



پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)

نقش ایرانداک در گسترش خط و زبان فارسی

۳	مقدمه
۵	سامانه‌ها
۱۴	پژوهش
۳۷	کتاب
۴۰	سخنرانی‌ها و میزگردهای علمی
۵۳	کارگاه‌ها و دوره‌های آموزشی

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

زبان فارسی، زبان دوم عالم اسلام و کلید بخش عظیمی از ذخایر ارزشمند علمی و ادبی تمدن اسلامی، و خود از ارکان هویت فرهنگی ملت ایران است. بنابر اصل پانزدهم قانون اساسی جمهوری اسلامی ایران، زبان و خط رسمی و مشترک مردم ایران فارسی است. اسناد و مکاتبات و متون رسمی و کتب درسی باید با این زبان و خط باشد ولی استفاده از زبان‌های محلی و قومی در مطبوعات و رسانه‌های گروهی و تدریس ادبیات آنها در مدارس، در کنار زبان فارسی آزاد است. پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) از پنج دهه پیش کار مدیریت اطلاعات علم و فناوری را در ایران آغاز کرده است. خط و زبان فارسی و کاربری آن در محیط رایانه‌ای یکی از حوزه‌های مورد توجه در اساسنامه و برنامه استراتژیک پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است. همچنین گسترش زبان فارسی برای علم در فضای مجازی یکی از مأموریت‌های ملی ایرانداک است. اسناد بالا دستی همچون نقشه جامع علمی کشور، سند چشم‌انداز، برنامه ششم توسعه اقتصادی، اجتماعی و فرهنگی جمهوری اسلامی و سیاست‌های کلی علم و فناوری (نظام آموزش عالی، تحقیقات، و فناوری)، همواره این حوزه را مورد توجه قرار داده‌اند. در این گزارش، چکیده‌ای از نقش ایرانداک در گسترش خط و زبان فارسی آمده است.



فناوری‌های نوین بر بازنمایی و بازیابی دانش در زمینه مستندسازی تأثیر بسیاری داشته‌اند. این فناوری‌ها تغییرات بزرگی را در بازنمایی اطلاعات دیجیتال و بازیابی آنها پدید آورده و بر رفتار کاربران نیز تأثیر بسیاری گذاشته‌اند. در سایه گسترش روزافزون متن‌های دیجیتال، ابزارهای بازنمایی اطلاعات نیز تغییر یافته و پای رشته‌های گوناگونی مانند زبان‌شناسی، هوش مصنوعی، و مهندسی زبان نیز به این حوزه باز شده است. بر این پایه، ابزارهایی مانند تاکسونومی‌ها یا سیستم‌های رده‌بندی، واژگان، اصطلاح‌نامه‌ها، نقشه‌های مفهومی، شبکه‌های معنایی، و هستان‌شناسی‌ها همچون ابزارهای بازنمایی دانش به کار می‌روند. هستان‌شناسی در بازنمایی دانش از کلیدی‌ترین ابزارهای سازمان‌دهی اطلاعات در وب به شمار می‌رود.

واژه‌ها و واژه‌نامه‌ها در گسترش و همگانی‌سازی علم و استانداردسازی زبان نقشی کلیدی دارند. ایرانداک در سازمان‌دهی اطلاعات، همواره با واژه‌های نوین علمی و فناورانه روبه‌رو بوده است و باید برای واژه‌های بیگانه، برابرهای در ست فارسی را بیابد یا برگزیند و هر گاه نیاز باشد، توصیف آنها را نیز همراه واژه‌ها سازد. از این رو، ساخت واژه‌نامه‌ها یکی از کارهای همیشگی در پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است و نخستین واژه‌نامه‌ای که منتشر ساخته به ۴۰ سال پیش باز می‌گردد. در پایگاه وب واژه‌نامه‌ها، همراه با باز نشر واژه‌نامه‌های چاپی ایرانداک در محیط وب، واژه‌هایی که در کار روزانه سازمان‌دهی اطلاعات، کنترل و برگزیده می‌شوند، در اختیار همه علاقه‌مندان گذاشته شده است. افزون بر این، واژه‌نامه‌های دیگر نهادها مانند فرهنگستان زبان و ادب فارسی نیز در این پایگاه وب در دسترس کاربران است. این پایگاه بیش از ده واژه‌نامه و نزدیک یکصد هزار واژه را در بر دارد.

اصطلاح‌نامه‌ها به مثابه گنجینه‌ای از واژه‌ها، افزون بر نظم الفبایی مانند فرهنگ‌ها، دارای نظام شبکه‌ای و مفهومی میان واژه‌های یک یا چند حوزه از دانش بشری هستند. این نظام مفهومی با چند گونه پیوند میان واژه‌ها؛ در سازمان‌دهی و بازیابی اطلاعات، تحلیل اطلاعات و علم‌سنجی، برنامه‌ریزی آموزشی، و مانند آنها کاربرد دارد و در ساخت آنتولوژی (هستان‌شناسی) همچون ابزاری برای وب معنایی، پیش‌نیازی کلیدی و بی‌بدیل است. در اصطلاح‌نامه، واژه‌ها ساختار می‌یابند و مفاهیم با اصطلاح‌ها بیان و به گونه‌ای مرتب می‌شوند که پیوند میان این مفاهیم، آشکار و روشن و اصطلاح‌های مرجع با سرمدخل‌هایی برای مترادف‌ها یا شبه‌مترادف‌ها همراه باشند. اصطلاح‌نامه، راهنمای نمایه‌ساز و کاربر برای گزینش اصطلاح یکسان یا ترکیب یکسانی از اصطلاح‌های مرجع برای بیان یک

موضوع است. با ساخت واژه‌های استاندارد برای یک موضوع، یک مدخل‌ها در نمایه‌سازی فراهم می‌شود. با تفکیک هم‌نویس‌ها/هم‌نگاره‌ها به گونه‌ای که هر یک تنها به یک معنی رهنمون شود و گزینش اصطلاح مرجع برای نمایه‌سازی از میان مجموعه‌ای از مترادف‌ها و شبه مترادف‌ها انجام گیرد، کاربر و نمایه‌ساز برای یک مفهوم، اصطلاح همانند یا نزدیک‌تری را بر خواهند گزید.

پردازش زبان طبیعی در پی ساخت و گسترش رویکردهای هوش مصنوعی برای ماشینی کردن فرایند درک و ساخت یک زبان طبیعی است که به ترجمه، خلاصه‌سازی، درست کردن املا، و مانند آنها به شکل خودکار می‌انجامد؛ این امر با کاربست الگوریتم‌ها و ساختارهای داده در علوم رایانه، به ویژه یادگیری ماشین انجام می‌شود. پردازش متن، انجام خودکار بسیاری از پردازش‌ها روی زبان مانند ترجمه، استخراج اطلاعات از متن‌ها، خلاصه‌سازی، استخراج واژه‌های کلیدی، یافتن دستبرد علمی، خوشه‌بندی متن و مدل‌سازی موضوع، نمایه‌سازی، و درست کردن املا، به مجموعه‌ای از ابزارها برای پیش‌پردازش و آماده‌سازی متن نیاز دارد. این ابزارها در دو دسته‌اند که گاهی آمیخته آنها نیز به کار می‌رود. دسته نخست، روش‌های وابسته به زبان هستند که بر پایه برخی از قواعد نحوی و ساختاری زبان انجام می‌شوند. روش‌های دسته دوم مستقل از زبان هستند و بیشتر بر پایه پیکره‌های زبانی و با کاربرد روش‌های یادگیری ماشینی به انجام می‌رسند. طراحی و ساخت این ابزارها برای زبان‌های گوناگون به شیوه‌های گوناگون و ویژه هر زبان انجام می‌شود. ابزارهای پردازش زبان طبیعی برای شناسایی جمله، جداسازی، شناسایی ماهیت اسم‌ها، شبکه واژه‌ها، ریشه‌یاب، شناسایی همانندی، شناسایی گروه‌های نحوی، برچسب‌دهنده نقش نحوی، نشانه‌گذاری در پیکره‌های متنی، تعیین‌کننده هم‌مرجع‌ها، و برچسب‌گذاری اجزای واژگانی کلام به کار می‌روند. کاربردهای مبتنی بر درک زبان طبیعی به واکاوی ژرف‌تری از درونداد (متن یا گفتار) نیاز دارد. کاربست تکنیک‌های کلان‌داده و یادگیری ماشین برای درک ماشینی متن‌های خبری و علمی، نمونه‌ای از این کاربردها هستند. ارتباط میان انسان و ماشین و گفت‌و شنود آنها دسته دیگری از کاربردهای پردازش زبان طبیعی هستند. برای نمونه، خدمات بانکی می‌تواند بر پایه ارتباط میان مشتری و ماشین ارائه شود. طراحی سیستم‌های تبدیل متن به گفتار و گفتار به متن، سیستم‌های ترجمه گفتاری، سیستم‌های گفت‌و شنود، و بازشناسی گفتار از دیگر کاربردهای پردازش زبان طبیعی هستند. انجام بسیاری از طرح‌های پژوهشی در زمینه پردازش زبان طبیعی و پردازش متن نیاز به کاربست پیکره‌های متنی تخصصی دارد. برای چنین کارهایی بهتر است به جای پیکره‌های متنی عمومی در بازار، پیکره‌هایی بر پایه داده‌های علمی و فنی پایگاه‌های داده ایرانداک ساخته شود. از آنجایی که این داده‌ها از پایان‌نامه‌ها و رساله‌های به‌روز به‌دست می‌آیند، می‌توانند در بسیاری از کارهای پژوهشی به کار روند و دستاوردهای بهتری را به‌بار آورند.

سامانه‌ها

ایرانداک برای پیشبرد اهداف خود در زمینه فناوری اطلاعات، سامانه‌های گوناگونی را فراهم آورده است. که از ابزارهای پردازش زبان طبیعی جهت بازیابی اطلاعات و پردازش متون فارسی استفاده می‌کنند.

سامانه ملی ثبت پایان‌نامه، رساله، و پیشنهاد

هر چند سامانه ملی ثبت پایان‌نامه، رساله، و پیشنهاد از سال ۱۳۸۶ و بر پایه بخش‌نامه وزارت عتف (شماره ۴۳۸۹/۱۲۲۳۸ تاریخ ۱۳۸۶/۸/۲۰) به‌راه افتاد، ولی این سامانه در آبان ۱۳۹۵ برای پشتیبانی از آیین‌نامه پیش‌گفته در نشانی sabt.irandoc.ac.ir درست شد. دانشجویان پس از تصویب پیشنهاد خود، به این سامانه مراجعه و اطلاعات پیشنهاد خود را ثبت و شناسه ره‌گیری دریافت می‌کنند. «پارسا»ها نیز پس از ثبت و بارگذاری فایل تمام‌متن و تأیید ایرانداک، آماده پذیرش دانشگاه می‌شوند. دریافت شناسه ره‌گیری به‌منزله بارگذاری پایان‌نامه یا رساله در سامانه است. سپس دانشگاه وارد سامانه می‌شود و اطلاعات دانشجو را بررسی و اگر تأیید کند، در سامانه ثبت می‌شود. اطلاعات ثبت شده در این سامانه، برای همانندجویی به‌کار می‌روند و در دسترس همگان نیز گذارده می‌شوند.

ورود به سامانه

رایانامه

گذرواژه

ورود به سامانه

مرا به یاد بسپار گذرواژه‌ام را فراموش کرده‌ام. [ویرایش رایانامه یا شماره ملی](#)

آگهی‌ها

ثبت پیشنهاد، رساله، و پایان‌نامه:

از این پس نیازی به فرستادن لوح فشرده (سی‌دی) پایان‌نامه‌ها و رساله‌ها و نسخه چاپی آنها به ایرانداک نیست.

آیین‌نامه اجرایی ثبت پایان‌نامه و رساله در ایرانداک ابلاغ شد.

درباره

آیین‌نامه ثبت و اشاعه پیشنهادها، پایان‌نامه‌ها، و رساله‌های تحصیلات تکمیلی و صیانت از حقوق پدیدآوران در آنها (شماره ۱۹۵۹۲۹ و تاریخ ۱۳۹۵/۹/۴) از همه دانشگاه‌ها، پژوهشگاه‌ها، مراکز آموزش عالی، پژوهشی، و فناوری دولتی و غیردولتی زیر نظر وزارت علوم، تحقیقات، و فناوری خواسته است که قابل تمام‌متن این مدارک را در پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) ثبت کنند. برای آسان‌سازی این فرایند، ایرانداک سامانه ملی ثبت پایان‌نامه، رساله، و پیشنهاد را در اختیار دانشجویان گرامی قرار داده است...

پایگاه اطلاعات علمی ایران (گنج)

پایگاه اطلاعات علمی ایران (گنج)، دستاورد کاربرد فناوری اطلاعات در خدمات مدیریت اطلاعات علم و فناوری در ایران است. پیشینه این بخش از خدمات ایرانداک به دهه چهل خورشیدی بازمی‌گردد که کار گردآوری، سازمان‌دهی، و اشاعه اسناد علمی و فناوری در مرکز اسناد علمی آغاز شد. بیشترین داده‌های روزآمد گنج از پایان‌نامه‌ها و رساله‌ها و پیشنهادها، علمی قلمرو آن گسترده است و طرح‌های پژوهشی، گزارش‌های دولتی، و مقاله‌های علمی را نیز در بر دارد. کار پایگاه اطلاعات علمی ایران، نخستین بار در سال ۱۳۷۰ در رایانه‌های «مین فریم» و محیط نرم‌افزار «سی‌دی‌اس‌آی‌اس‌آی‌اس» شروع شد. در سال ۱۳۷۳ این پایگاه با فناوری «دایال‌آپ» در شبکه‌ای با نام «صبا» از دور در دسترس کاربران قرار گرفت. سال ۱۳۷۸ بود که اینترنت به ایرانداک وارد شد و دسترسی به گنج نیز در نشانی ganj.irandoc.ac.ir در وب فراهم شد. آخرین ویرایش گنج در آذر ۱۳۹۵ رونمایی گردید. این پایگاه با صدها هزار پیشینه از بزرگ‌ترین گنجینه‌های علمی کشور و هم‌اکنون مرجع بسیاری از پژوهشگران ایران و جهان است. این سامانه بنیان سامانه‌های دیگری در ایرانداک همچون سامانه پیشینه پژوهش نیز هست.

