

به نام خداوند بخشنده مهربان



وزارت علوم، تحقیقات و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

طرح پژوهشی

ساخت پیکره از کتاب‌های دیجیتال پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

مجری

الهام علایی ابوذری

همکار

علی اصغر حجت پناه

آذر ۱۴۰۱

حقوق: پژوهشگاه علوم و فناوری اطلاعات ایران

طرح پژوهشی «ساخت پیکره از کتاب‌های دیجیتال پژوهشگاه علوم و فناوری اطلاعات ایران (ایرنداک)» پیرو قرارداد شماره ۳۳/۵۴۶۶ مورخ ۱۴۰۰/۰۷/۱۴ میان پژوهشگاه علوم و فناوری اطلاعات ایران (کارفرما) و خانم دکتر الهام علایی ابوذر اجرا شده است. گزارش حاضر مستند این طرح است. این گزارش و تمامی حقوق مادی آن بر اساس «قانون حمایت از حقوق مؤلفان و مصنفان و هنرمندان» مصوب سال ۱۳۸۴ و اصلاحیه‌های بعدی آن و همچنین آیین‌نامه‌های اجرایی این قانون، متعلق به پژوهشگاه علوم و فناوری اطلاعات ایران است و هرگونه استفاده از تمام یا پاره‌ای از آن شامل نقل‌قول، تکثیر، انتشار، کاربرد نتایج، تکمیل و مانند آن به صورت چاپی، الکترونیکی یا وسایل دیگر تنها با اجازه کتبی پژوهشگاه امکان‌پذیر است. نقل‌قول در حد هزار واژه در انتشارات علمی مانند کتاب و مقاله با درج اطلاعات کامل کتابشناختی، نیازی به مجوز پژوهشگاه ندارد.

شماره: ۳۳۹۵۶۹
تاریخ: ۱۴۰۱/۹/۱۴
پست:

برنام‌خدا

وزارت علوم، تحقیقات و فناوری
پژوهشگاه علوم و فناوری اطلاعات ایران
ایرانداک



پژوهش، مدیریت دانش، آموزش

گواهی

انجام طرح پژوهشی:

- ◇ عنوان: ساخت پیکره از کتاب‌های دیجیتال پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
- ◇ مجری: دکتر الهام علایی ابوذر
- ◇ همکار: مهندس علی اصغر حجت پناه
- ◇ شماره (و تاریخ) قرارداد: ۳۳/۵۴۶۶ (۱۴۰۰/۷/۱۴)
- ◇ تاریخ شروع: ۱۴۰۰/۷/۱۴
- ◇ تاریخ پایان (تصویب گزارش پایانی در شورای پژوهش): ۱۴۰۱/۹/۱۳ (نشست ۴۷۹)
- ◇ ناظران: دکتر محمود بی‌جن‌خان (ناظر علمی)، دکتر محمد حسن زاده (ناظر بهره‌بردار)، نساء رضائی (ناظر فنی)
- در پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) گواهی می‌شود. از سوی ایرانداک، از مجری و همکار و ناظران گرامی این طرح ارزشمند سپاسگزارم.

تندرست باشید

لیلا نامداریان

سرپرست معاونت پژوهش و آموزش

**ساخت پیکره از کتاب‌های دیجیتالی پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)**

مدیر طرح پژوهشی: الهام علایی ابوذر، عضو هیات علمی پژوهشکده علوم اطلاعات پژوهشگاه علوم و فناوری
اطلاعات ایران (ایرانداک)

نشانی: تهران، خیابان انقلاب، تقاطع فلسطین، شماره ۱۰۹۰

تلفن: ۶۶۹۵۱۴۳۰ (۰۲۱) **دورنگار:** ۶۶۴۶۲۲۵۴ (۰۲۱)

رایانامه: alayi@irandoc.a.ir

همکار: علی اصغر حجت‌پناه

رایانامه: hojjatpanah@irandoc.ac.ir

چکیده

ما در دنیای اطلاعات زندگی می‌کنیم؛ گردآوری منسجم اطلاعات، کاری زیرساختی و مهم محسوب می‌شود. یکی از روش‌های گردآوری منسجم داده‌ها، تهیه پیکره‌های زبانی است. تا کنون پیکره‌های عمومی بسیاری در زبان‌های گوناگون، از جمله زبان فارسی، ساخته شده است. اما تعداد پیکره‌های تخصصی که برای اهداف خاص و پردازش‌های خاص زبانی استفاده می‌شود، به اندازه پیکره‌های عمومی نیست. بنابراین، ساخت پیکره‌هایی که شامل متون تخصصی و میان‌رشته‌ای است، بسیار کاربردی است. در این طرح پژوهشی پیکره‌ای ساخته شده است که محتوای آن متون کتاب‌های دیجیتال پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است. برای ساخت این پیکره، پس از طی فرآیند نمونه‌گیری ۶۵ کتاب انتخاب شدند و فراداده مربوط به کتاب‌ها (عنوان مقاله، نام نویسنده/نویسندگان و موضوع مقاله) به طور جداگانه تهیه شد؛ نزدیک به نیمی از کتاب‌ها به صورت فایل پی.دی.اف (PDF) بودند که از طریق استفاده از ابزار آ.سی.آر (OCR) به فایل ورد (WORD) تبدیل شدند و برای وارد شدن در پیکره ویرایش‌های نگارشی و ماشینی روی آن‌ها انجام گرفت. برای هنجارسازی داده‌های وارد شده در پیکره از ابزار هضم، اطلاعات ساختوازی زبان فارسی (وندهای اشتقاقی و تصریفی و واژه‌بست‌ها) و برنامه‌نویسی ماشینی استفاده شد. پس از هنجارسازی، ابتدا برچسب‌گذاری ماشینی داده‌ها (برچسب‌گذاری اجزای واژگانی کلام (POS)) انجام شد و در نهایت برچسب‌ها به صورت دستی نیز کنترل شدند و اطلاعات در پایگاه داده ذخیره شد و پیکره در سامانه پیکره‌های ایرانداک (ساپا) قرار گرفت. این پیکره شامل ۳۳۲۹۲۸۱ واژه است که حوزه‌های میان‌رشته‌ای کتابداری، مدیریت، زبان‌شناسی، مدیریت اطلاعات و غیره را پوشش می‌دهد.

واژه‌های کلیدی: پیکره، پیکره تخصصی، هنجارسازی، اطلاعات ساختوازی زبان فارسی، برچسب‌گذاری اجزای واژگانی کلام

فهرست مطالب

عنوان	شماره صفحه
فصل اول: مقدمه	۷
۱-۱- مقدمه	۸
۲-۱- ضرورت انجام پژوهش	۱۲
۳-۱- پرسش پژوهش	۱۳
فصل دوم: پیشینه پژوهش	۱۷
۱-۲- مقدمه	۱۸
۲-۲- پژوهش‌های پیشین ساخت پیکره زبانی	۱۸
۳-۲- برخی از پیکره‌های زبان فارسی	۲۲
۴-۲- خلاصه فصل	۲۷
فصل سوم: نمونه‌گیری و وارد کردن داده‌ها در پیکره	۲۸
۱-۳- مقدمه	۲۹

۳-۲- نمونه گیری.....	۳۰
۳-۲-۱- جامعه آماری.....	۳۱
۳-۲-۲- نمونه گیری از جامعه آماری.....	۳۱
۳-۳- به کار بردن ابزار اُ.سی.آر (OCR).....	۳۷
۳-۴- تمیز و مرتب کردن متن خوانده شده.....	۳۹
۳-۵- خلاصه فصل.....	۴۳
فصل چهارم: هنجارسازی و نشانه گذاری داده های پیکره.....	۴۴
۴-۱- مقدمه.....	۴۵
۴-۲- هنجارسازی داده ها.....	۴۵
۴-۳- نشانه گذاری داده ها.....	۶۲
۴-۴- کنترل دستی برچسب ها.....	۶۶
۴-۴- نمودار جریانی ساخت پیکره و برچسب گذاری داده ها.....	۷۰
۴-۵- ارزیابی.....	۷۲
۴-۶- ارتقاء سامانه پیکره های ایرانداک (سایا).....	۸۰
۴-۷- خلاصه فصل.....	۸۰
فصل پنجم: نتیجه گیری.....	۸۱
۵-۱- جمع بندی.....	۸۲
۵-۲- پیشنهادات برای پژوهش های آینده.....	۹۰
فهرست منابع فارسی.....	۹۱
فهرست منابع لاتین.....	۹۳

فهرست جدول‌ها

جدول ۳-۱: مشکلات پس از اُ.سی.آر..... ۳۹

جدول ۴-۱: یونیکدهای «ی»، «ک» و «ه»..... ۴۷

جدول ۴-۲: فهرست وندها و واژه‌بست‌های فارسی..... ۴۷

جدول ۴-۳: فهرست پیشوندها..... ۵۸

جدول ۴-۴: فهرست پسوندها..... ۵۸

جدول ۴-۵: فهرست برجسب‌ها..... ۶۵

فهرست شکل‌ها

- شکل ۱-۱: نمودار جریانی از پیمانه‌های کلی ساخت پیکره..... ۱۵
- شکل ۱-۳: شاخص‌های اصلی تهیه پیکره زبانی (برگرفته از علایی: ۱۳۹۸)..... ۳۰
- شکل ۱-۴: نمودار جریانی مراحل ساخت پیکره از کتاب‌های دیجیتال ایرانداک..... ۷۱

سپاسگزاری

به این وسیله از ناظر محترم این طرح پژوهشی، آقای دکتر محمود بی‌جن‌خان که از تجربه و نظرات ارزشمند ایشان در ساخت پیکره بسیار بهره‌مند گشتم، سپاسگزارم.

فصل اول

مقدمه

۱-۱- مقدمه

امروزه انبوهی از اطلاعات، محیط زندگی ما را احاطه کرده است؛ در این راستا، گردآوری منسجم اطلاعات، کاری زیرساختی و مهم محسوب می‌شود. یکی از روش‌های گردآوری منسجم داده‌ها، تهیه پیکره‌های زبانی است. به لحاظ سنتی، زبان‌شناسان واژه پیکره^۱ را برای مجموعه داده‌های زبانی به کار می‌برند که می‌تواند پایه و اساس پژوهش‌های زبانی باشد؛ این داده‌های زبانی می‌تواند شامل متون نوشتاری، گفتاری یا نمونه‌هایی از زبان نوشتاری یا گفتاری باشد. پیکره اغلب به منظور نمایش زبان یا گونه خاص زبانی طراحی می‌شود. در دهه‌ها اخیر واژه پیکره را برای مجموعه داده‌های زبانی به کار می‌برند که به شکل الکترونیکی موجود باشد و قابلیت پردازش رایانه‌ای برای اهداف گوناگون، مانند پژوهش‌های زبان‌شناسی و مهندسی زبان، را داشته باشد. سینکلر (۲۰۰۴) پیکره را اینگونه تعریف می‌کند: پیکره، مجموعه‌ای نظام‌مند، رایانه‌ای شده، معتبر و موثق از زبان است که برای مطالعات زبان‌شناسی مورد استفاده قرار می‌گیرد. مک‌انری و ویلسون (۲۰۰۱) نیز پیکره را مجموعه‌ای کمابیش بزرگ از متون الکترونیکی می‌خوانند که حاشیه‌نویسی، برچسب‌گذاری و دسته‌بندی سنجیده‌ای دارند و از همین رو بررسی‌های بسیار گوناگونی را برای کاربر شدنی می‌سازند. کریستال (۱۹۹۴) می‌نویسد پیکره زبانی، مجموعه‌ای از داده‌های زبانی است که بصورت متون نوشتاری یا آوا نوشته‌های گفتاری گردآوری شده‌اند. از دهه ۱۹۶۰

¹ corpus

به این سو، پژوهشگران توانسته‌اند با کمک رایانه، مجموعه‌های بسیار بزرگی از متون معتبر زبانی را فراهم کنند که آغازگر همه این‌ها، پیکره زبانی «براون»^۱ بود. این پیکره، نمونه‌ای معیار از زبان انگلیسی امروز برای استفاده در رایانه‌های رقمی به شمار می‌رود (اسوارت ویکل: ۱۹۹۲، نقل از عاصی: ۱۳۸۲). در آغاز دهه ۱۹۸۰ شاید می‌شد پیکره‌های الکترونیکی را با انگشتان یک دست شمرد. اما در آغاز دهه ۱۹۹۰ شمار طرح‌هایی که در پی ساخت پیکره‌های زبانی بودند به ده‌ها می‌رسید. اکنون پس از گذشت تنها چند دهه، با جست‌وجویی در اینترنت تعداد بیشمار پیکره بزرگ و کوچک برای بسیاری از زبان‌های جهان یافت می‌شوند (عاصی: ۱۳۸۲). برخی از زبان‌شناسان حوزه نظری زبان، اعتقادی به پیکره ندارند. به عنوان نمونه، چامسکی معتقد است برخی از جمله‌های زبان در پیکره وجود ندارند، از طرفی برخی دیگر از داده‌های موجود در پیکره، جمله‌های غلط یا غیرمودبانه هستند که نباید در پیکره گنجانده شوند؛ وی بر این باور است که پیکره نمی‌تواند نماینده زبان باشد. زبان‌شناسی چامسکیایی روی طبقه‌بندی جمله‌ها بر حسب خوش‌ساختی یا بدساختی، صرف‌نظر از کاربرد آن‌ها، تمرکز می‌کند؛ در حالیکه در پردازش زبان طبیعی (NLP)^۲ احتمالات و فراوانی ساخت‌های زبانی از اهمیت بالایی برخوردار است که چنین بررسی‌های آماری و احتمالاتی را می‌توان در پیکره‌ها انجام داد. پردازش آماری زبان طبیعی به توصیف ساختارهای ترجیح داده شده با فراوانی بیشتر در کل نظام زبان می‌پردازد. به عنوان مثال، گویشوران غیربومی یک زبان چیزهایی را می‌گویند یا می‌نویسند که به هیچ وجه از لحاظ نحوی غیردستوری نیستند، اما تا حدی عجیب به نظر می‌رسند. این گونه جملات برای گویشوران بومی قابل فهم است اما می‌توان آن‌ها را کمی تغییر داده و منظور گوینده را بهتر نشان داد. اگر گویشوران غیربومی یک زبان، ساختارهای پربسامد را بدانند، این مشکل پیش نخواهد آمد. البته می‌دانیم هیچ پیکره‌ای وجود ندارد که همه بخش‌های زبان (آواها، ساختار، ساختار نحوی، معنی و غیره) را پوشش دهد و بنابراین، همیشه پیکره ناقص است و کامل نیست. اما هر پیکره‌ای که مورد بررسی قرار گرفته است، حتی کوچک، حقایقی در آن وجود داشته که در جای دیگر با آن برخورد نکرده‌ایم و بنابراین از هر پیکره، می‌توان نکته‌هایی را آموخت.

^۱ Brown Corpus

^۲ natural language processing

پردازش زبان طبیعی، به مجموعه روش‌های رایانه‌ای گفته می‌شود که جنبه نظری دارد و برای تجزیه و تحلیل متون در یک یا چند لایه تحلیل زبان - شناختی با هدف یافتن چگونگی پردازش انسان گونه زبان در محیط رایانه مورد استفاده قرار می‌گیرد (لیدی: ۲۰۰۱).

با افزایش قدرت و ظرفیت رایانه‌ها، پیکره‌ها نیز از نظر اندازه، تنوع و سهولت دسترسی، افزایش چشم‌گیری داشته‌اند و هم‌زمان با این تحولات، نرم‌افزارهای بسیاری نیز برای پردازش پیکره‌ها و دسترسی به اطلاعات درون پیکره‌ها، توسعه داده شدند. طی سال‌های اخیر تهیه پیکره‌های زبانی و تجزیه و تحلیل پیکره-بنیاد زبان بسیار مورد توجه پژوهشگران در حوزه زبان‌شناسی، زبان‌شناسی رایانشی و هوش مصنوعی قرار گرفته است. گرچه در گذشته، بسیاری از زبان‌شناسان بر اهمیت پیکره زبانی در بیشتر بررسی‌هایشان تأکید کرده‌اند، اما در دوران جدید است که تکیه بر داده‌های واقعی زبان به صورت گسترده‌ای رواج یافته و شرط اساسی بسیاری از پژوهش‌های نظری و کاربردی مانند: نظریه‌پردازی و توصیف ساختمان زبان، گویش‌شناسی، دستورنویسی و فرهنگ‌نگاری به شمار می‌آید. بحث پیکره‌ها و استفاده از پیکره‌ها در پژوهش‌ها و کارهای گوناگون، سابقه طولانی دارد. امکان سازمان‌دهی، تنظیم، تفکیک، جست‌وجو و دستیابی سریع داده‌های زبانی، افق‌های تازه‌ای در برابر پژوهشگران گشوده و باعث پیدایش شاخه‌ای تخصصی در حوزه زبان‌شناسی رایانه‌ای گردیده است؛ این شاخه با نام زبان‌شناسی پیکره‌ای^۱، تنها در آخرین دهه‌های قرن بیستم ایجاد شده است و در همین زمان کوتاه تبدیل به یکی از فعال‌ترین و پرکاربردترین زمینه‌ها شده است. ویژگی این رشته این است که به همه حوزه‌های زبان‌شناسی خدمات می‌دهد و در حقیقت، زبان‌شناسی پیکره‌ای در خدمت همه بررسی‌های زبانی است. همین مسئله دلیل پویایی و گسترش این رشته است. زبان‌شناسی پیکره‌ای به مطالعه فقره‌های واژگانی، ساخت‌های دستوری، یا پیوند این دو با دیگر مشخصه‌های زبانی و غیرزبانی می‌پردازد. در حقیقت، پژوهش‌های پیکره‌ای بر الگوهای واقعی کاربرد زبان در پایگاه‌های حجیم دادگان یا همان پیکره‌ها تمرکز دارد و رویکردی مکمل برای رویکردهای سنتی‌تر به بررسی زبان در نظر گرفته می‌شود. زبان‌شناسی پیکره‌ای را می‌توان گذار از رویکردهای ذهن‌گرایانه به پژوهش‌های زبانی دانست. در این میان، هم از روش‌های کمی و هم از روش‌های کیفی برای تحلیل زبان بهره می‌برد و رایانه را برای انجام تحلیل‌های پیچیده به کار می‌گیرد. آنچه زبان‌شناسی پیکره‌ای را به کار زبان می‌آورد، توانمندی آن در بررسی کاربرد واقعی زبان است تا پژوهشگر از زیاده‌روی در نظریه‌پردازی دور نگه داشته شود. همین توانمندی نیز واکاوی پایگاه‌های بزرگ داده‌های زبانی را به زبان‌شناسی پیکره‌ای می‌سپارد (کوهستانی: ۱۳۸۹). تیوبرت (۲۰۰۵) برای روشن شدن زبان‌شناسی پیکره‌ای در برابر سنتی، ویژگی‌هایی از شیوه پژوهش‌های زبانی در چارچوب زبان‌شناسی پیکره‌ای را برشمرده است که تأکید روی معنا، تأکید روی متن، کاربرد داده‌های

^۱ corpus linguistics

واقعی زبان و توجه به بعد در زمانی کلام را در بر دارد. زبان‌شناسی پیکره‌ای، ابزارهایی بنیادی برای بررسی‌ها، پژوهش‌ها، و کاربردهای زبانی گوناگون فراهم می‌آورد. بسیاری از پژوهش‌های زبان‌شناختی و تصمیم‌گیری‌ها در برنامه‌ریزی زبانی، تنها با استفاده از یک پیکره زبانی امکان‌پذیر است. برخی از کاربردهای پیکره شامل این موارد است: پردازش زبان طبیعی و درک و بازشناسی گفتار، تبدیل متن به گفتار و برعکس، تدوین فرهنگ‌ها، آموزش و پژوهش، ساخت پایگاه‌های داده زبانی، بررسی هم‌آیندها^۱ در زبان‌های گوناگون، پایشگری زبان برای پیگیری و ردگیری دگرگونی‌های زبانی، ترجمه ماشینی، توسعه مفاهیم و منابع در پیوند با واژگان، نگارش و گسترش مهارت‌های نوشتاری، آموزش و یادگیری زبان با شناخت گویش‌ها و گوناگونی زبانی، معناشناسی، تحلیل کلام، زبان‌شناسی اجتماعی، زبان‌شناسی حقوقی، واکاوی ژانرهای ادبی و پژوهش‌های دستوری.

پیکره انواع گوناگونی دارد؛ اتکینز و همکاران (۱۹۹۲) انواع پیکره را از چند دیدگاه مختلف بررسی کرده‌اند و بر آن اساس پیکره‌ها را به انواع «متن کامل»^۲، «نمونه‌ای»^۳، «نظارتی»^۴، «بسته»^۵، «باز»^۶، «هم-زمانی»^۷، «در زمانی»^۸، «عمومی»^۹، «اصطلاح‌شناسی»^{۱۰}، «یک‌زبانه»^{۱۱}، «دو‌زبانه»^{۱۲}، «چندزبانه»^{۱۳}، «منفرد»^{۱۴}، «موازی»^{۱۵}، «مرکزی»^{۱۶}، «پوسته‌ای»^{۱۷}، «هسته‌ای»^{۱۸} و «پیرامونی»^{۱۹} طبقه‌بندی کرده‌اند.

¹ collocations

² whole text

³ samples

⁴ monitor

⁵ open

⁶ closed

⁷ synchronic

⁸ diachronic

⁹ general

¹⁰ thesaurus

¹¹ monolingual

¹² bilingual

¹³ multilingual

¹⁴ single

¹⁵ parallel

¹⁶ central

¹⁷ cluster

¹⁸ nuclear

¹⁹ Perimeter

۲-۱- ضرورت انجام پژوهش

یکی از الویت‌های پژوهشی ایرانداک، ساخت پیکره از داده‌های علمی ایرانداک است. تا کنون پیکره‌های عمومی بسیاری در زبان‌های گوناگون، از جمله زبان فارسی، ساخته شده است. اما تعداد پیکره‌های تخصصی که برای اهداف خاص و پردازش‌های خاص زبانی استفاده می‌شود، به اندازه پیکره‌های عمومی نیست. بنابراین، ساخت پیکره‌هایی که شامل متون تخصصی و میان‌رشته‌ای است، بسیار ارزشمند است. در طرح پژوهشی پیشین اینجانب تحت عنوان «ساخت پیکره متنی از مقاله‌های پژوهش‌نامه پردازش و مدیریت اطلاعات»، سامانه‌ای نیز ساخته شد به نام «سامانه پیکره‌های ایرانداک (ساپا)» که پیکره‌ها در این سامانه قرار می‌گیرند. اکنون «پیکره پژوهش‌نامه» که در طرح پژوهشی «ساخت پیکره متنی از مقاله‌های پژوهش‌نامه پردازش و مدیریت اطلاعات» ساخته شده بود، در این سامانه قرار دارد. در پژوهش حاضر قصد داریم پیکره تخصصی دیگری را به سامانه اضافه کنیم تا امکانات بیشتری برای جامعه علمی برای پژوهش‌هایی که مستلزم بهره‌گرفتن از پیکره‌های تخصصی است، فراهم گردد. محتوای این پیکره، متون کتاب‌های دیجیتال ایرانداک است. این پیکره همان ملاحظات فنی را دنبال می‌کند که در ساخت «پیکره پژوهش‌نامه» در نظر گرفته شده است. تفاوت این دو پیکره در محتوای آن‌هاست. محتوای پیکره پژوهش‌نامه، متون مقاله‌های پژوهش‌نامه پردازش و مدیریت اطلاعات است، حال آنکه، محتوای پیکره‌ای که در طرح پژوهشی پیش رو ساخته خواهد شد، متون کتاب‌های چاپ‌شده در ایرانداک (کتاب‌های دیجیتال ایرانداک) است. البته ناگفته نماند که می‌شد پیکره‌ای بزرگ در ایرانداک ساخت که شامل مقاله‌های پژوهش‌نامه پردازش و مدیریت اطلاعات، کتاب‌های دیجیتال ایرانداک، گزارش طرح‌های پژوهشی انجام شده در ایرانداک و حتی پایان‌نامه‌ها و رساله‌ها (پارساهای ثبت شده در ایرانداک باشد. اما به دلیل محدودیت‌های حقوقی، مالی، فنی، زمانی و غیره، چنین پیکره‌ای ساخته نشده است و تصمیم گرفته شد که پیکره‌های گوناگونی در سامانه پیکره‌های ایرانداک قرار گیرد. شایان ذکر است برخی از پیکره‌هایی که ساخته شده‌اند نیز همین ویژگی را دارند. به عنوان نمونه، ابتدا «پیکره بی‌جن‌خان» ساخته شد (که دو میلیون و ۶۰۰ هزار واژه دارد)، سپس، این پیکره توسعه داده شد و «پیکره متنی فارسی» ساخته شد (که بیش از ۱۰۰ میلیون واژه دارد)؛ اکنون پژوهشگران، بسته به ماهیت پژوهش خود، از هر دو پیکره استفاده می‌کنند. آزمایشگاه پردازش زبان طبیعی دانشکده مهندسی و علوم کامپیوتر دانشگاه شهید بهشتی نیز «وردنت عمومی زبان فارسی» (فارس‌نت ۱، فارس‌نت ۲ و فارس‌نت ۳) را از سال ۸۷ تا کنون در قالب سه

پروژه ساخته است. با توجه به اینکه هنگام ثبت نام در سامانه فارس نت، امکان دریافت نسخه‌های فارس-نت ۲,۵ و نسخه‌های قبل تر و همچنین فارس نت ۳ وجود دارد، بنابراین، این نسخه‌ها مستقل از هم هستند.

ساخت پیکره‌های تخصصی گوناگون از منابع مختلف از این نظر نیز دارای اهمیت است که کاربرانی که می‌خواهند یک واژه یا عبارت خاص را تنها در مجموعه مقاله‌ها یا فقط در کتاب‌های میان‌رشته‌ای، جستجو کنند، می‌توانند پیکره مورد نظر را انتخاب کنند و در آن به جستجو پردازند؛ زیرا در سامانه پیکره‌های ایرانداک امکانی فراهم شده است که کاربر می‌تواند واژه مورد جستجو را در یک پیکره خاص یا همه پیکره‌های موجود در سامانه جستجو کند. به عنوان نمونه، برای بررسی نحوی زبان علم در مقاله‌های تخصصی، می‌توان فقط از پیکره پژوهش نامه استفاده کرد.

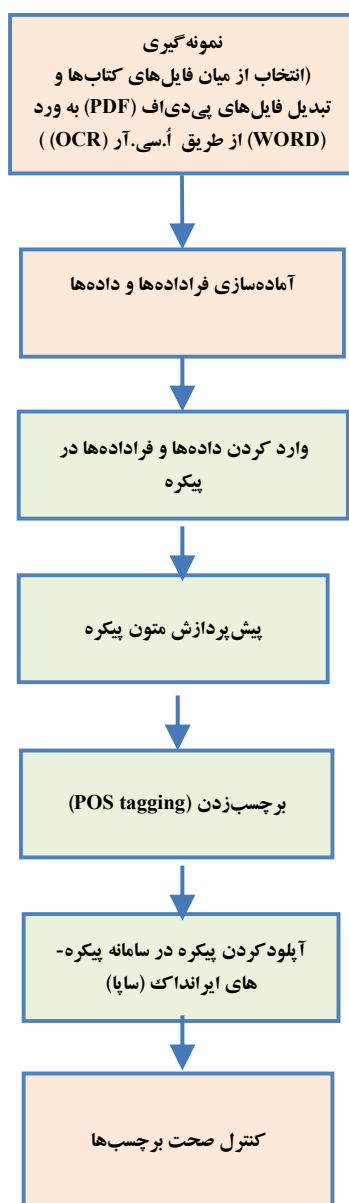
۳-۱ پرسش پژوهش

این طرح پژوهشی سعی دارد هنگام ساخت پیکره به سوال‌های زیر نیز توجه داشته باشد:

- مهمترین ملاک‌ها و معیارهای انتخاب کتاب‌های دیجیتال برای ساخت پیکره کدامند؟
 - کارآمدترین ابزارهای پیش پردازش متون در ساخت پیکره از کتاب‌های دیجیتالی کدامند؟
- برای ساخت پیکره از کتاب‌های دیجیتال پژوهشگاه علوم و فناوری اطلاعات، به طور کلی، مراحل زیر باید طی شود که در نمودار جریانی (شکل ۱-۱) نمایش داده شده است:
- دریافت متن کامل کتاب‌های دیجیتال ایرانداک
 - انتخاب کتاب‌هایی که فایل ورد (word) آن‌ها موجود است.
 - بررسی کتاب‌هایی که فایل ورد آن‌ها موجود نیست (فایل پی.دی.اف (pdf) آن‌ها موجود است) و انتخاب از میان این کتاب‌ها برای استفاده از ا.سی.آر (OCR) و تبدیل فایل‌های انتخاب شده به ورد.
 - تمیز و مرتب کردن فایل‌هایی که از طریق ا.سی.آر به ورد تبدیل شده‌اند (توضیح: هنگام ا.سی.-آر، ممکن است در برخی موارد، علامت سوال دیده شود یا برخی از واژه‌ها به هم چسبیده شوند، این موارد لازم است بررسی شوند و تا حد امکان فایل‌های به دست آمده با فایل پی.دی.اف آن‌ها مطابقت داشته باشد)
 - آماده سازی فراداده فایل‌ها برای وارد کردن در پیکره

- وارد کردن داده‌ها و فراداده مربوط به آن‌ها در پیکره
- پیش‌پردازش متون: این مرحله به کمک نرم‌افزارهای موجود انجام خواهد گرفت. (توضیح: در ساخت پیکره متنی، پس از انتخاب متونی که در پیکره متنی مورد استفاده قرار می‌گیرند، باید پیش‌پردازش‌هایی روی متن انجام شود که به اصطلاح «هنجارسازی متن» نامیده می‌شود. این پیش‌پردازش‌ها در متون فارسی می‌تواند شامل، یک‌دست کردن فاصله‌ها و نشانه‌گذاری‌های درون متن و یکسان کردن یونیکد نویسه‌های استفاده شده در متون (مانند انواع «ی»، «ک» و همزه) باشد.
- برچسب‌زدن به متون (نوعاً POS tagging): این مرحله به کمک نرم‌افزارهای موجود انجام خواهد گرفت.
- بارگذاری پیکره در سامانه پیکره‌های ایرانداک.
- کنترل صحت برچسب‌ها: این کار به صورت دستی و ماشینی و به کمک ابزارهای توسعه داده شده در سامانه پیکره‌های ایرانداک انجام خواهد شد.

ساخت پیکره از کتاب‌های دیجیتال پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) □ ۱۵



شکل ۱-۱: نمودار جریانی^۱ از پیمان‌های کلی ساخت پیکره

در طرح پژوهشی پیش رو، به منظور افزایش کارایی (ارتقاء) سامانه پیکره‌های ایرانداک (ساپا) در زمینه جست‌وجو، ویرایش و بازنمایی اطلاعات، قابلیت‌هایی نیز به سامانه اضافه خواهد شد که شامل موارد زیر است:

^۱ block diagram

- بهینه‌سازی اصلاح برچسب‌ها: امکان ویرایش کردن^۱ برچسب‌ها یا واژه‌ها به صورت گروهی یا صفحه‌ای و ذخیره (save) کردن تغییرات بعد از ویرایش هر صفحه/مقاله/کتاب. امکان ویرایش برچسب‌ها و واژه‌ها بر اساس الگوها (بر اساس جدول تهیه شده در طرح پیکره پژوهش‌نامه).
- بهینه‌سازی فیلتر کردن برچسب‌ها که شامل این موارد است: ۱- امکان خروجی گرفتن از هر یک از (Verb)V, (Noun)N, (Adjective)ADJ؛ مثلاً در پیکره چند فعل /v/ داریم و فراوانی هر فعل چقدر است و ۲- امکان ویرایش همزمان برچسب‌ها (مثلاً واژه‌ای را جست‌وجو (search) کرده‌ایم و تصادفاً متوجه شدیم به اشتباه برچسب /N/ خورده، بتوانیم آنجایی که /N/ خورده را اصلاح کنیم و برچسب درست بدهیم.
- اضافه کردن بخشی تحت عنوان «راهنمای برچسب‌ها» در صفحه متن برچسب‌گذاری شده.
- بهینه‌سازی هایلایت کردن واژه جست‌وجو و برچسب آن: هایلایت واژه جست‌وجو در صفحه متن بدون برچسب، متن با برچسب و همچنین در بخش فیلتر برچسب‌ها. به عنوان مثال، اگر دنبال واژه «اسلامی» با برچسب /N/ هستیم، فقط آن مقاله/کتاب‌هایی بیاید که این واژه به عنوان اسم به کار رفته است و هایلایت شده باشد، به عبارتی، موارد دیگر واژه «اسلامی» که برچسب صفت (ADJ) یا غیره دارد، هایلایت نشود.
- جستجوی پیشرفته

در این طرح پژوهشی ابتدا در فصل دوم، به پیشینه پژوهش، که در واقع مرور مطالعات انجام شده در حوزه ساخت پیکره زبانی است، پرداخته می‌شود و در پایان برخی از پیکره‌های زبان فارسی معرفی می‌شوند. در فصل سوم به چگونگی فرآیند نمونه‌گیری و وارد کردن داده‌ها در پیکره پرداخته می‌شود. فصل چهارم، چگونگی هنجارسازی و حاشیه‌نویسی پیکره و ابزارهای مربوطه و ارتقاء سامانه پیکره‌های ایرانداک توضیح داده می‌شود و در نهایت فصل پنجم، نتیجه‌گیری و ارائه پیشنهادات برای پژوهش‌های آینده است.

