Natural Language Text Corpus: Design, Construction and Management

Hamideh Asadi*
PhD in Library and Information Science-Information Retrieval; University of Tehran; Tehran, Iran Email: Asadi1366@gmail.com

Nader Naghshineh
PhD in Library and Information Science; Associate Professor in Library and Information Science; University of Tehran; Tehran, Iran; Email: nnaghsh@ut.ac.ir

Molouk Sadat Hosseini Beheshti
PhD in Linguistics; Associated Professor in Terminology & Ontology Research Group; Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran; Email: beheshti@irandoc.ac.ir

Iranian Journal of Information Processing and Management

Received: 28, Jun. 2023 **Accepted: 25, Nov. 2023**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 41 | No. 1 | pp. 71-98
Autumn 2025
https://doi.org/10.22034/jipm.2025.709151

Abstract: Considering the role of corpora in various fields of study and the need to construct a general corpus to increase efficiency and effectiveness in processes that require the extraction/use of natural language text, the purpose of this study is to focus on design and automatic construction of natural language text corpus and software for its management.

In this research, a technology-based method has been used to construct a monolingual corpus in Persian language. This corpus is produced automatically by collecting web data and its sources are news texts included in Persian language news agencies.

In the study, a corpus of natural language texts in Persian language was made. Due to the automaticity of the construction process, software is needed to manage it both in the construction stage and in the information extraction stage, which was designed, construct and implemented in this study.

The construction of general corpus of natural language texts is used for various research purposes, and the proposed method and the use of introduced tools in this study can facilitate the construction of corpus. Also, software design for corpus management will save time and cost of construction and will provide the possibility of extracting information from it.

Keywords: Context-Aware Recommender, Dissertation and Thesis Database, Selective Dissemination of Information

* Corresponding Author

پیکره متون زبان طبیعی
(طراحی، ساخت و مدیریت)

حمیده اسدی
دکتری علم اطلاعات و دانش‌شناسی؛ بازیابی اطلاعات؛ دانشگاه تهران؛ تهران، ایران؛
پدیدآور رابط Asadi1366@gmail.com

نادر نقشینه
دکتری علم اطلاعات و دانش‌شناسی؛ دانشیار؛ دانشگاه تهران، تهران، ایران nnaghsh@ut.ac.ir

ملوک‌السادات حسینی بهشتی
دکتری زبان‌شناسی همگانی؛ دانشیار؛ پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛ تهران، ایران؛
beheshti@irandoc.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۴۰۲/۰۴/۰۷ پذیرش: ۱۴۰۲/۰۹/۰۴ مقاله برای اصلاح به مدت ۴ روز نزد پدیدآوران بوده است.

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۴۱ | شمارهٔ ۱ | ص‌ص ۷۱-۹۸
پاییز ۱۴۰۴
https://doi.org/10.22034/jipm.2025.709151

چکیده: با توجه به نقش پیکره‌ها در حوزه‌های مطالعاتی گوناگون و لزوم ساخت یک پیکره عمومی برای افزایش کارایی و اثربخشی در پردازش‌هایی که مستلزم بهره‌جویی و استفاده از متن زبان طبیعی است، هدف این مطالعه تمرکز بر طراحی و ساخت خودکار پیکره متون زبان طبیعی و نرم‌افزاری برای مدیریت آن است.

در این پژوهش از روش مبتنی بر فناوری برای ساخت پیکره تک-زبانه و به زبان فارسی استفاده شده است. این پیکره به‌صورت خودکار و با گردآوری داده‌های وبی تولید شده و منابع آن را متون خبری مندرج در خبرگزاری‌های فارسی‌زبان تشکیل داده است.

در این مطالعه پیکره‌ای از متون زبان طبیعی به زبان فارسی ساخته شده است. با توجه به خودکار بودن فرایند ساخت پیکره، نرم‌افزاری برای مدیریت آن، چه در مرحله ساخت و چه در مرحله استخراج اطلاعات نیاز است که در این مطالعه طراحی، ساخته و پیاده‌سازی شده است.

ساخت پیکره‌ای عمومی از متون زبان طبیعی، برای اهداف پژوهشی گوناگون کاربرد دارد و روش پیشنهادی و استفاده از ابزارهای معرفی‌شده در این مطالعه می‌تواند ساخت پیکره را تسهیل کند. همچنین طراحی نرم‌افزاری برای مدیریت پیکره، صرفه‌جویی در زمان و هزینه ساخت را به همراه خواهد داشت و امکان استخراج اطلاعات از آن را فراهم خواهد آوُرد.

کلیدواژه‌ها: پیکره، دادگان، پردازش زبان طبیعی، زبان‌شناسی پیکره‌ای، هوش مصنوعی

۱. مقدمه

برای انجام وظایف مختلف پردازش زبان طبیعی، حجم زیادی از داده‌های متنی نیاز است که بتواند با استفاده از هوش مصنوعی تقلید درک و دریافت انسانی را بهبود بخشیده و استفاده از آن، فناوری را به توانایی انسانی نزدیک سازد (بحرانی و همکاران ۱۳۸۶؛ Verma and Khandelwal 2019). بدین‌منظور مهم‌ترین کار، تهیه و ساخت پیکره‌ای است که به‌صورت خودکار انجام می‌شود.

پیکره‌ها را مجموعه‌ای از داده‌های متنی سازمان‌یافته می‌دانند که می‌تواند حاوی متن، نقل قول، فهرست و حتی لغات باشند که برای اهداف گوناگون تهیه می‌شوند (دشتبانی، منصوری‌زاده و نصیری ۱۳۹۱).

در گذشته، پیکره‌های متنی سنتی برای پژوهش‌های زبان‌شناسی با استفاده از منابع چاپی مانند مقالات روزنامه‌ای و کتاب‌ها ساخته می‌شد. با رشد شبکه جهانی وب به‌عنوان یک منبع اطلاعاتی، استفاده از داده‌های آن برای وظایف گوناگون در پردازش زبان طبیعی افزایش یافت و مزایایی برای ساخت پیکره از داده‌های وبی نسبت به متن چاپی وجود دارد:

۱. داده‌های وبی به شکل الکترونیک و برای رایانه‌های قابل خواندن است؛ در حالی که همه داده‌های چاپی به شکل الکترونیک در دسترس نیست؛
۲. حجم داده‌های وبی زیاد است و برای آموزش داده تخمین بهتری را رقم می‌زند؛
۳. گردآوری داده‌های وبی با استفاده از موتورهای جست‌وجو انجام می‌شود و نیازی به دانلود گرانقیمت و پُرهزینه محاسباتی ندارد (Liu and James Curran 2006).

در واقع، ظهور رایانه/ فناوری‌های رایانه‌ای به خلق آنچه امروزه پیکره نامیده می‌شود، کمک فراوانی نمود و زبان‌شناسی پیکره‌ای به‌عنوان رویکردی عملی در پژوهش‌های زبان‌شناسی مورد توجه بسیاری از پژوهشگران قرار گرفته و سبب شده پیکره‌های گوناگونی با اهداف متنوع تهیه شوند؛ اما هدف مشترک آن‌ها مطالعات علمی بر روی زبان طبیعی است که گاه در یک حوزه خاص انجام می‌شود (دشتبانی، منصوری‌زاده و نصیری ۱۳۹۱؛ صفری ۱۳۹۴؛ Bennett 2010).

هرچند قابلیت انجام و کارایی مطالعات پیکره‌ای به روش و حجم پیکره وابسته بوده و بیشترِ این مطالعات نیز در سطح پیکره‌های کوچک طراحی شده و انجام گرفته است، اما پیکره‌هایی که در مقیاس بزرگ تهیه می‌شوند، افزون بر اینکه در پیشبرد پژوهش‌های زبان‌شناسی مؤثر هستند، برای توسعه نظام/ سامانه‌های پردازش زبان طبیعی هم کارایی

دارنـد و پیکره‌هـای متـنی بـا حجـم زیـاد اساساً بـرای آزمـون روش‌هـای گوناگـون نیـاز هسـتند (دشـتبانی، منصـوری‌زاده و نصیـری ۱۳۹۱؛ ذوالفقـار و همـکاران ۱۳۹۹؛ Sabeti et al. 2018).

امـروزه، به‌کارگیـری و توجـه بـه پیکره‌هـا و زبان‌شـناسی پیکـره‌ای بـرای تحلیـل منابـع در حوزه‌هـای مختلـف و به‌ویـژه مطالعـات زبان‌شـناختی، پژوهش‌هـای زبـانی را اعتبـار دیگـری بخشیـده اسـت و مزایـای اسـتفاده از پیکره‌هـا در تحلیل‌هـا و اسـتدلال‌های زبـانی همچـون بهره‌گیـری از حجـم زیـاد داده‌هـا، گـردآوری نظام‌منـد و صرفه‌جـویی در زمـان سـبب شـده کـه بـا کمـک آن‌هـا بتـوان اطلاعـات نهفتـه در متـون را شناسـایی و اسـتخراج کـرد و اهـداف پژوهـشی گوناگـونی را دنبـال نمـود (عـاصی و قنـدی ۱۳۹۴؛ میـرزایی و صفری ۱۳۹۴؛ صفری ۱۳۹۵؛ قدردوسـت نخچی و همـکاران ۱۳۹۵).

مبنـای تجـربی پیکره‌هـا، سـبب پدیـد آمـدن بنیـاد و بسـتری بـرای شناسـایی الگوهـای زبـانی و چگونـگی اسـتفاده از آن‌هـا فراهـم می‌آوَرد و بـررسی توزیـعی پیکره‌هـا نشـان می‌دهـد کـه ویژگی‌هـای آوایی، واژگانی، دسـتوری، گفتمـانی یـا کاربـرد شـناختی و معنـایی زبـان چگونـه اسـت (رضایی‌پنـاه و شـوکتی مقـرب ۱۳۹۵).

در مجمـوع، می‌تـوان گفـت کـه پیکره‌هـای زبـانی بـا هـدف اسـتفاده در یادگیـری ماشیـنی بـرای ترکیـب توانمندی‌هـای انسـانی و سـرعت پـردازش اطلاعـات ماشیـنی در راسـتای دسـتیابی بـه لایه‌هـای مختلـف زبـانی تهیـه می‌شـوند و بـا فراهـم آوردن امـکان اسـتخراج، پـردازش و دسـته‌بندی اطلاعـات، کاربردهـای گوناگـونی دارنـد. بـا ایـن حـال، می‌تـوان گفـت کـه هـدف نهـایی اسـتفاده از آن بـرای دسـتیابی بـه یـک مفهـوم اسـت؛ یعـنی از آشـکارترین لایـه زبـانی شـروع کـرده و ادامـه می‌دهـد تـا در انتهـا بـه معنـا و مفهـوم پنهـان در متـن دسـت یابـد. پیکره‌هـا به‌راسـتی نقطـه آزمـون نظام‌هـا و روش‌هـای ابهام‌زدایی و از ابزارهـای ضـروری ایـن حـوزه بـوده و ظهـور اینترنـت و پیشـرفت‌های رایانـه‌ای بـر توانایی‌هـای ایـن حـوزه افـزوده و نتیجـهٔ آن را نیـز بهبـود بخشیـده اسـت (دفتری‌نـژاد ۱۳۸۵؛ میـرزایی و مولـودی ۱۳۹۳؛ عـاصی و قنـدی ۱۳۹۴).

