

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

Received: 28, Jun. 2023

Accepted: 25, Nov. 2023

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

**Iranian Journal of
Information
Processing and
Management**

**Iranian Research Institute
for Information Science and Technology
(IranDoc)**

ISSN 2251-8223

eISSN 2251-8231

Indexed by SCOPUS, ISC, & LISTA

Vol. 41 | No. 1 | pp. ???-???

Autumn 2025

<https://doi.org/10.22034/jipm.2025.?????.????>



* Corresponding Author

پیکره متون زبان طبیعی (طراحی، ساخت و مدیریت)

حمیده اسدی

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

نادر نقشبند

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

ملوک‌السادات حسینی بهشتی

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



مقاله برای اصلاح به مدت ۴ روز نزد پدیدآوران بوده است.

پذیرش: ۱۴۰۲/۰۹/۰۴

دریافت: ۱۴۰۲/۰۴/۰۷

نشریه علمی | رتبه بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)

شاپا (جایی) ۲۲۵۱-۸۲۲۳

شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱

نمایه در SCOPUS، ISI، LISTA و
jipm.irandoc.ac.ir

دوره ۴۱ | شماره ۱ | صص ۷۵-۱۰۲

پاییز ۱۴۰۴

<https://doi.org/jipm.41.1>



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 فارسی زبان 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 معرفی شده 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 آورد.

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). بدین منظور مهم‌ترین کار، تهیه و ساخت پیکره‌ای است که به صورت 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 به عنوان 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 پرهزینه 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 به عنوان 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 بیشتر 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

دارند و پیکره‌های متنی با حجم زیاد اساساً برای آزمون روش‌های گوناگون نیاز هستند (دشتبانی، منصوری‌زاده و نصیری ۱۳۹۱؛ ذوالفقار و همکاران ۱۳۹۹؛ Sabeti et al. 2018). 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). مبنای تجربی پیکره‌ها، سبب پدید آمدن بنیاد و بستری برای شناسایی الگوهای زبانی و چگونگی استفاده از آن‌ها فراهم می‌آورد و بررسی توزیعی پیکره‌ها نشان می‌دهد که ویژگی‌های آوایی، واژگانی، دستوری، گفتمانی یا کاربرد شناختی و معنایی زبان چگونه است (رضایی‌پناه و شوکتی مقرب ۱۳۹۵).

در مجموع، می‌توان گفت که پیکره‌های زبانی با هدف استفاده در یادگیری ماشینی برای ترکیب توانمندی‌های انسانی و سرعت پردازش اطلاعات ماشینی در راستای دستیابی به لایه‌های مختلف زبانی تهیه می‌شوند و با فراهم آوردن امکان استخراج، پردازش و دسته‌بندی اطلاعات، کاربردهای گوناگونی دارند. با این حال، می‌توان گفت که هدف نهایی استفاده از آن برای دستیابی به یک مفهوم است؛ یعنی از آشکارترین لایه زبانی شروع کرده و ادامه می‌دهد تا در انتها به معنا و مفهوم پنهان در متن دست یابد. پیکره‌ها به‌راستی نقطه آزمون نظام‌ها و روش‌های ابهام‌زدایی و از ابزارهای ضروری این حوزه بوده و ظهور اینترنت و پیشرفت‌های رایانه‌ای بر توانایی‌های این حوزه افزوده و نتیجه آن را نیز بهبود بخشیده است (دفتری‌نژاد ۱۳۸۵؛ میرزایی و مولودی ۱۳۹۳؛ عاصی و قندی ۱۳۹۴).

با این حال باید یادآوری کرد که گردآوری داده و ساخت پیکره به‌تنهایی چندان ارزشمند نیست و موفقیت استفاده از آن، در به‌کارگیری و پیشرفت انواع ابزارها و روش‌های پردازش زبان طبیعی، یادگیری ماشینی، یادگیری آماری و یادگیری عمیق است (کامیابی گل و همکاران ۱۳۹۷؛ Li et al. 2019).

