحی و ساخت یک پیکره واژگانی از آن‌ها در هر سطح از منابع «آزفا» بود، اما می‌توان همپوشانی واژگان بین سطوح مختلف را نیز در هر جزء کلام ارزیابی نمود. در همین راستا یافته‌های زیر به‌دست آمد

واژگان همپوشان بین سطح مبتدی و میانه ۱۸۱۰ عدد، و شامل ۱۱۷۴ اسم (۶۶ درصد)، سپس، ۲۸۹ فعل (۱۵ درصد) و ۱۶۰ صفت (۸ درصد) بودند. بقیه شامل ۲۲ حرف اضافه، ۷ حرف ربط، ۳۲ قید و ۵۰ اسم خاص بود. واژگان همپوشان بین سطح میانه و پیشرفته ۱۲۴۰ عدد، و شامل ۸۳۳ اسم (۶۶ درصد) و سپس، ۲۲۰ فعل (۱۵ درصد) و ۱۱۷ صفت (۸ درصد)

بـود. بقیـه شـامل ۳۷ حـرف اضافـه، ۵ حـرف ربـط، ۱۸ قیـد و ۱۱ اسـم خـاص بـود. واژگان همپوشـان بیـن سـطح مبتـدی و پیشـرفته ۶۶۰ عـدد، و شـامل ۴۸۴ اسـم (۶۶ درصـد) و سـپس، ۸۶ فعـل (۱۵ درصـد) و ۴۸ صفـت ۸ درصـد) بـود. بقیـه شـامل ۱۵ حـرف اضافـه، ۴ حـرف ربـط، ۱۰ قیـد و ۵۰ اسـم خـاص بـود. واژگان همپوشـان در هـر سـه سـطح ۴۸۵ عـدد، و شـامل ۳۴۲ اسـم (۷۰ درصـد) و سـپس، ۷۲ فعـل (۱۵ درصـد) و ۳۹ صفـت (۸ درصـد) بـود. بقیـه شـامل ۱۵ حـرف اضافـه، ۸ حـرف ربـط، ۷ قیـد و ۲ اسـم خـاص بـوده. یافته‌هـای ایـن قسـمت در نمـودار ۱، قابـل مشـاهده اسـت



نمودار ۱. همپوشانی اجزای کلام بین سطوح مختلف

همان‌طـور کـه در نمـودار ۱، مشـخص اسـت، همپوشـانی بـالایی بیـن سـطوح مختلـف در اجـزای کلام اسـم، فعـل، صفـت و اسـم خـاص دیـده می‌شـود. به‌طـور کلی، بیشـترین میـزان همپوشـانی در سـطح مبتـدی بـا میانـه دیـده شـده و کمتریـن میـزان همپوشـانی بیـن هـر سـه سـطح مشـاهده می‌شـود. نکتـه جالـب توجـه در شـکل فـوق، همپوشـانی بـالای اسـامی خـاص در سـطح مبتـدی بـا میانـه و همچنیـن سـطح مبتـدی بـا پیشـرفته اسـت کـه ایـن تناظـر بیـن سـطح میانـه بـا پیشـرفته دیـده نمی‌شـود.

٦. تحلیل یافته‌ها

همان‌طـور کـه در یافته‌هـای مربـوط بـه ابـر بسـامدی واژگان اشـاره گردیـد، می‌تـوان از نسـبت

نوع‌واژه به واحد پیرامون غنای واژگانی بحث نمود. با افزایش سطح از مبتدی به میانه، انتظار افزایش این نسبت می‌رود، اما طبق یافته‌ها کاهش آن (از ۵۷/۵ درصد به ۴۷/۹ درصد) مشاهده شده است. این یافته مغایر با افزایش غنای واژگانی در سطح میانه نسبت به سطح مبتدی است و نشان‌دهنده افزایش تکراری بودن واژگان در سطح میانه است. از سوی دیگر، این نسبت در سطح پیشرفته افزایش داشته (۵۹/۵ درصد) که حاکی از کاهش تکراری بودن واژگان و افزایش مجدد غنای واژگانی است.

