Journal of Teaching Persian to Speakers of Other Languages
(JTPSOL)
Biannual Scientific Journal
Print Issn:5394-2322
Electronic Issn:2676-3354

Preparation of a Corpus Linguistics of Non-Iranian Learners of Persian: the case of the writings of Chinese learners of Persian

Mohammad Bagher Mirzaei Hesarian*
Corresponding author, Assistant Professor of Teaching Persian Language to speakers of other languages, Imam Khomeini International University, Qazvin, Iran.
mb.mirzaei@hum.ikiu.ac.ir
Leila Golpour
Assistant Professor of Teaching Persian Language to speakers of other languages, Imam Khomeini International University, Qazvin, Iran.
golpour@hum.ikiu.ac.ir
Amirreza Vakilifard
Associate Professor of Teaching Persian Language to speakers of other languages, Imam Khomeini International University, Qazvin, Iran.
vakilifard@hum.ikiu.ac.ir

Abstract:

The distinction between speech and writing as two types language skills is a controversial area in modern linguistic research. From this perspective, the Persian language, due to the considerable differences between the spoken and written forms, can be of particular complexity in education, especially in non-Persian languages.
Therefore, in this research we intend to study the use of speech forms in Farsi productions of foreign learners and Besides we want to know to what extent language learners employ spoken words properly without provide them with direct Persian spoken instruction and relying solely on the native environment?. besides these products are subject to a particular pattern or not.
Samples of this research were collected from intermediate level Persian students at Dehkhoda Institute. The results of this study indicate that language learners could produce spoken Persian without direct instruction of spoken form, and by relying only on the native language environment. They were taught only academically and with a focus on reading skills and grammar at Dehkhoda institute, furthermore spoken forms are used more accurately in vocabulary rather than in grammer. It is worth noting that in explaining the data of this study, the Input processing theory in general and the principle of content word priority in particular have been used. Analysis of the data has shown that according to this theory, products follow a certain pattern.

***Cite this article:** Mirzaei Hesarian, Mohammad Bagher. Golpour, Leila. Vakilifard, Amirreza. Preparation of a Corpus Linguistics of Non-Iranian Learners of Persian: the case of the writings of Chinese learners of Persian .*Vol. 12, No. 1 (Tome 25), April 2023*,23-43.

DOI: 10.30479/JTPSOL.2021.14990.1518
Received on: 30/01/2021 Accepted on: 15/08/2021

Publisher: Imam Khomeini International University

Extended Abstract:

Teaching Persian to Foreigners (TPF) is at the beginning of its ups and downs; therefore, one of the basic and necessary steps is to collect and record raw linguistic data and prepare a corpus for Non-Iranian Learners of Persian (CNLP) and describe its data using linguistic theories. The present study is the result of an in-university research project that has been carried out with the support of Imam Khomeini International University (IKIU) of Qazvin with the aim of taking a step towards creating a CNLP, identifying and resolving potential problems and meeting some of the needs of researchers .

The research is based on the book describing the grammatical structure of the Persian language based on the theory of Category and Scale grammar (CSG). In the CSG, four categories have been discussed. These four categories are "unit", "structure", "class" and "system". "Unit" and “structure" belong to the syntagmatic axis, which represents the sequence of the constituent or elements of language over time, while "class" and "system" belong to the paradigmatic axis, which represents a variety of possibilities at each point in the speech chain for the speaker to choose from.

The corpus of the research is taken from one of the final writing tests of the General and supplementary Persian language Course of the Persian Language Teaching Center (PLC) at IKIU. 90 Chinese Persian learners at the general level and 36 Chinese Persian learners at the supplementary level participated in the mentioned test; hence, a total of 126 test sheets were used as raw data .

To prepare the corpora, the writing sheets of Chinese Learners of Persian (CLP) were first typed in the Word software. Attempts were made to type as much as possible what the CPLs had written in their composition. Then, the grammatical tagging of the typed content was done within the framework of CSG. At this stage, 9 grammatical tags such as sentence, clause, ranked and rank shifted clause, finite and non-finite clauses, verbal group, the nominal and adverbial group were recorded in the writings of CLPs.

Since the dots are intended as the boundary between the end of one sentence and the beginning of another, the punctuation has been revised by scholars and, if necessary, corrected or supplemented. Next, the work of identifying and separating the sentences has been done. While analyzing corpora, the components of the rank-shifted clauses are identified and calculated as the constituent elements of the clause (verbal group, nominal group, and adverbial group). In the analysis of nominal groups with other nominal dependents, only the main nominal group is considered. Adverbial groups are also identified as a unit; this means that nominal groups are not labeled separately within adverbial groups. Also, due to the subject pronoun

dropping feature of Persian, in a significant number of sentences of CLP's writings, the subject is not specified in the form of a noun group. In the labeling of the corpus, an attempt was made to analyze the components of the text by Persian Learner's writings and the written text to be labeled without applying linguistic corrections as much as possible.

The CPLs at the general level were students who participated in the PLTC at IKIU for 16 weeks and 20 hours per week for the four skills of listening, reading, speaking and writing skills. So, they participated in a total of 320 training hours in face-to-face classes. Considering the quality and quantity of the educational program and the individual characteristics of CPL, the GPLC can be considered equivalent to the pre-intermediate level (A2) in the Common European Framework of Reference for Languages (CEFRL). The CLPs at the supplementary level were students who participated in the PLTC at IKIU for 32 weeks and 20 hours per week for the four skills of listening, reading, speaking and writing skills. So, they participated in a total of 640 training hours in face-to-face classes. Considering the quality and quantity of the educational program and the individual characteristics of CPL, the GPLC can be considered equivalent to the upper-intermediate level (B2) in the (CEFRL)

The most important achievement of the research is the preparation of the initial version of the CNLP with the characteristics that will be mentioned below: A total of 126 writings of CLPs were used as raw data for the CNLP at two levels (90 writings at the general level and 36 writings at the supplementary level). Therefore, the corpus is composed of a total of 126 written texts including 212 paragraphs and 29,857 words. Also, the corpus contains a total of 3175 sentences, 4912 clauses, 19369 groups (including 4912 current groups, 8760 noun groups and 4912 adverb groups including adjectives and preposition groups). The study proves the effectiveness of CSG in accurately describing the CLP's writings.

Keywords: Language learners corpora, Chinese learners of Persian, Category and Scale Grammar, Writing.

پژوهش‌نامهٔ آموزش زبان فارسی به غیر فارسی‌زبانان (علمی- پژوهشی)
شاپا چاپی: ۵۳۹۴-۲۳۲۲
شاپا الکترونیکی: ۲۳۵۴-۲۶۷۶

تهیهٔ پیکرهٔ زبان‌آموز فارسی‌آموزان غیرایرانی (مورد نوشتار فارسی‌آموزان چینی) (پژوهشی)

محمدباقر میرزایی حصاریان*
نویسندهٔ مسئول، استادیار گروه آموزش زبان فارسی به غیرفارسی‌زبانان، دانشگاه بین‌المللی امام خمینی(ره)، قزوین. ایران.
mb.mirzaei@hum.ikiu.ac.ir
لیلا گل‌پور
استادیار گروه آموزش زبان فارسی به غیرفارسی‌زبانان، دانشگاه بین‌المللی امام خمینی(ره)، قزوین. ایران.
golpour@hum.ikiu.ac.ir
امیررضا وکیلی‌فرد
دانشیار گروه آموزش زبان فارسی به غیرفارسی‌زبانان، دانشگاه بین‌المللی امام خمینی(ره)، قزوین. ایران.
vakilifard@hum.ikiu.ac.ir

چکیده

یکی از گام‌های اساسی و ضروری در آموزش زبان فارسی به غیرفارسی زبانان (آزفا)، جمع‌آوری و ثبت داده‌های خام زبانی و تهیهٔ پیکرهٔ زبان‌آموز فارسی‌آموزان غیرایرانی و توصیف داده‌های آن با استفاده از نظریه‌های زبان‌شناسی است. پژوهش حاضر با هدف برداشتن گامی در راستای ایجاد پیکره زبانی فارسی‌آموزان غیرایرانی انجام شده است. داده‌های خام زبانی پژوهش برای تهیه پیکره مورد نظر برگرفته از آزمون نگارش پایان دوره فارسی‌آموزان چینی مرکز آموزش زبان فارسی دانشگاه بین المللی امام خمینی(ره) (۹۰ فارسی آموز سطح فراپایه (A2) و ۳۶ فارسی‌آموز سطح فرامیانی(B2)) است و برچسب‌گذاری نحوی برپایه دستور مقوله و میزان و توصیف باطنی از دستور زبان فارسی و به صورت دستی انجام شده است. تعداد ده برچسب دستوری شامل جمله، بند؛ بند مرتبه‌بندی‌شده و بند واژگون‌مرتبه؛ بند خودایستا و بند ناخودایستا، گروه فعلی، گروه اسمی(متمم و مسند الیه) و گروه قیدی در نوشتار فارسی‌آموزان ثبت شده است. پیکره تهیه شده مجموعاً از ۱۲۶ متن نوشتاری شامل ۲۱۲ پاراگراف و ۲۹۸۵۷ واژه، ۳۱۷۵ جمله، ۴۹۱۲ بند، ۱۹۳۶۹ گروه شامل ۴۹۱۲ گروه فعلی، ۸۷۶۰ گروه اسمی و ۴۹۱۲ گروه قیدی شامل ادات و گروه‌های حرف اضافه‌ای تشکیل شده است. پژوهش همچنین کارایی دستور توصیفی باطنی را که مبتنی بر دستور مقوله و میزان است، در برچسب‌گذاری نحوی نوشتار فارسی‌آموزان چینی تایید می‌کند.

