Language and Linguistics
Spring & Summer 2023, 18 (35)
DOI: 10.30465/LSI.2023.8502

Original Research
Received: 2023-1-8
Accepted: 2023 - 3 – 28
ISSN: 23223847

The Corpus of ATU Papers, Theses and Dissertations Abstracts

Mirzaei, Azadeh[1]
Sedghi, Fatemeh[2]

Associate Professor of Linguistics, Allameh Tabtaba'I University, Tehran, Iran
Master of Artificial Intellignece, Alzahra university, Tehran, Iran

Abstract

This study explains how to develop the corpus of "ATU Papers, Theses and Dissertations Abstracts" and introduces the different characteristics and features of the corpus. The corpus contains ten thousand thesis abstracts and 9538 article abstracts from the scientific journals of Allameh Tabatabai University with a volume of more than three and a half million tokens. Academic abstracts as brief authored texts with scientific content can depict special linguistic features and therefore, they are valuable documents. In this article, to express the importance of access to such data and to examine some features of the corpus, the word content of a part of the data has been examined and presented according to the concept of keyness and n-grams. The results showed that the lexical content of this corpus could lead researchers to propose some hypotheses. Also, the exploring n-grams of this corpus showed that the language of science has specific word clusters that can depict a particular type of language.

Keywords: corpus, Language of science, keyness, N-gram, Persian Language

1. mirzaei.azade@gmail.com 2. fatemeh.sedghi@gmail.com
How to Cite: Mirzaei, A., & Sedghi, F. (2023). The Corpus of ATU Papers, Theses and Dissertations Abstracts. *Language and Linguistics*, *18*(35), 127-145. doi: 10.30465/lsi.2023.850213:28

زبان و زبان‌شناسی
دوره ۱۸، شماره ۳۵، بهار و تابستان، سال ۱۴۰۱، ۱۲۷- ۱۴۵
DOI: 10.30465/LSI.2023.8502

مقاله پژوهشی
تاریخ دریافت: ۱۴۰۱/۱۰/۱۸؛ تاریخ پذیرش: ۱۴۰۲/۰۱/۰۸؛ تاریخ انتشار: بهار ۱۴۰۲
ISSN: 23223847

پیکرهٔ چکیده‌های مقالات و پایان‌نامه‌های دانشگاهی دانشگاه علامه طباطبائی

میرزائی، آزاده دانشیار زبان‌شناسی، دانشگاه علامه طباطبائی، تهران، ایران
صدقی، فاطمه کارشناس ارشد هوش مصنوعی، دانشگاه الزهرا، تهران، ایران

چکیده: این مقاله از نحوهٔ شکل‌گیری پیکرهٔ «چکیده‌های مقالات و پایان‌نامه‌های دانشگاهی دانشگاه علامه طباطبائی» و همچنین از ویژگی‌ها و امکانات آن می‌گوید. داده‌های این پیکره شامل ده هزار چکیده پایان‌نامه و ۹۵۳۸ چکیده مقاله (برگرفته از نشریات علمی دانشگاه علامه طباطبائی) با حجمی در حدود سه و نیم میلیون موردواژه است که در قالب طرح پژوهشی گردآوری شده‌اند. اهمیت داده‌های این پیکره، یعنی چکیده‌های دانشگاهی، از آن جهت است که این نوع داده‌ها به‌عنوان متون تألیفیِ فشرده و با محتوای علمی می‌توانند تصویرگر ویژگی‌های خاص زبان علم به‌عنوان گونه‌ای از زبان باشند. در این نوشتار برای بیان اهمیت دسترسی به چنین داده‌هایی و با هدف بررسی امکانات پیکره، محتوایِ واژه‌ایِ بخشی از داده‌ها با توجه به مفهوم کلیدی‌بودگی و فهرست چندپشته‌ها مورد بررسی قرار گرفت. بررسی‌ها نشان داد محتوای واژگانیِ این پیکره می‌تواند پژوهشگران را به‌سوی طرح برخی فرضیه‌ها سوق دهد. همچنین بررسی چندپشته‌های داده‌های علمی نشان داد که زبان علم دارای توالی‌های واژه‌ای مشخصی است که می‌تواند تصویرگر نوع خاصی از زبان باشد.

کلیدواژه‌ها: پیکره، زبان علم، کلیدی‌بودگی، چندپشته، زبان فارسی.

۱ مقدمه

دسترسی به داده‌ها و منابع زبانی به‌عنوان متون اصیل و واقعی زبان از ملزومات پژوهش‌های

زبانی به‌شمار می‌رود. بدیهی است که محتوای درونی پیکره‌های زبانی در نوع جستجو و پژوهش‌هایی که برمبنای آنها صورت می‌گیرد تأثیرگذار است. متون دانشگاهی و علمی به‌لحاظ زبانی ویژگی‌های خاصی دارند که همواره در بررسی‌های زبانی مورد توجه بوده است. امکانات دستوری، ویژگی‌های گفتمانی و محتوای واژگانیِ این داده‌ها اطلاعات ارزشمندی در اختیار قرار می‌دهد که هم در مطالعات نظری زبان و هم در پژوهش‌های کاربردی کارآمد و تأثیرگذارند. این نوشتار از نحوهٔ شکل‌گیری و محتوای درونیِ پیکرهٔ چکیده‌های مقالات و پایان‌نامه‌های دانشگاه علامه طباطبائی و مشخصات آن‌ها اطلاعاتی به‌دست می‌دهد.

گفتنی است که پیکره‌های زبانی در دو شکل کلیِ پیکره‌های برچسب‌خورده و برچسب‌نخورده موجود هستند. پیکره‌های برچسب‌خورده به‌منظور تهیه محتوای آموزشیِ مناسب برای یادگیری ماشینی با الصاق برچسب‌های زبانی و گاهی غیرزبانی تهیه می‌شوند. برای تهیهٔ این پیکره‌ها پس از گردآوری دادهٔ اولیه لازم است که برچسب‌گذاری به‌صورت دستی یا نیمه‌دستی صورت بگیرد. اما در تولید پیکره‌های برچسب‌نخورده، تهیهٔ داده مهم‌ترین بخش تولید پیکره است. امروزه این داده‌ها معمولاً به‌شکل خزش ماشینی و براساس محتوای موجود در صفحات وب به‌دست می‌آیند. پس از گردآوری داده‌ها، پیش‌پردازش‌های متنی و استانداردسازیِ داده از الزامات کار است. بدیهی است که در صورت وجود بسته‌های داده‌ای، فرایند تولید پیکره از مرحلهٔ استانداردسازی، دسته‌بندی متنی و ثبت اطلاعات فراداده[1] آغاز می‌شود.