با بررسی پراستفاده‌ترین اجزای کلام در هر سطح می‌توان نتیجه گرفت که به‌طور کلی، اسم‌ها، افعال و صفت‌ها در همه سطوح پرتکرارترین هستند. البته بررسی میزان نسبتاً اندک همپوشانی واژگان در هر سطح (مبتدی حدود ۳۶ درصد، میانه حدود ۳۶/۵ درصد و پیشرفته حدود ۱۳ درصد) نشان می‌دهد که مؤلفان نسبت به تکرار و یادآوری واژگان، توجه ویژه‌ای مبذول نداشته‌اند. این مهم به‌ویژه در سطح مبتدی که نیاز به تکرار بیشتر واژگان دارد، مورد غفلت واقع شده است. علت همپوشانی اندک واژگان در سطح پیشرفته نیز به‌احتمال، دال بر کاربست بیشتر واژگان تخصصی است که نیاز به بافت خاصی داشتند و در نتیجه، شرایط و امکان تکرار نمودن آن‌ها را در بین کتاب‌ها محدودتر نموده است.

یک جنبه قابل سنجش دیگر از یافته‌های پژوهش اخیر، بررسی اجزای کلام است. در بررسی اجزای کلام بین سطوح مختلف چنانکه مشاهده گردید، پرتکرارترین مقوله‌ها شامل اسم، فعل و صفت هستند. با توجه به این نکته چنین دریافت می‌شود که تعداد صفت‌ها در تمامی سطوح کمتر از تعداد افعال است. Cervetti et al. (2015) معتقدند که اسم و صفت نسبت به فعل واژگان ساده‌تری محسوب می‌شوند. بنابراین انتظار می‌رود که این موارد در سطوح مبتدی واژه‌آموزی بیشتر از فعل دیده شوند. بنابراین انتظار می‌رود در هر سطحِ بالاتر نسبت به سطحِ پایین‌تر آن، افزایش جزء کلام فعلی و کاهش جزء کلام اسمی و صفتی مشاهده گردد، اما در هر سطح مقوله‌های اسمی همواره بیشتر از فعل بوده و مقوله‌های فعلی نیز همواره بیشتر از صفت بودند.

همان‌طور که در بررسی اجزای کلام واژگان همپوشان در هر سطح نیز اشاره گردید، بیشترین میزان همپوشانی (چارک آخر؛ دامنه نه و هشت) در سطح مبتدی شامل اسم، حرف اضافه و فعل بود. در صورتی که در سطح میانه، بیشترین میزان همپوشانی (چارک آخر؛ دامنه نه و هشت) تنها شامل اسم است که نشان از تکرار پایین‌تر سایر مقوله‌ها در این سطح نسبت به سطح مبتدی دارد. از سوی دیگر، در سطح پیشرفته بیشترین میزان

همپوشانی (چارک آخر؛ دامنه چهار) شامل مقوله‌های اسم، صفت، فعل و حرف ربط است. طبق بررسی Cervetti et al. (2015) پیرامون بالاتر بودن میزان اسم و صفت در سطوح مبتدی، نه‌تنها در پرتکرارترین واژگان سطح میانه نسبت به سطح مبتدی، مقوله‌های فعلی بیشتری دیده نشده که برعکس، تنها مقوله اسم که ساده‌ترین است، پرتکرارترین بوده است. این یافته می‌تواند بر عدم دقت نظر پیرامون تکرار نمودن مقوله پیچیده فعل در سطح میانه تأکید کند.

طبق آرای Nagy & Hiebert (2011) و Cervetti et al. (2015) بسامد واژگان توأم با طول کلمات از مهم‌ترین معیارهای سنجش دشواری واژگان و انتخاب آنان در هر سطح هستند. طول کلمات با بسامد وقوع رابطه‌ای نسبتاً معکوس دارد و انتظار می‌رود طول کلمات در سطوح ابتدایی واژه‌آموزی کمتر از سطوح پیشرفته و بسامد واژگان در سطوح ابتدایی بالاتر از سطوح پیشرفته باشد. در پژوهش حاضر به‌علت محدودیت زمانی فقط طول کلمات سنجیده شد. میانگین طول کلمات در سطح مبتدی ۴/۷۹، در سطح میانه ۵/۲۵ و در سطح پیشرفته ۵/۵۰ حرف محاسبه گردید. نظر به میانگین طول کلمات هر سطح و بنا بر بحث Cervetti et al. (2015)، در هر سطح شاهد افزایش طول کلمات نسبت به سطح قبلی هستیم و این یافته می‌تواند دال بر کاربست واژگان از ساده به پیچیده در سطوح منابع «آزفا» باشد.