¹ edit

فصل دوم

پیشینه پژوهش

۲-۱- مقدمه

در این فصل پژوهش‌های انجام شده در زمینه ساخت پیکره زبانی، مورد بررسی قرار خواهند گرفت و در ادامه به معرفی برخی از پیکره‌های زبان فارسی پرداخته خواهد شد.

۲-۲- پژوهش‌های پیشین ساخت پیکره زبانی

بسیاری از پژوهشگران معتقدند پیش از انجام هر پژوهشی که قرار است برونداد آن ملموس، کاربردی و عملیاتی باشد، بهتر است ابتدا مطالعه امکان‌سنجی انجام شود. درباره ساخت پیکره تنها علایی و علیدوستی (۱۳۹۹) به امکان‌سنجی ساخت پیکره پرداخته‌اند. امکان‌سنجی، گام‌ها، هزینه‌ها، نیروی انسانی، حقوق مادی و معنوی، و مانند آن‌ها را برای یک پروژه، بررسی و به مدیریت پروژه کمک می‌کند تا آن را با آمادگی و آینده‌نگری بیشتری به سرانجام برساند. در این راستا و به منظور ارائه مدلی برای امکان‌سنجی ساخت پیکره، علایی و علیدوستی پس از بررسی پژوهش‌های پیشین، شاخص‌های کلی ساخت پیکره‌های زبانی را استخراج کرده‌اند. سپس، فرآیند ساخت پیکره را به تفصیل مورد بررسی قرار داده‌اند. پس از بررسی پژوهش‌های پیشین در زمینه امکان‌سنجی، بر پایه اطلاعات به دست آمده از مطالعات امکان‌سنجی، بررسی شاخص‌های ساخت پیکره و فرآیند ساخت پیکره، ابعاد و مولفه‌های امکان‌سنجی ساخت پیکره استخراج شده است و مدلی کلی برای ساخت پیکره پیشنهاد شده است. سپس با روش دلفی و در دو دور اعتباریابی، مدل پایانی به دست آمده است. این مدل دارای هفت بعد فنی، اقتصادی، زمان-بندی، قانونی حقوقی، عملیاتی، تأمین مالی، و بازاریابی و روی هم رفته ۳۳ مؤلفه است.

برزینا و همکاران (۲۰۲۰) درباره پیکره ملی بریتانیا (نوع متنی)^۱ ۲۰۱۴ توضیح می‌دهند. پیکره ملی بریتانیا ۲۰۱۴، در قالب پروژه‌ای در دانشگاه لنکستر^۲ ساخته شده است که شامل ۱۰۰ میلیون واژه انگلیسی معاصر است. این پیکره، در حقیقت به عنوان همتای مقایسه‌ای با اصل پیکره ملی بریتانیا، که در آغاز دهه ۱۹۹۰ گردآوری شد، ساخته شده است. در این مقاله، علاوه بر توجیه علت ساخت چنین پیکره‌ای، نویسندگان تصویری کلی از این پیکره ارائه می‌دهند. همچنین درباره چگونگی گردآوری داده‌ها نیز توضیحاتی ارائه می‌دهند. آن‌ها بر این مسئله تاکید دارند که ساخت پیکره‌های بزرگ مستلزم همکاری تیمی و کمک افراد و سازمان‌های است که در مسیر گردآوری داده‌ها همراهی می‌کنند.

وین (۲۰۰۵) آراء نویسندگان و پژوهشگران سرشناس حوزه ساخت پیکره را ارائه می‌کند؛ آن‌ها پیشنهادات کاربردی برای ساخت پیکره ارائه می‌دهند. راهنمایی‌های این پژوهشگران می‌تواند مورد استفاده کسانی قرار گیرد که در مرحله ساخت پیکره زبانی هستند. از موضوعاتی که در این کتاب مورد بررسی قرار می‌گیرد تعاریف کلی مربوط به پیکره، اضافه کردن حاشیه‌نویسی زبانی، نوع داده‌های مورد استفاده و پیکره آوایی را می‌توان ذکر کرد. این کتاب، کتابی کاربردی است که به زبان ساده به بررسی مراحل ساخت پیکره زبانی می‌پردازد.

برخی از پژوهشگران اقدام به ساخت پیکره‌های تخصصی کرده‌اند که از میان آن‌ها می‌توان به این موارد اشاره کرد: دیویس (۲۰۲۱)، بلوسو (۲۰۱۵)، کلاد توریدا (۲۰۱۶) و علایی و همکاران (۱۴۰۰). دیویس (۲۰۲۱) چگونگی ساخت پیکره ویروس کرونا^۳ را توضیح می‌دهد. در حال حاضر (ماه مارچ ۲۰۲۱) این پیکره شامل ۹۰۰ میلیون واژه است. این پیکره، زیرمجموعه‌ای است از پیکره NOW^۴ که شامل ۱۲ میلیارد واژه است. این دو پیکره مرتب به‌روزرسانی می‌شوند. هر روز کاربران می‌توانند فراوانی واژه‌ها و عبارت‌ها را در پیکره ویروس کرونا ببینند. همچنین پیکره این امکان را فراهم آورده است که کاربران بتوانند باهم‌آیی‌های واژه‌ها و عبارت‌ها را ببینند و آن‌ها را با هم مقایسه کنند تا به این وسیله امکان بررسی این مسئله فراهم می‌شود که طی زمان درباره یک موضوع خاص چه مطالبی بیان می‌شود. بلوسو (۲۰۱۵) درباره پیکره تخصصی CADCE^۵ توضیح می‌دهد. این پیکره شامل مجموعه‌ای از ۵۰۰ هزار واژه زبان نوشتاری از منابع گوناگون است که نماینده زبان معماری در انگلیسی معاصر است و برای بررسی واژگان

¹ written British National corpus

² Lancaster

³ Coronavirus corpus

⁴ News On the Web (NOW)

⁵ Corpus of Architecture Discourse in Contemporary English (CADCE)

این حوزه ساخته شده است. این پیکره تک‌زبانه است، نشانه‌گذاری/حاشیه‌نویسی نشده است و شامل متون منتشر شده از گونه‌های زبانی انگلیسی آمریکای شمالی، بریتانیا، ایرلندی، کانادایی و استرالیایی است. از آنجائیکه شامل متون منتشر شده در سال‌های اخیر (۲۰۰۸-۲۰۰۷) است. این پیکره تخصصی، هم‌زمانی است و دارای متونی در حوزه‌های مربوط به معماری، شامل ساخت‌وساز، شهرسازی، مواد ساخت‌وساز، معماری سبز، طراحی داخلی و سایر حوزه‌های این رشته است. این پیکره روی نمایندگی، معاصر بودن و قابلیت در دسترس بودن به عنوان اصول مهم ساخت پیکره تاکید دارد. کلاد توریدا (۲۰۱۶) ساخت پیکره تخصصی و تهیه فهرست واژگان حاشیه‌نویسی شده و مبتنی بر فراوانی واژگان در پیکره را گام‌به‌گام توضیح می‌دهد. وی از پروژه «آموزش زبان انگلیسی برای اهداف دانشگاهی» که در دانشگاهی در خاورمیانه توسعه یافته است، نمونه‌هایی را ذکر می‌کند. مراحل ساخت پیکره تخصصی از نظر کلاد توریدا شامل انتخاب متون آموزشی، حذف واژگانی که قاموسی نیستند (مانند حروف اضافه، حروف ربط و....)، تجزیه و تحلیل متن با استفاده از نرم‌افزار AntConc، ایجاد فهرست فراوانی واژه‌ها، توسعه فهرست واژگان حاشیه‌نویسی شده که خود شامل: تعیین مقوله نحوی (POS) واژگان موجود در فهرست، اضافه کردن تعریف واژگان، باهم‌آیی واژگان و نمونه‌ای از جمله‌ای که واژه در آن به کار رفته، است. علایی و همکاران (۱۴۰۰) نیز پیکره‌ای ساخته‌اند که شامل متون مقاله‌های موجود در پژوهش‌نامه پردازش و مدیریت اطلاعات است. این پیکره شامل بیش از ۶۰۰ مقاله با بیش از چهار میلیون واژه است. موضوع این مقاله‌ها شامل کتابداری و اطلاع‌رسانی، علم اطلاعات و دانش‌شناسی، فناوری اطلاعات، مدیریت اطلاعات، مدیریت دانش، زبان‌شناسی، زبان‌شناسی رایانشی، اصطلاح‌شناسی، هستان‌شناسی و سایر حوزه‌های مرتبط با پردازش اطلاعات است. بنابراین، محتوای این پیکره عمومی نیست و حاوی متون بسیار تخصصی و میان‌رشته‌ای است و از این نظر برای پردازش‌هایی که مستلزم بهره‌گرفتن از متون تخصصی است، ارزشمند است.

از جمله تلاش‌هایی که در زمینه ساخت پیکره‌های دوزبانه فارسی-انگلیسی انجام گرفته است، می‌توان به پیکره دشتبانی و همکاران (۱۳۹۳) اشاره کرد که پیکره‌ای است دو زبانه در حوزه فاوا (حوزه فناوری ارتباطات و اطلاعات). این پیکره به صورت خودکار ساخته شده است و منابع آن، اسناد تخصصی حوزه فاوا است. در این پژوهش، نرم‌افزاری برای ساخت پیکره طراحی شده است که هزینه و مدت زمان ساخت پیکره را کاهش می‌دهد. علاوه بر این، نرم‌افزار ارائه شده قابلیت مدیریت پیکره را برای کاربران فراهم می‌کند. سیستم مدیریت پیکره دارای دو بخش اصلی است که بخش اول مربوط به ساخت پیکره است و

بخش دوم مربوط به استخراج اطلاعات از پیکره است. نرم‌افزاری که برای مدیریت پیکره ایجاد شده است، حاشیه‌نویسی اسناد، جستجو در پیکره و تصحیح خطا را آسان می‌کند. قبل از شروع فرآیند پردازش اصلی، هر سند توسط نرم‌افزار پیش‌پردازش می‌شود؛ این کار به منظور انتخاب جمله‌های درست و معنی‌دار برای پردازش اصلی است؛ علاوه بر آن، در صورتیکه در سند کاراکترهای بی‌معنی وجود داشته باشد، در فرآیند پیش‌پردازش از سند حذف می‌شوند. از جمله کارهایی که این سیستم انجام می‌دهد می‌توان به این موارد اشاره کرد: ویرایش پیکره، اضافه کردن متن‌های جدید به پیکره، اندیس‌گذاری و حاشیه‌نویسی پیکره. بخش دوم، یک موتور جستجوی پیکره است که برای مدیریت مجموعه بزرگی از متون طراحی شده است. پردازش متون به کمک سیستمی انجام شد که دارای این بخش‌ها است: طبقه‌بند^۱ برای پذیرش اسناد حوزه فاوا، برچسب‌گذار نقش دستوری واژگان (برچسب‌گذاری اجزای واژگانی کلام) و تجزیه‌کننده^۲ برای اسناد فارسی و یک تجزیه‌کننده، برچسب‌گذار نقش دستوری واژگان و ریشه‌یاب^۳ برای اسناد انگلیسی است. اسناد حوزه فاوا به کمک این سیستم حاشیه‌نویسی می‌شوند و اطلاعات پردازش‌شده اسناد در پایگاه داده پیکره ذخیره می‌شوند. مهم‌ترین مرحله ساخت پیکره‌های چندزبانی، ترازبندی داده‌های پیکره است. در این پروژه روشی برای ترازبندی جمله‌های پیکره فارسی تخصصی حوزه فاوا و جملات انگلیسی پیکره تخصصی حوزه فاوا ارائه شده است. الگوریتم پیشنهادی آن‌ها از مدل ترجمه کلمه‌به‌کلمه و تکنیک بلندترین زیردنباله مشترک^۴ برای ترازبندی استفاده می‌کند و در نهایت امتیاز نشان دهنده شباهت دو جمله، محاسبه می‌شود و اطلاعات مربوط به نگاشت جمله‌های دو مجموعه انگلیسی و فارسی در پایگاه داده پیکره، ذخیره می‌گردد.

تهیه پیکره متنی برای زبان فارسی با مشکلاتی همراه است؛ برخی از این مشکلات توسط قیومی و همکاران (۲۰۱۳) مورد بررسی قرار گرفته است. آن‌ها معتقدند منبع مشکلات مربوط به عوامل متعددی می‌شود، از جمله: نظام نوشتاری فارسی، تلفیق نظام نوشتاری فارسی با خط عربی، شیوه تایپ کردن تایپست‌ها و خلاقیت زبانی. این مشکلات به صورت خودکار و از طریق برنامه‌نویسی یا به صورت دستی حل شده است. بی‌جن‌خان و همکاران (۲۰۱۱) نیز به بررسی مسائلی می‌پردازند که طی ساخت یک منبع مکتوب زبانی، تحت عنوان «پیکره»، مورد توجه قرار گرفته است. این پیکره حاوی ۳۵/۰۵۸ فایل متنی

¹ classifier

² parser

³ stemmer

⁴ longest common subsequence (LCS) هدف این روش، مقایسه دو رشته و پیدا کردن شباهت بین آن‌هاست

است که هر کدام حاوی متن کامل یا نمونه تصادفی از یک فایل متنی هستند. اندازه هر متن زنجیره‌ای است با حداقل ۱۰۰۰ کلمه. برای حاشیه‌نویسی فارسی در پیکره، از دستورالعمل EAGLES استفاده شده است. این دستورالعمل برای زبان‌های اروپایی ساخته شده است؛ البته برای زبان‌های غیراروپایی مانند عربی هم استفاده شده است. از این دستورالعمل برای فارسی نیز می‌توان استفاده کرد؛ زیرا علاوه بر اینکه فارسی هم جزء زبان‌های هندواروپایی است، در مقایسه با انگلیسی، ساخت‌واژه تصریفی فارسی در مورد افعال و اسامی به حد کافی غنی است و می‌توان از این دستورالعمل استفاده کرد. نتیجه به دست آمده از به کار گرفتن این دستورالعمل، به صورت سلسله‌مراتبی از برجسب‌ها می‌باشد. برای حاشیه‌نویسی از روش نیمه-خودکار استفاده شده است. در تهیه این پیکره، به ساخت اضافه و هم‌نگاره‌ها که در تجزیه و تحلیل متن مشکلاتی ایجاد می‌کنند، توجه خاصی شده است.

همانگونه که مرور ادبیات نشان می‌دهد، پیکره‌های زبانی و پژوهش‌های پیکره‌بنیاد در زبان‌شناسی پیکره‌ای نقشی کلیدی دارند. ساخت پیکره‌ها کاری پیچیده، زمان‌بر، دارای گام‌های گوناگون، و میان‌رشته است. در این راستا علایی و علیدوستی (۱۳۹۹) مدلی برای امکان‌سنجی ساخت پیکره ارائه داده‌اند. برخی از پژوهشگران مانند برزینا و همکاران (۲۰۲۰)، دیویس (۲۰۲۱)، کلاد توریدا (۲۰۱۶)، بلوسو (۲۰۱۵)، وین (۲۰۰۵)، علایی و همکاران (۱۴۰۰) و دشتبانی و همکاران (۱۳۹۳) مراحل ساخت پیکره‌ها را به تفصیل بیان کرده‌اند. پژوهشگران دیگری مانند قیومی و همکاران (۲۰۱۳) و بی‌جن‌خان و همکاران (۲۰۱۱) نیز به بررسی چالش‌ها و مسائلی پرداخته‌اند که در مسیر ساخت پیکره‌های متنی فارسی مورد توجه است.

۲-۳- برخی از پیکره‌های زبانی فارسی

- پیکره همشهری (نسخه ۲) پیکره‌ای است حاوی ۳۱۸ هزار سند مربوط به اخبار سال‌های ۱۳۷۵ تا ۱۳۸۶ که با خزش (Crawl) وب‌سایت همشهری و چندین مرحله پیش‌پردازش و برجسب‌زنی حاصل آمده است. همه اسناد مجموعه همشهری دارای برجسب «Cat» هستند که نشان می‌دهد هر سند در چه رده‌ای است (اقتصادی، سیاسی و...). نسخه دوم پیکره همشهری توسط گروه تحقیقاتی پایگاه داده دانشگاه تهران و با حمایت مرکز تحقیقات مخابرات ایران تهیه شده است (آل‌احمد و همکاران: ۲۰۰۹).

- پیکره تطبیقی فارسی-انگلیسی دانشگاه تهران (UTPECC): این پیکره از دو مجموعه اخبار متفاوت از دو مبدأ مجزا یعنی اخبار فارسی خبرگزاری همشهری و اخبار انگلیسی خبرگزاری بی‌بی‌سی ساخته شده است. همچنین برای هم‌ترازی اسناد این دو زبان، علاوه بر تاریخ انتشار اخبار، شباهت محتوای اسناد نیز در نظر گرفته شده است. این پیکره در آزمایشگاه پژوهشی سیستم‌های هوشمند اطلاعات دانشگاه تهران تهیه شده است (برادران‌هاشمی و همکاران: ۲۰۱۰).
- پیکره متنی زبان فارسی: این پیکره، در واقع توسعه‌یافته «پیکره بی‌جن‌خان» است. پیکره بی‌جن‌خان، مجموعه‌ای است از متون فارسی شامل بیش از ۲ میلیون و ۶۰۰ هزار کلمه که با ۵۵۰ نوع برجسب اجزای واژگانی کلام، برجسب‌گذاری شده‌اند. این پیکره که در پژوهشکده پردازش هوشمند علائم تهیه شده است، همچنین شامل بیش از ۴۳۰۰ برجسب موضوعی چون سیاسی، تاریخی و ... برای متون است. پیکره متنی استاندارد زبان فارسی مجموعه‌ای از متون نوشتاری و گفتاری زبان فارسی به صورت رسمی است که از منابع واقعی همچون روزنامه‌ها، سایت‌ها و مستندات از قبل تایپ شده، جمع‌آوری شده، تصحیح گردیده و برجسب خورده است. حجم این دادگان حدوداً ۱۰۰ میلیون کلمه است و از منابع مختلف تهیه گردیده و دارای تنوعات بسیار زیادی است. پیکره متنی زبان فارسی دارای قابلیت‌ها و ویژگی‌هایی است که شامل: جمع‌آوری و سازمان‌دهی متون نوشتاری و گفتاری رسمی زبان فارسی با حجم ۱۰۰ میلیون کلمه. ویرایش نیمه‌خودکار اولیه متون. برجسب‌دهی نحوی-معنایی کلمات برای ۱۰ میلیون کلمه با استفاده از ۸۸۲ برجسب به صورت دستی توسط دانشجویان رشته زبان‌شناسی بر اساس دستورالعمل. تهیه نویسه‌های Unicode و XML برای پرونده‌های متنی دادگان. امکان برجسب‌دهی گروه‌های نحوی. طبقه‌بندی هر پرونده بر حسب موضوع و منبع آن. پوشش موضوعات مختلف سیاسی، اجتماعی، اقتصادی، فرهنگی، ... مجهز به یک نرم افزار آماری برای محاسبه و استخراج ویژگی‌های زبانی از قبیل: توزیع احتمالی مشروط، واژگان بسامدی، شناسایی هم‌نگاره‌ها، هم‌آیندها، مطابقه‌ها و ترتیب قاموسی با امکان گزارش‌گیری. طراحی یک زبان جستجوی هوشمند. اولین و تنها پیکره متنی حجیم در زبان فارسی، است (بی‌جن‌خان و همکاران: ۲۰۱۱).
- پیکره واژگان نحوی و معنایی افعال مرکب فارسی (نسخه ۱,۰): پیکره واژگان نحوی و معنایی افعال مرکب فارسی (PersPred) مجموعه‌ای است چندزبانه شامل اطلاعات نحوی و معنایی

افعال مرکب زبان فارسی، ترجمه انگلیسی و فرانسوی افعال، و حداقل یک جمله مثال برای هر فعل. اطلاعات نحوی بر اساس دیدگاه گروس و اطلاعات معنایی بر اساس دیدگاه لوین تهیه شده‌اند. نسخه اول این مجموعه که توسط پژوهشگران دانشگاه سوربن جدید فرانسه تهیه و عرضه شده شامل اطلاعات مربوط به بیش از ۶۰۰ فعل مرکب شامل همکرد «زدن» است (سامولین و فقیری: ۲۰۱۳).

- پیکره وابستگی نحوی زبان فارسی: این پیکره، مجموعه‌ای از ۳۰۰۰۰ جمله برچسب خورده با اطلاعات نحوی و ساخت‌واژی است. این پیکره که نخستین پیکره وابستگی زبان فارسی است، می‌تواند به عنوان زیرساختی اساسی در پردازش رایانه‌ای زبان فارسی به کار رود. همچنین اطلاعات این پیکره می‌تواند در پژوهش‌های زبان‌شناختی و آموزش و یادگیری زبان فارسی به کار رود. مهم‌ترین دلیل استفاده از دستور وابستگی در این پیکره نحوی نتایج رضایت‌بخش در یادگیری خودکار و سازگاری مناسب با طبیعت زبان‌های بی‌ترتیب، همچون زبان فارسی است. بی‌ترتیب بودن زبان فارسی بدین معناست که بیشتر جمله‌ها، ترتیبی ندارند. برای نمونه، جمله «من در مدرسه کتاب را به علی دادم» را به گونه‌های دیگر هم می‌توان گفت، مانند «من در مدرسه به علی کتاب را دادم»، «من به علی در مدرسه کتاب را دادم»، یا «من کتاب را به علی در مدرسه دادم». پیکره وابستگی نحوی زبان فارسی، جملات پیکره برگرفته از منابع گوناگون از متون فارسی معاصر را در بر دارد. همه جمله‌ها دارای برچسب روابط نحوی (بر مبنای دستور وابستگی) مانند فاعل، مفعول، مسند، مضاف‌الیه، بدل، و... هستند. همه جمله‌ها برچسب اطلاعات ساخت‌واژی (برچسب اجزای واژگانی کلام (POS)) مانند فعل، اسم، صفت، قید، ضمیر، و... دارند. جمله‌ها را تیمی از زبان‌شناسان باتجربه برچسب زده که در چند گام بازبینی شده‌اند. داده‌های پیکره به صورت تصادفی، به داده‌های یادگیری (۸۰٪)، آزمون (۱۰٪)، و ارزیابی (۱۰٪) تقسیم شده است. همه جمله‌ها ۲۹۹۸۲، همه واژه‌ها ۴۹۸۰۸۱، همه واژه‌های یکتا ۳۷۶۱۸، میانگین درازی هر جمله ۶۱/۱۶، شمار فعل‌های یکتا ۴۷۸۲، و میانگین حضور هر فعل ۶۷/۱۲ است (رسولی و همکاران: ۲۰۱۳).

- پیکره تطبیقی انگلیسی-فارسی دانشگاه الزهرا: منبع گردآوری داده‌ها در این پیکره، اسناد فارسی پیکره همشهری بوده است که شامل تعداد ۱۹۱/۴۴۰ سند از اخبار سال‌های ۱۳۸۱ تا ۱۳۸۵ است و برای اسناد انگلیسی، تعداد ۵/۳۶۹۷ سند از خزش در وبگاه خبری B.B.C در همان بازه زمانی

(۲۰۰۲ تا ۲۰۰۶) به دست آمد. پنج معیاری که برای دقت گردآوری این پیکره مورد توجه قرار گرفته است عبارتند از: برابری موضوعات دو سند تطبیقی، موضوعات مرتبط، تشابه ظاهری از نظر رویدادها، واژه‌های مشترک و عدم یکسان بودن اسناد تطبیقی است. از پیکره‌های چندزبانه می‌توان در حوزه مطالعات بازیابی اطلاعات استفاده نمود و کیفیت ترجمه ماشینی نیز به کیفیت این پیکره‌ها وابسته است (محمدی و ریاحی: ۲۰۱۶).

- پیکره فرهنگ جامع زبان فارسی: این پیکره در فرهنگستان زبان و ادب فارسی، زیر نظر دکتر علی اشرف صادقی در حال ساخت است. این پیکره واژه‌های عمومی زبان فارسی را دربردارد، هرچند واژه‌های تخصصی و علمی نیز در آن به‌وفور یافت می‌شود. برای اطمینان بیشتر از اینکه حتی‌الامکان هیچ واحد واژگانی‌ای از دست نرود، برای متون کهن و معاصر از هر قرن یا دوره تاریخی چند کتاب شاخص برگزیده شده و همه واحدهای واژگانی و شواهد آن‌ها استخراج شده‌اند. بدین‌سان، کوشش شده تا پیکره موجود همه واحدهای واژگانی این متون را دربرداشته باشد. برخی متون برجسته منظوم قدیمی به‌طور کامل در پیکره گنجانده شده‌اند، یعنی دست‌کم یک واحد واژگانی هر بیت گزینش شده تا همه ابیات این منظومه‌ها در پیکره وجود داشته باشند (شاهنامه، همه آثار منظوم سعدی، مولوی، دیوان حافظ و...). در مورد متون قدیم نیز تلاش شده است از همه انواع منظوم و منثور، از جمله متون حماسی، دینی، عرفانی، ادبی، علمی، سفرنامه‌ها، متون «عامیانه»، و در مورد منابع معاصر، از داستان‌ها، رمان‌ها، پژوهش‌های گوناگون، مقالات و سایر نوشته‌های مجله‌ها، روزنامه‌ها و وبگاه‌های اینترنتی و جز آنها مدخل‌گزینی شود (peykaregan.ir/dataset)

- واژگان زبانی فارسی: پیکره‌ای است شامل حدود ۵۵ هزار مدخل که هر مدخل دارای اطلاعات مربوط به صورت نوشتاری واژه در خط فارسی، ساخت‌واجی، مقوله واژگانی، الگوی تکیه و بسامد واژه می‌باشد. برای تهیه واژگان زبانی، یک پیکره متنی ۱۰ میلیون کلمه‌ای ملاک استخراج واژه‌ها قرار گرفته است. واژگان زبانی از حدود ۱۰۰ هزار کلمه با بسامدهای متفاوت تشکیل شده است. بعد از حذف صورت‌های تصریفی از فهرست فوق، حدود ۴۴ هزار واژه به مفهوم علمی آن به دست آمد. در بررسی فهرست واژه‌های حاصل از پیکره متنی معلوم شد که برخی واژه‌های عامیانه و برخی واژه‌های کاملاً علمی در فهرست ۴۴ هزار مدخلی غایب هستند. برای رفع این کاستی فهرست فوق با فرهنگ فارسی امروز (صدری‌افشار، ۱۳۸۱) مقایسه شد و

حدود ۱۱ هزار مدخل جدید به فهرست واژه‌ها اضافه شد و واژگان ۵۵ هزار مدخلی به دست آمد (اسلامی و همکاران: ۱۳۸۳).

– پایگاه دادگان زبان فارسی: این پیکره، مجموعه‌ای است برای ذخیره، پردازش، و ارائه داده‌های زبانی فارسی. این پایگاه دربرگیرنده پیکره‌هایی گوناگون در زبان فارسی است؛ پیکره‌های این پایگاه می‌توانند همواره روزآیند شوند. پایگاه دادگان زبان فارسی فراگیر و متنوع و در واقع فراتر از یک یا چند پیکره خاص است و کاربران بر پایه نیاز و هدف پژوهشی خود می‌توانند پیکره مناسب را از آن برگزینند. حتی پژوهندگان می‌توانند پیکره‌های اختصاصی خود را وارد پایگاه کنند و تحلیل‌ها و فهرست‌گیری‌های مورد نظر خود را انجام دهند. پایگاه دادگان زبان فارسی دارای متن‌های نشانه‌گذاری شده از جمله شناسنامه متن و برجسب‌های دستوری، آوایی، ریشه‌ای، و معنایی است. این دادگان مجهز به نرم‌افزارهای اختصاصی جست‌وجو، تقطیع، و تحلیل متن است که می‌تواند انواع فهرست‌های واژگانی، بسامدی، و آماری را ارائه کند. منابع به کار رفته در این پایگاه از گونه‌های نوشتاری با استفاده از متن‌های معتبر و با رعایت معیارهای مختلف نمونه‌گیری شده‌اند و البته هیچ‌گونه محدودیت و امساک‌های درباره آثار مهم ادبی و نویسندگان سرشناس و به ویژه صاحب سبک و تاثیرگذار اعمال نشده است. فهرست‌های مفصلی از همه منابع مهم نظم و نثر فارسی معاصر فراهم شده است. این فهرست‌ها جداگانه برای آثار شعری، داستانی، غیرداستانی، نمایشنامه و فیلمنامه، ادبیات کودکان، و نشریه‌های ادواری و مجلات علمی، تخصصی، و ادبی فراهم گردیده است. شمار آثاری که در این فهرست‌ها قرار گرفته‌اند، بیش از ۱۵۰۰ عنوان شده که پس از بررسی و کنار گذاشتن موارد مشابه، بیش از ۵۰۰ عنوان برای درونداد پایگاه داده‌ها برگزیده شده‌اند. نزدیک به ۴۵۰ اثر داستانی و غیرداستانی نثر؛ ۲۵۰ اثر شعری از شاعران معاصر؛ بیش از ۸۰ عنوان مجله و نشریه علمی، ادبی، و تخصصی؛ نزدیک به ۳۰۰ عنوان نمایش‌نامه و فیلم‌نامه؛ ۲۰۰ عنوان ادبیات کودک؛ چندین عنوان روزنامه و نشریه خبری؛ برخی از کتاب‌های درسی دانشگاهی و دبیرستانی؛ برخی از کتاب‌های دبستانی؛ نامه‌های اداری و بخشنامه‌ها؛ مجموعه کامل قوانین و مقررات؛ نشریه‌ها و جزوه‌های پراکنده؛ پوسترها؛ دیوارنوشته‌ها؛ و مانند این‌ها از جمله این متون هستند (عاصی: ۱۳۸۴).

– پیکره پژوهش‌نامه: این پیکره در پژوهشگاه علوم و فناوری اطلاعات ایران ساخته شده و دارای نزدیک به چهار میلیون و ۷۸۰ هزار واژه است. داده‌های پیکره از مقاله‌های پژوهش‌نامه پردازش و

مدیریت اطلاعات گرفته شده‌اند. درون‌مایه این پیکره همگانی نیست و دارای نوشته‌های بسیار تخصصی و میان‌رشته مانند علم اطلاعات و دانش‌شناسی، فناوری اطلاعات، مدیریت دانش، زبان‌شناسی رایانشی، اصطلاح‌شناسی، و مانند آن‌هاست. از این رو نیز، برای پردازش‌هایی که نیازمند بهره‌گیری از نوشته‌های تخصصی باشند، بسیار ارزشمند است (علایی و همکاران: ۱۴۰۰).

۲-۴- خلاصه فصل

در این فصل به بررسی پژوهش‌هایی پرداخته شد که در حوزه ساخت پیکره انجام شده است. در این پژوهش‌ها، چکیده‌ای از تجربیات پژوهشگران این حوزه درباره ساخت پیکره‌ها و چالش‌هایی که هنگام ساخت پیکره‌های متنی یا نوشتاری با آن‌ها مواجه شده‌اند، ارائه شده است. برخی دیگر از این پژوهش‌ها شامل توصیف انواع روش‌های ساخت پیکره، نوع پیکره ساخته شده و حاشیه‌نویسی‌های استفاده شده در پیکره است. در بخش انتهایی این فصل به معرفی برخی از پیکره‌های زبان فارسی پرداخته شد.

فصل سوم

نمونه‌گیری و وارد کردن داده‌ها در پیکره

۳-۱- مقدمه

در ساخت هر نوع پیکره‌ای، چه نوشتاری/متنی و چه گفتاری/آوایی، شاخص‌هایی کلی باید در نظر گرفته شود. شاخص‌های ساخت پیکره زبانی شامل این موارد است: نمونه‌گیری^۱، نمایندگی^۲/نماینده بودن حجم داده‌های زبانی، توازن^۳، اندازه^۴، نوع پیکره: عمومی^۵ یا تخصصی^۶، یک‌دستی^۷ (سینک‌لر: ۲۰۰۴). ویژگی این شاخص‌ها در شکل ۳-۱ نمایش داده شده است.