بـا ایـن حـال بایـد یـادآوری کـرد کـه گـردآوری داده و سـاخت پیکـره به‌تنهـایی چنـدان ارزشـمند نیسـت و موفقیـت اسـتفاده از آن، در بـه کارگیـری و پیشـرفت انـواع ابزارهـا و روش‌هـای پـردازش زبـان طبیـعی، یادگیـری ماشیـنی، یادگیـری آمـاری و یادگیـری عمیـق اسـت (کامیـابی گل و همـکاران ۱۳۹۷؛ Li et al. 2019).

۲. پیشینه پژوهش

در ایـن قسـمت، مطالعـات مرتبـط بـا پژوهـش حاضـر ارائـه می‌شـود. بـرای ارائـه پیشینـه پژوهـش

از الگوی پیشنهادی «نظری» (۱۳۹۲) استفاده شده است. بر اساس این الگو پیشینه پژوهش یا «نقشه پژوهش» با استفاده از دو دیدگاه «موضوعی» و «روش‌شناختی» که محصول دریافت‌های پژوهشگر از مبانی نظری موضوع هستند، تحلیل و دسته‌بندی می‌شوند. محصول مطالعهٔ پیشینه پژوهش با این رویکرد ترسیم گسست دانشی است که پژوهشگر بناست در پژوهش خود آن را پر نماید. بدین‌منظور مطالعات پیشین از دو دیدگاه موضوعی و روش‌شناختی تحلیل و ارائه می‌شوند.

الف) تحلیل پیشینه از دیدگاه موضوعی

پژوهش‌های بسیاری درباره پیکره‌ها انجام شده که به بررسی ساخت انواع پیکره‌های زبانی پرداخته یا با ساخت انواع پیکره‌های زبانی اهداف دیگری را مورد تحقیق و ارزیابی قرار داده‌اند. وجه اشتراک هر دو دسته پژوهش، توجه به معیارها و شاخص‌هایی برای ساخت پیکره بوده که آن‌ها را از دیدگاه‌های مختلف می‌توان بررسی کرد. از میان آن‌ها آنچه را که با موضوع این پژوهش مرتبط است، می‌توان در چهار گروه کلی دسته‌بندی کرد (نمودار ۱):

گروه اول پیکره‌هایی هستند که داده‌های خود را در اشکال متنی از جمله نوشتاری (ذوالفقار و همکاران ۱۳۹۹؛ افراشی، عاصی و جولایی ۱۳۹۴؛ میرزایی و مولودی ۱۳۹۳؛ دشتبانی، منصوری‌زاده و نصیری ۱۳۹۱؛ Pustejovsky, Bergler & Anick 1993) و الکترونیک (علایی ابوذر و همکاران، ۱۴۰۰؛ سلامی و همکاران، ۱۳۹۴؛ نظارات و موسوی میانگاه ۱۳۹۰؛ Mihalcea, Corley & Strapparava 2006؛ Sabeti et al. 2018؛ Sokolova & Bobicev 2018) انتخاب کرده‌اند.

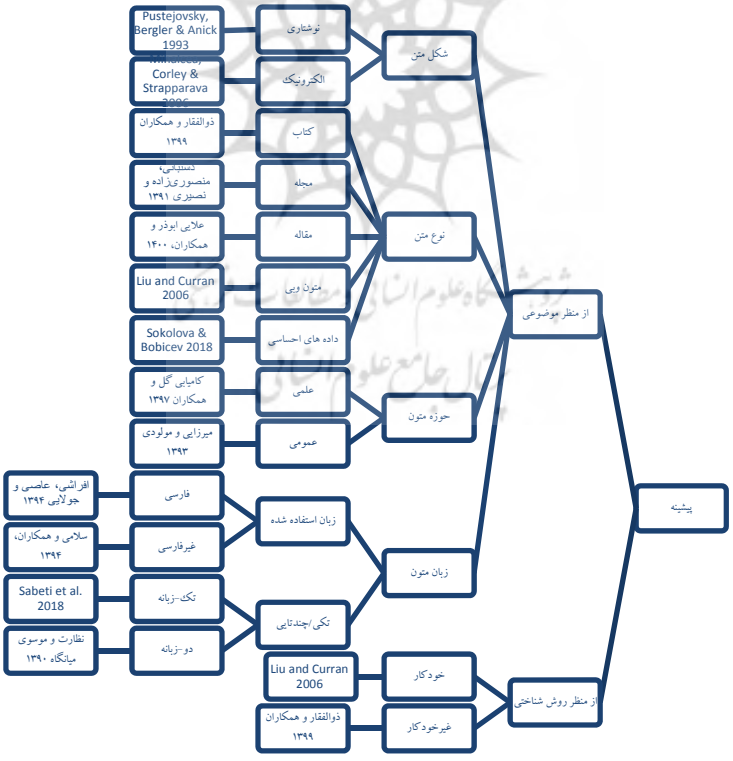
گروه دوم پیکره‌هایی هستند که بر اساس نوع متن می‌توانند به کتاب (ذوالفقار و همکاران ۱۳۹۹؛ میرزایی و مولودی ۱۳۹۳؛ دشتبانی، منصوری‌زاده و نصیری ۱۳۹۱؛ Pustejovsky, Bergler & Anick 1993)، مجله (دشتبانی، منصوری‌زاده و نصیری ۱۳۹۱)، مقاله (علایی ابوذر و همکاران ۱۴۰۰؛ کامیابی گل و همکاران ۱۳۹۷؛ سلامی و همکاران ۱۳۹۴؛ دشتبانی، منصوری‌زاده و نصیری ۱۳۹۱؛ Sabeti et al. 2018؛ Pustejovsky, Bergler & Anick 1993)، متون وبی (افراشی، عاصی و جولایی ۱۳۹۴؛ دشتبانی، منصوری‌زاده و نصیری ۱۳۹۱؛ نظارت و موسوی میانگاه ۱۳۹۰؛ Liu and Curran 2006) و داده‌های احساسی (Sokolova & Bobicev 2018) تقسیم‌بندی شوند.

گروه سوم پیکره‌هایی هستند که حوزه متون آن‌ها علمی (علایی ابوذر و همکاران

۱۴۰۰؛ ذوالفقار و همکاران ۱۳۹۹؛ کامیابی گل و همکاران ۱۳۹۷؛ افراشی، عاصی و جولایی ۱۳۹۴؛ سلامی و همکاران ۱۳۹۴؛ دشتبانی، منصوری‌زاده و نصیری ۱۳۹۱؛ Sabeti et al. 2018) یا عمومی (میرزایی و مولودی ۱۳۹۳؛ دشتبانی، منصوری‌زاده و نصیری ۱۳۹۱؛ نظارت و موسوی میانگاه ۱۳۹۰؛ Sokolova & Bobicev 2018؛ Liu and Curran 2006) است.

گروه چهارم پیکره‌هایی هستند که زبان متون آن، فارسی (علایی ابوذر و همکاران ۱۴۰۰؛ ذوالفقار و همکاران ۱۳۹۹؛ کامیابی گل و همکاران ۱۳۹۷؛ افراشی، عاصی و جولایی ۱۳۹۴؛ میرزایی و مولودی ۱۳۹۳؛ دشتبانی، منصوری‌زاده و نصیری ۱۳۹۱؛ Sabeti et al. 2018؛ Sokolova & Bobicev 2018) و غیرفارسی (سلامی و همکاران ۱۳۹۴) بوده و همچنین از لحاظ تک-زبانه (کامیابی گل و همکاران ۱۳۹۷؛ افراشی، عاصی و جولایی ۱۳۹۴؛ میرزایی و مولودی ۱۳۹۳؛ Sabeti et al. 2018؛ Sokolova & Bobicev 2018) و دو-زبانه بودن (نظارت و موسوی میانگاه ۱۳۹۰؛ Pustejovsky, Bergler & Anick 1993) قابل تفکیک هستند.

در ادامه، این پژوهش‌ها معرفی و دستاورهای آن‌ها ارائه می‌شود.



نمودار ۱. نقشه پژوهش حاضر

ب) تحلیل پیشینه از دیدگاه روش‌شناختی

در پژوهش‌های بررسی‌شده، شیوه ساخت پیکره‌ها به روشنی بیان نشده و از شواهد موجود در پژوهش چنین برمی‌آید که پیکره‌ها به‌صورت خودکار (Liu and Curran 2006) و یا غیرخودکار (ذوالفقار و همکاران ۱۳۹۹) تهیه شده‌اند.

به‌طور کلی، از تحلیل مطالعات پیشین چنین برمی‌آید که در پژوهش‌های گوناگون بسته به هدف آن، یک یا چند معیار در نظر گرفته شده و بدان اشاره شده است. اما در پژوهش حاضر و برای ساخت پیکره از متون خبری خبرگزاری‌های برخط فارسی‌زبان به زبان طبیعی بهره گرفته شده است. شایان توجه است که بر اساس این معیارها، قابلیت به‌روزرسانی پیکره وجود دارد و از نکات قابل تأمل آن به‌شمار می‌رود.

۳. روش پژوهش

هدف این پژوهش، پیشنهاد روشی خودکار برای تهیه یک پیکره زبانی از متون زبان طبیعی است. بر این اساس، از روش پژوهش مبتنی بر فناوری بهره‌گیری می‌شود. در این پژوهش‌ها، استفاده از تجهیزات فناورانه و به‌خدمت گرفتن کارکنانی که زمینه و سابقه فنی دارند، در اولویت است و بهره‌گیری از آن برای سهولت در انجام آن است. با این حال، مسئله‌ای که در این روش وجود دارد، منطبق ساختن روش علمی با فعالیت‌هایی است که شباهت بیشتری به تولید محصول دارند تا پژوهش بنیادی (پاول[1] ۱۹۹۷).

۳-۱. جامعه پژوهش و داده‌ها

این پیکره متنی، مجموعه‌ای از متون زبان طبیعی زبان فارسی است که از خبرگزاری‌ها جمع‌آوری می‌شود و مورد استفاده قرار می‌گیرد. این پیکره متنی، از نوع تک-زبانه و پویاست که قابلیت روزآمدسازی نیز دارد.