۷. نتیجه‌گیری

پژوهش حاضر با هدف بررسی همپوشانی واژگانی در سطوح مبتدی، میانه و پیشرفته منابع آموزشی به‌روز در مراکز «آزفا»ی داخل کشور و تدوین پیکره‌ای سطح‌بندی‌شده از واژگان با رویکردی پیکره‌آگاه و روشی رایانشی صورت پذیرفت که در نتیجهٔ آن، اولین پیکره واژگانی سطح‌بندی‌شده در حوزه «آزفا» توسعه یافت. نتایج پژوهش به‌طور کلی، به غنای واژگانی، همپوشانی واژگان کتاب‌های یک سطح، همپوشانی اجزای کلام در یک سطح، اجزای کلام پرتکرار در هر سطح، همپوشانی اجزای کلام بین سطوح مختلف و سرانجام، طول کلمات اشاره دارند.

بررسی غنای واژگانی نشان داد که احتمالاً در سطح میانه، واژگان تکراری بیشتری نسبت به سطوح مبتدی و پیشرفته معرفی گردیده است. علت این امر را می‌توان به سطح دشوارتر این واژگان و در نتیجه، نیاز به تکرار بیشتر آنان در این سطح نسبت داد. در تحلیل همپوشانی واژگان منابع در هر سطح نیز روندی متغیر دیده شد؛ به نحوی که در سطح

پیشرفته میزان همپوشانی واژگان کاهش نسبتاً شدیدی نسبت به سطوح پیشین داشته است که می‌توان علت آن را در عدم هماهنگی در انتخاب واژگان این سطح جست‌وجو نمود. به‌گفته دیگر، زمانی که بین مؤلفان سطح پیشرفته توافق نظر بالایی نسبت به انتخاب واژگان باشد، بایستی همپوشانی منابع در این سطح نیز بالاتر رود.

بررسی اجزای کلام در پیکره نشان داد که مقوله اسم، فعل و صفت پرتکرارترین اجزای کلام در تمامی سطوح هستند. به‌طور کلی، انتظار می‌رفت که بر اساس نظر Cervetti et al. (2015) در سطوح مبتدی، مقوله‌های اسم و صفت غالب باشند و با پیشرفت به سطوح بالاتر، مقوله فعل افزایش یابد. اما یافته‌ها نشان داد که در تمامی سطوح، اسم غالب است و فعل و صفت به‌ترتیب، در رتبه‌های بعدی قرار دارند. در سطح میانی، بر خلاف انتظار، نه‌تنها تعداد افعال افزایش نیافت، بلکه تنها اسم در دامنه‌های بالای همپوشانی غالب بود. این عدم توجه به تکرار افعال پیچیده‌تر می‌تواند یادگیری را در سطح میانی مختل کند. بنابراین، برای بهبود تألیف منابع آموزشی پیشنهاد می‌شود که مؤلفان به افزایش تکرار و تنوع افعال در سطح میانی توجه بیشتری کنند و از واژگان تخصصی‌تر در سطح پیشرفته به‌گونه‌ای استفاده کنند که امکان تکرار آن‌ها فراهم شود.

سنجش طول کلمات نشان داد که این مؤلفه از سطح مبتدی به پیشرفته روندی افزایشی داشته است. این امر با نظر Cervetti et al. (2015) که طول واژگان را معیاری برای پیچیدگی واژگان می‌دانند، در یک راستا قرار دارد. این نتیجه حاکی از آن است که منابع «آزفا» از واژگان ساده‌تر در سطوح ابتدایی به واژگان پیچیده‌تر در سطوح بالاتر حرکت کرده‌اند که رویکردی مناسب برای پیشرفت یادگیری است. با این حال، برای تعادل مناسب‌تر، بهتر است بسامد واژگان نیز بررسی شود تا اطمینان حاصل شود که واژگان طولانی‌تر با تکرار کافی همراه هستند

همان‌گونه که در این پژوهش مشاهده گردید، محدودیت‌هایی مانند اشتباه در تشخیص نقش‌های دستوری افعال در ابزارهای رایانشی وجود داشت. بنابراین، پیشنهاد می‌شود پژوهشگران زبان‌شناسی رایانشی به بهبود ابزارهای پردازش رایانشی زبان فارسی بپردازند تا شاهد دقت و عملکرد بهتری در این زمینه باشیم. نکته‌ای که به نظر پژوهشگرانِ مقاله کنونی رسید، کاهش دقت این ابزارها در شرایط فقدان بافت کامل واژگان بود؛ بدین‌معنا که واژگان گرچه چند-جزئی تایپ شده بودند، اما در بافت جمله یا متن به‌کار نرفته بودند. «استنزا» و «دادماتولز» هر دو ابزارهای آماری پردازش زبان هستند. بهبود دقت