منظور از اطلاعات فراداده، ویژگی‌ها و مشخصاتی از داده‌ها و پیکره‌های زبانی است که با امکان دسته‌بندی بخش‌های مختلف پیکره، در تسهیل فرایندهای پژوهشی مؤثر هستند. برای مثال در پیکره‌های زبان‌آموز که از دست‌نوشته‌های زبان‌آموزان گردآوری می‌شوند، تعیین سن، جنسیت، سطح زبانی، ملیت و زبان مادری زبان‌آموز تولیدکنندهٔ دست‌نوشته‌ها، همچنین اطلاعات هر سند مانند موضوع آن، نوع تکلیف، زمان تولید سند، نمونه‌هایی از اطلاعات فراداده هستند.

در این مقاله به توصیف مراحل تولید و همچنین ویژگی‌ها و امکانات پیکرهٔ «چکیده‌های مقالات و پایان‌نامه‌های دانشگاهی دانشگاه علامه طباطبائی» که از این پس و در ادامه با نام پیکرهٔ «چکیده‌های علمی» معرفی می‌شود، می‌پردازیم. همان‌طور که از نام

1. metadata

پیکره مشخص است، اسناد درون این پیکره همگی چکیده‌های علمی هستند. چکیدهٔ علمی دارای ویژگی‌های ارزشمندی است که براساس آن می‌تواند نمایندهٔ نوع خاصی از زبان باشد و همین موضوع اهمیت ساخت پیکرهٔ حاضر را مشخص می‌کند. برخی از این ویژگی‌ها به‌شرح ذیل هستند:

۱. چکیده‌های علمی غالباً متون تألیفی هستند و برخلاف بخش‌های دیگر مقاله‌ها و پایان‌نامه‌ها که معمولاً تحت تأثیر متون غیرفارسی قرار می‌گیرند، ویژگی‌های زبان فارسی علمی را بیشتر در خود جای می‌دهند.

۲. چکیده‌ها متون فشرده‌ای هستند که اگر اصولی نوشته شوند اطلاعات زیادی را دربرمی‌گیرند.

۳. بدیهی است که خلاصه‌نویسی و فشرده‌نویسی تابع شرایط و امکانات زبانی است و با بررسی ساختاری این گونه دست‌نوشته‌ها اطلاعات ارزشمندی حاصل می‌شود.

۴. به‌دلیل وجود کلیدواژه‌های ذیل چکیده‌ها امکان وزن‌دهی به واژه‌های درون چکیده و آمارگیری مناسب از محتوای واژگانی چکیده‌ها فراهم است.

۵. چون چکیده آغازگر و معرف متن علمی و محتوای درونی آن است، دقت زیادی برای نگارش و ویرایش آنها صورت می‌گیرد و به همین دلیل می‌تواند نمایندهٔ نوشتار ویراستهٔ علمی باشد.

با این توصیف، در ادامه ابتدا، در بخش پیشینهٔ پژوهش، فهرستی از برخی پیکره‌های زبان فارسی ارائه می‌شود و پس از آن و در بخش بعدی، مراحل تولید پیکرهٔ «چکیده‌های علمی» ارائه می‌شود. سپس در بخش پایانی محتوای واژگانی بخشی از پیکره با توجه به مفهوم کلیدی‌بودگی[1] بررسی می‌شود.

۲ پیشینهٔ پژوهش

پژوهش‌های زبان‌شناسی پیکره‌ای در یک معنا و در آنجا که این مفهوم با رایانه و تحولات پیدایشی و پیشرفت آن همراه می‌شود، مفهومی جدید به‌نظر می‌رسد. در این معنا پیدایش نخستین پیکره‌ها با تولد نخستین رایانه‌ها همزمان است و هرچه در علوم رایانه پیشرفت حاصل می‌شود، پیشرفت مطالعات زبان‌شناسی پیکره‌ای هم سرعت می‌گیرد.

1. keyness

در زبان فارسی پیکره‌های برچسب‌خورده و برچسب‌نخوردهٔ متنوعی وجود دارد که در نوع اول می‌توان به درخت‌بانک[1] نحوی زبان فارسی با رویکرد دستور ساخت سازه‌ای هسته‌بنیان (قیومی، ۲۰۱۲ و قیومی و کوهن[2]، ۲۰۱۴)، پیکرهٔ وابستگی نحوی زبان فارسی (رسولی و همکاران، ۲۰۱۳)، پیکرهٔ درختی وابستگی فارسی اوپسالا (سراجی، ۲۰۱۵)، پیکره گزاره‌های معنایی زبان فارسی (میرزائی و مولودی، ۲۰۱۶)، پیکرهٔ گفتمانی زبان فارسی (میرزائی و صفری، ۲۰۱۸)، پیکرهٔ مرجع‌دارهای زبان فارسی (میرزائی و صفری، ۲۰۱۸) و مانند آن اشاره کرد. همچنین از پیکره‌های برچسب‌نخورده می‌توان مواردی به‌شرح زیر را برشمرد:

• پیکرهٔ متنی زبان فارسی: با حجمی بالغ بر صد میلیون کلمه برگرفته از منابع متنوع نوشتاری فارسی رسمی معاصر که موردواژه‌یابی[3] شده و بخشی از آن (حدود هشت میلیون کلمه) برچسب مقولهٔ دستوری هم دریافت کرده است (بی‌جن‌خان و همکاران، ۲۰۱۱).

• پیکره دادگان زبان فارسی: مجموعه‌دادهٔ تهیه‌شده در پژوهشگاه علوم انسانی و مطالعات فرهنگی که از متون داستانی، غیرداستانی، نثر، نظم، متون آموزشی، متون روزنامه‌ها و مانند آن گردآوری شده است (عاصی، ۱۳۸۴). بخشی از این مجموعه داده حاوی متون تاریخی مربوط به قرن‌های ۵ تا ۷ هجری شمسی است که به‌عنوان فاز چهارم کار در سال ۱۳۹۵ به بهره‌برداری رسیده است.

• پیکرهٔ همشهری: توسعه‌یافته با خزش روزنامهٔ همشهری سال‌های ۱۳۷۶ تا ۱۳۸۱ با حجمی بالغ بر ۱۶۰ هزار مقالهٔ خبری در ۸۲ دستهٔ موضوعی که به‌صورت رایگان در دسترس است (آل‌احمد و همکاران، ۲۰۰۹).

• پیکرهٔ ناب: مجموعه داده‌ای شامل حدود ۱۳۰ گیگابایت دیتا متنی و معادل ۲۵۰ میلیون پاراگراف و ۱۵ میلیارد کلمه. این مجموعه استاندارد و پاک‌سازی‌شده و برای محققان حوزهٔ پردازش زبان طبیعی در فارسی دردسترس است (صبوری و همکاران، ۲۰۲۲).

• پیکرهٔ علمی فارسی: مجموعه‌داده‌ای از متون علمی فارسی با حجمی در حدود ۵۱ میلیون واژه از نشریات علمی زبان فارسی است که با خزش ماشینی و به‌صورت خودکار به‌منظور تهیه فهرست بسامدی واژه‌های علمی گردآوری شده است. فهرستی بسامدی

1 treebank 2. J. Kuhn 3. tokenization

مستخرج از این پیکره شامل ۳۰۷ بن‌واژه، تعیین و معرفی شده که هم به‌لحاظ تعداد و هم پراکندگی پوشش مناسبی داشته‌اند (رضائی شریف‌آبادی، ۱۴۰۱).