امروزه متون خبری در جامعه اهمیت بسیاری یافته و جایگاه برجسته آن نزد آحاد افراد جامعه و نیز مسئولان، مدیران، سیاست‌گذاران و ... قابل انکار نیست؛ چرا که:

- ◇ با توجه به سرعت و حجم بالای انتشار و میزان تأثیرگذاری آن، بررسی‌های جامع و دقیق از این منابع می‌تواند اطلاعات کاربردی از جامعه در اختیار قرار دهد (مظاهری و دل‌آرا ۱۳۹۸)؛

1. Ronald R. Powell

- در حوزه متن، خبرگزاری‌هایی که در تولید خبر نقش دارند، به‌طور معمول به سبک روزنامه‌ای خاص این کار را انجام داده و بازنشر می‌دهند و افزون بر اینکه نوع خاصی از اطلاعات به‌شمار می‌روند، از نظر موضوع، مضامین مرتبط، مالکیت، سرقت علمی و سایر فعالیت‌ها نیز قابل بررسی هستند (Kilgarriff and Grefenstette 2003)؛
- از آنجا که تنوع بیشتری در متنِ داده‌ها وجود دارد و متنِ داده از بازه زمانی مختلف انتخاب می‌شود، رفتار این اسناد به اطلاعات دنیای واقعی نزدیک‌تر است (شهشهانی و همکاران ۱۳۹۸)؛
- از سوی دیگر، امروزه خبرگزاری‌ها تنها مسئولیت نشر و اطلاع‌رسانی خبر را ندارند، بلکه هر خبرگزاریِ برخط بایستی انواع نیازهای کاربران خود را پاسخ دهد. از جمله آن‌ها می‌توان به دسته‌بندی و سازماندهی درست متون خبری، بازیابی متون خبری با استفاده از کلیدواژه‌های صحیح، سازماندهی و امکان دستیابی به سوابق خبری و ... اشاره کرد. این فعالیت‌ها افزون بر اینکه رضایت کاربران خبرگزاری‌ها را افزایش می‌دهد، باقی ماندن آن را در عرصه رقابت تضمین می‌کند (رباطی ۱۳۹۳).

لازمه بهره‌گیری از روش‌های یادگیری عمیق، حجم زیاد داده‌های متنی، زمان و منبع کافی برای آموزش و استخراج مدل است و هر قدر داده‌ها حجیم‌تر باشد، تخمین به‌دست‌آمده از مدل نیز بهتر خواهد بود. همچنین در بررسی‌های پیکره‌ای هر حوزه دانشی، توجه به انواع پایگاه‌های اطلاع‌رسانی و تولید دانش ضروری است و نیز تنوع مواد اطلاعاتی که محتوای پیکره را تشکیل می‌دهد، نیاز به ارزش‌گذاری و دفاع از صحت داده‌ها دارد (بحرانی و همکاران ۱۳۸۶؛ شهشهانی و همکاران ۱۳۹۸؛ روحانیان و همکاران ۱۳۹۹).

خبرگزاری‌ها با توجه به ویژگی‌هایی که پیشتر اشاره شد، به‌عنوان یکی از بسترهای اصلی تولید دانش زبان فارسی و به شکل زبان طبیعی به‌عنوان داده این پیکره انتخاب می‌شوند و می‌توانند در تحقق پژوهش‌های گوناگون بر اساس اهداف، روش و ... نقش بسیار مهمی ایفا کنند.

۳-۲. ساخت پیکره و نرم‌افزار مدیریت آن

پیکره‌ای که به‌صورت خودکار ساخته می‌شود، نیازمند نرم‌افزاری است که بتوان به کمک آن، پیکره را مدیریت کرد. به‌طور عام، نرم‌افزاری که برای مدیریت پیکره

طراحی می‌شود، هم در مرحله ساخت پیکره و هم در مرحله استخراج اطلاعات از پیکره کاربرد دارد و استفاده از پیکره را تسهیل می‌کند.

نرم‌افزار مدیریت پیکره موسوم به «پیکره‌نما»[1]، برای مدیریت پیکره نیز در دو مرحله ساخت و استخراج اطلاعات طراحی، ساخته و پیاده‌سازی شده است. هدف از طراحی این نرم‌افزار مدیریتی، ذخیره‌سازی، پردازش و جست‌وجوی اطلاعات با سرعتی قابل تحمل است که به‌سبب حجم زیاد اطلاعات گردآوری‌شده در رایانه‌های خانگی، بسیار وقتگیر خواهد بود.

۳-۲-۱. مشخصات فنی نرم‌افزار

این نرم‌افزار در بستر نرم‌افزاری «دات‌نت»[2] محصول شرکت «مایکروسافت»[3]، نسخه ۴/۵/۲ و با استفاده از زبان برنامه‌نویسی «سی»[4] و در محیط «استودیو ویژوال ۲۰۱۷»[5] پیاده‌سازی شده و در سیستم عامل «ویندوز»[6] قابل استفاده و اجراست. این نسخه از نرم‌افزار به‌صورت ۶۴ بیتی[7] همگردانی[8] شده و با توجه به وجود متن برنامه، این قابلیت را داراست که برای رایانه‌های ۳۲ بیتی هم بازطراحی شود. بدین‌ترتیب، نرم‌افزار تولیدشده قابلیت اجرا در بسیاری از رایانه‌های متداول امروزی را دارد. شایان ذکر است که با توجه به حجم داده‌ها، استفاده از نرم‌افزار روی رایانه‌هایی با حافظه کم و پردازشگر ضعیف با کندی اجتناب‌ناپذیری مواجه خواهد شد.

بانک اطلاعاتی پیکره، در بستر «اِس‌کیوِال‌لایت»[9] طراحی و پیاده‌سازی شده است. «اِس‌کیوِال‌لایت»، یک سامانه مدیریت پایگاه داده کم‌حجم و قابل جابجایی است که می‌تواند بانک‌های اطلاعاتی رابطه‌ای را تولید و مدیریت کند. ویژگی ممتاز آن این است که می‌تواند داده‌های حجیم با ساختار ساده را با کارایی و سرعت زیاد پردازش کند، قابلیت تلفیق با برنامه‌های اجرایی را دارد و برای استفاده از آن به نصب نرم‌افزار مستقلی نیاز نیست.

با تخمین حجم زیاد و ساختار ساده داده‌های پیکره متون زبان طبیعی، «اِس‌کیوِال‌لایت»

1. corpus viewer
2. .Net
3. Microsoft
4. C
5. Visual Studio 2017
6. Windows
7. bit
8. compile
9. SQL Lite

می‌توانـد بسـتر مناسـبی بـرای ذخیره‌سـازی و پـردازش داده‌هـای پیکـره باشـد و بـا توجـه بـه متـن بـاز بـودن بسـتر «اِس‌کیوال‌لایـت»، امـکان اسـتفاده یـا انتقـال داده‌هـا بـه هـر نرم‌افـزار دیگـر وجـود خواهـد داشـت.

۳-۲-۲. جست‌وجو و ذخیره داده‌ها

اولیـن مرحلـه سـاخت پیکـره، گـردآوری داده اسـت کـه از طریـق خـزش اینترنـتی و موتورهـای جسـت‌وجو انجـام می‌شـود. در ایـن روش، گـردآوری داده‌هـا بـر اسـاس موضـوع یـا پرسـش خـاص کاربـر صـورت می‌پذیـرد و مزیـت آن ایـن اسـت کـه قابـل گسـترش بـوده و می‌توانـد اطلاعـات جـاری و بـه‌روز را نیـز بازیـابی کنـد (Bennett 2010).

بـرای سـاخت پیکـره متـون زبـان طبیـعی، خبرگزاری‌هـای فارسی‌زبـان به‌سـبب دسـترس‌پذیری و خوانش‌پذیـری انتخـاب شـدند. ایـن خبرگزاری‌هـا دامنـه گسـترده‌ای از مطالـب و مقـالات خبـری را شـامل می‌شـوند کـه برگرفتـه از میلیون‌هـا صفحـه وب خبرگزاری‌هاسـت.

بـرای جسـت‌وجوی ایـن صفحـات خزشـگر/ ربـوت اینترنـتی «کوفکـس کاپـو»[1] انتخـاب شـد. «کوفکـس کاپـو»، یـک ربـوت نرم‌افـزاری هوشـمند و پلتفـرم/ پایـگاه یکپارچـه اسـت کـه بـا طـراحی یـک اسـتودیوی بصـری، داده‌هـا را بـه شـکل هوشـمند گـردآوری و یکپارچـه می‌کنـد (Kofax 2016, 2017).

پلتفـرم «کوفکـس کاپـو»، بـا روشـی سـریع‌تر و کارآمدتـر دسـترسی بـه داده‌هـای سـاختاریافته و سـاختارنیافتهٔ یـک برنامـه کاربـردی یـا منبـع داده مجـازی ماننـد پایـگاه داده‌هـا، نظـام داده‌ای و ایمیـل، وبگاه‌هـا، درگاه‌هـای سـامانه‌های نرم‌افـزاری، سـامانه‌های کسـب‌وکار، برنامه‌هـای رومیـزی و دیگـر منابـع داده‌ای را فراهـم می‌آوَرَد و بـا پشـتیبانی از انـواع برنامه‌هـای مبتنی‌بـر «ویندوز»، «جاوااسـکریپ»[2] و «آژاکـس»[3]، داده‌هـا را در قالـب «اکسـل»[4]، «ایکـس‌اِم‌اِل»[5]، «ایکـس‌اِل‌اِس»[6]، «پی‌دی‌اف»[7]، «آراِس‌اِس»[8] و «اِی‌پی‌آی»[9] و ... اسـتخراج می‌کنـد و داده‌هـای

1. Kofax Kapow
2. JavaScript
3. AJAX
4. Excel
5. XML
6. XLS
7. PDF
8. RSS
9. APIs

استخراج‌شده را بدون کد، ترکیب و یکپارچه می‌سازد (Kofax 2016, 2017).

البته، این امکان وجود داشت که داده‌ها را با کمک سرویس‌های «آراِس‌اِس» که بیشتر خبرگزاری‌ها ارائه می‌کنند نیز استخراج و ذخیره کرد، اما این ربوت خودکارسازی فرایندها، یک ظرفیت هوشمند دیجیتال است که در کنار نیروی انسانی، با غلبه بر مهم‌ترین چالش‌های اطلاعاتی یعنی پراکندگی اطلاعات در یک نظام‌ها و سامانه‌های اطلاعاتی، کارایی بیشتر و بهتری ایجاد می‌کند و با انجام تمام وظایف پردازش اطلاعات و در نتیجهٔ آن گردآوری و یکپارچه‌سازی داده‌ها، در چند ثانیه و به‌صورت خودکار در هزینه، زمان و تلاش صرفه‌جویی بسیار کرده و عملکرد و دسترس‌پذیری را بهینه می‌سازد (Kofax 2016, 2017).

