

A Corpus-Based Analysis of Middle Persian Texts Based on the Pārsīg Database

Farzaneh Goshtasb^{*}, Masood Ghayoomi^{}**

Nadia Hajipour^{*}**

Abstract

Recent attitude towards studying a language or a linguistic phenomenon is based on the existence of a collection of linguistic data. Therefore, it is required to develop a linguistic corpus that has been compiled from the real, naturally-occurred speech and not on the basis of an individual's intuition. This research methodology is highly important for studying historical linguistic data, as part of the dead languages which have no speakers now. The present research aims to develop a linguistic corpus of Middle (Pahlavi) Persian and to organize it in a database. To this end, six information layers are defined in the annotation process, including transliteration of the Pahlavi texts, transcription of the words along with their Persian translation, defining fine-grained syntactic category of the words, lemmatizing the words, and identifying whether the word is huzwāreš or not. To define fine-grained syntactic categories, the tag set for contemporary Persian developed by Bijankhan et al. (2011) and organized by Ghayoomi (2004) are modified and adapted to the Pahlavi language according to the requirements. The new tag set is used to label Pahlavi

* Associate Professor, Department of Ancient Culture and Languages, Research Institute of Linguistics, Institute for Humanities and Cultural Studies, Tehran, Iran (Corresponding Author), f.goshtasb@ihcs.ac.ir

** Assistant Professor, Department of Linguistics, Research Institute of Linguistics, Institute for Humanities and Cultural Studies, Tehran, Iran, m.ghayoomi@ihcs.ac.ir

*** Research Expert, Department of Ancient Culture and Languages, Research Institute of Linguistics, Institute for Humanities and Cultural Studies, Tehran, Iran, nadiahajipour@yahoo.com

Date received: 15/02/2021, Date of acceptance: 14/05/2021

Copyright © 2010, IHCS (Institute for Humanities and Cultural Studies). This is an Open Access article. This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/> or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

words. After annotating words and organizing the information, extracting the statistical information is possible to deepen the insight over the text content.

Keywords: Pahlavi Language, Middle Persian, Corpus Linguistics, Database, Annotation.

تحلیل پیکره بنیان متون فارسی میانه بر مبنای پایگاه داده پارسیگ

فرزانه گشتاسب*

مسعود قیومی**، نادیا حاجی‌پور***

چکیده

رویکرد نوین در مطالعات زبان‌شناختی یا یک پدیده زبانی بر اصل وجود مجموعه‌ای از داده‌های زبانی گردآوری شده نهادینه شده است؛ بنابراین به تهیه یک پیکره زبانی نیاز است که از تولیدات واقعی گویشوران و نه براساس شمّ زبانی فردی گردآوری شده است. این شیوه پژوهشی برای بررسی داده‌های زبانی تاریخی که جزء زبان‌های مرده است و اکنون هیچگونه گویشوری ندارد از اهمیت به‌سزایی برخوردار است. هدف از انجام این پژوهش، تهیه پیکره زبان پهلوی ساسانی (فارسی میانه) و ساماندهی آن در یک پایگاه است. برای هر واژه، شش لایه اطلاعاتی، اعم از حرف‌نویسی متن پهلوی، آوانویسی واژه‌ها به‌همراه ترجمه فارسی آنها، تعیین مقوله دستوری دانه‌ریز واژه‌ها، بن‌واژه‌سازی واژه‌ها و تعیین هزوارش بودن آنها، تعریف شده است. برای مقوله دستوری دانه‌ریز واژه‌ها، مجموعه برچسب مقولات دستوری فارسی معاصر تهیه شده توسط بی‌جن‌خان و همکاران (۲۰۱۱) و ساختارمندشده توسط قیومی (۲۰۱۴) با توجه به نیازهای زبان پهلوی جرح و تعدیل شده است و از مجموعه جدید برای برچسب‌گذاری واژه‌های پهلوی استفاده شده است. پس از نشانه‌گذاری واژه‌ها و ساماندهی اطلاعات، امکان استخراج اطلاعات

* دانشیار گروه فرهنگ و زبان‌های باستانی، پژوهشکده زبان‌شناسی، پژوهشگاه علوم انسانی و مطالعات فرهنگی (نویسنده مسئول)، f.goshtasb@ihcs.ac.ir

** استادیار گروه زبان‌شناسی، پژوهشکده زبان‌شناسی، پژوهشگاه علوم انسانی و مطالعات فرهنگی، m.ghayoomi@ihcs.ac.ir

*** کارشناس پژوهشی گروه فرهنگ و زبان‌های باستانی، پژوهشکده زبان‌شناسی، پژوهشگاه علوم انسانی و مطالعات فرهنگی، nadijahajipour@yahoo.com

تاریخ دریافت: ۱۳۹۹/۱۱/۲۷، تاریخ پذیرش: ۱۴۰۰/۰۲/۲۴

آماري وجود دارد که می‌تواند بینش عمیق‌تری از محتوای متن منتقل نماید. از این‌رو، اطلاعات آماری از پیکره به‌دست‌آمده استخراج شده و توضیح داده می‌شود تا دورنمای کلی نسبت به منابع تشکیل‌دهنده این پیکره به‌دست‌آید.

کلیدواژه‌ها: زبان پهلوی، زبان‌شناسی پیکره‌ای، پایگاه داده، نشانه‌گذاری داده

۱. مقدمه

دورهٔ میانهٔ زبان‌های ایرانی، شامل دوره‌ای نسبتاً طولانی (حدود ۱۲۰۰ سال) از پایان حکومت هخامنشیان تا قرن سوم هجری است. از این دوره، اسنادی مکتوب از شش زبان فارسی میانه (پهلوی ساسانی)، پهلوی اشکانی، سغدی، بلخی، ختنی و خوارزمی به دست ما رسیده‌است. زبان فارسی میانه یکی از مهم‌ترین این زبان‌ها و مادر زبان فارسی معاصر به شمار می‌آید، و تنها زبان ایرانی است که از دورهٔ باستانی آن اسناد نوشتاری که شامل کتیبه‌های فارسی باستان نیز می‌شود به‌دست آمده‌است.

زبان فارسی میانه (پهلوی ساسانی) که از این پس در این مقاله آن را «پهلوی» می‌خوانیم، زبان نوشته‌های فارسی میانه زردشتی، کتیبه‌های ساسانی، متون مانوی و متون فارسی میانه مسیحی است که البته هر یک از این متون با خط‌های متفاوت نوشته شده‌است. بیشترین آثاری که از زبان پهلوی به‌جای مانده‌است نوشته‌های دینی زرتشتی است که به خط فارسی میانه کتابی یا شکسته نوشته شده‌است و پس از آن (از نظر تعداد آثار) متون فارسی میانه مانوی نوشته‌شده به خط مانوی قرار می‌گیرد.

آثار به‌جامانده از متون پهلوی دربارهٔ موضوعات مختلفی به دست ما رسیده‌است. بیشترین کتاب‌ها مربوط به متون زند یعنی ترجمه و تفسیر اوستا به زبان پهلوی و کتاب‌هایی است که بر اساس متون زند و ترجمه‌های اوستا تدوین شده‌است، مانند «دینکرد»، «بندش»، «گزیده‌های زادسپرم»، «دادستان دینی»، «روایات پهلوی» و جز آن. دیگر موضوعات کتاب‌های پهلوی عبارت است از: الف) متون فلسفی-کلامی، مانند «شکند گمانیک و زار»، «پس دانش کامگ و گجستک ابالیس»؛ ب) متون کشف و شهود و پیشگویی، مانند «ارداویراف‌نامه»، «زند و همن یسن»، «پیشگویی‌های جاماسب»، «جاماسب‌نامه»، «یادگار جاماسبی» و «بهرام ورجاوند»؛ پ) اخلاقیات و اندرزنامه‌ها شامل تعداد زیادی اندرزنامه، مانند «یادگار بزرگمهر»، «اندرز پوریوتکیشان»، «دادستان مینوی خرد» و جز آن؛ ت) مناظره و مفاخره، مانند «درخت آسوری»؛ ث) تاریخ و جغرافیا، مانند

«کارنامه اردشیر بابکان» و «شهرستان‌های ایران»؛ ج) حماسه، مانند «یادگار زیران»؛ چ) فقه و حقوق، مانند «شایست نشایست»، «روایت امید اشووهیشتان»، «روایت آذر فرنبغ فرخزادان»، «روایت آذر فرنبغ سروش»، «مادیان هزار دادستان»، «پرسش‌های هیربد اسفندیار»؛ ح) رسالات کوچک تعلیمی، مانند «خسرو و ریدگ»، «شگفتی‌های سیستان»، «گزارش شطرنج و وضع نرد»، «ماه فرودین و روز خرداد»، «سور سخن» و «آیین نامه‌نویسی»؛ خ) فرهنگ‌ها و واژه‌نامه‌ها، مانند «فرهنگ پهلوی» و «فرهنگ اویم ایوک». این متون جز این که گنجینه‌ای ارزشمند برای شناخت پیشینه فرهنگی، دینی و تاریخی سرزمین ایران محسوب می‌شود، مطالعه دقیق زبان‌شناختی آنها، برای مطالعات زبان فارسی، بررسی‌های تاریخی این زبان و واژه‌گزینی بسیار سودمند است.