۲. پیشینه پژوهش

در این قسمت، مطالعات مرتبط با پژوهش حاضر ارائه می‌شود. برای ارائه پیشینه پژوهش

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 مطالعه پیشینه پژوهش با این رویکرد ترسیم گسست دانشی است که پژوهشگر بناست در پژوهش خود آن را پر نماید. بدین منظور 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 ابوزر و همکاران، ۱۴۰۰؛ سلامی و همکاران، ۱۳۹۴؛ نظارات و موسوی میانگاه ۱۳۹۰؛ Sokolova & Bobicev 2018؛ Sabeti et al. 2018؛ Mihalcea, Corley & Strapparava 2006) انتخاب کرده‌اند.

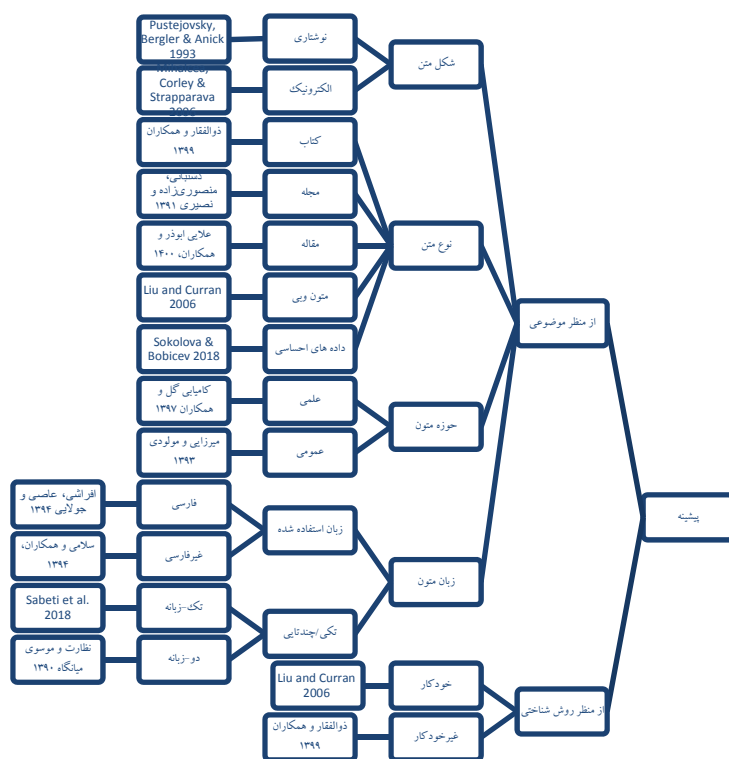
گروه دوم پیکره‌هایی هستند که بر اساس نوع متن می‌توانند به کتاب (ذوالفقار و همکاران ۱۳۹۹؛ میرزایی و مولودی ۱۳۹۳؛ دشتبانی، منصوری‌زاده و نصیری ۱۳۹۱؛ Pustejovsky, Bergler & Anick 1993)، مجله (دشتبانی، منصوری‌زاده و نصیری ۱۳۹۱)، مقاله (علایی ابوزر و همکاران ۱۴۰۰؛ کامیابی گل و همکاران ۱۳۹۷؛ سلامی و همکاران 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 ابوزر 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 دستاوردهای 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^۱ 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 تک‌زبان 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، سرعت علمی و سایر فعالیت‌ها نیز قابل بررسی هستند (Kilgariff and Grefenstette 2003)؛

◇ از آنجا که تنوع بیشتری در متن داده‌ها وجود دارد و متن 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 خبرگزاری 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 به‌دست آمده 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.

۲-۳. 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»، 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 به سبب حجم زیاد اطلاعات گردآوری شده در رایانه‌های خانگی، بسیار وقتگیر خواهد بود.