با توجه به «یوآرال»های[1] معتبر، «کوفکس کاپو» محتوای صفحات را بررسی و محتوای آن را با استفاده از عملگرهایی پردازش می‌کند. مهم‌ترین مراحل فرایندِ خزش بدین‌شرح است:

۱. صفحات «اچ‌تی‌اِم‌اِل»[2] واکشی[3] شده و لینک ورودی به این صفحات برای خزش استخراج می‌شود؛
۲. محتوای صفحات بازیابی‌شده برای استخراج اطلاعات مورد نیاز تجزیه می‌شود؛
۳. در آخرین مرحله، حوزه‌های/ اطلاعات استخراج‌شدهٔ مرتبط با هر صفحه در یک پایگاه داده نمایه (ذخیره) شده و مجموعه جدیدی از «یوآرال» برای ادامه خزش انتخاب می‌شود.

در طی فرایندِ خزش این امکان وجود دارد که صفحات خاص با محتوای یکسان نیز بازیابی شوند. این به‌دلیل ابهام در دسته‌بندی‌ها و کلیدواژه‌های جست‌وجوست. برای حذف محتوای تکراری در طول فرایند، فیلترهایی اعمال می‌شود. در نهایت، تمام اطلاعات استخراج‌شده در قالب یک پایگاه داده ذخیره شده و برای اعمال پیش‌پردازش‌های گوناگون آماده می‌شود.

بازه زمانی برای گردآوری داده‌ها سال ۹۸-۱۳۹۷ است. انتهای بازه بر اساس حجم داده‌هایی است که بتوان با استفاده از رایانه‌های شخصی قابل پردازش باشد؛ وگرنه در

1. URL
2. HTML
3. fetching

صورت وجود رایانه‌هایی با قابلیت‌های بالاتر، امکان ذخیره حجم بیشتری از داده‌ها و به‌روزرسانی پیکره وجود دارد.

۳-۲-۳. ساختار پایگاه داده پیکره

ساختار پایگاه داده به‌گونه‌ای طراحی شده که در عین سادگی، امکان گزارش‌گیری سریع از داده‌ها را میسر سازد. ساختار ساده داده‌ها همچنین موجب می‌شود که اطلاعات موجود در پایگاه داده قابل استفاده در سایر پردازش‌ها و نرم‌افزارها نیز باشد.

پایگاه داده نرم‌افزار پیکره از سه جدول تشکیل شده است. این امکان وجود دارد که تمامی این جداول در یک پایگاه داده ذخیره شوند. اما با هدف افزایش سرعت پردازش و گزارش‌گیری، هر یک از جداول در قالب یک پایگاه داده و در یک فایل مستقل ذخیره‌سازی شده است.

جدول واژه‌ها: این جدول شامل واژه‌های به‌کاررفته در تمامی متون پیکره است. همچنین تعداد تکرار یا بسامد واژه‌ها نیز در این جدول ذخیره می‌شود.

جدول محتوا: این جدول شامل عناوین محتویات اسناد متنی پیکره، آدرس یکتای اینترنتی، تاریخ انتشار سند و متن محتوای سند است. این جدول بیشترین حجم داده پیکره زبانی را در خود جای داده است.

جدول ارجاعات: در این جدول، ارجاعات واژه‌ها به متن اسناد پیکره نگهداری می‌شود. به‌گفته ساده‌تر، در این جدول بیان می‌شود که یک واژه در چه اسنادی تکرار شده است.

مشخصات فیلدهای این پایگاه داده و جدول‌های آن در جدول ۱، آمده است.

جدول ۱. جداول پایگاه داده پیکره متون زبان طبیعی و مشخصات هر جدول

جداول	محتوا	فایل‌های کتابخانه‌ای/ داده‌ای	نام فیلد	کاربرد	نوع
واژه‌ها	واژه‌های پیکره زبانی	Words.db	Rowid	کلید اصلی	Integer
			Word	واژه	رشته یونی‌کد با طول متغیر حداکثر ۵۰ کاراکتر
			Ferq	تعداد کل تکرار واژه در متون	Integer

جداول	محتوا	فایل‌های کتابخانه‌ای/ داده‌ای	نام فیلد	کاربرد	نوع
محتوا	صفحات داده‌های متنی	Content.db	Rowid	کلید اصلی	Integer
			Title	عنوان سند	رشته یونی‌کد با طول متغیر نامحدود
			Link	آدرس یکتای اینترنتی	رشته یونی‌کد با طول متغیر نامحدود
			FaDate	تاریخ انتشار سند	رشته یونی‌کد با طول متغیر حداکثر ۲۰ کاراکتر
			Content	متن سند	رشته یونی‌کد با طول متغیر نامحدود
ارجاعات	ارجاع واژه‌ها به صفحات	Refs.db	Rowid	کلید اصلی	Integer
			WordID	شناسه واژه در جدول Words	Integer
			RefBlock	یک بلاک از ارجاعات	Blob

با توجه به تعداد زیاد ارجاعات و اسناد به‌کاررفته در این نرم‌افزار، از ساختار خاصی به نام فیلد رِف‌بلاک[1] برای ذخیره‌سازی ارجاعات استفاده شده است. این ساختار در ادامه شرح داده می‌شود:

فیلد رِف‌بلاک شامل بلوکی از داده‌هاست که ارجاعات به یک واژه در متن‌های مختلف را نشان می‌دهد. از آنجا که تعداد ارجاعات به یک واژه می‌تواند در مجموعه بزرگی از اسناد اتفاق بیفتد، در پیاده‌سازی این طرح از یک ساختار رشته‌ای برای ذخیره‌سازی ارجاعات استفاده شده است. استفاده از این ساختار دارای مزایای زیر است:

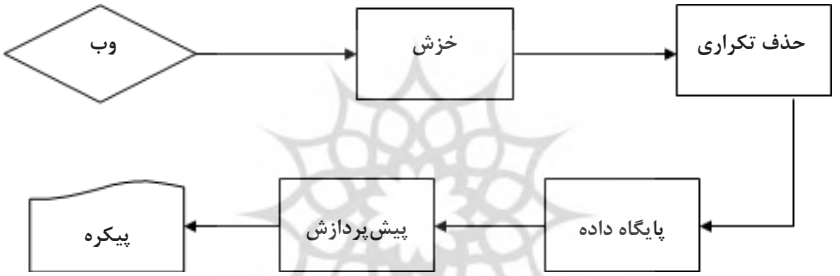
- ◇ تعداد رکوردهای پایگاه داده ارجاعات کمتر شده و سربار[2] حافظه مصرفی به ازای هر ارجاع به یک واژه کاهش می‌یابد؛
- ◇ با واکشی هر رکورد از پایگاه داده ارجاعات، تعداد بیشتری از ارجاعات قابل پردازش است. در طراحی پایگاه داده ارجاعات، در هر رِف‌بلاک تعداد ۱۰۰۰ ارجاع ذخیره می‌شود.

۳-۲-٤. پیش‌پردازش داده‌ها

برای اینکه اطلاعات استخراج‌شده قابلیت قرار گرفتن در پایگاه داده پیکره را پیدا کرده و

1. RefBlock
2. overload

برای پردازش اصلی آماده شود، با کمک نرم‌افزار مدیریت پیکره پیش‌پردازش می‌شود. از آنجا که داده‌های این پژوهش از وب‌سایت خبرگزاری‌ها گردآوری شده، ممکن است برخی اطلاعات غیرمرتبط را نیز در بر داشته باشد و چون برای پردازش، تنها متن خبر کارآمد است، در مرحله پیش‌پردازش این موارد شناسایی و حذف می‌شود تا جملات درست و با مفهوم برای پردازش اصلی تهیه و انتخاب گردد. همچنین با کمک الگوریتم و برخی توابع، سایر پیش‌پردازش‌ها از جمله نرمال‌سازی، توکن‌بندی، حذف علائم نگارشی و ایست‌واژگان[1] انجام می‌شود. بدین‌ترتیب و با کمک این نرم‌افزار، ساخت پیکره به‌صورت خودکار و زیر نظر متخصص انجام می‌شود و داده‌ها برای استفاده و تجزیه‌وتحلیل در پژوهش‌های گوناگون آماده می‌شود.



نمودار ۲. فرایند ساخت پیکره

از آنجا که پیکره به‌صورت خودکار تهیه و برای گردآوری داده‌های آن از خزشگر/ ربوت اینترنتی استفاده شده، و همچنین برخی متون این پیکره بسیار مفصل و برخی دیگر بسیار موجز و مختصر است، از این‌رو، تنها تعداد کل واژه‌ها/ تعداد واژه‌ها–سند به‌عنوان یک کل در نظر گرفته شده است.

پیش‌فرض الگوریتم تولید رِف‌بلاک و ساختارهای داده‌ای مورد نیاز این است که متن‌های مورد نظر از منابع استخراج و توسط هر تابع جداکننده[2] به رشته‌ای از کلمات تبدیل شده باشند. برای استخراج متن‌ها می‌توان از روش‌های مختلفی استفاده کرد.

الگوریتم‌های جداکننده متعددی در محیط‌های برنامه‌نویسی وجود دارند. در این طرح نوعی از این الگوریتم برای جداسازی واژه‌ها استفاده شده است.