آن‌چنانکه گفته شد، هدف مقالهٔ حاضر معرفی پیکرهٔ «چکیده‌های علمی» زبان فارسی است که در دستهٔ پیکره‌های برچسب‌نخورده قرار می‌گیرد و تنها موردواژه‌یابی و بن‌واژه‌یابی بر روی آن صورت گرفته است. در معرفی محتوای درونیِ این پیکره دسته‌بندی‌های موضوعی ارائه می‌شود که به‌شکل دستی براساس طبقه‌بندی موضوعی نشریات و پایان‌نامه‌های دانشگاه علامه طباطبائی به‌دست داده شده‌اند. بدیهی است که دسته‌بندی موضوعی متون به‌صورت خودکار، زمینهٔ پژوهشی‌ای است که به‌عنوان نمونه‌ای از آن و در موضوع زبان علم در زبان فارسی می‌توان به قیومی و موسویان (۱۴۰۱) اشاره کرد.

۳ ساختار داده

پیکرهٔ «چکیده‌های علمی» مجموعه‌داده‌ای در ژانر علمی است که ازنظر انجام اصلاحات و پردازش‌های دستی و با نظارت انسانی، داده‌های متمایز و ارزشمندی را به‌دست داده است. تعداد کلی چکیده‌های موجود در پیکره شامل ده‌هزار چکیدهٔ پایان‌نامه و ۹۵۳۸ چکیدهٔ مقاله است. این چکیده‌ها درراستای اجرای طرح «ساخت پیکرهٔ چکیده‌های مقالات و پایان‌نامه‌های فارسی دانشگاه علامه طباطبائی» از چکیده‌های پایان‌نامه‌های کارشناسی ارشد و رساله‌های دکتری دفاع‌شده در دانشگاه علامه طباطبائی و چکیده‌های مقالات نشریات علمی دانشگاه علامه طباطبائی گردآوری شده‌اند. همهٔ چکیده‌های مورداشاره ازطریق پردازش ماشینی آماده شده و در پایگاه داده SQL server ذخیره شده است. از میان چکیده‌های مورداشاره تاکنون تعداد ۴۵۸۸ چکیدهٔ پایان‌نامه و ۴۹۵۰ چکیدهٔ مقاله پس از ویرایش و استانداردسازی و موردواژه‌یابی توسط کتابخانه هضم[1] ازطریق نظارت انسانی هم ویرایش شده است.

ساختار دادهٔ اولیه به این شکل است که برای هر چکیدهٔ مقاله یا پایان‌نامه، علاوه بر متن چکیده، عنوان پژوهش، کلیدواژه‌ها، نام نویسنده، تاریخ تولید یا انتشار سند و دستهٔ موضوعی سند که برای مقالات با عنوان نشریه و برای پایان‌نامه با فیلدی به نام «رشته» مشخص می‌شود، موجود است. برای بسیاری از چکیده‌ها، معادل انگلیسی چکیده و

1. https://github.com/sobhe/hazm

کلیدواژه‌های انگلیسی هم در اختیار است. برای پایان‌نامه‌ها علاوه بر این مشخصات، نام استاد راهنما هم موجود است.

به‌دلیل نایکدستی‌های زیادی که در اطلاعات فرادادهٔ مرتبط با هر سند و همچنین در متن چکیده‌ها وجود داشت، در فاز اول چکیده‌هایی که مشکلات کمتری دارند بررسی شدند. با وجود این انتخاب، نیاز به ویرایش و پیش‌پردازش اسناد و اطلاعات فرادادهٔ مرتبط با آن همچنان وجود داشت. برای ترسیم فرایند اجرایی کار، برخی از چالش‌های پیش رو و اقدامات پردازشی صورت‌گرفته به‌شرح زیر قابل معرفی هستند:

۱) اصلاح اشتباهات تایپی و املایی؛

۲) اصلاح تداخل‌های متنی: گاهی چکیده‌های انگلیسی با چکیده‌های فارسی یا نام نویسنده و ایمیل‌آدرس نویسنده مخلوط شده بود که باید اصلاح می‌شد؛

۳) اصلاح علائم نگارشی: کتابخانهٔ هضم نقطه را علامت پایان جمله درنظر می‌گیرد. در نتیجه، در مواردی که نویسنده از نقطه به‌عنوان علامت اعشار استفاده کرده یا در اختصارهایی چون «ه. ق.» یا در شماره‌گذاری‌هایی مانند «۱. ۲. ۳.»، در علائم دو نقطه (..) و سه نقطه (...) به‌معنی «غیره»[1] به مشکل می‌خورد؛

٤) یکدستیِ نویسه: تنوع نوشتاری در یک واژهٔ یکسان مانند «سوال»، «سؤال» و «سئوال» که همگی باید به یک شکل تبدیل شوند. این وضعیت درخصوص واژه‌های زیادی مانند مولفه، مسئله، مسایل، ارایه و مانند آن وجود داشت.

۵) جداکردن فعل «است» از کلمه قبلی در مواردی مانند «کتابیست»؛

٦) شناسایی پایان جمله و درج نقطه در مواردی که نویسنده از نقطه و علائم نگارشی استفاده نکرده است؛

۷) حذف فاصله‌های اضافی و اصلاح اشکالات نیم‌فاصله‌نویسی: برای نمونه اصلاح «بیشترین‌میزان» به «بیشترین میزان»؛

۸) اصلاح علائم و کاراکترهای اشتباه: گاهی به‌جای علامت نیم‌فاصله از خط تیره یا برخی حروف استفاده شده است؛

۹) اصلاح اشکالات بن‌واژه‌یاب کتابخانهٔ هضم

۱۰) اصلاح برخی واژه‌ها و اصطلاحات تخصصی: به‌دلیل نوع داده‌های آموزشی در زیرساخت کتابخانهٔ هضم برخی اصطلاحات تخصصی متون علمی برای این کتابخانه ناآشنا

۱. هضم سه نقطه را به‌عنوان «غیره» تشخیص می‌دهد ولی دو نقطه را علامت پایان جمله در نظر می‌گیرد.

بود و درنتیجه غیرقابل تشخیص. این موارد باید شناسایی و به‌صورت دستی اصلاح می‌شد؛ برای مثال تبدیل «سره نویسی» به «سره‌نویسی» که در متون عمومی کم‌کاربرد است و به رشته‌ای خاص تعلق دارد و بنابراین هضم آن را نمی‌شناسد و خطایش را تشخیص نمی‌دهد.

پس از استانداردسازی داده امکان آمارگیری از پیکره و ارائهٔ اطلاعات آماری فراهم است. حجم کلی پیکره براساس موردواژه‌یابیِ ماشینی در حدود سه‌ونیم میلیون موردواژه و یکتاواژه/ نوع‌واژه[1] تقریباً ۱۴۰ هزار مورد است. بدیهی است که پس از نظارت انسانی این آمار تاحدودی تغییر می‌کند. حجم بخش ویرایش‌شدهٔ پیکره (با نظارت انسانی) در حدود یک میلیون و نهصد موردواژه و کلمات یکتای این بخش در حدود ۷۳ هزار مورد است. در این بخش از پیکره ۵۷۲۳۵ بن‌واژه قابل شناسایی و شمارش است.