پایگاه اطلاعات علمی ایران

گنج

دانشگاه تهران

پرونده علمی و فناوری اطلاعات ایران (ایرانداک)

صفحه نخست | درباره | آمار | پارسا در جهان | پرسش‌های همیشگی | راهنما | مسئولیت محتوا | تماس با ما | نام‌نویسی / ورود

جست‌وجو

همه فیلدها

برای جست‌وجو، واژه یا عبارتی بنویسید.

جست‌وجوی پیشرفته

پربازدید

موضوع‌های پرشمار

مدیریت ۸۲۱۲

مهندسی عمران ۶۷۱۳

مهندسی مکانیک ۵۳۲۵

ریاضیات ۵۱۳۸

مهندسی برق و الکترونیک ۵۰۸۹

کشاورزی، عمومی و میان رشته‌ای ۴۹۸۶

معارف اسلامی ۴۴۲۶

کلیدواژه‌های پرشمار

ایران (مدیریت بحران) ۳۱۴۶

نانوذره (فیزیک) ۲۱۵۲

یورس اوراق بهادار تهران ۱۹۵۰

تهران (شهر) ۱۵۶۷

شبکه عصبی (فنی و مهندسی) ۱۴۸۹

ارزیابی عملکرد (مدیریت بحران) ۱۲۳۳

مطالعه تطبیقی (ارتقای بهداشت) ۱۱۸۹

پژوهشگران پراثر

ابوالفضل فراهانی ۱۹۸

غلامرضا بخشی خانیکی ۱۷۴

علی ربیعی ۱۷۰

عادل آذر ۱۶۴

علی رستمی ۱۵۹

احمد برجعلی ۱۵۴

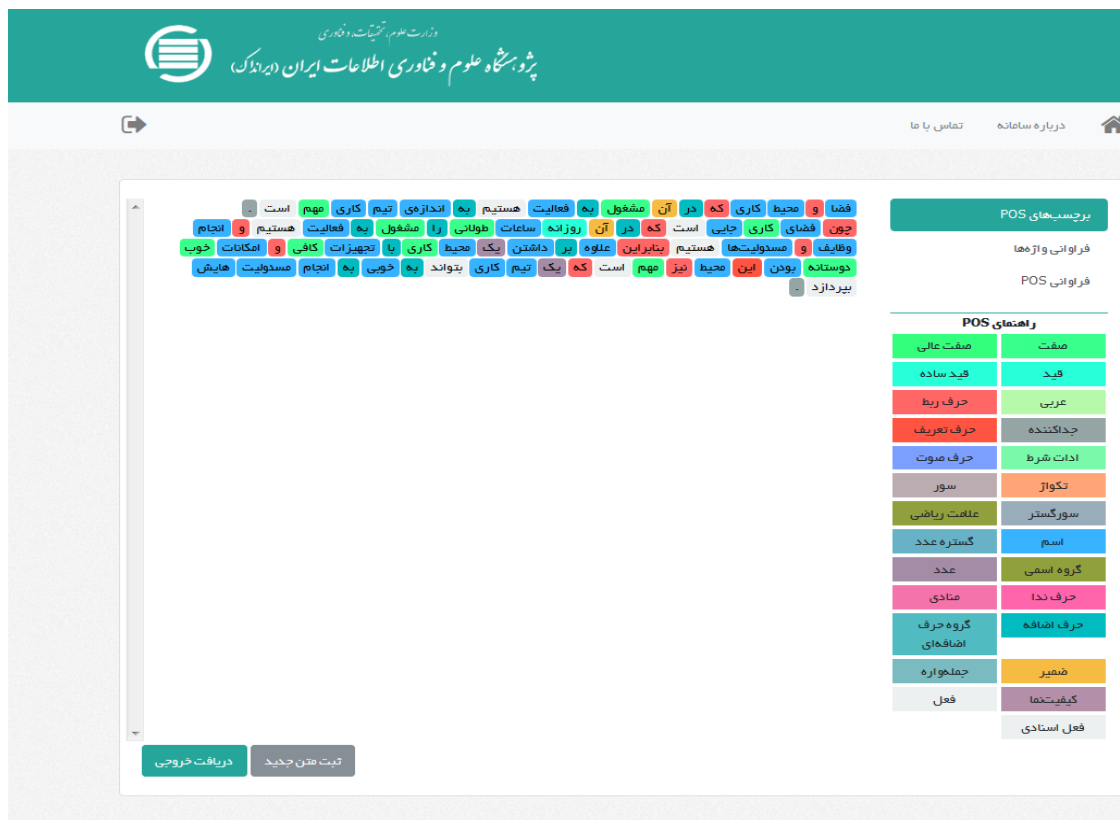
رضا رسولی ۱۵۳

سامانه پیشینه پژوهش

در پاسخ به خواست جامعه علمی کشور برای هدفمندسازی و افزایش کارایی و اثربخشی پژوهش‌ها، ایرانداک سامانه پیشینه پژوهش را با پشتوانه در حال افزایش صدها هزار «پارسا» از آبان ۱۳۹۳ راه‌اندازی کرده و در نشانی pishineh.irandoc.ac.ir در اختیار همه پژوهشگران کشور قرار داده است. راه‌اندازی سامانه پیشینه پژوهش، گامی در کمک به پیشبرد پژوهش و گسترش علم و فناوری با آگاهی از پژوهش‌های انجام شده و زمینه‌سازی برای پیش‌گیری از دوباره‌کاری در پژوهش است. در این سامانه، با دریافت عنوان و کلیدواژه‌های پژوهش از کاربر، با جست‌وجوی کارشناسان آموزش‌دیده در پایگاه اطلاعات گنج، اطلاعات پایان‌نامه‌ها و رساله‌های انجام شده به آگاهی کاربر می‌رسد.

سامانه همانندجو

برای اجرای سیاست‌های کلی علم و فناوری و در پاسخ به خواست جامعه علمی کشور برای گسترش اخلاق پژوهش و پشتیبانی از مالکیت فکری و معنوی و هم‌چنین پیش‌گیری از بدرفتاری‌های علمی، به‌ویژه سایه‌نویسی و دستبرد علمی در نگارش پایان‌نامه‌ها و رساله‌ها، ایرانداک سامانه «همانندجو» را در نشانی tik.irandoc.ac.ir با پشتوانه تمام‌متن و فراینده ۳۰۰ هزار «پارسا» ۲۰ آذر ۱۳۹۵ راه‌اندازی کرد. همانندجویی در نوشتار پایان‌نامه‌ها و رساله‌ها، گامی در کمک به نگه‌داشت حقوق پدیدآوران و گسترش علم و فناوری و زمینه‌سازی برای دسترسی آزاد همگان به اطلاعات است.



سامانه هوشمند استخراج پژوهشگران مشابه

در پایگاه داده‌های ثبت و گنج پژوهشگاه، یک پژوهشگر مشخص به اشکال مختلف ثبت شده است. دلایلی مانند چند املائی بودن برخی نام‌ها، عدم دقت دانشجو، غلط‌های املائی و ... را می‌توان برای علت گوناگونی نام پژوهشگر نام برد. همچنین اقلام اطلاعاتی شناسه ملی، شماره شناسنامه، تاریخ تولد در حجم عظیمی از داده‌ها معتبر نمی‌باشند. سامانه هوشمند استخراج پژوهشگران مشابه با استفاده از الگوریتم‌های زبان‌شناسی رایانشی و تطبیق هوشمند رشته، پژوهشگران یکتایی که به اشکال گوناگون ثبت شده‌اند را استخراج می‌کند.

این سامانه نقش بسزایی در پاکسازی و تجمیع پژوهشگران داشته و امکان طراحی و پیاده‌سازی سرویس‌های ارزش افزوده را میسر می‌سازد. هم‌اکنون حدود ۸۰۰ هزار پژوهشگر در این سامانه در حال تجمیع و پاکسازی می‌باشد. این سامانه جهت استفاده معاونت اطلاعات علم و فناوری ایران در پژوهشگاه طراحی و پیاده‌سازی گردیده است.

ردیف	شماره پروژه	کد های	نام	نام خانوادگی	فنی و مهندسی	علوم انسانی	علوم پایه	هنر و معماری	کشاورزی	علوم پزشکی	نام پزشکی	نام خانوادگی
1	14319	95.0	نصراله	عالم چرکری	49	0	1	0	0	0	0	1
2	119772	95.0	نصراله	عالم چرکری	50	0	1	0	0	0	0	0
3	128125	95.0	نصراله	عالم چرکری	1	0	0	0	0	0	0	0
4	349969	95.0	نصراله	عالم چرکری	7	0	0	0	0	0	0	0
5	347202	95.0	نصراله	عالم چرکری	10	0	0	0	0	0	0	0
6	422765	95.0	نصراله	عالم چرکری	0	0	2	0	0	0	0	0
7	469954	95.0	نصراله	عالم چرکری	3	0	0	0	0	0	0	0
8	469548	95.0	نصراله	عالم چرکری	1	0	0	0	0	0	0	0
9	675790	95.0	نصراله	عالم چرکری	1	0	0	0	0	0	0	0
10	692023	95.0	نصراله	عالم چرکری	1	0	0	0	0	0	0	0
11	698555	95.0	نصراله	عالم چرکری	1	0	0	0	0	0	0	0

سامانه جامع اصطلاح‌نامه‌های علمی و فنی

در این سامانه؛ esn.irandoc.ac.ir؛ تمامی اصطلاح‌نامه‌هایی که تاکنون توسط ایرانداک منتشر شده‌اند بارگذاری شده و امکان جستجوی واژه‌ها به صورت جداگانه در هر اصطلاح‌نامه و جستجو در چندین اصطلاح‌نامه به صورت همزمان، وجود دارد. همچنین امکان مشاهده ساختار درختی واژگان به صورت نمایه الفبایی به دو زبان فارسی و انگلیسی برای واژه‌ها وجود دارد. اصطلاح‌نامه‌ها گنجینه‌ای از واژه‌ها هستند. میان این واژه‌ها، نظامی شبکه‌ای و مفهومی برقرار است. اصطلاح‌نامه‌ها در سازماندهی اطلاعات، بازیابی از پایگاه‌های اطلاعات، تولید هستان‌شناسی‌ها (آنتولوژی‌ها)، تحلیل اطلاعات، ترجمه ماشینی و نمایه‌سازی ماشینی، و تولید محتوای آموزشی کاربرد دارند. ایرانداک کار روی اصطلاح‌نامه‌ها را از سال ۱۳۵۰ با نگارش دست‌نامه «قواعد و قراردادهای ساختن اصطلاح‌نامه» آغاز کرد و در سال ۱۳۷۵ نخستین اصطلاح‌نامه خود را با نام «نظام مبادله اطلاعات علمی فنی (نما)» به چاپ رساند. در سال‌های ۱۳۷۶ تا ۱۳۸۶ نیز طراحی و نگارش اصطلاح‌نامه دو زبانه فارسی-انگلیسی فراگیر در حوزه‌های گوناگون علم و فناوری مانند ریاضی، فیزیک، شیمی، علوم زیستی، علوم زمین، مهندسی، کشاورزی، مدیریت بحران، ارتقای بهداشت، و جامعه‌شناسی با بیش از ۱۰۰ هزار واژه انجام شد. در سال ۱۳۹۸، چند ده هزار واژه دیگر به واژه‌های پیشین افزوده شد و روابط معنایی میان اصطلاحات آنها گسترش و بهبود یافت و نسخه دوم سامانه اصطلاح‌نامه‌ها نیز با فارسی‌سازی نرم‌افزار منبع‌باز «اسکاسموس» در دسترس همگان گذارده شد. دستاورد این خدمت ایرانداک هم‌اکنون به شانزده اصطلاح‌نامه رسیده است. این وبگاه می‌تواند راهنمای مناسبی برای کاربران و نمایه‌سازان در سازماندهی و بازیابی اطلاعات باشد. ایرانداک به عنوان یکی از مراجع ملی در این حوزه، آمادگی دارد تا نسبت به طراحی، تولید،

تدوین و روزآمدسازی اصطلاح‌نامه‌ها، تدوین فرا اصطلاح‌نامه‌ها (متاتزاروس‌ها)، و ساخت وب‌سرویس برای اصطلاح‌نامه‌ها اقدام نماید. در تابستان ۱۳۹۸، ده‌ها هزار واژه دیگر به واژه‌های پیشین در اصطلاح‌نامه‌ها افزوده شد و روابط معنایی میان آنها گسترش و بهبود یافت و نسخه دوم سامانه اصطلاح‌نامه‌ها نیز با امکانات بسیار در نشانی esn.irandoc.ac.ir به دست معاون پژوهش و فناوری وزارت عتف رونمایی شد.

پژوهش

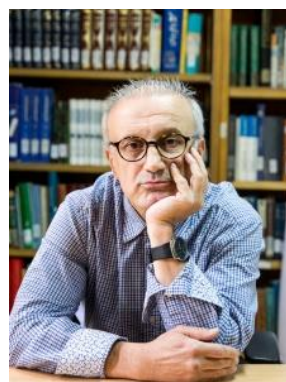
یکی از رویکردهای کلیدی ایرانداک به گسترش خط و زبان فارسی، راهاندازی گروه پژوهشی و انجام پژوهش در این زمینه بوده است.