با وجود پیشینه غنی اطلاعات از متون پهلوی، اهمیت مطالعه دقیق زبان‌شناختی این متون و نقش آن در مطالعات زبان فارسی امری بدیهی و آشکار است. نخستین قدم در این راه، ساماندهی این متون و فراهم‌آوردن همه اطلاعاتی است که برای مطالعه این زبان به کار می‌آید، از جمله املائی واژه براساس خط فارسی میانه کتابی، حرف‌نویسی و آوانویسی واژه‌ها، معنا، همه جزئیات درباره مقوله دستوری واژه، بن‌واژه، ارجاع متنی و ارجاع به دستنویس‌های موجود.

آنچه در شیوه پژوهش‌های پیشین می‌توان مشاهده کرد این است که تمامی تحلیل‌ها به‌صورت دستی بوده و تمام اطلاعات استخراج‌شده از متن در قالب واژه‌نامه‌ای که خارج از بافت زبانی است در انتهای کتاب ذخیره شده‌است.

روشن است که گردآوری همه این اطلاعات برای همه متون پهلوی، به شیوه‌های سنتی کاری تقریباً ناممکن است. به همین دلیل به نظر می‌رسد با استفاده از فناوری اطلاعات می‌توان یک پیکره زبانی از متون پهلوی تهیه کرد و ضمن گردآوری داده‌های زبان پهلوی به‌صورت الکترونیکی و غنی‌سازی متون با اضافه‌کردن اطلاعات زبان‌شناختی، این حجم از اطلاعات را در قالب یک پایگاه داده ساماندهی نمود. وجود این اطلاعات زبان‌شناختی و استخراج اطلاعات آماری به پژوهشگر کمک می‌کند تا وی به درک عمیق‌تر و خوانش جدیدتری از متون دست یابد.

۲. پیشینه پژوهش

برای زبان پهلوی فرهنگ جامع واژگان و واژه‌نامه‌ای که شامل واژه‌های تمامی متون پهلوی با ارجاع به متن باشد، وجود ندارد. «فرهنگ زبان پهلوی» نوشته فرهوشی (۱۳۸۶)، «فرهنگ کوچک پهلوی» نوشته مکزی (۱۹۷۱) و ترجمه فارسی آن در (۱۳۷۹) و «دست‌نامه پهلوی» نوشته نیبرگ (۱۹۶۴-۱۹۷۴) تنها واژه‌نامه‌هایی هستند که برای متون فارسی میانه زردشتی تدوین شده‌است و شامل آوانویسی، حرف‌نویسی و معنای واژه می‌باشد.

همچنین بسیاری از متون فارسی میانه آوانویسی شده و به زبان‌های گوناگون ترجمه شده‌است و برای برخی از آنها واژه‌نامه نیز تهیه شده‌است. از جمله شاخص‌ترین پژوهش‌هایی که می‌توان به آنها اشاره کرد عبارت است از کتاب‌های «گزیده‌های زادسپرم» و «دینکرد هفتم» از راشد‌محصل (۱۳۸۵؛ ۱۳۸۹)، «روایت پهلوی» از ویلیامز (۱۹۹۰) و «میرفخرایی (۱۳۶۷)، و «بررسی دینکرد ششم» از شاکد (۱۹۷۹) و «میرفخرایی (۱۳۹۳) و «ارداویراف‌نامه» از ژینیو (ترجمه فارسی در ۱۳۸۲) که همگی دارای واژه‌نامه به زبان پهلوی بوده و بجز املای پهلوی واژه و آوانویسی آن، معنای فارسی و ارجاع به متن نیز در واژه‌نامه آنها آمده‌است. متون دیگری که از زبان فارسی میانه منتشر شده‌است اغلب دارای ترجمه فارسی و تعلیقات و گاه آوانویسی و فاقد واژه‌نامه است مانند «شایست‌ناشایست» و «بررسی دستنویس م. او ۲۹» از مزداپور (۱۳۶۹؛ ۱۳۷۸)، «روایت امید اشوهِیشتان» از صفای اصفهانی (۱۳۷۶)، «بندش» از انکلساریا (۱۹۰۸) و بهار (۱۳۶۹)، «مینوی خرد» از انکلساریا (۱۹۱۳) و تفضلی (۱۳۷۹)؛ «دادستان دینی» از انکلساریا (۱۹۱۱)، جعفری دهقی (۱۹۹۸) و میرفخرایی (۱۳۹۷)، «متون پهلوی» از جاماسب‌آسانا (۱۹۱۳) و عریان (۱۳۷۱) جز آن (برای آگاهی از پژوهش‌هایی که روی متون پهلوی انجام شده‌است نک: تفضلی، ۱۳۷۸، صص ۱۱۱-۳۲۴؛ ماکوچ، ۲۰۰۹، صص ۱۱۶-۱۹۶). بجز واژه‌نامه‌هایی که در انتهای این کتاب‌ها فراهم آمده‌است، برای متونی چون «گزیده‌های زادسپرم» (بهار، ۱۳۵۱)، «شایست‌ناشایست» (طاووسی، ۱۳۶۵) و «بندش» (بهار، ۱۳۴۵) واژه‌نامه جداگانه نیز تهیه شده‌است. برای بخش‌هایی از متون فارسی میانه، همین پژوهش‌ها نیز انجام نشده‌است و هنوز آوانویسی و ترجمه آنها فراهم نیامده‌است که مهم‌ترین آنها، متون «زند اوستا» و فصل‌هایی از دینکرد است.^۱

همه این پژوهش‌ها تاکنون فقط به صورت کتاب منتشر شده‌است و اقدامی برای جمع‌آوری داده‌های زبانی آنها در یک پایگاه اطلاعاتی واحد انجام نشده‌است؛ حتی به دلیل

حجم زیاد داده‌ها، واژه‌نامه‌ای جامع به صورت کتاب نیز برای متون فارسی میانه تهیه نشده‌است. «پهلواژ» (فرهنگ جامع فارسی میانه- فارسی نو) را می‌توان نخستین واژه‌نامه برخط موجود بر شمرده که در آن اطلاعات واژه‌های نه متن فارسی میانه مشتمل بر آوانویسی، معنای فارسی و ارجاع به متن پهلوی ثبت شده‌است. پیکره‌های استفاده‌شده در شکل‌گیری این فرهنگ عبارت است از: «زند بهمن‌یسن»، «روایت پهلوی»، «گزیده‌های زادسپرم»، «ماتیگان یوشت فریان»، «ارداویراف‌نامه»، «شایست ناشایست»، «کارنامه اردشیر بابکان»، «متن‌های پهلوی» و «مینوی خرد». این واژه‌نامه را گروهی از استادان پژوهشگاه علوم انسانی در طول سال‌های ۱۳۶۰-۱۳۷۰ تهیه کردند ولی به دلیل مشکلاتی که برای حروف چینی دشوار آن در آن زمان وجود داشت، هیچگاه به مرحله چاپ نرسید و برگیزه‌های آن بایگانی شد. از سال ۱۳۸۵ مجدداً کار تصحیح و بازبینی برگیزه‌ها آغاز شد. اکنون حدود نیمی از واژه‌های این فرهنگ پس از بازنگری مجدد، حروف چینی شده و در وبگاهی به همین نام بارگذاری شده‌است.^۲ این واژه‌نامه، چنان که گفته شد، دارای اطلاعات بسیار محدودی درباره هر واژه است و امکان توسعه آن به صورت پایگاه داده میسر نیست (نک: مزدپور و دیگران، ۱۳۹۰، صص ۹۷-۱۱۷).

مهم‌ترین اقدامی که برای تهیه پایگاه داده زبان پهلوی انجام شده‌است، طرحی به نام «واژه‌نامه فارسی میانه» Middle Persian Dictionary Project است که حدود بیست سال پیش، در سال ۲۰۰۱ در همایشی که در رم برگزار شد، برای نخستین بار مطرح گردید. این طرح شامل جمع‌آوری تمام متون فارسی میانه کتابی، مانوی و کتیبه‌ای، متون پازند و همچنین وام‌واژه‌های فارسی میانه در متون دیگر بود که هم به صورت کتاب و هم واژه‌نامه دیجیتال و پایگاه داده در دسترس قرار می‌گرفت. این طرح قرار بود به سرپرستی شائول شاکد و همکار او کارلو چرتی و به صورت یک طرح مشترک با همکاری آکادمی علوم تجربی و انسانی اسرائیل The Israel Academy of Sciences and Humanities، آکادمی لینیسی رم The Accademia dei Lincei، مؤسسه ایتالیایی ایسانو (مؤسسه مطالعات آفریقا و شرق) Istituto Italiano per l'Africa e l'Oriente (ISIAO)، دانشگاه سایانزای رم Sapienza Università degli Studi di Napoli، دانشگاه شرق‌شناسی ناپل Università di Roma "L'Orientale" و دانشگاه عبری اورشلیم Hebrew University of Jerusalem انجام شود (شاکد و چرتی، ۲۰۰۵: ۱۸۹). تا امروز داده‌هایی که برای این پایگاه جمع‌آوری شده‌است، در دسترس عموم قرار نگرفته است و تنها اطلاعاتی که از این طرح در دست است، مطلبی

کوتاه در سایت آکادمی علوم تجربی و انسانی اسرائیل^۳ و مقاله‌ای در معرفی طرح‌نامه این پایگاه است که در سال ۲۰۰۵ در مجموعه مقالات همایش رم منتشر شد (همان: ۱۸۱-۱۹۰).