همان‌طور که در جدول زیر مشخص است، در بخش ویرایش‌شدهٔ پیکره، تعداد ۴۹۵۰ چکیده مقاله و ۴۵۸۸ چکیدهٔ پایان‌نامه ثبت شده است. تعداد بن‌واژه‌ها و موردواژه‌های هر بخش نیز در جدول ذیل مشخص است.

جدول ۱- اطلاعات آماری از بخش ویرایش‌شده پیکره

	تعداد چکیده‌ها	موردواژه‌ها	بن‌واژه‌ها
مقاله	۴۹۵۰	۱۰۳۴۴۰۳	۳۲۴۰۶
رساله	۴۵۸۸	۸۵۵۶۷۷	۲۴۸۲۹
مجموع	۹۵۳۸	۱۸۹۰۴۷۳	۵۷۲۳۵

همان‌طور که گفته شد، پیش‌پردازش‌های ماشینی ازطریق کتابخانهٔ هضم صورت گرفته است. هضم کتابخانهٔ پایتونی برای پردازش زبان فارسی است که امکان استانداردسازی داده، قطعه‌بندی[2]، موردواژه‌یابی، بن‌واژه‌یابی، تحلیل صرفی و نحوی داده‌های زبان فارسی را فراهم می‌آورد.

در میان اطلاعات فرادادهٔ این پیکره شامل زمان تولید سند و جنسیت نویسنده، موضوع[3] و حوزهٔ پژوهشی هر چکیده از مهم‌ترین اطلاعاتی است که امکان دسته‌بندی متون و پژوهش در متون علمی را فراهم می‌کند. داده‌های این پیکره به‌لحاظ موضوعی در بخش

1. type
2. segmentation
۳. اطلاعات آماری این بخش و در بحث موضوع برروی کل داده‌های پیکره است.

نشریات در ۳۹ عنوان موضوع قرار می‌گیرند که در جدول ۲ قابل مشاهده است. در این جدول در کنار نام نشریه (و به‌بیانی دیگر دستهٔ موضوعی) تعداد چکیده‌های موجود در هر بخش نیز مشخص شده است.

جدول ۲- عنوان نشریات دانشگاه علامه طباطبائی به‌عنوان منبع دریافت چکیده‌های پیکره

فصلنامه اقتصاد محیط زیست و منابع طبیعی	۱۸	سراج منیر	۲۰۲
پژوهش‌های بیمه‌ای	۱۳	مطالعات روان‌شناسی بالینی	۲۷۷
دولت‌پژوهی	۱۲۵	پژوهش‌های کیفی در برنامهٔ درسی	۵۴
فصلنامه برنامه‌ریزی توسعه شهری و منطقه‌ای	۲۷	پژوهش‌نامهٔ مددکاری اجتماعی	۱۰۳
پژوهش در مدیریت ورزشی	۵۴	اندیشه علامه طباطبائی	۴۴
فناوری آموزش و یادگیری	۸۶	علم زبان	۷۰
فرهنگ مشاوره و روان درمانی	۳۱۷	مطالعات پیش‌دبستان و دبستان	۳۶
پژوهشنامه اقتصاد انرژی ایران	۲۰۳	پژوهش‌های اقتصادی ایران	۵۵۹
روان‌شناسی افراد استثنایی	۳۱۰	مطالعات تجربی حسابداری مالی	۴۰۳
پژوهش حقوق عمومی	۵۱۷	مطالعات مدیریت صنعتی	۴۱۴
فصلنامه علوم اجتماعی	۵۲۰	فصلنامه مطالعات دانش‌شناسی	۱۳۷
فصلنامه روان‌شناسی تربیتی	۴۴۲	پژوهشنامه اقتصادی	۶۵۴
پژوهش‌های رهبری و مدیریت آموزشی	۹۶	دو فصلنامه دانش های بومی ایران	۷۳
پژوهش‌های راهبردی سیاست	۲۳۸	مطالعات مدیریت کسب و کار هوشمند	۱۹۸
مطالعات مدیریت گردشگری	۳۲۳	پژوهشنامه معارف قرآنی	۲۶۷
پژوهش‌های ترجمه در زبان و ادبیات عربی	۱۴۵	مطالعات رسانه‌های نوین	۱۶۰
پژوهش حقوق خصوصی	۲۰۹	حکمت و فلسفه	۳۹۷
مطالعات مدیریت (بهبود و تحول)	۵۹۰	فصلنامه پژوهش حقوق کیفری	۲۰۸
فصلنامه اندازه‌گیری تربیتی	۳۱۲	متن پژوهی ادبی	۶۸۶
برنامه‌ریزی رفاه و توسعهٔ اجتماعی	۲۷۸		

بدیهی است که برخی از نشریاتی که در جدول ۲ وجود دارند، با توجه به شباهت موضوعی (همانند آنچه در جدول ۳ قابل مشاهده است) می‌توانند در دسته‌های کلان‌تری گروه‌بندی شوند. البته گفتنی است که در هنگام استفادهٔ پژوهشی از پیکره، هم می‌توان داده‌ها را با دسته‌بندی موضوعیِ کلان آن مورد توجه و بررسی قرار داد و هم اینکه در زیربخش‌ها و با توجه به دسته‌بندی موضوعی که در جدول ۲ وجود دارد.

جدول ۳- موضوعات کلی چکیده‌های مقالات پیکره

اقتصاد	۱۴۳۴	زبان و ادبیات عربی	۱۴۵
علوم سیاسی	۱۲۹۷	مطالعات رسانه‌های نوین	۱۶۰
علوم اجتماعی	۷۲۳	حکمت و فلسفه	۳۹۷
مدیریت	۱۹۸۲	زبان‌شناسی	۷۰
روان‌شناسی و علوم تربیتی	۲۰۶۷	زبان فارسی و متن پژوهی ادبی	۶۸۶
معارف	۵۱۳	پژوهش‌های بیمه‌ای	۱۳

در ارتباط با چکیدهٔ پایان‌نامه‌ها نیز اطلاعاتی چون عنوان چکیده، نام پژوهشگر (و درنتیجه جنسیت نویسنده)، سال انتشار، رشته و گرایش تحصیلی، نام استاد راهنما، درجهٔ پایان‌نامه (ارشد یا دکتری)، کلیدواژه و چکیدهٔ انگلیسی موجود است. البته برخی از چکیده‌ها معادل انگلیسی و کلیدواژهٔ انگلیسی ندارند. اما با توجه به وجود چکیده‌هایی که معادل انگلیسیِ آنها هم موجود هستند می‌توان ایدهٔ ایجاد پیکرهٔ موازیِ ترجمه‌آموز را هم برای مرحلهٔ بعدی درنظر داشت. در هر حال در نسخهٔ ۱ پیکره، برای هر چکیده اطلاعات فراداده‌ای چون سال انتشار، رشتهٔ تحصیلی و نام پژوهشگر (و درنتیجه جنسیت نویسنده)، ثبت و استانداردسازی شده است.