پژوهشکده، گروه و اعضای هیئت علمی

ایرانداک در تیر ۱۳۹۶ دو گروه پژوهشی را تحت عناوین «زبان‌شناسی رایانشی» و «اصطلاح‌شناسی و هستان‌شناسی» در شورای گسترش آموزش عالی به تصویب رساند که هر دو در پژوهشکده «علوم اطلاعات» قرار گرفته‌اند. این دو گروه هم‌اکنون **شش** عضو هیئت علمی دارد.



دکتر ملوک السادات حسینی بهشتی
دانش‌آموخته دکتری تخصصی زبان‌شناسی از دانشگاه تهران
پژوهشکده علوم اطلاعات
گروه پژوهشی اصطلاح‌شناسی و هستان‌شناسی



دکتر سیدمهدی سمائی
دانش‌آموخته دکتری تخصصی زبان‌شناسی از دانشگاه تهران
پژوهشکده علوم اطلاعات
گروه پژوهشی اصطلاح‌شناسی و هستان‌شناسی



دکتر الهام علایی ابوذّر
دانش‌آموخته دکتری تخصصی زبان‌شناسی همگانی از
دانشگاه تهران
پژوهشکده علوم اطلاعات
گروه پژوهشی زبان‌شناسی رایانشی



تقی رجایی
دانش‌آموخته کارشناسی ارشد شیمی فیزیک از
دانشگاه تربیت مدرس
پژوهشکده علوم اطلاعات
گروه پژوهشی اصطلاح‌شناسی و هستان‌شناسی



دکتر نصراله پاک‌نیت

دانش‌آموخته دکتری تخصصی ریاضی (علوم کامپیوتر) از
دانشگاه شهید بهشتی
پژوهشکده علوم اطلاعات
گروه پژوهشی زبان‌شناسی رایانشی



دکتر جلال‌الدین نصیری

دانش‌آموخته دکتری تخصصی مهندسی کامپیوتر از
دانشگاه تربیت مدرس
پژوهشکده علوم اطلاعات
گروه پژوهشی زبان‌شناسی رایانشی

پژوهش

اصطلاح‌نامه نگاری

اصطلاح‌نامه گنجینه‌ای از واژه‌هاست که افزون بر نظم الفبایی موجود در فرهنگ‌ها، دارای نظامی شبکه‌ای و مفهومی میان واژه‌های یک یا چند حوزه از دانش بشری است. این نظام مفهومی، ارتباطات معنایی گوناگون میان واژه‌ها را نشان می‌دهد. اصطلاح‌نامه‌ها در سازماندهی اطلاعات برای ذخیره و نمایه‌سازی، بازیابی اطلاعات به کمک شبکه معنایی، تحلیل اطلاعات و علم‌سنجی، مصورسازی پژوهش، تحلیل متن و پردازش زبان علمی، طراحی نقشه‌های مفهومی علم در حوزه‌های گوناگون، طراحی ساختار و محتوای آموزشی و آسان‌سازی یاددهی و یادگیری، داده‌کاوی، ترجمه ماشینی و نمایه‌سازی ماشینی، معادل‌یابی فارسی واژه‌های بیگانه و شناخت اصطلاح‌های استاندارد به کار می‌روند. اصطلاح‌نامه در ساخت آنتولوژی (هستان‌شناسی) به مثابه ابزاری برای وب معنایی، نیازی کلیدی و بدون جایگزین است. ایرانداک در سال ۱۳۷۶ نگارش نخستین اصطلاح‌نامه خود را با نام اصطلاح‌نامه فنی و مهندسی آغاز و تا سال ۱۳۸۶، ده اصطلاح‌نامه دو زبانه با نام‌های؛ ارتقای بهداشت، جامعه‌شناسی، ریاضیات، شیمی، علوم زمین، علوم زیستی، علوم کشاورزی، فنی و مهندسی، فیزیک و مدیریت بحران؛ را با بیش از یکصد هزار واژه، تدوین و به جامعه علمی کشور تقدیم کرده است. ایرانداک با سازمان‌های جهانی اصطلاح‌شناسی مانند «ترم‌نت» و «اینفوترم» نیز همکاری دارد و توانسته است با کمک یونسکو برخی از اصطلاح‌نامه‌های خود را در سایت‌های جهانی نیز بگذارد.

توسعه و روزآمدسازی اصطلاح‌نامه‌های ایرانداک

توسعه علم و فناوری، تولید واژه‌های جدید، افزایش انتشارات علمی، واژه‌هایی که توسط کاربران وارد نظام سازماندهی اطلاعات می‌شود، و تغییراتی که در روابط بین مفاهیم و موضوعات ایجاد می‌شود از مجموعه دلایل روزآمدسازی اصطلاح‌نامه‌ها هستند. در این پژوهش، افزایش کمی، ویرایش و توسعه روابط معنایی میان واژه‌های اصطلاح‌نامه‌های علمی و فنی تدوین شده در پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، صورت گرفته است. فرایند روزآمدسازی و توسعه، شامل مراحل مرور ادبیات و تهیه منابع، طراحی و بررسی نقشه‌های مفهومی، ترجمه واژگان، تعیین روابط معنایی، یکسان‌سازی، واپایش روابط و خطایابی، تولید خروجی‌های اصطلاح‌نامه‌ای، ویرایش، و انتشار است. در سال ۱۳۹۸، چند ده هزار واژه دیگر به واژه‌های پیشین افزوده شد و روابط معنایی میان اصطلاحات آنها گسترش و بهبود یافت و نسخه دوم سامانه اصطلاح‌نامه‌ها نیز با فارسی‌سازی نرم‌افزار منبع‌باز «اسکاسموس» در دسترس همگان گذارده شد. دستاورد این خدمت ایرانداک هم‌اکنون به شانزده اصطلاح‌نامه رسیده است. این وبگاه می‌تواند راهنمای مناسبی برای کاربران و نمایه‌سازان در سازماندهی و بازیابی اطلاعات باشد.

توسعه و روزآمدسازی اصطلاح‌نامه شیمی

روزآمدسازی و ویرایش مجموعه واژگان اصطلاح‌نامه شیمی تدوین شده در ایرانداک موضوع این پژوهش است. در روزآمدسازی، اکثر مراحل که در تألیف و تدوین و یا ترجمه اصطلاح‌نامه وجود دارد طی خواهد شد. این مراحل در سطح کلان عبارتند از: ۱- مرور ادبیات و تهیه منابع؛ ۲- طراحی و یا بررسی شبکه مفهومی؛ ۳- ترجمه واژگان؛ ۴- تعیین روابط معنایی؛ ۵- یکسان‌سازی؛ ۶- کنترل روابط و خطایابی؛ ۷- تولید خروجی‌های اصطلاح‌نامه‌ای؛ ۸- ویرایش؛ ۹- چاپ و انتشار. نمایه‌سازان در بسیاری از موارد هنگام استفاده از اصطلاح‌نامه در سازماندهی اطلاعات، با اصطلاحاتی مواجه می‌شوند که آنها را در اصطلاح‌نامه‌های موجود نمی‌یابند. رشد علم و فناوری و افزایش انتشارات علمی و انتخاب توصیفگرهای مناسب‌تر هنگام سازماندهی اطلاعات از یک سو و نیز دیدگاه‌های متفاوت کاربران نهایی نظام‌های بازیابی اطلاعات از سویی دیگر، روزآمدسازی و گسترش روابط معنایی در اصطلاح‌نامه را ضروری می‌نماید. روزآمدسازی اصطلاح‌نامه شیمی با استفاده از واژگان کنترل شده پایگاه اطلاعات علمی ایران (گنج) و مجموعه واژگان شیمی مرکز نشر دانشگاهی و نیز مجموعه واژه‌های شیمی مصوب فرهنگستان زبان و ادب فارسی انجام شد.

طراحی ساختار درختی و تدوین اصطلاح‌نامه شیمی از واژگان مصوب فرهنگستان

واژه‌های فارسی مصوب فرهنگستان زبان و ادب فارسی، فاقد ساختار درختی و تزاروسی است. در این پروژه، این واژه‌ها در یک ساختار اصطلاح‌نامه‌ای ارائه شد. عرضه الگویی به فرهنگستان برای استفاده از شبکه معنایی جهت سازماندهی واژه‌های مصوب، در کلیه گرایش‌ها یکی از اهداف این پژوهش است. واژگان شیمی مصوب فرهنگستان از طریق سایت فرهنگستان و نیز از کتاب "هزار واژه شیمی" فرهنگستان قابل دستیابی است. در طراحی ساختار درختی از روش استقرایی (حرکت از جز به کل) استفاده شد. در این مرحله اشراف موضوعی مهمترین عامل در چینش واژگان بود. بنابراین در حوزه‌های مختلف شیمی از مشاوره متخصصین بهره‌مند شده‌ایم. در تدوین اصطلاح‌نامه و تولید خروجی‌های اصطلاح‌نامه‌ای از نرم‌افزار "تزاروس بیلدر" استفاده شد و تمام روش‌های کلاسیک تدوین اصطلاح‌نامه مورد استفاده قرار گرفت. تولید اصطلاح‌نامه‌ها به دلیل کاربرد مستقیم آن در استاندارد کردن و بازیابی اطلاعات در راستای برنامه استراتژیک و نیز اساس‌نامه ایرانداک است.

اصطلاح‌نامه دبیرستان و آموزش نوین (اصطلاح‌نامه آموزشی)

در اکثر پژوهش‌های انجام شده در ارتباط با روان‌شناسی یادگیری و آموزش، الگوها و روش‌های تدریس، برنامه‌ریزی آموزشی، آموزش استراتژیک تکنولوژی آموزشی در کنار صدها اصطلاح و مبحث تخصصی همواره با اصطلاح مواد درسی (ماده درسی) و یا محتوای آموزشی برخورد می‌کنیم. مواد یا محتوای آموزشی در شیوه‌های سنتی و مدرن در مبحث تدریس و یادگیری انواع و اقسامی دارند ولی شاکله همه آنها متون کتابی است. ارائه مواد درسی (محتوای آموزشی) در قالب اصطلاح‌نامه تصویری در کنار (یا به جای) متون کتابی به منظور افزایش یادگیری فراگیران فرضیه‌ای است که در این طرح مورد آزمون قرار گرفته است. برای اصطلاح‌نامه (تزاروس) دو کاربرد اصلی: ۱. استاندارد کردن اطلاعات؛ ۲. بازیابی اطلاعات ذکر می‌کنند. در این طرح نشان داده شده که اولاً می‌توان محتوای آموزشی (ماده درسی) را در قالب ساختاری اصطلاح‌نامه‌ای ارائه کرد، دوماً: این قالب ساختاری به عنوان یک ابزار موثر آموزشی و یا کمک آموزشی در آموزش علوم کاربرد داشته و موجب افزایش یادگیری فراگیران خواهد شد. در این طرح، در ابتدا به عنوان مطالعه موردی کتاب شیمی سال اول دبیرستان انتخاب شده و کل مفاهیم و موضوعاتی مطرح شده در این طرح، در قالب یک اصطلاح‌نامه تصویری ارائه شده است. سپس، با استفاده از دو روش آزمون و پیمایش به سنجش نمرات و نظرات گروه‌های آزمون و کنترل در چند دبیرستان تهران اقدام شده است. این طرح از هر دو منظر بنیادی و کاربردی می‌تواند از مصادیق رابط بین علوم اطلاع‌رسانی، ارتباطات، مهندسی و مدیریت دانش،

روان‌شناسی، تکنولوژی آموزشی و نیز رایانه قلمداد گردد.

تدوین روش‌شناسی تحلیل اطلاعات علمی شیمی (علوم پایه) و ترسیم نقشه علمی شیمی در ایران بر اساس آن (با بررسی پایگاه اطلاعاتی پایان‌نامه‌های پژوهشگاه)

در این پژوهش ابتدا الگویی جهت تحلیل اطلاعات علمی پایان‌نامه‌های شیمی (علوم پایه) ارائه شد، سپس با استفاده از این الگو به تحلیل پایان‌نامه‌های شیمی موجود در پایگاه اطلاعات پژوهشگاه علوم و فناوری اطلاعات ایران در یک دهه اخیر پرداخته شد. با استفاده از منابع معتبر شیمی در سطح کارشناسی و تحصیلات تکمیلی و نیز اصطلاح‌نامه شیمی و با تأیید نظر متخصصان در هر یک از گرایش‌های چهارگانه شیمی علوم پایه (شیمی فیزیک، شیمی آلی، شیمی معدنی، و شیمی تجزیه) به تهیه نقشه علم شیمی همت گماشته شد. در حقیقت تحلیل اطلاعات علمی شیمی، بدون در اختیار داشتن الگوی مدونی از علم شیمی بسیار مشکل خواهد بود. جست‌وجوی فراوان برای یافتن چنین الگویی بی‌نتیجه بود، لذا الگوی مدونی از علم شیمی (علوم پایه) برای اولین بار به این صورت ارائه شد، که علاوه بر استفاده در تحلیل اطلاعات علمی، قابلیت استفاده آموزشی هم دارد و البته با کمک متخصصان می‌توان به بسط و غنای علمی آن افزود. تحلیل علمی اطلاعات پایان‌نامه‌های شیمی که یک تحلیل توصیفی است می‌تواند در سیاست‌گذاری‌های تحقیقاتی نقشی کلیدی ایفا نماید.