ابزارهـای رایانـشی می‌توانـد بـه روش تلفیـقی یعـنی قاعده‌بنیـاد در کنـار آمـاری انجـام شـود؛ به‌گونـه‌ای کـه اگـر در شـرایطی ماننـد ایـن پژوهـش دسـترسی بـه بافـت کامـل متـون ممکـن نبـود، ابـزار بتوانـد بـا قواعـد ازپیش‌تعریف‌شـده به‌درسـتی عمـل کنـد. در پژوهـش حاضـر متـن کامـل کتاب‌هـا به‌رغـم درخواسـت پژوهشـگران، در دسـترس آن‌هـا قـرار نگرفـت. بنابرایـن، یـکی از چالش‌هـای اسـاسی، عـدم دسـترسی بـه فایـل منابـع اسـت؛ گرچـه متـن کامـل کتاب‌هـا از نظـر مؤلفـان یـا ناشـران نبایـد در دسـترس عمـوم قـرار گیـرد، بـا ایـن حـال، بهتـر اسـت دسـت کـم بـرای پژوهش‌هـای دانشـگاهی ایـن امـکان فراهـم شـود. ایـن اسـت کـه در ایـن زمینـه پیشـنهاد می‌شـود در کارهـای آتی ضمـن پرداختـن بـه بهبـود ابزارهـای رایانـشی و بهبـود ابزارهـای نویسـه‌خوان نـوری[1]، متـن کامـل کتاب‌هـا تحلیـل شـوند.

بـا توجـه بـه اینکـه در منابـع آمـوزش فـارسی بـه غیرفارسی‌زبانـان، هـم صـورت رسـمی واژگان و اصطلاحـات لحـاظ شـده‌اند و هـم صـورت محـاوره و غیررسـمی، نظـر بـه محدودیـت زمـانی در انجـام پژوهـش کنـونی، فرصـت لحـاظ نمـودن صـورت غیررسـمی کلمـات در پیکـره فراهـم نشـد. بنابرایـن، پیشـنهاد می‌شـود در پژوهش‌هـای آتی بـه ایـن موضـوع، یعـنی صورت‌هـای غیررسـمی واژگان نیـز توجـه شـود. افـزون بـر ایـن، همان‌طـور کـه مشـاهده شـد در ایـن پژوهـش مؤلفـه طـول کلمـات در منابـع آمـوزشی در سـطوح مختلـف زبان‌آمـوزی سـنجیده شـد و توجـه بـه سـایر مؤلفه‌هـا کـه در نظریـه انتخـاب واژگان Nagy & Hiebert (2011) و صـورت بسـط‌یافته آن کـه توسـط Cervetti et al. (2015) مطـرح شـد، از اهمیـت بسـیاری برخـوردار اسـت. به‌عنـوان مثـال، در سـنجش بسـامد واژگان می‌تـوان کلمـات هـر سـطح از کتاب‌هـا را نسـبت بـه یـک فرهنـگ لغـت بسـامدی ماننـد «راتلـج»[2] سـنجید. همچنیـن، در سـنجش پیچیـدگی معنـایی (چند-معنـایی) می‌تـوان از «فارس-نـت» فـارسی بهـره جسـت. بـرای سـنجش پیچیـدگی تـک‌واژی می‌تـوان از تحلیلگرهـای صـرفی رایانـشی بـرای شـمارش وندهـا و یافتـن ریشـه‌ها اسـتفاده کـرد. بـرای بـررسی طـول کلمـات بـا احتسـاب حالت‌هـای بسـیط از غیربسـیط نیـز مشـابه بـا پژوهـش کنـونی می‌تـوان از برنامه‌نویسـی «پایتـون» بهـره گرفـت. ایـن امـر می‌توانـد مؤلفـان منابـع آمـوزش زبـان فـارسی بـه غیرفارسی‌زبانـان را به‌سـوی معیـاری سـوق دهـد کـه مطابـق بـا آن و بـه‌دور از سـلیقه شـخصی، بـه دسـته‌بندی و گزینـش صحیـح واژه‌هـا و در صـورت نیـاز بـه

1. optical character reader (OCR)

2. Routledge Frequency Dictionary of Persian

اصلاح منابع آموزشی بپردازند؛ بدین‌معنا که واژگان بسامد بالا در سطح ۱ معرفی گردند، از نظر صرفی به‌هم مرتبط باشند، و البته پیچیدگی تک‌واژی بالایی نداشته باشند، و از نظر معنایی نیز به‌نسبت ساده و تک‌معنا باشند