^۱ sampling

^۲ representativeness

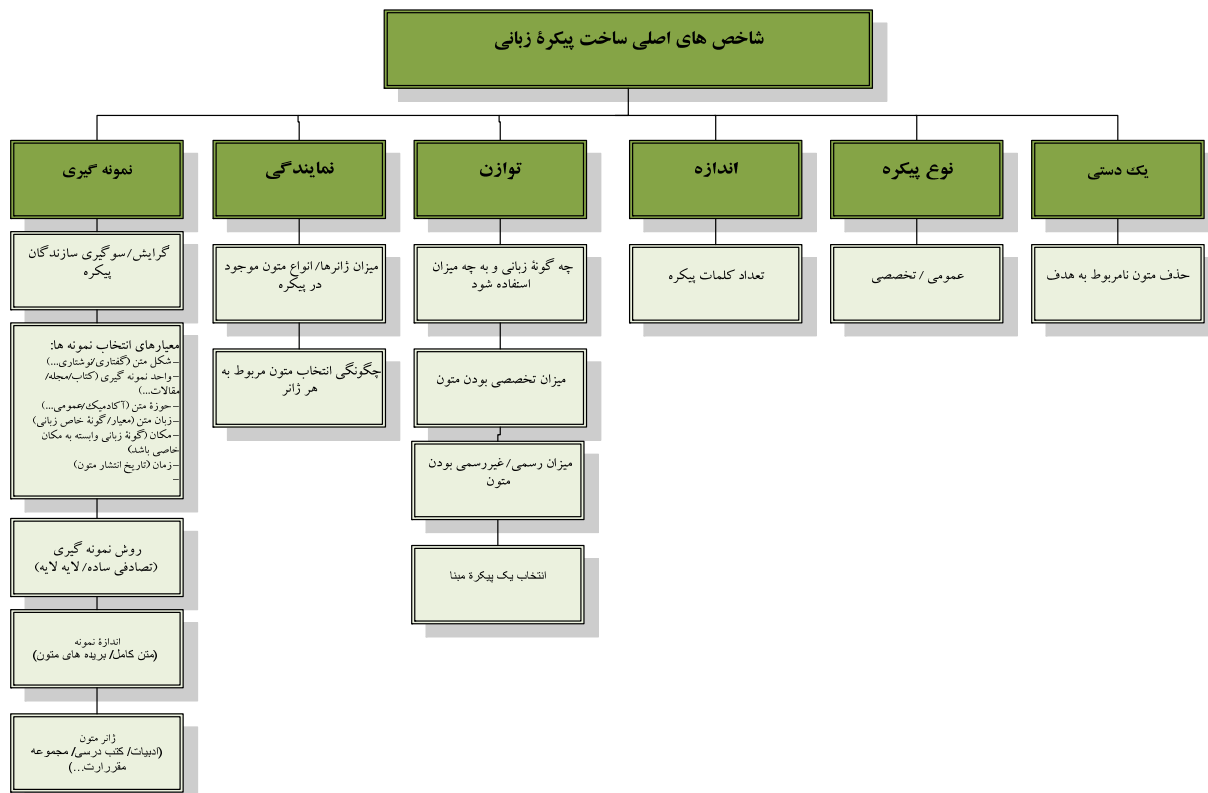
^۳ balance

^۴ size

^۵ general corpus

^۶ specialized corpus

^۷ homogeneity



شکل ۳-۱: شاخص‌های اصلی تهیه پیکره زبانی (برگرفته از علایی: ۱۳۹۸)

۳-۲- نمونه‌گیری

در طراحی کلی پیکره، ملاحظات از قبیل نوع متون مورد استفاده، تعداد متون، تعداد واژگان پیکره، توازن، نمایندگی و یکدستی وجود دارد که هر کدام مستلزم تصمیم‌گیری درباره نمونه‌گیری است. تصمیماتی که در این زمینه گرفته می‌شود بر مناسب بودن پیکره برای انواع تجزیه و تحلیل زبانی تاثیر می‌گذارد. نمونه‌گیری در واقع عمل انتخاب متون مربوط به هر ژانر با توجه به هدف تهیه پیکره است. برخی از معیارهایی که بر اساس آن‌ها نمونه‌گیری صورت می‌پذیرد شامل این موارد است: وجه متن^۱ (گفتاری/نوشتاری/الکترونیکی)، نوع متن (کتاب/مجله/نامه)، حوزه متن (آکادمیک/عمومی)، زبان متن (زبان‌ها یا گونه‌های زبانی پیکره) و مکان متن (به عنوان مثال، انگلیسی بریتانیا باشد یا استرالیا) است. داده‌های پیکره در این پژوهش، کتاب‌های دیجیتال پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است که

¹ mode

ساخت پیکره از کتاب‌های دیجیتال پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) □ ۳۱

شامل متون تخصصی میان‌رشته‌ای است. شکل متون: نوشتاری (الکترونیکی)، نوع متون: کتاب، حوزه متون: آکادمیک و گونه زبانی متون کتاب‌ها: گونه نوشتاری و رسمی است.

۳-۲-۱- جامعه آماری

پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) از سال ۱۳۴۸ تا سال ۱۴۰۰، تعداد ۲۸۱ کتاب منتشر کرده است. حوزه‌های موضوعی این کتاب‌ها بسیار متنوع است؛ عمدتاً زیرمجموعه حوزه کتابداری و اطلاع‌رسانی است و موضوعات زیر را شامل می‌شود (شایان ذکر است که برخی از کتاب‌ها با توجه به محتوای آن‌ها در بیش از یک گروه قرار گرفته‌اند و بیش از یک موضوع به آن‌ها اختصاص داده شده است):

ایران (۱۰۶ کتاب)، نشریات ادواری (۴۱ کتاب)، علوم اجتماعی (۴۱ کتاب)، نشریات ادواری فارسی (۳۴ کتاب)، کتابخانه‌ها (۳۰ کتاب)، کتابشناسی (۲۱ کتاب)، آمار (۱۹ کتاب)، همکاری‌ها (۱۷ کتاب)، فهرست‌ها (۱۶ کتاب)، ارزشیابی (۱۹ کتاب)، کتابخانه‌های دانشگاهی (۱۶ کتاب)، خدمات اطلاع‌رسانی (۱۲ کتاب)، علوم (۱۱ کتاب)، اصطلاح‌نامه‌ها (۱۰ کتاب)، مدیریت (۹ کتاب)، تحقیق (۱۰ کتاب)، سازمان‌های دولتی (۹ کتاب)، کنگره‌ها (۹ کتاب)، کتابداری (۸ کتاب)، اطلاع‌رسانی (۸ کتاب)، کتابخانه‌ها و مراجعان (۸ کتاب)، اصطلاح‌ها و تعبیرها (۸ کتاب)، انتشارات علمی (۸ کتاب)، سیاست علمی (۹ کتاب)، نوآوری (۹ کتاب)، استناد کتابشناختی (۱۰ کتاب)، راهنماها (۷ کتاب)، چکیده‌ها (۷ کتاب)، دانشگاه‌ها و مدارس عالی (۹ کتاب) و رتبه‌گذاری (۷ کتاب).

۳-۲-۲- نمونه‌گیری از جامعه آماری

با توجه به این نکته که هدف اصلی ساخت پیکره، فراهم کردن امکان بررسی واژه‌ها، عبارت‌ها و جمله‌ها در بافت است، نمی‌توان از همه کتاب‌های دیجیتال ایرانداک برای ساخت این پیکره استفاده کرد. زیرا محتوای بسیاری از این کتاب‌ها صرفاً متن نیست. به عنوان نمونه، مجموعه کتاب‌های «دانش ایران»، «ارزشیابی» «انتشارات علمی» و بخش عمده «استناد کتاب‌شناختی» بیشتر آمارهای فراگیر انتشارات علمی

ایران، مشارکت موسسه‌ها در انتشارات علمی، نقش ایران در پدیدآوری انتشارات علمی جهان و منطقه، نسبت انتشارات علمی به شاخص‌های پایه و تاثیر انتشارات علمی ایرانیان را ارائه می‌کند و بیشتر حجم کتاب دربرگیرنده اطلاعات آماری است. «نشریات ادواری/نشریات ادواری فارسی/علوم اجتماعی» نیز عمدتاً از فهرست مندرجات مجله‌های دیگر تشکیل شده‌اند که بیشتر شامل فهرست مندرجات مجله‌های حوزه علوم و علوم اجتماعی است. «آمار» و «همکاری‌ها» و «کتابخانه‌های دانشگاهی» نیز بیشتر چکیده وضعیت کتابخانه‌ها را بر پایه استاندارد NISO Z39.7 نشان می‌دهد؛ این اطلاعات به صورت آماری و نموداری ارائه شده است. همچنین «کتاب‌شناسی» شامل مجموعه‌هایی از مشخصات و مدخل کتاب‌ها، گزارش‌ها، پایان‌نامه‌ها و مقالات مجله‌ها و روزنامه‌ها است. مجموعه کتاب‌ها در گروه «فهرست‌ها» به ارائه فهرست انتشارات سازمان‌های دولتی، فهرست پایان‌نامه‌های فارغ‌التحصیلان ایرانی خارج از کشور یا فهرست برخی از پایان‌نامه‌های داخل کشور می‌پردازد و «اصطلاح‌نامه‌ها» و «اصطلاح‌ها و تعبیرها» مجموعه اصطلاحات یک رشته است که میان آن‌ها روابط معنایی، رده‌ای، و سلسله‌مراتبی برقرار شده و توانایی آن را دارد که موضوع آن رشته را با همه جنبه‌های اصلی و فرعی و وابسته، به گونه‌ای نظام‌یافته و به‌منظور ذخیره و بازیابی اطلاعات و مقاصد جنبی ارائه دهد. بنابراین، وارد کردن متون همه کتاب‌های دیجیتال ایرانداک در پیکره کاربردی نیست و باید از میان این کتاب‌ها، کتاب‌هایی را انتخاب نمود که عمدتاً متن هستند و نه گزارش‌های آماری و مجموعه واژگان (بدون آنکه در بافت زبانی باشند). در نهایت پس از بررسی محتوای کتاب‌های دیجیتال ایرانداک، در مرحله نمونه‌گیری، بسیاری از کتاب‌ها با موضوعات راهنما، گزارش، کتاب‌نامه، نمایه‌نامه، فهرست، آمار، کتاب‌شناسی، نشریات، همکاری، اصطلاح‌نامه و غیره از فهرست کتاب‌هایی که متن آن‌ها وارد پیکره خواهد شد، حذف شدند و این گونه متون در جامعه آماری پیکره قرار نگرفتند، زیرا در نمونه‌گیری شرکت داده نشدند و عضو قالب نمونه‌برداری^۱ نیستند.

نکته دیگری که در نمونه‌گیری باید مدنظر قرار گیرد، تاریخ انتشار کتاب‌ها است. از آنجائیکه فایل ورد (word) بسیاری از کتاب‌ها موجود نیست و تنها فایل پی‌دی‌اف (pdf) آن‌ها موجود است، ناچاریم با ابزار اُ.سی.آر (OCR) بسیاری از فایل‌های پی‌دی‌اف را به ورد تبدیل کنیم. در این میان به نظر می‌رسد فایل‌های قدیمی (مربوط به قبل از دهه ۷۰ شمسی) با توجه به نوع قلم استفاده شده در آن‌ها، قابل تبدیل به فایل ورد (word) با صحت بالا نباشند، بنابراین، این فایل‌ها نیز در نمونه‌گیری انتخاب نمی‌شوند.

^۱ sampling frame

با در نظر گرفتن ملاحظات فوق، از ۲۸۱ کتاب دیجیتال ایرانداک (پیوست ۱)، ۶۵ (۳۱ فایل ورد (word) و ۳۴ فایل پی.دی.اف (pdf)) کتاب برای وارد شدن در پیکره انتخاب شدند. شایان ذکر است که کتاب‌ها از نظر تکراری نبودن بررسی شدند؛ بنابراین، هیچ فایلی دو بار انتخاب نشده است. فهرست کتاب‌ها به شرح زیر است:

- بازی‌وارسازی، مفاهیم و کاربردها
- ساخت‌واژه، اصطلاح‌شناسی و مهندسی دانش
- فرایند داوری در مجلات علمی؛ نقاط قوت و ضعف آن
- جهانی‌شدن و مدل توسعه نظام آموزش عالی در ایران با رویکرد فناوری اطلاعات و ارتباطات
- دورکاری در سازمان‌ها و مراکز اطلاع‌رسانی
- آشنایی با حق مؤلف؛ مالکیت فکری در عصر دیجیتال
- کتابخانه‌های دیجیتال: یکپارچه‌سازی محتوا و سیستم‌ها
- داوری و هم‌ترازخوانی: مبانی، شیوه‌ها و رهنمودهای اخلاقی
- سواد دیداری برای متخصصان علم اطلاعات
- انقلاب جدید کتاب در محیط الکترونیکی
- اصول طرح تحقیق و روش‌شناسی
- سیستم اطلاعات مکانی (جی‌آی‌اس) و کاربردهای آن در علم اطلاعات و دانش‌شناسی
- مبانی اطلافرینی؛ کارآفرینی با اطلاعات و در اطلاع‌رسانی
- سیستم‌های اطلاعات مدیریت
- ارزیابی علم، فناوری و نوآوری: مروری بر شاخص‌ها و سازمان‌های فعال این حوزه
- راهنمای نگارش گزارش‌های علمی و فنی
- شیوه‌نامه ایران (۱۳۹۵)
- شیوه‌نامه ایران (۱۳۹۳)
- سنجه‌های وبی برای متخصصان کتابداری و اطلاع‌رسانی
- سنجش پژوهش‌های دانشگاهی: چگونگی انجام یک مطالعه کتاب‌سنجی

- آواشناسی
- آینده‌نگاری علم و فناوری و اثرات آن در سیاستگذاری
- تجاری‌سازی نتایج پژوهش: الگوها، سازوکارها و راهبردها
- سنجش و مصورسازی پراکندگی جغرافیایی ثروت علمی مفاهیم و کاربردها
- دگرسنگ‌ها برای متخصصان اطلاعات؛ گذشته، حال، آینده
- زبان مخفی چیست؟
- دستاوردهای پنجمین همایش ملی مدیران فناوری اطلاعات
- دستاوردهای هفتمین همایش ملی مدیران فناوری اطلاعات
- دستاوردهای چهارمین همایش ملی مدیران فناوری اطلاعات
- دستاوردهای سومین همایش ملی مدیران فناوری اطلاعات
- دستاوردهای ششمین همایش ملی مدیران فناوری اطلاعات
- مدل‌ها و روش‌های مدیریت کیفیت داده: نگرشی کاربردی بر داده‌های علم و فناوری
- نیازسنجی اطلاعات
- مدیریت اطلاعات (مجموعه دوم)
- اشتراک منابع و همکاری بین کتابخانه‌ها
- انتخاب منابع اطلاعاتی
- آواشناسی آکوستیک و شنیداری
- برنامه‌ریزی استراتژیک برای خدمات کتابخانه‌ای و اطلاعاتی
- برنامه‌ریزی حوادث در خدمات اطلاع‌رسانی و کتابداری
- بخش پوستر در کنفرانس‌ها
- تحصیلات تکمیلی علوم کتابداری و اطلاع‌رسانی در انگلستان، آمریکا، هند و ایران
- تهیه راهنما برای خدمات کتابداری و اطلاع‌رسانی
- تبلیغ کارآمد سایت وب
- دورکاری برای کتابداران و اطلاع‌رسانان
- دورکاری؛ مفاهیم و فرآیند برنامه‌ریزی

- رهنمودهایی برای سیاست‌های اصطلاح‌شناسی؛ تدوین و اجرای سیاست اصطلاح‌شناسی در جوامع زبانی
- ریشه‌های فرهنگی ارتباط در ایران
- سواد اطلاعاتی
- عوامل موفقیت همکاری
- فرآیند تدوین اصطلاح‌نامه
- فناوری اطلاعات و کتابخانه‌ها
- کاربرد سامانه اطلاعات جغرافیایی (جی‌آی‌اس) در خدمات کتابداری و اطلاع‌رسانی
- کاربرد سیستم‌های اطلاعات جغرافیایی (جی‌آی‌اس) در پژوهش‌های سازمانی
- مدیریت تغییر در کتابخانه‌ها و خدمات اطلاعاتی
- مدیریت کمیته‌ها در سازمان
- مدیریت مجموعه‌های فیلم و ویدیو
- مبانی نظام‌ها و خدمات اطلاعاتی
- مجموعه مقالات نخستین کنفرانس ملی کتابخانه دیجیتال
- مجموعه مقالات همایش ملی هر کتابخانه یک رصدخانه
- مجموعه‌سازی مشارکتی؛ راهنمای عملی برای کتابخانه شما
- مدل تعالی موسسه‌های پژوهشی
- مدیریت ارتباطات علمی
- مدیریت خودکارسازی کتابخانه‌ها
- مدیریت نظام علوم، تحقیقات و فناوری: پرسش‌های اساسی
- مقدمه‌ای بر یادگیری الکترونیکی
- نشر الکترونیکی

در یک فایل اکسل (excel)، فراداده کتاب‌های انتخاب شده به همراه نام فایل (شماره فایل) و نوع فایل (ورد یا پی.دی.اف) وارد شد تا بتوان در پیکره وارد کرد. فراداده^۱، به داده‌هایی گفته می‌شود که

^۱ metadata

جزئیات یک داده را تشریح می‌کند. به عبارت دیگر فراداده‌ها، داده‌هایی هستند درباره داده‌های دیگر. برخی این واژه را آبر داده یا فوق داده هم ترجمه کرده‌اند.

در کتابداری، برگه‌های داده‌نما معمولاً شامل اطلاعاتی همچون نام نویسنده کتاب، تاریخ چاپ، تعداد صفحات، شمارگان، زبان کتاب، نام مترجم، کلیدواژه‌ها و غیره هستند. هدف از داده‌نمایی، دستیابی آسان‌تر و بهتر به داده‌های موجود است. اساس کاربرد فراداده، کاوش، بازیابی، دسترسی، کشف، مستندسازی، ارزیابی، و انتخاب منابع الکترونیکی به ویژه منابع شبکه‌ای است که باعث افزایش دقت در بازیابی می‌گردد. مهم‌ترین کاربردهای فراداده شامل موارد زیر است:

- فراهم نمودن امکان نمایه‌سازی طیف گسترده‌ای از نوع منابع اطلاعاتی شبکه‌ای
- کمک به کشف و بازیابی مؤثر اطلاعات
- امکان‌پذیر ساختن تشخیص مطابقت منابع بازیابی شده با نیاز کاربران
- کنترل اطلاعات محدود از لحاظ دسترسی و جلوگیری از دسترسی برخی کاربران به داده‌ها (مانند کودکان)
- ارائه تاریخ داده‌های تأثیرگذار مانند: شرایط قانونی، سن، اندازه و مانند آن
- ارائه تاریخ داده‌ها مانند: منبع اصلی و هرگونه تغییر شکل بعدی
- ارائه اطلاعاتی در مورد مالک یا پدیدآورنده منبع مانند نشانی رایانامه
- نشان دادن روابط منابع با دیگر منابع مانند: پیوندهایی به نسخه‌های قبلی و بعدی
- کمک به تصمیم‌گیری در مورد اینکه چه شکل و قالبی از داده‌ها باید بازیابی شوند (در صورت وجود قالب‌های متعدد)
- تطبیق، اشتراک و یکپارچه‌سازی منابع اطلاعاتی ناهمگن شبکه‌ای
- کنترل مدیریت داده‌ها

در طرح پژوهشی حاضر به منظور رعایت یکدستی در بازیابی اطلاعات موجود در پیکره‌ها، نوع فراداده‌ها همان‌هایی بود که در پیکره پژوهش‌نامه (طرح قبلی نگارنده) انتخاب شده بودند؛ بنابراین،

ساخت پیکره از کتاب‌های دیجیتال پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) □ ۳۷

«عنوان کتاب»، «موضوع» و «پدیدآور/مترجم» به عنوان فراداده وارد شدند. قسمتی از این فایل برای نمونه آورده شده است:

ردیف	عنوان کتاب	موضوع	پدیدآور/مترجم	فایل	نوع فایل	1
2۴	۲۵	دگرمنجه‌ها برای متخصصان اطلاعات؛ گذشته، حال، آینده	کتابداری و اطلاع‌رسانی	کیم هولبرگه، ترجمه: مه‌ری صدیقی	۲۵	word
2۵	۲۶	زبان مخفی چیست	زبان‌شناسی	ژان لویی کالود، ترجمه: شهرزاد پزشکی	۲۶	word
2۶	۲۷	دستاوردهای پنجمین همایش ملی مدیران فناوری اطلاعات	فناوری اطلاعات	محمدجواد ارشادی، آرمان ساجدی نژاد، مرفیبه زرین‌بال ماسوله، علیرضا نقه‌الاسلامی، علی نعیمی صدیق، آزاده محبی، احمد گنجی، آذین رحمتی، ندا حیدریان‌فرد، فاطمه نعیمی دارستانی	۲۷	word
2۷	۲۸	دستاوردهای هفتمین همایش ملی مدیران فناوری اطلاعات	فناوری اطلاعات	علی نعیمی صدیق، محمد ربیعی، آرمان ساجدی نژاد، سمیه لبافی	۲۸	word
2۸	۲۹	دستاوردهای چهارمین همایش ملی مدیران فناوری اطلاعات	فناوری اطلاعات	علیرضا نقه‌الاسلامی، بهزاد دوران، مرفیبه زرین‌بال ماسوله، آزاده محبی، آزاده نامداریان، علی نعیمی صدیق، احمد گنجی، آذین رحمتی، ندا حیدریان‌فرد، اصغر بختیار زاده	۲۹	word
2۹	۳۰	دستاوردهای سومین همایش ملی مدیران فناوری اطلاعات مراکز علمی کشور	فناوری اطلاعات	آذین رحمتی، آزاده محبی، احمد گنجی، ندا حیدریان‌فرد	۳۰	word
3۰	۳۱	مدل‌ها و روش‌های مدیریت کیفیت داده: تگرشی کاربردی بر داده‌های علم و فناوری	مدیریت اطلاعات	محمدجواد ارشادی، محمد مهدی ارشادی	۳۱	word
3۱	۳۲	نیازسنجی اطلاعات	کتابداری و اطلاع‌رسانی	محمود بابائی	۳۲	pdf
3۲	۳۳	مدیریت اطلاعات؛ مجموعه دوم	مدیریت اطلاعات	محمد تقی مهدوی	۳۳	pdf

در مرحله بعد باید فایل‌های پی.دی.اف به ورد تبدیل شوند و در نهایت، همه فایل‌های ورد وارد پیکره می‌شوند.

۳-۳ به کار بردن ابزار اُ.سی.آر (OCR)

نرم‌افزار اُ.سی.آر (OCR)^۱ که معادل فارسی آن «نویسه خوان نوری» است، نوعی فناوری است که می‌تواند انواع مختلف اسناد (مانند اسناد اسکن شده و فایل‌های پی.دی.اف) را به داده‌های قابل ویرایش و جست‌وجو تبدیل کند. اُ.سی.آر ابتدا تنها در مورد بازشناسی ارقام و حروف چاپی بکار گرفته می‌شد.

^۱ optical character recognition (OCR)

پسوند «نوری» در این عبارت در مقابل عبارت «مرکب مغناطیسی»^۱ قرار داده شد تا این روش از روش قدیمی‌تر بازشناسی نویسه‌ها با مرکب مغناطیسی^۲ (MICR)، متمایز شود. با گذشت زمان و پیشرفت قابل توجه در این زمینه، روش‌های بازشناسی دست‌نوشته و متون چاپی مطرح شدند که دامنه کار را به واژه‌ها و عبارت‌ها گسترش دادند. با وجود عدم تطبیق دقیق اُ.سی.آر با این موارد، این نام برای این روش‌ها هم استفاده شد و رواج پیدا کرد. هم‌اکنون اُ.سی.آر را بیشتر برای بازشناسی مستندات چاپی مثل صفحات کتاب‌ها، مجله‌ها و نامه‌های چاپی به کار می‌برند. پیش از اُ.سی.آر، برای تغییر محتوای اسناد، مجبور بودیم آن‌ها را مجدداً تایپ کنیم. اما امروزه با استفاده از نرم‌افزار اُ.سی.آر، دیگر نیازی به تایپ مجدد متون نیست و به راحتی می‌توان اسناد اسکن شده را به فایل‌های قابل ویرایش تبدیل کرد. در این فرآیند تصاویر مربوط به یک متن یا سند از طریق تصویربرداری دیجیتال با دوربین دیجیتال‌های معمولی یا اسکنر گرفته شده و به رایانه وارد می‌شود؛ سپس، این تصاویر به عنوان فایل‌های گرافیکی مورد استفاده قرار می‌گیرند. طی این فرآیند تصویر متن که بر پایه فرمت‌های تصویری است به فایل متنی، مانند txt یا docs و هر فرمت دیگری، می‌تواند تبدیل شود. از مهم‌ترین مزایای استفاده از این نرم‌افزار می‌توان به کاهش خطا هنگام تایپ، صرفه‌جویی در زمان و جست‌وجو در متون اشاره کرد. همچنین این ابزار در نرم‌افزارهای تبدیل دست خط به نوشته‌های تایپی نیز به کار می‌رود. ابزارهای اُ.سی.آر مانند بسیاری از ابزارهای هوشمند دیگر، پیچیدگی‌های زیادی دارد. پردازش تصویر و بازشناسی الگو دو پایه اصلی این ابزارها هستند. پیچیدگی این ابزارها برای زبان‌های گوناگون، متفاوت است. به عنوان نمونه نوشتن اُ.سی.آر برای زبان‌های لاتین به دلیل اینکه حروف آن‌ها به طور مجزا نوشته می‌شود آسان‌تر از زبان‌هایی مانند فارسی و عربی است.

همانگونه که ذکر شد، در مرحله نمونه‌گیری، بسیاری از کتاب‌های انتخاب شده برای پیکره، به صورت فایل پی.دی.اف بودند و لازم بود از طریق ابزار اُ.سی.آر به فایل‌های ورد تبدیل شوند. برای این منظور از کتابخانه IronOCR در دات نت استفاده شده است. یک برنامه در زبان سی‌شارپ تهیه شد که فولدر فایل‌های پی.دی.اف کتاب‌ها را گرفته و با کمک بسته IronOCR به همراه کتابخانه زبان فارسی یعنی IronOCR.Languages.Persian عملیات نویسه‌خوانی را انجام می‌دهد. با توجه به آنکه این کتابخانه در دو نسخه تجاری (پولی) و نسخه تبلیغاتی موجود است و امکان تهیه نسخه تجاری فراهم نبود،

^۱ magnetic ink

^۲ Magnetic ink character recognition (MICR)

لذا از نسخه تبلیغاتی آن استفاده شد. این نسخه تمامی قابلیت های نسخه تجاری را دارد و تنها محدودیت آن این است که فایل پی.دی.اف باید کمتر از ۱۰ صفحه باشد. برای رفع این محدودیت یک برنامه جانبی در دات نت تهیه شد که فایل های پی.دی.اف را گرفته و آن ها را به مجموعه های ۱۰ صفحه ای تبدیل می کند. سپس، در گام بعد این فایل ها با کمک کتابخانه IronOCR به متن تبدیل می شوند.

۳-۴- تمیز و مرتب کردن متن خوانده شده

در مجموعه فایل های مورد نظر برای این طرح که تحت عنوان کتاب های پژوهشگاه در اختیار تیم توسعه قرار گرفته بود با محتوای غیرمتنی نیز مواجه شدیم؛ به عبارتی، محتوای آن ها تصاویری از نمایشگاه ها و اطلاعاتی بودند که مناسب OCR نبودند؛ در حقیقت، برنامه OCR متون انگلیسی و یا فارسی موجود در تصاویر را هم استخراج کرده بود و این موضوع باعث آشفتگی در واژه ها و اضافه شدن برخی کلمات به متون شده بود. همچنین برخی مدارک از فونت های بسیار قدیمی و یا برخی کتاب ها شامل نویز زیاد بودند که باعث خرابی در داده ها شده بود. برای رفع این موارد عملیات بسیار متنوعی صورت گرفت که عبارتند از: ۱- استخراج الگو، با بررسی متون بدست آمده برخی اشتباهات دارای الگوی مشخصی بودند که این موارد با عملیات Replace بصورت یکجا تصحیح شدند. ۲- استفاده از ابزار «پاکنویس» در ms-word: این افزونه محصول مرکز تحقیقات کامپیوتری علوم اسلامی است و بصورت خودکار با کمک پایگاه داده، واژه ها را تا حد امکان تصحیح می کند. ۳- تصحیح داده ها بصورت دستی. همانگونه که اشاره شده، یکی از مهم ترین مسائل تغییراتی بود که در ظاهر واژه ها ایجاد می شد، برخی از این موارد در جدول ۳-۱ قابل مشاهده است:

جدول ۳-۱: مشکلات پس از ا.سی.آر:

مسائل نگارشی واژه ها پس از ا.سی.آر	مثال
۱- به هم چسبیدن واژه ها	- این مراکز اغلب (این مراکز اغلب)

- الکترونیک دولت (الکترونیک دولت) - تابا استفاده از تسهیلاتی (تا با استفاده از تسهیلاتی) - تختها ککن (تخته پاک کن) - زبان کنترل شده (زبان کنترل شده) - سالها بیر نامهریزی (سال های برنامه ریزی)	
- طلاع (اطلاع) - کیده نویسی (چکیده نویسی) - اختیار (اختیار) - طلاح (اصطلاح) - نعطاف (انعطاف) - یف (توصیف) - فجه بندی (صفحه بندی) - فحه (صفحه) - سپتام (سپتامبر)	۲- ناقص شدن واژه در اثر حذف بخش ابتدایی (یا گاهی انتهایی) واژه
- خدمات (خدمات)	۳- درج حرف «ا» یا «و» (یا هر دو) در میان واژه

- زیار (زیر) - زیسات (زیست) - عصور (عصر) - علاوم (علوم) - علمای (علمی) - نشوان (نشان) - نشوده (نشده) - جملاه (جمله) - نظورات (نظرات) - نورمافزارهای (نرم‌افزارهای) - نورمافوزار (نرم‌افزار) - نمونوهگیوری (نمونه‌گیری) - نمویکننود (نمی‌کنند)	
- تتپایهگذاری (پایه‌گذاری) - هخصیصه‌های (خصیصه‌های)	۴- تکرار حروف یا اضافه شدن برخی از حروف در ابتدای واژه

- سسجمع‌بندی (جمع‌بندی) - سسبرجسته (برجسته)	
- تپین (تعیین) - تپداد (تعداد) - فپالسازی (فعال‌سازی) - فپالیوت (فعالیت) - فپیلی (فعلی)	۵- درج حرف «پ» به جای حرف «ع»
- تشکیل‌دهنده‌ی (تشکیل‌دهنده) - نکته‌یابی (نکته‌یابی) - نکته‌سنجی (نکته‌سنجی) - نیمه‌مترکز (نیمه‌مترکز) - شاخه‌ای (شاخه‌ای)	۶- چسبیدن حرف «ه» آخر واژه به واژه یا وند بعد از خود
- نشسته‌اید (نشسته‌اید) - نوشته‌اش (نوشته‌اش) - نکرده‌است (نکرده‌است)	۷- چسبیدن شناسه‌ها به ستاک

۸- چسبیدن «آ» به واژه یا وند قبل از خود، بدون نیم- فاصله	- نیم‌آیم (نمی‌آیم) - نیم‌آموزد (نمی‌آموزد) - نهایتاً‌مادگی (نهایى آمادگی) - سرسام‌آور (سرسام‌آور)
---	---

در نهایت با صرف زمان بسیار، مجموعه متون مورد بازنگری قرار گرفت و موارد مشکل‌دار، که تعداد آنها بسیار بود، در چندین مرحله اصلاح شدند.

۳-۵- خلاصه فصل

در این فصل ابتدا به چگونگی فرآیند نمونه‌گیری پرداخته شد. سپس، تهیه فراداده‌ها توضیح داده شد. در ادامه به فرآیند تبدیل فایل‌های پی.دی.اف به ورد از طریق استفاده از ابزار اُ.سی.آر پرداخته شد و در نهایت چالش‌های این روش و چگونگی تمیز و مرتب کردن فایل خروجی توضیح داده شد.

فصل چهارم

هنجارسازی و نشانه‌گذاری داده‌های پیکره

۴-۱- مقدمه

در این فصل ابتدا به تعریف «هنجارسازی» و چگونگی هنجارسازی داده‌های وارد شده در پیکره پرداخته می‌شود. سپس مفهوم «حاشیه‌نویسی» توضیح داده خواهد شد و در نهایت، نوع حاشیه-نویسی/برچسب‌گذاری استفاده شده در پیکره و ابزار استفاده شده برای این منظور، معرفی می‌گردد و نحوه کنترل دستی برچسب‌ها توضیح داده می‌شود.

۴-۲- هنجارسازی داده‌ها

پیش از وارد کردن داده‌ها در پیکره و برچسب‌گذاری، لازم است متون پیکره از نظر یک‌دست بودن وضعیت نگارشی آن‌ها، کنترل شوند. این کار هنجارسازی^۱ نامیده می‌شود. در حقیقت، هنجارسازی اینگونه تعریف می‌شود: یک‌دست‌سازی واحدهای متنی به طوری که قابل پردازش توسط ماشین باشند.