برای دسته‌بندی موضوعیِ چکیده‌ها از رشتهٔ تحصیلیِ مرتبط با هر چکیده استفاده شد. بر این اساس، چکیده‌ها در ۱۱ دستهٔ موضوعی به ترتیبی که در جدول ۴ مشخص است، دسته‌بندی شدند. بدیهی است که این کلان‌دسته‌ها دارای زیربخش‌هایی هستند که در اطلاعات فراداده‌ای هر چکیده موجود است و بنابراین در هنگام استفادهٔ پژوهشی از پیکره می‌تواند مورد استفاده و استناد قرار بگیرد. برای مثال در شاخهٔ روان‌شناسی و علوم تربیتی، زیرشاخه‌های مشاورهٔ خانواده، روان‌شناسی عمومی، روان‌شناسی بالینی، مشاورهٔ تحصیلی و مانند آن وجود دارد.

جدول ۴- دسته‌های موضوعی چکیده‌های پایان‌نامه‌ها

موضوع	فلسفه	ادبیات عرب	زبان‌شناسی	زبان و ادبیات فارسی	علوم اجتماعی	حقوق و علوم سیاسی	اقتصاد	روان‌شناسی و علوم تربیتی	علوم قرآنی	مدیریت و حسابداری
تعداد	۳۳۸	۲۳۸	۳۹۷	۵۶۷	۷۴۳	۱۰۶۳	۴۶۶	۲۷۱۰	۵۰	۱۶۲۵

در جدول ۵ مجموع چکیده‌های پایان‌نامه‌ها و نشریات علمی با توجه به شباهت موضوعی یا یکسانی دقیق مشخص شده است. بر این اساس، پیکرهٔ چکیده‌های علمی در ۱۲ دستهٔ موضوعی قابل استفاده است[1].

جدول ۵- مجموعه چکیده‌های مقالات و پایان‌نامه‌ها

روان‌شناسی و علوم تربیتی	۴۷۷۷	حکمت و فلسفه	۷۳۵
مدیریت و حسابداری	۳۶۰۷	معارف	۵۶۳
حقوق و علوم سیاسی	۲۳۶۰	زبان‌شناسی	۴۶۷
اقتصاد	۱۹۰۰	زبان و ادبیات عربی	۳۸۳
علوم اجتماعی	۱۴۶۶	مطالعات رسانه‌های نوین	۱۶۰
زبان فارسی و متن پژوهی ادبی	۱۲۵۳	پژوهش‌های بیمه‌ای	۱۳

همان‌طور که از فراوانی چکیده‌ها مشخص است، پیکره به‌لحاظ موضوعی متوازن[2] نیست. یعنی تعداد اسناد هر دسته متفاوت است و این اختلاف در برخی موارد خیلی هم زیاد است. مثلاً چکیده‌های رشتهٔ زبان‌شناسی ۴۶۷ مورد و چکیده‌های حوزهٔ مدیریت ۳۶۰۷ مورد است. بدیهی است که گردآورندگان این پیکره به‌منظور ایجاد توازن موضوعی، چکیده و داده‌ای را از مجموعهٔ داده‌ها خارج نمی‌کنند و انتظار می‌رود که در هر پژوهشی که قرار است دسته‌های موضوعیِ مختلف در ارتباط با هم مورد بررسی و توجه قرار گیرند، با امکان فیلترکردن و با توجه به تعداد موردواژه‌های در هر موضوع، دادهٔ متوازنی از پیکره استخراج شود و براساس آن دادهٔ متوازن بررسی و پژوهش مورد نظر صورت بگیرد.

۴ پژوهش موردی در پیکره

در این بخش برای بیان اهمیت دسترسی به داده‌های موجود در پیکرهٔ چکیده‌های علمی و

۱. آن‌چنان‌که مشخص است، در این مقاله و برای ثبت اطلاعات فراداده، «موضوع» هر سند به‌صورت دستی و با توجه به اطلاعات اولیهٔ ثبت‌شده برای دادهٔ خام تعیین می‌شود. به این ترتیب، هر سند با توجه به اطلاعات اولیه، در یک طبقهٔ موضوعی قرار می‌گیرد. پس از آن، موضوع‌های مشابه در یک کلان‌دسته قرار می‌گیرند و بنابراین برای هر سند یک کلان‌موضوع هم ثبت می‌شود. برای مثال، مقاله‌ای با عنوان «کارایی فناوری و نرم‌افزارهای آموزشی در یادگیری واژگان زبان خارجی» در موضوع «آزفا» و کلان‌موضوع «زبان‌شناسی» طبقه‌بندی می‌شود.

2. balanced

به جهت بررسی امکانات آن، محتوای واژه‌ای بخشی از داده‌ها با توجه به مفهوم کلیدی‌بودگی مورد بررسی قرار گرفته و ارائه شده است. همچنین بررسی توالی‌ها و خوشه‌های واژه‌ای و موردپژوهی دیگری است که در ادامه به آن می‌پردازیم.

۴-۱ واژه‌های کلیدی و مفهوم کلیدی‌بودگی

در این بخش برای بررسی محتوای درونی پیکره، واژه‌های کلیدی دو حوزهٔ زبان‌شناسی و فلسفه را براساس مفهوم کلیدی‌بودگی استخراج کرده‌ایم. کلیدی‌بودن یک شاخص آماری است که براساس آن کلمه‌های تخصصی و کلیدی متعلق به هر متن مشخص می‌شود. برای تعیین میزان کلیدی‌بودگی محتوای واژگانی پیکره و دادهٔ آزمون، بسامد واژه‌ها در این داده تعیین می‌شود و سپس فهرست بسامدی پیکرهٔ موردنظر با فهرست بسامدی پیکرهٔ مرجع[1] مقایسه می‌شود (میرزائی، ۱۴۰۰: ۱۸۵).

پیکرهٔ مرجع در بحث کلیدی‌بودگی، پیکره‌ای است با ویژگی‌های خاص که براساس آن می‌توان درخصوص وزن و اهمیت محتوای واژگانی پیکرهٔ آزمون قضاوت کرد. یکی از این ویژگی‌ها اندازهٔ پیکره است؛ یعنی پیکرهٔ مرجع باید از پیکرهٔ آزمون بزرگتر باشد. بربر ساردینا[2] (۲۰۰۴: ۳۰۱-۳۰۳) معتقد است که حجم این پیکره باید حداقل پنج برابر پیکرهٔ مورد مطالعه باشد. علاوه بر اندازه، تنوع ژانر، تنوع موضوع، تنوع در زمان تولید متون و گونه‌های زبانی نیز برای پیکرهٔ مرجع مهم است و برخی پژوهش‌ها حتی تنوع ژانری را مهم‌تر و تأثیرگذارتر از حجم پیکره می‌دانند. با توجه به این موضوع در اینجا و در این بررسی برای پیکرهٔ مرجع از مجموع داده‌های ۴ بخش زبان‌شناسی، مدیریت، حقوق و علوم سیاسی و فلسفه استفاده شده است. این انتخاب سبب شد که حجم بیش از پنج برابر و تنوع موضوعی تأمین شود، سپس دادهٔ دیگری بالغ بر ۲۱۳۳۸۷ موردواژه که از متون روایی- داستانی و متون ژورنالیستی تهیه شده بود به این مجموعه اضافه شد. به این ترتیب حجم پیکرهٔ مرجع در این بررسی به ۷۲۱۵۴۵ موردواژه رسید.