شناسایی مسائل و ارائه راهکارهای زبان‌شناختی سازمان‌دهی و بازیابی اطلاعات در سامانه اطلاعات علمی ایران (گنج)

یکی از چالش‌هایی ساماندهی و بازیابی اطلاعات ساختار زبانی متن است. ساختار زبانی متن شامل ویژگی‌های ساختارزی-نحوی-معنایی و لغوی متن است که تأثیر بسزایی بر دقت بازیابی دارد. در سامانه گنج به عنوان مثال، مسائل خط فارسی از جمله عواملی است که سبب کاهش دقت بازیابی می‌شود. در این پژوهش، در ابتدا عوامل زبان‌شناختی سازمان‌دهی و بازیابی اطلاعات در سامانه اطلاعات علمی ایران (گنج) به روش میدانی شناسایی و بررسی شده و دسته‌بندی آنها بر اساس ادبیات موجود انجام شده است. در ادامه، بر پایه مطالعات و روش‌های بکار گرفته شده در پژوهش‌های موجود در این زمینه، راه‌کارها یا راه‌حل‌هایی برای رفع یا بهبود موارد شناسایی شده، ارائه شده است.

طراحی یک روش هوشمند تجزیه رشته‌های مرجع در زبان فارسی

یک رشته مرجع را می‌توان به عنوان مجموعه‌ای از مولفه‌ها مانند نام نویسندگان، عنوان، محل نشر، سال نشر، شماره صفحات و ... در نظر گرفت. در حالی که تجزیه رشته‌های مرجع موجود در انتهای یک

مدرک علمی توسط کاربر انسانی به راحتی انجام پذیر است، تنوع موجود در فرمت های نگارش رشته های مرجع در کنار اشتباهات رخ داده توسط نویسندگان در نگارش این رشته ها، خود کارسازی انجام این عملیات را دشوار نموده است. روش های زیادی برای خود کارسازی تجزیه رشته های مرجع ارائه شده است اما این روش ها وابسته به زبان بوده و استفاده از روشی ارائه شده برای یک زبان در زبانی دیگر منجر به نتایجی اشتباه می شود. تحقیقات صورت گرفته بیانگر این است که تاکنون هیچ روشی برای خود کارسازی تجزیه رشته های مرجع در زبان فارسی ارائه نشده است. با توجه به این مهم و نقش گسترده این مسأله در ساخت خود کار شبکه های استنادی مدارک علمی و فرایندهای بازیابی اطلاعات، در این طرح پژوهشی به این مسأله پرداخته شده است. در این راستا، پس از بررسی پیشینه مسأله، از روش یادگیری ماشین بردار پشتیبان به عنوان یک دسته بند چند دسته ای استفاده شده و یک روش هوشمند برای مسأله تجزیه رشته های مرجع در زبان فارسی ارائه شده است. با توجه به اهمیت انتخاب ویژگی های مناسب برای استفاده در دسته بند ماشین بردار پشتیبان، در این پژوهش این مهم با توجه به ویژگی های استفاده شده در زبان انگلیسی و ویژگی های زبان فارسی و ارجاع دهی در این زبان صورت گرفته است. در ادامه، روش پیشنهادی پیاده سازی شده و با استفاده از مجموعه داده ای شامل ۹۲۲ رشته مرجع فارسی آموزش داده شده و مورد آزمایش قرار گرفت. نتایج به دست آمده نشانگر مقدار ۹۲ درصد برای پارامترهای دقت، فراخوانی و معیار "ف" است.

ارائه یک ماژول نرم افزاری در راستای ایجاد شبکه استنادی پارساهای موجود در پایگاه اطلاعات علمی ایران (گنج)

پایگاه اطلاعات علمی ایران (گنج) از مهم ترین منابع علمی زبان فارسی بوده که توسط پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) توسعه یافته و دربرگیرنده اغلب پارساهای خاتمه یافته پژوهشگران ایرانی است. با توجه به اهمیت این پایگاه اطلاعاتی و عدم وجود روش هایی برای ساخت خود کار شبکه های تحلیل استنادی در زبان فارسی، در این پژوهش به مسأله ساخت خود کار شبکه های تحلیل استنادی در پارساهای فارسی زبان پرداخته شده است. در این راستا، در ابتدا مسأله ساخت خود کار شبکه تحلیل استنادی را به زیرمسأله هایی چون استخراج مراجع از تمام متن، تعیین نوع مراجع، تجزیه مراجع و بررسی تطابق مراجع و مولفه های آن ها تقسیم شده و در ادامه، با در نظر گرفتن این مسائل به عنوان مسائلی از جنس دسته بندی، از روش های یادگیری ماشینی و الگوریتم های تطبیق رشته تقریبی برای حل آن ها استفاده شده است. گفتنی است که با توجه به فراوانی ارجاعات به منابع انگلیسی زبان در مدارک علمی فارسی، راه کارهای ارائه شده علاوه بر مراجع فارسی زبان، مراجع انگلیسی زبان را نیز

در نظر می گیرند. روش یادگیری ماشینی استفاده شده در این پژوهش، روش بردار ماشین پشتیبان بوده که از نوع با نظارت است. پس از ارائه راه کار برای زیر مساله های موردنظر، آن ها با یکدیگر ترکیب و تحت یک ماژول نرم افزاری ارائه شده است. ماژول ارائه شده، با دریافت تمام متن و فراداده های متناظر مجموعه ای از پارساها، شبکه تحلیل استنادی مربوطه را ایجاد کرده و در قالب Json خروجی می دهد. نتایج ارزیابی راه کار نهایی ایجاد شده نشان دهنده مقادیر ۱ به عنوان پارامتر دقت، ۰,۷۸ به عنوان پارامتر فراخوانی و ۰,۸۸ به عنوان پارامتر F۱ در بازیابی و تشخیص تطابق رشته های مرجع است.

طراحی سامانه برچسب دهی به اجزای کلام برای متون فارسی

در این پژوهش به این مسأله پرداخته شد که آیا با رفع ابهام از برچسب نحوی هم نگاره های اسمی و صفتی مختوم به «-ی»، که فراوانی بالایی در پیکره های متنی فارسی دارند، کارایی یک سیستم برچسب زنی خود کار، افزایش می یابد و در نهایت می توان سامانه ای طراحی کرد که عمل برچسب دهی خود کار را با در نظر گرفتن رفع ابهام از برچسب هم نگاره های اسمی و صفتی مختوم به «-ی» در فارسی، با کارایی بهتری انجام دهد؟ سیستم مورد مطالعه در این پژوهش، سیستم «هضم» بود. نرم افزاری جهت رفع ابهام از برچسب نحوی هم نگاره های اسمی و صفتی مختوم به «-ی» در فارسی، تهیه شد که خود مبتنی بر الگوهای حساس به بافت نحوی است که بر اساس این الگوها می توان برچسب درست را به هم نگاره های مذکور اختصاص داد. ارزیابی کلی نرم افزار تهیه شده جهت رفع ابهام از برچسب نحوی هم نگاره های اسمی و صفتی مختوم به «-ی» در فارسی، نشان می دهد اگر تنها الگوهای حساس به بافت نحوی که تأثیر مثبت در برچسب زنی داشته اند را به برچسب زن «هضم» اضافه کنیم، صحت (Accuracy) کلی برچسب زن ۹۵,۶۹۱ درصد می شود که ۱,۳۴ درصد نسبت به حالتی که از تمام الگوهای حساس به بافت نحوی استفاده شود، بالاتر است.

ارایه روشی هوشمند برای استخراج کلیدواژه از مستندات علمی زبان فارسی بر اساس سیستم های پیشنهاددهنده

با توجه به افزایش حجم تولید و ثبت مستندات علمی، به خصوص در پایگاه گنج، نیاز است که فرایند نمایه سازی با سرعت بیشتری انجام شود. هم اکنون علاوه بر متوسط نرخ ثبت روزانه ۲۰۰ عنوان پایان نامه در روز در گنج، حدود ۱۴۰ هزار سند موجود است که هنوز نمایه نشده اند. در حال حاضر فرایند نمایه سازی در ایرانداک بر اساس دانش و تخصص نمایه سازان انجام می شود و روش های ماشینی برای پردازش خود کار متون و نمایه سازی آنها استفاده نمی شود. در بسیاری از پایگاه های علمی در دنیا

از روش‌های ماشینی و خودکار در کلیه فعالیت‌های فرایند نمایه‌سازی یا بخشی از آنها استفاده می‌شود. تعدادی از این روش‌ها بر مبنای تحلیل آماری متون و استفاده از روش‌های یادگیری ماشین هستند، تعدادی بر مبنای تحلیل معنایی متون به واسطه اصطلاح‌نامه‌های تخصصی و هستان‌شناسی، و در تعدادی دیگر از این روش‌ها از تلفیق هر دو استفاده می‌شود. در این پژوهش، با هدف تسریع فرایند نمایه‌سازی، با استفاده از روش‌های یادگیری ماشین و تحلیل آماری متون روشی ارائه شد که از طریق آن برای نمایه‌سازی یک سند جدید، مجموعه‌ای از کلیدواژه‌های مرتبط به نمایه‌ساز پیشنهاد شود تا نمایه‌ساز سریع‌تر بتواند از بین آنها کلیدواژه‌های مناسب را انتخاب کند. پیشنهاد کلیدواژه به نمایه‌ساز، بر اساس رویکرد سیستم‌های پیشنهاددهنده و با اتکا به دانش ضمنی موجود در مجموعه مستندات نمایه شده قبلی صورت گرفت.

طراحی و ساخت سامانه‌ی استانداردساز و خطایاب متون فارسی

روزانه هزاران مستند متنی متنوع در حوزه‌های مختلف علمی بر روی وب جهان گستر قرار می‌گیرد. این مستندات می‌تواند شامل پایان‌نامه‌ها، مقاله‌ها، گزارش‌های علمی و مواردی از این قبیل باشد. نگارش متن این مستندات علمی جهت حفظ یکنواختی باید بر اساس اصول ثابت انجام گیرد، اما همواره به‌طور غیرعمدی دستخوش سلیقه‌های مختلفی در طول تاریخ می‌شود. اگرچه این تغییرات ناشی از پویا بودن زبان و خلاقیت ذهن بشری است، اما این پویایی و خلاقیت پردازش ماشینی متن را با چالش‌های متعددی روبرو می‌کند و دقت پردازش داده‌ها را به میزان چشمگیری پایین می‌آورد. علاوه بر تنوع نگارشی، غلط‌های سهوی املائی نیز وجود دارد که فحوای گفتمانی متن را منحرف کرده و درک آن را با مشکل مواجه می‌کند. بنابراین، کلیه نویسه‌های متن باید به حالت استاندارد تبدیل شوند و عاری از هرگونه خطاهای املائی گردند. در این پژوهش، سامانه‌ای برای استانداردسازی و خطایابی متون علمی فارسی طراحی شده که متون نوشتاری علمی و تخصصی فارسی را به‌لحاظ صحت نگارشی و املائی بررسی کرده و متن را به‌شکل استاندارد درمی‌آورد.

عرضه قواعد سرواژه‌سازی در زبان فارسی

سرواژه‌سازی یکی از فرایندهای واژه‌سازی و یکی از انواع اختصارسازی است. در این نوع واژه‌سازی، با حروف آغازین کلمات یا با حروف آغازین و برخی حروف دیگر کلمات و اصطلاحات صورت‌های جدیدی ساخته می‌شود. قواعد کلی سرواژه‌سازی در زبان فارسی شبیه به قواعد سرواژه‌سازی در زبان‌های غربی است اما برخی ویژگی‌های خط فارسی و ویژگی‌های آوایی زبان

فارسی و حتی مسائل فرهنگی نیز قواعد سرواژه‌سازی در زبان فارسی موثرند. در این پژوهش تلاش شده با بررسی مسائل سرواژه‌سازی در زبان فارسی، قواعدی برای آن عرضه شود.

تعیین قواعد صوری همگون گروه‌های دستوری برای کاربرد در پردازش زبان

پردازش زبان یکی از حوزه‌های پژوهشی هوش مصنوعی است. ترجمه ماشینی و درک متن دو زمینه اصلی پژوهش در حوزه پردازش زبان‌اند. در دست داشتن قواعد صوری زبان و تبدیل آنها به قواعد و کدهای قراردادی از ملزومات و داده‌های اصلی در این گونه پژوهش‌ها است. استخراج قواعد دستوری گروه‌های دستوری و عرضه کدها و قواعد قراردادی مقدمه توصیف صوری دستور زبان است. تاکنون گروه‌های دستوری زبان فارسی به طور جداگانه بررسی و برای آنها کدها و قواعد قراردادی عرضه شده است. در این پژوهش کدهایی که تاکنون برای گروه‌های دستوری در طرح‌های جداگانه انتخاب شده بررسی و همگون شده و فهرست درهم کرد و جامع آنها عرضه شده است.

طراحی یک الگوریتم همانندجو برای تشخیص متون بازنویسی شده در زبان فارسی

همانندجویی ابزاری است که از آن برای تشخیص سرقت علمی/ادبی استفاده می‌شود. در یک روش همانندجویی، هدف تشخیص تمام قسمت‌های همانند موجود در یک متن مشکوک با توجه به تعدادی متن منبع احتمالی است. روش‌های زیادی برای همانندجویی ارائه شده اما از یک طرف، استفاده از روش‌های همانندجوی موجود برای سایر زبان‌ها به منظور همانندجویی در زبان فارسی مناسب نیست و از طرف دیگر، اغلب روش‌های ارائه شده برای همانندجویی در زبان فارسی قادر به تشخیص متون بازنویسی شده نیستند. با توجه به این مهم، در این پژوهش، دو روش همانندجویی جدید با هدف تشخیص متون فارسی بازنویسی شده، ارائه شده است. روش اول پیشنهادی روشی معنایی است و از لغت‌نامه جهت بررسی همانندی جملات متون استفاده می‌کند. روش دوم پیشنهادی روشی احتمالاتی است و از اطلاعات آماری به دست آمده از پیکره‌ای عظیم از متون برای همانندجویی استفاده می‌کند. علاوه بر این، در حالیکه در سایر روش‌های موجود، همانندی هر دو جمله از متون موردنظر به صورت مستقل بررسی می‌شود، در روش‌های پیشنهادی همانندی جملات همسایه نیز در بررسی همانندی دو جمله در نظر گرفته شده است. نتایج پیاده‌سازی و آزمایشات صورت گرفته بر روی روش‌های پیشنهادی نشان می‌دهد که در حالیکه هر دو روش از کیفیت مناسب و تقریباً یکسانی برخوردار هستند، روش همانندجوی احتمالاتی پیشنهادی بسیار کاراتر است.