کلیدواژه‌ها

پیکره زبان‌آموز، دستور مقوله و میزان، نوشتار، فارسی‌آموز چینی، فراپایه، فرامیانی.

---

* **استناد:** میرزائی حصاریان، محمدباقر. گل‌پور، لیلا. وکیلی‌فرد، امیررضا. تهیهٔ پیکرهٔ زبان‌آموز فارسی‌آموزان غیرایرانی (مورد نوشتار فارسی‌آموزان چینی)، سال دوازدهم، شمارهٔ اول (پیاپی ۲۵)، بهار و تابستان ۱۴۰۲، ۲۳-۴۳.
شناسه دیجیتال (DOI): 10.30479/JTPSOL.2021.14990.1518
تاریخ دریافت مقاله: ۱۳۹۹/۱۱/۱۱ تاریخ پذیرش نهایی مقاله: ۱۴۰۰/۰۲/۲۸
ناشر: دانشگاه بین‌المللی امام خمینی(ره)

۱. مقدمه و بیان مسأله

امروزه، پیکره‌های زبان‌آموز در مباحث آموزش زبان و به‌ویژه برای برطرف کردن مشکلات یادگیری زبان دوم که بر اثر تداخل زبانی حاصل می‌شوند، جایگاه برجسته‌ای دارند. آموزش زبان فارسی به غیرفارسی زبانان (آزفا)، در ابتدای مسیر پرفراز و نشیب خود قرار دارد. در این مسیر پرفراز و نشیب، یکی از گام‌های اساسی و ضروری، جمع‌آوری و ثبت داده‌های خام زبانی و تهیهٔ پیکرهٔ زبان‌آموز فارسی‌آموزان غیرایرانی و توصیف داده‌های آن با استفاده از نظریه‌های زبان‌شناسی است. کمترین فایدهٔ چنین اقدامی آن است که چنانچه نظریات زبان‌شناسی در گذر زمان، ارزش و اعتبار خود را از دست بدهند، پیکرهٔ زبانی در همه حال، ارزش خود را حفظ کرده و ماندگار می‌ماند.. نظریهٔ زبان‌شناسی و پیکرهٔ زبانی لازم و ملزوم یکدیگرند؛ بدین معنا که نظریات زبان‌شناسی، با مطالعهٔ پیکره‌های زبانی محک زده می‌شوند و میزان اعتبارشان بر اساس داده‌های واقعی سنجیده می‌شوند. از سوی دیگر، پیکرهٔ زبانی، زمینهٔ اصلاح یا تغییر نظریات زبان‌شناسی موجود و یا طرح نظریات جدید را فراهم آورد (Halliday& Matthiessen,2004,p.33-36). پیکره‌های زبانی، به‌عنوان انباره‌ای از واژه‌های زبان، ویژگی و ظرفیت‌هایی دارند که با کشف آن‌ها می‌توان به اطلاعات پنهان و پیدای متون دست یافت. بسامدگیری از این پیکره‌ها به شیوه‌های مختلف و با ارزش‌گذاری‌های متفاوت، اهداف متنوعی را دنبال می‌کنند (Mirzaei&Safari,2015).

نبود الگوی عملی دربارهٔ کاربرد نظریات زبانی در واکاوی داده‌های زبانی فارسی‌آموزان غیرایرانی، یکی از مشکلات موجود در استفادهٔ هرچه بهتر از داده‌های خام زبانی فارسی‌آموزان غیرایرانی است؛ به‌طوری که می‌توان ادعا کرد، تقریباً تمام داده‌های زبانی فارسی‌آموزان در مراکز آموزش زبان فارسی داخل و کرسی‌های زبان و ادبیات فارسی و ایران‌شناسی خارج از کشور، بدون هیچ برنامه‌ریزی خاصی از بین می‌روند؛ در صورتی که ایجاد پیکرهٔ خام زبانی فارسی‌آموزان در گام نخست و واکاوی داده‌های زبانی از جنبه‌های گوناگون و براساس متغیرهایی؛ همچون ملیّت، جنسیت، زبان مادری، سطح زبانی و محیط یادگیری زبان فارسی (زبان دوم/ زبان خارجی)، کمک بزرگی در جهت علمی‌تر شدن فرایند آموزش زبان فارسی به غیرفارسی زبانان خواهد شد.

پیکره‌های زبانی از یک‌سو برای پژوهش‌گران، مدرسان و دانشجویان مقطع کارشناسی ارشد و دکتری رشتهٔ آموزش زبان فارسی به غیرفارسی زبانان، به‌عنوان منبع داده‌های واقعی در امور پژوهشی مورد استفاده قرار خواهد گرفت و از سوی دیگر، برای مراکز پژوهشی داخل و خارج کشور نیز در انجام امور پژوهشی پیکره‌بنیاد در حوزهٔ متناسب با عنوان طرح، کاربرد خواهد داشت. بر این اساس، پژوهش حاضر در قالب طرح پژوهشی و با حمایت دانشگاه بین المللی امام خمینی(ره) قزوین، با هدف برداشتن گامی در راستای ایجاد پیکرهٔ زبانی فارسی‌آموزان غیرایرانی و رفع بخشی از نیازهای پژوهشی پژوهش‌گران انجام گرفته است. انتظار می‌رود؛ با انجام طرح پیشنهادی، بنیان ابتدایی تهیهٔ پیکرهٔ زبانی فارسی‌آموزان غیرایرانی پی‌ریزی شود و موانع و مشکلات احتمالی شناسایی گردد

و نیز پژوهش‌گران امکان استفاده از داده‌های پیکره را پیدا کنند. در این نوشتار تلاش می‌شود، از مراحل انجام طرح تحقیقاتی فوق و دستاوردهای حاصل از آن گزارشی بیان شود

۲. پیشینه پژوهش

۲. ۱. پیکره بی‌جن‌خان

پیکره بی‌جن‌خان (Bijankhan,2004) مجموعه‌ای است از متون فارسی؛ شامل بیش از ۲ میلیون و ۶۰۰ هزار کلمه که با ۵۵۰ نوع برچسب، برچسب‌گذاری شده‌اند. این پیکره که در پژوهشکدهٔ پردازش هوشمند علائم تهیه شده است، همچنین شامل بیش از ۴۳۰۰ برچسب موضوعی همچون سیاسی، تاریخی و ... برای متون است.

۲. ۲. پیکرهٔ گفتار محاوره‌ای زبان فارسی

پیکرهٔ گفتار محاوره‌ای زبان فارسی (Bijankhan,2016)، شامل ۳۵۰ ساعت دادهٔ گفتاری است که به‌صورت سیگنال اکوستیکی از فارسی‌زبانان داوطلب در موقعیت‌های مختلف ارتباطی، با هدف تحقیقات کاربردی و آموزش و آزمون سامانه‌های محاوره‌ای ضبط شده است و در سطح نوبت و پاره‌گفتار نشانه‌گذاری می‌شود. خروجی‌های پیکره؛ شامل پروندهٔ شرکت‌کنندگان بر حسب سن، جنسیت، لهجه، میزان تحصیلات، نوع سیاق، مدت‌زمان گفتمان، پرونده‌های صوتی و شبکهٔ متنی متناظر، پروندهٔ متنی نشانه‌گذاری نوشتاری، پروندهٔ واژگان پیکره و مستندات پیکره است.

۲. ۳. پیکره فارس دات

فارس‌دات (Bijankhan & Others,2015)، مجموعه‌ای از عبارات و جملات است که توسط گویندگان فارسی‌زبان، از مناطق مختلف کشور بیان شده است. این دادگان در سطح واج (آوا) با دقت میلی‌ثانیه تقطیع و برچسب‌دهی شده و بصورت فایل‌های مجزا ذخیره شده است. این دادگان، به عنوان دادگان استاندارد گفتاری زبان فارسی در داخل و خارج کشور شناخته شده و برای آموزش سیستم‌های هوشمند تشخیص گفتار استفاده می‌شود.