چگونگی پردازش متون فارسی با پردازش متون انگلیسی متفاوت است. در زبان انگلیسی حروف جدا جدا نوشته می‌شوند و هیچ حرفی به حرف دیگر نمی‌چسبد؛ در حالیکه در زبان فارسی بسیاری از وندها به ستاک متصل می‌شوند و چگونگی اتصال وندها به ستاک، می‌تواند به صورت با فاصله، نیم‌فاصله

^۱ normalization

یا حتی بدون فاصله باشد. به عنوان نمونه، هر سه حالت «کتاب‌ها»، «کتاب‌ها» و «کتابها»، صورت نوشتاری درستی در نظر گرفته می‌شوند. در پردازش متون، چنانچه حروف، نشانه‌های نگارشی و واژه‌ها به شکل یکسانی نوشته نشده باشند، پردازش متون به درستی انجام نخواهد شد و در بازیابی اطلاعات به نتایج درستی نخواهیم رسید. از این رو پیش از انجام هر گونه پردازش زبانی و ساخت پیکره، لازم است متون هنجارسازی شوند که هنجارسازی در متون فارسی می‌تواند شامل این موارد باشد: یک‌دست کردن فاصله‌ها و نشانه‌گذاری‌های درون متن، یکسان کردن یونیکد کاراکترهای استفاده شده در متون (مانند انواع «ی»، «ک»، «همزه و...»، یکسان کردن روش اتصال وندهای گوناگون به ستاک، اصلاح غلط‌های املایی، ارتباط دادن کلمات چنداملایی و یکسان در نظر گرفتن آن‌ها و غیره.

برای هنجارسازی داده‌ها ابتدا از یکی از ابزارهای آماده استفاده شد: ابزار «هضم». هضم ابزاری است برای پردازش زبان فارسی در پایتون (sobhe/hazm). ویژگی‌های این ابزار شامل موارد زیر است:

- تمیز و مرتب کردن متن
- تقطیع جمله‌ها و واژه‌ها
- ریشه‌یابی واژه‌ها
- تحلیل صرفی جمله
- تجزیه نحوی جمله
- واسط استفاده از داده‌های زبان فارسی
- سازگاری با بسته NLTK
- پشتیبانی از پایتون نسخه ۲ و ۳
- آزمون مداوم کدها

از کدهای ابزار «هضم» که به رایگان در اختیار کاربران قرار گرفته است، کد واحدسازی/جداسازی/توکنایزر (tokenizer) در پیکره استفاده شده است. ابزار واحدسازی/جداسازی، مرز واژه‌ها را در متون تشخیص می‌دهد و متن را به دنباله‌ای از واژه‌ها تبدیل می‌کند و آن را برای تحلیل-های بعدی آماده می‌کند. در حقیقت، واحدسازی، تکه‌تکه متن به قسمت‌های کوچکی به نام واحد است. واحدسازی در سطح واژه رخ می‌دهد و واحدهای استخراج شده به عنوان ورودی پیمانه‌های دیگر مانند

ریشه‌یاب، برچسب‌گذاری و غیره استفاده می‌شوند. بنابراین، بخشی از پیش‌پردازش متن، واحدسازی/جداسازی است.

برای هنجارسازی، برخی موارد بصورت خودکار توسط ابزار «هضم» انجام شد. این موارد عبارتند از حروف عربی «ی» و «ک» و برخی موارد مانند نیم‌فاصله که با دو یونیکد مختلف بودند و حرف «ه» بصورت کدهای افزوده در فاز هنجارسازی به کار گرفته شدند. برای یکدست کردن انواع «ی» و «ک» و «کسره اضافه روی «ه» / «ه» در حالت اضافی («ه»)، از دستورات TSQL در برنامه پایگاه داده SQL Server نیز استفاده شده است (جدول ۴-۱).

جدول ۴-۱: یونیکدهای «ی»، «ک» و «ه»

Arabic letter Unicode	U+1610 for <ی>	U+1603 for <ک>	U+1577 for <ه>
Persian letter Unicode	U+1740 for <ی>	U+1705 for <ک>	U+1607 for <ه>

بررسی داده‌ها نشان داد که علیرغم اصلاحاتی که در مورد نیم‌فاصله انجام شد، کماکان بسیاری از وندها جدا از ستاک بودند که این مسئله قطعاً مشکلاتی را در مرحله برچسب‌گذاری ایجاد خواهد کرد. برای اجتناب از این مشکل، از اطلاعات ساختوازی فارسی (انواع وند) استفاده شد. این فهرست برگرفته از مجموعه ۱۲ مقاله دکتر علی‌اشرف صادقی (شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر ۱ تا ۱۲: ۱۳۷۰-۱۳۷۲)؛ «اشتقاق پسوندی در زبان فارسی/امروز» اثر کشانی (۱۳۷۱)، «دستور زبان فارسی معاصر» اثر لازار (۱۳۸۹) و «مشخصه‌های تصریفی در زبان فارسی/امروز» اثر قطره (۱۳۸۶) است. این اطلاعات در جدول ۴-۲ ارائه شده است.

جدول ۴-۲: فهرست وندها و واژه‌بست‌های فارسی (برگرفته از علایی و همکاران: ۱۴۰۰)

وندها	نوع وند	مثال
ها	پسوند تصریفی (شمار)	کتاب‌ها
ان	پسوند تصریفی (شمار)	درختان
ین	پسوند تصریفی (شمار)	معلمین

ات	پسونند تصریفی (شمار)	درجات
ی	پسونند تصریفی (یای نکره)	کتابی (در عبارت «کتابی خوب»)
ه	پسونند تصریفی (معرفه)	دختره
ش	پسونند تصریفی (معرفه)	آشش (در جمله «آشش خوبه»)
تر/ترین	پسونند تصریفی (نشانه صفت تفصیلی و عالی)	خوب‌ترین/بهتر
ت/اد/د/ید/ست	پسونند تصریفی (نشانه ماضی)	کشت/ایستاد/خورد/رسید/آراست
ان	پسونند تصریفی (متعدی ساز)	رساند (رس+ان+د)
می/ب/ان/م	پیشوند تصریفی	می‌گفت/بروم/نرو/مشنو
با	پیشوند اشتقاقی (صفت ساز)	باادب
بی	پیشوند اشتقاقی (صفت ساز)	بیکار
نا	پیشوند اشتقاقی (صفت ساز)	ناکام
هم	پیشوند اشتقاقی (صفت ساز)	همکار
ن	پیشوند اشتقاقی (صفت ساز)	نشکن
ام/ای/ه/س/ست//ایم/اید/اند (صورت‌های وابسته فعل بودن)	واژه‌بست	زنده‌ام/زنده‌ای/زنده‌ست/زنده‌ایم/زنده-اید/زنده‌اند
«ی»/نکره/«ی» تخصیص نکره	واژه‌بست	«کتابی» در جمله «کتابی خریدم»
و (صورت پی‌بستی «را»)	واژه‌بست	«و» در جمله «منو می‌بری؟» (من را می‌بری؟)
و (حرف هم‌پایگی یا عطف)	واژه‌بست	«و» در «من و تو»
م (صورت پی‌بستی «هم»)	واژه‌بست	«م» در «منم میام» (من هم می‌آیم)
ها و صورت محاوره آن (آ) (برای تاکید بعد از فعل می‌آید)	واژه‌بست	او مدم آ
م/ت/ش/مان/تان/شان (ضمایر پی-بستی/پیوسته)	واژه‌بست	مدادم/مدادت/مدادمان/مدادتان/مدادشان
گاه	پسونند اشتقاقی (اسم مکان)	دانشگاه
ستان	پسونند اشتقاقی (اسم مکان)	کردستان
کده	پسونند اشتقاقی (اسم مکان)	دانشکده
ئیه	پسونند اشتقاقی (اسم مکان)	جمشیدیه

ک	پسونند اشتقاقی (اسم مکان)	قلهک
دان/دانی	پسونند اشتقاقی (اسم مکان)	نمکدان / سگ‌دانی
زار	پسونند اشتقاقی (اسم مکان)	کشتزار
آباد	پسونند اشتقاقی (اسم مکان)	حسن آباد
ی	پسونند اشتقاقی (اسم مکان)	پنچری
سار	پسونند اشتقاقی (اسم مکان)	کوهسار
بار	پسونند اشتقاقی (اسم مکان)	جویبار
لاخ	پسونند اشتقاقی (اسم مکان)	سنگ‌لاخ
ان	پسونند اشتقاقی (اسم مکان)	دیلمان
بان	پسونند اشتقاقی (دال بر شغل، محافظت یا دارندگی)	نگهبان
چی	پسونند اشتقاقی (دال بر شغل، محافظت یا دارندگی)	کافه‌چی
دار	پسونند اشتقاقی (دال بر شغل، محافظت یا دارندگی)	صندوق‌دار
ی	پسونند اشتقاقی (دال بر شغل، محافظت یا دارندگی)	بازاری
یار	پسونند اشتقاقی (دال بر شغل، محافظت یا دارندگی)	استادیار
گر	پسونند اشتقاقی (دال بر شغل، محافظت یا دارندگی)	آهنگر
ساز	پسونند اشتقاقی (دال بر شغل، محافظت یا دارندگی) / پسونندواره فعلی	آهنگ‌ساز
دوز	پسونند اشتقاقی (دال بر شغل، محافظت یا دارندگی) / پسونندواره فعلی	کفش‌دوز
باف	پسونند اشتقاقی (دال بر شغل، محافظت یا دارندگی) / پسونندواره فعلی	فرش‌باف

پز	پسوند اشتقاقی (دال بر شغل، محافظت یا دارندگی) / پسوندواره فعلی	شیرینی‌پز
ه	پسوند اشتقاقی (اسم اشیاء و اسم معنی)	پایه
ک	پسوند اشتقاقی (اسم اشیاء و اسم معنی)	عروسک
ئه	پسوند اشتقاقی (اسم اشیاء و اسم معنی)	فطریه
ی	پسوند اشتقاقی (اسم اشیاء و اسم معنی)	گوشی
گان	پسوند اشتقاقی (اسم اشیاء و اسم معنی)	ناوگان
واره	پسوند اشتقاقی (اسم اشیاء و اسم معنی)	جشنواره
ئینه	پسوند اشتقاقی (اسم اشیاء و اسم معنی)	گنجینه
ک	پسوند اشتقاقی (تصغیر و تحبیب)	طفلک
چه	پسوند اشتقاقی (تصغیر و تحبیب)	تاریخچه
ی	پسوند اشتقاقی (تصغیر و تحبیب)	مامانی
ه	پسوند اشتقاقی (تأنیث)	همشیره
ا	پسوند اشتقاقی (تأنیث)	آتوسا
یه	پسوند اشتقاقی (تأنیث)	خیریه
ی	پسوند اشتقاقی (مصدر/حاصل)	آزادی
ن	پسوند اشتقاقی (مصدر/حاصل)	رفتن
ا	پسوند اشتقاقی (مصدر/حاصل)	ژرفا
کار	پسوند اشتقاقی (اسم‌ساز)	فراموش‌کار

بندی	پسونند اشتقاقی (اسم‌ساز)	طبقه‌بندی
گری	پسونند اشتقاقی (اسم‌ساز)	وحشی‌گری
کاری	پسونند اشتقاقی (اسم‌ساز)	کنک‌کاری
بازی	پسونند اشتقاقی (اسم‌ساز)	کاغذبازی
ش	پسونند اشتقاقی (اسم‌ساز)	آرایش
ار	پسونند اشتقاقی (اسم‌ساز)	خریدار
ه	پسونند اشتقاقی (اسم‌ساز)	خنده
مان	پسونند اشتقاقی (اسم‌ساز)	زایمان
ان	پسونند اشتقاقی (اسم‌ساز)	آشتی‌کنان
ئیت	پسونند اشتقاقی (اسم‌ساز)	حساسیت
مند	پسونند اشتقاقی (دال بر دارندگی و اتصاف)	ثروتمند
ور/اور/وار	پسونند اشتقاقی (دال بر دارندگی و اتصاف)	بارور/سزاوار
ناک	پسونند اشتقاقی (دال بر دارندگی و اتصاف)	اندوهناک
ه	پسونند اشتقاقی (دال بر دارندگی و اتصاف)	همه‌کاره
نده	پسونند اشتقاقی (فاعلی و مفعولی)	فرساینده
ان	پسونند اشتقاقی (فاعلی و مفعولی)	لرزان
ا	پسونند اشتقاقی (فاعلی و مفعولی)	توانا
ار	پسونند اشتقاقی (فاعلی و مفعولی)	پرستار
ی	پسونند اشتقاقی (فاعلی و مفعولی)	شکاری
گار	پسونند اشتقاقی (فاعلی و مفعولی)	خواستگار
ند	پسونند اشتقاقی (فاعلی و مفعولی)	خوشایند
ه	پسونند اشتقاقی (فاعلی و مفعولی)	پخته
ی	پسونند اشتقاقی (دال بر نسبت)	دنده‌ای
ین	پسونند اشتقاقی (دال بر نسبت)	آهنین
ئینه	پسونند اشتقاقی (دال بر نسبت)	دیرینه
گان	پسونند اشتقاقی (دال بر نسبت)	خدایگان

گانه	پسوند اشتقاقی (دال بر نسبت)	پنج گانه
وند	پسوند اشتقاقی (دال بر نسبت)	شهر وند
ئیه	پسوند اشتقاقی (دال بر نسبت)	تحریریه
و	پسوند اشتقاقی (دال بر نسبت)	اخمو
گین	پسوند اشتقاقی (دال بر نسبت)	غمگین
گون	پسوند اشتقاقی (دال بر شباهت)	گندم گون
فام	پسوند اشتقاقی (دال بر شباهت)	گل فام
آسا	پسوند اشتقاقی (دال بر شباهت) / پسوندواره فعلی	رعد آسا
سان	پسوند اشتقاقی (دال بر شباهت)	گر به سانان
وار	پسوند اشتقاقی (دال بر شباهت)	دیوانه وار
وش	پسوند اشتقاقی (دال بر شباهت)	مehوش
انه	پسوند اشتقاقی (دال بر شباهت)	دلیرانه
سا	پسوند اشتقاقی (دال بر شباهت)	پریسا
وی	پسوند اشتقاقی (صفت ساز)	دنیوی
باره	پسوند اشتقاقی (دال بر کیفیت جسمی یا روحی)	شکمباره
آگین	پسوند اشتقاقی (دال بر کیفیت)	زهر آگین
آنی	پسوند اشتقاقی (صفت ساز)	عصبانی
گرا	پسوند اشتقاقی (صفت ساز) / پسوندواره فعلی	سنت گرا
نما	پسوند اشتقاقی (اسم ساز) / پسوندواره فعلی	سال نما
طلب	پسوند اشتقاقی (صفت ساز) / پسوندواره فعلی	استقلال طلب
پرست	پسوند اشتقاقی (صفت ساز) / پسوندواره فعلی	خدا پرست
سنج	پسوند اشتقاقی (اسم ابزار) / پسوندواره فعلی	فشار سنج
بر	پسوند اشتقاقی (اسم دارو) /	تب بر

	پسونداره فعلی	
پذیر	پسوند اشتقاقی (صفت‌ساز) / امکان‌پذیر پسونداره فعلی	
خیز	پسوند اشتقاقی (صفت‌ساز) / نفت‌خیز پسونداره فعلی	
شناس	پسوند اشتقاقی (صفت‌ساز) / هواشناس پسونداره فعلی	
دان	پسوند اشتقاقی (صفت‌ساز) / ریاضیدان پسونداره فعلی	
نشین	پسوند اشتقاقی (اسم‌ساز، معرف نژاد و قوم و طبقه اجتماعی) / پسونداره فعلی کارگرنشین	
خواه	پسوند اشتقاقی (صفت‌ساز / اسم‌ساز، معرف یک رژیم یا نظام اجتماعی) / مشروطه‌خواه پسونداره فعلی	
دار	پسونداره فعلی	امانتدار
گیر	پسونداره فعلی	دندانگیر
زن	پسونداره فعلی	چشمک‌زن
کش	پسونداره فعلی	زجرکش
بند	پسونداره فعلی	ماست‌بند
خوار	پسونداره فعلی	گوشت‌خوار
فروش	پسونداره فعلی	دستفروش
گو	پسونداره فعلی	حقیقتگو
پوش	پسونداره فعلی	ژنده‌پوش
خوان	پسونداره فعلی	آوازه‌خوان
آور	پسونداره فعلی	سودآور
جو	پسونداره فعلی	حقیقت‌جو
انداز	پسونداره فعلی	چشم‌انداز
رو	پسونداره فعلی	تندرو
کن	پسونداره فعلی	خردکن

آمیز	پسونداره فعلی	تمسخر آمیز
کوب	پسونداره فعلی	دیوار کوب
انگیز	پسونداره فعلی	رقت انگیز
تاب	پسونداره فعلی	بازتاب
چین	پسونداره فعلی	دورچین
نویس	پسونداره فعلی	خوشنویس
پرور	پسونداره فعلی	شهیدپرور
شکن	پسونداره فعلی	دندان شکن
بخش	پسونداره فعلی	فرح بخش
پیچ	پسونداره فعلی	طناب پیچ
افکن	پسونداره فعلی	مردافکن
ریز	پسونداره فعلی	سرریز
آرا	پسونداره فعلی	زینت آرا
بر	پسونداره فعلی	بیرون بر
گداز	پسونداره فعلی	جانگداز
بار	پسونداره فعلی	سبک بار
نواز	پسونداره فعلی	گوش نواز
پرواز	پسونداره فعلی	بلند پرواز
نشین	پسونداره فعلی	خوش نشین
افشان	پسونداره فعلی	زرافشان
کش	پسونداره فعلی	دخترکش
گرد	پسونداره فعلی	بیابان گرد
گذار	پسونداره فعلی	خدمتگذار
تراش	پسونداره فعلی	پیکر تراش
جوش	پسونداره فعلی	خودجوش
کن	پسونداره فعلی	پوست کن
مال	پسونداره فعلی	لگدمال
یاب	پسونداره فعلی	طلایاب
اندیش	پسونداره فعلی	دوراندیش
زا	پسونداره فعلی	بیماری‌زا

شمار	پسونددواره فعلی	گاه‌شمار
افروز	پسونددواره فعلی	جنگ‌افروز
پر	پسونددواره فعلی	؟
گشا	پسونددواره فعلی	مشکل‌گشا
رس	پسونددواره فعلی	فریادرس
گستر	پسونددواره فعلی	مهر‌گستر
خواب	پسونددواره فعلی	خوش‌خواب
نورد	پسونددواره فعلی	دریانورد
پسند	پسونددواره فعلی	مشکل‌پسند
دزد	پسونددواره فعلی	بچه‌دزد
ران	پسونددواره فعلی	لو‌کوموتیوران
فریب	پسونددواره فعلی	دلفریب
نام	پسونددواره فعلی	خوش‌نام
نگار	پسونددواره فعلی	چهره‌نگار
افزا	پسونددواره فعلی	روح‌افزا
آموز	پسونددواره فعلی	دست‌آموز
پاش	پسونددواره فعلی	نمک‌پاش
بو	پسونددواره فعلی	خوش‌بو
دوش	پسونددواره فعلی	ماردوش
شو	پسونددواره فعلی	تاشو
آفرین	پسونددواره فعلی	جان‌آفرین
تراز	پسونددواره فعلی	هم‌تراز
خند	پسونددواره فعلی	پوز‌خند
سرا	پسونددواره فعلی	ترانه‌سرا
آویز	پسونددواره فعلی	رخت‌آویز
دو	پسونددواره فعلی	سگ‌دو
چران	پسونددواره فعلی	گاو‌چران
نوش	پسونددواره فعلی	دردنوش
ربا	پسونددواره فعلی	آهن‌ربا
پیوند	پسونددواره فعلی	؟

تاز	پسونندواره فعلی	یکه تاز
رسان	پسونندواره فعلی	پیام رسان
روب	پسونندواره فعلی	خاک روب
انجام	پسونندواره فعلی	سرانجام
پیرا	پسونندواره فعلی	چمن پیرا
خا	پسونندواره فعلی	شکر خا
دم	پسونندواره فعلی	آتشین دم
گزین	پسونندواره فعلی	خلوت گزین
آزار	پسونندواره فعلی	مردم آزار
بیز	پسونندواره فعلی	آردبیز
خر	پسونندواره فعلی	بزخر
فرسا	پسونندواره فعلی	طاقت فرسا
ورز	پسونندواره فعلی	دادورز
آگند	پسونندواره فعلی	کج آگند
باش	پسونندواره فعلی	شادباش
ترس	پسونندواره فعلی	خدا ترس
چسب	پسونندواره فعلی	دلچسب
اندوز	پسونندواره فعلی	مال اندوز
چرخ	پسونندواره فعلی	تک چرخ
زدا	پسونندواره فعلی	گندزدا
گسار	پسونندواره فعلی	غم گسار
توان	پسونندواره فعلی	کم توان
چر	پسونندواره فعلی	علف چر
آزما	پسونندواره فعلی	بخت آزما
آشوب	پسونندواره فعلی	دل آشوب
پران	پسونندواره فعلی	مگس پران
توز	پسونندواره فعلی	کینه توز
خراش	پسونندواره فعلی	گوش خراش
شتاب	پسونندواره فعلی	؟
گوار	پسونندواره فعلی	خوش گوار

آشام	پسوندواره فعلی	خون آشام
ی	قیدساز	آخر عمری/آمدنی/برگشتی
اَ	قیدساز	اخیراً/دائماً
ک	قیدساز	کم کمک/نم‌نمک
کی	قیدساز	یواشکی/هول‌هولکی
ه	قیدساز	امروزه
ان	قیدساز	سحرگاهان
انه	قیدساز	پیروزمندانه
وار	قیدساز	دیوانه‌وار
بر	پیشوندی که در معنی فعل تغییر ایجاد می‌کند	برداشت
در	پیشوندی که در معنی فعل تغییر ایجاد می‌کند	درافتاد
باز	پیشوندی که در معنی فعل تغییر ایجاد می‌کند	بازداشت
فرا	پیشوندی که در معنی فعل تغییر ایجاد می‌کند	فراگرفت
فرو	پیشوندی که در معنی فعل تغییر ایجاد می‌کند	فرورفت

پیش از استفاده از اطلاعات ساختوازی موجود در جدول ۴-۲، چند فایل از مجموعه فایل‌های کتاب‌ها انتخاب شد و نگارنده به بررسی این مسئله پرداخت که چه میزان می‌توان از اطلاعات موجود در جدول ۴-۲ استفاده کرد؛ بررسی اجمالی نشان داد که نیازی به اعمال همه اطلاعات موجود در جدول در مرحله پیش‌پردازش نیست و برخی واژه‌ها هستند که در این فایل‌ها از نظر رعایت نیم‌فاصله مسئله دارند و باید در این مرحله این واژه‌ها اصلاح شوند. در نتیجه دو جدول از جدول فهرست‌وندها استخراج شد (جدول ۴-۳ و ۴-۴) و اطلاعات موجود در این جدول برای پیش‌پردازش متون پیکره مورد استفاده قرار گرفت.

جدول ۴-۳: فهرست پیشوندها

پیشوندهایی که باید با نیم‌فاصله به واژه <u>بعد</u> از خود متصل شوند.	مثال
می-	می گفت / می توان
نمی-	نمی گفت / نمی توان
بی-	بی بنیه / بی ریشه
فرا-	فراگرفت
فرو-	فرورفت

جدول ۴-۴: فهرست پسوندها

پسوندهایی که باید با نیم‌فاصله به واژه <u>قبل</u> از خود متصل شوند.	مثال
-ها / -های / -هایی	کتاب‌ها / کتاب‌های / کتاب‌هایی
-تر / -ترین	مومن‌تر / خوب‌ترین
-ام	زنده‌ام / خورده‌ام
-ای	زنده‌ای / خواننده‌ای
-ایم	زنده‌ایم / خسته‌ایم
-اید	زنده‌اید / رفته‌اید
-اند	رفته‌اند / خوانده‌اند
-مان	طرحمان / پایگاهمان / پایگاه‌مان
-تان	لباستان / روسری‌تان / دانشگاه‌تان
-شان	لباسشان / دانشگاه‌شان
-ستان	کردستان
-کده	دانشکده
-دان / -دانی	نمک‌دان / سگ‌دانی / ریاضیدان

کشتزار	-زار / -زاری
سنگ‌لاخ	-لاخ / -لاخی
نگهبان	-بان
آهنگر / وحشی‌گری	-گر / -گری
آهنگ‌سازی	سازی
فرش‌باف	-باف / -بافی
شیرینی‌پزی	-پز / -پزی
ناوگان	-گان
جشنواره	-واره
طبقه‌بندی	-بندی
ثروت‌مند	-مند / -مندی
اندوه‌ناک	-ناک
کتابدار / کتابداری	-دار / -داری
فرساینده	-نده
خواستگار	-گار
خوشایند	-ند
پنج‌گانه	-گانه
شهروند	-وند
غمگین	-گین
گندم‌گون	-گون
گل‌فام	-فام
رعد‌آسا	-آسا / -آسایی
گربه‌سانان	-سان / -سانان
دیوانه‌وار	-وار
مهوش	-وش
دلیرانه	-انه
شکمبار	-باره
زهر‌آگین	-آگین
سنت‌گرا	-گرا / -گرایی
خداپرست	-پرست / -پرستی

فشار سنج	-سنج / -سنجی
امکان پذیر	-پذیر / -پذیری
هواشناس	-شناس / -شناسی
کارگرنشین	-نشین
دندانگیر	-گیر / -گیری
گوشته‌خوار / گیاه‌خواری	-خوار / -خواری
زنده‌پوش	-پوش
آوازه‌خوان	-خوان / -خوانی
سودآور / جمع‌آوری	-آور / -آوری
حقیقت‌جو	-جو / -جویی
چشم‌انداز	-انداز
خردکن	-کن
تمسخرآمیز	-آمیز / -آمیزی
دیوارکوب	-کوب / -کوبی
رقت‌انگیز	-انگیز / -انگیزی
خوشنویس	-نویس / -نویسی
شهیدپرور	-پرور / -پروری
دندان‌شکن	-شکن / -شکنی
مردافکن	-افکن / -افکنی
نرم‌افزار / نرم‌افزاری	-افزار / -افزاری
جانگذا	-گذا / -گذازی
گوش‌نواز / مهمان‌نوازی	-نواز / -نوازی
زرافشان	-افشان
دخترکش	-کش
خدمتگذار / برچسب‌گذاری	گذار / -گذاری
طلایاب / دستیابی	-یاب / -یابی
دوراندیش	-اندیش
بیماری‌زا	-زا
مشکل‌گشا	-گشا / -گشایی
فریادرس / دادرسی	-رس / -رسی

مهر گستر / داد گستری	- گستر / - گستری
دریانورد	-نورد / -نوردی
مشکل پسند	-پسند / -پسندی
روح افزا	-افزا / -افزایی
دست آموز	-آموز / -آموزی
نمک پاش	-پاش
پوز خند	-خند
یکه تاز	-تاز
پیام رسان / اطلاع رسانی	-رسان / -رسانی
طاقت فرسا	-فرسا / -فرسایی
دادورز	-ورز
مال اندوز / ثروت اندوزی	-اندوز / -اندوزی
گندزدا	-زدا / -زدایی
غم گسار	-گسار
بخت آزما	-آزما / -آزمایی
کینه توز	-توز / -توزی
خوش گوار	-گوار / -گواری
خون آشام	-آشام
دیوانه وار	-وار / -واری

از اطلاعات جدول‌های ۴-۳ و ۴-۴ به این صورت استفاده شده است: برنامه‌ای نوشته شد که بر اساس آن اگر وندها (پیشوند یا پسوند) جدا از ستاک باشند، با نیم فاصله به ستاک متصل شوند. به عنوان نمونه، هر جا تکواژ «-ها»، «-های» و «-هایی» جدا از تکواژ قبل است و یک واژه مجزا در نظر گرفته شده است، با نیم فاصله به واژه/تکواژ قبل از آن متصل گردد یا هر جا تکواژ «-ی» و «-ای» جدا از تکواژ قبل است و یک واژه مجزا در نظر گرفته شده است، با نیم فاصله به واژه/تکواژ قبل از آن متصل گردد.

۴-۳- نشانه‌گذاری داده‌ها

پس از گردآوری داده‌ها و هنجارسازی آن‌ها، گام بعدی در ساخت پیکره، نشانه‌گذاری داده‌های پیکره است. داده‌های گردآوری شده، داده خام هستند که هیچگونه نشانه‌گذاری ندارند. نشانه‌گذاری/حاشیه‌نویسی^۱ پیکره، افزودن اطلاعات زبانی توضیحی-تفسیری به پیکره است. برخی از پژوهشگران مانند سینکلر (۲۰۰۴، نقل از لیچ: ۲۰۰۴)، ترجیح می‌دهند وارد مقوله حاشیه‌نویسی پیکره نشوند؛ زیرا معتقد هستند پیکره حاشیه‌نویسی نشده، پیکره‌ای خالص است و برای پژوهش‌های زبان‌شناسی چنین پیکره‌هایی را مناسب‌تر می‌دانند. این پژوهشگران به حاشیه‌نویسی اعتماد ندارند و با شک به آن می‌نگرند و معتقد هستند حاشیه‌نویسی نمی‌تواند بدون خطا باشد. اما بیشتر پژوهشگران این حوزه بر این باورند که پیکره‌ها تنها زمانی مفید هستند که بتوان دانش یا اطلاعات را از آن‌ها استخراج کرد که این کار خود مستلزم ساخت اطلاعات درون پیکره است؛ به عبارتی دیگر، این کار به وسیله اضافه کردن نشانه‌گذاری انجام می‌شود. پیکره خام^۲ حاوی هیچ اطلاعات مستقیمی، در دستور یا سایر زمینه‌ها، نیست و این مسئله برای بسیاری از زمینه‌هایی که پیکره برای آن به کار گرفته می‌شود، عاملی بازدارنده برای به کار گرفتن پیکره محسوب می‌شود. به عنوان مثال، واژه «left» در انگلیسی را در نظر بگیرید: این واژه در تضاد با «right» می‌تواند صفت باشد (my left hand) یا می‌تواند قید باشد (turn left) و یا اسم باشد (on your left). همچنین این واژه می‌تواند گذشته یا شکل حال کامل فعل «leave» باشد (I left early). بنابراین، معانی و کاربردهای گوناگون واژه «left» را نمی‌توان تنها از طریق شکل نوشتاری این واژه تشخیص داد. اما اگر یک پیکره از نظر دستوری، خوب و موفق برچسب‌گذاری شده باشد، هر کدام از حالت‌های واژه «left» همراه با یک برچسب که طبقه آن واژه را نشان می‌دهد، همراه می‌شود (لیچ: ۲۰۱۳). بنابراین برای بسیاری از زبان‌شناسان برای کارهای خاصی که در زمینه‌های پژوهشی انجام می‌شود، داده خام عملاً کاربردی نیست و این نیاز وجود دارد که اطلاعات زبان‌شناختی به آن اضافه شود که بتواند در پژوهش‌های زبان‌شناختی پیکره‌ای به پژوهشگر کمک کند. آن‌ها معتقدند نشانه‌گذاری/حاشیه‌نویسی، پیکره را بسیار مفیدتر و غنی‌تر از پیکره خام می‌کند و ارزشی افزوده به پیکره می‌دهد. به عنوان نمونه، عمل «برچسب‌گذاری اجزای واژگانی کلام»^۳ که برای «پیکره براون» انجام شده است، در سطح گسترده‌ای توزیع شده و مورد استفاده افرادی قرار گرفته است که در راستای پژوهش‌های خود نیازمند برچسب-

¹ annotation

² raw corpus

³ part of speech (POS) tagging

گذاری اجزای واژگانی کلام یا پیکره برچسب گذاری شده بودند. غیر از برچسب گذاری اجزای واژگانی کلام، که بسیار رایج و کاربردی است، حاشیه‌نویسی‌های دیگری نیز وجود دارد که به سطوح مختلف تجزیه و تحلیل‌های زبانی یک پیکره مربوط می‌شود. این حاشیه‌نویسی‌ها شامل: حاشیه‌نویسی/نشانه گذاری آوایی^۱، حاشیه‌نویسی/نشانه گذاری نحوی^۲، حاشیه‌نویسی/نشانه گذاری معنایی^۳، حاشیه‌نویسی/نشانه گذاری کاربردی^۴، حاشیه‌نویسی/نشانه گذاری گفتمانی^۵، حاشیه‌نویسی/نشانه گذاری سبکی^۶ و حاشیه‌نویسی/نشانه-گذاری واژگانی^۷ است. موارد نام‌برده شده، معروف‌ترین و متداول‌ترین نوع حاشیه‌نویسی هستند؛ می‌توان انواع دیگر حاشیه‌نویسی را از ادبیات مربوطه استخراج کرد یا حتی حاشیه‌نویسی‌های جدیدی را به فهرست حاشیه‌نویسی‌ها اضافه کرد.