1. stopwords
2. tokenizer

ساختار داده‌ای ایست‌واژگان شامل کلمات و افعال رابطه‌ای و توصیفی است که در تولید و پردازش پیکره زبانی نقشی ندارند؛ زیرا به‌دلیل استفاده متواتر در متون، پردازش و اعمال آن‌ها در پیکره زبانی موجب افزایش حجم بانک‌های اطلاعاتی و همچنین تأثیر بر آمار سایر واژه‌های زبانی می‌شود. به‌منظور کاهش بار پایگاه داده، این کلمات در پایگاه داده واژگان ثبت نمی‌شود و تنها در پایگاه اصلی محتوا و متن اسناد قابل مشاهده هستند. ایست‌واژگانی که در این طرح از بانک اطلاعاتی واژگان حذف شده‌اند، شامل موارد زیر است:

جدول ۲. فهرست ایست‌واژگان

آن	اگر	بودم	دارد	شدن	ما	می‌کنند	هستند
آمد	با	بودند	دارند	شده	من	می‌گردد	همه
آنرا	باشد	بگردد	داریم	شما	مورد	مگر	همین
آنها	باشند	تو	داشت	شود	می‌باشیم	نباید	هنوز
آنچه	باشیم	تا	داشتن	شوند	می‌کند	ندارد	و
آیا	باشید	تواند	داشتند	طی	می‌کنند	نشده	وی
از	باید	حتی	در	که	می‌داد	نشدم	یا
است	بایست	خواهد	در این	کرد	می‌دارد	نمی	پس
اکنون	بر	خواهند	دهد	کرده	می‌شود	نمودیم	چه
اما	برای	خواهیم	دهیم	کردند	می‌گردد	نمی‌شدم	چون
اند	بشود	خود	را	کرده‌اند	می‌برد	نمی‌شدم	چونانکه
او	به	خورد	رسد	کرده‌اند	می‌خواهد	نیز	چونکه
ای	بلکه	داد	رسید	کند	می‌داد	هم	گردد
این	بهتر	دادن	زیرا	کنند	می‌شود	هر	گردید
ایشان	بود	دادند	سپس	کنیم	می‌کردند	ها	گرفت
اینرا	بودن	داده	شد	لذا	می‌کنند	های	گشت
گفت	گیرد	گیرند					

در پایگاه داده پیکره، هر واژه‌ای دارای تعدادی ارجاع است که شماره سند و آدرس محل‌های استفاده از آن واژه را در سند دربرمی‌گیرد. ارجاعات به‌صورت بسته‌هایی از ۱۰۰۰

ارجاع در بانک اطلاعاتی ارجاعات و در فیلد رِف‌بلاک ذخیره شده‌اند. ساختار داده مورد استفاده مطابق با جدول زیر تعریف و استفاده شده است.

جدول ۳. ساختار داده واژگان پیکره

شماره سند (۳۲ بیت)	تعداد آدرس‌ها (۱۶ بیت)	آدرس ۱	آدرس ۲	...
شماره سند (۳۲ بیت)	تعداد آدرس‌ها (۱۶ بیت)	آدرس ۱	آدرس ۲	...
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

اگر تعداد ارجاعات یک واژه در یک رِف‌بلاک از ۱۰۰۰ آدرس بیشتر شود، یک رکورد رِف‌بلاک جدید برای آن واژه ایجاد خواهد شد. از آنجا که نوع داده‌ای رِف‌بلاک در پایگاه داده از نوع بلاب[1] (داده‌های بزرگ باینری) تعریف شده است، برای ذخیره‌سازی ارجاعات در این قالب لازم است که ارجاعات، مطابق ساختار فوق در یک نوع داده‌ای با نام «ممووری استریم»[2] ذخیره‌سازی شوند. بنابراین الگوریتم تبدیل باید به‌گونه‌ای عمل کند که ابتدا داده‌های مربوط به یک واژه را از یک سند استخراج کرده و آن را تبدیل به نوع داده‌ای «ممووری استریم» کند. در صورتی که تعداد ارجاعات از حد مشخصی بیشتر شد (مثلاً ۱۰۰۰ آدرس) «ممووری استریم» به دو شیء شکسته می‌شود و «ممووری استریم» جدید در یک رکورد جدید ذخیره می‌شود.

برای تعریف ارجاعات به یک واژه در یک سند، کلاسی با نام ارجاعات تعریف شده است.

«نیوآی‌دی»[3] شناسه سند و آدرس[4]، نشانی (آفست یا فاصله به تعداد کاراکتر)های واژه مورد نظر در آن سند است که در یک لیست ذخیره شده‌اند و برای مجموعه اسناد پردازش‌شده از کلاسی با نام جدول‌واژه[5] استفاده می‌شود.

1. Blob
2. Memory Stream
3. NewID
4. Address
5. WordTable

در ایـن کلاس فهرسـتی از واژه‌هـا وجـود دارد کـه بـه ازای هـر واژه مجموعـه‌ای از کلاس‌هـای ارجاعـات (شـماره سـند بـه همـراه آدرس‌هـای تکـرار واژه در سـند) تولیـد خواهـد شـد.

بـا فـرض اینکـه مجموعـه کلمـات یـک سـند توسـط الگوریتـم جداکننـده‌ای ماننـد «سـپریت تـرم»[1] تجزیـه شـده و به‌صـورت یـک لیسـت سـاده، مـثلاً آرایـه‌ای از کلمـات تی‌اِل[2] بـاز، گردانـده می‌شود.

دو تابـع مهـم دیگـر توابـع تبدیـل سـاختار داده‌ای از نـوع جـدول‌واژه بـه نـوع «ممـوری اسـتریم» اسـت. از آنجـا کـه نـوع داده‌هـا «ممـوری اسـتریم» به‌طـور مسـتقیم، قابـل ذخیره‌سـازی در فیلدهـایی از نـوع بلاب در بانـک اطلاعـاتی اسـت، ایـن توابـع می‌توانـد ارتبـاط کامـل بانـک اطلاعـاتی و الگوریتم‌هـای کدنویسـی بـه زبـان «سـی» را بـرای ارجاعـات پایـگاه داده پیکـره زبـانی برقـرار سـازد.

تابـع شمارشـگر «ممـوری اسـتریم» نیـز یـکی از توابـع مهـم و مفیـد کتابخانـه نرم‌افـزار پیکـره اسـت. بـه کمـک ایـن تابـع می‌تـوان تعـداد ارجاعـات موجـود در یـک شیء «ممـوری اسـتریم» را کـه بـر پایـه سـاختار رِف‌بلاک سـاخته شـده باشـد، مشـخص کـرد.

روش خوانـدن داده‌هـا از پایـگاه داده نرم‌افـزار پیکـره در قالـب الگوریتـمی اعمـال می‌شـود. فـرض بـر آن اسـت کـه داده‌هـا بـا توجـه بـه گزارش‌گیـری بـا زبـان «اِس‌کیـوال» در شیء «رِفزریـدر»[3] بـه برنامـه بازگردانـده شـده‌اند. ایـن شیء دربرگیرنـده تمـام رکوردهـای جـدول ارجاعـات مرتبـط بـا یـک واژه اسـت.

همچنیـن فـرض دیگـر آن اسـت کـه داده‌هـای واکـشی شـده از پایـگاه داده بایـد در شیء «رفزدیتـا»[4] از نـوع جـدول‌داده ذخیـره شـوند تـا قابـل پـردازش در هـر برنامـه بـه زبـان «سـی» باشـند.

بـه کمـک مجموعـه الگوریتم‌هـای شرح‌داده‌شـده می‌تـوان اطلاعـات پایـگاه داده پیکـره را توسـعه داد یـا داده‌هـای موجـود در آن را بازیـابی نمـود. بدیـهی اسـت کـه سـایر الگوریتم‌هـای مـورد نیـاز بـا توجـه بـه کاربـرد نرم‌افزارهـای اسـتفاده‌کننده از داده‌هـای پیکـره زبـانی طـراحی می‌گردنـد.

1. Separate Term
2. TL
3. Refs Reader
4. RefsData

۳-۲-۵. جست‌وجو و استخراج اطلاعات از پیکره

با توجه به تخمینی که در مورد حجم این منابع برآورد می‌شود، نیاز به ابزاری برای مدیریت اطلاعات آن پیکره وجود دارد که نخستین آن، طراحی یک موتور جست‌وجوست که بتواند مجموعه حجیم متون را در کل پایگاه داده، جست‌وجو و اطلاعات لازم را سریع‌تر و آسان‌تر از پیکره استخراج کند و استفاده از آن با کمک رایانه‌های معمول امکان‌پذیر باشد. نرم‌افزار طراحی‌شده ناظر پیکره، افزون بر جست‌وجوها، قابلیت‌های دیگری خواهد داشت که استفاده از پیکره را تسهیل کند. از جمله آن‌ها می‌توان به ارائه فهرست واژگان پیکره و بسامد آن‌ها، ارجاعات مربوط به صفحات منبع داده‌های متنی شامل آدرس اینترنتی و تاریخ انتشار سند و ... اشاره کرد.

۳-۲-۶. نحوه اجرا و فایل‌های مهم نرم‌افزار پیکره نما

نرم‌افزار از مسیر Corpus Viewer] bin Release] فایل Corpus Viewer.exe اجرا می‌شود. برای این منظور، تمامی فایل‌های کتابخانه‌ای که در این مسیر قرار دارند، مورد استفاده قرار می‌گیرند. بنابراین، وجود تمامی فایل‌های این مسیر برای اجرای صحیح نرم‌افزار ضروری است.

چنانکه پیشتر هم اشاره شد (نگاه کنید به جدول ۱)، داده‌های نرم‌افزار در فایل‌های زیر ذخیره شده‌اند:

۱. جدول واژگان؛
۲. جدول محتوا؛
۳. جدول ارجاعات.

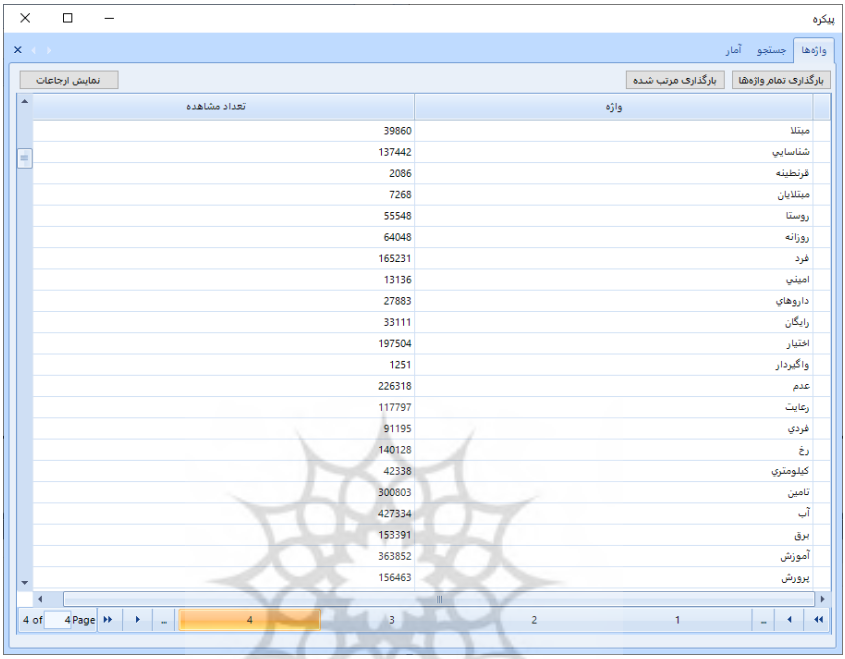
تمامی فایل‌های داده‌ای با ساختار پایگاه داده «اِس کیو اِل لایت»، نسخه ۳ سازگار هستند. برای بازخوانی یا هر گونه اعمال تغییر در داده‌های پایگاه‌های داده می‌توان از نرم‌افزارهای ویرایش پایگاه داده «اِس کیو اِل لایت» استفاده نمود. برای این پژوهش از نرم‌افزار حرفه‌ای «اِس کیو لایت»[1] نسخه ۵/۳ استفاده شده است.