۲. ۴. پیکرهٔ فارسی‌نت

فارس‌نت (Shamsfard & Others,2010) نخستین، دقیق‌ترین و بزرگترین وردنت فارسی است که در آزمایشگاه پردازش زبان طبیعی دانشگاه شهید بهشتی و با حمایت مرکز تحقیقات مخابرات ایران توسعه یافته است. در فارس‌نت، واژگان فارسی توسط فرد خبره در قالب دسته‌های هم‌معنا قرار گرفته و بین آن‌ها روابط معنایی ایجاد شده است. همچنین برای معناهای واژگان، برخی اطلاعات نحوی با توجه به مقولهٔ نحوی تکمیل

شده و بین معانی روابطی برقرار شده است. تعداد واژگان در نسخهٔ فارس‌نت ۳ بیش از ۱۰۰۰۰۰ و روابط بین معنا بیش از ۳۱۰۰۰ رابطه است.

۲. ۵. پیکرهٔ متنی زبان فارسی

پیکره متنی زبان فارسی (Bijankhan & Others,2011) مجموعه‌ای از متون نوشتاری و گفتاری رسمی زبان فارسی است که از منابع واقعی؛ همچون روزنامه‌ها، سایت‌ها و مستنداتِ از قبل تایپ شده، جمع‌آوری شده، تصحیح گردیده و برچسب خورده است. حجم این دادگان، حدود ۱۰۰ میلیون کلمه است و از منابع مختلف تهیه گردیده و دارای تنوع بسیار زیادی است. ۱۰ میلیون کلمه از این پیکره، با استفاده از ۸۸۲ برچسب نحوی – معنایی به‌صورت دستی توسط دانشجویان رشتهٔ زبان‌شناسی برچسب‌دهی شده‌اند و هر پرونده، بر حسب موضوع و منبع آن طبقه‌بندی شده است. این پیکره که توسط پژوهشکدهٔ پردازش هوشمند علائم تهیه شده است، برای استفاده در آموزش مدل زبانی و سایر پروژه‌های مربوط به پردازش زبان طبیعی مناسب است.

۲. ۶. پیکره متون خبری

پیکرهٔ متون خبری یا پرسیکا (Eghbalzadeh & Others,2012)، پیکره‌ای حاوی متون خبری برگرفته از خبرگزاری ایسنا است. متون این پیکره، در یازده طبقهٔ موضوعی؛ شامل ورزشی، اقتصادی، فرهنگی، مذهبی، تاریخی، سیاسی، علمی، اجتماعی، آموزشی، حقوق قضایی و بهداشت طبقه‌بندی شده‌اند و پیش‌پردازش‌هایی به‌منظور قابل استفاده بودن در کاربردهای مختلف پردازش زبان طبیعی و داده‌کاوی بر روی آن‌ها انجام گرفته است.

۲. ۷. پیکرهٔ وابستگی نحوی زبان فارسی

پیکرهٔ وابستگی نحوی زبان فارسی (Rasool & Kouhestani & Moloodi,2013)، شامل حدود ۳۰ هزار جملهٔ برچسب خورده است که اطلاعات نحوی و ساخت‌واژی را بر مبنای دستور وابستگی تهیه و عرضه نموده است. این پیکره، از منابع مختلفی از متون فارسی معاصر تهیه شده است؛ تمامی جملات آن دارای برچسب روابط نحوی (بر مبنای دستور وابستگی) از قبیل: فاعل، مفعول، مسند، مضاف‌الیه، بدل و ... هستند. همچنین تمامی جملات دارای برچسب اطلاعات ساخت‌واژی؛ از قبیل فعل، اسم، صفت، قید، ضمیر ... هستند. جملات، توسط تیمی از زبان‌شناسان مجرب برچسب خورده‌اند و در چند مرحله بازبینی شده‌اند. پیکره، شامل ۲۹،۹۸۲ جمله، ۴۹۸،۰۸۱ واژه، ۳۷،۶۱۸ واژهٔ منحصر به فرد است. میانگین طول هر جمله، ۱۶/۶۱ و تعداد افعال منحصر به فرد، ۴،۷۸۲ است.

۲ .۸. پیکره زبان‌آموز فارسی

پیکرهٔ زبان‌آموز فارسی (Safari,2015) مجموعه‌ای است؛ شامل تعداد ۱۵۰ متن نگارشی که به‌صورت نمونه و تصادفی از میان بایگانی انشاهای فارسی‌آموزانی که از کشورهای مختلف با سابقهٔ زبان اول مختلف (ترکی، هندی، انگلیسی، عربی، چینی و ...) بوده‌اند، انتخاب شده است و خطاهای زبانی در آن برچسب خورده است. کاربرد اصلی این پیکره، بررسی خطاهای فارسی‌آموزان، با توجه به نوع زبان اول ایشان است. داده‌های مورد نظر این پیکره، از متون نگارشی فارسی‌آموزان سطوح میانی و پیشرفته، از مرکز آموزش زبان فارسی المهدی، وابسته به جامعةالمصطفی (ص) جمع‌آوری و تهیه شده و موضوع متن‌های نگارشی متفاوت و عمدتاً در راستای اهداف زبان‌آموزی ایشان؛ یعنی یادگیری فارسی برای تحصیل علوم دینی بوده است.

۲ .۹. پیکره فارسی روز

پیکره فارسی روز (Ghorbanzadeh,2015) پیکره‌ای خام، یک‌زبانه و پیوسته؛ شامل ۱۲۷ متن فارسی است که تمام آن بین سال‌های ۱۳۸۰ تا ۱۳۹۲ نوشته شده و اغلب از میان داستان‌های کوتاه و رمان‌ها انتخاب شده است. این پیکره، برای تألیف اثری با عنوان «فرهنگ فارسی روز» گردآوری شده و برای استفاده از آن، نرم‌افزاری ویژهٔ فرهنگ‌نویسی با عنوان پِرلِکس طراحی و آماده شده است. تعداد کل واژه‌های موجود در پیکرهٔ فارسی روز ۴,۲۷۴,۳۶۰ واژه است و ۱۴۴,۲۱۰ واژه بدون تکرار در آن وجود دارد.

۲ .۱۰. پیکرهٔ گفتار محاوره‌ای زبان فارسی

پیکرهٔ گفتار محاوره‌ای زبان فارسی (Bijankhan & Others,2016)؛ شامل ۳۵۰ ساعت دادهٔ گفتاری است که به‌صورت سیگنال اکوستیکی از فارسی‌زبانان داوطلب در موقعیت‌های مختلف ارتباطی با هدف تحقیقات کاربردی و آموزش و آزمون سامانه‌های محاوره‌ای ضبط شده است و در سطح نوبت و پاره‌گفتار نشانه‌گذاری می‌شود. خروجی‌های پیکره؛ شامل پروندهٔ شرکت‌کنندگان بر حسب سن، جنسیت، لهجه، میزان تحصیلات، نوع سیاق، مدت‌زمان گفتمان، پرونده‌های صوتی و شبکهٔ متنی متناظر، پروندهٔ متنی نشانه‌گذاری نوشتاری، پروندهٔ واژگان پیکره و مستندات پیکره است.

۲,۱۱. پایگاه داده‌های زبان فارسی

پایگاه داده‌های زبان فارسی (Assi,2018)، مجموعه‌ای است از متون مختلف فارسی که بخشی از آن دارای نشانه‌گذاری‌هایی؛ از جمله شناسنامهٔ متن، برچسب‌های دستوری، آوایی، ریشه‌ای و معنایی است. این دادگان که در پژوهشگاه علوم انسانی و مطالعات فرهنگی تهیه شده است، مجهز به نرم‌افزارهای اختصاصی جستجو، تقطیع و تحلیل متن است که می‌تواند انواع فهرست‌های واژگانی، بسامدی و آماری را ارائه کند. از جمله متون موجود

در پیکرهٔ حدود ۴۵۰ اثر داستانی و غیرداستانی نثر، ۲۵۰ اثر شعری از شاعران معاصر، بیش از ۸۰ عنوان مجله و نشریه علمی، ادبی و تخصصی، نزدیک به ۳۰۰ عنوان نمایشنامه و فیلم‌نامه و ۲۰۰ عنوان ادبیات کودک، چندین عنوان روزنامه و نشریهٔ خبری، برخی از کتاب‌های درسی دانشگاهی و دبیرستانی، برخی از کتاب‌های دبستانی، نامه‌های اداری و بخشنامه‌ها، مجموعهٔ کامل قوانین و مقرّرات، نشریه‌ها و جزوه‌های پراکنده، پوسترها و دیوارنوشته‌ها است.