برای نشانه گذاری پیکره، دو روش وجود دارد: نشانه گذاری دستی و نشانه گذاری ماشینی. نشانه-گذاری دستی، در حقیقت استفاده از متخصص برای نشانه گذاری است؛ این روش زمانی به کار می‌رود که هیچگونه داده برچسب گذاری شده با اهداف مورد نظر موجود نباشد. این روش نقاط قوت و ضعف نیز دارد که به آن‌ها اشاره می‌شود:

ویژگی مثبت روش نشانه گذاری دستی: ساختن چیزی است برای اولین بار. کیفیت بالا در صحت برچسب‌ها از دیگر مزیت‌های این نوع نشانه گذاری شمرده می‌شود. در حقیقت، درصد خطا در این روش برچسب گذاری به دلیل انجام آن توسط فرد خبره، بسیار پایین است.

نقطه ضعف روش نشانه گذاری دستی: زمان‌بر بودن و طاقت‌فرسا بودن، پرهزینه بودن (نیروی انسانی هزینه انجام این کار را طلب خواهد کرد زیرا وقت زیادی از ناظر انسانی گرفته می‌شود)، حجم محدود داده (آنچه تهیه خواهد شد، از نظر حجم، محدود است. مثلاً برچسب گذاری دو میلیون واژه، ممکن است یک سال زمان لازم داشته باشد و نتوان آن را ادامه داد. البته می‌توان زمان را کم کرد و نیروی انسانی را بیشتر کرد که خود مستلزم صرف هزینه بیشتر است)، تنوع در نشانه گذاری (اگر از افراد مختلف برای نشانه گذاری داده استفاده شود، ممکن است عدم انسجام در برچسب گذاری پیش آید و منجر به کاهش کیفیت نشانه گذاری شود. به عنوان نمونه، واژه «دوشنبه» را یک فرد تحت عنوان «اسم مکان» (اسم شهر)

¹ phonetic annotation

² syntactic annotation

³ semantic annotation

⁴ pragmatic annotation

⁵ discourse annotation

⁶ stylistic annotation

⁷ lexical annotation

برچسب‌گذاری کند و فرد دیگر «اسم زمان» (روز دوشنبه) و فرد دیگری اسم خاص (اسم یک فرد) و این مسئله باعث می‌شود، یک واژه در بافت یکسان، بیش از یک برچسب داشته باشد و بنابراین، نمی‌توان آمار دقیقی از واژه‌ها و برچسب‌هایشان داشت. حتی اگر فرد خبره فقط یک نفر باشد، باز هم این مشکل ممکن است به وجود آید، زیرا یک فرد در یک زمان درباره‌ی برچسب یک کلمه یک تصمیم می‌گیرد و در زمانی دیگر برچسب دیگری به همان واژه در بافت مشابه اختصاص می‌دهد).

روش دوم، نشانه‌گذاری داده‌های پیکره با کمک ماشین است که از ابزارهای خاصی برای این منظور استفاده می‌شود.

روش سوم، نشانه‌گذاری تلفیقی است، به این معنا که ابتدا پیکره با کمک ابزارهای خاص توسعه‌داده شده برای برچسب‌گذاری، نشانه‌گذاری می‌شود و سپس، نیروی انسانی برچسب‌ها را کنترل می‌کند. در این روش، در صورت به کارگیری ابزارهای کارآمد، زحمت و هزینه‌ی نیروی انسانی را به میزان قابل توجهی کم می‌کند.

برچسب‌گذاری اجزای واژگانی کلام، که برچسب‌گذاری دستوری^۱ و نشانه‌گذاری نحوی-واژگانی^۲ نیز خوانده می‌شود، به معنی تخصیص دادن برچسب اجزای واژگانی کلام، مانند اسم، فعل، قید، صفت، حروف اضافه و غیره به هر واژه در پیکره است. این نوع برچسب‌گذاری، همانگونه که ذکر شد، یکی از اولین گونه‌های برچسب‌گذاری است که در سطح وسیعی در پیکره‌ها به کار گرفته شده است و در میان انواع برچسب‌گذاری، رایج‌ترین آن‌ها در نظر گرفته می‌شود. شایان ذکر است که برچسب‌گذاری اجزای واژگانی کلام، ابتدایی‌ترین و اساسی‌ترین نوع برچسب‌گذاری نیز در نظر گرفته می‌شود که به عنوان پایه و اساس تجزیه و تحلیل‌های بیشتر، مانند قطعه‌قطعه کردن^۳ و نشانه‌گذاری معنایی محسوب می‌شود. پیکره-ای که به این صورت نشانه‌گذاری شده باشد، کاربردهای فراوانی دارد که می‌تواند شامل طیف وسیعی، از رفع ابهام از هم‌نگاره‌ها گرفته تا کاربردهای پیچیده‌تر مانند محاسبه‌ی وقوع طبقات واژگانی در پیکره، شود. بسیاری از تجزیه و تحلیل‌های زبان‌شناسی مانند باهم‌آیی یک واژه با واژه‌های دیگر نیز به این نوع برچسب‌گذاری وابسته هستند. با توجه به کاربرد وسیع برچسب اجزای واژگانی کلام در پردازش متن، تصمیم گرفته شد این نوع برچسب‌گذاری نیز به پیکره اضافه شود. در این راستا تصمیم گرفته شد ابتدا از

^۱ grammatical tagging

^۲ morpho-syntactic annotation

^۳ parsing

یکی از ابزارهای آماده برای برچسب‌گذاری ماشینی اجزای واژگانی کلام استفاده شود و سپس، برچسب‌ها دستی کنترل شوند. ابزار استفاده شده برای برچسب‌گذاری ماشینی، ابزار «هضم» است که برای هنجارسازی متون پیکره نیز از این ابزار استفاده شد.

فهرست برچسب‌ها، شامل برچسب‌هایی است که در مقاله بی‌جن‌خان و همکاران (۲۰۱۱) معرفی شده است. این فهرست با مثال در جدول ۴-۵ آورده شده است:

جدول ۴-۵: فهرست برچسب‌ها

مثال	Tag Name	POS tag
دانشجو، کتابخانه، میز	Noun	N
از، به، برای، در	Preposition	PREP/P
نقطه، ویرگول، علامت سوال	Punction	PUNC
زشت، اجتماعی، کوچک	Adjective	AJ
می‌شمرد، خورد، خوابید	Verb	V
که، چنانکه	Conjunction	CON
۵۰، ۱۰۰، ۱۲	Number	NUM
من، تو، او، ما، شما، ایشان	Pronoun	PRO
این، آن، هر	Determiner	DET
یقیناً، مطمئناً، خوب، حتماً، بسیار	Adverb	ADV
را	Postposition	POSTP
?	Residual	RES
نوع، دست، چنین	Classifier	CL
ای، یا	Interjection	INT

نکته: در فهرستی که در جدول ۴-۵ آورده شده است، کسره اضافه نمایش داده نشده است؛ «هضم» به منظور نمایش کسره اضافه، هر گاه پس از واژه‌ای، کسره اضافه وجود داشته، کنار برچسب، e گذاشته است که نمایانگر کسره اضافه است. به عنوان نمونه، برچسب عبارت «شناسایی ساختار محتوایی مطالعات» به صورت: شناسایی (Ne) ساختار (Ne) محتوایی (ADJe) مطالعات (N) است.

شایان ذکر است، دقت برچسب‌گذاری ابزار «هضم» ۹۷,۱٪ گزارش شده است (sobhe/hazm).

جزئیات فنی هنجارسازی و برچسب‌گذاری ماشینی داده‌های پیکره در پیوست ۲ آورده شده است.

۴-۴- کنترل دستی برچسب‌ها

کنترل دستی برچسب‌ها در چند مرحله انجام گرفت: ابتدا فهرستی از واژه‌های پیکره به صورت فایل اکسل تهیه شد تا نگارنده به بررسی برچسب‌های واژه‌هایی که می‌توان بدون بافت هم برچسب آن‌ها را تشخیص داد، بپردازد؛ به عنوان نمونه واژه «می‌توان» فارغ از اینکه در چه بافت زبانی به کار رود، همیشه برچسب «فعل (V)» دارد، یا واژه «علمی» می‌تواند «اسم» یا «صفت» باشد و نمی‌تواند حرف اضافه (P) باشد. بخشی از این فایل اکسل به صورت زیر به عنوان نمونه آورده شده است:

درتحقیق	Ne		
درتحقیقات	N		
درتحلیل	Ne		
درترکیب	Ne		
درتصویر	Ne		
درتعامل	N		
درتعامل	PUNC		
درتعداد	V		
درتعرف	Ne		
درتعریف	N		
درتعیین	AJ		
درتعیین	N		
درتعیین	Ne		
درتلاشند	V		
درتمام	DET		
درتمایی	AJ		
درتولی‌د	V		
درتولیدات	Ne		
درج	N		
درج	Ne		
درج	PUNC		
درجات	N		

در این مرحله مشخص شد، علیرغم همه کارهای انجام شده در مرحله هنجارسازی داده‌ها، گویا هنگام انتقال داده‌ها از فایل‌ها به پیکره برای برچسب‌گذاری نیز برخی از داده‌ها مجدداً تغییراتی در آن‌ها اتفاق افتاده است و در بسیاری موارد فاصله‌ها درست تشخیص داده نشده است و بسیاری واژه‌ها به هم چسبیده شده‌اند و در نتیجه برچسب‌گذاری به درستی انجام نشده است. شایان ذکر است که این امر حاصل چندین بار چک کردن و تطبیق واژه‌ها در فایل اصلی (اصل فایل word کتاب‌ها) بوده است.

در این مرحله تصمیم گرفته شد خروجی دیگری از داده‌های پیکره در فایل اکسل گرفته شود که در آن واژه‌ها برچسب‌گذاری نشده‌اند (مانند نمونه زیر) تا بتوان هر جا واژه‌ای مشکل املایی دارد یا به دلیل عدم درج فاصله به واژه‌های قبل و بعد خود چسبیده است، اصلاح گردد.

A	B	C
اجتماعی‌ای	اجتماعی‌ای	
اجتماعی‌تر	اجتماعی‌تر	
اجتماعی‌تر	اجتماعی‌تر	1
اجتماعی-تکنیکی	اجتماعی-تکنیکی	
اجتماعی-تکنیک‌نگریست	اجتماعی-تکنیک‌نگریست	1
اجتماعی‌خاصی	اجتماعی‌خاصی	1
اجتماعی‌سازی	اجتماعی‌سازی	
اجتماعی‌شان	اجتماعی‌شان	1
اجتماعی-فرهنگی	اجتماعی-فرهنگی	
اجتماعی-فناورانه	اجتماعی-فناورانه	
اجتماعی-فنی	اجتماعی-فنی	
اجتماعی‌کردن	اجتماعی‌کردن	
اجتماعی‌مان	اجتماعی‌مان	
اجتماعی	اجتماعی	1
اجتناب	اجتناب	
اجتناب‌ناپذیر	اجتناب‌ناپذیر	1
اجتناب‌ناپذیر	اجتناب‌ناپذیر	
اجتناب‌ناپذیر	اجتناب‌ناپذیر	1
اجتناب‌ناپذیری	اجتناب‌ناپذیری	
اجتناب‌ناپذیری	اجتناب‌ناپذیری	1

ستون A، واژه‌ها به شکل موجود در فایل‌هاست. ستون B، شکل اصلاح‌شدهٔ واژه‌هایی است که غلط تایپی دارند، یا ناقص تایپ شده‌اند اما می‌توان حدس زد چه واژه‌هایی هستند، یا اشکال از عدم رعایت فاصله است که ممکن است در فایل اصلی کتاب هم نویسنده فاصله را به درستی رعایت نکرده باشد یا هنگام انتقال فایل به پیکره فاصله‌ها به هم خورده باشد؛ به هر صورت شکل درست نوشتاری واژه‌ها در ستون B وارد شده است و چنانچه این ویرایش انجام گرفته است، عدد ۱، در ستون C مقابل واژهٔ اصلاح شده درج شده است که بتوانیم در مرحلهٔ بعد این واژه‌ها را شناسایی کنیم.

در مرحلهٔ بعد برنامه‌ای ماشینی نوشته شد که خروجی فایل اکسل را روی همهٔ فایل‌ها اعمال کند؛ به این صورت که برنامه ابتدا همهٔ واژه‌های مسئله‌دار موجود در فایل اکسل (آن‌هایی که عدد ۱ مقابل صورت اصلاح‌شدهٔ آن‌هاست) را در فایل‌های متنی (فایل کتاب‌ها) پیدا می‌کند و سپس، همهٔ صورت‌های غلط را با صورت‌های نوشتاری صحیح جایگزین می‌کند. به عنوان نمونه، «اجتماعی-تکنیک‌نگریست» را

در یک یا چند فایل پیدا می‌کند و صورت درست نوشتاری آن، که در حقیقت شامل سه واژه است (اجتماعی-تکنیکی نگریست)، را جایگزین صورت نوشتاری غلط می‌کند. این مرحله خود نوعی پیش-پردازش محسوب می‌شود؛ در حقیقت، پیش‌پردازشی که پس از مرحله برچسب‌گذاری به ضرورت آن پی برده شد و باید به صورت دستی واژه‌ها اصلاح می‌شد. شایان ذکر است که این مرحله بسیار وقت‌گیر بود اما باعث می‌شد فرآیند برچسب‌گذاری ماشینی با دقت بیشتری انجام شود. خروجی این مرحله مجدداً وارد برچسب‌گذاری ماشینی شد تا بتوان در نهایت مجدداً برچسب‌ها را کنترل کرد.

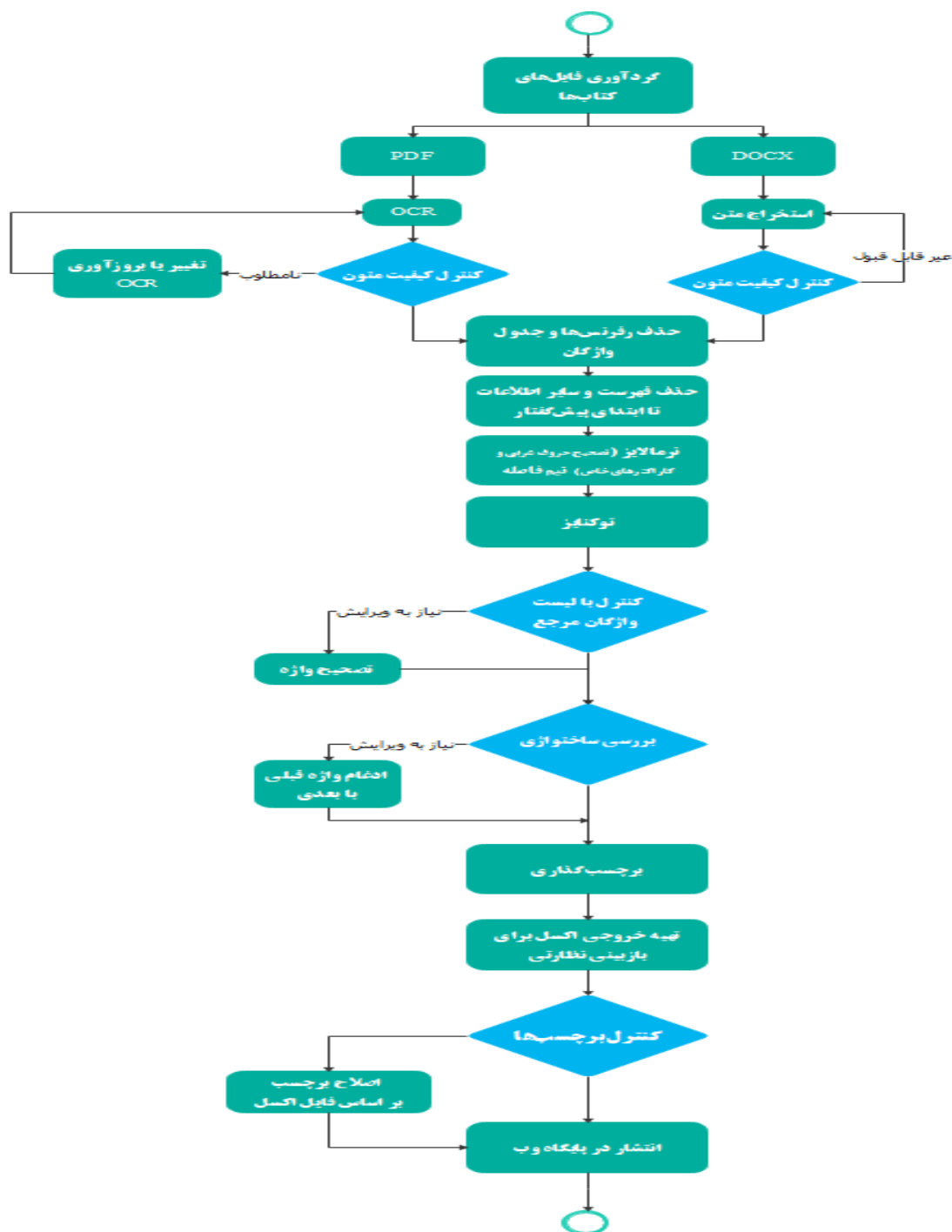
در نهایت خروجی دیگری به صورت فایل اکسل در اختیار نگارنده قرار گرفت که در آن واژه‌ها همراه برچسب‌های آن‌ها نمایش داده شده است. در این مرحله نیز با مشکلات املائی مواجه شدیم که حاصل اضافه کردن فایل‌های جدید به فایل‌های قبلی بود؛ مشکلات از جنس همان مشکلات فایل قبلی از نظر فاصله و نیم‌فاصله یا تایپ اشتباه واژه بود. این فایل را نگارنده به دقت مورد بررسی قرار داد و در دو ستون (C و D) مجدداً واژه‌ها و برچسب‌های آن‌ها را کپی کرد تا بتوان غلط‌های املائی و برچسب‌های غلط را اصلاح نمود، در این فرآیند واژه‌هایی که پس از اصلاح املائی بیش از یک واژه شده‌اند، مجدداً در فرآیند برچسب‌گذاری قرار گرفتند تا به هر یک از واژه‌های تشکیل‌دهنده آن‌ها برچسب درست داده شود. نمونه‌های زیر قسمتی از این فرآیند را نمایش می‌دهند:

A	B	C	D
گسترده‌ها	N	گسترده‌ها	N
گسترده‌های	Ne	گسترده‌های	Ne
گسترده‌هزینه	ADV	گسترده هزینه	
گسترده‌هوال‌مارتینین	Ne	گسترده وال‌مارتینین	
گسترده‌ی	V	گسترده‌ی	AJe
گسترده‌ی	AJe	گسترده‌ی	AJe
گسترده‌ی	AJe	گسترده‌ی	AJe
گسترده‌ی	ADVe	گسترده‌ی	AJe
گسترده‌ی	AJe	گسترده‌ی	AJe
گسترده‌ی	Ne	گسترده‌ی	AJe
گسترده‌ی	V	گسترده‌ی	AJe
گسترش	N	گسترش	N
گسترش	Ne	گسترش	Ne
گسترش	Ne	گسترش	Ne
گسترش‌پذیر	AJ	گسترش‌پذیر	AJ
گسترش‌پذیر	N	گسترش‌پذیر	AJ
گسترش‌پذیری	N	گسترش‌پذیری	N
گسترش‌پذیری	V	گسترش‌پذیری	N
گسترش‌های	AJ	گسترش‌های	N
گسترش‌های	Ne	گسترش‌های	Ne
گسترش‌های	Ne	گسترش‌های	Ne
گسترشی	N	گسترشی	N
گسترشی	Ne	گسترشی	Ne
گسترش‌یابنده	AJ	گسترش‌یابنده	AJ

شایان ذکر است، از آنجائیکه داده‌های موجود در فایل‌های ورد (word) و پی.دی.اف (pdf) به صورت یکجا وارد فرآیند برچسب‌گذاری نشدند و مرحله به مرحله اضافه شدند، فرآیند اصلاح واژگان و برچسب‌ها نیز در چند مرحله انجام شد و خروجی هر مرحله (نوعاً واژه‌هایی که اصلاحات املایی و فاصله در مورد آن‌ها انجام شده است) به عنوان منبع / فرهنگ لغت برای فایل‌های جدید مورد استفاده قرار گرفت؛ به عنوان نمونه اگر در فایل‌های اول دو واژه «تکنیکی نگریست» به هم چسبیده بودند (تکنیکینگریست) یا واژه «سیستم» به صورت «سیسیستم» تایپ شده است و نگارنده در فایل اکسل اول اصلاح املایی انجام داده است، دیگر این صورت اصلاح شده در همه فایل‌ها جایگزین صورت‌های غلط به هم چسبیده یا غلط شده است و نیازی به اصلاح دستی همه این صورت‌ها غلط در فایل‌های جدید نبوده است.

۴-۵- نمودار جریانی ساخت پیکره و برچسب‌گذاری داده‌ها

در ساخت این پیکره مراحل مختلفی بکار گرفته شد که عبارتند از: ۱- هنجارسازی (نرمالایز کردن): این مورد با کتابخانه «هضم» انجام شد و مشکلات مربوط به نیم فاصله، برخی یونیکدهای عربی و حروف «ی» و «ک» در این مرحله اصلاح شد. ۲- توکنایز کردن: برای شکستن جمله‌ها به واژه‌ها به منظور انجام پردازش‌های بعدی. در این مرحله نیز از کتابخانه «هضم» استفاده شد و مجموعه بزرگی از واژگان تهیه شد. ۳- اصلاحات ساخت‌واژی: برخی واژه‌هایی که بیش از یک تکواژ داشتند (مانند «اطلاع‌رسانی»، «نتیجه‌گیری»، «نداشته‌اند» و غیره)، در گام اول به درستی شناسایی نشده بودند و سیستم ماشینی آن‌ها را یک واژه در نظر نگرفته بود («اطلاع‌رسانی» را دو واژه در نظر گرفته بود، به دلیل وجود فاصله بین دو بخش این واژه). برای رفع این مشکل از اطلاعات موجود در فرهنگ لغت و تجربیات طرح پیشین و اصلاحات املایی که در مرحله هنجارسازی انجام شده بود (به عنوان منبع) استفاده شد و تکواژهایی که با فاصله از ستاک قرار داشتند، با نیم‌فاصله به ستاک متصل شدند. در گام بعدی این مجموعه واژگان در اختیار هضم قرار گرفت و عملیات برچسب زنی روی آن‌ها بصورت ماشینی انجام شد. در گام آخر بصورت دستی خروجی‌ها بررسی شدند. نمودار جریانی مراحل ساخت پیکره همراه با جزئیات هر مرحله به صورت زیر است:



شکل ۴-۱: نمودار جریان‌ی مراحل ساخت پیکره از کتاب‌های دیجیتال ایرانداک

۴-۶- ارزیابی

برای ارزیابی دو بخش هنجارسازی و برجسب‌گذاری، یکی از فایل‌های ورد انتخاب شد و پاراگراف اول فایل برای بررسی جدا شد (متن زیر)، (توضیح: برای تمایز این متن از متن گزارش طرح، متن زیر با قلم دیگری نمایش داده شده است).

یکی از نخستین تلاش‌های علمی بشر برای شناخت آینده، استفاده از روش «پیشبینی ۱» بوده است. پیشبینی سعی میکند تا به یک پیشگویی ۲ دقیق از آینده دست یابد. با گذشت زمان، پیشبینی مورد انتقادات فراوانی قرار گرفت. بیشترین انتقادات مطرح شده، در خصوص تکیه پیشبینی بر گذشته و افق زمانی کوتاه مدت آن بوده است. پیشبینی، آینده را بر طبق گذشته کشف میکرد و همین امر باعث تشکیک افراد بسیاری درباره آن شد؛ زیرا هیچ دلیلی وجود نداشت که الزاماً آینده از گذشته پیروی نماید. از سویی دیگر و به واسطه مدلسازیهای افراطی که با استفاده از ابزار پیشبینی انجام میگرفت، ناگزیر بسیاری از متغیرهای تأثیرگذار را نادیده میانگاشتند. چون مدلسازی نمیتوانست عیناً به تمامی متغیرها توجه کند و در مدل خود از همه آنها استفاده نماید، لذا تنها به متغیرهای اصلیتر و مهمتر میپرداخت. این امر منجر به نوعی تقلیلگرایی ۳ میشد و پدیدههای پیچیده را به متغیرهای عددی ساده‌شدهای تبدیل مینمود. به این ترتیب، مطرح شدن مسائلی از این دست، پیشبینی را به چالشهای عمیقی کشاند. چالشهایی که به ظهور دانشهای جدیدی در خصوص آینده و نیز به شکلگیری حوزه آیندهنگاری انجامید (بورژویس ۴، ۲۰۰۱). بنابر تعریف «مارتین ۵» یکی از صاحب‌نظران این حوزه، آیندهنگاری فرایندی نظاممند با نگاه به آینده بلندمدت در زمینه‌های علمی، فناوریانه، اقتصادی و اجتماعی است که هدف آن تعیین حوزه‌های تحقیقات راهبردی و پیدایش فناوریهای نوظهور با بیشترین فواید اقتصادی و اجتماعی است (مارتین، ۱۹۹۵). به طور کلی، آیندهنگاری در پی شناسایی آینده‌های بدیل و در نتیجه ایجاد فرصتی برای شکلهای آینده، افزایش انعطافپذیری در سیاستگذاری و ایجاد دیدگاههای کلنگر است. آیندهنگاری میتواند اولاً داده‌های ضروری را برای سیاستگذاری در زمینه‌های متفاوت ارائه نماید، ثانیاً نشانه‌های ضعیف اما مهمی که منجر به ارزیابی و تنظیم مجدد سیاستها خواهد شد را شناسایی کند. امروزه، بسیاری از حکومتها به اهمیت فعالیتهای آیندهنگاری پیبرده‌اند و از اینرو این ابزار نسبتاً جدید و نوآورانه کمک به سیاست علم، فناوری در سراسر جهان در حال گسترش است. آیندهنگاری اولین بار به عنوان یک ابزار سیاستگذاری در اواخر دهه ۵۰ و اوایل دهه ۶۰ میلادی در آمریکا به ویژه در بخش دفاعی به کار رفت و بعد از آن ژاپن با یک افق زمانی ۳۰ ساله آن را انجام داد و سپس به تدریج در بین سایر ملتها رواج یافت. لذا با توجه به اهمیتی که موضوع آیندهنگاری در مباحث سیاستگذاری دارد، تاکید اصلی کتاب حاضر بر این است که زمانی میتوان یک فعالیت آیندهنگاری را اثربخش تلقی نمود که بتوان آثار و نتایج آن را به گونهای مناسب در سیاستگذاری و تصمیم‌گیریهای سیاستی ملاحظه نمود (یونیدو ۶، ۲۰۰۵). از این رو این کتاب بر نقش آیندهنگاری علم و فناوری، به عنوان ابزار هوشمندی سیاستی تاکید دارد و بر پایه همین مفهوم، به دنبال آن است که ارتباط و تأثیرگذاری آیندهنگاری را بر سیاستگذاری علم، فناوری و نوآوری در سطح ملی مورد بررسی قرار دهد و همین امر مهمترین تمایز این اثر با آثار پیشین است. لازم به ذکر است، نوع نگاه این کتاب، فرصت‌هایی را برای سیاستگذاران فراهم میکند تا سیاستهای راهبردی را با بهره‌گیری از بهترین شواهد حاصل از تحقیقات آیندهنگاری تدوین نمایند و میتواند اثرات مهمی بر بازیگران سیاستی داشته باشد؛ همچنین، کتاب حاضر تلاش

نموده است تا با ارائه پیشنهاداتی، به سیاستگذاران، دانشگاهیان، مشاوران و سایر افرادی که بر تولید نتایج اثر بخش آینده‌نگاری تأثیرگذارند، کمک نماید. بر همین اساس، محتوی این کتاب در هفت فصل تنظیم گردیده است.

مواردی که باید اصلاح شوند، شکل درست نگارشی آن‌ها در متن زیر وارد شده است:

یکی از نخستین تلاش‌های علمی بشر برای شناخت آینده، استفاده از روش «پیش‌بینی ۱» بوده است. پیش‌بینی سعی می‌کند تا به یک پیشگویی ۲ دقیق از آینده دست یابد. با گذشت زمان، پیش‌بینی مورد انتقادات فراوانی قرار گرفت. بیشترین انتقادات مطرح شده، در خصوص تکیه پیش‌بینی بر گذشته و افق زمانی کوتاه‌مدت آن بوده است. پیش‌بینی، آینده را بر طبق گذشته کشف می‌کرد و همین امر باعث تشکیک افراد بسیاری درباره آن شد؛ زیرا هیچ دلیلی وجود نداشت که الزاماً آینده از گذشته پیروی نماید. از سویی دیگر و به واسطه مدل‌سازیهای افراطی که با استفاده از ابزار پیش‌بینی انجام می‌گرفت، ناگزیر بسیاری از متغیرهای تأثیرگذار را نادیده می‌انگاشتند. چون مدل‌سازی نمی‌توانست عیناً به تمامی متغیرها توجه کند و در مدل خود از همه آنها استفاده نماید، لذا تنها به متغیرهای اصلی‌تر و مهم‌تر می‌پرداخت. این امر منجر به نوعی تقلیل‌گرایی ۳ می‌شد و پدیده‌های پیچیده را به متغیرهای عددی ساده‌شده‌ای تبدیل می‌نمود. به این ترتیب، مطرح شدن مسائلی از این دست، پیش‌بینی را به چالش‌های عمیقی کشاند. چالش‌هایی که به ظهور دانش‌های جدیدی در خصوص آینده و نیز به شکل‌گیری حوزه آینده‌نگاری انجامید (بورژویس ۴، ۲۰۰۱). بنابر تعریف «مارتین ۵» یکی از صاحب‌نظران این حوزه، آینده‌نگاری فرایندی نظام‌مند با نگاه به آینده بلندمدت در زمینه‌های علمی، فناوری، اقتصادی و اجتماعی است که هدف آن تعیین حوزه‌های تحقیقات راهبردی و پیدایش فناوریهای نوظهور با بیشترین فواید اقتصادی و اجتماعی است (مارتین، ۱۹۹۵). به طور کلی، آینده‌نگاری در پی شناسایی آینده‌های بدیل و در نتیجه ایجاد فرصتی برای شکل‌دهی آینده، افزایش انعطاف‌پذیری در سیاستگذاری و ایجاد دیدگاه‌های کل‌نگر است. آینده‌نگاری می‌تواند اولاً داده‌های ضروری را برای سیاستگذاری در زمینه‌های متفاوت ارائه نماید، ثانیاً نشانه‌های ضعیف اما مهمی که منجر به ارزیابی و تنظیم مجدد سیاستها خواهد شد را شناسایی کند. امروزه، بسیاری از حکومتها به اهمیت فعالیتهای آینده‌نگاری پی‌برده‌اند و از اینرو این ابزار نسبتاً جدید و نوآورانه کمک به سیاست علم، فناوری در سراسر جهان در حال گسترش است. آینده‌نگاری اولین بار به عنوان یک ابزار سیاستگذاری در اواخر دهه ۵۰ و اوایل دهه ۶۰ میلادی در آمریکا به ویژه در بخش دفاعی به کار رفت و بعد از آن ژاپن با یک افق زمانی ۳۰ ساله آن را انجام داد و سپس به تدریج در بین سایر ملتها رواج یافت. لذا با توجه به اهمیتی که موضوع آینده‌نگاری در مباحث سیاستگذاری دارد، تاکید اصلی کتاب حاضر بر این است که زمانی می‌توان یک فعالیت آینده‌نگاری را اثربخش تلقی نمود که بتوان آثار و نتایج آن را به گونه‌ای مناسب در سیاستگذاری و تصمیم‌گیریهای سیاستی ملاحظه نمود (یونیدو ۶، ۲۰۰۵). از این رو این کتاب بر نقش آینده‌نگاری علم و فناوری، به عنوان ابزار هوشمندی سیاستی تاکید دارد و بر پایه همین مفهوم، به دنبال آن است که ارتباط و تأثیرگذاری آینده‌نگاری را بر سیاستگذاری علم، فناوری و نوآوری در سطح ملی مورد بررسی قرار دهد و همین امر مهمترین تمایز این اثر با آثار پیشین است. لازم به ذکر است، نوع نگاه این کتاب، فرصت‌هایی را برای سیاستگذاران فراهم می‌کند تا سیاستهای راهبردی را با بهره‌گیری از بهترین شواهد حاصل از تحقیقات آینده‌نگاری تدوین نمایند و می‌تواند اثرات مهمی بر بازیگران سیاستی داشته باشد؛ همچنین، کتاب حاضر تلاش نموده است تا با ارائه پیشنهاداتی، به سیاستگذاران، دانشگاهیان، مشاوران و سایر افرادی که بر

تولید نتایج اثر بخش آینده‌نگاری تأثیرگذارند، کمک نماید. بر همین اساس، محتوی این کتاب در هفت فصل تنظیم گردیده است

تعداد واژه‌های متن ورودی ۶۰۳ واژه بوده است. تعداد مواردی که نیاز به اصلاح نگارشی داشتند، ۵۷ مورد بوده است. ۳۲ مورد توسط ابزار هضم و بررسی ساختوازی اصلاح شده است؛ در حقیقت، ۳۲ مورد به صورت ماشینی اصلاح شده است. بنابراین، ۵۶٪ به صورت ماشینی اصلاح شده است و بقیه موارد به صورت دستی اصلاح شده است. سپس، متن اصلاح شده وارد بخش برچسب گذاری شد.