هم‌زمان با مطرح شدن طرح «واژه‌نامه فارسی میانه»، در سال ۱۹۹۹، طرحی با عنوان «واژه‌نامه متون مانوی» The Dictionary of Manichaean Texts در دانشگاه لندن، در مدرسه مطالعات شرقی و آفریقایی SOAS School of Oriental and African Studies (SOAS) با بودجه بیش از نیم میلیون پوند معرفی شد. زمان این طرح پنج سال تعیین شده بود و قرار بود پایگاه داده و واژه‌نامه‌ای کامل از تمام واژه‌های متون مانوی تهیه و در دسترس قرار گیرد. خبر کوتاهی درباره این طرح در سایت دانشگاه لندن منتشر شده است.^۴

در حال حاضر سایت «تزاروس زبان‌های هندواروپایی و مواد متنی» (تیتوس) Thesaurus indogermanischer Sprach- und Textmaterialien (TITUS) تنها پایگاه داده‌ای است که تعدادی از متون پهلوی را به صورت برخط در اختیار پژوهشگران قرار می‌دهد. در این سایت متون بسیاری از زبان‌های زنده و مرده دنیا بارگذاری شده‌است.^۵ ایده ایجاد این پایگاه در سال ۱۹۸۷ در کنفرانس هندی‌اروپایی در لایدن مطرح شد و در همان سال در جلد ۲ شماره ۳۲ مجله «زبان» Sprache در فراخوانی با عنوان «تزاروس زبان‌های هندواروپایی روی دیسکت» Thesaurus indogermanischer Textmaterialien auf Datenträgern مطرح شد. این ایده هشت سال پروارنده شد و در نهایت منجر به ایجاد «پایگاه تیتوس» گردید. هدف این پایگاه چنانکه گفته شد، در دسترس قراردادن تمام متن‌های کهن زبان‌های هندی‌اروپایی مانند هندی باستان، ایرانی باستان، یونانی باستان، لاتین، زبان‌های آناتولی و جز آن بود. البته ایده ایجاد پایگاه داده برای متون زبانی به سال ۱۹۸۷ محدود نمی‌شود و به‌ویژه در ایالت متحده آمریکا از دهه ۱۹۶۰ میلادی این پیشنهاد و فکر مطرح بوده‌است. درباره زبان‌های هندی‌اروپایی، قبل از این، سایت «گنجینه زبان یونانی» Thesaurus Linguae Graecae (TLG) در سال ۱۹۷۱ توسط ماریان مکدونالد Marianne McDonald راه‌اندازی شد^۶ و آن را باید نخستین پایگاه متون زبان‌های هندی‌ایرانی به‌شمار آورد که پس از تلاش‌های نافرجام قبلی مانند تلاش برونو اسنل Bruno Snell در سال ۱۹۵۰ به نتیجه رسید و به بار نشست. در بخش زبان‌های ایرانی این پایگاه، متونی از زبان‌های اوستایی، فارسی باستان، ختنی، بلخی، سغدی، پهلوی اشکانی، فارسی میانه (پهلوی)، فارسی نو و آسی در دسترس قرار گرفته‌است. در بخش پهلوی، بجز متن‌های

فارسی میانه مانوی، متون «ارداویراف‌نامه»، «اندرز اوشنر دانا»، «بندش»، «دینکرد هفتم»، «کارنامه اردشیر بابکان»، «مینوی خرد»، «مادیان هزار دادستان»، «شایست نشایست»، «گزیده‌های زادسپرم»، «زند بهمن یسن» و «متون پهلوی» در دسترس قرار دارد. این متون تنها شامل آوانویسی متن (بدون ترجمه و گاهی همراه با حرف‌نویسی) است و برای واژه‌های متن نیز آوانویسی و معنای واژه بر اساس «فرهنگ کوچک زبان پهلوی» (تألیف دیوید نیل مکنزی، ۱۹۷۱) ثبت شده‌است. به‌جز این موارد، در این دادگان تحلیل زبان‌شناختی دیگری از واژه‌ها انجام نشده‌است.

از کارهای دیگری که بخشی از داده‌های زبانی و متون ایرانی باستان را دیجیتال و رقمی کرده‌است، می‌توان به وبگاه دانشگاه تگزاس و وبگاه Wikisource اشاره کرد. در سایت دانشگاه تگزاس در بخش آموزش آنلاین زبان‌های هندی‌اروپایی، درس‌نامه‌هایی شامل گرامر و نمونه متن و واژه‌نامه^۶ تحلیلی برای آموزش زبان‌های مختلف هندواروپایی تدوین شده‌است. از زبان‌های ایرانی، درس‌نامه‌هایی برای زبان فارسی باستان و اوستا تهیه شده و چند متن فارسی باستان (شامل بخش‌هایی از کتیبه داریوش در بیستون و نقش رستم و کتیبه خشایارشا در تخت‌جمشید) و چند متن اوستایی (شامل بخش‌هایی از هوم‌یشت، مهریشت و گاهان) همراه با واژه‌نامه و تحلیل زبان‌شناختی کلمات این چند متن بر روی سایت بارگذاری شده‌است.^۷ در این وبگاه، تنها متن آوانویسی شده تمام اوستا بجز خرده‌اوستا (شامل نندیداد، یشت‌ها، یسنا و ویسپرد) و ستون اول کتیبه بیستون به خط میخی فارسی باستان بارگذاری شده‌است.^۸ هیچ تحلیل زبان‌شناختی از واژه‌های این متون در این سایت ارائه نشده‌است.

«پهلویکا»^۹ جدیدترین فرهنگ برخطی است که برای زبان‌های ایرانی راه‌اندازی شده‌است. این وبگاه که در اواخر سال ۲۰۱۹ در دسترس پژوهشگران قرار گرفت، شامل واژه‌های پهلوی با حرف‌نویسی، آوانویسی، معنای انگلیسی و فارسی و نیز ریشه‌شناسی etymology واژه‌ها است. همچنین نمونه‌های کاربرد واژه‌ها در متون به‌صورت مثال ذیل هر مدخل در حال بارگذاری و تکمیل است. در این فرهنگ برخط نیز هیچگونه تحلیل زبان‌شناختی از واژه‌های پهلوی ارائه نشده‌است.

علی‌رغم اینکه ایده ایجاد پایگاه داده برای زبان‌های ایرانی از نیمه قرن بیستم آغاز شده‌است، غیر از پایگاه تیتوس، پروژه دیگری موفق به ایجاد پایگاه داده برای زبان‌های ایرانی نشده‌است؛ و چنانکه گفته شد، پایگاه تیتوس نیز تحلیل زبان‌شناختی واژه‌ها را در

برنامه خود قرار نداده است. مهم‌ترین دلیل این امر، گستردگی این متون بوده است که نیازمند طرحی طولانی‌مدت، با همکاری گروهی از پژوهشگران توانمند و آشنا به زبان پهلوی بوده است. در این مقاله تلاش می‌شود ضمن معرفی طرحی پژوهشی برای ایجاد و گسترش پیکره دادگان متون فارسی میانه کتابی و تحلیل اطلاعات زبان‌شناختی آنها، چگونگی تحلیل واژه‌ها در پایگاه و نیز ذکر نمونه‌هایی از تحلیل اطلاعات زبان‌شناختی و هزوارش‌ها به همراه تحلیل آماری هشت متن از متون این پایگاه و همچنین اهمیت، ویژگی و امکانات این پایگاه داده توضیح داده شود.

۳. پایگاه داده پارسیگ

هدف اصلی در این پژوهش، گردآوری داده‌های زبانی از متون فارسی میانه به صورت یک پایگاه داده، تحلیل زبان‌شناختی واژه‌های متون نوشته‌شده به زبان پهلوی و خط فارسی میانه کتابی و همچنین به دست آوردن دانش جدید با استفاده از امکانات فناوری اطلاعات در این پایگاه داده است. در انجام این کار، رایانه به عنوان یک ابزار مفید مورد استفاده پژوهشگر قرار می‌گیرد تا دانش جدیدی که پیش از این دست‌نیافتنی می‌نمود به دست آورد. حوزه وسیعی از سازمان‌ها و افراد حقیقی جزء کاربران این پایگاه قرار می‌گیرد. این پیکره می‌تواند در تهیه فرهنگ‌های تاریخی و نیز فرهنگ‌های گوناگون زبان فارسی و فارسی میانه و دیگر زبان‌های ایرانی به کار آید، برای واژه‌گزینی علمی مورد استفاده قرار گیرد. همچنین این پایگاه داده می‌تواند در بررسی‌های معنی‌شناسی، واج‌شناسی، تحلیل نحوی و بررسی سیر تحول تاریخی واژگان فارسی به کار برده شود.

چارچوب این پایگاه داده بر دو اصل پایه‌ریزی شده است. یکی از این اصول مربوط به ساختارمندسازی داده‌های متنی فارسی میانه در پایگاه داده است و دیگری تحلیل‌هایی است که بر روی متون این پایگاه انجام شده است.

۱.۳ ساختارمندسازی داده

یکی از ویژگی‌های کاربرد فناوری اطلاعات در علوم انسانی این است که داده و اطلاعات پراکنده و غیرمنسجم باید به گونه‌ای که برای رایانه قابل فهم باشد منظم گردیده و سازمان یابد (قیومی، ۱۳۹۷). طی این فرایند، داده بدون ساختار به داده ساختارمند تبدیل می‌شود.