طراحی و پیاده‌سازی آزمایشی موتور سامانه هوشمند استخراج پژوهشگران مشابه

پژوهشگاه علوم و فناوری اطلاعات ایران با جمع‌آوری و پردازش تمام پایان‌نامه‌ها و رساله‌های وزارت علوم، تحقیقات و فناوری از جایگاه ممتاز و رو به رشدی در میان دانشجویان، اساتید و مدیران برخوردار است. ارائه سرویس‌های ارزش افزوده برای سیاست‌گذاران و تصمیم‌گیرندگان حوزه علم و فناوری و امکان تجاری‌سازی دانش در جهت تثبیت نقش مرجعیت پژوهشگاه می‌تواند مفید باشد. استخراج تخصص پژوهشگران، مطالعه روند فعالیت علمی پژوهشگران، پژوهشگر در یک نگاه و ... نمونه‌ای از سرویس‌های ارزش افزوده با محوریت پژوهشگر است. در پایگاه داده ایرانداک، یک پژوهشگر مشخص به اشکال مختلف در پایگاه داده ثبت شده است. دلایلی مانند چند املائی بودن برخی نام‌ها، عدم دقت دانشجو، غلط‌های املائی و ... را می‌توان برای علت گوناگونی نام پژوهشگر نام برد. از طرف دیگر تا زمانی که زیرساخت یکتاسازی پژوهشگران ترمیم نگردد طراحی و پیاده‌سازی سامانه‌های ارزش افزوده در حوزه پژوهشگران با شکست روبرو خواهد شد. با توجه به تعداد بالای پژوهشگران، در این پژوهش، ابتدا با استفاده از روش‌های زبان‌شناسی رایانشی و مفاهیم کلان داده یک معماری برای استخراج پژوهشگران مشابه طراحی شده و در ادامه نسخه آزمایشی آن پیاده‌سازی شده است.

بررسی ساختار گفتمانی چکیده‌های پایان‌نامه‌های فارسی در برخی رشته‌های علوم انسانی، فنی‌مهندسی و علوم پایه

رشته‌های مختلف علمی در نگارش چکیده دارای تنوع گفتمانی هستند. شناسایی این ساختار گفتمانی برای سازماندهی اطلاعات در پایگاه‌های اطلاعاتی، تدوین استانداردهای چکیده‌نویسی و نیز سامانه‌های چکیده‌ساز خودکار متن از اهمیت زیادی برخوردار است. هدف از این پژوهش، تحلیل ساختار گفتمانی چکیده‌ی پایان‌نامه‌های فارسی در رشته‌های مختلف، تعیین حرکت‌های بلاغی آنها و به‌دست‌دادن تشابهات/تمایزات آنهاست. برای این منظور، سه شاخه‌ی علمی علوم انسانی، فنی‌مهندسی و علوم پایه تعیین و از هر شاخه نیز دو رشته‌ی نماینده برای تحلیل چکیده‌هایشان انتخاب شدند. رشته‌های نماینده عبارتند از: شیمی و فیزیک برای شاخه‌ی علوم پایه، مهندسی برق و مکانیک برای شاخه‌ی فنی‌مهندسی و فلسفه و زبان‌شناسی برای شاخه‌ی علوم انسانی. برای هر کدام از شش رشته، ۴۰ چکیده از پایگاه گنج ایرانداک جمع‌آوری و ساختار گفتمانی آنها در قالب حرکت‌های بلاغی و مدل پنج‌حرکتی هایلند کدگذاری شد. در مرحله‌ی بعد، حرکت‌های بلاغی هر کدام از رشته‌ها تحلیل و

الگوی کلان برای شاخه‌های علمی به دست آمد. نتایج پژوهش نشان داده الگوی غالب برای اکثر رشته‌ها الگوی I-P-M-Pr و P-M-Pr است و در این میان دو رشته‌ی فلسفه و شیمی تفاوت‌هایی با دیگر رشته‌ها دارند.

شناسایی هویت نویسنده بر اساس زبان فردی: پژوهشی پیکره‌ای در زبان فارسی

زبان فردی به شیوه منحصر به فرد هر گویشور در به کارگیری عناصر زبانی از جمله واژه‌ها، ساختار دستوری و آواها گفته می‌شود که اخیراً موضوع پژوهش‌های رایانشی و زبان‌شناختی در شناسایی نویسنده متون فاقد هویت واقع شده است. یکی از مولفه‌های زبان‌شناختی که گفته می‌شود تجلی‌گاه زبان فردی واقع می‌شود، واژه‌های دستوری در زبان است. واژه‌های دستوری از آن جهت که به‌طور ناخودآگاه در تولید زبان به کار گرفته می‌شوند، مستقل از موضوع متن به کار می‌روند و بسامد بالایی در متون کوتاه دارند، همواره مورد توجه پژوهشگران در حوزه شناسایی سبک نویسنده بوده‌اند. در این پژوهش امکان تفکیک متون متعلق به یک نویسنده از متون دیگر با استفاده از واژه‌های دستوری زبان فارسی بررسی شده است. نتایج به دست آمده نشان می‌دهد که واژه‌های دستوری در زبان فارسی قابلیت تفکیک متون متعلق به یک نویسنده را دارند و عملکرد واژه‌های تک‌نگاشتی بهتر از دونگاشتی و سه‌نگاشتی‌ها در متون کم‌حجم است. همچنین نتایج پژوهش نشان می‌دهد که حجم کمینه متن برای شناسایی موفقیت‌آمیز نویسنده در متون فارسی حدود ۴۰۰۰ واژه بر اساس ۲۰ واژه دستوری پربسامد است.

رفع ابهام از برجسب نحوی هم‌نگاره‌های اسمی و صفتی فارسی

بررسی هم‌نگاره‌ها در پیکره‌های متنی فارسی نشان می‌دهد اکثر این هم‌نگاره‌ها، در اثر یکسان بودن نمود نوشتاری تکواژ یا یای نکره، یای اسم‌ساز، شناسهٔ دوم شخص مفرد، یای صفت‌ساز و یای متصل به گروه اسمی ایجاد شده‌اند. در صورت رفع ابهام از برجسب نحوی این هم‌نگاره‌ها (هم‌نگاره‌های اسمی و صفتی مختوم به «-ی»)، رفع ابهام معنایی از کلمات با سهولت بیشتری انجام خواهد شد. در این پژوهش ابتدا فهرست مبسوطی از هم‌نگاره‌های اسمی و صفتی مختوم به «-ی» تهیه شده است؛ قواعد حساس به بافت نحوی جهت رفع ابهام از برجسب نحوی هم‌نگاره‌های مختوم به «ی» استخراج شده است. سپس جهت بررسی صحت قواعد مذکور، برنامهٔ ماشینی تهیه شده که صحت قواعد مستخرج از بررسی هم‌نگاره‌های مختوم به «ی» را بررسی می‌کند. بررسی ماشینی صحت قواعد نشان می‌دهد، صحت بیش از نیمی از قواعد بالای ۷۰٪ است.

بررسی ساخت‌واژی هم‌نگارهای اسمی و صفتی به منظور کمک به برچسب‌دهی «اسم» به کلیدواژه‌ها در پیکره‌های علمی

در این تحقیق ساخت‌واژه اسم‌ها و صفت‌ها در فارسی به تفصیل بررسی شده است. نظام نوشتاری زبان فارسی نیز مورد بررسی دقیق قرار گرفت تا از این رهگذر بتوان به شناسایی انواع هم‌نگارها در زبان فارسی پرداخت. سپس، انواع هم‌نگارها در زبان فارسی مورد مطالعه دقیق قرار گرفت و در نهایت از طریق جستجو به دو روش ماشینی و دستی، فهرست مبسوطی از هم‌نگارها از پیکره‌های «پیکره متنی زبان فارسی»، «پایگاه دادگان زبان فارسی» و «پیکره وابستگی نحوی زبان فارسی» تهیه شده است. بررسی کلی هم‌نگارها در پیکره‌های مورد مطالعه نشان می‌دهد که بیشتر هم‌نگارها، فراوانی بالایی در پیکره‌های متنی فارسی دارند و اکثر آنها در اثر یکسان بودن نمود نوشتاری تکواژ یاء نکره، یاء اسم‌ساز، شناسه دوم شخص مفرد، یاء صفت‌ساز و یاء متصل به گروه اسمی، ایجاد شده‌اند.

غنی‌سازی روابط معنایی هستان‌شناسی علوم پایه (شیمی)

هستان‌شناسی در واقع ابزار باز نمودن دانش در حوزه سازماندهی اطلاعات و هوش مصنوعی است و در دهه اخیر کاربرد آن در وب معنایی نیز مورد توجه قرار گرفته است. در تدوین هستان‌شناسی مفهومی، ایجاد یک ساختار سلسله‌مراتبی از مفاهیم، اولین ضرورت و ایجاد و تعیین نوع روابط بین مفاهیم در اولویت بعدی قرار دارد. در این پژوهش، به منظور توسعه روابط معنایی هستان‌شناسی علوم پایه، که بر مبنای اصطلاح‌نامه‌های تخصصی تدوین شده در پژوهشگاه علوم و فناوری اطلاعات ایران تولید شده است، روشی نیمه‌خودکار برای غنی‌سازی روابط معنایی در حوزه شیمی ارائه شده است. در روش، ابتدا روابط سلسله‌مراتبی موجود در هستان‌شناسی مورد بازبینی و پالایش قرار می‌گیرد تا انواع روابط سلسله‌مراتبی از یکدیگر متمایز شوند. در مرحله بعد مفاهیم موجود در هستان‌شناسی طبقه‌بندی و بر اساس روابط وابسته در اصطلاح‌نامه و الگوهای روابط معنایی استخراج شده از یک هستان‌شناسی سطح بالا، روابط معنایی بین مفاهیم تعیین شده و به هستان‌شناسی اضافه می‌گردد.

رویکردی پیکره‌ای و ساخت‌بنیاد به طبقه‌بندی معانی واژگانی در زبان فارسی

در این پژوهش با استفاده از رویکرد پیکره‌بنیاد، به بررسی افعال مرکب فارسی با جزء فعلی «خوردن» پرداخته شده است. در فعل مرکب، جزء فعلی نظیر «خوردن» که به آن فعل سبک نیز می‌گویند با عناصر غیر فعلی متعددی نظیر «ضربه»، «غذا»، «زمین» و غیره هم‌نشین می‌شود و این هم‌نشینی در افعال

مرکب مختلف سبب می شود که فعل سبک «خوردن» معانی مختلفی در کنار اجزاء غیرفعلی داشته باشد. آنچه مد نظر این پژوهش بوده، استفاده از روشی داده بنیاد برای طبقه بندی معانی مختلف فعل سبک «خوردن» بوده است. برای انجام پژوهش، پیکره ای ۵۰ میلیون کلمه ای، که به صورت تصادفی از پیکره همشهری نمونه گیری شده، مورد استفاده قرار گرفته است. با استفاده از نرم افزار تحلیل پیکره ای AntConc عناصر غیرفعلی فعل سبک «شدن» استخراج شد. سپس ۳۵ فعل به ترتیب بسامد پیکره انتخاب و برای هر فعل ۲۰ نمونه کاربردی از پیکره استخراج و با استفاده از روش تحلیل نمای رفتاری، مشخصه های ساخت واژ، نحوی و معنایی آن برچسب دهی شد. سپس جدول نمای رفتاری افعال با استفاده از تحلیل خوشه ای سلسه مراتبی در نرم افزار R تجزیه و تحلیل شد. خوشه های معنایی به دست آمده از تحلیل خوشه ای نشان می دهد که معانی بر اساس بافت کاربردی افعال به خوشه های متمایز تقسیم می شوند و افعال مرکبی که اجزاء غیر فعلی آنها به لحاظ حوزه معنایی مشترک هستند، در خوشه های مشابه و نزدیک به هم جای می گیرند.

تعیین قاعده وابسته های پسین گروه اسمی

منظور از پردازش زبان فراهم کردن تمهیداتی رایانه بنیاد است که بر اساس آن بتوان زبان را به شیوه ای ماشینی و در حد امکان بدون دخالت انسان فهم و درک کرد. در این پژوهش، ابتدا پیشینه بررسی گروه اسمی و مشخصا وابسته های پسین آن بررسی شده و سپس قواعد زبانی وابسته های پسین با ذکر مثال عرضه شده است. در نهایت این قواعد به صورت فرمول های قراردادی ارائه شده است.