وجود	N	AJe	زمانی	N	آینده	NUM	یکی
نداشت	V	AJe	کوتاه‌مدت	N	دست	P	از
که	CONJ	PRO	آن	V	یابد	Ne	نخستین
الزاما	ADV	V	بوده_است	PUNC	.	AJe	تلاشهای
آینده	N	PUNC	.	P	با	AJe	علمی
از	P	N	پیش‌بینی	Ne	گذشت	N	بشر
گذشته	N	PUNC	،	N	زمان	Pe	برای
پیروی	N	N	آینده	PUNC	،	Ne	شناخت
نماید	V	POSTP	را	Ne	پیش‌بینی	N	آینده
	PUNC	P	بر	Ne	مورد	PUNC	،
از	P	Ne	طبق	Ne	انتقادات	N	استفاده
سوی	N	N	گذشته	AJ	فراوانی	P	از
دیگر	PRO	N	کشف	N	قرار	Ne	روش
و	CONJ	V	می‌کرد	V	گرفت	PUNC	«
به	P	CONJ	و	PUNC	.	N	پیش‌بینی ۱
واسطه	Ne	DET	همین	AJ	بیشترین	PUNC	»
مدل‌سازیهای	Ne	N	امر	N	انتقادات	V	بوده_است
افراطی	AJ	AJe	باعث	AJ	مطرح	PUNC	.
که	CONJ	Ne	تشکیک	V	شده	N	پیش‌بینی
با	P	Ne	افراد	PUNC	،	N	سعی
استفاده	N	AJ	بسیاری	P	در	V	می‌کند
از	P	Pe	درباره	Ne	خصوص	CONJ	تا
ابزار	Ne	PRO	آن	N	تکیه	P	به
پیش‌بینی	N	V	شد	N	پیش‌بینی	NUM	یک
انجام	N	PUNC	؛	P	بر	Ne	پیشگویی ۲
می‌گرفت	V	CONJ	زیرا	N	گذشته	AJ	دقیق
	PUNC	DET	هیچ	CONJ	و	P	از
ناگزیر	ADV	N	دلیلی	Ne	افق		

CONJ	که	CONJ	نیز	N	تقلیل‌گرایی ^۳	DET	بسیاری
Ne	هدف	P	به	V	می‌شد	P	از
PRO	آن	Ne	شکل‌گیری	CONJ	و	Ne	متغیرهای
Ne	تعیین	Ne	حوزه	Ne	پدیده‌های	AJ	تأثیرگذار
Ne	حوزه‌های	N	آینده‌نگاری	AJ	پیچیده	POSTP	را
Ne	تحقیقات	V	انجامید	POSTP	را	AJ	نادیده
AJ	راهبردی	PUNC	(	P	به	V	می‌انگاشتند
CONJ	و	N	بورژوازی ^۴	Ne	متغیرهای	PUNC	.
Ne	پیدایش	PUNC	,	AJe	عددی	CONJ	چون
Ne	فناوریهای	NUM	۲۰۰۱	AJ	ساده‌شده‌ای	N	مدل‌سازی
AJ	نوظهور	PUNC	)	N	تبدیل	V	نمی‌توانست
P	با	PUNC	.	V	می‌نمود	ADV	عینا
AJ	بیشترین	P	بنابر	PUNC	.	P	به
Ne	فواید	N	تعریف	P	به	DETe	تمامی
AJ	اقتصادی	PUNC	«	DET	این	N	متغیرها
CONJ	و	N	مارتین ^۵	N	ترتیب	N	توجه
AJ	اجتماعی	PUNC	»	PUNC	,	V	کند
V	است	NUM	یکی	AJ	مطرح	CONJ	و
PUNC	(	P	از	Ne	شدن	P	در
N	مارتین	Ne	صاحب‌نظران	N	مسائلی	Ne	مدل
PUNC	,	DET	این	P	از	PRO	خود
NUM	۱۹۹۵	N	حوزه	DET	این	P	از
PUNC	)	PUNC	,	N	دست	Ne	همه
PUNC	.	N	آینده‌نگاری	PUNC	,	PRO	آنها
P	به	N	فرایندی	N	پیش‌بینی	N	استفاده
Ne	طور	AJ	نظام‌مند	POSTP	را	V	نماید
AJ	کلی	P	با	P	به	PUNC	,
PUNC	,	N	نگاه	Ne	چالشهای	CONJ	لذا
N	آینده‌نگاری	P	به	AJ	عمیقی	ADV	تنها
P	در	Ne	آینده	V	کشاند	P	به
Ne	بی	AJ	بلندمدت	PUNC	.	Ne	متغیرهای
Ne	شناسایی	P	در	N	چالشهایی	AJ	اصلی‌تر
Ne	آینده‌های	Ne	زمینه‌های	CONJ	که	CONJ	و
AJ	بدیل	AJ	علمی	P	به	AJ	مهم‌تر
CONJ	و	PUNC	,	Ne	ظهور	V	می‌پرداخت
P	در	AJ	فناورانه	Ne	دانشهای	PUNC	.
N	نتیجه	PUNC	,	AJ	جدیدی	DET	این
Ne	ایجاد	AJ	اقتصادی	P	در	N	امر
N	فرصتی	CONJ	و	Ne	خصوص	AJ	منجر
Pe	برای	AJ	اجتماعی	N	آینده	P	به
Ne	شکل‌دهی	V	است	CONJ	و	N	نوعی

آینده	AJ	شناسایی	N	یک	NUM	در	P
،	PUNC	کند	V	ابزار	Ne	بین	Ne
افزایش	Ne		PUNC	سیاستگذاری	N	سایر	Ne
انعطاف‌پذیری	N	امروزه	ADV	در	P	ملتها	N
در	P	،	PUNC	اواخر	Ne	رواج	N
سیاستگذاری	N	بسیاری	DET	دهه	Ne	یافت	V
و	CONJ	از	P	۵۰	NUM		PUNC
ایجاد	Ne	حکومتها	N	و	CONJ	لذا	CONJ
دیدگاه‌های	Ne	به	P	اوایل	Ne	با	P
کل‌نگر	AJ	اهمیت	Ne	دهه	Ne	توجه	N
است	V	فعالیت‌های	Ne	۶۰	NUMe	به	P
	PUNC	آینده‌نگاری	AJ	میلادی	AJ	اهمیتی	N
آینده‌نگاری	N	پی‌برده‌اند	V	در	P	که	CONJ
می‌تواند	V	و	CONJ	آمریکا	N	موضوع	Ne
اولا	ADV	از	P	به	P	آینده‌نگاری	AJ
داده‌های	Ne	اینرو	N	ویژه	AJ	در	P
ضروری	AJ	این	DET	در	P	مباحث	Ne
را	POSTP	ابزار	Ne	بخش	Ne	سیاستگذاری	N
برای	Pe	نسبتا	ADV	دفاعی	AJ	دارد	V
سیاستگذاری	N	جدید	AJ	به	P	،	PUNC
در	P	و	CONJ	کار	N	تاکید	Ne
زمینه‌های	Ne	نواورانه	Ne	رفت	V	اصلی	AJe
متفاوت	AJ	کمک	N	و	CONJ	کتاب	Ne
ارائه	N	به	P	بعد	P	حاضر	AJ
نماید	V	سیاست	Ne	از	P	بر	P
،	PUNC	علم	N	آن	PRO	این	PRO
ثانیا	ADV	،	PUNC	ژاپن	N	است	V
نشانه‌های	Ne	فناوری	N	با	P	که	CONJ
ضعیف	AJ	در	P	یک	NUM	زمانی	N
اما	CONJ	سراسر	Ne	افقی	Ne	می‌توان	V
مهمی	AJ	جهان	N	زمانی	AJe	یک	NUM
که	CONJ	در	P	۳۰	NUM	فعالیت	Ne
منجر	AJ	حال	Ne	ساله	RESe	آینده‌نگاری	AJ
به	P	گسترش	N	آن	PRO	را	POSTP
ارزیابی	N	است	V	را	POSTP	اثربخش	AJ
و	CONJ		PUNC	انجام	N	تلقی	N
تنظیم	Ne	آینده‌نگاری	N	داد	V	نمود	V
مجدد	AJe	اولین	NUM	و	CONJ	که	CONJ
سیاستها	N	بار	N	سپس	CONJ	بتوان	V
خواهد شد	V	به	P	به	P	آثار	N
را	POSTP	عنوان	Ne	تدریج	N	و	CONJ

نتایج	Ne	همین	DET	ذکر	N	حاضر	AJ
آن	PRO	مفهوم	N	است	V	تلاش	N
را	POSTP	،	PUNC	،	PUNC	نموده_است	V
به	P	به	P	نوع	Ne	تا	CONJ
گونه‌ای	N	دنبال	Ne	نگاه	Ne	با	P
مناسب	AJ	آن	PRO	این	DET	ارائه	Ne
در	P	است	V	کتاب	N	پیشنهاداتی	N
سیاستگذاری	N	که	CONJ	،	PUNC	،	PUNC
و	CONJ	ارتباط	N	فرصتهایی	N	به	P
تصمیم‌گیریهای	Ne	و	CONJ	را	POSTP	سیاستگذاران	N
سیاستی	N	تأثیرگذاری	Ne	برای	Pe	،	PUNC
ملاحظه	N	آینده‌نگاری	N	سیاستگذاران	N	دانشگاهیان	N
نمود	V	را	POSTP	فراهم	N	،	PUNC
(	PUNC	بر	P	می‌کند	V	مشاوران	N
یونیدو۶	N	سیاستگذاری	Ne	تا	CONJ	و	CONJ
،	PUNC	علم	N	سیاستهای	Ne	سایر	Ne
۲۰۰۵	NUM	،	PUNC	راهبردی	AJ	افرادی	N
)	PUNC	فناوری	N	را	POSTP	که	CONJ
.	PUNC	و	CONJ	با	P	بر	P
از	P	نوآوری	N	بهره‌گیری	N	تولید	Ne
این	DET	در	P	از	P	نتایج	Ne
رو	N	سطح	Ne	بهترین	AJ	اثر	Ne
این	DET	ملی	AJ	شواهد	Ne	بخش	Ne
کتاب	N	مورد	Ne	حاصل	N	آینده‌نگاری	N
بر	P	بررسی	N	از	P	تأثیرگذارند	V
نقش	Ne	قرار	N	تحقیقات	Ne	،	PUNC
آینده‌نگاری	Ne	دهد	V	آینده‌نگاری	N	کمک	N
علم	N	و	CONJ	تدوین	N	نماید	V
و	CONJ	همین	DET	نمایند	V	.	PUNC
فناوری	N	امر	N	و	CONJ	برهمن	DET
،	PUNC	مهمترین	AJ	می‌تواند	V	اساس	N
به	P	تمایز	Ne	اثرات	Ne	،	PUNC
عنوان	Ne	این	DET	مهمی	AJ	محتوی	Ne
ابزار	Ne	اثر	N	بر	P	این	DET
هوشمندی	AJe	با	P	بازیگران	Ne	کتاب	N
سیاستی	N	آثار	Ne	سیاستی	N	در	P
تاکید	N	پیشین	AJ	داشته_باشد	V	هفت	NUM
دارد	V	است	V	؛	PUNC	فصل	N
و	CONJ	.	PUNC	همچنین	CONJ	تنظیم	N
بر	P	لازم	AJ	،	PUNC	گردیده	V
پایه	Ne	به	P	کتاب	Ne	است	V

در فرآیند ساخت پیکره سعی شده است تا حد امکان کنترل دستی به صورت دقیق انجام شود و زمان بسیاری صرف این بخش شده است. البته علیرغم کنترل ماشینی و دستی برچسب‌ها، گم‌اگان خطاهایی ممکن است در برچسب‌گذاری مشاهده شود. اما این خطاها بسیار کم و قابل چشم‌پوشی است. به عنوان نمونه، در بریده‌ای از متن برچسب‌گذاری شده که از سامانه پیکره‌های ایرانداک (سایا) گرفته شده است، حاصل جستجوی یک واژه در پیکره کتاب‌های دیجیتال ایرانداک (پکا) است، تنها سه خطای برچسب‌گذاری مشاهده می‌شود که با توجه به امکانی که در سامانه پیکره‌های ایرانداک (سایا) فراهم شده است، می‌توان خطاهای نگارشی یا برچسب‌گذاری را اصلاح نمود.

پیشگفتار بازی‌وارسازی (Ne) با gamification (CON) مفهومی جدید (AJ) در (P) کسب (Ne) و (CON) کار (Ne) بوده (CON) در (P) تلاش (Ne) است (P) با (P) استفاده (Ne) از (P) سرگرمی (Ne) و (CON) مفاهیم (Ne) مطرح (Ne) در (P) بازیها (PUNO) ، (Ne) کاربران (Ne) را (POSTP) به (P) آغاز (Ne) و (CON) یا (CON) ادامه (Ne) فعالیت (Ne) تشویق (Ne) کرده (PUNO) ، (AJ) مشارکت (Ne) آنان (PRO) را (POSTP) افزایش (Ne) داده (PUNO) ، (Ne) برخی (Ne) از (DET) چالشها (Ne) را (POSTP) حل (Ne) نموده (Ne) و (CON) یا (CON) رفتاری (Ne) خاص (AJ) را (POSTP) در (P) جامعه (Ne) هدف (Ne) خود (PRO) ایجاد (Ne) نماید (PUNO) . (Ne) به (P) سخن (Ne) دیگر (PUNO) ، (PRO) بازی‌وارسازی (Ne) ابزاری (AJ) برای (Ne) حل (Ne) چالشهای (Ne) کسب (Ne) و (CON) کار (Ne) است (Ne) که (CON) یا (P) کاربران (Ne) و (CON) مشارکت (Ne) آنان (PRO) در (P) ارتباط (Ne) بوده (Ne) و (CON) از (P) طریق (Ne) بکارگیری (Ne) سرگرمی (Ne) انجام (Ne) می‌شود (PUNO) . بازی‌وارسازی (Ne) به (P) معنای (Ne) تبدیل (Ne) تمام (AJ) کسب (Ne) و (CON) کار (Ne) به (P) بازی (Ne) نیست (Ne) بلکه (CON) به (P) معنای (Ne) استفاده (Ne) از (P) ابزارها (Ne) و (CON) روشهای (Ne) مورد (Ne) استفاده (Ne) در (P) بازیها (Ne) برای (Ne) حل (Ne) چالشهای (Ne) مرتبط (AJ) با (P) مشارکت (Ne) کاربران (Ne) در (P) کسب (Ne) و (CON) کار (Ne) است (Ne) حتی (ADV) اگر (CON) راه‌حلهای (Ne) ارائه (Ne) شده (Ne) شبیه (AJ) بازی (Ne) نباشند (PUNO) . (Ne) بر (P) این (DET) اساس (Ne) برای (Ne) طراحی (Ne) و (CON) اجرای (Ne) موفق (AJ) بازی‌وارسازی (Ne) تسلط (Ne) بر (P) دو (NUM) حوزه (Ne) دانش (Ne) مورد (Ne) نیاز (Ne) است (PUNO) : (Ne) دانش (Ne) حوزه (Ne) کسب (Ne) و (CON) کار (Ne) و (CON) دانش (Ne) حوزه (Ne) طراحی (Ne) بازی (PUNO) . (Ne) با (P) ترکیب (Ne) این (DET) دو (NUM) حوزه (Ne) و (CON) تعامل (Ne) مناسب (AJ) آنها (PRO) می‌توان (Ne) روشیابی (Ne) را (POSTP) طراحی (Ne) و (CON) پیاده (AJ) نمود (Ne) که (CON) با (P) کمک (Ne) آنها (PRO) چالشهای (Ne) کسب (Ne) و (CON) کار (Ne) برطرف (AJ) شده (Ne) و (CON) اشتیاق (Ne) و (CON) انگیزه (Ne) کاربران (Ne) افزایش (Ne) یابد (PUNO) . (Ne) با (P) این (DET) همه (PUNO) ، (PRO) بدون (Ne) داشتن (Ne) الگویی (Ne) جامع (AJ) و (CON) کل‌نگر (AJ) و (CON) بدون (Ne) بهره‌گیری (Ne) از (P) روشی (Ne) ساختارمند (AJ) نمیتوان (Ne) بازی‌وارسازی (Ne) را (POSTP) پیاده‌سازی (Ne) کرد (PUNO) . (Ne) کتاب (Ne) حاضر (AJ) به (P) دنبال (Ne) ارائه (Ne) چنین (Ne) الگویی (Ne) است (Ne) که (CON) در (P) طی (Ne) آن (PRO) به (P) کاربران (PUNO) ، (Ne) نیازمندین (Ne) و (CON) الزامات (Ne) آنان (PRO) توجه (Ne) شده (Ne) و (CON) روشهای (Ne) بازی‌وارسازی (Ne) به (P) شکلی (Ne) ساختارمند (AJ) برای (Ne) پاسخ (Ne) به (P) این (DET) نیازمندین (Ne) در (P) تعامل (Ne) با (P) یکدیگر (PRO) اجرا (Ne) شوند (PUNO) . (Ne) این (DET) الگو (Ne) از (P) سه (NUM) عامل (Ne) کلیدی (AJ) طرح (Ne) به (P) عنوان (Ne) شکل (Ne) دهنده (AJ) بازی‌وارسازی (Ne) و (CON) هدایت‌کننده (Ne) فعالیتها (PUNO) ، (Ne) سیستم (Ne) به (P) عنوان (Ne) آن‌چیزی (Ne) که (CON) به (P) دنبال (Ne) بازی‌وارسازی (Ne) آن (PRO) هستیم (Ne) و (CON) کاربران (Ne) به (P) عنوان (Ne) مشارکت‌کنندگان (Ne) و (CON) تعامل‌کنندگان (Ne) با (P) سیستم (Ne) تشکیل (Ne) شده (Ne) است (PUNO) . (Ne) در (P) طراحی (Ne) سیستم (Ne) عناصری (Ne) وجود (Ne) دارند (Ne) که (CON) در (P) تعامل (Ne) با (P) کاربران (Ne) بوده (Ne) و (CON) آنان (PRO) را (POSTP) درگیر (AJ) می‌سازد (PUNO) . (Ne) امتیازات (PUNO) ، (Ne) نشانها (Ne) و (CON) تالیلی (Ne) امتیازات (Ne) از (P) پربکارترین (Ne) این (DET) عناصر (Ne) هستند (Ne) که (CON) برای (Ne) اجرای (Ne) موفق (AJ) بازی‌وارسازی (Ne) می‌توان (Ne) در (P) طی (Ne) الگویی (Ne) هدفمند (AJ) طراحی (Ne) شده (Ne) و (CON) مورد (Ne) استفاده (Ne) قرار (Ne) گیرد (PUNO) . (Ne) در (P) کتاب (Ne) حاضر (AJ) این (DET) الگو (Ne) در (P) قالب (Ne) محرکهای (Ne) هشگانه (AJ) مطرح (Ne) شده (Ne) و (CON) عناصر (Ne) مطرح (Ne) در (P) بازی‌وارسازی (Ne) در (P) قالب (Ne) آنها (PRO) بازتابی (Ne) میشوند (PUNO) . همچنین (Ne) با (P) بهره‌گیری (Ne) از (P) این (DET) ساختار (Ne) تلاش (Ne) شده (Ne) است (Ne) تا (CON) برخی (Ne) از (DET) نمونه‌های (Ne) موفق (AJ) کسب (Ne) و (CON) کار (Ne) در (P) حوزه‌های (Ne) آموزش (Ne) و (CON) یادگیری (PUNO) ، (Ne) سلامت (PUNO) ، (Ne) گردش‌گری (PUNO) ، (Ne) حمل (Ne) و (CON) نقل (Ne) و (CON) کسب (Ne) و (CON) کار (Ne) الکترونیک (Ne) بررسی (Ne) گردند (PUNO) . (Ne) بر (P) این (DET) اساس (PUNO) ، (Ne) در (P) فصل (Ne) اول (NUM) تعاریف (Ne) و (CON) مفاهیم (Ne) اولیه (AJ) بازی‌وارسازی (PUNO) ، (Ne) برخی (Ne) از (DET) نمونه‌های (Ne) موفق (AJ) و (CON) ناموفق (AJ) اجرای (Ne) آن (PRO) بررسی (Ne) و (CON) عوامل (Ne) موثر (Ne) بر (P) بازی‌وارسازی (Ne) ارائه (Ne) میشوند (PUNO) . فصلهای (Ne) دوم (NUM) و (CON) سوم (NUM) نگاهی (Ne) دقیقتر (AJ) به (P) عوامل (Ne) موثر (AJ) بر (P) بازی‌وارسازی (Ne) شامل (AJ) طرح (Ne) بازی‌وارسازی (PUNO) ، (Ne) کاربر (Ne) بازی‌وارسازی (Ne) و (CON) سیستم (Ne) دارد (PUNO) . طرح (Ne) شکل‌دهنده (AJ) بازی‌وارسازی (Ne) بوده (Ne) و (CON) نسبت (Ne) به (P) الزامات (Ne) کاربران (Ne) از (P) دانش (Ne) کافی (AJ) برخوردار (AJ) است (PUNO) . طرح (Ne) باید (Ne) بداند (Ne) که (CON) چه (DET) خدماتی (PUNO) ، (Ne) چه (DET) محصولاتی (Ne) و (CON) چه (DET) مفاهیمی (Ne) در (P) حوزه (Ne) کسب (Ne) و (CON) کار (Ne) وجود (Ne) دارند (Ne) که (CON) با (P) کمک (Ne) آنان (PRO) سیستم (Ne) قادر (AJ) خواهد بود (Ne) کاربرانی (Ne) وفادارتر (Ne) و (CON) مشتاق‌تر (AJ) داشته باشد (PUNO) . ویژگی‌های (Ne) اصلی (AJ) کاربران (Ne) و (CON) نیازمندین (Ne) آنان (PRO) به (P) همراه (Ne) تجارب (Ne) و (CON) چرخه (Ne) عمر (Ne) آنان (PRO) در (P) سیستم (Ne) در (P) بخش (Ne) مربوط (AJ) به (P) کاربر (Ne) بررسی (Ne) شده (Ne) و (CON) بخش (Ne) مربوط (AJ) به (P) سیستم (Ne) به (P) بررسی (Ne) و (CON) الزاماتی (Ne) می‌پردازد (Ne) که (CON) در (P) انتقال (Ne) تجارب (Ne) مورد (Ne) نظر (Ne) طرح (Ne) به (P) کاربر (Ne) موثرند (PUNO) . محرکها (Ne) یکی (NUM) از (P) نکات (Ne) کلیدی (AJ) در (P) این (DET) فصل (Ne) است (Ne) که (CON) به (P) طور (Ne) مفصل (Ne) بدان (P) پرداخته (Ne) شده (Ne) و (CON) الگویی (Ne) در (P) قالب (Ne) محرکهای (Ne) هشگانه (AJ) برای (Ne) اطمینان (Ne) از (P) پیاده‌سازی (Ne) درست (AJ) آنها (PRO) در (P) فصل (Ne) سوم (NUM) ارائه (Ne) شده (Ne) است (PUNO) . فصل (Ne) چهارم (NUM) کتاب (Ne) حاضر (AJ) نیز (ADV) به (P) بررسی (Ne) کاربردهای (Ne) بازی‌وارسازی (Ne) در (P) حوزه‌های (Ne) آموزش (Ne) و (CON) یادگیری (PUNO) ، (Ne) سلامت (PUNO) ، (Ne) گردشگری (PUNO) ، (Ne) حمل (Ne) و (CON) نقل (Ne) و (CON) کسب (Ne) و (CON) کار (Ne) الکترونیک (AJ) در (P) قالب (Ne) الگویی (Ne) ارائه (Ne) شده (Ne) اختصاصی (Ne) دارد (PUNO) . فصل (Ne) اول (NUM) : بازی‌وارسازی (Ne) چیست (Ne) ؟ پیش (ADV) از (P) آنکه (CON) به (P) بازی‌وارسازی (PUNO) ، (Ne) مفاهیم (Ne) و (CON) ابزارهای (Ne) آن (PRO) پرداخته شود (Ne) ضروری (AJ) است (Ne) تا (CON) درباره (Ne) ویژگیها (Ne) و (CON) خصوصیات (Ne) بازی (Ne) ، (Ne) به (P) عنوان (Ne) محور (Ne) بازی‌وارسازی (Ne) بحث (Ne) شود (PUNO) . بازی (Ne) عموماً (ADV) به (P) فعالیت (Ne) گفته (Ne) می‌شود (Ne) که (CON) با (P) مسا (Ne) و (CON) اختار (Ne) و (CON)

۷-۴- ارتقاء سامانه پیکره‌های ایرانداک (ساپا)

- کارهای انجام شده به منظور ارتقاء سامانه پیکره‌های ایرانداک شامل موارد زیر است:
- یکپارچه شدن پایگاه داده، که مهمترین اقدام در ارتقاء سامانه است: به دلیل شرایط خاص استفاده از کتابخانه هضم در طرح قبلی که در آن پایگاه داده mysql و sqlserver استفاده شده بود، مشکلاتی در یونیکدها در هنگام انتقال داده از پایگاهی به پایگاه دیگر ایجاد می‌شد. در این طرح پژوهشی، تنها از یک پایگاه داده استفاده شد و مشکلات انتقال داده بین دو پایگاه برطرف شد.
 - در طرح پیشین نگارنده (ساخت پیکره پژوهش‌نامه)، اصلاح برچسب‌ها از طریق سامانه بسیار زمان‌بر بود؛ به این صورت که به ازای هر تغییر، عملیات ذخیره‌سازی صورت می‌گرفت. در این طرح پژوهشی، بروزرسانی در سامانه به نحوی انجام می‌شود که تغییرات فقط در انتهای کار بوده و می‌توان چندین ویرایش را انجام داد و سپس، یکجا همه تغییرات ذخیره شوند.
 - افزودن بخش اصلاح کلی برچسب‌ها: بدین صورت که اگر از طریق این بخش اقدام به اصلاح برچسبی صورت گیرد، آن برچسب برای کل پایگاه بروزرسانی می‌شود.
 - با افزودن جستجوی واژه به همراه برچسب قابلیت جستجوی تخصصی به مجموعه اضافه شد.
 - اضافه شدن راهنمای برچسب‌ها
 - بروزرسانی ابزار برجسته کردن واژه جستجو شده و رفع نواقص قبلی

۷-۴- خلاصه فصل

در این فصل چگونگی هنجارسازی داده‌های وارد شده در پیکره توضیح داده شد که در این راستا از ابزار هضم، اطلاعات ساختوازی زبان فارسی (وندهای اشتقاقی و تصریفی و واژه‌بست‌ها) و برنامه‌نویسی ماشینی استفاده شد. سپس مفهوم «حاشیه‌نویسی» توضیح داده شد و نوع حاشیه‌نویسی/برچسب‌گذاری استفاده شده در پیکره و ابزار استفاده شده برای این منظور، معرفی شد و در نهایت نحوه کنترل دستی برچسب‌ها توضیح داده شد.

فصل پنجم

نتیجه گیری

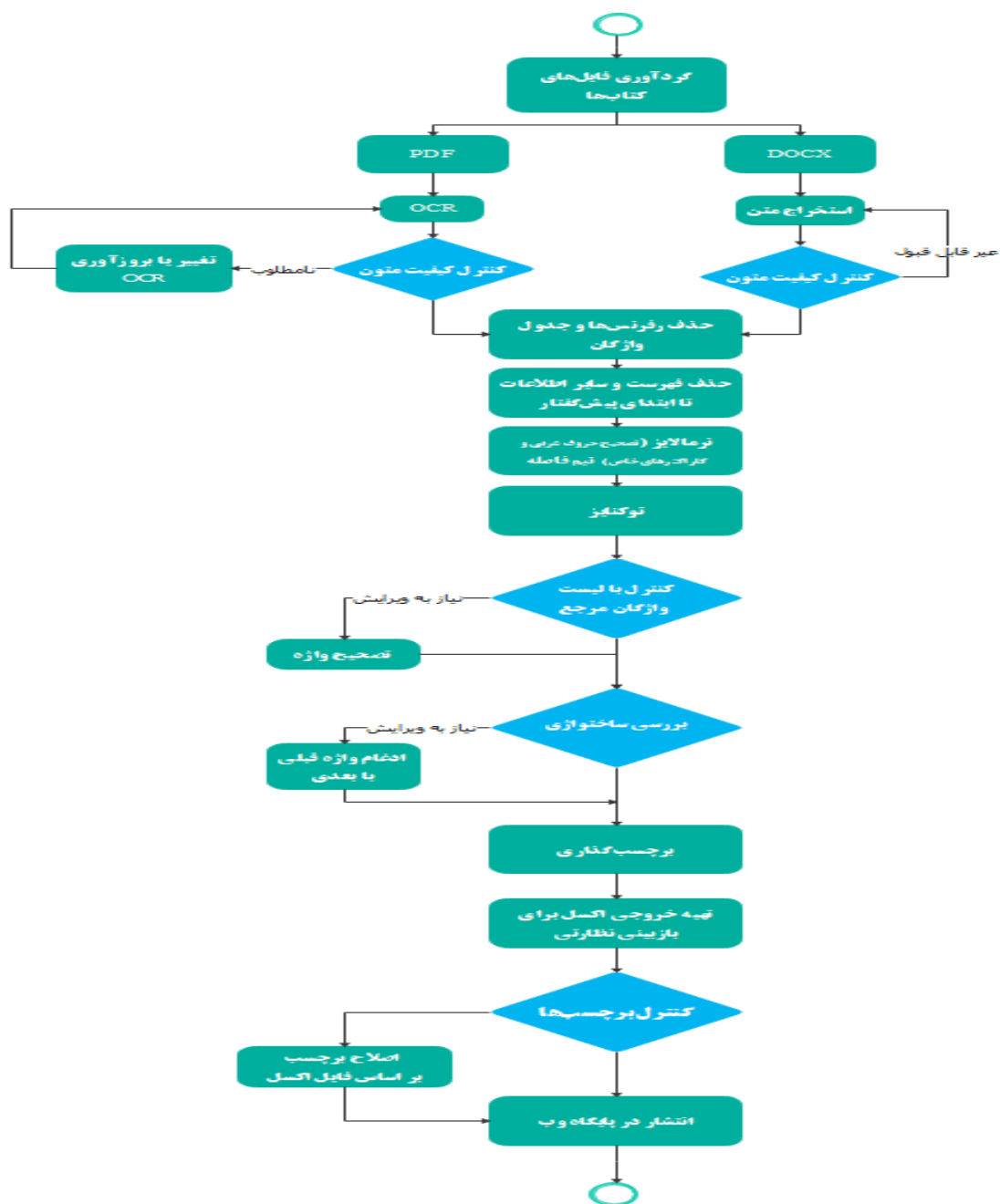
۵-۱- جمع‌بندی

یکی از روش‌های گردآوری منسجم داده‌ها، تهیه پیکره‌های زبانی است. تا کنون پیکره‌های عمومی بسیاری در زبان‌های گوناگون، از جمله زبان فارسی، ساخته شده است. اما تعداد پیکره‌های تخصصی که برای اهداف خاص و پردازش‌های خاص زبانی استفاده می‌شود، به اندازه پیکره‌های عمومی نیست. بنابراین، ساخت پیکره‌هایی که شامل متون تخصصی و میان‌رشته‌ای است، بسیار ارزشمند است. گزارش این طرح پژوهشی شامل پنج فصل است: پس از مقدمه، در فصل دوم به بررسی پژوهش‌هایی پرداخته شد که در حوزه ساخت پیکره انجام شده است. در این پژوهش‌ها، چکیده‌ای از تجربیات پژوهشگران این حوزه درباره ساخت پیکره‌ها و چالش‌هایی که هنگام ساخت پیکره‌های متنی یا نوشتاری با آن‌ها مواجه شده‌اند، ارائه شده است. برخی دیگر از این پژوهش‌ها شامل توصیف انواع روش‌های ساخت پیکره، نوع پیکره ساخته شده و حاشیه‌نویسی‌های استفاده شده در پیکره است. در بخش انتهایی این فصل به معرفی برخی از پیکره‌های زبان نیز فارسی پرداخته شد. در فصل سوم ابتدا به چگونگی فرآیند نمونه‌گیری پرداخته شد. سپس، تهیه فراداده‌ها توضیح داده شد. در ادامه به فرآیند تبدیل فایل‌های پی.دی.اف به ورد از طریق استفاده از ابزار اُ.سی.آر پرداخته شد و در نهایت چالش‌های این روش و چگونگی تمیز و مرتب کردن فایل خروجی پرداخته شد و در فصل چهارم، چگونگی هنجارسازی داده‌های وارد شده در پیکره توضیح داده شد که در این راستا از ابزار هضم، اطلاعات ساختوازی زبان فارسی (وندهای اشتقاقی و تصریفی و واژه‌بست‌ها) و برنامه‌نویسی ماشینی استفاده شد. سپس مفهوم «حاشیه‌نویسی» توضیح داده شد و نوع حاشیه‌نویسی/برچسب‌گذاری استفاده شده در پیکره و ابزار استفاده شده برای این منظور، معرفی شد و

ساخت پیکره از کتاب‌های دیجیتال پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) □ ۸۳

در نهایت نحوه کنترل دستی برچسب‌ها توضیح داده شد. نمودار جریانی^۱ از پیمانه‌های ساخت پیکره به این صورت است:

^۱ block diagram



ساخت پیکره از کتاب‌های دیجیتال پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) □ ۸۵

نمونه‌ای از جستجوی یک واژه در پیکره متنی ایرانداک به شکل زیر است:

ابتدا وارد لینک زیر می‌شویم:

<https://sapa.irandoc.ac.ir/>

این صفحه نمایش داده می‌شود:



The screenshot shows the homepage of the Sapa.irandoc.ac.ir website. The header is blue with the logo of the National Library and Archives of Iran on the left and the text 'سامانه پیکره‌های ایرانداک' (Sapa.irandoc.ac.ir) on the right. Below the header is a navigation bar with links for 'تلفن ما' (Our Phone), 'نام‌نویسی' (Registration), and 'بایگ‌ها' (Archives). The main content area features a large search bar with a magnifying glass icon and the text 'برای جست‌وجو در پیکره واژه‌ای را وارد کنید.' (Enter a word to search in the corpus). Below the search bar is a word cloud titled 'درباره سامانه' (About the System). The word cloud contains various words related to linguistics and corpus analysis, with 'corpus' being the most prominent. To the right of the word cloud is a text block titled 'درباره سامانه' (About the System) which describes the system's purpose and features.