۳-۲-۱. مشخصات فنی نرم‌افزار

این نرم‌افزار در بستر نرم‌افزاری «دات‌نت»^۲ محصول شرکت «مایکروسافت»^۳، نسخه ۴/۵/۲ و با استفاده از زبان برنامه‌نویسی «سی»^۴ و در محیط «استودیو ویژوال ۲۰۱۷»^۵ پیاده‌سازی شده و در سیستم عامل «ویندوز»^۶ قابل استفاده و اجراست. این نسخه از نرم‌افزار به صورت ۶۴ بیتی^۷ همگردانی^۸ شده و با توجه به وجود متن برنامه، این قابلیت را داراست که برای رایانه‌های ۳۲ بیتی هم بازطراحی شود. بدین ترتیب، 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 «اس کیوال لایت»^۹ طراحی و پیاده‌سازی شده است. «اس کیوال لایت»، 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، «اس کیوال لایت»

1. corpus viewer

2. .Net

3. Microsoft

4. C

5. Visual Studio 2017

6. Windows

7. bit

8. compile

9. SQL Lite

می‌تواند بستر مناسبی برای ذخیره‌سازی و پردازش داده‌های پیکره باشد و با توجه به متن باز بودن بستر «اس کیوال لایت»، امکان استفاده یا انتقال داده‌ها به هر نرم‌افزار دیگر وجود خواهد داشت.

۳-۲-۲. جست‌وجو و ذخیره داده‌ها

اولین مرحله ساخت پیکره، گردآوری داده است که از طریق خزش اینترنتی و موتورهای جست‌وجو انجام می‌شود. در این روش، گردآوری داده‌ها بر اساس موضوع یا پرسش خاص کاربر صورت می‌پذیرد و مزیت آن این است که قابل گسترش بوده و می‌تواند اطلاعات جاری و به‌روز را نیز بازیابی کند (Bennett 2010).

برای ساخت پیکره متون زبان طبیعی، خبرگزاری‌های فارسی‌زبان به‌سبب دسترس‌پذیری و خوانش‌پذیری انتخاب شدند. این خبرگزاری‌ها دامنه گسترده‌ای از مطالب و مقالات خبری را شامل می‌شوند که برگرفته از میلیون‌ها صفحه وب خبرگزاری‌هاست. برای جست‌وجوی این صفحات خزشگر/ربوت اینترنتی «کوفکس کاپو»^۱ انتخاب شد. «کوفکس کاپو»، یک ربوت نرم‌افزاری هوشمند و پلتفرم/ پایگاه یکپارچه است که با طراحی یک استودیوی بصری، داده‌ها را به شکل هوشمند گردآوری و یکپارچه می‌کند (Kofax 2016, 2017).

پلتفرم «کوفکس کاپو»، با روشی سریع‌تر و کارآمدتر دسترسی به داده‌های ساختاریافته و ساختارنیافته یک برنامه کاربردی یا منبع داده مجازی مانند پایگاه داده‌ها، نظام داده‌ای و ایمیل، وبگاه‌ها، درگاه‌های سامانه‌های نرم‌افزاری، سامانه‌های کسب‌وکار، برنامه‌های رومیزی و دیگر منابع داده‌ای را فراهم می‌آورد و با پشتیبانی از انواع برنامه‌های مبتنی بر «ویندوز»، «جاوااسکریپت»^۲ و «آژاکس»^۳، داده‌ها را در قالب «اکسل»^۴، «ایکس‌ام‌ال»^۵، «ایکس‌ال‌اس»^۶، «پی‌دی‌اف»^۷، «آر‌اس‌اس»^۸ و «ای‌پی‌آی»^۹ و ... استخراج می‌کند و داده‌های

1. Kofax Kapow

2. JavaScript

3. AJAX

4. Excel

5. XML

6. XLS

7. PDF

8. RSS

9. APIs

استخراج شده را بدون کد، ترکیب و یکپارچه می‌سازد (Kofax 2016, 2017). البته، این امکان وجود داشت که داده‌ها را با کمک سرویس‌های «آراس‌اس» که بیشتر خبرگزاری‌ها ارائه می‌کنند نیز استخراج و ذخیره کرد، اما این ربات 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 نتیجه 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^۱ معتبر، «کوفکس کاپو» محتوای صفحات را بررسی و محتوای آن را با استفاده از عملگرهایی پردازش می‌کند. مهم‌ترین مراحل فرایند خزش بدین شرح است:

۱. صفحات «اچ‌تی‌ام‌ال»^۲ واکنشی^۳ شده و لینک ورودی به این صفحات برای خزش استخراج می‌شود؛

۲. محتوای صفحات بازبازی شده برای استخراج اطلاعات مورد نیاز تجزیه می‌شود؛

۳. در آخرین مرحله، حوزه‌های / اطلاعات استخراج شده مرتبط با هر صفحه در یک پایگاه داده نمایه (ذخیره) شده و مجموعه جدیدی از «یوآرال» برای ادامه خزش انتخاب می‌شود.

در طی فرایند خزش این امکان وجود دارد که صفحات خاص با محتوای یکسان نیز بازبازی شوند. این به دلیل ابهام در دسته‌بندی‌ها و کلیدواژه‌های جست‌وجوست. برای حذف محتوای تکراری در طول فرایند، فیلترهایی اعمال می‌شود. در نهایت، تمام اطلاعات استخراج شده 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 جداول‌های آن در جدول ۱، آمده است.

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

جدول	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 Title	کلید اصلی عنوان سند	Integer رشته یونی‌کد با طول متغیر نامحدود
			Link	آدرس یکنای 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 WordID	کلید اصلی شناسه واژه در جدول Words	Integer Integer
			RefBlock	یک بلاک از ارجاعات	Blob

با توجه به تعداد زیاد ارجاعات و اسناد به کاررفته در این نرم‌افزار، از ساختار خاصی به نام فیلد رِف بلاک^۱ برای ذخیره‌سازی ارجاعات استفاده شده است. این ساختار در ادامه شرح داده می‌شود:

فیلد رِف بلاک 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 سر بار^۲ 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 رِف بلاک تعداد ۱۰۰۰ ارجاع ذخیره می‌شود.

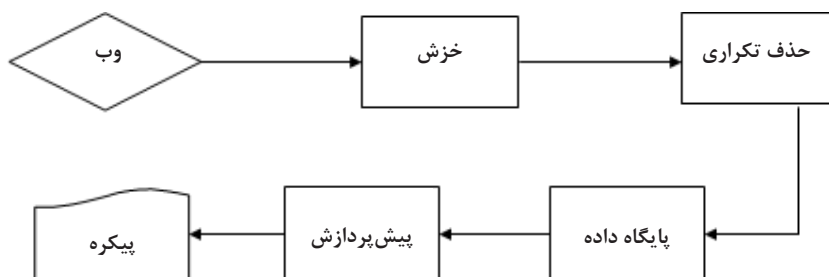
۳-۲-۴. پیش‌پردازش داده‌ها

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^۱ انجام می‌شود. بدین ترتیب و با کمک این نرم‌افزار، ساخت پیکره به‌صورت خودکار و زیر نظر متخصص انجام می‌شود و داده‌ها برای استفاده و تجزیه و تحلیل در پژوهش‌های گوناگون آماده می‌شود.



نمودار ۲. فرایند ساخت پیکره

از آنجا که پیکره به‌صورت خودکار تهیه و برای گردآوری داده‌های آن از خزشگر/ ربات 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 رِف بلاک و ساختارهای داده‌ای مورد نیاز این است که متن‌های مورد نظر از منابع استخراج و توسط هر تابع جداکننده^۲ 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 به دلیل 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. به منظور 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
آنها	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	بهرتر	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 به صورت بسته‌هایی از ۱۰۰۰

ارجاع در بانک اطلاعاتی ارجاعات و در فیلد رِف بلاک 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)	آدرس ۱	آدرس ۲	...
(بیت)				
شماره سند (۳۲ بیت)	تعداد آدرس‌ها (۱۶)	آدرس ۱	آدرس ۲	...
(بیت)				
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