وجود ساختار در داده سبب می شود بتوان با کاربرد الگوریتم های رایانشی، از این داده برای کاربردهای متنوع استفاده نمود. یکی از مهم ترین کاربردهای شیوه های ساختارمندسازی داده، پایگاه داده است. پس از ساماندهی داده در پایگاه داده می توان کار جستجو در پایگاه را انجام داد و با توجه به سؤالات مورد نظر کاربر، خروجی های متنوعی را از پایگاه استخراج کرد و از این خروجی برای پاسخ به فرضیات پژوهشگر استفاده نمود.

در این پژوهش برای هر واژه فارسی میانه، اطلاعاتی نظیر آوانویسی، حرف نویسی، ترجمه فارسی، مقوله دستوری واژه، بن واژه و تعیین هزوارش بودن واژه را فراهم می آوریم. برای رسیدن به این اهداف، در مرحله اول، تمام اطلاعات زبان شناختی مورد نیاز را در یک فایل اکسل وارد کرده و اطلاعات را ساماندهی می کنیم؛ پس از تحلیل زبان شناختی واژه، این داده را به صورت ساختارمند به ساختار زبان نشاننداری XML eXtensible Markup Language تبدیل کرده و به پایگاه داده منتقل می نماییم.

ویژگی این روش ساختارمندسازی داده این است که می توان تمام یک متن را به ساختار یک درخت تبدیل کرد که بندها و جملات شاخه های آن درخت بوده و واژه ها نیز برگ های هر ساقه تلقی گردد. هر برگ این درخت می تواند در لایه های مختلف اطلاعات جزئی تری را در بر گیرد. این اطلاعات در قالب خصیصه attribute و ارزش value برای هر واژه که در یک برگ است تعریف می گردد.

۲.۳ تحلیل داده های پایگاه داده پارسیگ

پایگاه داده پهلوی که در چارچوب پژوهش حاضر تهیه شده است چندین ویژگی مورد توجه قرار گرفته است که در ادامه توضیح داده می شود.

۱.۲.۳ اندازه و حجم

اندازه و حجم هر پیکره، با هدف تشکیل پیکره و برخی ملاحظات کاربردی ارتباط دارد. با توجه به هدف تشکیل پایگاه داده زبان پهلوی، حجم پیکره این پایگاه به اندازه واژه های تمام متون پهلوی است. با توجه به امکانات رایانه ای، محدودیتی برای رسیدن به هدف وجود ندارد و می توان اندازه پیکره ها را بدون محدودیت افزایش داد. موانعی که باعث می شود تهیه این پیکره به سختی و به کندی پیشرفت رود، این است که پژوهشگر در خواندن متن پهلوی و قرائت های مختلف کلمات، حرف نویسی و آوانویسی متن باید بدون

واسطه رایانه، مداخله مستقیم داشته باشد و این امر سرعت ورود اطلاعات را کند می‌نماید. همچنین به دلیل پیچیدگی‌های ساخت واژه در عبارات پهلوی نسبت به جملات فارسی و نبود داده اولیه برای طراحی یک مدل پردازش خودکار برای تحلیل متون پهلوی نیاز است تمامی فرایندهای جمع‌آوری و تحلیل داده به صورت دستی و با نظارت نیروی انسانی خبره انجام پذیرد.

۲.۲.۳ اولویت‌دهی محتوایی به متن‌ها

یکی از ویژگی‌هایی که در ایجاد پیکره‌های گوناگون مورد توجه قرار می‌گیرد، دسته‌بندی متن‌ها براساس گونه و سبک ادبی آنها و انتخاب این متن‌ها بر اساس زمان تولید، نوع تولید (کتاب، نشریه و جزآن) و اعتبار مؤلف است. با توجه به اینکه قرار است در پایگاه داده زبان پهلوی پیکره‌ای برای تحقیقات زبان‌شناختی زبان‌های باستانی ایران فراهم شود، نیاز است با یک اولویت‌بندی محتوایی، تمامی متون پهلوی در این پایگاه قرار گیرد؛ به این صورت که نخست متن‌های ساده مانند اندرنامه‌های پهلوی و متون تعلیمی و تاریخی و سپس متن‌های دشوار مانند متون فقهی و سپس حقوقی و فلسفی در پایگاه بارگذاری خواهد شد. از این رو، این پایگاه داده در طول زمان با اضافه شدن متون جدید براساس معیار مطرح شده توسعه و گسترش می‌گردد.

۳.۲.۳ تعریف اطلاعات برای هر واژه

چنان که پیشتر گفته شد، اطلاعاتی که برای هر واژه در پایگاه تعریف می‌شود، شامل حرف‌نویسی، آوانویسی و ترجمه فارسی هر واژه، تعیین هزوارش بودن یا نبودن واژه، برچسب مقوله دستوری واژه بر اساس جدولی که برای این کار تهیه شده است، تعیین بن‌واژه، ارجاع به صفحه و سطر دستنویس پهلوی و ارجاع به شماره بند هر متن است. به جز این اطلاعات که برای هر واژه ثبت می‌شود، آوانویسی کل متن و ترجمه آن جداگانه بارگذاری می‌شود. این اطلاعات که به صورت فراداده به داده اولیه اضافه می‌شود می‌تواند در جستجو کاربرد داشته باشد. بر این اساس، وجود امکانات جستجوی متنوع بر کارایی این پایگاه داده مؤثر بوده و می‌تواند در تعمیق دانش به دست آمده از این پایگاه کمک شایانی نماید. شایان ذکر است در برای تعیین بودن یا نبودن واژه به عنوان هزوارش، از دو عدد

صفر و یک استفاده می‌شود که عدد یک به مفهوم هزوارش بودن و صفر به مفهوم عدم هزوارش بودن آن واژه است.

۳.۳ امکانات پایگاه برای جستجوی متن و واژه

در حال حاضر اطلاعاتی که از متون و واژه‌های فارسی میانه در پایگاه داده پارسیگ فراهم آمده است از طریق وب در دسترس است.^۱ جستجوی در متن، جستجوی واژه، جستجوی برچسب مقوله دستوری و جستجوی باهم‌آیی واژه‌ها در متون از جمله امکانات جستجویی است که امکان استخراج اطلاعات از پایگاه داده فارسی میانه را میسر می‌سازد. در بخش جستجوی واژه، می‌توان آوانویسی، حرف‌نویسی، معنا و بن هر واژه را جستجو کرد. نمایش نتیجه جستجو، فهرستی از تمام متونی است که واژه هدف در آنها به کار رفته است را با تمام تحلیل‌هایی که برای آن انجام شده است، به همراه جمله شاهدهی که واژه هدف در جمله به کار رفته است، مشاهده نمود.

۴. درونداد و تحلیل داده

۱.۴ درونداد و ساماندهی اطلاعات

قبل از انجام تحلیل، نیاز است این متن‌ها در قالب یک پیکره زبانی از شکل کاغذی به شکل الکترونیکی تبدیل گردد تا بتوان متن‌ها را در پایگاه داده ساماندهی و تحلیل نمود. در این ساماندهی، هر واژه در یک ردیف قرار دارد و در ستون‌های مقابل هر واژه، پنج لایه اطلاعات تحلیلی وارد می‌شود. ترتیب اطلاعات در پایگاه داده به این صورت است که نخست آوانویسی واژه و سپس حرف‌نویسی، اطلاعات مربوط به هزوارش بودن واژه، معنی فارسی واژه، مقوله دستوری واژه، بن واژه، ارجاع به دستنویس و ارجاع متنی مشخص می‌شود. در جدول (۱) نحوه ساماندهی اطلاعات و تحلیل عبارت «به‌نام یزدان» از متن «اندرز پیشینیان» نشان داده شده است.

جدول ۱: نحوه تحلیل عبارت «به نام یزدان»

ارجاع متنی	ارجاع به دستنویس	بن واژه	مقوله دستوری	معنی فارسی	هزوارش	حرف نویسی	آوانویسی
HP1: 0	JA: 39/1	pad	E-----	به	l	PWN	pad
HP1: 0	JA: 39/1	nām	Ncsp-----	نام	l	šM	nām
HP1: 0	JA: 39/1	ī	K	—	l	Y	ī
HP1: 0	JA: 39/1	yazad	Nasp-----	یزدان	0	yzd'n	yazdān

۲.۴ تحلیل نحوی داده‌های ورودی

برای تحلیل نحوی در سطح واژه‌ها باید مقوله دستوری واژه‌ها تعیین گردد. نبود داده برچسب‌خورده اولیه برای ساخت یک مدل برچسب‌زنی خودکار سبب شد این کار توسط دو فرد خبره به ساخت زبان فارسی میانه به کار برچسب‌زنی دستی آن پردازند. برای رسیدن به این هدف به این صورت عمل شد که ابتدا جدول استاندارد شده قیومی (۲۰۱۴) که براساس جدول مقولات دستوری بی‌جن‌خان و همکاران (۲۰۱۱) برای فارسی معاصر تهیه شده است مورد بررسی قرار گرفت و سپس باتوجه به ویژگی‌ها و نیازهای زبان پهلوی تغییر و توسعه یافت. بر این اساس، ۱۲ مقوله دستوری اصلی برای واژه‌های متون پهلوی تعریف شد.