۱۲,۲. پیکره گفتمانی زبان فارسی

پیکرهٔ گفتمانی زبان فارسی (Mirzaei & Safari,2018) جزو معدود پیکره‌های برچسب‌خوردهٔ جهان است که اطلاعات گفتمانی را به‌صورت دستی بر روی داده‌های زبانی قرار داده است تا زیرساخت مناسبی را برای هوشمندی‌سازی ماشین و پردازش زبان طبیعی فراهم آورد. این پیکره، توسط پژوهش‌گران گروه پیکره و دادگان مرکز تحقیقات کامپیوتری علوم اسلامی (نور) و با حمایت سازمان فناوری اطلاعات تولید شده است. داده‌ای که برچسب‌گذاری گفتمانی روی آن صورت گرفته است، پیکرهٔ وابستگی نحوی زبان فارسی با حجم حدود ۳۰ هزار جمله است. حدود ۲۰ هزار جمله، به‌صورت دوبرچسبی برچسب‌زنی شده است تا توافق میان برچسب‌زنان قابل بررسی باشد.در این پیکرهٔ گفتمانی، تمام روابط منطقی درون‌جمله‌ای برچسب‌زنی شده است.

۱۳,۲. پیکره آموزشی زبان فارسی

پیکرهٔ آموزشی زبان فارسی (Jahangardi,2016) که به‌عنوان بخشی از کار پژوهشی رسالهٔ دکتری در پژوهشگاه علوم انسانی و مطالعات فرهنگی تهیه شده است، دارای ۵ میلیون واژه است و در طراحی و ایجاد آن از داده‌های کتاب‌های آموزش زبان فارسی به غیرفارسی زبانان، در مهم‌ترین مراکز حال حاضر آموزش زبان فارسی داخل کشور استفاده شده است.

پیشینهٔ پژوهش‌ها، نشانگر آن است که تاکنون پیکرهٔ زبان‌آموز فارسی‌آموزان غیرایرانی و به‌طور ویژه پیکره‌ای ویژهٔ نوشتار فارسی‌آموزان چینی تهیه نشده است و این موضوع بر اهمیت و ضرورت انجام پژوهش می‌افزاید.

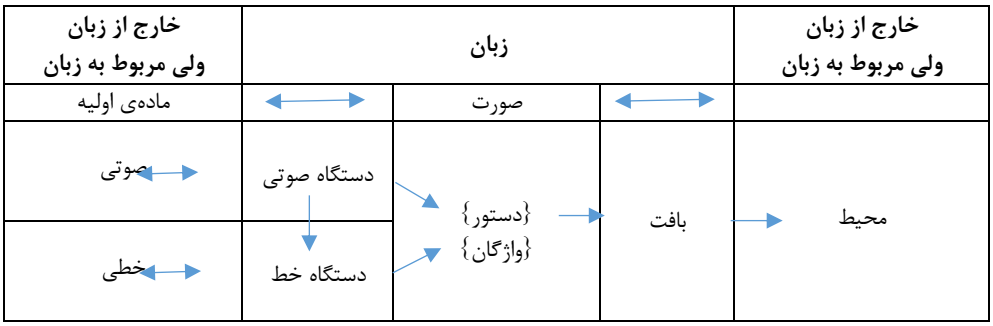
۳. چهارچوب نظری

کتاب توصیف ساختمان دستوری زبان فارسی، بر بنیاد نظریهٔ مقوله و میزان که در سال ۱۳۴۸، توسط محمدرضا باطنی به چاپ رسیده و تاکنون بیش از سی بار تجدید چاپ شده و هزاران نسخه از آن از سوی علاقه‌مندان زبان فارسی مورد استفاده قرار گرفته است، مبنای علمی پژوهش بوده است. از ویژگی‌های بارز کتاب باطنی، توصیف نسبتاً جامع، روشن و دقیق زبان فارسی، در چهارچوب دستوری مقوله و میزان است. این موضوع سبب شده تا آن چهارچوب به ظاهر قدیمی، مبنای پیکره‌پژوهی حاضر قرار گیرد تا هم کارآمدی دستور مقوله و میزان، در

توصیف و واکاوی پیکرهٔ نوشتار فارسی‌آموزان چینی سطح فرامیانی سنجیده شود و هم تقدیر و سپاسی از باطنی به خاطر پایه‌ریزی سنت توصیف علمی زبان فارسی باشد.

۳. ۱. نظریهٔ زبانی مقوله و میزان

نظریهٔ زبانی مقوله و میزان، یکی از نظریه‌های عمومی زبان است که به وسیلهٔ ام. ا. کی. هلیدی، استاد زبان‌شناسی دانشگاه لندن عرضه شده است و به میزان وسیعی از نظریات جی.آر. فرث، زبان‌شناس فقید انگلیسی، متأثر است. طبق این نظریه، همهٔ زبان‌های جهان، از امواج صوتی در گفتار و از نشانه‌های دیداری در نوشتار، به‌عنوان مادهٔ اولیه استفاده می‌کنند تا دربارهٔ جهان بیرون و دریافت انسان از آن بحث کنند. زبان‌های مختلف، به‌طور متفاوت بین مادهٔ اولیهٔ خود و جهان بیرون رابطه برقرار می‌کنند؛ یعنی طرح‌ها و الگوهای متفاوتی بر مادهٔ اولیه تحمیل می‌نمایند (Bateni,2013,p.21). شبکهٔ روابطی را که بین مادهٔ اولیه و جهان بیرون وجود دارد، می‌توان به شکل زیر نشان داد:



شکل ۱: نمودار شبکه روابط بین ماده اولیه و جهان بیرون

در دستور مقوله و میزان، از چهار مقوله سخن به میان آمده است. این چهار مقوله عبارتنداز «واحد»، «ساختمان»، «طبقه» و «دستگاه». دو مقولهٔ «واحد» و «ساختمان»، متعلّق به محور زنجیری هستند که نمایندهٔ تسلسل یا توالی عنصرهای سازندهٔ زبان در روی بُعد زمان است و دو مقولهٔ «طبقه» و «دستگاه»، متعلّق به محور انتخابی هستند که نمایندهٔ امکانات گوناگونی است که در هر نقطه از زنجیر گفتار، در اختیار اهل زبان قرار می‌دهد تا از میان آن‌ها، یکی را انتخاب کند (Ibid.p.37).

«میزان»، یعنی رابطه‌ی پیوسته بین چند عنصر یا چند عامل که در یک طرف نشان‌دهنده‌ی حداقل و در طرف دیگر نشان دهندهٔ حداکثر یک خصوصیت باشد. نظریهٔ مقوله و میزان، در دستگاه دستوری زبان به سه میزان قائل است. این سه میزان دستوری عبارتنداز: میزان مراتب، میزان نمود و میزان تحلیل. میزان «مراتب» یا سلسله‌مراتب، نشان‌دهندهٔ رابطه واحدها از بزرگتر به کوچکتر یا از کوچکتر به بزرگتر است. میزان «نمود»؛ یعنی

رابطهٔ عنصر ساختمانی، طبقه و مورد که به ترتیب از مفهومی ذهنی به مفهومی عینی سیر می‌کند. میزان «تحلیل» هم نشان‌دهندهٔ ظرافتی است که ما در تجزیه و تحلیل یک قطعهٔ زبان نشان می‌دهیم (Ibid.p.51).

۳. ۲. ساختمان دستوری زبان فارسی

۳ . ۲ . ۱. ساختمان جمله

جمله، آن واحد زبان فارسی است که از یک بند یا بیشتر ساخته شده است. جمله، در زبان فارسی به دو گونهٔ: هسته‌ای و خوشه‌ای تقسیم می‌شود. جملهٔ هسته‌ای، جمله‌ای است که از یک هستهٔ مرکزی تشکیل شده است که وجود آن اجباری است و از تعدادی وابسته که وجود آن‌ها اختیاری است. جملهٔ (۱) از چهار بند ساخته شده است که یک هسته و سه وابسته (یک وابستهٔ پیشرو و دو وابستهٔ پیرو) دارد (Ibid.p.60-61).

(۱) اگر او تلفن کرد، تو حتما برو؛ والا او می‌رنجد، چون آدم خیلی حساسی است.

وابسته **هسته** **وابسته** **وابسته**

جملهٔ هسته‌ای، دارای دو عنصر ساختمانی بند هسته و بند وابسته است. بند هسته، می‌تواند مستقل از وابسته وجود داشته باشد؛ ولی بندهای وابسته‌ها نمی‌توانند مستقل از هسته وجود داشته باشند. جایگاه هسته، توسط طبقه‌ای از بندها اشغال می‌شود که به آن طبقهٔ بند آزاد گفته می‌شود و جایگاه بند وابسته، به‌وسیلهٔ طبقه‌ای از بندها که طبقهٔ بند مقید نامیده می‌شوند، اشغال می‌گردد؛ برای نمونه بند «تو حتما برو» می‌تواند به‌جای یک جمله بنشیند، بدون این که احتیاج به بندهای دیگر در پیش یا پی خود داشته باشد. ولی بندهایی مانند «چون خیلی آدم حساسی است»، «اگر او تلفن کرد»، «و الا او می‌رنجد» و مانند آن، نمی‌توانند به تنهایی در جای یک جمله بنشینند.