۳-۲-۷. شرح محیط نرم‌افزار

در طراحی نرم‌افزار سعی شده که از پیچیدگی‌های کاربری اجتناب گردد و داده‌ها و

1. SQLite

جست‌وجو بر روی آن‌ها به‌راحتی در اختیار کاربر قرار گیرد. پنجره ورودی نرم‌افزار به شکل زیر است:



شکل ۱. پنجره ورودی نرم‌افزار

پنجره اصلی نرم‌افزار دارای سه برگه: «واژه‌ها»، «جست‌وجو» و «آمار» است.

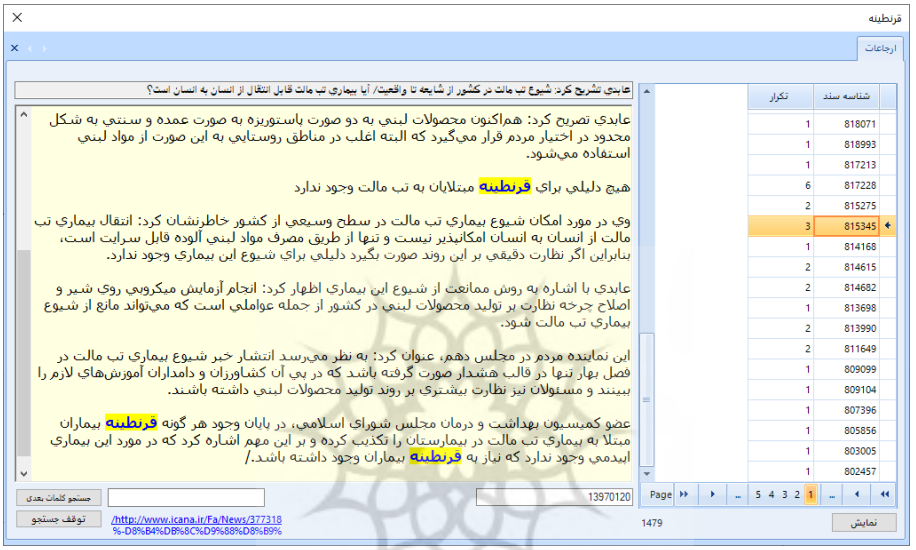
برگه واژه‌ها: در برگه واژه‌ها، جدولی از واژه‌های موجود در پایگاه داده پیکره نمایش داده می‌شود. این جدول دارای دو ستون است. ستون اول به خود واژه اختصاص دارد و ستون دوم بسامد تکرار واژه در مجموعه متن‌های پیکره نمایش داده شده است.

در هنگام اجرای اولیه نرم‌افزار، ۲۰۰۰ واژه اول پایگاه داده در قالب چهار صفحه ۵۰۰ واژه‌ای به نمایش درمی‌آید. علت محدودسازی تعداد در نمایش اولیه، افزایش سرعت بارگذاری داده‌ها و کاهش زمانی انتظار کاربر برای بارگذاری نرم‌افزار است؛ ضمن آنکه بارگذاری تمامی داده‌ها در حافظه بر روی رایانه‌هایی که حافظه «رَم»[1] کافی نداشته باشند، باعث صرفه‌جویی در بخش عمده‌ای از حافظه رایانه و کندی اجرای سایر نرم‌افزارها خواهد

1. RAM

شـد. کاربـران در صـورت نیـاز بـه بارگـذاری تمـامی داده‌هـا می‌تواننـد از کلیدهـای «بارگـذاری تمـام واژه‌هـا» و «بارگـذاری مرتب‌شـده» اسـتفاده نماینـد کـه اجـرای ایـن دسـتورات مدت‌زمـان قابـل توجـهی بـه طـول خواهـد انجامیـد.

بـا انتخـاب هـر واژه می‌تـوان از کلیـد «نمایـش ارجاعـات» اسـتفاده نمـود. در صـورت اسـتفاده از ایـن گزینـه، پنجـره‌ای از ارجاعـات واژه مـورد نظـر نمایـش داده می‌شـود:



شکل ۲. پنجره ارجاعات

اجزای پنجره ارجاعات

در پنجـره ارجاعـات، تمـامی متن‌هـایی کـه واژه مـورد بـررسی در آن‌هـا بـه‌کار رفتـه، نمایـش داده می‌شـود. در جـدول سـمت چـپ، فهرسـت اسـنادی کـه از واژه مـورد جسـت‌وجو در آن‌هـا اسـتفاده شـده، قـرار دارد. در مقابـل هـر سـند تعـداد تکـرار واژه مـورد جسـت‌وجو نمایـش داده شـده اسـت. در سـمت راسـت پنجـره، متـن سـند نمایـش داده می‌شـود. در متـن واژه مـورد جسـت‌وجو بـا رنـگ متفـاوت نمایـش داده شـده اسـت. همچنیـن آدرس اینترنـتی محـل اسـتخراج متـن سـند و تاریـخ انتشـار در زیـر متـن نمایـش داده شـده اسـت.

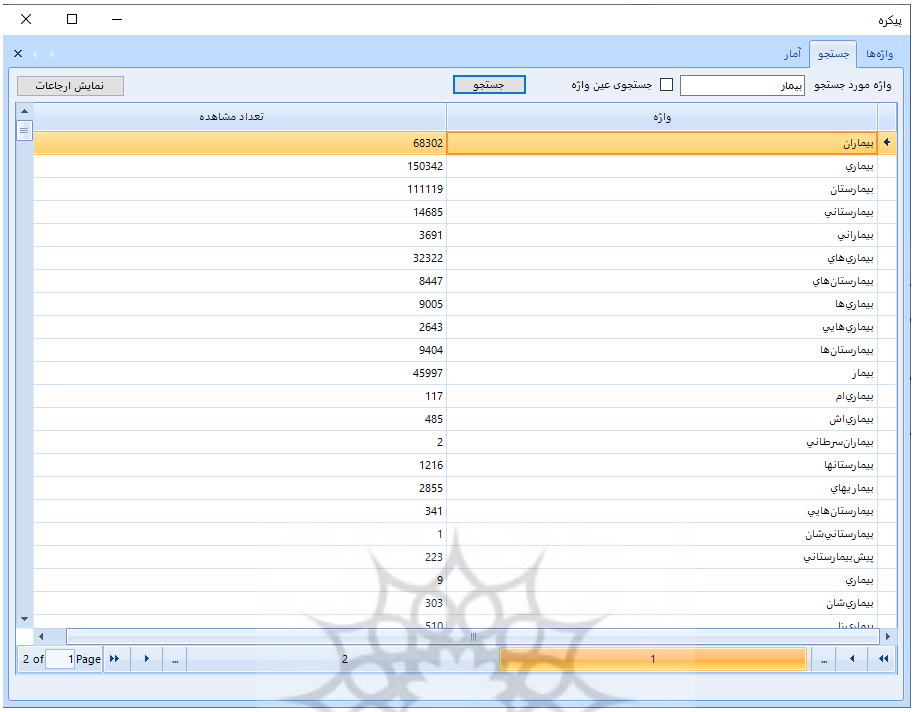
امـکان جسـت‌وجوی کلمـات بعـد از واژه اصـلی نیـز در ایـن پنجـره ایجـاد شـده اسـت. به‌عنـوان مثـال، بـا جسـت‌وجوی واژه «سـنتی» در ارجاعـات مربـوط بـه واژه «درمـان» نتایـج زیـر حاصـل می‌شـود:

شکل ۳. پنجره جست‌وجوی واژه در ارجاعات مربوط

در زیرِ جدولِ سمت چپ، آمار تعداد کلمات بعدی یافت‌شده به نسبت تعداد کل اسناد نمایش داده می‌شود. همچنین واژه دوم جست‌وجو در اسناد یافت‌شده به رنگ متفاوت نمایش داده می‌شود.

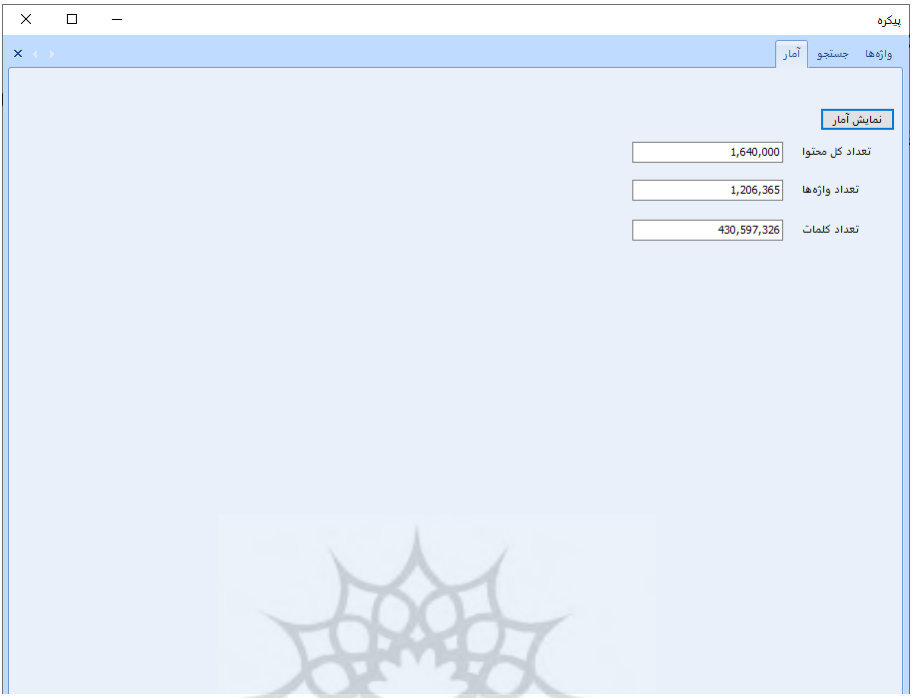
برگه جست‌وجو: در این برگه امکان جست‌وجوی یک واژه در کل پایگاه داده وجود دارد. پس از ورود واژه یا بخشی از واژه مورد نظر و زدن کلید جست‌وجو، فهرستی از رکوردهایی که شامل آن واژه باشند، نمایش داده می‌شود. در صورتی که گزینه «جست‌وجوی عین واژه» انتخاب شود، تنها رکورد مربوط به واژه مورد جست‌وجو نمایش داده می‌شود و واژه‌های مشابه نمایش داده نمی‌شوند.