ساختار جدول‌های برچسب‌ها به این صورت است که ابتدا مقوله دستوری واژه مشخص می‌شود و براساس این مقوله، اطلاعات صرفی و نحوی و حتی معنایی واژه در یک ساختار لایه‌ای مشخص می‌گردد و در هر لایه آن مقوله، اطلاعات مشخص قرار می‌گیرد. با توجه به استفاده از استاندارد Multi-East Text توسط قیومی (۲۰۱۴)، در این پژوهش نیز از این استاندارد استفاده شده است؛ بنابراین هریک از ویژگی‌های مقولات دستوری با یک حرف نشان داده می‌شود. این نوع نمایش دو ویژگی دارد: اولاً طول برچسب‌های مربوط به یک مقوله یکسان است؛ و ثانیاً در هر جایگاه از برچسب، یک اطلاعات مشخص درج می‌شود که به سهولت در جستجو و کاربردهای رایانشی می‌انجامد. به دلیل کمبود فضا برای شرح تمامی اطلاعات صرفی-نحوی-معنایی که در برچسب‌های مقوله‌های مختلف تعریف شده است، در جدول (۲) شرح مختصری از تعریف مقوله دستوری «صفت» ارائه می‌شود.

جدول (۲): جدول مقوله دستوری صفت

1 st	2 nd	3 rd	4 th	5 th
CATEGORY (مقوله اصلی)	TYPE (نوع)	CLITIC (واژه‌بست)	POLARITY (قطبیت)	NUMBER (شمار)
Adjective (A) (صفت)	Simple (s) (ساده) Comparative (c) (مقایسه‌ای) Superlative (u) (تفضیلی)	Ezafe (z) (اضافه)	Positive (p) (مثبت) Negative (n) (منفی)	Singular (s) (مفرد) Plural (l) (جمع)
		Ya (y) (نشانه نکره ی)		
		Adverb (d) (قید)		
		Pronominal 1ST,SG (1) (ضمیری اول شخص مفرد)		
		Pronominal 2ND,SG (2) (ضمیری دوم شخص مفرد)		
		Pronominal 3rd,SG (3) (ضمیری سوم شخص مفرد)		
		Pronominal 1ST,PL (4) (ضمیری اول شخص جمع)		
Pronominal 2ND,PL (5) (ضمیری دوم شخص جمع)				
Pronominal 3RD,PL (6) (ضمیری سوم شخص جمع)				

براساس اطلاعات موجود در جدول (۲)، مقوله «صفت» حاوی ۵ ستون اطلاعات است. در ستون اول، مقوله دستوری اصلی که همان صفت است مشخص می‌شود. در ستون دوم، نوع صفت اعم از ساده، مقایسه‌ای و تفضیلی، تعیین می‌گردد. ممکن است واژه یک واژه‌بست شامل کسره اضافه، یای نکره یا یکی از صورت‌های ضمیر متصل ملکی داشته باشد که این اطلاعات در ستون سوم مشخص می‌شود. در ستون چهارم قطبیت واژه اعم از مثبت، منفی یا خنثی بودن واژه تعیین می‌گردد. در ستون پنجم شمار صفت از نظر مفرد یا جمع بودن تعیین می‌شود. برای مثال، در جدول (۳) اطلاعات مربوط به واژه‌های صفت در مثال (۱) نمایش داده شده‌است:

(1) cē afsōsgar mard afsōs-bar zad-xwarrah nifřīdag bawēnd.

(۱) ... زیرا مرد افسوس‌گر (= ریشخندکننده)، افسوس‌بر (= موجب ریشخند خود)، زد- فره (= ضربه‌زننده به فره خود)، نفرین‌شده باشند (چیده اندرز پوریوتکیشان، بند ۴۳).

جدول (۳): نحوه تعیین اطلاعات برای هر واژه صفت

مقوله دستوری	معنی فارسی	حرف نویسی	آوانویسی
Ac-ns	افسوس‌گر، مسخره‌کننده	ʔpswskl	afsōsgar
Ac-ns	افسوس‌بر، موجب ریشخند خود	ʔpsws-bl	afsōs-bar
Ac-ns	زدفَره، دارای فره نابودشده	zt-GDE	zad-xwarrah
Ac-ns	نفرین‌شده	nplytk	nifřīdag

نکاتی که در برجسب‌گذاری صفات پهلوی می‌توان متذکر شد، به شرح زیر است:

الف) ستونی که مشخص‌کننده شمار صفت است، به جدول برچسب‌گذاری قیومی (۲۰۱۴) اضافه شد، زیرا در زبان پهلوی گاهی صفت با موصوف از نظر شمار مطابقت دارد (ابوالقاسمی، ۱۳۷۵، ص ۵۸)، مانند مثال (۲):

(2) pad ušahin gāh abar ēstēd ošōmandān mardōmān

(۲) در گاه اُشهین برخیزید ای مردمان فانی (اندرز دانایان به مزدیسنان، بند ۲)
ب) قطبیت صفت بر اساس مفهوم محتوایی آن تعیین شده‌است نه ساخت واژه. برای نمونه در مثال‌های زیر که دارای پیشوند abē- «بی-، بدون» است، واژه‌های abēwināh و abēbīm با توجه به بافت و معنای جمله، مانند مثال‌های (۳) و (۴)، صفت مثبت و abēšnōhr صفت منفی در نظر گرفته شده‌است.

(3) abēwināh bāš kū abēbīm bawē.

(۳) بی‌گناه باش تا بی‌بیم باشی (اندرز آذرباد ماراسپندان، بند ۷۲).

(4) abēšnōhr mardōm ma dār čē-t spās nē dārēd.

(۴) مردم ناسپاس را نپرور، زیرا تو را سپاس ندارد (اندرز آذرباد ماراسپندان، بند ۸۳).
پ) مسلماً مقوله واژه‌ها بر اساس نقش آنها در جمله تعیین شده‌است. در زبان پهلوی نیز مانند فارسی صفت در جمله ممکن است در مقوله اسم یا قید ظاهر شود.
در آوانویسی اسم‌ها و صفت‌های مرکب از خط تیره (مانند afsōs-bar و zad-xwarrah در مثال (۱)) استفاده شد و پیشوندها و پسوندها بدون فاصله با بن واژه نوشته شده‌است (مانند afsōsgar و nifridag در مثال (۱)).

۵. تحلیل آماری از داده‌های پایگاه

پیشتر در بخش ۳-۳ امکانات پایگاه برای انجام انواع جستجوهای تعریف‌شده توضیح داده شد. از دیگر امکانات این پایگاه، استخراج اطلاعات آماری از پیکره است. این اطلاعات آماری می‌تواند از یک متن مشخص یا مجموعه‌ای از متون مورد نیاز استخراج گردد و دید کلی‌تری از مجموعه متون یا دید عمیق‌تری نسبت به هر متن ایجاد نماید. در این پژوهش، با رویکرد آماری به دو نوع تحلیل می‌پردازیم. یکی از تحلیل‌ها استخراج اطلاعات آماری توصیفی از واژه‌های پهلوی ۸ متن موجود در این پایگاه است که شامل متون (۱) اندرز آذرباد مهرسپندان؛ (۲) اندرز پیشینیان؛ (۳) اندرز خسرو قبادان؛ (۴) اندرز دانایان به مزدیسنان؛

(۵) بهرام ورجاوند؛ (۶) گزارش شطرنج و وضع نرد؛ (۷) سور سخن؛ و (۸) یادگار بزرگمهر است و تحلیل آماری دیگر مربوط به هزوارش‌های این مجموعه داده است.

۱.۵ تحلیل آماری متون پهلوی

هشت متن پهلوی که در بخش اول گزارشی از آنها ارائه شده است مشتمل بر ۵ اندرزنامه، یک رساله کوچک تعلیمی، یک داستان و یک متن پیشگویی است. اطلاعات آماری از هر متن استخراج شد که در جدول (۴) گزارش شده است. اطلاعات آماری مورد نظری که استخراج شده است عبارت است از تعداد بندهای متن، تعداد واژه‌ها با تکرار متن تعداد واژه‌ها بدون تکرار متن، تعداد بن واژه در هر متن، طول متوسط هر بند، تعداد واژه‌های با بسامد ۱ در هر متن، تعداد واژه‌های محتوایی در هر متن، تعداد واژه‌های دستوری در هر متن، تعداد هزوارش‌ها در هر متن و تعداد هزوارش‌هایی که جزء واژه‌های دستوری نبوده و واژه محتوایی محسوب می‌شود.

همان‌گونه که در جدول (۴) قابل مشاهده است «بهرام ورجاوند» کوتاه‌ترین متن و «یادگار بزرگمهر» بلندترین متن از نظر تعداد بند و واژه است که در حال حاضر در پایگاه موجود است. اگرچه متن «بهرام ورجاوند» بسیار کوتاه است ولی از نظر طول متوسط بندها و تعداد واژه، حاوی طولانی‌ترین بندها با متوسط حدود ۴۱ واژه در هر بند است. در حالی که متن‌های «یادگار بزرگمهر» و «اندرز خسرو قبادان» حاوی کوتاه‌ترین بندها با متوسط حدود ۷ واژه است. تنوع واژگانی در متن «بهرام ورجاوند» از تمام متن‌های موجود بیشتر است؛ در حالی که متن‌های «یادگار بزرگمهر» و «اندرز آذرباد» علیرغم داشتن متن‌های طولانی تقریباً به یک میزان از کمترین تنوع واژگانی برخوردار است. در همین دو متن، نسبت بُن واژه‌ها به تمام واژه‌های به‌کاررفته در متن از کمترین نسبت برخوردار است. این امتیاز بیانگر پایین بودن تنوع واژه‌های به‌کاررفته در متن‌ها است که از جمله مهم‌ترین دلایل، استفاده از واژه‌های تکراری و کم بودن سهم کاربرد واژه‌های با بسامد ۱ نسبت به سایر واژه‌ها است.