پس از تهیه و تعیین پیکرهٔ مرجع امکان بررسی محتوای واژگانی پیکرهٔ آزمون فراهم است. اگر در مقایسه با پیکرهٔ مرجع، یک واژه از پیکرهٔ آزمون به‌شکل غیرمعمولی پربسامد یا کم‌بسامد باشد آن واژه برای آن متن، کلیدی است.

1. reference corpus
2. A. P. Berber Sardinha

در اینجا برای بررسی محتوای واژگانیِ بخش‌های مورداشاره، ابتدا خروجیِ متنیِ این بخش‌ها از پیکره استخراج شد. سپس فایل‌های متنیِ حاصل به‌عنوان دادهٔ آزمون در نرم‌افزار انت‌کانک[1] وارد شد. ازسوی دیگر، پیکرهٔ مرجع تعیین‌شده در بخش موردنظر اضافه شد. نتیجهٔ بررسی آماریِ صورت‌گرفته در تصاویر ۱ تا ۲ قابل مشاهده است.

Keyword List Results 3

Keyword Types: 299 **Keyword Tokens:** 16844 **Search Hits:** 0

Rank	Freq	Keyness (LL4)	Effect (DICE)	Keyword
1	772	+ 2099.14	0.0319	زبان
2	525	+ 1360.43	0.0219	فارسی
3	196	+ 563.44	0.0083	زبانی
4	171	+ 490	0.0073	واژه
5	176	+ 478.98	0.0075	نحوی
6	137	+ 415.58	0.0058	گویش
7	152	+ 382.89	0.0065	معنایی
8	133	+ 372.83	0.0057	افعال
9	114	+ 339.54	0.0049	واژگانی
10	150	+ 325.95	0.0064	فعل
11	155	+ 251.57	0.0066	ساخت
12	89	+ 235.26	0.0038	مرکب
13	109	+ 220.86	0.0046	کاربرد
14	71	+ 215.29	0.003	آوایی
15	82	+ 213.39	0.0035	گفتار
16	404	+ 205.03	0.0162	بررسی
17	90	+ 176.43	0.0038	شناختی
18	77	+ 166.38	0.0033	کودکان
19	55	+ 165.5	0.0024	بسامد
20	54	+ 163.72	0.0023	واج

شکل ۱- بیست واژهٔ کلیدی دادهٔ زبان‌شناسی

از بررسی ۲۰ واژهٔ پربسامد حوزهٔ زبان‌شناسی می‌توان به برخی نتایج نائل آمد، سمت و سوی پژوهش‌های زبان‌شناختیِ صورت‌گرفته را مشخص کرد و پیش‌فرض‌های پژوهشی را به‌دست داد. برای مثال، طبق نتایج نمایش‌داده‌شده در تصویر بالا و در بیست واژهٔ کلیدیِ پربسامد حاصل از بررسی پیکرهٔ چکیده‌های علمی حوزه زبان‌شناسی، مشخص می‌شود که از حوزه‌های اصلیِ زبان‌شناسی شامل صرف، نحو، معنی‌شناسی و واج‌شناسی، ابتدا مفاهیم «نحوی» و بعد «معنایی» بسامد وقوع بیشتری داشته و دو حوزهٔ دیگر در این ۲۰ جستجو دیده نمی‌شوند.

در مقابل، از میان واحدهای زبانی شامل آوا و واج، تکواژ، واژه، جمله و متن، «واژه» و

1. https://www.laurenceanthony.net/software/antconc/

بعد از آن «واج» پربسامدتر است. همچنین در میان مقوله‌های دستوری شامل اسم، صفت، فعل، قید و حرف اضافه، فقط «فعل» در ۲۰ واژهٔ پربسامد این حوزه جای گرفته است. از جمع این نتایج می‌توان این فرض را مطرح کرد که عمدهٔ پژوهش‌های صورت‌گرفته در این حوزه و در پیکرهٔ چکیده‌های علمی حاضر، صرفی–نحوی هستند. همچنین با توجه به آنکه زبان فارسی در رتبهٔ دوم و بدون رقیب در ۲۰ جستجوی اولیه ظاهر شده است، می‌توان گفت که زبان بررسی در این پژوهش‌ها زبان فارسی بوده است.

Keyword List Results 2

Keyword Types: 306 Keyword Tokens: 17266 Search Hits: 0

Rank	Freq	Keyness (LL4)	Effect (DICE)	Keyword
1	365	+ 977.24	0.0135	فلسفه
2	181	+ 486.81	0.0068	فلسفی
3	177	+ 445.95	0.0066	دین
4	181	+ 354.28	0.0067	انسان
5	125	+ 338.12	0.0047	معرفت
6	114	+ 319.4	0.0043	کانت
7	109	+ 291.01	0.0041	ابن
8	131	+ 283.48	0.0049	علم
9	109	+ 274.62	0.0041	عقل
10	101	+ 267.49	0.0038	عالم
11	109	+ 264.52	0.0041	دینی
12	156	+ 262.73	0.0058	نظریه
13	107	+ 238.88	0.004	نفس
14	122	+ 218.44	0.0046	اندیشه
15	79	+ 203.85	0.003	منطق
16	96	+ 201.16	0.0036	کتاب
17	71	+ 197.64	0.0027	ارسطو
18	68	+ 188	0.0025	فلاسفه
19	70	+ 185.37	0.0026	سینا
20	116	+ 182.95	0.0043	بحث
21	63	+ 176.46	0.0024	ملاصدرا
22	74	+ 175.81	0.0028	آراء

شکل ۲- بیست واژهٔ کلیدی دادهٔ فلسفه

از بررسی ۲۰ واژهٔ کلیدیِ حوزهٔ فلسفه در پیکرهٔ چکیده‌های علمی نیز می‌توان پیش‌فرض‌هایی را مطرح کرد. در واژه‌های کلیدیِ پربسامد فلسفه، «انسان» از پربسامدترین واژه‌ها است. این فراوانی این فرض را ایجاد می‌کند که مهم‌ترین موضوع بررسی در این حوزه انسان است و با توجه به پربسامدبودن «معرفت»، «عقل» و «اندیشه» می‌توان گفت که در موضوع انسان، تفکر و اندیشهٔ انسانی مهم‌ترین موضوع بحث است.