در صورت انتخاب هر یک از واژه‌های پیداشده و فشار کلید «نمایش ارجاعات»، صفحه ارجاعات مربوط به واژه مورد جست‌وجو نمایش داده خواهد شد.



شکل ٤. پنجره ارجاعات مربوط به واژه مورد جست‌وجو

برگه آمار: در این برگه آمار واژه‌ها و ارجاعات پایگاه داده استخراج و نمایش داده می‌شود. با توجه به حجم قابل توجه پایگاه داده و تعداد رکوردهای ذخیره‌شده در آن، استخراج آمار، به‌خصوص در رایانه‌هایی با مشخصات سخت‌افزاری ضعیف، بسیار زمان‌بر و طولانی خواهد بود. تعداد کل اسناد محتوا، تعداد کل واژه‌ها و مجموع کلمات ذخیره‌شده در پایگاه داده مطابق با تصویر زیر است. قابل ذکر است که افعال و کلمات رابطه‌ای متون مورد استفاده در محاسبه این آمار پیکره لحاظ نشده است.



شکل ۵. پنجره آمار واژه‌ها و ارجاعات پایگاه داده

۴. تجزیه‌وتحلیل یافته‌ها

برخی از داده‌های آماری به‌دست‌آمده از پایگاه داده پیکره به‌شرح زیر است:

- تاریخ انتشار جدیدترین سند: ۱۳۹۷/۱۱/۰۶
- تاریخ انتشار قدیمی‌ترین سند: ۱۳۸۹/۱۲/۱۰
- بیشترین بسامد استفاده از یک واژه: واژه «ایران» با ۲۶۱۹۲۶۵ ارجاع
- تعداد کل محتواهای تمام‌متن در بانک اطلاعاتی: ۱۶۴۰۰۰۰ سند
- تعداد واژه‌های استفاده‌شده در متن اسناد: ۱۲۰۶۳۶۵ واژه
- تعداد کل کلمات استفاده‌شده در متن اسناد: ۴۳۰۵۹۷۳۲۶ کلمه

۵. بحث و نتیجه‌گیری

در این مقاله معرفی و شیوه ساخت پیکره‌ای خودکار به‌تفصیل مورد بحث قرار گرفت. پژوهش‌ها نشان داده که ساخت پیکره در مقیاس بزرگ ساده نیست و به زمان و هزینه

زیادی نیاز دارد. بنابراین، برای کارایی و اثربخشی بیشتر آن‌ها باید در مرحله ساخت پیکره به نکاتی توجه نمود و آن‌ها را در جریان ساخت پیکره‌ها لحاظ کرد. از جمله آن‌ها می‌توان به روش‌های خودکار تهیه پیکره اشاره نمود.

مطالعه حاضر نشان داد که در زمینه ساخت و تهیه پیکره می‌توان از ابزار و شیوه‌های گوناگون بهره گرفت که بیش از همه هدف پژوهش، ساخت و استفاده از آن را توجیه می‌کند. با این حال، بدیهی است که تکیهٔ صرف بر آن کافی نیست و بایستی در هر دوره زمانی و با توجه به کارکردها، از جنبه‌های گوناگون بررسی و به‌روزرسانی شود.

امروزه به‌کارگیری و توجه به پیکره‌ها و زبان‌شناسی پیکره‌ای برای تحلیل منابع در حوزه‌های مختلف و به‌ویژه مطالعات زبان‌شناختی، پژوهش‌های زبانی را اعتبار دیگری بخشیده است و مزایای استفاده از پیکره‌ها در تحلیل‌ها و استدلال‌های زبانی همچون بهره‌گیری از حجم زیاد داده‌ها، گردآوری نظام‌مند و صرفه‌جویی در زمان سبب شده که با کمک آن‌ها بتوان اطلاعات نهفته در متون را شناسایی و استخراج کرد و اهداف پژوهشی گوناگونی را دنبال نمود (عاصی و قندی ۱۳۹۴؛ میرزایی و صفری ۱۳۹۴؛ صفری ۱۳۹۵؛ قدردوست نخچی و همکاران ۱۳۹۵).

از آنجا که این پیکره حاوی مقالات و متون خبریِ خبرگزاری‌های فارسی زبان است، یک پیکره عمومی زبان طبیعی به شمار می‌رود و برای پردازش‌هایی که مستلزم بهره‌جویی و استفاده از متون زبان طبیعی است، مناسب و ارزشمند خواهد بود.

فهرست منابع

افراشی، آزیتا، مصطفی عاصی، و کامیار جولایی. ۱۳۹۴. استعاره‌های مفهومی در زبان فارسی؛ تحلیلی شناختی و پیکره‌مدار. *زبان‌شناخت* ۶ (۲): ۳۹-۶۱.

بحرانی، محمد، حسین صامتی، نازیلا حافظی، و سعیده ممتازی. ۱۳۸۶، اسفند ۱۹-۲۱. خوشه‌بندی خودکار کلمات بر اساس مقوله‌های نحوی برای سیستم‌های بازشناسی گفتار پیوسته فارسی. مقاله ارائه شده در *سیزدهمین کنفرانس ملی انجمن کامپیوتر ایران*. جزیره کیش، ایران

پاول، رونالد ار. ۱۹۹۷. *روش‌های اساسی پژوهش برای کتابداران*. مترجم: نجلا حریری ۱۳۸۹. [تهران]: آثار نفیس

دشتبانی، شکوفه، محرم منصوری‌زاده، و محمد نصیری. ۱۳۹۱. طراحی و ساخت پیکرهٔ متنی برای حوزه تخصصی فاوا. مقاله ارائه‌شده در *نخستین کنفرانس بین‌المللی پردازش خط و زبان فارسی*. سمنان

دفتری‌نژاد، الهه. ۱۳۸۵. فرایند مارکوف، الگوی احتمالاتی رفع ابهام در زبان‌شناسی رایانه‌ای. *علوم انسانی دانشگاه الزهرا (س)* ۱۶-۱۷ (۶۳-۶۴): ۱۰۷-۱۳۹.

ذوالفقار، زهره، طیبه موسوی میانگاه، بلقیس روشن، و امیررضا وکیلی‌فرد. ۱۳۹۹. بررسی تکنیک‌های بهبود عملکرد روش‌های بسامدشماری پیکره‌بنیاد در استخراج خودکار واژگان (مورد مطالعه: واژگان پایه علوم پزشکی). *پژوهشنامه پردازش و مدیریت اطلاعات* ۳۵ (۴): ۱۰۳۹-۱۰۶۴.

رباطی، زهرا. ۱۳۹۳. دسته‌بندی اخبار فارسی با استفاده از تکنیک‌های هوش مصنوعی. پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی شاهرود. [شاهرود].

رضایی‌پناه، امیر، و سمیه شوکتی مقرب. ۱۳۹۵. تحلیل پیکره‌بنیاد مدارهای هویت در سند استراتژی امنیت ملی ۲۰۱۵ بریتانیا. در *مجموعه مقالات دومین همایش ملی زبان‌شناسی پیکره‌ای*، ویراسته آزاده میرزایی، ۶۹-۹۱. تهران: نشر نویسه پارسی

روحانیان، مرتضی، مصطفی صالحی، علی درزی، و وحید رنجبر. ۱۳۹۹. تحلیل احساس در رسانه‌های اجتماعی فارسی با رویکرد شبکه عصبی پیچشی. *مهندسی برق و مهندسی کامپیوتر ایران* ۱۸ (۱): ۵۹-۶۶.

سلامی، مریم، زهرا سادات جلالی، مریم پاکدامن نائینی، و محمد علائی آرانی. ۱۳۹۴. تحلیل محتوای مقالات علوم پزشکی بر اساس مطالعه پیکره زبانی. *مدیریت اطلاعات سلامت* ۱۲ (۵): ۵۹۵-۶۰۷.

شهشهانی، مهسا، مهدی محسنی، آزاده شاکری، و هشام فیلی. ۱۳۹۸. پیکره برچسب خورده موجودیت‌های اسمی زبان فارسی. *پردازش علائم و داده‌ها* ۱۶ (۱): ۹۱-۱۰۹.

صفری، سعید. ۱۳۹۴. از زبان‌شناسی پیکره‌ای تا پیکره زبان‌آموز. در *مجموعه مقالات نخستین همایش ملی زبان‌شناسی پیکره‌ای*، ویراسته آزاده میرزایی، ۱۳۱-۱۵۲. تهران: نشر نویسه پارسی

_____. ۱۳۹۵. پیکره زبان‌آموز: مبانی، روش‌شناسی، الگوی طراحی و تولید. در *مجموعه مقالات دومین همایش ملی زبان‌شناسی پیکره‌ای*، ویراسته آزاده میرزایی، ۹۳-۱۲۳. تهران: نشر نویسه پارسی

عاصی، مصطفی، و سعیده قندی. ۱۳۹۴. پایگاه داده‌های زبان فارسی و پیکره تاریخی آن. در *مجموعه مقالات نخستین همایش ملی زبان‌شناسی پیکره‌ای*، ویراسته آزاده میرزایی، ۱۹۳-۲۲۰. تهران: نشر نویسه پارسی

علایی ابوذر، الهام، نصراله پاک‌نیت، علی‌اصغر حجت‌پناه، مجتبی زالی، و محمدهادی آقالویی آغمیونی. ۱۴۰۰. معرفی یک پیکره متنی تخصصی: پیکره پژوهشنامه. *پژوهش‌های زبان‌شناسی تطبیقی* ۱۱ (۲۲): ۲۷۱-۲۸۹.

قدردوست نخچی، سعیده، ندا پورمرتضی خامنه، پری‌ناز دادرس، و سلیمه زمانی. ۱۳۹۵. بررسی پیکره‌بنیاد مقوله قید. در *مجموعه مقالات دومین همایش ملی زبان‌شناسی پیکره‌ای*، ویراسته آزاده میرزایی، ۱۴۷-۱۶۵. تهران: نشر نویسه پارسی

کامیابی گل، عطیه، الهام اخلاقی باقوجری، احسان عسگریان، و هانیه حبیبی. ۱۳۹۷. استخراج اطلاعات از پیکره زبانی: معرفی پیکره مقاله‌های علمی پژوهشی دانشگاه فردوسی مشهد. *کتابداری و اطلاع‌رسانی* ۲۱ (۲): ۳-۲۵.