ساخت پیکره گفتاری هجای CV زبان فارسی مبتنی بر الگوی نوایی گفتار برای استفاده در برنامه تبدیل متن به گفتار

در برنامه های تبدیل متن به گفتار معمولا پردازش نوایی با تغییر فرکانس پایه و دیرش واحدهای بازسازی صورت می گیرد، که گفتار حاصل تا حدی ویژگی های نوایی جمله را دارد اما با توجه به آزمون های ادراکی و پژوهش های جدیدی که در زمینه بازسازی واحدهای گفتاری صورت گرفته به نظر می رسد بهتر است واحدهای بازسازی از بافت های نوایی مختلف و حساس به ساختار نوایی انتخاب شده و سپس برای تولید گفتار هم گذاری شوند. از این رو، در این پژوهش، تهیه واحدهای بازسازی یعنی هجا بر این اساس مدنظر قرار گرفت و پیکره نوشتاری هجای CV فارسی بر این اساس تهیه شده است. با توجه به پیچیدگی ها و دشواری های موجود، هجاهای مورد نظر تنها از چهار بافت نوایی (۱) هجای بی تکیه واژگانی؛ (۲) هجای تکیه بر واژگانی، تکیه زیر و بمی غیرهسته؛ (۳) هجای تکیه بر

واژگانی، تکیه زیر و بمی هسته، نواخت مرزی افتان؛ ۴) هجای تکیه بر واژگانی، تکیه زیر و بمی هسته، نواخت مرزی خیزان استخراج شده و تولید هجاها در گفتار پیوسته و در درون جمله صورت گرفته است. ضبط و تهیه پیکره گفتاری هجاهای ساده و مرکب بر اساس سیگنال آکوستیکی و طیف نگاشت و با لحاظ معیارهای مشخص صورت گرفته و سپس هجاهای حاصل در پرونده‌های مختلف مبتنی بر جایگاه نوایی ذخیره شده‌اند. داده‌های حاصل می‌توانند برای استفاده در برنامه تبدیل متن به گفتار مورد استفاده قرار گیرند.

طراحی هستی‌شناسی علوم پایه و پایگاه اطلاعات مفهوم- بنیان آن

تعامل میان هستی‌شناسی و زبان‌های طبیعی مقدم بر فرایند پردازش زبان طبیعی و مهندسی دانش است. هستی‌شناسی‌ها که عمدتاً از الگوهای معناشناختی بهره می‌برند وجوه مشترکی با واژگان دارند. یک مجموعه واژگان اصطلاحات را بر اساس مفاهیم مرتبط سازماندهی می‌کند، در حالی که هستی‌شناسی‌ها مفهوم‌سازی کرده و روابط مفاهیم را تعیین می‌نمایند. یک واژگان مشترک پیش‌زمینه و ابزاری است در به اشتراک گذاشتن دانش جمعی و یک هستی‌شناسی مشترک علاوه بر این ابزاری است در به اشتراک گذاشتن دانش از طریق فناوری اطلاعات. در ساخت الگوهای مفهومی و ساختار زبانی، این حوزه زبان‌شناسی رایانه‌ای است که در ترسیم روابط میان مفاهیم و واژگان به منظور کاربرد در مهندسی دانش و سازماندهی اطلاعات نقش ایفا می‌کند. به منظور شکل‌گیری هستی‌شناسی مشترک به تلفیق و همسان‌سازی اصطلاح‌نامه‌های موجود در پژوهشگاه علوم و فناوری اطلاعات ایران پرداخته، مغایرت‌ها و هم‌پوشانی‌های واژگان علوم پایه را برطرف و شبکه مفاهیم این گرایش را با زیرمجموعه‌های آن به شکل هستی‌شناسی واژگانی علوم پایه طراحی نمودیم تا پس از تبدیل روابط اصطلاح‌نامه‌ای به مفهوم- بنیاد به زبان OWL^۲ و بر اساس استاندارد ۲۵۹۶۴ در پایگاه اطلاعات مفهوم- بنیاد به عنوان ابزاری قوی جهت ذخیره و بازیابی اطلاعات به کار گرفته شود.

شناسایی طرح‌های مرتبط با تحقق مترجم ماشینی در زبان فارسی

ترجمه ماشینی زیر شاخه‌ای از زبان‌شناسی محاسباتی است که عبارت است از ترجمه متنی از یک زبان طبیعی به زبانی دیگر توسط کامپیوتر. ترجمه ماشینی از دیرباز دغدغه بشر بوده و به مرور با گذشت سالیان تممادی توانسته به نقطه‌ای برسد که امروز شاهد آن هستیم. در این پژوهش، در ابتدا به بررسی پیشینه مساله پرداخته شده و پژوهش‌های انجام شده توسط دیگر متخصصان این حوزه و برخی نرم‌افزارهای موجود، معرفی شده است. در ادامه، به هدف اصلی این پژوهش پرداخته و طرح کلی

ساخت نرم افزار مترجم فارسی ارائه شده است.

بهبود سیستم بازشناسی اعداد تلفنی با استفاده از پارامترهای آکوستیکی آواهای زبان فارسی

در گفتار تلفنی عوامل متعددی سبب از بین رفتن تمایزات واجی و در نتیجه ابهام آوایی می شوند. اما این تمایزات توسط برخی از پارامترهای آکوستیکی باقیمانده در سیگنال گفتار، انتقال پیدا می کنند. تحلیل ۱۸۱ نمونه آوایی که توسط ۱۳ زن و ۱۴ مرد در بافت های مختلف آوایی تولید شده اند، نشان می دهد که پارامترهای نوفه سایش شامل دیرش سایش و مرکز ثقل طیف و نیز اطلاعات آوایی واکه شامل دیرش واکه، بسامد پایانه F۲ تمایز واجی [se]- [sel] را حتی در نمونه های کاهش یافته ای مانند [sef] در سیگنال گفتار انتقال می دهند. از آنجا که پارامترهای واکه، سرنخ های دقیق تر و بهتری برای انتقال تمایز واجی [se]- [sel] در سیگنال تلفنی هستند، سه پارامتر دیرش واکه، بسامد پایانه F۲ و شیب گذر F۲ در ۸۱۱ نمونه آوایی (چهار زن و نه مرد) مورد اندازه گیری قرار گرفت و به عنوان داده های تست برای افزایش دقت سیستم بازشناسی استفاده شد. برای افزایش دقت بازشناسی از روش تلفیق دو شیوه بازشناسی بر اساس مدل مخفی مارکوف و ضرایب مل کپستروم و نیز بازشناسی بر اساس پارامترهای استخراج شده از واکه استفاده شد. برای ترکیب نتایج حاصل از دو روش فوق، از معیار اطمینان استفاده شد. یکی از روش هایی که می توان برای سیستم مبتنی بر پارامتر، معیار اطمینان تعریف کرد، استفاده از درجه شباهت نرمال شده ای است که با استفاده از توزیع گوسی به دست می آید. دقت حدود یک درصد بیشتر از سیستم بازشناسی مبتنی بر HMM ۳۳/۹۳ درصد است و به ۳۱/۹۴ درصد می رسد. بنابراین می توان گفت روش ترکیبی مبتنی بر معیار اطمینان خطای بازشناسی ارقام سه و صفر به طور نسبی حدود ۱۵ درصد کاهش داده است.

طراحی مدل مفهومی یکپارچه نمایه سازی ماشینی منابع فارسی

در این پژوهش، در ابتدا به بررسی رابطه بین نمایه سازی و بازیابی اطلاعاتی (جست و جو)، مباحث نظری و سیر تحول نمایه سازی ماشینی از دیدگاه تئوریک، نمایه سازی ماشینی به صورت عملیاتی و از نگاه فرایندی، پروژه های موفق انجام شده در دنیا و برای زبان های مختلف و نیازمندی های نمایه ساز ماشینی در زبان فارسی پرداخته شده است. در ادامه، اجزای یک نمایه ساز زبان فارسی تشریح گردیده و مدل مفهومی نمایه ساز ماشینی به صورت دقیق ترسیم شده و روابط بین اجزاء، جزئیات زیرسیستم ها و مدل منطقی اولیه سیستم طراحی شده است. در انتها نیازمندی های توسعه نمایه ساز ماشینی برای پژوهشگاه علوم و فناوری اطلاعات ایران مورد بحث واقع شده و وضعیت پروژه هایی که در حال حاضر

انجام شده و یا تجربیاتی که وجود داشته و می‌تواند در اجرای این پروژه مفید باشد مورد توجه قرار گرفته است.

تعیین قاعده‌های صوری وابسته‌های پیشین گروه اسمی و کاربرد آن در ترجمه ماشینی

پردازش زبان یکی از حوزه‌های پژوهشی هوش مصنوعی بوده و ترجمه ماشینی نوعی پردازش زبان از زبان مبدا به زبان مقصد است. یکی از ملزومات ترجمه ماشینی استخراج قواعد صوری دستور زبان است که در این راستا باید برای تمامی حوزه‌های نحو زبان قواعدی صوری استخراج کرد. یکی از حوزه‌های نحو مربوط به گروه اسمی است. گروه اسمی مهمترین سازه در پردازش دستور زبان فارسی است. گروه اسمی شامل قواعد وابسته‌های پیشین و قواعد وابسته‌های پسین است. در این پژوهش به تعیین قاعده‌های صوری وابسته‌های پیشین گروه اسمی پرداخته شده و کاربرد آن در ترجمه ماشینی مورد بررسی قرار گرفته است.

فعل‌شکن فارسی

قواعد صوری دستور زبان یکی از ملزومات پردازش زبان است. پردازش نحوی دستور زبان یکی از حوزه‌های پردازش زبان است. برای پردازش نحوی باید قواعد صوری حوزه‌های مختلف دستور زبان را استخراج کرد. یکی از حوزه‌های نحو، حوزه افعال است. در این پژوهش، به استخراج قواعد برنامه‌ای رایانه‌ای که قادر به تشخیص و تجزیه فعل‌های ساده باشد، پرداخته شده است. برنامه خروجی که فعل‌شکن فارسی نامیده شده برنامه‌ای رایانه‌ای است که صیغگان فعل‌های زبان فارسی را تجزیه و شناسایی می‌کند و قادر به شناسایی و تجزیه ساختمان انواع فعل‌های ساده با قاعده و بی‌قاعده زبان فارسی است. در این طرح فعل‌های مرکب و پیشوندی بررسی نشده‌اند. زیرا هدف همکرد در فعل‌های مرکب و عنصر غیرفعلی در فعل‌های پیشوندی، آنچه باقی می‌ماند در واقع یک فعل ساده است.

عددشکن فارسی

زبان فارسی همچون بسیاری دیگر از زبان‌ها دارای قواعد ترکیب اعداد است. قواعد ترکیب اعداد بخشی از گروه اسمی در زبان فارسی‌اند. گروه اسمی دارای وابسته‌های پیشین و وابسته‌های پسینی است که قبل و بعد از هسته گروه اسمی قرار می‌گیرند. اعداد و صور ترکیبی آنها هم قبل از هسته گروه اسمی و هم بعد از آن می‌آیند. عددشکن فارسی برنامه‌ای رایانه‌ای است که صورت مکتوب اعداد و ترکیبات آن در زبان فارسی را به شکل عددی آن نمایش می‌دهد. این نرم‌افزار حافظه‌ای دارد که قواعد ترکیب

اعداد فارسی در آن تعبیه شده است. نرم افزار شکل مکتوب را می خواند و آن را با قواعدی که در حافظه دارد منطبق می کند و سپس آن را به صورت اعداد عرضه می کند.

واژه شکن فارسی

واژه شکن فارسی برنامه ای رایانه ای است که کلمات پیچیده را به اجزاء سازنده اش تجزیه و مقوله دستوری و معنایی آن را مشخص می کند. این نرم افزار حافظه ای دارد که کلمات بسیط و تکواژها و قواعد ساخت کلمات در آن قرار دارد. نرم افزار ابتدای کلمه ای را که به آن عرضه شده در حافظه جستجو و مقوله دستوری و معنایی آن را می یابد. چنانچه صورت کامل کلمه مورد نظر در حافظه وجود نداشته باشد نرم افزار در صدد تجزیه آن برآمده و در صورتی که اجزاء کلمه مفروض قاعده مند باشد، مقوله آن را مشخص می کند. چنانچه کلمه غیر دستوری باشد، نرم افزار تنها به تجزیه کلمه اکتفاء می کند.

طراحی الگوی نمایه استنادی نشریات فارسی ایران

نمایه استنادی مهمترین و اصلی ترین ابزار در تحلیل استنادی می باشد، از این طریق می توان ضمن آنکه امکان بازیاب اطلاعات را فراهم آورد، به ارزیابی کمی و کیفی تولیدات علمی، مطالعه تاریخ و ساختار علم، شناخت رابطه بین اسنادها و ترسیم ساختار موضوعی رشته های علمی دست یافت. این پژوهش به مطالعه و طراحی یک الگو جهت ایجاد نمایه استنادی نشریات فارسی ایران پرداخته است. نمایه استنادی یکی از بسترهایی است که می تواند به عرضه کتاب شناختی مقالات کلیدی و برجسته علمی دست بزند و علاوه بر آن ارزیابی پژوهش و پژوهشگران را در میان انبوه اطلاعات امکان پذیر سازد. در این پژوهش به مطالعه و بررسی تعاریف و مفاهیم نمایه استنادی، شاخص های انتخاب نشریات و سوابق سایر کشورها و ایران پرداخته شده است. علاوه بر آن ساختار پایگاه های استنادی موسسه اطلاعات علمی آمریکا مورد بررسی قرار گرفته و فرایند تولید و انجام کار در این نمایه های استناد ارائه گردیده است. نهایتاً بر اساس مجموعه اطلاعات جمع آوری شده، الگوی ایجاد یک نمایه استنادی برای نشریات فارسی در ایران ارائه شده است.