از نظر نحوی، متن‌های «اندرز دانایان به مزدیسنان» و «سور سخن» حاوی بیشترین واژه‌های نقشی نسبت به واژه‌های محتوایی است. این ویژگی به‌طور تلویحی بیانگر وجود ساخت‌های نحوی پیچیده‌تر در مقایسه با سایر متون موجود در پایگاه است که برای تأیید این نظر به بررسی‌های بیشتر نیاز است.

از نظر کاربرد هزوارش، متن «اندرز دانایان به مزدیستان» حاوی کمترین تعداد هزوارش است؛ درحالی‌که بیش از ۸۰٪ از واژه‌های متون «اندرز خسرو قبادان» و «گزارش شطرنج و وضع نرد» متشکل از این نوع واژه‌ها است.

جدول (۴): استخراج اطلاعات آماری از متن‌های هدف

نام متن	تعداد بند	تعداد واژه تکراری	تعداد واژه بی‌تکرار	تعداد بن‌واژه	طول متوسط بند	تعداد واژه با بسامد ۱	تعداد واژه محتوایی	تعداد واژه دستوری	تعداد هزوارش	تعداد هزوارش محتوایی
اندرز آذرباد	۱۵۸	۱۹۵۹	۶۳۰	۵۱۲	۱۲/۴۰	۴۶۵	۱۳۶۷	۵۹۲	۱۲۰۱	۱۳۶۶
اندرز پیشینان	۳۱	۴۴۳	۲۰۹	۱۷۸	۱۴/۲۹	۱۶۶	۲۷۸	۱۶۵	۱۹۲	۲۷۸
اندرز خسرو قبادان	۴۳	۳۳۵	۱۴۳	۱۳۳	۷/۵۶	۱۰۰	۲۲۰	۱۰۵	۲۵۷	۲۲۰
اندرز دانایان به مزدیستان	۶۰	۸۸۹	۳۶۷	۲۸۷	۱۴/۸۲	۲۹۱	۵۴۵	۳۴۴	۳۱۵	۵۴۵
بهرام ورجاورد	۵	۲۰۸	۱۳۲	۱۱۱	۴۱/۶۰	۱۱۴	۱۳۱	۷۷	۱۲۴	۱۳۱
گزارش شطرنج و وضع نرد	۴۱	۶۰۲	۲۷۷	۲۳۶	۱۴/۶۸	۲۰۸	۳۸۸	۲۱۴	۴۸۵	۳۸۸
سور سخن	۲۳	۶۳۴	۳۰۹	۲۴۷	۲۷/۵۷	۲۶۹	۳۸۹	۲۴۵	۳۳۴	۳۸۹
یادگار بزرگمهر	۲۷۴	۲۱۲۷	۶۳۹	۴۹۵	۷/۷۶	۵۶۰	۱۳۹۹	۷۲۸	۱۰۴۶	۱۳۹۹

هشت متنی که در این بخش بررسی شده‌اند، در مجموع شامل ۷۱۸۷ واژه با تکرار، ۲۷۶۰ واژه بدون تکرار و ۶۳۵ بند است. فهرست واژه‌های بدون تکرار براساس صورت واژه‌ها تهیه شده است که بن واژه‌های این تعداد واژه ۲۱۹۹ بن واژه است. طول متوسط بندها ۱۱/۳۲ واژه است که بیانگر بندهای نسبتاً طولانی است. از تعداد واژه‌های موجود در این متون، ۲۱۷۳ واژه با بسامد ۱ فقط یکبار در متون دیده شده است که حدود ۳۰ درصد از کل واژه‌های متون را شامل می‌شود. اغلب این واژه‌ها اسامی خاص است. به‌طور کلی، نسبت واژه‌های بدون تکرار به واژه‌های با تکرار که تنوع واژگانی (غنای واژگانی) نامیده می‌شود ۳۸/۴۰ درصد است. این عدد به این مفهوم است در مجموع متن‌ها از تنوع واژگانی نه‌چندان زیادی استفاده شده است. براساس این شاخص، چنانچه متون جدول (۴) را با یکدیگر مقایسه کنیم در بعضی از متن‌ها مانند «اندرز آذرباد» و «یادگار بزرگمهر» تنوع واژگانی از متوسط اعلام شده پایین‌تر است ولی در بعضی از متون مانند «اندرز خسرو قبادان»، «اندرز پیشینیان»، «اندرز دانایان به مزدیستان»، «بهرام ورجاوند»، «گزارش شطرنج و وضع نرد» و «سور سخن» تنوع واژگانی زیاد است. همان‌گونه که پیشتر مطرح شده بود از میان متن‌های بررسی شده، «بهرام ورجاوند» غنای واژگانی نسبتاً بالایی با ۶۳/۴۶ درصد را دارد. این تنوع واژگانی موجب دشواری درک مطلب می‌گردد.

با برچسب‌گذاری مقوله دستوری واژه‌ها می‌توان کار دسته‌بندی واژه‌ها به واژه‌های محتوایی و دستوری را به آسانی انجام داد. واژه‌های محتوایی عبارت است از فعل، اسم، صفت، قید، ضمیر و شبه جمله؛ و واژه‌های نقشی عبارت است از حرف ربط، حرف اضافه و حرف تعریف. از میان واژه‌های متون، ۴۷۱۷ واژه محتوایی است که ۶۵/۶۳ درصد کل واژه را شامل می‌شود. همچنین، این متون ۲۴۷۰ واژه نقشی داشته است که ۳۴/۳۷ درصد از کل واژه‌های متون را شامل می‌شود. به نظر می‌رسد یکی از دلایل طولانی بودن متوسط بندها، کاربرد همین واژه‌های دستوری است که به افزایش طول بند منجر می‌شود.

تعداد کل هزوارش‌ها در داده‌های تحلیل شده ۳۹۴۴ واژه است که این تعداد بیانگر این است که ۵۴/۸۸ درصد از کل واژه‌ها هزوارش است. شایان ذکر است که در متن «گزارش شطرنج و وضع نرد» ۸۰/۵۶ درصد از واژه‌های متن هزوارش است. این نکته بیانگر اهمیت موضوع هزوارش در متون پهلوی است. هزوارش‌ها شامل واژه‌های محتوایی و نقشی است به این صورت که از میان این حجم از هزوارش‌ها (با احتساب تکرار)، ۶۴/۴۵ درصدشان واژه‌های محتوایی و مابقی واژه‌های نقشی است.

۲.۵ تحلیل آماری هزوارش‌های متون پهلوی

در بخش پیشین به این نتیجه رسیدیم که بیش از نیمی از واژه‌های متون پهلوی که در پایگاه داده پارسیک موجود است هزوارش است. بنا بر اهمیت این دسته از واژه‌ها، در این بخش به تحلیل بیشتر هزوارش‌ها می‌پردازیم. در جدول شماره (۵) تعداد هزوارش‌های این متون برای تمامی مقولات دستوری شامل حرف ربط، اضافه، حرف اضافه پیشین و پسین، قید، اسم، ضمیر، فعل، حرف تعریف، صفت و عدد ارائه شده‌است. در این جدول علاوه بر شمارش واژه‌های ساده به‌عنوان هزوارش، واژه‌هایی که از ترکیب هزوارش و پهلوی نوشته شده‌است نیز در نظر گرفته شده‌است. همچنین بسامد نسبی هزوارش‌ها نسبت به تمام واژه‌های این پایگاه داده و همچنین نسبت به واژه‌های نماینده محاسبه شده‌است.

جدول ۵: اطلاعات آماری استخراج شده مربوط هزوارش‌ها

هزوارش و پهلوی (بی تکرار)	هزوارش و پهلوی (با تکرار)	هزوارش ساده (بی تکرار)	هزوارش ساده (با تکرار)	بسامد نسبی کلی بی تکرار (درصد)	بسامد نسبی کلی با تکرار (درصد)	
۲۲	۷۸	۱۷	۹۵۰	۱/۴۱	۱۴/۳۰	حرف ربط (ساده)
۱۸	۱۱	۲۲	۶۳۲	۱/۴۵	۸/۹۵	اضافه
۲	۲۴	۵	۴۳۷	۰/۲۵	۶/۴۱	حرف اضافه پیشین (ساده)
۲۱	۲۶	۲۳	۳۰۳	۱/۵۹	۴/۵۸	قید
۵۰	۱۰۲	۵۴	۲۷۱	۳/۷۷	۵/۱۹	اسم (ساده)
۲۴	۶۹	۱۸	۱۸۱	۱/۵۲	۳/۴۸	ضمیر
۱۶۳	۵۰۷	۳۲	۱۴۱	۷/۰۷	۹/۰۲	فعل
۲	۴	۷	۸۳	۰/۳۳	۱/۲۱	حرف تعریف
۳۱	۴۳	۱۰	۶۸	۱/۴۹	۱/۵۴	صفت
۰	۰	۴	۶۸	۰/۱۵	۰/۹۵	عدد
۲	۵	۲	۶۵	۰/۱۵	۰/۹۷	حرف اضافه پسین

در بخش پیشین گفته شد که ۵۴/۸۸ درصد از کل واژه‌ها هزوارش است که اطلاعات جزئی مربوط به هر مقوله دستوری در جدول ۵ قابل ملاحظه است. این مقدار دربرگیرنده هزوارش ساده و ترکیب هزوارش و واژه پهلوی است. چنانچه از اطلاعات این جدول مشاهده می‌شود، دو دسته واژه‌های محتوایی فعل و اسم و همچنین سه دسته واژه‌های نقشی حرف ربط، اضافه و حرف اضافه پیشین، با احتساب مرز پنج درصد نسبت به کل واژه‌های این پایگاه داده، پرکاربردترین هزوارش‌ها در این داده‌ها هستند.