از دیگر واژه‌های محتواییِ پربسامد «دین» و «دینی»، «علم»، «عالم»، «نظریه» و «منطق» است که در پژوهش‌های این حوزه معنادار است. «بحث» واژهٔ دیگری است که به‌نظر

می‌رسد نسبت به دیگر واژه‌ها از عمومیت بیشتری برخوردار است و حتی در داده‌های دیگر می‌تواند به‌عنوان فعلیار، در ردیف واژه‌های ایستا[1] قرار گیرد. اما اینجا مشخص است که با توجه به روح حاکم بر این حوزه، «بحث» از کلیدواژه‌های معنادار است و این خود نتیجهٔ درخور توجهی است.

موضوع قابل بررسیِ دیگر، بسامد وقوع و ترتیب حضور فلاسفهٔ غربی و شرقی در میان کلیدواژه‌های این حوزه است. در میان فلاسفهٔ غربی مانند کانت، نیجه، هگل، دکارت، فرگه و مانند آنها، «کانت» پربسامدترین است که با توجه به تأثیرگذاریِ کانت در مطالعات فلسفی و تغییر مسیر این مطالعات پس از او نتیجه جالبی است.

بررسی دیگری که در واژه‌های کلیدیِ این حوزه می‌تواند صورت بگیرد، بررسی بسامدِ وقوعِ فلاسفهٔ غربی و شرقی و در نتیجه تأثیرگذاری آنها و رویکردهایشان در مطالعاتِ فلسفی است. بررسی ۱۵۰ واژهٔ کلیدی حوزهٔ فلسفه نشان می‌دهد که فلاسفهٔ مطرح از پربسامد به کم‌بسامد به ترتیب زیر است:

کانت، ابن‌سینا، ارسطو، ملاصدرا، هگل، دکارت، افلاطون، هایدگر، سهروردی، نیچه، غزالی، فارابی، هیوم و ویتگنشتاین.

این توالی و فراوانی آنها احتمالاً در نشان دادن سمت و سوی پژوهش‌های فلسفی تأثیرگذار است.

۲-۴ چندپشته‌ها

موضوع قابل بررسیِ دیگر در داده‌های پیکره‌ای و اینجا در پیکرهٔ چکیده‌های علمی، چندپشته‌ها[2] است. چندپشته در زبان‌شناسی رایانشی و پیکره‌ای به توالی‌های واحدهای زبانیِ درون متن گفته می‌شود که می‌تواند هر تعداد عضوی داشته باشد و براساس همان تعدادِ عضو، مقدار چند در چندپشته مشخص می‌شود. در زبان‌شناسی رایانشی از چندپشته‌ها برای مدل کردن زبان و پیش‌بینی کلمات در یک دنبالهٔ داده استفاده می‌شود. در این حالت بسامد وقوع و فراوانی چندپشته‌ها محاسبه می‌شود تا ماشین بتواند براساس آنها، وقوع کلمه بعدی را در توالی پیش‌بینی کند.

بررسی دوپشته‌های دو حوزهٔ زبان‌شناسی و فلسفه و مقایسه آنها با دوپشته‌های پیکرهٔ

1. stop words 2. N-gram

مرجع، توالی‌هایی چون «زبان فارسی»، «ثبت واژگانی»، «فارسی زبان»، «قلب نحوی»، «زبان ترکی»، «گروه فعلی»، «گروه اسمی»، «نظام آوایی» و مانند آن را در زبان‌شناسی و «فلسفه اسلامی»، «حکمت متعالیه»، «هنر اسلامی»، «فلسفه اخلاق»، «اثبات وجود»، «عالم مثال»، «فلسفه غرب»، «اصل علیت» و مانند آن را در فلسفه نشان می‌دهد که همگی مرتبط با حوزه‌های پژوهشی مورد اشاره هستند.

اما بررسی سه و چهارپشته‌ها نتایج دیگری را نشان می‌دهد. بررسی داده‌ها و مقایسهٔ آنها با پیکرهٔ مرجع این پژوهش حاکی از آن است که خوشه‌های واژه‌ای سه و چهار عضوی، بیشتر چندپشته‌های پرکاربرد ژانر علمی هستند و نه مربوط به یک حوزهٔ خاص؛ یعنی فهرست‌گیری از این چندپشته‌ها، توالی‌های واژه‌ای را معرفی می‌کند که اصولاً در هریک از حوزه‌های علم پرکاربرد است و به‌نوعی ویژگی‌هایی از زبان علم و نوشتار علمی را آشکار می‌کند. توالی‌هایی چون «رابطهٔ معناداری وجود دارد»، «حاکی از آن است»، «تحقیق پیمایشی جامعه آماری»، «مورد بررسی قرار گرفت»، «بررسی عوامل مؤثر بر»، «تحقیق نشان می‌دهد که»، «هدف تحقیق بررسی عوامل» و مانند آن نمونه‌هایی از این نوع هستند.

۵ نتیجه‌گیری

کاوش در داده‌های طبیعیِ زبان ویژگی‌های پنهان و پیدایی از زبان را پیش روی هر پژوهشگر قرار می‌دهد که بدون بررسی داده‌ای و تنها با اتکا به مفروضات زبانی حاصل نمی‌شوند. بر همین اساس، امروزه پژوهشگران حوزهٔ زبان یا با پیش‌فرض‌های پژوهشی خود و به‌جهت اعتبار بخشیدن به نتایج پژوهشی به سراغ داده‌ها و پیکره‌های زبانی می‌روند یا اینکه بدون تعریف فرضیهٔ خاص به بررسی داده‌های زبانی می‌پردازند و از روابط پنهان درون‌داده‌ای به کشفیاتی نائل می‌شوند که دریچه‌های جدیدی را پیش روی آنها باز می‌کند. همهٔ اینها سبب شده است که امروزه بررسی و مطالعات زبانی بدون توجه و اتکا به داده‌های طبیعی زبان مقدور نباشد. این نوشتار به معرفی پیکرهٔ چکیده‌های علمی زبان فارسی و بیان ویژگی‌ها و مشخصات آن پرداخته است. این پیکره حدود سه‌ونیم میلیون موردواژه دارد و در ۱۲ عنوان کلی با زیربخش‌های موضوعی متعدد شامل روان‌شناسی و علوم تربیتی، مدیریت و حسابداری، حقوق و علوم سیاس، اقتصاد، علوم اجتماعی، زبان فارسی و متن پژوهی ادبی، حکمت و فلسفه، معارف، زبان‌شناسی، زبان و ادبیات عربی، مطالعات رسانه‌های نوین و پژوهش‌های بیمه‌ای دسته‌بندی شده است. داده‌های این پیکره

از چکیده‌های علمی مربوط به مقالات نشریات علمی و پایان‌نامه‌های دانشگاهی دانشگاه علامه طباطبائی و در حوزهٔ علوم انسانی تهیه شده است.