جملهٔ خوشه‌ای، از اجتماع جمله‌های هسته‌ای تشکیل می‌شود. عناصر سازندهٔ جملهٔ خوشه‌ای، جمله‌هایی هستند که معمولاً به‌وسیلهٔ «و» و «یا» به یکدیگر قلاب شده‌اند. هر جملهٔ خوشه‌ای، به تعداد جمله‌های هسته‌ای که آن را می‌سازد، هسته دارد و هرکدام از این هسته‌ها، می‌توانند خود دارای وابسته‌هایی باشند. حداقل ساختمان یک جملهٔ خوشه‌ای، دو جمله‌ی هسته‌ای است که یکی از آن‌ها باید الزاماً دارای وابسته باشد. مثال (۲) یک جملهٔ خوشه‌ای است که از اجتماع دو جملهٔ هسته‌ای تشکیل شده است. این دو جمله به‌وسیلهٔ «و» به هم قلاب شده‌اند (Ibid.p.60-73).

(۲) معمولا آن‌ها از چنین طرحی استقبال نخواهند کرد، اگرچه به نفع آن‌ها باشد؛ و اگر شما فایدهٔ این طرح را به طور عینی ثابت نکنید، آن را نخواهند پذیرفت.

هسته **وابسته** **وابسته** **هسته**

۳. ۲. ۲. ساختمان بند

«بند»، به آن واحد زبان فارسی گفته می‌شود که از یک گروه یا بیشتر ساخته شده است و خود در ساختمان واحد بالاتر، یعنی جمله به‌کار می‌رود. در سلسله مراتب واحدها، بند، پایین‌تر از جمله و بالاتر از گروه قرار می‌گیرد. بند، دارای چهار عنصر ساختمانی اسناد، مسندالیه، متمم و ادات است. اسناد، آن عنصر ساختمانی بند است که محل کارکرد گروه فعلی است. مسندالیه، آن عنصر ساختمانی است که محل کارکرد گروه اسمی به‌استثنای آن ریز طبقه از گروه اسمی است که دارای عنصر«را» می‌باشد. متمم، آن عنصر ساختمانی بند است که محل کارکرد گروه اسمی است. ادات، آن عنصر ساختمانی بند است که محل کارکرد گروه قیدی است. پس از شناختن اسناد، مسندالیه و متمم، آنچه در ساختمان بند باقی می‌ماند، ادات است.

بند را از نظر ساخت، می‌توان به دو نوع تقسیم کرد: بند مهین و بند کهین. در بند مهین، جایگاه اسناد حتماً اشغال شده است. بند مهین، بدون مسندالیه، متمم و ادات می‌تواند وجود داشته باشد؛ ولی بدون عنصر اسناد نمی‌تواند وجود داشته باشد. بند کهین، بندی است که در آن جایگاه اسناد اشغال نشده است. مانند: چشمم روشن، از تو حرکت از خدا برکت (Ibid.p.4-82).

۳. ۲. ۳. ساختمان گروه‌های فارسی

گروه، واحدی است که از یک کلمه یا بیشتر ساخته شده است و خود در ساختمان واحد بالاتر، یعنی بند به کار می‌رود. در سلسله‌مراتب واحدها، گروه، پایین‌تر از بند و بالاتر از کلمه قرار می‌گیرد.گروه‌های فارسی، به سه طبقه تقسیم می‌شوند: طبقهٔ گروه فعلی، طبقهٔ گروه اسمی و طبقهٔ گروه قیدی.

طبقهٔ گروه فعلی در ساختمان واحد بالاتر؛ یعنی بند، همیشه جایگاه اسناد را اشغال می‌کند و از یک کلمه یا بیشتر ساخته شده است. یکی از خصوصیات زبان فارسی، داشتن فعل‌های متعدّد ترکیبی است که معمولاً از یک اسم یا صفت یا عنصر دیگری به اضافهٔ فعل ساخته می‌شوند. زبان فارسی، از این ابزار فراوان استفاده می‌کند؛ به‌طوری که برای ساختن فعل‌های تازه، به‌ندرت از صرف مستقیم کلمه استفاده می‌کند؛ بلکه عموماً روش ترکیب را به‌کار می‌برد. مثلاً گفته می‌شود: «سرخ شدن» نه «سرخیدن»، «باز کردن» نه «بازیدن»، «کتک زدن» نه «کتکیدن» و غیره. این دسته از افعال فارسی، از نظر معنی، یک واحد هستند؛ ولی از نظر ساختمان دستوری، دو جزء هستند و دارای دو نوع رفتار متفاوت می‌باشند. مثلاً «فریفتن» از نظر معنی، معادل با «فریب دادن» است؛ ولی «فریب دادن» از نظر دستوری، قابل تجزیه به دو جزء است: جزء اول آن، «فریب»، می‌تواند مرکز یک گروه اسمی قرار گرفته و مانند یک گروه اسمی گسترش یابد. بدین ترتیب، می‌توان گفت؛ «او را فریب سختی داد» که در اینجا «فریب»، مرکز یک گروه اسمی قرار گرفته و وابسته‌ای در پی آن افزوده شده است. چون جزء اول این افعال ترکیبی، به‌صورت گروه اسمی قابل بسط هستند، بنابراین، از نظر طبقه‌بندی دستوری، جزء اول و دوم،

متعلّق به یک طبقه نیستند. به‌همین دلیل، افعال ترکیبی؛ مانند مثال (۳) به دو جزء تجزیه می‌شوند: قسمت غیرفعلی که متمم نامیده می‌شود و به‌عنوان واژهٔ جداگانه در پیکره شمارش می‌شود و قسمت فعلی که اسناد خوانده می‌شود.

(۳) تو او را فریب دادی

مسندالیه متمم متمم اسناد

در گروه فعلی زبان فارسی، سه دستگاه به‌طور هم‌زمان وجود دارد که عبارتنداز: زمان‌داری، نفی و اثبات و جهت. منظور از هم‌زمان بودن دستگاه‌ها، این است که از امکانات آن‌ها در آن واحد باید انتخاب شود (Ibid.p.11-136).

۳. ۲. ۴. ساختمان گروه اسمی

گروه اسمی زبان فارسی، از یک کلمه یا بیشتر ساخته شده است و در ساختمان واحد بالاتر؛ یعنی بند، جایگاه مسندالیه، متمم و گاهی نیز جایگاه ادات را اشغال می‌کند. هر گروه اسمی در فارسی، از یک هسته و تعدادی وابسته که در دو طرف هسته قرار می‌گیرند، ساخته می‌شود. وابسته‌هایی که پیش از هسته قرار می‌گیرند، وابسته‌های پیشرو و وابسته‌هایی که پس از هسته می‌آیند، وابسته‌های پیرو نامیده می‌شوند. ملاک تشخیص این سه عنصر (وابستهٔ پیشرو، هسته و وابستهٔ پیرو) از یکدیگر «اضافه» است. حداکثر چهار وابستهٔ پیشرو، می‌تواند در جلوی هسته قرار گیرد و بین آن‌ها و نیز بین وابستهٔ پیشرو و هسته، «اضافه» به کار نمی‌رود. حداکثر پنج عنصر هم می‌توانند جایگاه وابسته‌های پس از هسته را اشغال کنند. بر این اساس، به لحاظ نظری، هستهٔ گروه اسمی می‌تواند در مجموع، نه وابسته به خود بگیرد (Ibid.p.137-170).