از میان واژه‌های محتوایی، مقوله دستوری فعل ۹/۰۲ درصد از کل واژه‌های این پیکره را شکل داده‌است.^{۱۱} بدون در نظر گرفتن تکرار، ۶۴۸ فعل در این پیکره به کار رفته‌است که از این تعداد ۱/۶۹ درصد افعال همیشه هزوارش، ۷/۰۵٪ هم به صورت هزوارش و هم املائی پهلوی و حدود ۹۱ درصد نیز همیشه به صورت غیرهزوارش نوشته شده‌است. شایان ذکر است فعل ۷/۰۷ درصد از واژه‌های نماینده که واژگان این مجموعه داده را شکل می‌دهد را در بر گرفته‌است.

اسم مقوله دستوری دوم پربسامد در میان هزوارش‌هایی است که جزء واژه‌های محتوایی تلقی می‌شود. این مقوله دستوری ۵/۱۹ درصد از کل واژه‌ها و ۳/۷۷ درصد از واژگان را شکل داده‌است. اگر تعداد اسم‌های با تکرار را در نظر بگیریم، ۳/۷۷ درصد از اسم‌ها همیشه به صورت هزوارش، ۱/۴۹ درصد هم هزوارش و هم پهلوی و حدود ۹۵ درصد از اسم‌ها همیشه پهلوی نوشته شده‌است. ده اسم پربسامد که همواره به صورت هزوارش آمده‌است، عبارت‌است از: *dēw, šāh, dast, sāl, may, gāw, dar, pus, xwēš, asp*.

پربسامدترین مقوله دستوری در این پایگاه داده، حرف ربط است. این مقوله دستوری ۱۴/۳۰ درصد از واژه‌های پیکره را شکل می‌دهد ولی این گروه از واژه‌ها بسیار کم تنوع بوده و ۱/۴۱ درصد از واژگان را به خود اختصاص داده‌است. تعداد حروف ربط ساده، بدون در نظر گرفتن تکرار، ۱۴ مورد است، که از میان آنها دو حرف ربط *ayāb* و *ciyōn* همیشه با املائی پهلوی و بقیه حروف همواره به صورت هزوارش به کار رفته‌اند.

اضافه نیز که ۸/۹۵ درصد از واژه‌های پیکره و ۱/۴۵ درصد از واژگان را در بر گرفته است همانند حرف ربط از تنوع واژگانی بسیار کمی برخوردار است.

حروف اضافه پیشین سومین واژه نقشی پربسامد در این پیکره است. بدون در نظر گرفتن تکرار، تنها ۱۵ مورد پیش‌اضافه در متون به کار رفته‌است که ۰/۲۵ واژگان را تشکیل می‌دهد. از ۷ پیش‌اضافه پربسامد پهلوی (که بیش از پنجاه بار در پیکره به کار رفته‌است)، سه حرف اضافه *pad, az, ō* هم به صورت هزوارش و هم به صورت غیرهزوارش و حرف‌اضافه‌های *be, andar, abāg, abar* همیشه به صورت هزوارش نوشته شده‌است.

در بررسی این داده‌ها می‌بینیم که نقش هزوارش در اسم و فعل به عنوان واژه‌های محتوایی بسیار کمتر از واژه‌های نقشی مانند حرف ربط و پیش‌اضافه است. با در نظر گرفتن کل آمار هزوارش‌ها در تمام مقوله‌های دستوری^{۱۲} معلوم می‌شود که بیش از ۳۳ درصد واژه‌های نقشی به صورت هزوارش نوشته می‌شود و در واژه‌های محتوایی نقش

هزوارش در قید، ضمیر و صفت بسیار کمتر از فعل و اسم است. بنابراین می‌توان نتیجه گرفت با وجود آنکه حدود نیمی از هر متن پهلوی را هزوارش‌ها تشکیل می‌دهد، بیشترین تعداد آنها مربوط به واژه‌های نقشی است که بدون احتساب تکرار، تعدادشان محدود است. به عبارت دیگر، با یادگیری هزوارش تعداد محدودی واژه‌های نقشی و همچنین تعداد مشخصی واژه‌های محتوایی می‌توان نیمی از واژه‌های این پایگاه داده را خواند. شاید علی‌رغم اینکه هزوارش‌ها در ظاهر از دشواری‌های خط پهلوی به شمار می‌آید، با توجه به تعداد محدود و صورت نوشتاری خاص آنها، مشخص شدن نقش مقولات دستوری واژه‌ها در جمله، اعم از واژه‌های محتوایی و واژه‌های نقشی، می‌تواند به خوانش متن کمک به‌سزایی برساند.

۶. جمع‌بندی

در این مقاله تلاش کردیم به چگونگی ایجاد و تحلیل پیکره‌بنیان داده‌های متعلق به متون فارسی میانه کتابی پردازیم. ضمن معرفی این پایگاه و توضیح درباره چگونگی تحلیل واژه‌های فارسی میانه، به چگونگی ساماندهی اطلاعات در یک پایگاه داده و تحلیل آماری متون این پایگاه پرداختیم و امکانات این پایگاه داده را توضیح دادیم.

ویژگی مهم این پایگاه، گردآوری داده‌ها به‌صورت الکترونیکی در قالب یک پیکره زبانی و تحلیل زبان‌شناختی آنها در چندین سطح است. سطوح مختلف تحلیل‌های ارائه‌شده نیاز به سازماندهی و ساختارمندسازی دارد. از این رو، داده‌ها به‌همراه تحلیلشان به ساختار XML تبدیل شد. ویژگی الکترونیکی شدن متن‌ها این است که ضمن سادگی در کاربری مجدد آنها، به‌سادگی امکان جستجو با کمک یک زبان جست‌وجو مانند (XPath XML Path Language) برای یافتن اطلاعات مورد نظر و همچنین استخراج اطلاعات آماری را در پژوهش میسر می‌سازد.

حجم پیکره این پایگاه به اندازه واژه‌های تمام متون پهلوی است. داده‌های زبانی این متون با یک اولویت‌بندی محتوایی در این پایگاه قرار می‌گیرد؛ به این صورت که نخست متن‌های ساده مانند اندرزنامه‌های پهلوی و متون تعلیمی و تاریخی و سپس متن‌های دشوار مانند متون فقهی و سپس حقوقی و فلسفی در پایگاه بارگذاری خواهد شد. اطلاعات مربوط به هر واژه شامل حرف‌نویسی، آوانویسی و ترجمه فارسی هر واژه، تعیین هزوارش بودن یا نبودن واژه، برچسب مقوله دستوری واژه بر اساس جدولی که برای این کار تهیه

شده است، تعیین بن واژه، ارجاع به صفحه و سطر دستنویس پهلوی و ارجاع به شماره بند هر متن است. برای برجسب‌دهی دستوری واژه‌ها، از جدول استاندارد شده قیومی (۲۰۱۴) که براساس جدول مقولات دستوری بی‌جن‌خان و همکاران (۲۰۱۱) برای فارسی معاصر تهیه شده است، استفاده شد و بر اساس نیازهای زبان پهلوی تغییر و توسعه یافت.

با توجه به ساختار پایگاه داده و برجسب‌هایی که برای واژه‌های متون این پایگاه تعیین شده است می‌توان به تحلیل داده‌های موجود در این پیکره پرداخت. در این مقاله دو نوع تحلیل آماری از داده‌های این پایگاه ارائه و توضیح داده شد. در تحلیل نخست براساس ۸ متن منتخب موجود در این پایگاه داده، تعداد بندها، تعداد واژه‌ها با تکرار و بدون تکرار، تعداد بن‌واژه‌ها در هر متن، طول متوسط هر بند، تعداد واژه‌های با بسامد ۱ در هر متن، تعداد واژه‌های محتوایی و نقشی در هر متن و همچنین تعداد هزوارش‌ها (هم‌واژه‌های نقشی و هم‌محتوایی) در هر متن، شمارش و تحلیل شد. در تحلیل آماری دوم تحلیلی از هزوارش‌های این متون برای پنج مقوله دستوری پربسامد ارائه شد.

تحلیل توصیفی اطلاعات مذکور، نکات دقیق و مهمی را درباره گونه ادبی این متن‌ها، تنوع یا غنای واژگانی این متون، نسبت واژه‌های محتوایی به واژه‌های دستوری، همچنین نقش و اهمیت هزوارش‌ها در متن روشن نمود. بنابراین به دلیل این که هزوارش‌ها بیش از نیمی از واژه‌های متون پایگاه داده پارسیک را تشکیل می‌دهد اهمیت توجه به هزوارش در خوانش متن را مشخص می‌کند.