پیکره، مجموعه‌ای نظام‌مند، رایانه‌ای، و درست از زبان است. بسیاری از پژوهش‌های زبان‌شناختی و تصمیم‌گیری‌ها در برنامه‌ریزی زبانی، تنها با کاربرد یک پیکره زبانی شدنی است. پیکره متنی ایرانداک در پژوهشگاه علوم و فناوری اطلاعات ایران ساخته شده و دارای ۴۷۷۴۹۰ واژه است. داده‌های پیکره از مقاله‌های پژوهش‌نامه پردازش و مدیریت اطلاعات گرفته شده‌اند. درون‌های این پیکره همگانی نیست و دارای نوشته‌های بسیار تخصصی و میان‌رشته‌مانند علم اطلاعات و دانش‌شناسی، فناوری اطلاعات، مدیریت دانش، زبان‌شناسی رایانشی، اصطلاح‌شناسی، و مانند آن‌هاست. از این رو نیز، برای پردازش‌هایی که نیازمند بهره‌گیری از نوشته‌های تخصصی باشند، بسیار ارزشمند است.

ساخت پیکره از کتاب‌های دیجیتال پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) □ ۸۷

جستجو

پکا

اجتماعی

شمار همخوانی‌ها با واژه درخواستی: ۲۵۲۲ واژه در ۴۱ مدرک

پیکره	عنوان مدرک	باقت	موضوع	پدیدآور	اطلاعات بیشتر
پکا	بازی‌وارسازی: مفاهیم و کاربردها	بازی‌وارسازی با افزایش میزان رضایت از عملکرد - ایجاد حس موفقیت در دستیابی به اهداف - تسهیل تعاملات اجتماعی و افزایش معاشرت‌های گروهی - حس خودمختاری و تسلط را در فرد ایجاد کرده و با ایجاد سرگرمی، کاربرانی مشتاق‌تر و وفادارتر را فراهم می‌آورد (بوهم ۱۲ و لیمبستر ۱۳۰۱۳)	فناوری اطلاعات	مرضیه زرین‌بال ماسوله	 مشاهده
پکا	ساختواره، اصطلاح‌شناسی و مهندسی دانش	مقدمه [GV] در جهان کنونی: « دانش جمعی » به‌عنوان نیروی محرکه و هدایت‌کننده در تحولات اجتماعی، سیاسی	زبان‌شناسی	ملوک‌السادات حسینی بهشتی	 مشاهده

برچسب‌ها

AJ ☐

AJe ☐

همه

ترو

برای دیدن واژه در متن هر مقاله می‌توان روی گزینه «مشاهده» کلیک کرد و صفحه زیر نمایش داده می‌شود:

عنوان	جهانی شدن و مدل توسعه نظام آموزش عالی در ایران با رویکرد فناوری اطلاعات و ارتباطات
جست‌وجو با	اجتماعی
شمار همخوانی‌ها	۵۲
پیداآور	محبوبه بیاتانی، حمیدرضا ریاحی، داوود کریمزادگان مقدم

تمام متن با برچسب

تمام متن بدون برچسب

پیشگفتار در این عصر که متأثر از پیشرفت فناوری اطلاعات و ارتباطات (فاوا) است ، عالم هستی به گونه‌ای ژرف با تحولات جهانی شدن در آمیخته و با تبادلات جهانی دگرگون گردیده‌است . این امر ، بازنگری تمام سیاست‌ها و عادات سازمانی را در برتو چالش جهانی‌شدن الزامی می‌سازد . بازنگری در آموزش عالی نیز از این قاعده مستثنی نیست ، به‌ویژه آنکه آموزش و پژوهش ، این دو عنصر اصلی آموزش عالی ، عناصر کلیدی در شکل‌گیری فضای جهانی هستند . بنابراین ، ضروری است برای توسعه نظام آموزش عالی هر کشوری متناسب با تحولات جهانی شدن و نقش‌آفرینی فاوا در توسعه نظامها و مؤسسات آموزش عالی ، برنامه‌ریزی و سیاست‌گذاری مناسبی از سطح حاکمیت گرفته تا سطح دانشگاهها به طور هماهنگ صورت پذیرد . فناوری اطلاعات و ارتباطات در عصر حاضر اساس توسعه و منشأ تحولات عظیم در نظام آموزش عالی و ابزار توانمندساز مهمی است . با این حال ، نظام آموزش عالی ایران از بسیاری از کاربردهای این فناوری بی‌نصیب بوده و حتی فاقد برنامه استراتژیک در این زمینه است . از سوی دیگر ، پژوهشی هم که عوامل و موانع توسعه نظام آموزش عالی کشور را با رویکرد فاوا از دیدگاه‌های بخشی و فرابخشی به طور یکجا مورد مطالعه قرار دهد ، وجود ندارد . این کتاب با تکیه بر تحقیقاتی وسیع ، منابع معتبر خارجی و داخلی در زمینه توسعه نظامها و مؤسسات آموزش عالی عصر حاضر را به‌ویژه با رویکرد فاوا مورد مطالعه قرار داده و با توجه به تأثیرات جهانی‌شدن و فناوری اطلاعات و ارتباطات در توسعه نظامها و مؤسسات آموزش عالی ، موانع و عوامل توسعه نظام آموزش عالی کشور را در سه سطح حاکمیت ، وزارت علوم ، و دانشگاهها استخراج می‌کند ؛ سپس مدل توسعه را در هر یک از این سه سطح ترسیم نموده و در مرحله بعد با توجه به نتایج حاصل از مطالعات و مدل‌های ترسیم‌شده و ارتباطات مستخرج از این سه مدل ، مدل نهایی برای توسعه نظام آموزش عالی کشور را با رویکرد فاوا از سطح حاکمیت تا سطح مؤسسات آموزش عالی ترسیم مینماید . شایان ذکر است که چون ترسیم چارچوب مفهومی این کتاب در قالب مدل‌های شمانیک نیازمند تعیین ابعاد ، مؤلفه‌ها و شاخصهای توسعه نظام آموزش عالی با رویکرد فاوا بود ، در استخراج مدل و اجزای آن از مطالعه اکتشافی و روش‌های استقرایی و قیاسی استفاده شده‌است ، ضمن آنکه مؤلفه‌ها و شاخصهای مدل در ارزیابی خبرگان آزمون گردیده و مورد تأیید قرار گرفته‌است . به عبارت دیگر ، باید گفت که این کتاب با توجه به تأثیرات جهانی‌شدن و پیشرفت فناوری اطلاعات و ارتباطات بر روند توسعه نظامها و مؤسسات آموزش عالی ، به بررسی جامع عوامل و موانع توسعه نظام آموزش عالی کشور از سطح حاکمیت تا سطح دانشگاهها می‌پردازد و مدلی بومی را برای توسعه مطلوب نظام آموزش عالی در ایران با رویکرد فاوا ترسیم مینماید . محتوای این کتاب در هفت فصل تنظیم گردیده و رئوس مطالب آن در فصول مختلف به شرح زیر است : فصل اول : جهانی شدن و آموزش عالی این

برای دیدن برچسب واژه‌ها می‌توان روی گزینه «تمام متن با برچسب» کلیک نمود که در آنصورت

صفحه زیر نمایش داده می‌شود:

۵-۲- پیشنهادات برای پژوهش‌های آینده

- با توجه به موجود بودن چکیده‌های فارسی و انگلیسی پایان‌نامه‌ها و مقاله‌های پژوهش‌نامه پردازش و مدیریت اطلاعات در ایرانداک، می‌توان با بررسی ملاحظات حقوقی، ابتدا پیکره‌ای مقایسه‌ای حاوی چکیده‌های فارسی و انگلیسی مقاله‌ها یا پایان‌نامه‌ها ساخت و به سامانه پیکره‌های ایرانداک اضافه کرد و سپس در طرح پژوهشی دیگری پیکره مقایسه‌ای را به پیکره موازی تبدیل کرد.
- بررسی امکان ساخت پیکره از پایان‌نامه‌ها و رساله‌های ثبت‌شده در ایرانداک با در نظر گرفتن ملاحظات حقوقی.
- اضافه کردن برچسب‌های گفتمانی به داده‌های پیکره.

فهرست منابع فارسی

- آیت، سید سعید. ۱۳۸۹. طراحی و پیاده‌سازی دادگان دایفون زبان فارسی برای کاربرد زبان‌شناسی رایانه‌ای. پژوهش‌های زبان‌شناسی. ۲ (۳): ۱۱-۱.
- اسلامی، محرم، مسعود شریفی آتشگاه، صدیقه علیزاده لمجیری، و طاهره زندی. ۱۳۸۳. واژگان زبانی زبان فارسی. مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه. تهران.
- بی‌جن‌خان، محمود. ۱۳۸۳. نقش پیکره‌های زبانی در نوشتن دستور زبان: معرفی یک نرم‌افزار رایانه‌ای. مجله زبان‌شناسی ۱۹ (۲): ۶۷-۴۸.
- تابع بردبار، علیرضا. ۱۳۹۳. ایجاد پیکره زبانی موازی به وسیله پیکره‌های قیاس‌پذیر. کارشناسی ارشد. شیراز: دانشگاه شیراز.
- دشتبانی، شکوفه، محرم منصوری‌زاده، و محمد نصیری. ۱۳۹۳. پیکره متنی تطبیقی فارسی-انگلیسی حوزه تخصصی فاوا. نشریه پژوهش‌های زبان‌شناسی تطبیقی. سال چهارم. شماره ۸. پاییز و زمستان ۱۳۹۳: ۱۲۱-۱۴۱.
- صادقی، علی‌اشرف. ۱۳۷۰. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱)». تهران: نشر دانش. شماره ۶۴.
- صادقی، علی‌اشرف. ۱۳۷۰. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۲)». تهران: نشر دانش. شماره ۶۵.
- صادقی، علی‌اشرف. ۱۳۷۰. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۳)». تهران: نشر دانش. شماره ۶۷.
- صادقی، علی‌اشرف. ۱۳۷۱. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۴)». تهران: نشر دانش. شماره ۶۹.
- صادقی، علی‌اشرف. ۱۳۷۱. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۵)». تهران: نشر دانش. شماره ۷۰.
- صادقی، علی‌اشرف. ۱۳۷۱. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۶)». تهران: نشر دانش. شماره ۷۱.

- صادقی، علی‌اشرف. ۱۳۷۱. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۷)». تهران: نشر دانش. شماره ۷۲.
- صادقی، علی‌اشرف. ۱۳۷۱. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۸)». تهران: نشر دانش. شماره ۷۴.
- صادقی، علی‌اشرف. ۱۳۷۲. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۹)». تهران: نشر دانش. شماره ۷۵.
- صادقی، علی‌اشرف. ۱۳۷۲. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱۰)». تهران: نشر دانش. شماره ۷۶.
- صادقی، علی‌اشرف. ۱۳۷۲. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱۱)». تهران: نشر دانش. شماره ۷۷.
- صادقی، علی‌اشرف. ۱۳۷۲. «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱۲)». تهران: نشر دانش. شماره ۷۹ و ۸۰.
- عاصی، مصطفی. ۱۳۸۲. از پیکره زبانی تا زبان‌شناسی پیکره‌ای. مجموعه مقالات پنجمین کنفرانس زبان‌شناسی. دانشگاه علامه طباطبائی.
- عاصی، مصطفی. ۱۳۸۴. گزارش کوتاهی از شکل‌گیری پایگاه داده‌های زبان فارسی در اینترنت. مجله پژوهشگران شماره ۲. صفحه ۱۳.
<https://hawzah.net/fa/Magazine/View/6280/6282/69701>
- علایی، الهام. ۱۳۹۸. امکان‌سنجی ساخت پیکره زبانی از داده‌های علمی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک). پژوهشکده علوم اطلاعات. پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک).
- علایی، الهام، و سیروس علیدوستی. ۱۳۹۹. ساخت پیکره متنی: طراحی مدل امکان‌سنجی. پژوهش‌های زبان‌شناسی تطبیقی. پاییز و زمستان ۹۹.
- علایی، الهام، نصراله پاک‌نیت، علی‌اصغر حجت‌پناه، مجتبی زالی، و محمدهادی آقالویی آغمیونی. ۱۴۰۰. ساخت پیکره متنی از مقاله‌های پژوهش‌نامه پردازش و مدیریت اطلاعات. پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک).
- قطره، فریا. ۱۳۸۶. «مشخصه‌های تصریفی در زبان فارسی امروز». دستور. شماره ۳: ۵۲-۸۱.
- کشانی، خسرو. ۱۳۷۱. اشتقاق پسوندی در زبان فارسی امروز. تهران: مرکز نشر دانشگاهی.

- کوهستانی، منوچهر. ۱۳۸۹. بررسی خطاهای املائی و نگارشی در وبلاگ‌های فارسی و ماهیت زبان‌شناختی آن‌ها. کارشناسی ارشد، تهران. دانشگاه تهران.
- لازار، ژیلبر. ۱۳۸۹. دستور زبان فارسی معاصر. ترجمه مهستی بحرینی و توضیحات و حواشی هرمز میلانیان. تهران: انتشارات هرمس. چاپ دوم.
- محمدی، رویا. ۱۳۹۱. ساخت پیکره تطبیقی فارسی-انگلیسی و استخراج جملات موازی از آن. پایان‌نامه کارشناسی ارشد. دانشگاه الزهرا (س). دانشکده فنی و مهندسی.
- مرادی، مهدی، و محمد بحرانی. ۱۳۹۴. تشخیص خودکار جنسیت نویسنده در متون فارسی. پردازش علائم و داده‌ها ۱۲ (۴): ۸۳-۹۴.

فهرست منابع لاتین

- AleAhmad, A., H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian. 2009. Hamshahri: A Standard Persian Text Collection. *Knowledge-Based Systems*. Dubai 22(5): 382-387. Elsevier.
- Atkins, S., J. Clear, and N. Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing*. 7 (1). 1-16
- Baradaran Hashemi, H., Shakery, A., and H. Faili. 2010. Creating a Persian-English Comparable Corpus. In *Proceedings of Conference on Multilingual and Multimodal Information Access Evaluation (CLEF)*, 27-39, Padua, Italy.
- Beloso, Begona Soneira. 2015. Designing, describing and compiling a corpus of English for architecture. *7th International conference on corpus linguistics: current work in corpus linguistics: working with traditionally-conceived corpora and beyond (CILC)*. *Procedia-social and behavioral sciences* 198. 459-464. Elsevier.

- Bijankhan, M., J. Sheykhzadegan, M. Bahrani, and M. Ghayoomi. 2011. Lesson from building a Persian written corpus: Peykare. *Language resources and evolution* 45 (2): 143-164. Springer.
- Brezina, Vaclav., Hawtin, Abi., and Toney McEnery. 2020. The written British national corpus 2014-design and comparability. *Journal of text and talk*. Volum 41 (5-6). 595-615. De Gruyter Mouton.
- Claude Toriida, M. 2016. Steps for creating specialized corpus and developing an annotated frequency-based vocabulary list. *TESL Canada journal/ revue TESL du Canada* 34 (11): 87-105.
- Crystal, D., 1994. *An Encyclopedic Dictionary of Language and Languages*. Oxford, p. 85.
- Dehghani, M., A. Shakery, M. Asadpour, and A. Koushkestani. 2013. A learning approach for email conversation thread reconstruction. *Journal of Information Science* (JIS) 39(6): 846-863.
- Davies, Mark. 2021. The Coronavirus corpus: design, construction and use. *International journal of corpus linguistics*. John Benjamins publishing company.
- Durand, J., U. Gut, and K. Gjert. 2014. *The handbook of corpus phonology*. Oxford University Press (OUP).
- Ghayoomi, M., S. Momtazi, and M. Bijankhan. 2013. A study of corpus development for Persian. *International journal on Asian language processing* 20 (1): 17-33.
- <https://github.com/sobhe/hazm>
- <https://www.peykaregan.ir/dataset>
- Leech, G. 2004. *Developing Linguistic Corpora: a Guide to Good Practice. adding linguistic annotation*. Edited by Martin Wynne .ahds.literature, languages and linguistics. The Oxford Text Archive.
- Leech, Geoffrey. 2013. *Corpus annotation. Linguistic information from computer text corpora. Chapter 1: Introducing corpus annotation*. Routledge. Taylor and Francis group. London and Newyork.

- Liddy, E., D. 2001. Natural language. Proceeding in *Encyclopedia of library and information science*, Stanford, California: CSLI publications. Marcel Decker incorporation, 2nd ed.
- McEnery, T., & A. Wilson. 2001. *Corpus Linguistics: An Introduction*. Edinburgh University Press.
- Megerdooian, K. 2004. Developing a Persian part-of-speech tagger. *Proceedings of the 1st workshop on Persian language and computer*. Pp. 99-105.
- Mohammadi, S. R., and N. Riahi. 2016. Presenting an Optimal Method for Constructing an English-Persian Comparable Corpus. *International Journal of Intelligent Information Systems*. 5 (3): 42-47
- Rasooli, M., M. Kouhestani, and A. Moloodi. 2013. Development of a Persian Syntactic Dependency Treebank. In *The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*: 306-314. Atlanta, USA.
- Samvelian, P., and P. Faghiri. 2013. Introducing PersPred, A Syntactic and Semantic Database for Persian Complex Predicates. In *Proceedings of the 9th Workshop on Multiword Expressions, Atlanta, Georgia, USA. Association for Computational Linguistics*, 11-20.
- Sinclair, J. 2004. *Developing Linguistic Corpora: a Guide to Good Practice. Chapter 1: Corpus and Text — Basic Principles*. Edited by Martin Wynne .ahds.literature, languages and linguistics. The Oxford Text Archive.
- Teubert, W. 2005. My version of Corpus linguistics. *International journal of Corpus linguistics*, 10 (1), 1-13.
- Wayne, M. 2005. *Developing linguistic corpora: a guide to good practice*. Oxbow books. *Literary and linguistic computing* 22 (1).

پیوست ۱: فهرست کتاب‌های دیجیتال ایرانداک

عنوان
آواشناسی آکوستیک و شنیداری
سیستم‌های اطلاعات مکانی (جی‌آی‌اس) و کاربردهای آن در علم اطلاعات و دانش‌شناسی
اصطلاح‌نامه ریاضیات
سنجش پژوهش‌های دانشگاهی: روش انجام یک مطالعه کتاب‌سنجی
دورکاری در سازمان‌ها و مراکز اطلاع‌رسانی
سنجش و مصورسازی پراکندگی جغرافیایی ثروت علمی: مفاهیم و کاربردها
ساختواره، اصطلاح‌شناسی و مهندسی دانش
جهانی‌شدن و مدل توسعه نظام آموزش عالی در ایران با رویکرد فناوری اطلاعات و ارتباطات
راهنمای نگارش گزارش‌های علمی و فنی
کتابخانه‌های دیجیتال: یکپارچه‌سازی محتوا و سیستم‌ها
استاندارد بین‌المللی BS ISO 25964-2: اطلاعات و مستندسازی- اصطلاح‌نامه‌ها و تعامل‌پذیری آن‌ها با واژگان‌های دیگر: بخش دوم: تعامل‌پذیری با واژگان‌های دیگر
مبانی اطلافرینی: کارآفرینی با اطلاعات و در اطلاع‌رسانی
شیوه‌نامه ایران: راهنمای استناد به منابع اطلاعات فارسی و انگلیسی
راهنمای نگارش گزارش‌های علمی و فنی
اصول طرح تحقیق و روش‌شناسی
استاندارد بین‌المللی ISO25964-1: اطلاعات و مستندسازی- اصطلاح‌نامه‌ها و تعامل‌پذیری آن‌ها با واژگان دیگر بخش اول: اصطلاح‌نامه‌های بازیابی اطلاعات

ریشه‌های فرهنگی ارتباط در ایران
مبانی و الگوی ذخیره و سازماندهی اطلاعات پروژه‌های فنی و عمرانی
مدل تعالی موسسه‌های پژوهشی
فرایند داوری در مجلات علمی؛ نقاط قوت و ضعف آن
مقدمه‌ای بر یادگیری الکترونیکی
اصطلاح‌نامه علوم کشاورزی
طراحی پایگاه وب دستگاه‌های دولتی: اطلاعات و ارتباطات
مدیریت ارتباطات علمی
کاربرد سامانه اطلاعات جغرافیایی (جی‌آی‌اس) در خدمات کتابداری و اطلاع‌رسانی
مجموعه‌سازی مشارکتی: راهنمای عملی برای کتابخانه شما
آواشناسی
ریشه‌های فرهنگی ارتباط در ایران
آمار طرح غدیر: عضویت فراگیر کتابخانه‌ها سال تحصیلی ۸۶-۱۳۸۵
نمایه‌سازی همارا: مفاهیم و روش‌ها
نیازسنجی اطلاعات
فرایند تدوین اصطلاح‌نامه
کاربرد سیستم‌های اطلاعات جغرافیایی (GIS) در پژوهش‌های سازمانی
رهنمودهایی برای سیاست‌های اصطلاح‌شناسی: تدوین و اجرای سیاست اصطلاح‌شناسی در جوامع زبانی
اصول طرح تحقیق و روش‌شناسی
فهرست مستند اسامی نهادها، سازمان‌ها و موسسات دولتی ایران
مدیریت نظام علوم، تحقیقات، و فناوری: پرسش‌های اساسی

واژه‌شناسی
زبان مخفی چیست؟
مدیریت کمیته‌ها در سازمان
تدوین و کاربرد اصطلاح‌نامه
فناوری اطلاعات و کتابخانه‌ها
شیوه‌نامه ایران: راهنمای استناد به منابع اطلاعات فارسی و انگلیسی
واژه‌نامه مقایسه‌ای اختلافات املائی انگلیسی امریکایی و بریتانیایی
اصطلاح‌نامه فیزیک
اطلاعات علمی شیمی و مهندسی شیمی ایران
اصطلاح‌نامه علوم زمین
سواد اطلاعاتی
سازمان‌ها و مراکز بین‌المللی کتابداری و اطلاع‌رسانی
اصطلاح‌نامه علوم زیستی
اصطلاح‌نامه ارتقای بهداشت
دانش ایران در سطح بین‌المللی سال ۲۰۰۴
طراحی برنامه درسی کارشناسی ارشد علوم کتابداری و اطلاع‌رسانی (گرایش اطلاع‌رسانی)
دانش ایران در سطح بین‌المللی سال ۲۰۰۱
دانش ایران در سطح بین‌المللی سال ۲۰۰۲
دانش ایران در سطح بین‌المللی سال ۲۰۰۳
نمایه‌سازی ماشینی
اصطلاح‌نامه شیمی
مجموعه مدارک نشست و کارگاه توسعه راهبردی در زمینه فناوری اطلاعات با تاکید بر توسعه محتوای بومی

دور کاری: مفاهیم و فرایند برنامه ریزی
دانش ایران در سطح بین المللی
مبانی اصطلاح شناسی
دانش ایران در سطح بین المللی سال ۲۰۰۰
بخش پوستر در کنفرانس ها: راهنمای برگزار کنندگان، شرکت کنندگان، و کتابداران
فهرستگان دیسک های نوری در مراکز آموزش و تحقیقاتی ایران
کتابشناسی علوم پزشکی
مبانی نظام ها و خدمات اطلاعاتی
مدیریت همایش های علمی و حرفه ای: راهنمای برگزاری یک همایش موفق
کتابشناسی توصیفی لغت نامه های تخصصی انگلیسی- فارسی، فارسی- انگلیسی
فهرست نویسی کامپیوتری
کتابشناسی اصطلاح نامه ها
کتابشناسی کتابشناسی ها: (عمومی و موضوعی)
گزارش فعالیت های سه ماهه دی، بهمن و اسفند ۱۳۶۹ مرکز اسناد و مدارک علمی ایران
گزارش سفر شرکت در بیست و چهارمین کنگره جهانی دامپزشکی در ریودوژانیرو- برزیل (از ۲۶ مرداد لغایت ۴ شهریور ۱۳۷۰)
فهرست مستند اسامی موسسات و سازمان های دولتی ایران
گزارش بازدید از مراکز اطلاعاتی و بانک های اطلاعاتی امریکا (کالیفرنیا، آریزونا)
راهنمای ناشران ایران
راهنمای انتشارات سازمان های دولتی ایران (فهرست سالانه ۱۳۶۳)
راهنمای سمینارها، کنگره ها و سمپوزیوم های علمی و فرهنگی ایران (۱۳۶۴)
کتابشناسی نجوم

کتابشناسی فیزیک
طرح ایجاد مرکز مدارک آب‌شناسی وابسته به کمیته ملی آب‌شناسی یونسکو
فهرست فیلم‌ها و اسلایدهای علمی موجود در مراکز آموزشی و پژوهشی کشور
راهنمای نوشتن رساله‌های پژوهشی و کتابشناسی توصیفی روش تحقیق
کتابشناسی فرهنگ‌های دوزبانه و چندزبانه فارسی
کتابشناسی برق-الکترونیک-مخابرات
نمایه‌نامه مدارک غیر کتابی موجود در مرکز اسناد و مدارک علمی ایران
فهرست گزارش‌های علمی و فنی یونسکو (۱۹۷۷-۱۹۸۶) موجود در کتابخانه مرکز اسناد و مدارک علمی
مقاله‌نامه علوم: علوم پایه، فنی-مهندسی، کشاورزی و پزشکی، شش ماهه اول و دوم ۱۳۶۴، دوره سوم، شماره یک و دو
اصطلاح‌نامه فنی و مهندسی
مدیریت تغییر در کتابخانه‌ها و خدمات اطلاعاتی
تدوین و کاربرد اصطلاح‌نامه
اصطلاح‌نامه جامعه‌شناسی
نشر الکترونیکی
اشتراک منابع و همکاری بین کتابخانه‌ها
تحصیلات تکمیلی علوم کتابداری و اطلاع‌رسانی در انگلستان، آمریکا، هند و ایران، بررسی تطبیقی
چهل و هشتمین کنگره فید (گزیده مقالات) (اتریش: ۲۵-۲۰ اکتبر) ۱۹۹۶
تبلیغ کارآمد سایت وب
انتخاب منابع اطلاعاتی
مدیریت مجموعه‌های فیلم و ویدئو

برنامه‌ریزی استراتژیک برای خدمات کتابخانه‌ای و اطلاعاتی
دورکاری برای کتابداران و اطلاع‌رسانان
برنامه‌ریزی مقابله با حوادث در کتابخانه‌ها و مراکز خدمات اطلاع‌رسانی
مدیریت خودکارسازی کتابخانه‌ها
تهیه راهنما برای خدمات کتابداری و اطلاع‌رسانی
ارزشیابی خدمات کتابداری و اطلاع‌رسانی
نمایه‌نامه مدارک غیرکتابی موجود در مرکز اسناد و مدارک علمی
کتابنامه اطلاع‌رسانی
راهنمای سمینارها، کنگره‌ها و سمپوزیوم‌های علمی و فرهنگی ایران (۱۳۶۳)
فهرست تحلیلی مقالات مجله هنر و مردم (۱۳۴۲-۱۳۵۷)
مقاله‌نامه علوم: علوم پایه، فنی-مهندسی، کشاورزی و پزشکی، شش‌ماهه دوم سال ۱۳۶۳، دوره دوم، شماره دو
راهنمای تهیه و گسترش اصطلاح‌نامه یک‌زبانه
فهرست پایان‌نامه‌های فارغ‌التحصیلان ایرانی خارج از کشور (۱۳۶۱-۱۳۵۵): انگلیسی، آلمانی، فرانسه
کتابشناسی علوم زیستی
کتابشناسی منتخب فلاسفه اسلامی
راهنمای انتشارات سازمان‌های دولتی ایران (فهرست سالانه ۱۳۶۲)
کتابشناسی رودکی
نمایه‌نامه مدارک غیرکتابی موجود در مرکز اسناد و مدارک علمی
کتابشناسی جغرافیا: چاپی-خطی
راهنمای سمینارها، کنگره‌ها، و سمپوزیوم‌های برگزار شده در ایران (۱۳۶۲-۱۳۵۸)
کتابشناسی اعتیاد (مواد مخدر-دخانیات-الکل)

راهنمای مراکز اسناد و کتابخانه‌های تخصصی، اختصاصی و دانشگاهی
کتابشناسی روند پژوهش در ایران
کتابشناسی مدیریت
مقاله‌نامه علوم: علوم پایه، فنی - مهندسی، کشاورزی و پزشکی، شش ماهه دوم سال ۱۳۶۲، دوره اول، شماره پنجم
کتابشناسی کتابشناسی‌ها
نمایه‌نامه مدارک غیر کتابی موجود در مرکز اسناد و مدارک علمی
فهرست پایان‌نامه‌های فارغ‌التحصیلان ایرانی خارج از کشور (۱۳۵۰-۱۳۵۴): انگلیسی، آلمانی، فرانسه
فهرست گزارش‌های علمی و فنی یونسکو (۱۹۸۲-۱۹۷۶) موجود در کتابخانه مرکز اسناد و مدارک علمی
کتابشناسی تکنولوژی سنتی ایران
راهنمای ایجاد و توسعه مراکز ارجاعی
کتابشناسی جامعه‌شناسی
کتابشناسی تکنولوژی سنتی ایران
نمایه‌نامه مدارک غیر کتابی موجود در مرکز اسناد و مدارک علمی
فهرست پایان‌نامه‌های فارغ‌التحصیلان ایرانی خارج از کشور (۱۲۹۲-۱۳۴۹): انگلیسی، آلمانی، فرانسه
فهرست گزارش‌های علمی و فنی یونسکو (۱۹۷۵-۱۹۸۱) موجود در کتابخانه مرکز اسناد و مدارک علمی
مقاله‌نامه علوم: علوم پایه، فنی - مهندسی، کشاورزی و پزشکی، دوره اول، شماره ۲، پاییز ۱۳۶۱
مقاله‌نامه علوم: علوم پایه، فنی - مهندسی، کشاورزی و پزشکی، شش ماهه اول سال ۱۳۶۲، دوره اول، شماره ۴

فهرست انتشارات سازمان‌های دولتی ایران موجود در مرکز اسناد و مدارک علمی: ۷-وزارت فرهنگ و آموزش عالی، دانشگاه‌ها و موسسات آموزش عالی
فهرست مشترک نشریه‌های ادواری موجود در کتابخانه‌های ایران
طرح پیشنهادی بازسازی کتابخانه و مرکز اسناد وزارت مسکن و شهرسازی
مقاله‌نامه علوم: علوم پایه، فنی-مهندسی، کشاورزی و پزشکی، دوره اول، شماره ۳، زمستان ۱۳۶۱
راهنمای سازمان‌دادن و تشکیل دوره‌های آموزشی و سمینارها و دوره‌های علمی در زمینه اطلاعات علمی و فنی و دکومانتاسیون
نمایه‌نامه مدارک غیرکتابی موجود در مرکز اسناد و مدارک علمی
مقاله‌نامه علوم: علوم پایه، فنی و مهندسی، کشاورزی، پزشکی، دوره اول، شماره ۱، شش‌ماهه اول سال ۱۳۶۱
راهنمای وزارتخانه‌ها و سازمان‌ها
فهرست انتشارات سازمان‌های دولتی ایران موجود در مرکز اسناد و مدارک علمی
برنامه پیشنهادی برای بازسازی کتابخانه سازمان پژوهش‌های علمی و صنعتی ایران
فهرست انتشارات سازمان‌های دولتی ایران موجود در مرکز اسناد و مدارک علمی
نگرشی در وضعیت اطلاع‌رسانی ایران و چند کشور دیگر - بخش اول - قسمت سوم: مروری در نظام اطلاع‌رسانی ایران
فهرست انتشارات سازمان‌های دولتی ایران موجود در مرکز اسناد و مدارک علمی
فهرست انتشارات سازمان‌های دولتی ایران موجود در مرکز اسناد و مدارک علمی
فهرست انتشارات سازمان‌های دولتی ایران موجود در مرکز اسناد و مدارک علمی