۳. ۲. ۵. ساختمان گروه قیدی

گروه قید فارسی، از یک کلمه یا بیشتر ساخته شده و در ساختمان واحد بالاتر؛ یعنی بند، جایگاه ادات را اشغال می‌کند. علاوه بر گروه‌های قیدی که همیشه در جایگاه ادات در ساختمان بند قرار می‌گیرند، بسیاری از گروه‌های اسمی نیز می‌توانند این جایگاه را اشغال کنند. گروه‌های قیدی را می‌توان به دو دستهٔ کلی تقسیم کرد: گروه‌های قیدی، دارای علامت صوری و گروه‌های قیدی بی‌علامت. گروه‌های قیدی دارای علامت صوری، به دو ریزطبقه تقسیم می‌شوند: گروه‌های قیدی با علامت تنوین؛ مانند « ظاهراً»، «فعلاً» و غیره و گروه‌های قیدی، با حرف اضافه مانند «در هفته گذشته»، «دور از خانه» و غیره. گروه‌های قیدی بی‌علامت، دارای هیچ علامت صوری خاصی نیستند که ساختمان آن‌ها را متمایز کند. این طبقه نیز به دو ریزطبقه تقسیم می‌شود: ریزطبقهٔ باز واژگانی و ریزطبقهٔ بستهٔ دستوری. در ریزطبقهٔ باز واژگانی، گروه‌هایی قرار می‌گیرند که صورت ظاهر آن‌ها، تمایز به خصوصی ندارد؛ ولی به اعتبار معنی خود معمولاً جایگاه ادات را در ساختمان بند اشغال می‌کنند. مانند: «همیشه»،

«هنوز»، «هرگز» «نه» و غیره. در ریزطبقهٔ بستهٔ دستوری، گروه‌هایی قرار می‌گیرند که دارای کاربردهای دستوری ویژه‌ای هستند؛ مانند «چنانچه»، «وقتی که»، «چون» و غیره (Ibid.p.171-181).

۴. روش تهیهٔ پیکره

۴. ۱. گردآوری داده‌های زبانی

پیکرهٔ پژوهش، برگرفته از یکی از آزمون‌های نگارش پایان دورهٔ عمومی و دورهٔ تکمیلی زبان فارسی فارسی‌آموزان مرکز آموزش زبان فارسی دانشگاه بین‌المللی امام خمینی(ره) است که در آن، تعداد ۹۰ فارسی‌آموز چینی در سطح عمومی و تعداد ۳۶ فارسی‌آموز چینی در سطح تکمیلی شرکت داشتند؛ از این رو، در مجموع ۱۲۶ برگهٔ آزمون، به‌عنوان داده‌های خام مورد استفاده قرار گرفته است.

آزمون پایانی نگارش دوره عمومی، شامل دو بخش است: در بخش اول، چهار تصویر که با شماره‌های یک تا چهار مشخص شده است، داستان کوتاهی را به نمایش گذاشته و از فارسی‌آموزان خواسته شده تا داستان تصاویر را با دست‌کم ۱۰ سطر ۱۰ واژه‌ای بنویسند. در بخش دوم، از فارسی‌آموزان خواسته شده که از دو موضوع داده شده، یکی را انتخاب کنند و متنی دست‌کم ۱۰ تا ۱۲ سطر ۱۰ واژه‌ای بنویسند. در یکی از موضوعات پیشنهادی، از فارسی‌آموزان خواسته شده، دربارهٔ یکی از مسافرت‌هایی که تاکنون رفته‌اند، متنی بنویسند. این که «قبل از سفر چه کارهایی انجام دادند؟»، «چه وسایلی با خود بردند؟»، «از چه مکان‌هایی دیدن کردند؟» و در موضوع دوم، خواسته شده، یک نامه برای استاد ایرانی خود بنویسند و کشور خود را از نظر آب و هوا، جمعیت، فرهنگ، تمدن و غیره برای او توضیح دهند.

در آزمون پایانی نگارش دورهٔ تکمیلی، از فارسی‌آموزان خواسته شده بود، دربارهٔ یکی از سه موضوع پیشنهادی دست‌کم ۱۵ سطر ۱۰ واژه‌ای بنویسند. موضوعات پیشنهادی؛ شامل اعلام نظر دربارهٔ گزاره‌هایی همچون «اولین معلّم‌های زندگی فرزندان، بودن پدر و مادر»، «رواج ادامهٔ تحصیل در دانشگاه بین مردم جهان» و «استفاده از فناوری پیشرفته در سال‌های اخیر» بوده و هر برگهٔ آزمون نگارش؛ شامل یک متن نوشتاری دربارهٔ موضوع انتخابی بوده است.

۴. ۲. برچسب‌گذاری داده‌ها

برای آماده‌سازی پیکره، ابتدا برگه‌های نگارش فارسی‌آموزان چینی، در نرم‌افزار واژه‌پرداز پیاده‌سازی شد. تلاش گردید تا حد امکان مطالب به همان‌گونه که فارسی‌آموزان در برگه‌های خود نوشته‌اند، تایپ شود. پس از پالایش و یکسان‌سازی داده‌های خام، کار برچسب‌گذاری دستوری مطالب تایپ شده در چهارچوب دستور مقوله و میزان و به‌صورت دستی انجام گرفت. در این مرحله، ۹ برچسب دستوری؛ شامل جمله، بند؛ بند مرتبه‌بندی شده و بند

واژگون‌مرتبه؛ بند خودایستا و بند ناخودایستا، گروه فعلی، گروه اسمی (متمم و مسند الیه) و گروه قیدی در نوشتار فارسی‌آموزان ثبت شد. برای نشانه‌گذاری، برچسب‌های دستوری از نمادهای زیر استفاده شده است.

جدول۲. نمادهای به‌کاررفته برای برچسب‌گذاری پیکره

نماد	واحد دستوری	نشانه
lll…lll	جمله	مرزنمای جمله
ll… ll	بند	مرزنمای بند (هسته و وابسته)
[[ ]]	بند	مرزنمای بند واژگون‌مرتبه
حروف سیاه	گروه فعلی	اسناد
خط نقطه چین	گروه قیدی	ادات
خط ممتد	گروه اسمی	مسند الیه
دو خط ممتد		متمم

از آن‌جا که نقطه به‌عنوان مرز پایان یک جمله و شروع جملهٔ دیگر منظور شده است، نقطه‌گذاری نوشته‌ها، توسط پژوهش‌گران، بازبینی و در صورت نیاز اصلاح یا تکمیل شده است. درادامه، کار شناسایی و تفکیک جمله‌ها انجام گرفته است. در تحلیل پیکره، اجزای بند واژگون‌مرتبه، به‌عنوان عناصر تشکیل دهندهٔ بند (گروه فعلی، گروه اسمی و گروه قیدی) شناسایی و محاسبه شده است. به‌عنوان‌مثال؛ فارسی‌آموزی نوشته: lll«بعداز (ادات) [[قهوه(متمم) **خوردن** (گروه فعلی ناخودایستا)]] علی به مدرسه **می‌رود**.lll. در این جمله، در واقع یک اسناد «می‌رود»، یک مسندالیه «علی» و دو ادات «به مدرسه» و «بعد از قهوه خوردن» داریم. ادات «بعداز قهوه خوردن»، از ادات «بعد از» همراه با بند ناخودایستای «قهوه خوردن» که به‌صورت واژگون‌مرتبه در درون گروه قیدی به‌کار رفته، ساخته شده است؛ بنابراین، در تحلیل این جمله، متمم «قهوه» و اسناد ناخودایستای«خوردن» نیز شناسایی شده است. در تحلیل گروه‌های اسمی دارای وابسته، تنها به احتساب یک گروه اسمی اکتفا شده است. به‌عبارتی دیگر؛ گروه‌های اسمی، درون گروه‌های قیدی مانند مثال (۴)، به‌صورت مجزا تحلیل نشده است.

(۴) پر (هسته) از چیزهای مختلف (وابسته)

گروه‌های قیدی نیز به‌صورت یک واحد شناسایی شده‌اند؛ بدین‌معنا که گروه‌های اسمی، درون گروه‌های قیدی به‌صورت مجزا برچسب‌گذاری نشده‌اند. به‌عنوان‌مثال؛ گروه قیدی« به تدریس معلّم» از ادات «به» و گروه اسمی«تدریس معلم» ساخته شده است و در پیکره، به‌صورت یک گروه قیدی برچسب‌گذاری شده است. تنها استثنای قابل ذکر، گروه قیدی دارای بند واژگون‌مرتبه است که توضیح آن در بالا داده شد. همچنین به‌دلیل

ضمیرانداز بودن زبان فارسی، در تعداد قابل توجهی از جمله‌های نوشتار فارسی‌آموزان، مسندالیه در قالب گروه اسمی مشخص نشده است.