اگر تعداد ارجاعات یک واژه در یک رِف بلاک از ۱۰۰۰ آدرس بیشتر شود، یک رکورد رِف بلاک جدید برای آن واژه ایجاد خواهد شد. از آنجا که نوع داده‌ای رِف بلاک در پایگاه داده از نوع بلاب^۱ (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 «مموری استریم»^۲ ذخیره‌سازی شوند. بنابراین الگوریتم تبدیل باید به گونه‌ای عمل کند که ابتدا داده‌های مربوط به یک واژه را از یک سند استخراج کرده و آن را تبدیل به نوع داده‌ای «مموری استریم» کند. در صورتی که تعداد ارجاعات از حد مشخصی بیشتر شد (مثلاً ۱۰۰۰ آدرس) «مموری استریم» به دو شیء شکسته می‌شود و «مموری 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»^۳ شناسه سند و آدرس^۴، 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 پردازش شده از کلاسی با نام جدول واژه^۵ استفاده می‌شود.

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

در این کلاس فهرستی از واژه‌ها وجود دارد که به ازای هر واژه مجموعه‌ای از کلاس‌های ارجاعات (شماره سند به همراه آدرس‌های تکرار واژه در سند) تولید خواهد شد.

با فرض اینکه مجموعه کلمات یک سند توسط الگوریتم جداکننده‌ای مانند «سپریت ترم»^۱ تجزیه شده و به صورت یک لیست ساده، مثلاً آرایه‌ای از کلمات تی‌ال^۲ باز، گردانده می‌شود.

دو تابع مهم دیگر توابع تبدیل ساختار داده‌ای از نوع جدول واژه به نوع «مموری استریم» است. از آنجا که نوع داده‌ها «مموری استریم» به طور مستقیم، قابل ذخیره‌سازی در فیلدهایی از نوع بلاب در بانک اطلاعاتی است، این توابع می‌تواند ارتباط کامل بانک اطلاعاتی و الگوریتم‌های کدنویسی به زبان «سی» را برای ارجاعات پایگاه داده پیکره زبانی برقرار سازد.

تابع شمارشگر «مموری استریم» نیز یکی از توابع مهم و مفید کتابخانه نرم‌افزار پیکره است. به کمک این تابع می‌توان تعداد ارجاعات موجود در یک شیء «مموری استریم» را که بر پایه ساختار رف بلاک ساخته شده باشد، مشخص کرد.

روش خواندن داده‌ها از پایگاه داده نرم‌افزار پیکره در قالب الگوریتمی اعمال می‌شود. فرض بر آن است که داده‌ها با توجه به گزارش‌گیری با زبان «اس کیوال»^۳ در شیء «رفزریدر»^۴ به برنامه بازگردانده شده‌اند. این شیء دربرگیرنده تمام رکوردهای جدول ارجاعات مرتبط با یک واژه است.

همچنین فرض دیگر آن است که داده‌های واکشی شده از پایگاه داده باید در شیء «رفزدیتا»^۴ از نوع جدول داده ذخیره شوند تا قابل پردازش در هر برنامه به زبان «سی» باشند. به کمک مجموعه الگوریتم‌های شرح داده‌شده می‌توان اطلاعات پایگاه داده پیکره را توسعه داد یا داده‌های موجود در آن را بازیابی نمود. بدیهی است که سایر الگوریتم‌های مورد نیاز با توجه به کاربرد نرم‌افزارهای استفاده‌کننده از داده‌های پیکره زبانی طراحی می‌گردند.

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 طراحی شده 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 «اس کیوال لایت»، نسخه ۳ سازگار هستند. برای بازخوانی یا هر گونه اعمال تغییر در داده‌های پایگاه‌های داده می‌توان از نرم‌افزارهای ویرایش پایگاه داده «اس کیوال لایت» استفاده نمود. برای این پژوهش از نرم‌افزار حرفه‌ای «اس کیوال لایت»^۱ 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 به راحتی در اختیار کاربر قرار گیرد. پنجره ورودی نرم افزار به شکل زیر است:

واژه	تعداد مشاهده	پیکره
مبتلا	39860	
شناسایی	137442	
فرنیطینه	2086	
مبتلایان	7268	
روستا	55548	
روزانه	64048	
فرد	165231	
امیدی	13136	
داروهای	27883	
رایگان	33111	
اختیار	197504	
واگردار	1251	
عدم	226318	
رعایت	117797	
فردی	91195	
رغ	140128	
کیلوستری	42338	
تأمین	300803	
آب	427334	
بوق	153391	
آموزش	363852	
پرورش	156463	