فهرست مشترک نشریه‌های ادواری کتابخانه‌های ایران
نگرشی در وضعیت اطلاع‌رسانی ایران و چند کشور دیگر
طرح تشکیل مرکز اسناد و مدارک کتابخانه مرکزی دانشگاه پلی‌تکنیک تهران
راهنمای مراکز اسناد و کتابخانه‌های تخصصی و دانشگاهی (طرح مقدماتی)
نگرشی در وضعیت اطلاع‌رسانی ایران و چند کشور دیگر
فهرست انتشارات سازمان‌های دولتی ایران موجود در مرکز اسناد و مدارک علمی
راهنمای شوراهای عالی کشور
فهرست فیلم‌استریپ‌های موجود در کتابخانه مرکز مدارک علمی
لیست موضوعی نشریات ادواری موجود در مرکز مدارک علمی (۱۹۸۰-۱۹۷۹)، (لاتین)
راهنمای واحدهای تولیدکننده اطلاعات در ایران
فهرست فیلم‌های موجود در کتابخانه مرکز مدارک علمی
فهرست مشخصات برخی از مراکز اسناد، اطلاعات و کتابخانه‌های تخصصی مهم ایران: (طرح اولیه)
مقاله‌نامه اقتصادی
مطبوعات ایران: فهرست تحلیلی نشریات آتش، آذر، آرزو، آرشیکت، آزادگان، آسیا، ایران کنونی، نامه اراک (موجود در کتابخانه ملی)
مقاله‌نامه انقلاب فرهنگی
فهرست مقاله‌های اقتصادی و صنعتی روزنامه انقلاب اسلامی سال ۱۳۵۸
مقاله‌نامه علوم اجتماعی: جامعه‌شناسی، جمعیت‌شناسی، سیاست، آمار (۱۳۴۸)
چکیده مقاله‌های مجله‌های ایران در زمینه علوم اجتماعی دوره هفتم شماره ۱-۲: چکیده شش‌ماهه اول سال ۱۳۵۲
فهرست الفبایی توصیفگرهای پذیرفته شده در مرکز مدارک علمی (سال ۲۵۳۴ تا ۲۵۳۶)

فهرست مشترک نشریه‌های ادواری موجود در کتابخانه‌های ایران: علوم و علوم اجتماعی
فهرست مشترک نشریه‌های ادواری موجود در کتابخانه‌های ایران: علوم و علوم اجتماعی
فهرست مشترک نشریه‌های ادواری موجود در کتابخانه‌های ایران: علوم و علوم اجتماعی
نمایه‌نامه مدارک غیرکتابی موجود در مرکز اسناد و مدارک علمی: نشریه ۲۳۰ - قسمت اول شماره ۱ تا ۵۰۰
فهرست مشترک نشریه‌های ادواری موجود در کتابخانه‌های ایران: علوم و علوم اجتماعی
کتابشناسی گزیده خلیج فارس
چکیده مقاله‌های مجله‌های ایران در زمینه علوم اجتماعی دوره پنجم شماره ۳ و ۴: چکیده مقاله‌های شش ماهه دوم سال ۱۳۵۰
فهرست گزیده مقالات شیمی (از شماره ۱ تا ۱۱۴) دوره ۵ شماره ۱
مقاله‌نامه آموزش و پرورش شماره ۵ دوره ۱
کتابشناسی کتاب و کتابداری در ایران؛ به زبان انگلیسی سده نوزدهم و بیستم
آموزش و پرورش: چکیده‌نامه (از شماره ۱ تا ۱۷۷)
فهرست کتابشناسی‌های ادواری موجود در کتابخانه مرکز مدارک علمی
فهرست پایان‌نامه‌های پزشکی و داروسازی
چکیده مقاله‌های مجله‌های ایران در زمینه علوم اجتماعی دوره چهارم شماره ۳-۴: چکیده مقاله‌های شش ماهه دوم ۱۳۴۹
مقاله‌نامه زیست‌شناسی دریای خزر
فهرست نشریات وزارت بهداشت
مقاله‌نامه زیست‌شناسی دریای خزر: نشریه شماره ۵۶ (دوره چهارم - شماره ۱ - از شماره ۱-۵۹)

چکیده مقاله‌های مجله‌های ایران در زمینه علوم اجتماعی دوره چهارم شماره ۱-۲: چکیده مقاله‌های شش ماهه اول سال ۱۳۴۹
کتابنامه موضوعی دندانپزشکی (تا پایان سال ۱۳۵۰)
فهرست پایان‌نامه‌های علوم اجتماعی
چکیده مقاله‌های مجله‌های ایران در زمینه علوم اجتماعی دوره دوم شماره اول و دوم
کتابنامه کشاورزی (تا پایان سال ۱۳۴۹)
فهرست گزیده مقالات در شیمی از شماره ۱ تا ۹۷، دوره چهارم، شماره ۱ و ۲
مقاله‌نامه زیست‌شناسی دریای خزر: دوره سوم - شماره اول - از شماره ۱-۴۹
فهرست الفبایی نشریات ادواری خارجی موجود در کتابخانه مرکز مدارک علمی
چکیده مقاله‌های مجله‌های ایران در زمینه علوم دوره چهارم شماره ۱-۴: چکیده مقاله‌های سال ۱۳۴۹
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست موضوعی نشریات ادواری موجود در کتابخانه مرکز مدارک علمی - تجدید چاپ
کتابنامه شیمی
فهرست گزیده مقالات در شیمی: الف - شیمی نظری از شماره ۱۱۸ تا ۱۵۰، دوره سوم، شماره ۶
فهرست گزیده مقالات در شیمی: ب - شیمی عملی، از شماره ۲۳ تا ۴۷، دوره سوم، شماره ۳
گزیده فهرست مقالات در شیمی: الف - شیمی نظری از شماره ۱ تا ۲۱، دوره سوم، شماره اول و دوم
مقاله‌نامه زیست‌شناسی دریای خزر: دوره سوم - شماره اول - شماره ۱-۴۹

مقاله‌نامه زیست‌شناسی دریای خزر
فهرست مقالات زیست‌شناسی دریای خزر: از شماره ۱-۶۵
فهرست مجله‌های زیست‌شناسی، کشاورزی و تغذیه موجود در مرکز مدارك علمی
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
چکیده مقاله‌های مجله‌های ایران در زمینه علوم و علوم اجتماعی دوره سوم شماره اول و دوم
چکیده مقاله‌های مجله‌های ایران در زمینه علوم و علوم اجتماعی دوره سوم شماره سوم و چهارم
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
چکیده مقاله‌های مجله‌های ایران در زمینه علوم دوره پنجم شماره ۱-۲: چکیده مقاله‌های سال ۱۳۵۰
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست منتخب نشریات در زمینه آموزش و پرورش (به ویژه برنامه‌ریزی آموزشی)
فهرست مشترک نشریه‌های ادواری موجود در کتابخانه‌های ایران
فهرست مشترک نشریه‌های ادواری موجود در کتابخانه‌های ایران
فهرست مقالات زیست‌شناسی دریای خزر: از شماره ۱-۷۱
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران

فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
چکیده مقاله‌های مجله‌های ایران در زمینه علوم اجتماعی دوره اول شماره اول
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
فهرست مندرجات مجله‌های علمی و علوم اجتماعی ایران
گزارش بررسی مقدماتی درباره رابطه صنایع با موسسات آموزش عالی فنی
نمایه گردان: چکیده مقاله‌های مجله‌های ایران در زمینه علوم و علوم اجتماعی، دوره اول سال ۱۳۴۸
دانش ایران (مشارکت ایرانیان در دانش جهان سال ۲۰۱۳)
عوامل موفقیت همکاری: بررسی متون پژوهشی
شیوه‌نامه ایران: راهنمای استناد به منابع اطلاعات فارسی و انگلیسی بر پایه شیوه‌نامه شیکاگو ویرایش ۱۶
دست‌آورد: سومین همایش ملی مدیران فناوری اطلاعات مراکز علمی کشور
آینده‌نگاری علم و فناوری و اثرات آن در سیاست‌گذاری
طراحی فرایند دوره‌های آموزشی کوتاه‌مدت

اطلاعات علمی شیمی و مهندسی شیمی ایران
دانش ایران در سطح بین‌المللی سال ۲۰۰۵
مجموعه مقالات کنفرانس کتابخانه دیجیتال: یک دهه کتابخانه‌های دیجیتالی در ایران با نگاهی به آینده
نمای ۲۰۱۶: جایگاه علم، فناوری، و نوآوری ایران در جهان
ارزیابی علم، فناوری و نوآوری: مروری بر شاخص‌ها و سازمان‌های فعال این حوزه
سیاست اطلاعات علمی در کشور (بررسی و پیشنهادات)
پیش‌نما: جایگاه علم، فناوری، و نوآوری ایران در جهان (توصیف شاخص‌ها و روند سالانه آنها تا ۲۰۱۵)
نمای ۲۰۱۴: جایگاه علم، فناوری، و نوآوری ایران در جهان
نمای ۲۰۱۵: جایگاه علم، فناوری، و نوآوری ایران در جهان
دستاوردهای چهارمین همایش ملی مدیران فناوری اطلاعات
استاندارد بریتانیا: توصیه‌هایی در مورد تهیه فهرست‌های کتاب‌ها، نشریات ادواری و سایر انتشارات
اصطلاح‌نامه «نما» (نظام مبادله اطلاعات علمی - فنی)
اصول گزارش‌نویسی: راهنمای نوشتن گزارش‌های علمی و اداری
اصول نامه‌نویسی: راهنمای نوشتن نامه‌های اداری
اصول نمایه‌سازی
اصول نمایه‌سازی همارا: با تاکید بر علوم هسته‌ای
اطلاعات علمی و فنی در آسیا و اقیانوسیه: بررسی کلی، نیازها و پیشنهادها
امانت بین کتابخانه‌ها: اصول و رهنمودها
انتخاب کتاب: راهنمای کتابداران و کتابخانه‌ها
آفت‌ها و آسیب‌های کتاب: روش‌های علمی و عملی پیشگیری و ترمیم

آیین نامه طرح غدیر: عضویت فراگیر کتابخانه ها
تدارک پژوهش نامه
چکیده نویسی: مفاهیم و روش ها
خدمات مربوط به اطلاعات علمی در اتحاد شوروی
دستور ساختن واژه نامه استاندارد های بین المللی
راهنمای استفاده از مراجع، مراجع زیست شناسی: بایولوژیکال ابستراکتز و بایو ریسرچ ایندکس
راهنمای آماده سازی مقالات علمی برای چاپ - یونسکو یونی سیست
راهنمای دانش آموختگان کتابداری ایران
راهنمای روزنامه های ایران ۱۳۴۷
راهنمای کتابخانه های ایران
راهنمای کتابخانه های شمال شرق ایران
راهنمای گردآوری و آماده سازی مدارک در مرکز مدارک علمی
راهنمای مراکز اسناد و کتابخانه های تخصصی، اختصاصی و دانشگاهی
راهنمای مراکز آموزشی، پژوهشی و تولید کننده اطلاعات علمی و فنی در ایران: فارسی - انگلیسی
روش ساده ذخیره و بازیابی اطلاعات در مراکز اسناد
سیاست اطلاعاتی در خدمت رشد و توسعه وظایف و مسئولیت های ملی و بین المللی
شبکه های اطلاعاتی
فرهنگ اصطلاحات دکومانتاسیون
فرهنگ علوم سیاسی: فارسی - انگلیسی، انگلیسی - فارسی
فهرست نویسی کامپیوتری
کاربرد مینی کامپیوتر و میکرو کامپیوتر در عمل آوری اطلاعات
دانش ایران: مشارکت ایرانیان در دانش جهان سال ۲۰۱۴
بازخوانی انتقادی ادبیات حوزه اخلاق فناوری اطلاعات

کتابنامه علوم ایران
هوش تجاری: داده کاوی و بهینه سازی برای تصمیم گیری
مدیریت اطلاعات (مجموعه دوم)
واژه نامه کتابداری و دکومانتاسیون انگلیسی - فارسی
واژه نامه مخابرات و ارتباطات راه دور: انگلیسی - فارسی
ویراستاری مجلات علمی و فنی: رهنمودها و توصیه هایی چند به ویراستاران، سردبیران، نویسندگان، کتابداران و دکومانتالیست ها
هدف های سیاست علمی: پیشنهاد های یونی سیست
آمار طرح غدیر عضویت فراگیر کتابخانه ها سال تحصیلی ۱۳۷۸-۷۹
آمار طرح غدیر عضویت فراگیر کتابخانه ها سال تحصیلی ۱۳۷۹-۸۰
آمار طرح غدیر عضویت فراگیر کتابخانه ها سال تحصیلی ۱۳۸۰-۸۱
آمار طرح غدیر عضویت فراگیر کتابخانه ها سال تحصیلی ۱۳۸۱-۸۲
آمار طرح غدیر عضویت فراگیر کتابخانه ها سال تحصیلی ۱۳۸۳-۸۴
آمار طرح غدیر عضویت فراگیر کتابخانه ها سال تحصیلی ۱۳۸۴-۸۵
مجموعه مقالات همایش ملی هر کتابخانه یک رصدخانه
درآمدی بر اخلاق اطلاعات: مبانی و مفاهیم اخلاق اطلاعات از دیدگاه لوچیانو فلوریدی
استانداردهای توانمندی سواد دیداری انجمن کتابخانه های دانشکده ای و پژوهشی برای آموزش عالی
مجموعه قوانین صبا مشتمل بر: قانون اساسی، قانون مدنی، قوانین موجر و مستاجر، قانون تجارت... همراه با آرای دیوان عالی کشور، نظریات شورای نگهبان...
فرهنگ عبارات فقهی و تعبيرات عربی

آشنایی با حق مولف: مالکیت فکری در عصر دیجیتال
اصطلاح‌شناسی و واژه‌سازی در زبان آلمانی
سنجه‌های وی: برای متخصصان کتابداری و اطلاع‌رسانی
شرح قانون حمایت مولفان، مصنفان و هنرمندان
دستاوردهای پنجمین همایش ملی مدیران فناوری اطلاعات
خدمات اطلاعاتی و کاربرد آن: سنجش و آمار برای کتابخانه‌ها و فراهم‌آوردگان اطلاعات
آمار طرح امین: امانت بین کتابخانه‌ها دوره اول تا هفتم از مهر ۱۳۷۹ تا بهمن ۱۳۹۶
آمار کتابخانه‌ها: سال تحصیلی ۱۳۹۴-۹۵
نمای ۲۰۱۷: جایگاه علم، فناوری، و نوآوری ایران در جهان
بازی‌وارسازی: مفاهیم و کاربردها
بازنگری در مفهوم دستبرد ادبی
مبانی کپی‌رایت (از آغاز تا عصر دیجیتال): راهنمای مولفان، مترجمان، ناشران، حقوق‌دانان و متخصصان علم اطلاعات و دانش‌شناسی
مدیریت رسانه: مفاهیم، نظریه‌ها و رویکردها
هم‌گرایی رسانه‌ای
دگرسنجه‌ها برای متخصصان اطلاعات گذشته، حال، آینده
آمار کتابخانه‌ها: سال تحصیلی ۱۳۹۵-۹۶
چکیده آمار کتابخانه‌ها: سال تحصیلی ۱۳۹۴-۹۵
چکیده آمار کتابخانه‌ها: سال تحصیلی ۱۳۹۵-۹۶
تجاری‌سازی نتایج پژوهش: الگوها، سازوکارها و راهبردها
هدف‌های سیاست اطلاعات علمی: پیشنهادهای «یونی‌سیست»
تاریخ حقوق ایران: ماده‌ها و هخامنشیان
قواعد و قراردادهای ساختن اصطلاح‌نامه
شرح ماندنی بر قانون مدنی: نکته‌ها و گفته‌ها - شرح ماده به ماده قانون مدنی

مدل‌ها و روش‌های مدیریت کیفیت داده: نگرشی کاربردی بر داده‌های علم و فناوری
سامانه استانداردساز و خطایاب متون علمی فارسی
دستاوردهای ششمین همایش ملی مدیران فناوری اطلاعات
آمار کتابخانه‌ها: سال تحصیلی ۹۷-۱۳۹۶
چکیده آمار کتابخانه‌ها: سال تحصیلی ۹۷-۱۳۹۶
سرقت علمی و قوانین و مقررات حاکم بر آن در ایران (بررسی وضع موجود و ارائه راهکار)
نمای ۲۰۱۸: جایگاه علم، فناوری، و نوآوری ایران در جهان
قانون مدنی: بر اساس نخستین و کامل‌ترین نسخه چاپی قانون مدنی (چاپ شده در ۱۳۱۴ توسط مجلس شورای ملی)
چالش‌ها و رویکردهای امنیتی در اینترنت اشیا
نمای ۲۰۱۹: جایگاه علم، فناوری، و نوآوری ایران در جهان
دانش ایران: مشارکت ایرانیان در دانش جهان سال ۲۰۱۵
دستاوردهای هفتمین همایش ملی مدیران فناوری اطلاعات
کاربردهای معاصر تئوری کنشگر- شبکه (فصل ۷- سیاست‌گذاری محافظت از داده‌ها در رسانه‌های اجتماعی؛ رویکرد نظریه کنشگر- شبکه)
مدیریت دانش: تجارب بخشی عمومی و دولتی در ایران و جهان
آمار کتابخانه‌ها: سال تحصیلی ۹۸-۱۳۹۷
چکیده آمار کتابخانه‌ها: سال تحصیلی ۹۸-۱۳۹۷
دانش ایران: مشارکت ایرانیان در دانش جهان سال ۲۰۱۶
دانش ایران: مشارکت ایرانیان در دانش جهان سال ۲۰۱۷
سیستم‌های اطلاعات مدیریت
مدیریت اطلاعات (مجموعه اول)
اطلاعات و جامعه نوین جهانی

جایگاه کتاب، مطالعه، تفکر و دانش در بیان حضرت علی (ع)
مجموعه مقاله‌های همایش خدمات کتابخانه‌ای و افراد با آسیب بینایی
نقد منابع اطلاعاتی: شناخت مفاهیم، ساختار، ماهیت و معیارهای نقدی
مصورسازی اطلاعات مبانی و کاربردها
حاکمیت اطلاعات: مجموعه مقالات همایش ملی حاکمیت اطلاعات، آذر ۹۶

پیوست ۲: جزئیات فنی هنجارسازی و برجسب‌گذاری ماشینی داده‌های پیکره

برای اجرای کد برجسب‌گذاری ابزار «هضم»، ابتدا لازم است جدولی در پایگاه داده‌ای mysql ایجاد شود. به منظور ایجاد این جدول اسکریپت زیر باید اجرا شود. با اجرای این کد در پایگاه داده، جدولی با ساختار جدول ۱ ایجاد خواهد شد. از این جدول برای نگهداری داده‌های برجسب‌گذاری شده استفاده می‌شود.

```
SET SQL_MODE = "NO_AUTO_VALUE_ON_ZERO";
```

```
SET time_zone = "+00:00";
```

```
/*!40101 SET
```

```
@OLD_CHARACTER_SET_CLIENT=@@CHARACTER_SET_CLIENT */;
```

```
/*!40101 SET
```

```
@OLD_CHARACTER_SET_RESULTS=@@CHARACTER_SET_RESULTS */;
```

```
/*!40101 SET
```

```
@OLD_COLLATION_CONNECTION=@@COLLATION_CONNECTION */;
```

```
/*!40101 SET NAMES utf8mb4 */;
```

```
--
```

```
-- Database: `corpus`
```

```
--
CREATE TABLE IF NOT EXISTS `tags` (
  `id` bigint(20) NOT NULL AUTO_INCREMENT,
  `document_id` int(11) NOT NULL,
  `word` varchar(500) COLLATE utf8_unicode_ci NOT NULL,
  `tag` varchar(20) COLLATE utf8_unicode_ci NOT NULL,
  PRIMARY KEY (`id`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8 COLLATE=utf8_unicode_ci;

/*!40101 SET CHARACTER_SET_CLIENT=@OLD_CHARACTER_SET_CLIENT
*/;

/*!40101 SET
CHARACTER_SET_RESULTS=@OLD_CHARACTER_SET_RESULTS */;

/*!40101 SET COLLATION_CONNECTION=@OLD_COLLATION_CONNECTION
*/;
```

ساختار جدول برای نگهداری کلمات و برچسب‌ها

شماره	نام فیلد	توضیحات
۱	id	کلید اصلی این جدول است و دارای اعداد ترتیبی
۲	document_id	دارای کلید خارجی به شماره سندی است که برای برچسب‌گذاری استفاده می‌شود.
۳	word	برای نگهداری کلمات از این فیلد استفاده می‌شود.
۴	tag	برای برچسب هر کلمه از این فیلد استفاده می‌شود.

پس از ایجاد جدول، از کد زیر که به زبان پایتون و سازگار با نسخه python 3.x است، برای برچسب‌گذاری اسناد می‌توان استفاده کرد.

```
from __future__ import unicode_literals
from hazm import *
```

```
import mysql.connector
```

این فایل شامل مدل آموزش دیده برای برچسب گذاری است که توسط «هضم» توسعه داده شده است.

```
tagger = POSTagger(model='./resources-0.5/postagger.model')
```

```
db = mysql.connector.connect(user='user', password='password',  
                             host='host_name_or_ip',  
                             database='name_of_database')
```

```
cursor = db.cursor()
```

```
cursor.execute("SELECT * FROM hazm")
```

```
for document in cursor.fetchall():
```

```
    document_id = document[0]
```

```
    document_text = document[1]
```

```
    print(document_id)
```

```
    tags = tagger.tag(word_tokenize(document_text))
```

```
    c = 0
```

```
    for tag in tags:
```

```
        values = (document_id, tag[0], tag[1])
```

```
        cursor.execute("INSERT INTO `tags` (`id`, `document_id`, `word`, `tag`)  
VALUES (NULL, %s, %s, %s)", values)
```

```
        c += 1
```

```
    if c > 1000:
```

```
        c = 0
```

```
        db.commit()
```

```
db.close()
```

قطع کد زیر توسعه داده شده توسط «هضم» برای برچسب گذاری یک رشته (string) استفاده می شود.

```
# coding: utf-8
```

```
from __future__ import unicode_literals
from nltk.tag import stanford
from .SequenceTagger import SequenceTagger
```

```
class POSTagger(SequenceTagger):
```

```
    """
```

```
    >>> tagger = POSTagger(model='resources/postagger.model')
```

```
    >>> tagger.tag(['من', '.', 'به', 'مدرسه', 'رفته_بودم'])
```

```
    [('من', 'PRO'), ('به', 'P'), ('مدرسه', 'N'), ('رفته_بودم', 'V'), ('.', 'PUNC')]
```

```
    """
```

```
class StanfordPOSTagger(stanford.StanfordPOSTagger):
```

```
    """
```

```
    >>> tagger = StanfordPOSTagger(model_filename='resources/persian.tagger',
    path_to_jar='resources/stanford-postagger.jar')
```

```
    >>> tagger.tag(['من', '.', 'به', 'مدرسه', 'رفته_بودم'])
```

```
    [('من', 'PRO'), ('به', 'P'), ('مدرسه', 'N'), ('رفته_بودم', 'V'), ('.', 'PUNC')]
```

```
    """
```

```
    def __init__(self, model_filename, path_to_jar, *args, **kwargs):
```

```
        self._SEPARATOR = '/'
```

```
        super(stanford.StanfordPOSTagger,
self).__init__(model_filename=model_filename, path_to_jar=path_to_jar, *args,
**kwargs)
```

```
    def tag(self, tokens):
```

```
return self.tag_sents([tokens])[0]
```

```
def tag_sents(self, sentences):  
    refined = map(lambda s: [w.replace(' ', '_') for w in s], sentences)  
    return super(stanford.StanfordPOSTagger, self).tag_sents(refined)
```

قطع کد زیر توسعه داده شده توسط «هضم» برای هنجارسازی یک رشته (string) استفاده می‌شود.

```
# coding: utf-8
```

```
from __future__ import unicode_literals  
import re  
from .Lemmatizer import Lemmatizer  
from .WordTokenizer import WordTokenizer  
from .utils import maketrans
```

```
compile_patterns = lambda patterns: [(re.compile(pattern), repl) for pattern, repl in  
patterns]
```

```
class Normalizer(object):  
    def __init__(self, remove_extra_spaces=True, persian_style=True,  
persian_numbers=True, remove_diacritics=True, affix_spacing=True,  
token_based=False, punctuation_spacing=True):  
        self._punctuation_spacing = punctuation_spacing  
        self._affix_spacing = affix_spacing  
        self._token_based = token_based  
  
        translation_src, translation_dst = ' یکی ', ' یکی "'  
  
        if persian_numbers:  
            translation_src += '0123456789%'
```

```

translation_dst += '%.۱۲۳۴۵۶۷۸۹'

self.translations = maketrans(translation_src, translation_dst)

if self._token_based:
    lemmatizer = Lemmatizer()
    self.words = lemmatizer.words
    self.verbs = lemmatizer.verbs
    self.tokenizer = WordTokenizer(join_verb_parts=False)
    self.suffixes = {'ی', 'ای', 'ها', 'های', 'تر', 'تری', 'ترین', 'گر', 'گری', 'ام', 'ات', 'ش'}

self.character_refinement_patterns = []

if remove_extra_spaces:
    self.character_refinement_patterns.extend([
        (r' +', ' '), # remove extra spaces
        (r'\n\n+', '\n\n'), # remove extra newlines
        (r'[\r]', ''), # remove keshide, carriage returns
    ])

if persian_style:
    self.character_refinement_patterns.extend([
        ('"([^\n"]+)"', r'«\1»'), # replace quotation with gyoome
        ('([\d+])\.([\d+])', r'\1.\2'), # replace dot with momayez
        (r' ?\.\.\.', ' ...'), # replace 3 dots
    ])

if remove_diacritics:
    self.character_refinement_patterns.append(

```

```
(['\u064B\u064C\u064D\u064E\u064F\u0650\u0651\u0652'], ""), # remove
FATHATAN, DAMMATAN, KASRATAN, FATHA, DAMMA, KASRA, SHADDA,
SUKUN
```

```
)
```

```
self.character_refinement_patterns =
compile_patterns(self.character_refinement_patterns)
```

```
punc_after, punc_before = r'\.:\!؟؛»\}\}', r'«\[\{\'
```

```
if punctuation_spacing:
```

```
self.punctuation_spacing_patterns = compile_patterns([
    quotation (" ([^n"]+) ", r""1""), # remove space before and after

    (' ([+ punc_after +])', r'1'), # remove space before
    (' ([+ punc_before +]) ', r'1'), # remove space after
    (' ([+ punc_after[:3] +])' ([^ '+ punc_after +\d.۱۲۳۴۵۶۷۸۹]'),
r'1 \2'), # put space after . and :
    (' ([+ punc_after[3:] +])' ([^ '+ punc_after +])', r'1 \2'), #
put space after
    (' ([^ '+ punc_before +])' ([+ punc_before +])', r'1 \2'), #
put space before
])
```

```
if affix_spacing:
```

```
self.affix_spacing_patterns = compile_patterns([
    (r'([ ^ ]۰) ی', r'1 ی'), # fix ی space
    (r'(^|)(می?ن?)', r'1\2'), # put zwnj after می, نمی
    (r'(?<=[^ \n\d '+ punc_after + punc_before +]){2})
تر, (?=[ \n'+ punc_after + punc_before +])$)', r'1'), # put zwnj before
تری, ترین, گری, ها, های
```

(r'([^]ه)((م|یم|اش|ند|ی|ید|ت)))(?=[\n'+ punc_after +']|\$)', r'\1 \2'), # join ات, اید, ای, اند, اش, ایم, ام

```
def normalize(self, text):
    text = self.character_refinement(text)
    if self._affix_spacing:
        text = self.affix_spacing(text)

    if self._token_based:
        tokens = self.tokenizer.tokenize(text.translate(self.translations))
        text = ' '.join(self.token_spacing(tokens))

    if self._punctuation_spacing:
        text = self.punctuation_spacing(text)

    return text
```

```
def character_refinement(self, text):
    """
    >>> normalizer = Normalizer()
    >>> normalizer.character_refinement('اصلاح کاف و یای عربی')
    'اصلاح کاف و یای عربی'
    >>> normalizer.character_refinement('4.2 قراردادی به ارزش "2012عراق سال
    ('.میلیارد دلار" برای خرید تجهیزات نظامی با روسیه امضا کرد
    عراق سال ۲۰۱۲ قراردادی به ارزش «۴۰۲ میلیارد دلار» برای خرید تجهیزات نظامی با روسیه امضا
    'کرد.
    >>> normalizer.character_refinement('رمان')
```

'رمان'

```
>>> normalizer.character_refinement('بشقابِ من را بگیر')
```

'بشقاب من را بگیر'

"""

```
text = text.translate(self.translations)
```

```
for pattern, repl in self.character_refinement_patterns:
```

```
    text = pattern.sub(repl, text)
```

```
return text
```

```
def punctuation_spacing(self, text):
```

"""

```
>>> normalizer = Normalizer()
```

```
>>> normalizer.punctuation_spacing('اصلاح ( پرانتزها ) در متن')
```

'اصلاح (پرانتزها) در متن'

```
>>> normalizer.punctuation_spacing('1396 تهران، 22:00 در ساعت 0.5 نسخه')
```

'1396 تهران، 22:00 در ساعت 0.5 نسخه'

```
>>> normalizer.punctuation_spacing('۹.۰ میلیون.اتریش ۷')
```

'۹.۰ میلیون.اتریش ۷'

"""

```
for pattern, repl in self.punctuation_spacing_patterns:
```

```
    text = pattern.sub(repl, text)
```

```
return text
```

```
def affix_spacing(self, text):
```

```

"""
>>> normalizer = Normalizer()
>>> normalizer.affix_spacing('خانه ی پدری')
'خانه ی پدری'
>>> normalizer.affix_spacing('فاصله میان پیشوند ها و پسوند ها را اصلاح می کند')
'فاصله میان پیشوندها و پسوندها را اصلاح می کند'
>>> normalizer.affix_spacing('می روم')
'می روم'
>>> normalizer.affix_spacing('حرفه ای')
'حرفه ای'
>>> normalizer.affix_spacing('محبوب ترین ها')
'محبوب ترین ها'
"""

```

```

for pattern, repl in self.affix_spacing_patterns:
    text = pattern.sub(repl, text)
return text

```

```

def token_spacing(self, tokens):
    """
    >>> normalizer = Normalizer(token_based=True)
    >>> normalizer.token_spacing(['کتاب', 'ها'])
['کتاب ها']
>>> normalizer.token_spacing(['او', 'می', 'رود'])
['او', 'می رود']

```

```
>>> normalizer.token_spacing(['ماه', 'می', 'سال', 'جدید'])
```

```
['ماه', 'می', 'سال', 'جدید']
```

```
>>> normalizer.token_spacing(['اخلال', 'اگر'])
```

```
['اخلال', 'اگر']
```

```
>>> normalizer.token_spacing(['پرداخت', 'شده', 'است'])
```

```
['پرداخت', 'شده', 'است']
```

```
>>> normalizer.token_spacing(['زمین', 'لرزه', 'ای'])
```

```
['زمین', 'لرزه', 'ای']
```

```
""
```

```
result = []
```

```
for t, token in enumerate(tokens):
```

```
    joined = False
```

```
    if result:
```

```
        token_pair = result[-1] + token
```

```
        if token_pair in self.verbs or token_pair in self.words and  
self.words[token_pair][0] > 0:
```

```
            joined = True
```

```
self.verbs:                if t < len(tokens)-1 and token+'_'+tokens[t+1] in
```

```
                            joined = False
```

```
                            elif token in self.suffixes and result[-1] in self.words:
```

```
                                joined = True
```

```
    if joined:
```

```
        result.pop()
        result.append(token_pair)
    else:
        result.append(token)

return result
```

Abstract

We live in an information age. Systematic data gathering is considered as a fundamental task; in this regard, one of the systematic method for data gathering is building corpora. Many general corpora have been generated in different languages, including Persian, but not many specialized corpora have been built to be used in natural language processing. So, specialized corpora, which include specialized texts in interdisciplinary fields are valuable. In present study a specialized corpus has been built, which include IranDoc digital books. In sampling process 65 books were chosen and the related metadata (including the title of the books, the author(s) as well as the subject of the books) was extracted from the books. Around half of the books were in pdf format, which were converted to word format via OCR and the texts were edited (misspelling correction). In normalization phase, HAZM tool, Persian morphological information (regarding derivational and inflectional morphemes) as well as computer programming were used and finally the texts were annotated (POS tagging) and checked manually. All the information was saved in database and the corpus was inserted in Irandoc corpora website (SAPA). The corpus include 3329281 words, which covers different fields such as management, librarianship, linguistics, information management and the like.

Key words: corpus, specialized corpus, normalization, Persian morphological information, Part Of Speech (POS) tagging



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

Building a corpus from IranDoc digital books

By:

Elham Alayiaboozar

Aliasghar Hojjatpanah

November 2022