فرایند نمایه سازی

استفاده موثر از اطلاعات به سازماندهی آن نیاز دارد و سازماندهی به همراه فرایندهای متنوع بازنمایی محتوای مدرک برای بهبود جست و جو و بازیابی اطلاعات است. یکی از مهمترین شیوه های بازنمایی محتوا، نمایه سازی است که با تخصیص توصیفگرها به مدارک، محتوای موضوعی آنها نشان

داده می‌شود. نمایه‌سازی، یک مدرک را در میان انبوهی از منابع موجود در پایگاه‌ها و مخازن اطلاعات قابل بازیابی می‌سازد. نمایه‌سازی فرایندی پیچیده است و ارائه تعاریف، اصول، روش‌ها، ماهیت انواع نمایه‌ها، و روش‌های عملی نمایه‌سازی می‌تواند به استفاده بهینه از اطلاعات موجود کمکی شایان نماید. این پژوهش ابتدا به تعریف موجزی از اطلاعات، و سازماندهی اطلاعات و روش‌های آن پرداخته است. سپس بر نمایه‌سازی به عنوان یکی از روش‌های سازماندهی تاکید و تعاریف نمایه و نمایه‌سازی، انواع آن، ساختار نمایه، و مسائل مربوط به سیاست‌گذاری نمایه‌سازی را ارائه کرده است. در ادامه درباره نحوه ایجاد نمایه، ویژگی‌ها و فرایند نمایه‌سازی، و همچنین انواع زبان‌های نمایه‌سازی بحث شده است. در پایان ضمن ارائه مباحثی در ارزیابی نمایه‌سازی، نمایه‌سازی خودکار، و نقش نمایه در بازیابی اطلاعات پیشنهادهایی در زمینه بهبود نمایه‌ها مطرح شده است.

بهینه‌سازی اطلاعات علمی شیمی و مهندسی شیمی (سمینارها) در بانک جامع اطلاعات پژوهشگاه اطلاعات و مدارک علمی ایران و طراحی نرم‌افزار بانک ترکیبات شیمیایی

در این پژوهش اطلاعات علمی ۲۷۱۴ رکورد اطلاعاتی مربوط به سمینارهای شیمی و مهندسی شیمی که در فاصله زمانی سال‌های ۱۳۶۸ تا ۱۳۷۹ هجری شمسی در کشور برگزار شده‌اند به صورت یک مجموعه واحد و دارای نمایه‌نامه‌های تخصصی به صورت چاپی و نیز نرم‌افزاری ارائه شده است. کلیه اطلاعات مورد بازبینی، ویرایش و بازپردازش تخصصی قرار گرفتند. بخصوص کلیدواژه‌ها مورد استانداردسازی، یکسان‌سازی، ارجاع دهی و معادل‌یابی قرار گرفتند و در نهایت در یک مجلد به چاپ رسیدند، تا جهت تعیین اولویت‌های تحقیقاتی و جلوگیری از دوباره کاری توسط دانش‌پژوهان و محققان در بخش صنعت و دانشگاه مورد استفاده قرار گیرد. کنترل نامگذاری‌ها با توجه به استانداردهای ایوپاک انجام شده است. همچنین شیوه بازیابی اطلاعات شیمی افزون بر جستجوی موضوعی با استفاده از فرمول‌ها، اسامی ترکیبات شیمیایی و نیز شماره‌های ثبت شیمیایی عملی شده است. به منظور تقویت و اصلاح در ساختار تشکیلات نمایه‌سازی و چکیده‌نویسی در پژوهشگاه، دو نفر کارشناس شیمی مورد آموزش قرار گرفتند. طراحی نرم‌افزار تخصصی شیمی در ثبت، ذخیره و بازیابی اطلاعات ترکیبات شیمیایی، ایجاد نمایه‌نامه‌های تخصصی شیمی، طراحی کاربرگ‌های استاندارد ورود اطلاعات شیمی و استاندارد کردن شیوه‌های ذخیره و بازیابی اطلاعات شیمی از موارد مورد توجه در این پژوهش است. همچنین فهرست‌های تخصصی و واژه‌های یکدست شده در این طرح می‌توانند توسط دانشجویان، اساتید، محققان، نمایه‌سازان، مترجمان و فرهنگ‌نویسان مورد استفاده قرار گیرد.

بررسی اطلاعات علمی شیمی و مهندسی شیمی در بانک جامع اطلاعات مرکز به منظور بهینه‌سازی آنها

در این پژوهش، اطلاعات علمی ۱۳۰۰ پایان‌نامه مربوط به رشته‌های شیمی و مهندسی شیمی در مقطع کارشناسی ارشد و دکتری از سال ۱۳۴۱ تا سال ۱۳۷۵ مورد بازبینی، ویرایش و بازپردازش تخصصی قرار گرفته است. به خصوص کلیدواژه‌ها، مورد استاندارد کردن، یکسان‌سازی، ارجاع‌دهی و معادل‌یابی قرار گرفته و در نهایت در یک مجلد به چاپ رسیده، تا محققانی که در بخش پژوهش و صنعت شیمی فعال هستند جهت تعیین اولویت‌های تحقیقاتی و جلوگیری از دوباره‌کاری آن را مورد استفاده قرار دهند. کنترل نامگذاری‌ها با توجه به استانداردهای ایوپاک انجام شده است. همچنین شیوه بازیابی اطلاعات شیمی با توجه به فرمول‌ها و اسامی ترکیبات شیمیایی عملی شده است. ضرورت تغییر در ساختار تشکیلات نمایه‌سازی و چکیده‌نویسی در مرکز، نمایه‌سازی و چکیده‌نویسی مدارک شیمی توسط متخصصان موضوعی، استفاده از نرم‌افزارهای تخصصی شیمی در ثبت و ذخیره و بازیابی اطلاعات شیمی، ایجاد نمایه‌نامه‌های تخصصی شیمی، استاندارد کردن شیوه‌های ذخیره و بازیابی اطلاعات شیمی و نهایتاً ایجاد پایگاه تخصصی شیمی از مهمترین مطالب مورد توجه در این پژوهش است. همچنین فهرست‌های تخصصی و واژه‌های یک‌دست شده در این پژوهش می‌توانند بطور وسیعی توسط دانشجویان، اساتید، محققان، نمایه‌سازان، مترجمان و فرهنگ‌نویسان مورد استفاده قرار گیرد.

یکسان‌سازی شیوه رسم الخط اسامی ترکیبات شیمیایی در زبان فارسی

ضبط چندگانه صورت‌هایی واحد با خط فارسی باعث ایجاد مشکلاتی در کار تعلیم و تربیت شده و وقت بسیاری برای یکسان کردن شیوه نگارش متون از دست می‌رود. یک‌دست کردن خط از ابزار ضروری پردازش زبان نیز هست. یکی دیگر از پیامدهای همگون نبودن رسم الخط، ایجاد نابسامانی در کار اطلاع‌رسانی علوم است. ضبط اعلام و کلیدواژه‌ها به شکل‌های مختلف از مشکلات رایج در کتابخانه‌ها و مراکز اطلاع‌رسانی است. شیمی یکی از علوم است که اصطلاحات آن به گونه‌هایی متفاوت ثبت می‌شود و مشکلات سابق‌الذکر بیشتر در آن نمایان است. در این پژوهش سعی شده است که شیوه نگارش اسامی ترکیبات شیمیایی و بالاخص ترکیبات آلی آن در زبان فارسی و معضلات مربوط به آن بررسی و الگوهای کلی برای یکسان‌نویسی آنها پیشنهاد شود. این پیشنهادات منطبق با اصول و ضوابط رسم الخط پیشنهادی فرهنگستان زبان و ادب فارسی است.

پیکره‌های زبانی و واکاوی پیکره بنیاد زبان در زبان‌شناسی پیکره‌ای نقشی کلیدی دارند. ساخت پیکره‌ها کاری پیچیده، زمان‌بر، دارای گام‌های گوناگون، و میان‌رشته است. بنابراین برای کارایی و اثربخشی در ساخت آن‌ها نیاز است که پیش از آغاز کار، امکان‌سنجی انجام شود. 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. از آنجایی که مدلی برای امکان‌سنجی ساخت پیکره‌ی متنی همچون گونه‌ای از پیکره‌ها در پژوهش‌های پیشین نبود، این پژوهش در پی ساخت چنین مدلی انجام شد. برای این کار، پس از بررسی پژوهش‌های پیشین، شاخص‌های کلی ساخت پیکره‌های زبانی استخراج شد. سپس، فرآیند ساخت پیکره به تفصیل مورد بررسی قرار گرفت. 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 مؤلفه‌های امکان‌سنجی ساخت پیکره استخراج شد و مدلی کلی برای ساخت پیکره پیشنهاد شد. سپس با روش دلفی و در دو دور اعتباریابی، مدل پایانی به دست آمد. این مدل دارای هفت بُعد فنی، اقتصادی، زمان‌بندی، قانونی-حقوقی، عملیاتی، تأمین مالی، و بازاریابی و روی‌هم‌رفته ۳۳ مؤلفه است. این مدل و کاربرد آن در پروژه‌های ساخت پیکره‌ی متنی می‌تواند کمک بزرگی برای پژوهشگران و سازمان‌ها باشد. سپس در آخرین مرحله از پژوهش، مدل امکان‌سنجی ساخت پیکره در پنلی متشکل از متخصصان موضوعی و اجرایی در ایرانداک به شور گذاشته و اعتباریابی شد و مدل پایانی برای ساخت پیکره از داده‌های علمی ایرانداک به دست آمد و در پایان، امکان‌سنجی ساخت پیکره از داده‌های ایرانداک نیز انجام شد و چگونگی پرداختن به ابعاد و مؤلفه‌های مدل مشخص شد.

طراحی نقشه مفهومی برای اصطلاح‌نامه زبان‌شناسی

در زبان فارسی اصطلاح‌نامه دوزبانه فارسی-انگلیسی زبان‌شناسی موجود نیست. گام اول برای تدوین اصطلاح‌نامه زبان‌شناسی دوزبانه فارسی-انگلیسی، مشخص نمودن تقسیم‌بندی دقیق علم زبان‌شناسی به صورت نظری است. در این پروژه نقشه مفهومی فارسی برای اصطلاح‌نامه دو زبانه فارسی-انگلیسی زبان‌شناسی طراحی و تدوین شده است تا در پروژه‌ای دیگر بتوان بر مبنای آن، اصطلاح‌نامه زبان‌شناسی دو زبانه فارسی-انگلیسی را تدوین کرد. برای طراحی نقشه مفهومی یک حوزه علمی، معمولاً از روش قیاسی استفاده می‌شود و حوزه‌های مختلف علمی به صورت عام به خاص

و به صورت سلسله‌مراتبی مرتب‌سازی می‌شوند. این تقسیم‌بندی از یک طرف به اشراف موضوعی متخصص زبان‌شناس در حوزه‌های مختلف زبان‌شناسی وابسته است و از طرف دیگر نیازمند دانش و تجربه طراحی و تدوین اصطلاح‌نامه است. سطوح پیش‌روی واژگان در جهت عام به خاص در طول عملیات طراحی و با توجه به منابع مورد استفاده تعیین می‌شود و معمولاً در طراحی نقشه مفهومی حداکثر تا شش سطح پیش‌روی کافی است و حوزه‌های خردتر در صورت وجود، در عملیات تدوین اصطلاح‌نامه و با استفاده از نرم‌افزار تولید اصطلاح‌نامه قابل افزایش است. حدود ۱۰۰۰ واژه برای طراحی نقشه مفهومی زبان‌شناسی در نظر گرفته شده است. سرفصل دروس و کتب معتبر دانشگاهی در مقاطع تحصیلات تکمیلی زبان‌شناسی و منابع مرجع و فرهنگ‌های تخصصی زبان‌شناسی انگلیسی و فارسی، در گرایش‌های مختلف آواشناسی (Phonetics)، واج‌شناسی (Phonology)، صرف (Morphology)، نحو (Syntax) و معنی‌شناسی (Semantics) مورد استفاده خواهند بود. همچنین رویکردهای تاریخی علم زبان‌شناسی از منظر سنتی (Traditional)، ساختاری (Structural) و شناختی (Cognitive) مورد توجه خواهند بود. در این پروژه برای دیداری‌سازی نقشه‌های طراحی شده از نرم‌افزار visio استفاده شده است.

مفهوم‌شناسی و روش تدوین فرااصطلاح‌نامه

فرااصطلاح‌نامه یک پایگاه اطلاعات واژگانی جامع، بزرگ و گسترش‌یابنده چندمنظوره و چندزبانی حاصل از انطباق، اتصال، ادغام و تلفیق اصطلاح‌نامه‌ها و زبان‌های مقید موجود است که در اصل به قصد استفاده توسعه‌دهندگان آن تهیه و تدوین می‌شود. فرااصطلاح‌نامه همچنین می‌تواند اطلاعات موردنیاز برنامه‌های رایانه‌ای را برای ایجاد داده‌های استاندارد تأمین نموده، پرسش‌های کاربران را تفسیر و از طریق پالایش و جرح و تعدیل سؤالات با آن‌ها تعامل داشته و اصطلاحات کاربران را به واژگان مورد استفاده در منابع مرتبط برگرداند. فرااصطلاح‌نامه معمولاً در حوزه معینی از دانش شکل می‌گیرد. در ایران‌داک اصطلاح‌نامه‌های تخصصی در دهه ۱۳۸۰ تدوین شده است و در سال‌های ۹۷-۱۳۹۶ روزآمدسازی شده‌اند. ایران‌داک، علاوه بر اصطلاح‌نامه‌ها، مجموعه‌ای از واژه‌نامه‌ها و واژگان تخصصی را نیز تدوین و به چاپ رسانده است. برای جستجو و کاوش دقیق‌تر ضرورت دارد که این واژگان و اصطلاح‌نامه‌ها با یکدیگر تلفیق یابند تا از ریزش کاذب اطلاعات در جستجوی پایگاه گنج ممانعت به عمل آید و سرعت دستیابی هدفمند به اطلاعات تخصصی فراهم گردد. در این طرح به بررسی منابع گوناگون با موضوع فرااصطلاح‌نامه و شناسایی مفاهیم وابسته به آن، بیان کاربرد و نیز مراحل تدوین آن پرداخته شده است.