در ادامهٔ این پژوهش برای بررسی محتوای درونیِ پیکره و امکانات آن، بخش کوچکی از چکیده‌های پایان‌نامه‌ها از حیث مفهومِ کلیدی‌بودگی مورد بررسی قرار گرفت. سپس با توجه به بیست واژهٔ کلیدی اول در هر حوزه، تلاش شد بدون طرح هیچ پیش‌فرضی و تنها با توجه به فراوانی واژه‌های کلیدی نتیجه‌گیری‌هایی صورت بگیرد. کمترین نتیجه‌ای که از این بررسی به‌دست آمد آن بود که براساس مفهوم کلیدی‌بودگی می‌توان در موضوعات مختلف فهرستی از واژه‌های تخصصیِ آن حوزه به‌دست داد که این خود در فرهنگ‌نگاری و تهیهٔ نمایه‌های واژه‌ای اهمیت ویژه‌ای دارد. همچنین تلاش برای تحلیل محتوایی پیکره براساس بسامد واژه‌ای و فهرست واژه‌های به‌دست‌آمده نشان داد که امکان طرح برخی فرضیات براساس فهرست‌های واژه‌ای ممکن است و اصولاً فهرست واژه‌های کلیدی دارای اطلاعات درونی هستند که می‌توانند مورد توجه و بررسی باشند. همچنین در بخش موردپژوهی پیکره فهرست‌گیری از چندپشته‌های بخشی از پیکره انجام شد. بررسی چندپشته‌ها نشان داد که با فهرست‌گیری از این توالی‌ها می‌توان از یک سو به توالی‌های پرکاربرد و تخصصی هر حوزه دست پیدا کرد و از سوی دیگر با فهرست کردن سه و چهارپشته‌ها و به‌طور کلی چندپشته‌های با اعضای بالاتر و بررسی آنها به‌بیان ویژگی‌هایی از زبان علم و ارائهٔ تصویری از آن پرداخت.

در ادامه و برای آیندهٔ این پژوهش، ویرایش اسناد باقی‌مانده در دستور کار است. همچنین تکمیل داده‌های این حوزه با توجه به تنوع زیربخش‌های هر حوزه و ایجاد توازن در داده‌ها از دیگر برنامه‌های پیش روی این پژوهش است که از منابع علمی دیگر در علوم انسانی تأمین خواهد شد. در موضوع بررسی ویژگی‌های زبان علم بررسی توالی‌های پرکاربرد و ارائهٔ فهرستی از چندپشته‌های مشترک میان متون دانشگاهی در مقایسه با متون غیردانشگاهی پژوهش دیگری است که می‌تواند در ادامه انجام شود و در معرفی زبان علم و بیان ویژگی‌های محتوایی و دستوریِ آن راهگشا باشد.

۶ سپاسگزاری

مقالهٔ حاضر مستخرج از طرح پژوهشی با عنوان «ساخت پیکرهٔ چکیده‌های مقالات و پایان‌نامه‌های فارسی دانشگاه علامه طباطبائی» است و طبق قرارداد شمارهٔ ۸۹۴/د/ط مورخ

۱۳۹۸/۱/۱۹ با حمایت مالی و معنوی دانشگاه علامه طباطبائی انجام شده است. در انجام این طرح از همراهی و همکاری افراد متعددی بهره جستیم که از همه آنها سپاسگزاریم. قدردان معاونت پژوهشی دانشگاه علامه طباطبائی، کتابخانهٔ مرکزی دانشگاه، انتشارات دانشگاه به‌طور کلی و آقایان دکتر حمیدرضا علومی یزدی، دکتر علی خورسندی طاسکو، دکتر رضا ناظمیان، دکتر امیر زندمقدم، دکتر میرسعید موسوی رضوی، دکتر سید مهدی طاهری، آقای سیدمهدی سمیعی وفسی، خانم منیره قاسمی و خانم رکسانا شمسائی، به‌طور خاص هستیم که در تصویب طرح و در دریافت دادهٔ اولیهٔ پیکره ما را یاری کردند. همچنین از آقای دکتر سعید انواری سپاسگزار هستیم که در بخش «پژوهش موردی در پیکره» محتوای نوشتاری مربوط به فلسفه را از نظر گذراندند و پیشنهادهایی در خصوص آن ارائه کردند. همچنین از خانم میهن محقق‌زاده برای دریافت داده‌ها در مرحله اول کار سپاسگزاریم.

منابع

رضایی شریف‌آبادی، مرتضی (۱۴۰۱). *تولید پیکرهٔ علمی فارسی و فهرست بسامدی واژگان علمی*. پایان‌نامه دکتری، دانشگاه شیراز.

عاصی، مصطفی (۱۳۸٤). «پایگاه داده‌های زبان فارسی در اینترنت». *پژوهشگران* (پژوهشگاه علوم انسانی و مطالعات فرهنگی)، ش۲.

قیومی مسعود و مریم موسویان (۱۴۰۱). «کاربرد یادگیری ماشینی مبتنی بر شبکهٔ عصبی برای دسته‌بندی مستندات علمی». *پژوهشنامه پردازش و مدیریت اطلاعات*. ۳۷ (٤)، ۱۲۴۴–۱۲۱۷.

میرزائی، آزاده (۱۴۰۰). *فرهنگ توصیفی زبان‌شناسی پیکره‌ای*. تهران: انتشارات علمی.

AleAhmad, A., et al. (2009). "Hamshahri: A standard Persian text collection". *Knowledge-Based Systems*. 22(5), 382-387.

Berber Sardinha, A. P. (2004) Lingüística de Corpus. San Paulo: Manole.

Bijankhan, M., et al. (2011). "Lessons from building a Persian written corpus: Peykare". *Language Resources and Evaluation*. 45(2), 143-164.

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

Ghayoomi, M., & J. Kuhn (2014). "Converting an HPSG-based treebank into its parallel dependency-based treebank". *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland, 802-809.

Mirzaei, A. & A. Moloodi (2016). "Persian proposition Bank". *Proceedings of the 10th International Language Resources and Evaluation*. Portorož (Slovenia), 3828-3835.

Mirzaei, A., & P. Safari (2018). "Persian discourse treebank and coreference corpus". *Proceedings of the Eleventh International Conference on Language Resources and*

Evaluation (LREC 2018). URL https://www.aclweb.org/anthology/L18-1638.
Rasooli, M. S., & M. Kouhestani, & A. S. Moloodi (2013). "Development of a Persian Syntactic Dependency Treebank". *The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*. Alanta, USA.
Sabouri, S., et al. (2022). "naab: A ready-to-use plug-and-play corpus for Farsi". *arXiv preprint arXiv:2208.13486*.
Seraji, M. (2015). *Morphosyntactic Corpora and Tools for Persian. Doctoral dissertation, Uppsala University*. Studia Linguistica Upsaliensia 16.

استناد به این مقاله: میرزائی، آزاده و صدقی، فاطمه (۱۴۰۱). پیکره چکیده‌های مقالات و پایان‌نامه‌های دانشگاهی دانشگاه علامه طباطبائی. زبان و زبان‌شناسی ۱۸(۳۵)، ۱۲۷ـ۱۴۵. doi: 10.30465/lsi.2023.8502