شکل ۱. پنجره ورودی نرم افزار

پنجره اصلی نرم افزار 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 نرم افزار، 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 نرم افزار است؛ ضمن آنکه بارگذاری تمامی داده‌ها در حافظه بر روی رایانه‌هایی که حافظه «رم»^۱ کافی نداشته باشند، باعث صرفه جویی در بخش عمده‌ای از حافظه رایانه و کندی اجرای سایر نرم افزارها خواهد

1. RAM

شد. کاربران در صورت نیاز به بارگذاری تمامی داده‌ها می‌توانند از کلیدهای «بارگذاری تمام واژه‌ها» و «بارگذاری مرتب‌شده» استفاده نمایند که اجرای این دستورات مدت زمان قابل توجهی به طول خواهد انجامید.

با انتخاب هر واژه می‌توان از کلید «نمایش ارجاعات» استفاده نمود. در صورت استفاده از این گزینه، پنجره‌ای از ارجاعات واژه مورد نظر نمایش داده می‌شود:

فهرست ارجاعات

شماره سند	تکرار
818071	1
818993	1
817213	1
817228	6
815275	2
815345	3
814168	1
814615	2
814682	2
813698	1
813990	2
811649	2
809099	1
809104	1
807396	1
805856	1
803005	1
802457	1

عبارتی شرح کرد شیوع تب مالت در کشور از شایعه تا واقعیت/ آیا بیماری تب مالت قابل انتقال از انسان به انسان است؟

عایدی تصریح کرد: هم‌اکنون محصولات لبنی به دو صورت پاستوریزه به صورت عمده و سبکی به شکل محدود در اختیار مردم قرار می‌گیرد که البته اغلب در مناطق روستایی به این صورت از مواد لبنی استفاده می‌شود.

هیچ دلیلی برای **قرنطینه** مبتلایان به تب مالت وجود ندارد

وی در مورد امکان شیوع بیماری تب مالت در سطح وسیعی از کشور خاطرنشان کرد: انتقال بیماری تب مالت از انسان به انسان امکانپذیر نیست و تنها از طریق مصرف مواد لبنی آلوده قابل سرایت است، بنابراین اگر نظارت دقیقی بر این روند صورت بگیرد دلیلی برای شیوع این بیماری وجود ندارد.

عایدی با اشاره به روش ممانعت از شیوع این بیماری اظهار کرد: انجام آزمایش میکروبی روی شیر و اصلاح چرخه نظارت بر تولید محصولات لبنی در کشور از جمله عواملی است که می‌تواند مانع از شیوع بیماری تب مالت شود.

این نماینده مردم در مجلس دهم، عنوان کرد: به نظر می‌رسد انتشار خبر شیوع بیماری تب مالت در فصل بهار تنها در قالب هشدار صورت گرفته باشد که در پی آن کشاورزان و دامداران آموزش‌های لازم را ببینند و مسئولان نیز نظارت بیشتری بر روند تولید محصولات لبنی داشته باشند.

عضو کمیسیون بهداشت و درمان مجلس شورای اسلامی، در پایان وجود هر گونه **قرنطینه** بیماران مبتلا به بیماری تب مالت در بیمارستان را تکذیب کرده و بر این مهم اشاره کرد که در مورد این بیماری ابتدیی وجود ندارد که نیاز به **قرنطینه** بیماران وجود داشته باشد.