در برچسب‌گذاری پیکرهٔ زبانی، تا حد امکان تلاش شد تا ملاک تحلیل اجزای متن، نوشتار فارسی‌آموزان باشد و متن نوشتاری بدون اعمال اصلاحات زبانی برچسب‌گذاری شود. به‌عنوان‌مثال؛ در یکی از نوشته‌ها، فارسی‌آموزی نوشته: «علی با مینا در آشپزخانه صبح خانه را متنظر **بود**». این جمله، در تحلیل شامل یک اسناد «بود»، یک مسندالیه «علی»، دو متمم «صبح خانه» و «منتظر» و دو ادات «با مینا» و «در آشپزخانه» ثبت شده است؛ در صورتی‌که جملهٔ درست و مورد نظر فارسی‌آموز، عبارت «علی با مینا در آشپزخانه منتظر صبحانه **بود**.» است و دربردارندهٔ یک اسناد «بود»، یک مسندالیه «علی»، یک متمم «منتظرِ صبحانه» و دو ادات «با مینا» و «در آشپزخانه» است. اشتباهات املایی؛ مانند« صبح خانه»، تأثیری بر تحلیل دستوری نداشته است. در ادامه، نمونه‌ای از برچسب‌گذاری نوشته نخستین فارسی‌آموز چینی سطح فرامیانی ارائه می‌گردد:

فارسی‌آموز نخست:|||من *موافق* **هستم،**|| پدر و مادر، *اولین معلم‌های زندگی فرزندان‌شان* **هستند.**||| وقتی‌که ما *بچه* **بودیم،**|| ما بیشتر وقت با مادر و پدرمان *زندگی* **می‌کردیم.**||| وقتی‌که ما به مدرسه **نرفتیم،**|| ما بیشتر با پدر و مادرمان با هم *درس* **خواندیم.**||| پدر و مادرمان **توانستند**|| به ما *معنی دانش‌های زندگی یاد* **بدادند.**||| ما نیز **توانستیم**|| زودتر *دانش‌های زندگی یاد* **بگرفتیم.**||| *دانش‌های زندگی* مانند: چطور *اتاق را تمیز* **می‌کند،**|| چطور با معلم و هم کلاس *صحبت* **می‌کند،**|| چطور *درس* **می‌خواند** و....||| وقتی‌که ما *اتاق را تمیز* **می‌کردیم**|| پدر و مادرمان به ما **گفت:**|| تو **باید** *زباله داخل سطل زباله* **گذاشتی،**|| کتاب روی میز **باید** *مرتب* **بود**|| و پنجره **باید** *باز* **کرد** و....||| وقتی‌که ما به مدرسه **رفتیم،**|| پدر و مادرمان به ما **گفت:**|| در کلاس *درس را توجه* **کردی،**|| بیشتر با معلم و هم کلاس *صحبت* **کردی**|| و حتماً **نباید** با هم کلاسی *دعوا* **کردی** و...||| وقتی‌که ما *درس* **خواندیم،**|| پدر و مادرمان به ما **گفت:**|| **باید** دقیق *درس* **خواندی،**|| دقیق *جمله* **خواندی**|| و **نباید** *دانش‌های اشتباه حفظ* **کردی**|| و الان پدر و مادرمان هم *یکسان* **هستند.**||| گاهی ما *عصبانی* **می‌شویم**|| چون *** ما *بچه* **نیستیم.**||| اما پدر و مادرمان هنوز به ما زیاد *حرف* **می‌گویند.**||| مانند ما *بچه* **بودیم؛**|| از جاده *گذر* **می‌کنی،**|| **باید** *قواعد را رعایت* **کنی**|| و وقتی‌که چراغ *قرمز* **است**|| **باید بمانی؛**|| چراغ *سبز*|| از جاده *گذر* **کنی.**||| اگر پیر مرد و زن از جاده *گذر* **می‌کنند،**|| **باید** به آنها *کمک* **کنی.**||| والین *زیاد حرف* به ما **می‌گویند**|| جون آنها *ما را دوست* **دارند.**||| انها هم *معلم خوب* **هستند.**|||

۴. ۳. سطح فارسی‌آموزان

فارسی‌آموزانی که در آزمون پایان دورهٔ عمومی زبان فارسی مرکز آموزش زبان فارسی دانشگاه بین‌المللی امام

خمینی (ره) حضور داشتند، مدت ۱۶ هفته و هر هفته ۲۰ ساعت، درمجموع ۳۲۰ ساعت، در کلاس‌های آموزش حضوری برای مهارت‌های چهارگانهٔ شنیدن، خواندن، صحبت کردن و نوشتن شرکت کرده بودند. با در نظر گرفتن کیفیت و کمیت برنامهٔ آموزشی و ویژگی فردی فارسی‌آموزان چینی‌زبان، می‌توان سطح عمومی را معادل سطح فراپایه(A2) در چهارچوب مرجع مشترک اروپا در نظر گرفت. در این چهارچوب، زبان‌آموز سطح فراپایه می‌تواند جمله‌ها و اصطلاحات پرکاربرد مربوط به کاربردهای ضروری؛ مانند اطلاعات ابتدایی فردی و خانواده، خرید کردن، مکان، مشاغل و غیره را درک کند. می‌تواند در فعالیت‌های روزمره و ساده که نیازمند تبادل اطلاعات ساده و مستقیم دربارهٔ خانواده و مسائل روزمره است، با دیگران ارتباط برقرار کند. می‌تواند وجوه سادهٔ گذشتهٔ خود و نیز محیط اطراف و مسائل را در زمینه نیازهای ضروری، توصیف کند (Sahraei & Marsoos,2016).

فارسی‌آموزانی که در آزمون پایان دورهٔ تکمیلی زبان فارسی مرکز آموزش زبان فارسی دانشگاه بین‌المللی امام خمینی(ره) نیز حضور داشتند، مدت ۳۲ هفتهٔ ۲۰ ساعته و درمجموع ۶۴۰ ساعت در کلاس‌های آموزش حضوری برای مهارت‌های چهارگانهٔ شنیدن، خواندن، صحبت کردن و نوشتن شرکت کرده بودند. با در نظر گرفتن کیفیت و کمیت برنامهٔ آموزشی و ویژگی فردی فارسی‌آموزان چینی‌زبان، می‌توان سطح تکمیلی را معادل سطح فرامیانی(B2)در چهارچوب مرجع مشترک اروپا در نظر گرفت. در این چهارچوب، زبان‌آموز سطح فرامیانی می‌تواند مفهوم اصلی متن پیچیده با موضوعات عینی یا انتزاعی را درک کند. در بحث و گفت‌وگو در حوزهٔ تخصصی خود شرکت کند. می‌تواند با میزانی از روانی کلام و فی‌البداهگی در تعاملات معمولی با گویشور زبان بدون مشکل، شرکت کند. می‌تواند متن واضح و با جزئیات در دامنهٔ وسیعی از موضوعات تولید کند و دیدگاه خود را دربارهٔ موضوع تبیین و نقاط قوت و ضعف مختلف دربارهٔ موضوع را توضیح دهد (Sahraei & Marsoos,2016).

۵. یافته‌های پژوهش

تهیهٔ پیکره‌ای از نوشتار فارسی‌آموزان غیرایرانی، به‌دلیل ویژگی‌های خاص نوشته‌ها؛ همچون وجود خطاهای زبانی و نگارشی و نیز وجود کاستی‌هایی در انسجام‌بخشی به نوشته و نشانه‌گذاری آن، نسبت به نوشتار فارسی‌زبانان، نیازمند تخصص و تلاش مضاعف است؛ از این رو می‌توان گفت؛ مهم‌ترین دستاورد پژوهش، تهیهٔ نسخهٔ اولیهٔ پیکره زبان‌آموز فارسی‌آموزان غیرایرانی، با مشخصاتی است که در ادامه به آن اشاره می‌شود. درمجموع، ۱۲۶ نوشتهٔ فارسی‌آموزان چینی در دو سطح فراپایه (۹۰ نوشته) و فرامیانی (۳۶ نوشته)، به‌عنوان داده‌های خام پیکرهٔ نوشتار فارسی‌آموزان چینی مورد استفاده قرار گرفت؛ بنابراین، پیکره مجموعاً از ۱۲۶ متن نوشتاری؛ شامل ۲۱۲ پاراگراف و ۲۹۸۵۷ واژه تشکیل شده است. همچنین پیکرهٔ ایجاد شده، مجموعاً حاوی ۳۱۷۵ جمله، ۴۹۱۲ بند، ۱۹۳۶۹ گروه؛ شامل ۴۹۱۲ گروه فعلی، ۸۷۶۰ گروه اسمی و ۴۹۱۲ گروه قیدی؛ شامل ادات و گروه‌های حرف اضافه‌ای است. در جدول زیر، واحدهای دستوری موجود در پیکره، به تفکیک هر واحد دستوری و نیز هر گروه از فارسی‌آموزان چینی مشاهده می‌گردد.