در پایان، این پژوهش با ارائه یک پیکره واژگانی سطح‌بندی‌شده، گامی مهم در جهت استانداردسازی آموزش زبان فارسی به غیرفارسی‌زبانان برداشته است. پیشنهاد می‌شود در پژوهش‌های آینده، این پیکره با داده‌های بیشتر از منابع متنوع‌تر گسترش یابد و همچنین نرم‌افزارهای تخصصی برای پردازش واژگان فارسی توسعه داده شود تا بتوان با دقت بیشتری به تحلیل واژگان فارسی پرداخت

فهرست منابع

بی‌جن‌خان، محمود، و مهدی محسنی. ۱۳۹۱. *فرهنگ بسامدی بر اساس پیکره متنی زبان فارسی امروز*. تهران: مؤسسه انتشارات دانشگاه تهران

ترابی، منیره. ۱۳۸۹. بررسی روش‌ها و معیارهای کاربرد پیکره‌ها در آموزش زبان. با توجه ویژه به زبان فارسی. پایان‌نامه کارشناسی ارشد. دانشگاه علامه طباطبائی

جهانگردی، کیومرث. ۱۳۹۵. تحلیل کتاب‌های آموزش زبان فارسی به غیرفارسی‌زبانان: رویکرد پیکره‌ای-شناختی به آموزش واژگان. رساله دکتری. پژوهشگاه علوم انسانی و مطالعات فرهنگی

حسنی، حمید. ۱۳۸۴. *واژه‌های پرکاربرد فارسی امروز (بر مبنای پیکره یک میلیون لغتی)*. تهران: کانون زبان ایران

شمس‌فرد، مهرنوش. ۱۴۰۱. دادگان‌ها و منابع زبان فارسی: از متن تا واژه. مهرنوش شمس‌فرد و محمود بی‌جن‌خان (ویراستاران)، *پردازش متن و گفتار فارسی: مروری بر مبانی نظری و آخرین یافته‌های پژوهشی* (۱-۲۵). تهران: سمت

صحرائی، رضامراد، و سمیرا میرزائی. ۱۴۰۲. کاربردهای زبان شناسی پیکره‌ای در آموزش زبان فارسی به غیرفارسی‌زبانان. *مطالعات زبان‌ها و گویش‌های غرب ایران* (۴) ۱۱: ۱۱۳-۱۴۰.

عبادی، سامان، امیررضا وکیلی‌فرد، و خسرو بهراملو. ۱۳۹۳. تدوین فهرست واژگان پایه برای زبان فارسی: رویکردی تلفیقی. *پژوهشنامه آموزش زبان فارسی به غیرفارسی‌زبانان* ۳ (۸): ۳-۲۳.

علایی ابوذر، الهام. ۱۳۹۷. بررسی پیکره-بنیاد هم‌نگاره‌های اسمی و صفتی فارسی جهت کمک به برچسب‌گذاری صحیح اجزای کلام. *پژوهشنامه پردازش و مدیریت اطلاعات* ۳۴ (۲): ۸۹۷-۹۲۲.

فرهنگستان زبان و ادب فارسی. ۱۴۰۱. *دستور خطّ فارسی*. تهران: نشر آثار.

قیومی، مسعود. ۱۴۰۱. پیش‌پردازش و ابزارهای پایه. در مهرنوش شمس‌فرد، و محمود بی‌جن‌خان (ویراستاران)، کتاب *پردازش متن و گفتار فارسی: مروری بر مبانی نظری و آخرین یافته‌های پژوهشی* (۸۶-۱۱۳). تهران: سمت

قیومی، مسعود. ۱۳۹۶. مسئله چندواژگی در پردازش نحو رایانشی زبان فارسی. *در مجموعه مقالات چهارمین همایش ملی زبان‌شناسی رایانشی*، ۱۱-۴۰ . تهران: *نشر نویسه پارسی*

نعمت‌زاده، شهین، محمد دادرس، مهدی دستجردی کاظمی، و محرم منصوری‌زاده. ۱۳۹۰. *واژگان پایه فارسی از زبان کودکان ایرانی*. تهران: مؤسسه فرهنگی مدرسه برهان

وکیلی‌فرد، امیررضا. ۱۳۷۸. کدام زبان فارسی را به غیرفارسی‌زبانان آموزش دهیم؟ *نامه پارسی* ۴ (۳): ۲۱۲-۲۱۹.

References

Academy of Persian Language and Literature. 2023. *Dastour-e-khat*. Tehran: Asar Publication. [In Persian]

Ahmad, A., I. Ahmed Abbasi, R. Hussain Abbasi, & B. Rasheed. 2025. Exploring the intricate relationship between semantics and computational linguistics. *Liberal Journal of Language and Literature Review* 3 (1): 164-181.