13970120

Page 5 4 3 2 1

1479

نمایش

مستجو کلمات بعدی

توقف جستجو

<http://www.icana.ir/Fa/News/377318-%D8%B4%D8%8C%D9%88%D8%B9%>

شکل ۲. پنجره ارجاعات

اجزای پنجره ارجاعات

در پنجره ارجاعات، تمامی متن‌هایی که واژه مورد بررسی در آن‌ها به کار رفته، نمایش داده می‌شود. در جدول سمت چپ، فهرست اسنادی که از واژه مورد جست‌وجو در آن‌ها استفاده شده، قرار دارد. در مقابل هر سند تعداد تکرار واژه مورد جست‌وجو نمایش داده شده است. در سمت راست پنجره، متن سند نمایش داده می‌شود. در متن واژه مورد جست‌وجو با رنگ متفاوت نمایش داده شده است. همچنین آدرس اینترنتی محل استخراج متن سند و تاریخ انتشار در زیر متن نمایش داده شده است.

امکان جست‌وجوی کلمات بعد از واژه اصلی نیز در این پنجره ایجاد شده است. به‌عنوان مثال، با جست‌وجوی واژه «سنتی» در ارجاعات مربوط به واژه «درمان» نتایج زیر حاصل می‌شود:

بهری فروشگاه اینترنتی تجهیزات پزشکی بازار طب

درومان خود داستانی باستاند.

مشکلات **درومان** بستگی زخم:

با توجه به اطلاعات گسب شده از مراجعین در حوزه زخم، اکثر **درومان**های مربوط به زخم بستر و زخم دیابتی یا روشهای **بستگی** و قدیمی در منازل توسط افراد غیر متخصص انجام شده که مناسباته باعث طی شدن عدم **درومان** صحیح می شود و معمولاً بیمار را در مراحل تشدید بیماری قرار داده و باید به مراکز تخصصی و نوین **درومان** زخم مراجعه نمایند. این موضوع به تنهایی باعث کاهش احتمال بهبودی و بالا رفتن هزینه های **درومان**ی آن ها می شود. همچنین لازم به ذکر است مشاورین صورت و محرب شرکت آماده خدمت رسانی و ارائه مشاوره علمی به تمام بیماران می باشد.

بهبود زخم:

همانطور که می دانید روند بهبود زخم به فعالیت های پیچیده ای بستگی دارد. تمام تلاش فروشگاه اینترنتی بازار طب فراهم نمودن خدمت رسانی آسان بوده، تا بیماران تمام لوازم، تجهیزات و بانسمان زخم مورد نیاز خود را به آسانی تهیه نمایند. همچنین شما می توانید با تماس تلفنی و یا مراجعه حضوری به بازار طب تمام سوالات خود را در مورد **درومان** زخم دیابتی، زخم پای دیابتی و **درومان** زخم بستر که شایع ترین نوع زخم های مرم برای بیماران دیابتی هستند درمیان گذاشته و روند بهبود علمی به شما توضیح داده خواهد شد.

نیم فروش شرکت بازار طب نوتیکا تمام امور مثل استرس بیمار، مرحله زخم، محصولات مورد نیاز برای کمک به بهبود زخم و اسودگی خاطر بیمار را در مد نظر قرار داده و به عنوان یک نماینده معتبر در ایران اقدام به خدمت رسانی به بیماران می کند.

13970921

صفحه 34 از 69648

https://www.alef.ir/news/3970921150

شکل ۳. پنجره جست‌وجوی واژه در ارجاعات مربوط

در زیر جدول 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.