تقدیر و تشکر

این پژوهش بر اساس قرارداد شماره ۹۷۰۰۴۲۹۴ که توسط صندوق حمایت از پژوهشگران و نوآوران کشور و پژوهشگاه علوم انسانی و مطالعات فرهنگی حمایت مالی شده است، انجام گرفته است.

پی‌نوشت‌ها

۱. واژه‌نامه‌هایی که معرفی شد، مربوط به متون فارسی میانه زردشتی هستند. «واژه‌نامه فارسی میانه و پارتی مانوی» نوشته دورکین-مایسترانست (۲۰۰۴) پس از واژه‌نامه‌ای که بویس (۱۹۷۷) برای متون مانوی تهیه کرده بود، امکان دسترسی پژوهشگران به واژه‌های متون مانوی را میسر ساخت.

درباره کتیبه‌های فارسی میانه نیز چند واژه‌نامه با ارجاع به متن وجود دارد که با توجه به محدود و کم‌تعداد بودن کتیبه‌ها، تقریباً پاسخگوی پژوهش‌های فارسی میانه کتیبه‌ای است.

2. <http://pahlavazh.ihcs.ac.ir>

3. <https://www.academy.ac.il/Branches/Branch.aspx?nodeId=830&branchId=353>

4. <https://www.soas.ac.uk/nme/research/manichaeandict>

5. <http://titus.uni-frankfurt.de>

6. <http://stephanus.tlg.uci.edu/history.php>

7. <https://lrc.la.utexas.edu/eieol/aveol>

8. <https://wikisource.org/wiki/Category:Avesta> ,

https://wikisource.org/wiki/Behistun_Inscription/Column_I

9. <http://pahlavica.org/>

10. <https://www.parsigdatabase.com>

۱۱. می‌دانیم که همواره فعل‌های کمکی (h-, ēstādan, būdan) که در صرف ماضی ساده، ماضی نقلی، ماضی بعید و مجهول به کار می‌روند، به صورت هزوارش نوشته می‌شوند. در پیکره ما نیز همواره به همین صورت بود، بجز یک مورد فعل bast hēm در متن «نیرنگ زهر بستن» که با املاي <bst hm> نوشته شده بود (Jamasp-Asana, 1913, p.84).

۱۲. کل آمار استخراج شده درباره هزوارش‌ها در ۱۲ مقوله دستوری که در پایگاه در برجسب‌گذاری واژه‌ها استفاده شده است، در پژوهش دیگری به تفصیل ارائه خواهد شد.

کتاب‌نامه

- بهار، مهرداد. ۱۳۴۵. *واژه‌نامه بندهش*. تهران: بنیاد فرهنگ ایران.
- بهار، مهرداد. ۱۳۵۱. *واژه‌نامه گزیده‌های زادسپرم*. تهران: بنیاد فرهنگ ایران.
- بهار، مهرداد. ۱۳۶۹. *بندهش*. فرنیغ دادگی. تهران: توس.
- تفضلی، احمد. ۱۳۷۸. *تاریخ ادبیات ایران پیش از اسلام*. تهران: سخن.
- تفضلی، احمد. ۱۳۷۹. *مینوی خرد*. به کوشش ژاله آموزگار. تهران: توس.
- راشدمحصل، محمدتقی. ۱۳۸۵. *وزیدگی‌های زادسپرم*. تهران: پژوهشگاه علوم انسانی و مطالعات فرهنگی.
- راشدمحصل، محمدتقی. ۱۳۸۹. *دینکرد هفتم*. تهران: پژوهشگاه علوم انسانی و مطالعات فرهنگی.

تحلیل پیکره بنیان متون فارسی میانه ... (فرزانه گشتاسب و دیگران) ۲۷۹

ژینو، فیلیپ. ۱۳۸۲. *ارداویراف نامه (ارداویرازنامه)*. ترجمه و تحقیق ژاله آموزگار. تهران: معین و انجمن ایرانشناسی فرانسه.

صفای اصفهانی، زهت. ۱۳۷۶. *روایت امید اشوهیستان*. تهران: نشر مرکز.

طاووسی، محمود. ۱۳۶۵. *واژه‌نامه شایست نشایست*. شیراز: دانشگاه شیراز.

عریان، سعید. ۱۳۷۱. *متون پهلوی (ترجمه، آوانوشت)*، تهران: کتابخانه ملی جمهوری اسلامی ایران.

فروشی، بهرام. ۱۳۸۶. *فرهنگ زبان پهلوی*. تهران: انتشارات دانشگاه تهران.

قیومی، مسعود. ۱۳۹۷. *پیکره زبانی و ضرورت ساختارمندسازی داده*. در مجموعه چکیده‌های سخنرانی‌های دهمین همایش بین‌المللی زبان‌شناسی ایران، دانشگاه علامه طباطبایی، ص: ۱۰۶.

مزدپور، کتابون. ۱۳۶۹. *شایست نا شایست*. تهران: مؤسسه مطالعات و تحقیقات فرهنگی.

مزدپور، کتابون. ۱۳۷۸. *بررسی دستنویس م/او ۲۹، داستان گرشاسب، تهمورث و جمشید گلشاه و متن‌های دیگر*. تهران: آگاه.

مزدپور، کتابون و فرزانه گشتاسب و نادیا حاجی‌پور، ۱۳۹۰. «معرفی فرهنگ فارسی میانه»، *مقاله‌های نخستین همایش فرهنگ نویسی علامه دهخدا*، صص ۹۷-۱۱۷.

مکنزی، دیوید نیل. ۱۳۷۹. *فرهنگ کوچک زبان پهلوی*. ترجمه مهشید میرفخرایی. تهران: پژوهشگاه علوم انسانی و مطالعات فرهنگی.

میرفخرایی، مهشید. ۱۳۶۷. *روایت پهلوی*. تهران: مؤسسه مطالعات و تحقیقات فرهنگی.

میرفخرایی، مهشید. ۱۳۹۳. *بررسی دینکردششم*. تهران: پژوهشگاه علوم انسانی و مطالعات فرهنگی.

میرفخرایی، مهشید. ۱۳۹۷. *دادستان دینی*. تهران: پژوهشگاه علوم انسانی و مطالعات فرهنگی.

Anklesaria, T.D. 1908. *Bundahishn*. Bombay.

Anklesaria, T.D. 1911. *Datistan-i-Dinik*, Part I. Purnišn I-XL. Bombay.

Anklesaria, T.D. 1913. *Dānāk-u Mainyō-I Khard*. Bombay.

Anklesaria, B.T. 1957. *Zand-i Vohuman Yasn and two pahlavi fragments with text*. Bombay.

Bijankhan, M., J. Sheykhzadegan, M. Bahrani, & M. Ghayoomi. 2011. "Lessons from building a Persian written corpus: Peykare", *Language resources and evaluation*, 45 (2): 143-164.

Boyce, M. 1977. *A Word-List of Manichaean Middle Persian and Parthian* (Acta Iranica 9a). Tehran-Liege.

Durkin-Meisterernst, D. 2004. *Dictionary of Manichaean Middle Persian and Parthian*. Belgium.

Ghayoomi, M. 2012. "Bootstrapping the Development of an HPSG-based Treebank for Persian", *Linguistic Issues in Language Technology* 7 (1): 1-13.

Ghayoomi, M. 2014. *From HPSG-based Persian Treebanking to Parsing: Machine Learning for Data Annotation*. PhD Dissertation, Department of Mathematics and Computer Science, Freie Universität Berlin, Germany.

- Jaafari-Dehaghi, Mahmoud. 1998. *Dādestān ī Dēnīg*. Studia Iranica, Cahier 20. Paris.
- Jamasp-Asana, J.M. 1913. *Pahlavi Texts*. Bombay.
- MacKenzie, D.N. 1971. *A Concise Pahlavi Dictionary*. London.
- Macuch, M. 2009. "Pahlavi Literature", *A History of Persian Literature*, vol. XVII. ed. R.E. Emmerick & M. Macuch. New York
- Nyberg, H.S. 1964-1974. *A Manual of Pahlavi*. 2.vols (1964, vol.1; 1974, vol.2). Wiesbaden.
- Shaked, Sh. 1979. *The Wisdom of the Sasanian Sages, (Dēnkart VI)*. (Persian Heritage Series, 34). Boulder. Colorado.
- Shaked, Sh. & Carlo G. Cereti, 2005, "A Middle Persian Dictionary: Project Proposal", *Orientalia Romana 8: Middle Iranian Lexicography*, Proceedings of the Conference held in Rome, 9-11 April 2001, C. G. Cereti and M. Maggi (eds.), pp. 181-190.
- Williams, A.V., 1990, *The Pahlavi Rivāyāt Accompanying the Dādestān ī Dēnīg*.

منابع اینترنتی

- <http://pahlavazh.ihcs.ac.ir>
- <http://pldb.ihcs.ac.ir/>
- <http://titus.uni-frankfurt.de>
- <http://stephanus.tlg.uci.edu/history.php>
- <https://lrc.la.utexas.edu/eieol/aveo>
- <https://wikisource.org/wiki/Category:Avesta>
- <http://pahlavica.org/>