Alayiaboozar, E. 2019. A corpus-based study of Persian noun and adjective homographs to help correct pos tagging. *Iranian Journal of Information Processing and Management* 34 (2): 897-922. [In Persian]

Alenizi, A., & R. Adawi. 2024. Investigating the Effectiveness of Using Corpus-Based Developed Materials in Vocabulary Learning for Saudi EFL Students. *Forum for Linguistic Studies* 6 (3): 721–745.

Anthony, L. 2023. *AntConc* (Version 4.3.1) [Computer software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.html (accessed Jan 5, 2025)

Barth, D., & S. Schnell. 2022. *Understanding Corpus Linguistics*. London & New York: Routledge.

Biber, D., & E. Finegan. 1991. *English Corpus Linguistics* London & New York: Routledge.

Bijankhan, M., & M. Mohseni. 2012. *Frequency dictionary according to a written corpus of today Persian language*. Tehran: University of Tehran Press. [In Persian]

Brezina, V., & D. Gablasova. 2015. Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics* 36 (1): 1-22.

Çalışkan, G., & S. I. Kuru Gönen. 2018. Training teachers on corpus-based language pedagogy: Perceptions on vocabulary instruction. *Journal of Language and Linguistic Studies* 14 (4):190-210.

Cervetti, G. N., E. H. Hiebert, P. D. Pearson & N. A. McClung. 2015. Factors that influence the difficulty of Science Words. *Journal of Literacy Research* 47 (2): 153–185. https://doi.org/10.1177/1086296X15615363

Chan, T. P., & H. C. Liou. 2005. Effects of web-based concordancing instruction on EFL students' learning of verb-noun collocations. *Computer Assisted Language Learning* 18 (3): 231-251.

Chen, H. J. H. 2011. Developing and evaluating a web-based collocational retrieval tool for EFL students and teachers. *Computer Assisted Language Learning* 24 (1): 59-76.

Cheng, W. 2012. *Exploring Corpus Linguistics: Language in action*. London: Routledge.

Cobb, T. 1999. Breadth and depth of lexical acquisition with hands-on concordancing. *ComputerAssisted Language Learning* 12 (4): 345–360.

Council of Europe. 2001. *The common European framework of reference for languages: Learning, teaching and assessment*. Cambridge: Cambridge University Press.

Daskalovska, N. 2015. Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning* 28 (2): 130-144.

Ebadi, S., A. R. Vakilifard & Kh. Bahramlu. 2014. Developing a General Service Wordlist for Persian Language: An Integrated Approach. *Journal of Teaching Persian to Speakers of Other Languages* 3 (8): 3-23. [In Persian]

Etezadi, R., M. Karrabi, N. Zare, M. B. Sajadi & M. T. Pilehvar. 2022. Dadmatools: Natural language processing toolkit for Persian language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations* (pp. 124-130). Seattle, Washington.

Gardner, D. & M. Davies. 2014. A new academic vocabulary list. *Applied Linguistics* 35 (3): 305-327.

Ghayoomi, M. 2018. The problem of multi-words in syntactic processing of Persian*, In Proceedings of the Fourth National Conference on Computational Linguistics* (pp.11-40). Tehran: Neveeseh. [In Persian]

Ghayoomi, M. 2022. Pre-processing and basic tools. In Shamsfard, M and Bijankhan, M (Eds), *Text and speech processing for Persian language: The state of the art and a brief review of the theoretical foundations* (pp. 86-113). Samt. [In Persian]

Hassani, H. 2005. *The Most Frequent Words of Today Persian*. Tehran: Iran Language Institute. [In Persian]

Hunston, S. 2002. *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Ide, N., & C. M. Sperberg-McQueen. 2023. *XML in theory and practice*.: Addison-Wesley Longman.

Indurkhya, N., & F. C. Damerau. 2010. *Handbook of natural language processing.* New York: CRC Press.

Jahangardi, K. 2016. *An Analysis of Textbooks for Teaching Persian to Non-Persians: A Corpus-Cognitive Approach to Teaching Vocabulary*. Doctoral dissertation. Ministry of Science, Research & Technology, Institute for Humanities & Cultural Studies. Iran. [In Persian]

Jurafsky, D., & J. H. Martin. 2024. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models.* Online manuscript. https://web.stanford.edu/~jurafsky/slp3. (accessed Jan 13, 2025)

Keck, C. M. 2004. Corpus linguistics and language teaching research: bridging the gap. *Language Teaching Research* (1): 83-109.