جدول۱. آمار واحدهای دستوری موجود در پیکرهٔ ایجاد شده

واحد دستوری	سطح فراپایه	سطح فرامیانی	مجموع
جمله	۲۶۶۴	۵۱۱	۳۱۷۵
بند	۳۶۷۳	۱۲۳۹	۴۹۱۲
گروه	۱۴۵۲۹	۴۸۴۰	۱۹۳۶۹
گروه فعلی	۳۶۷۳	۱۲۳۹	۴۹۱۲
گروه اسمی	۶۶۰۹	۲۱۵۱	۸۷۶۰
گروه قیدی	۴۲۴۷	۱۴۵۰	۵۶۹۷

پژوهش حاضر همچنین نشان داد که دستور مقوله و میزان، با وجود این که در بین نظریه‌های دستوری معاصر، جایگاه برجسته‌ای ندارد و دستور نظام‌مند – نقش‌گرا که نسخهٔ جدیدتر آن است، توجهات زیادی را به خود جلب کرده است، کارایی لازم برای توصیف دقیق نوشتار فارسی‌آموزان چینی را داراست. یکی از مهم‌ترین علل آن، ویژگی بارز کتاب «توصیف ساختمان دستوری زبان فارسی» (Bateni,2013) است که در آن، توصیفی نسبتاً جامع، روشن و دقیق زبان فارسی در چهارچوب دستوری مقوله و میزان ارائه نموده است و این موضوع سبب گردیده، با وجود این که کاستی‌هایی در نظریهٔ مقوله و میزان و به‌دنبال آن، ضعف‌هایی در توصیف ساختمان زبان فارسی باطنی قابل ملاحظه و نقد است؛ ولی آن چهارچوب به ظاهر قدیمی، در برچسب‌گذاری پیکرهٔ حاضر، کارآمدی لازم را نشان داد. توصیف دستوری نوشته‌های فارسی‌آموزان چینی سطح فراپایه و فرامیانی، دلیلی بر این مدعاست.

۶. جمع‌بندی و نتیجه‌گیری

پژوهش حاضر، گامی در راستای ایجاد پیکرهٔ زبان‌آموز فارسی‌آموزان غیرایرانی در چهارچوب دستور مقوله و میزان است که به‌صورت دستی، ده نوع برچسب دستوری بر داده‌های زبانی آن اعمال شده است. تهیهٔ پیکرهٔ زبانی فارسی‌آموزان، با استفاده از نوشته‌های فارسی‌آموزان چینی سطح فراپایه و فرامیانی، نشانگر آن است که با وجود ویژگی‌های خاص نوشتار فارسی‌آموزان، امکان عملیاتی کردن طرح تهیهٔ پیکرهٔ زبانی فارسی‌آموزان غیرایرانی وجود دارد. این پژوهش، همچنین کارایی دستور توصیفی باطنی را که مبتنی بر دستور مقوله و میزان است، در برچسب‌گذاری نحوی نوشتار فارسی‌آموزان چینی تأیید کرد. تهیهٔ چنین پیکره‌ای، به پژوهشگران، مدرسان و دانشجویان مقطع کارشناسی ارشد و دکتری رشتهٔ آموزش زبان فارسی به غیرفارسی زبانان کمک می‌کند تا پژوهش‌های گوناگونی را با اتّکا بر آن به سرانجام برسانند و به دستاوردهای علمی باارزشی برسند. از آنجا که پیکرهٔ پژوهشی حاضر، با حمایت بخش معاونت پژوهشی دانشگاه بین‌المللی امام خمینی(ره) انجام گرفته است، کلیهٔ حقوق مادی و معنوی حاصل از آن، متعلّق به آنجاست؛ ولی امید است، پیکرهٔ زبان‌آموز حاضر در قالب کتاب

و یا نرم‌افزار، در دسترس پژوهش‌گران و علاقه‌مندان قرار گیرد. انتظار می‌رود؛ در ادامه تهیهٔ نسخه نخست پیکرهٔ زبانی فارسی‌آموزان غیرایرانی، داده‌های نوشتاری و گفتاری بیشتری از فارسی‌آموزان ملل گوناگون به پیکره افزوده شود و همچنین امید آن است که نسخه‌های آتی پیکره به‌صورت نرم‌افزاری آماده بهره‌بردای گردد تا برچسب‌گذاری اطلاعات دستوری و تهیهٔ گزارش‌های لازم مطابق استانداردهای جهانی صورت پذیرد.

فهرست منابع:

باطنی، محمد رضا. (۱۳۹۲). *توصیف ساختمان دستوری زبان فارسی*. چاپ سی‌ام، تهران، انتشارات امیرکبیر.

بی‌جن‌خان، محمود. (۱۳۸۳). *نقش پیکره‌های زبانی در نوشتن دستور زبان: معرفی یک نرم‌افزار رایانه‌ای*. مجله زبان‌شناسی. ۱۹(۲)، ۴۸-۶۷.

بی‌جن‌خان، محمود. (۱۳۹۵). *پیکرهٔ گفتار محاوره‌ای زبان فارسی امروز*. مجموعه مقالات دومین همایش ملی زبان‌شناسی پیکره‌ای. تهران: نشر نویسه پارسی.

جهانگردی، کیومرث. (۱۳۹۵). *تحلیل محتوای کتاب‌های آموزش زبان فارسی به غیرفارسی زبانان*. رساله دکتری. تهران: پژوهشگاه علوم انسانی و مطالعات فرهنگی.

صحرایی، رضامراد؛ مرصوص، فائزه. (۱۳۹۵). *استاندارد مرجع آموزش زبان فارسی*. تهران: انتشارات دانشگاه علّامه طباطبایی.

صفری، سعید (۱۳۹۴). *از زبان‌شناسی پیکره‌ای تا پیکرهٔ زبان‌آموز*. مجموعه مقالات نخستین همایش ملی زبان‌شناسی پیکره‌ای. تهران: نشر نویسه پارسی.

قربان‌زاده، ف. (۱۳۹۴). معرفی پیکره *فارسی روز*. مجموعه مقالات نخستین همایش ملی زبان‌شناسی پیکره‌ای. تهران: نشر نویسه پارسی.

میرزایی، آزاده؛ صفری، پگاه. (۱۳۹۴). *ساخت‌واژه – متن‌های تخصصی و عمومی زبان فارسی، بر اساس بسامدگیری واژه‌های نقشی و محتوایی*. در مجموعه مقالات نخستین همایش ملی زبان‌شناسی پیکره‌ای. تهران: نشر نویسه پارسی.

References:

Assi, S. M. (1997). Farsi linguistic database (FLDB). International Journal of Lexicography, 10(3), 5.

Bateni , M.R.(2013). Description of Persian Grammatical Structure (30rd Ed). Tehran:Amir kabir.[In Persian]

Bi Jen Khan, M.(2016). The Corpus of Contemporary Colloquial Persian. Proceedings of the Second National Conference on Corpus Linguistics. Tehran: Nevise-e-Parsi.

Bijankhan, M., Sheikhzadegan, J., Roohani, M. R., Samareh, Y., Lucas, C., & Tebyani, M. (1994). FARSDAT-The Speech Database of Farsi Spoken Language. The Proceedings of the Australian Conference on Speech Science and Technology, 2, ۸۲۶–۸۳۱.

Bijankhan, M., Sheykhzadegan, J., Bahrani, M., & Ghayoomi, M. (2011). Lessons from building a Persian written corpus: Peykare. Language Resources and Evaluation, 45(2), ۱۴۳–۱۶۴.

Eghbalzadeh, H., Hosseini, B., Khadivi, S., and Khodabakhsh, A. (2012, November). Persica: A Persian Corpus for Multipurpose Text Mining and Natural Language Processing. In Sixth International Symposium on Telecommunications (IST). IEEE. Tehran.

Ghorbanzadeh, F. (2015). Introducing the Contemporary Persian Corpus. Proceedings of the First National Conference on Corpus Linguistics. Tehran: Nevise-e-Parsi.

Halliday, M.A.K., & Matthiessen, C. M. I. M. (2004). An introduction to functional grammar (3rd ed.). London: Arnold.

Jahangardi, K. (2016). An Analysis of Textbooks for Teaching Persian to Non-Persians: A Corpus-Cognitive Approach to Teaching Vocabulary. Ph.D.Thesis. Tehran: Institute for Humanities & Cultural Studies.

Rasooli, M. S. Kouhestani, M. and Moloodi, A. S. (2013). Development of a Persian Syntactic Dependency Treebank. In The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT), Atlanta, USA.

Safari, S. (2015). From Corpus Linguistics to Learner Corpus. Proceedings of the First National Conference on Corpus Linguistics. Tehran: Nevise-e-Parsi.

Sahraei, R.M.& Marsoos,F.(2016). *Persian Teaching Reference Standard.* Tehran: Allameh Tabatabai publications.[In Persian]

Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Famian, A., Bagherbeigi, S., Fekri, E. and et al. (2010). Semi Automatic Development of Farsnet; the Persian Wordnet. Proceedings of 5th Global WordNet Conference (GWA2010). Mumbai, India.

Mirzaei,A.& Safari,P.(2014). Building specialized and general documents in Persian based on the frequency of function and content words. Mirzaei,A., *proceeding of 1st National Conference on Corpus Linguistics*(175-192), Tehran: Neviseh Parsi. [In Persian]

Mirzaei, A., and Safari, P. (2018). Persian Discourse Treebank and Coreference Corpus. In LREC 2018, 4049-4055.