پیکره

واژه‌ها جستجو آمار

واژه مورد جستجو بیمار جستجوی عین واژه جستجو

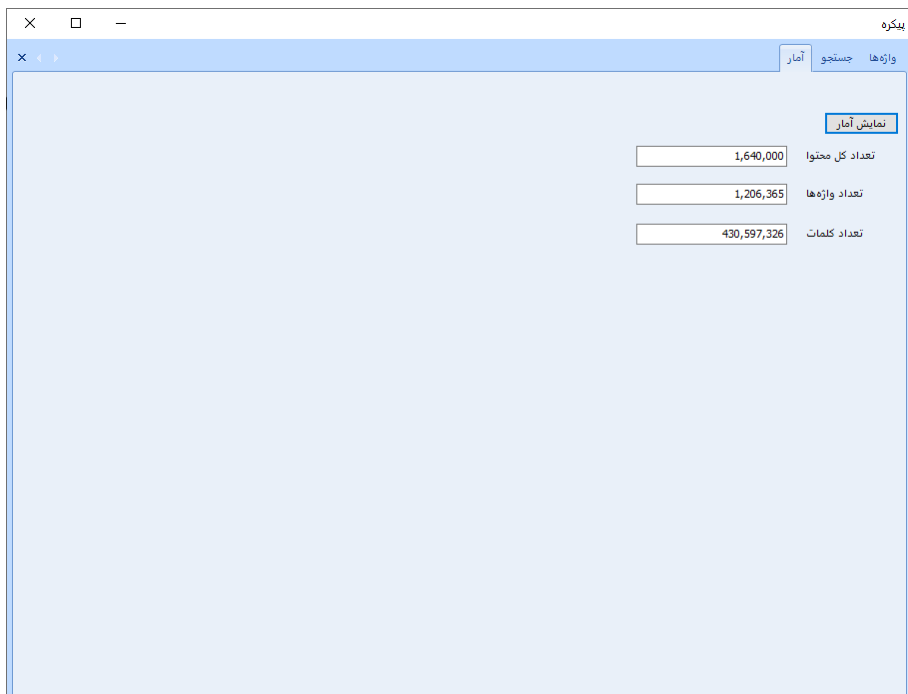
نمایش ارجاعات

تعداد مشاهده	واژه
68302	بیماران
150342	بیماری
111119	بیمارستان
14685	بیمارستانی
3691	بیمارانی
32322	بیماری‌های
8447	بیمارستان‌های
9005	بیماری‌ها
2643	بیماری‌هایی
9404	بیمارستان‌ها
45997	بیمار
117	بیماری‌ام
485	بیماری‌اش
2	بیماران سرطانی
1216	بیمارستانها
2855	بیمار بهائی
341	بیمارستان‌هایی
1	بیمارستانی‌شان
223	پیش بیمارستانی
9	بیماری
303	بیماری‌شان
510	بیمارستان

2 of 1 Page 2 1

شکل ۴: پنجره ارجاعات مربوط به واژه مورد جستجو

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. تجزیه و تحلیل یافته‌ها

برخی از داده‌های آماری به‌دست آمده از پایگاه داده پیکره به شرح 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 به تفصیل 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 تکیه صرف بر آن کافی نیست و بایستی در هر دوره زمانی و با توجه به کارکردها، از جنبه‌های گوناگون بررسی و به‌روزرسانی شود.

امروزه به کارگیری 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 خبری 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، آریتا، 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 پیکره متنی برای حوزه تخصصی فاوا. مقاله ارائه شده در *نخستین کنفرانس بین‌المللی پردازش خط و زبان فارسی*. سمنان

- دفتري نژاد، الهه. ۱۳۸۵. فرایند مارکوف، الگوی احتمالاتی رفع ابهام در زبان‌شناسی رایانه‌ای. علوم انسانی دانشگاه الزهراء (س) ۱۶-۱۷ (۶۳-۶۴): ۱۰۷-۱۳۹.
- ذوالفقار، زهره، طیبه موسوی میانگاه، بلقیس روشن، و امیررضا وکیلی فرد. 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 علانی 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 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، علی اصغر 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 و فناوری اطلاعات. مازندران، ایران

میرزائی، آزاده، و پگاه صفری. ۱۳۹۴. ساخت 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): ۴۷-۲۹.

نظارات، امین؛ طیه 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 Groups 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]
- Alayiaboobar, E., N. Pakniat, A. Hoojatpanah, 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. Mansoorzadeh, & 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 Baghuji, 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]

- Kilgariff, 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=AOvVaw2aAIEADX7IGrhnmULWN85g (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=AOvVaw0Oxy4hd6MjYY_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 Groups linguistics* (pp. 175-191). Tehran: Nevisheh 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: Nevisheh 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 Conference 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 Groups 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.
- Zolfaghari, Z., T. Mosavi Miangah, B. Rovshan, and A. R. Vakili-fard. 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 نقشبند

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، داده‌کاوی، سیمینتیک 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.