مظاهری، ویدا، و چنگیز دل‌آرا. ۱۳۹۸، مرداد. استخراج اطلاعات از وب‌سایت‌های خبری با استفاده از روش مبتنی بر آنتولوژی. مقاله ارائه‌شده در *هفتمین کنفرانس ملی علوم و مهندسی کامپیوتر وفناوری اطلاعات*. مازندران، ایران

میرزائی، آزاده، و پگاه صفری. ۱۳۹۴. ساختِ واژه- متن‌های تخصصی و عمومی زبان فارسی بر اساس بسامدگیری واژه‌های نقشی و محتوایی. در *مجموعه مقالات نخستین همایش ملی زبان‌شناسی پیکره‌ای*. ویراسته آزاده میرزایی، ۱۷۵-۱۹۱. تهران: نشر نویسه پارسی

میرزائی، آزاده، و امیرسعید مولودی. ۱۳۹۳. نخستین پیکره نقش‌های معنایی زبان فارسی. *علم زبان* ۲ (۳): ۲۹-۴۷.

نظارات، امین؛ طیبه موسوی میانگاه. ۱۳۹۰. طراحی و پیاده‌سازی یک سامانه بازیابی اطلاعات دو زبانه با استفاده از پیکره‌های زبانی. *پژوهشنامه پردازش و مدیریت اطلاعات*، ویژه‌نامه ذخیره، *بازیابی و مدیریت اطلاعات*: ۱۹۷-۲۱۲.

نظری، مریم. ۱۳۹۲. گسست دانشی در پژوهش‌های مولد چگونه رصد می‌شود؟ پیشنهاد ترسیم دو نقشه: نقشه دانش و نقشه پژوهش. *تحقیقات کتابداری و اطلاع‌رسانی دانشگاهی* ۴۷ (۱): ۲۷-۴۸.

References

Aasi, M., & S. Ghandi. 2015. Persian language databases and their historical corpora. In A. Mirzaei (ed.), Proceedings of the 1st. national conference on Croups linguistics (pp. 193–220). Tehran: Neviseh Parsi Publishing. [In Persian]

Afrashi, A., S. M. Asi, and K. Joulaei. 2016. Conceptual metaphors in Persian: A cognitive perspective and a corpus driven Analysis. *Language Studies* 6 (12): 39-61. [In Persian]

Alayiaboozar, E., N. Pakniat, A. Hojjatpanah, M. Zali, and M. Aghalouyaghmiyouni. 2021. introducing a specialized corpus: Pazhooheshname. *Comparative Linguistic Research* 11 (22): 271-289. [In Persian]

Bahrani, M., H. Sameti, N. Hafezi, & S. Momtazi. 2007. Automatic word clustering based on syntactic categories for continuous Persian speech recognition systems. In 13th Annual National Conference of the Iranian Computer Society. Kish Island, Iran. [In Persian]

Bennett, Gena R. 2010. *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers*. [Michigan]: University of Michigan Press.

Daftarinezhad, E. 2007. Markof Model: A Probability Model for Disambiguation in Computational Linguistics. Journal of Humanities 16-17 (63-64): 107-139. [In Persian]

Dashtbani, Sh., M. Mansoorizadeh, & M. Nassiri. 2012. Design and construction of a textual corpus for the ICT domain. Paper presented at the 1st International Conference on Persian Script and Language Processing, Semnan, Iran. [In Persian]

Ghadardoust Nakhchi, S., N. Pourmortazavi Khameneh, P. Dadras, & S. Zamani. 2016. A corpus-based study of the adverbial category. In A. Mirzaei (Ed.), Proceedings of the 2nd. national conference on corpus linguistics (pp. 147–165). Tehran: Neviseh Parsi Publishing. [In Persian]

Kamyabi Gol, A., E. Akhlaghi Baghujeri, E. Asgarian, and H. Habibi. 2018. Extracting information from language corpus: introducing the corpus of scientific articles of Ferdowsi University of Mashhad. *Library and Information Sciences* 21 (2): 3-25. [In Persian]

Kilgarriff, Adam, and Gregory Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics* 29 (3): 333-347.

Kofax. 2016. Kofax Kapow. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwid7NjGuJ__AhVgSfEDHeNbAGIQFnoECAwQAQ&url=https%3A%2F%2Fcobwebb.com%2Fwp-content%2Fuploads%2F2021%2F11%2Fds-kofax-kapow-en.pdf&usg=AOvVaw2aAIEADX7lGrhnmULWN85g (accessed March 7, 2023).

Kofax. 2017. Kofax Kapow. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjck7GEs5__AhV1RvEDHdJ9B6cQFnoECAwQAQ&url=https%3A%2F%2Fbpas.dk%2Fwp-content%2Fuploads%2F2018%2F02%2FKapow-datasheet.pdf&usg=AOvVaw0Oxy4hId6MjYY_a0D-MAGm (accessed March 7, 2023).

Li, Qin, Shaobo Li, Sen Zhang, Jie Hu, and Jianjun Hu. 2019. A Review of Text Corpus-Based Tourism Big Data Mining. *Applied Sciences* 9: 3300.

Liu, Vinci, and James R. Curran. 2006, April 3-7. Web Text Corpus for Natural Language Processing. Paper presented at *11th Conference of EACL: The European Chapter of the Association for Computational Linguistics*. Trento, Italy.

Mazaheri, V., & Ch. Delara. 2019. Information extraction from news websites using an ontology-based method. Paper presented at the 7th National Conference on Computer Science and Information Technology Engineering, Mazandaran, Iran. [In Persian]

Mihalcea, Rada, Courtney Corley, Carlo Strapparava. 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *AAAI'06: Proceeding of the 21st National Conference on Artificial Intelligence*, (Vol.1, P: 775-780). Boston, Massachusetts.

Mirzaei, A., & P. Safari. 2015. Lexical construction in specialized and general Persian texts based on the frequency of functional and content words. In A. Mirzaei (Ed.), Proceedings of the 1st. national conference on Croups linguistics (pp. 175–191). Tehran: Neviseh Parsi Publishing. [In Persian]

Mirzaei, A. and A. S. Moloodi. 2014. The First Semantic Role Corpus in Persian Language. *Language Science* 2 (3): 48-29. [In Persian]

Nazari, M. 2013. How Knowledge Gap Is Captured in Generative Research? A Proposal for Developing Two Maps: Knowledge Map and Research Map. *Academic Librarianship and Information Research*, 47 (1): 27-48. [In Persian]

Nezarat, A. and T. Mosavi Miangah. 2012. Designing and Implementing a Cross-Language Information Retrieval System Using Linguistic Corpora. Iranian Journal of Information Processing and Management 27 (2): 798-813. [In Persian]

Powell, Ronald R. 2010. *Basic Research Methods for Librarians*. (Hariri, N. translator). Tehran: Naafis Publications. [In Persian]

Pustejovsky, James, Sabine Bergler, Peter Anick. 1993. Lexical Semantic Techniques for Corpus. *Computational Linguistics* 19 (2): 331-358.

Rezaeipanah, A., & S. Shokati-Mogharab. 2016. A corpus-based analysis of identity circuits in the UK National Security Strategy 2015. In A. Mirzaei (Ed.), Proceedings of the 2nd. national conference on corpus linguistics (pp. 69–91). Tehran: Neviseh Parsi Publishing. [In Persian]

Robati, Zahra. 2014. Persian News Classification Using Artificial Intelligence. MA Thesis, Shahrood University of Technology. [Shahrood]. [In Persian]

Rohanian, M., M. Salehi, A. Darzi, Vahid Ranjbar. 2020. Convolutional Neural Networks for Sentiment Analysis in Persian Social Media. *Iranian Journal of Electrical and Computer Engineering* 8 (1): 59-66. [In Persian]

Sabeti, Behnam, Hossein Abedi Firouzjaee, Ali Janalizadeh Choobasti, S.H.E. Mortazavi Najafabadi, & Amir Vaheb. 2018. MirasText: An Automatically Generated Text Corpus for Persian. In *Proceedings of the Eleventh International Conferenece on Language Resources and Evaluation (LREC 2018)*, 1174-1177. Japan: European Language Resources Association (ELRA).

Safari, S. 2015. From corpus linguistics to learner corpus. In A. Mirzaei (ed.), Proceedings of the 1st. national conference on Croups linguistics (pp. 131–152). Tehran: Neviseh Parsi Publishing. [In Persian]

Safari, S. 2016. Learner corpus: Foundations, methodology, design and production model. In A. Mirzaei (ed.), Proceedings of the 2nd. national conference on corpus linguistics (pp. 93–123). Tehran: Neviseh Parsi Publishing. [In Persian]

Salami, M., Z. S. Jalali, M. Pakdaman Naeini, and M. Alaei Arani. 2015. Content Analysis of Medical Research Articles: A corpus-based study. *Health Information Management* 12 (5): 595-607. [In Persian]

Shahshahani M. S., M. Mohseni, A. Shakery, H. Faili. 2019. PAYMA: A Tagged Corpus of Persian Named Entities. *Journal of Signal and Data Processing* 16 (1): 91-110. [In Persian]

Sokolova, Marina, & Victoria Bobicev. 2018. Corpus Statistics in Text Classification of Online Data. *Arxiv*: 1803.06390.

Verma, Parul, and Brijesh Khandelwal. 2019. Word Embeddings and Its Application in Deep Learning. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* 8 (11): 337-341.

Zolfaghar, Z., T. Mosavi Miangah, B. Rovshan, and A. R. Vakilifard. 2020. A Study on the Improved Techniques of Corpus-based Frequency Approaches in Automatic Term Extraction (ATE) (The Case Study: Basic Medicine Vocabulary). *Iranian Journal of Information Processing and Management* 35 (4): 1039-1064. [In Persian]

حمیده اسدی

دکتری رشته علم اطلاعات و دانش‌شناسی با گرایش بازیابی اطلاعات از دانشگاه تهران است.

حوزه‌های روش‌شناسی پژوهش، بازیابی اطلاعات و علم‌سنجی از جمله علایق پژوهشی وی است.

نادر نقشینه

دارای مدرک تحصیلی دکتری در رشته علم اطلاعات و دانش‌شناسی از دانشگاه تهران است. ایشان هم‌اکنون دانشیار گروه علم اطلاعات و دانش‌شناسی دانشگاه تهران است.

حوزه‌های فناوری اطلاعات، داده کاوی، سیبرنتیک از جمله علایق پژوهشی وی است

ملوک‌السادات حسینی بهشتی

دارای مدرک تحصیلی دکتری در رشته زبان‌شناسی با گرایش همگانی از دانشگاه تهران است. ایشان دانشیار پژوهشکده علوم اطلاعات، گروه اصطلاح‌شناسی و هستان‌شناسی پژوهشگاه علوم و فناوری اطلاعات (ایرانداک) است.

مطالعه اصطلاح‌شناسی، مدیریت دانش، مدیریت اطلاعات، سازماندهی اطلاعات و پردازش زبان طبیعی از جمله علایق پژوهشی وی است.