Kübler, S., & H. Zinsmeister. 2014. *Corpus linguistics and linguistically annotated corpora*. New York: Bloomsbury Publishing.

Leech, G. 1992. Corpora and Theories of Linguistic Performance. In J. Startvik (Ed.), *Directions in Corpus Linguistics* (pp. 105-122). Mouton de Gruyter. https://doi.org/10.1515/9783110867275.105.

Leech, G. 1997. Teaching and Language Corpora: A Convergence, in A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds) *Teaching and Language Corpora*, Harlow: Addison Wesley Longman, pp. 11–23.

Li, D., N. Noordin, L. Ismail & D. Cao. 2025. A systematic review of corpus-based instruction in EFL classroom. *Heliyon*, *11* (2), e42016. https://doi.org/10.1016/j.heliyon.2025.e42016

_____, S. 2017. Using corpora to develop learners' collocational competence. *Language Learning & Technology* 21 (3): 153–171.

Ma, Q., F. Mei & B. Qian. 2024. Exploring EFL students' pronunciation learning supported by corpus-based language pedagogy. *Computer Assisted Language Learning* ?: 1– 27. https://doi.org/10.1080/09588221.2024.2432965.

Ma, Q., R. Yuan, (Eric), L. M.E. Cheung, & J. Yang, J. 2022. Teacher paths for developing corpus-based language pedagogy: a case study. *Computer Assisted Language Learning* 37 (3): 461–492. McCarthy, M., and A. O'Keeffe. 2010. Historical perspective: What are corpora and how have they evolved? in A. O'Keeffe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 3-13). London: Routledge.

McEnery, T., & G. Brookes. 2022. Building a written corpus: What are the basics? In A. O'Keeffe and M. J. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 35–47). London: Routledge.

McEnery, T., & A. Hardie. 2011. *Corpus Linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

McEnery, T., R. Xiao. & Y. Tono. 2006. *Corpus- based language Studies: An advanced resource book.* London and New York: Routledge.

Meunier, F. & R. Reppen. 2015. Corpus versus non-corpus-informed pedagogical materials: Grammar as the focus, in D. Biber and R. Reppen (eds.) *The Cambridge Handbook of English Corpus Linguistics* (pp. 498-514). Cambridge University Press. https://doi.org/10.1007/9781139764377.028

Meyer, Ch. F. 2004. *English corpus linguistics: An introduction.* Cambridge: Cambridge University Press.

Nagy, W. E., & E. H. Hiebert. 2011. Toward a theory of word selection. In M. L. Kamil, P. D. Pearson, E. B. Moje, & P. P. Afflerbach (Eds.), *Handbook of reading research* (Vol. 4, pp. 388-404). New York, NY: Longman.

Nation, P. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nematzadeh, Sh., M. Dadras, M. Dastjerdi Kazemi, M. & Mansorizadeh. 2011. *Persian core vocabulary based on Iranian children.*Tehran: Borhan Cultural Institute. [In Persian]

Qi, P., Y. Zhang, Y. Zhang, J. Bolton, & C. D. Manning. 2020. *Stanza: A Python natural language processing toolkit for many human languages*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pages 101–108,), Association for Computational Linguistics.

Rasooli, M. S., M. Kouhestani, & A. Moloodi. 2013. Development of a Persian syntactic dependency treebank. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 306-314). Atlanta, Georgia.

Reppen, R. 2022. Building a corpus: what are key considerations? In *The Routledge handbook of corpus linguistics* (pp. 13-20). London & New York: Taylor and Francis.

Sahraei, R. M., & Samira Mirzaei. 2023. Applications of Corpus Linguistics in Teaching Persian to Non-Persian Speakers. *Journal of Research in Western Iranian Languages and Dialects* 11 (4): 113-140. [In Persian]

Seraji, M. 2015. Morphosyntactic corpora and tools for Persian. Doctoral dissertation. Uppsala University.

_____, B. Megyesi, & J. Nivre. 2012. A basic language resource kit for Persian. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2245-2252), European Language Resources Association.

Shamsfard, M. 2022. Data and Persian resources: from text to word. In Shamsfard, M and Bijankhan, M (Eds), *Text and speech processing for Persian language: The state of the art and a brief review of the theoretical foundations* (pp. 25-1). Samt. [In Persian]

_____, H. S. Jafari, & M. Ilbeygi. 2010. STeP-1: A Set of Fundamental Tools for Persian Text Processing. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (pp.859-865), European Language Resources Association.

Sharifi Atashgah, M., & M. Bijankhan. 2009. Corpus-Based Analysis for Multi-Token Units in Persian, *Proceedings of the 3rd Workshop on Computational Approaches to Arabic Script-Based Languages* [at] MT, Ottawa, Canada.

Sinclair, J. M. 1987. *Looking Up*. London: Collins Publication and The University of Birmingham.

_____. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

_____. 2004. *Trust the text: Language, corpus and discourse*.: Routledge.

_____, & A. Renouf. 1988. A lexical syllabus for language learning, in R. Carter and M. McCarthy (eds.), Vocabulary and Language Teaching (140-160). London: Longman.

Szudarski, P. 2018. *Corpus Linguistics for Vocabulary.* London & New York: Routledge.

Tiansoodeenon, M., B. Meeporm, N. Kaewrattanapat, & S. Tarapond. 2023. Enhancing vocabulary acquisition through progressive word increments in English language learning. *Journal of Liberal Arts RMUTT* 4 (2): 88–100.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Torabi, M. 2010. Study of methods and criteria for the application of corpora in language teaching, with special reference to Persian language. Master thesis. Allameh Tabatabai University. [In Persian]

Vakilifard, Amirreza. 1999. Which Persian language should we teach to Non-Persian speakers?. *Name-ye-Farsi* 4 (3): 212-219. [In Persian]

Varley, S. 2009. I'll just look that up in the concordancer: integrating corpus consultation into the language learning environment. *Computer Assisted Language Learning* 22 (2): 133-152.

West, M. 1953. *A general service list of English words*. London: Longman, Green & co.

Yan, J. & Q. Ma. 2025. Theory-supported corpus pedagogy for ESL pre-service teachers: using Parallel EAP Corpora for language learning. *Journal of China Computer-Assisted Language Learning*. https://doi.org/10.1515/jccall-2024-0016

Youngblood, A. M., & K. S. Folse. 2017. Survey of corpus-based vocabulary lists for tesol classes. *MEXTESOL Journal* 41 (1): 1-15.

Yu Liu, Ch. 2023. A corpus-based study of vocabulary in massive open online courses (MOOCs). *English for Specific Purposes* 72 (1): 40-50.

زهرا علیزاده معمار

دارای مدرک کارشناسی ارشد زبان‌شناسی همگانی از دانشگاه فردوسی مشهد است. ایشان هم‌اکنون دانشجوی دکتری زبان‌شناسی همگانی در دانشگاه فردوسی مشهد است.

زبان‌شناسی کاربردی، زبان‌شناسی پیکره‌ای و زبان‌شناسی فرهنگی از جمله علایق پژوهشی وی است.

عطیه کامیابی گل

دارای مدرک تحصیلی دکتری در رشته زبان‌شناسی کاربردی از دانشگاه مالایا مالزی است. ایشان هم‌اکنون دانشیار زبان‌شناسی کاربردی در گروه زبان‌شناسی و دانشیار وابسته گروه زبان و ادبیات فارسی دانشگاه فردوسی مشهد است. پژوهش‌های وی بر تأثیر متغیرهای روان‌شناختی در آموزش زبان فارسی و انگلیسی، رویکردی بین رشته‌ای دارد.

زبان‌شناسی کاربردی، آموزش زبان فارسی به غیرفارسی‌زبانان، روان‌شناسی زبان، زبان‌شناسی پیکره‌ای، آزمون‌سازی و تهیه و تدوین و ارزیابی منابع آموزشی مبتنی بر فناوری از جمله علایق پژوهشی وی است.

شهلا شریفی

متولد سال ۱۳۵۱، دارای مدرک تحصیلی دکتری در رشته زبان‌شناسی همگانی از دانشگاه فردوسی مشهد است. ایشان هم‌اکنون دانشیار گروه زبان‌شناسی دانشگاه فردوسی مشهد است.

رده‌شناسی زبان، عصب–روان‌شناسی زبان، زبان‌شناسی فرهنگی و زبان‌شناسی حسی از جمله علایق پژوهشی وی است.

امیرمسعود ایروانی

متولد سال ۱۳۷۲، دارای مدرک کارشناسی ارشد زبان‌شناسی رایانشی از مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری شیراز است. ایشان هم‌اکنون دانشجوی دکتری زبان‌شناسی همگانی در دانشگاه فردوسی مشهد است.

تحلیل احساسات، زبان‌شناسی پیکره‌ای، معناشناسی رایانشی و معناشناسی واژگانی از جمله علایق پژوهشی وی است.