زبان فارسی نیز همانند دیگر زبان‌ها به گونه‌های اجتماعی متفاوتی تقسیم می‌شود. این گونه‌ها تابع متغیرهای سن و جنسیت و حرفه و جایگاه اجتماعی گویشوران در جامعه زبانی است. زبان علم یکی از گونه‌های اجتماعی زبان فارسی است که طبقات تحصیل کرده در مجامع و بافت‌های علمی آن را به کار می‌بندند. در این پژوهش، ویژگی‌های صوری و زبانی بخشی از مدارک علمی فارسی در رشته‌های زبان‌شناسی و زمین‌شناسی و شیمی استخراج شده است. هدف از انجام این پژوهش عرضه تصویری کلی از ویژگی‌های زبان علم فارسی بوده است. بدین معنی که مقولات و واژگانی نظیر اسم و صفت و قید و حروف اضافه و برخی ساخت‌های نحوی نظیر ساخت مجهول و زمان افعال از متون مورد بررسی استخراج و آمار آنها عرضه و محتوای آنها تحلیل شده است. از ویژگی‌های مشترک زبان علم فارسی استفاده از الفاظ فخیم فارسی و عربی در مقوله‌های اسم و صفت و قید و حروف اضافه و کاربرد فراوان ساخت مجهول و استفاده بیشتر از زمان افعال بوده است.

بررسی مفهوم و ابعاد خلاصه همگان‌فهم

ترویج علم، شرایطی را فراهم می‌سازد که گذشته از نخبگان و دانشمندان حوزه‌های مختلف علوم، سایر افراد جامعه نیز درک و بینشی از علم به دست آورند. یکی از ابزارهای ترویج علم، تنظیم و انتشار خلاصه همگان‌فهم است. خلاصه همگان‌فهم، خلاصه کوتاهی از پیشنهادیه یا طرح پژوهشی است که غیر از پژوهشگران و متخصصان همان حوزه، برای عموم به نگارش درآمده است. هدف این پژوهش تشریح مفهوم خلاصه همگان‌فهم، شباهت‌ها و تفاوت‌های مفاهیم همانند با آن، سازمان‌های فعال، الگوهای پیشنهادی، زمینه‌های موضوعی، و اشکال به کارگیری آن در تولیدات علمی بوده و شباهت‌ها و تفاوت‌های مفاهیم مورد بررسی شامل «چکیده»، «خلاصه»، «خلاصه اجرایی» و «خلاصه همگان‌فهم» از چهار جنبه «مخاطب یا مخاطبان»، «ساختار»، «هدف» و «کاربرد در نوع مدرک»، به صورت مقایسه‌ای ارائه شده است. برای پاسخگویی به پرسش‌های پژوهش، پژوهشگران، اسناد در پیوند با خلاصه همگان‌فهم را جست‌وجو، بازیابی، مطالعه، و از منظر تخصصی تجزیه و تحلیل کرده‌اند. با توجه به نتایج، ۱۷ مؤلفه اساسی و مهم که در تدوین خلاصه همگان‌فهم نقش داشتند، استخراج و بر اساس شاخص‌های استخراجی از الگوهای ارائه شده برای تدوین این نوع خلاصه‌ها، در سه بُعد دسته‌بندی شدند. این ابعاد شامل «شکلی و صوری»، «محتوایی» و «ارزیابی» است. نتایج بررسی‌های صورت گرفته نشان می‌دهد که مقالات علمی پژوهشی، طرح‌های پژوهشی پایان‌یافته، پیشنهادیه‌های پژوهشی، پایان‌نامه‌ها و

رساله‌های دانشگاهی از جمله منابع و پزشکی، سلامت و مراقبت‌های بهداشتی از جمله موضوعاتی هستند که به منظور تهیه و انتشار خلاصه‌های همگان فهم در نظر گرفته شده‌اند.

طراحی روشی هوشمند برای دسته‌بندی نوشتارهای علمی فارسی

با توجه به حجم بالای متون علمی و سرعت پایین کنترل اقلام اطلاعاتی در سامانه ثبت و در نتیجه عدم نمایش به‌روز در سامانه گنج، افزایش سرعت نمایه‌سازی را می‌طلبند. نمایه‌سازی نیازمند کنترل و تصحیح اقلام اطلاعاتی ثبت شده است. موضوع اصلی شامل علوم انسانی، فنی و مهندسی، علوم پایه، هنر و معماری، علوم پزشکی و غیره یکی از فراداده‌هایی است که در ثبت پایان‌نامه‌ها و رساله‌ها در پایگاه داده ثبت پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) دریافت می‌شود. استخراج هوشمند حوزه پارساها به وسیله الگوریتم‌های یادگیری ماشین افزایش سرعت فرایند نمایه‌سازی و افزایش کیفیت داده‌ها را در پی خواهد داشت. با توجه به این مهم، در این طرح پژوهشی مساله دسته‌بندی هوشمند محتوایی متون به یکی از پنج دسته موضوع اصلی، یعنی علوم انسانی، فنی و مهندسی، علوم پایه، هنر و معماری، و علوم پزشکی، در نظر گرفته شده است. این طرح، با خودکارسازی فرایند تایید یا عدم تایید موضوع اصلی ثبت شده از طرف پژوهشگر، افزایش سرعت نمایه‌سازی را در پی خواهد داشت. در این پژوهش، ابتدا پیش‌پردازش‌های معمول در پردازش زبان طبیعی انجام شده و ویژگی‌های متمایزکننده استخراج شده است. در ادامه، با استفاده از دسته‌بندی متنوع، محتوای پایان‌نامه‌ها به پنج دسته آموزش داده شد. نتایج ارزیابی‌ها بر روی متون پایان‌نامه و رساله‌های فارسی نشان می‌دهد روش پیشنهادی از دقت و سرعت خیلی خوبی برای تشخیص موضوع اصلی پارساها برخوردار است.

بررسی موتور جستجوی گنج به منظور بهبود بازیابی اطلاعات در زبان فارسی

بازیابی اطلاعات به معنای یافتن محتوایی با طبیعت غیرساخت یافته (معمولاً متن) از مجموعه‌هایی به منظور برآورده ساختن نیازهای اطلاعاتی کاربران است. با بهره‌گیری از روش‌های بازیابی اطلاعات موجود در ادبیات پلتفرم‌های متن‌باز متعددی طراحی شده‌اند. کتابخانه آپاچی لوسن و موتور جست‌وجوی آپاچی سولر از این نوع پلتفرم‌ها هستند. سامانه گنج نیز به منظور جست‌وجو در اسناد و مدارک ذخیره شده در پایگاه داده خود، از این کتابخانه و موتور جست‌وجو به عنوان هسته اصلی فرآیند بازیابی اطلاعات استفاده می‌کند. علی‌رغم کاربردهای گسترده، سولر دارای کاستی‌هایی بوده، تاکنون نیز هیچ‌گونه پژوهشی در زمینه شناسایی و واکاوی این کتابخانه و موتور جست‌وجو در پژوهشگاه انجام نشده و تنها عملکرد رابط کاربری آن و مشکلات موجود بررسی شده است. لذا علی‌رغم ضرورت

بررسی کارایی رابط کاربری گنج، تا زمانی که هسته اصلی آن (خصوصاً با تمرکز بر زبان فارسی) بررسی و اصلاح نشود، نمی‌توان انتظار داشت عملکرد سامانه مطلوب باشد. لذا ضروری است تا با بررسی محدودیت‌ها و مشکلات موجود در روش‌های مورد استفاده در موتور جست‌وجوی آپاچی سولر (از منظر بازیابی اطلاعات) دیدگاهی کلان نسبت به آن در پژوهشگاه ایجاد شده تا از این طریق بتوان پیشنهادهایی به منظور ارتقا عملکرد هسته اصلی فرآیند بازیابی اطلاعات سامانه گنج ارائه داد. بر این اساس، در این طرح دانش فنی درمورد این موتور جست‌وجو و روش‌های مورد استفاده آن کسب شده و راهکارهایی به منظور ارتقای نتایج بازیابی در زبان فارسی و در سامانه گنج ارائه شد. راهکارهای پیشنهاد شده در قالب دو دسته راهکارهای نیازمند پژوهش و راهکارهای فنی و عملیاتی دسته‌بندی شدند.

تاریخچه آواشناسی و سهم ایرانیان

آواشناسی عمدتاً در چارچوب مطالعه زبان بررسی شده است. به این سبب دستاوردهای آواشناسی در این کتاب از میان آثاری که درباره مسائل کلی زبان است، استخراج و با تسلسلی تاریخی عرضه شده است. محدوده زمانی آثار از حدود شش قرن قبل از میلاد تا قرن پانزدهم میلادی است. نویسنده سعی کرده داور عادل باشد و سهم واقعی دانشمندان ایرانی را در آواشناسی مشخص کند، نه آنکه دانشمندان هموطنش را از سر احساس در بهترین جای یک قصه علمی تخیلی قرار دهد. در این کتاب آثار و بقایای آواشناسی از خرابه‌های هند و یونان و روم و سرزمین‌های اسلامی جمع شده و در بافت اجتماعی دیاری که دانشمندان در آن به بررسی این علم پرداخته‌اند، عرضه شده است.

آواشناسی آکوستیک و شنیداری

این کتاب مقدمه‌ای کوتاه و غیرفنی (مناسب به عنوان یک منبع مکمل برای درس آواشناسی عمومی یا علم گفتار) در چهار مبحث مهم آواشناسی آکوستیک شامل: (۱) ویژگی‌های آکوستیکی طبقات عمده آواهای گفتاری؛ (۲) نظریه آکوستیکی تولید گفتار؛ (۳) بازنمایی شنیداری گفتار؛ و (۴) این کتاب برای دانشجویان در درس‌های مقدماتی آواشناسی زبان‌شناسی، علوم گفتاری و شنیداری و شاخه‌هایی از مهندسی برق و روان‌شناسی شناختی مورد استفاده خواهد بود. پنج فصل اول به معرفی مبانی آکوستیک، نظریه آکوستیکی تولید گفتار، پردازش دیجیتالی سیگنال، شنوایی و درک گفتار می‌پردازد و در چهار فصل باقیمانده، طبقات عمده آوایی مورد مطالعه قرار می‌گیرد و مشخصه‌های آکوستیکی آواهای گفتاری طبق پیش‌بینی نظریه آکوستیکی تولید گفتار، ویژگی‌های شنیداری و مشخصه‌های درکی آنها مرور می‌شود. در پایان هر فصل، فهرستی از منابع پیشنهادی برای مطالعه بیشتر به همراه تعدادی تمرین آمده است. تمرین‌ها با تأکید بر اصطلاحاتی که در هر فصل با حروف سیاه مشخص گردیده (و در قسمت «واژگان تخصصی» فهرست شده‌اند)، خواننده را به به کار بردن مفاهیم مطرح شده در فصل ترغیب می‌نمایند.

دستور زبان فارسی از دیدگاه رده‌شناسی

این کتاب نخستین دستور زبان فارسی است که از دیدگاه رده‌شناسی زبان که یکی از گرایش‌های عمده در زبان‌شناسی امروز است، نوشته شده است. اصل کتاب جزو مجموعه‌ای است که به قصد فراهم

آوردن داده‌هایی واحد و دارای نظام برای محققان زبان‌شناسی نظری و پژوهشگران تدوین شده و به این هدف تاکنون شصت زبان بررسی و برای آنها دستور نوشته شده است. این دستور شامل سه بخش صرف، نحو و واج‌شناسی است و در مباحث آن مقایسه‌هایی نیز بین ویژگی‌های دستوری فارسی و دیگر زبان‌ها انجام شده است. کتاب، دستوری توصیفی است و در چارچوب بررسی «هم‌زمانی»، زبان محاوره‌ای فارسی تهرانی افراد تحصیل کرده را انتخاب و آن را توصیف کرده است. نکته‌های بدیع، ساختار، هدف خاص و مباحث جدید کتاب بر جامعیت و سودمندی آن افزوده است.

ساختواژه، اصطلاح‌شناسی و مهندسی دانش

مهندسی دانش، به منظور سازماندهی اطلاعات علمی، به برنامه‌ریزی رایانه‌ای و ایجاد شبکه اطلاعاتی می‌پردازد. ساماندهی نام‌گذاری مفاهیم علمی، یکی از مهم‌ترین مباحث علمی اصطلاح‌شناسی محسوب می‌گردد که در آن، «مفهوم» به عنوان واحد دانش فنی، تعریف شده است. به منظور ترسیم روابط مفهومی یک رشته تخصصی باید نظام مفاهیم زیررشته‌های مرتبط با آن را، در یک ساختار کلی تلفیق نمود. بر این اساس، اصطلاح‌نامه‌های تخصصی تدوین می‌شوند و در نهایت با یکدیگر ادغام می‌گردند. فرااصطلاح‌نامه نوعی اصطلاح‌نامه کلان است که موجب تلفیق اصطلاح‌نامه‌های متعدد موجود در سطح ملی و بین‌المللی می‌گردد و در بازایی، ارجاعات متقابل میان پایگاه‌های اطلاعات علمی گوناگون برقرار می‌نماید. در اثر حاضر با استفاده از بانک اطلاعات واژگان (تلفیق واژگان علمی اصطلاح‌نامه‌های علوم پایه) پژوهشگاه علوم و فناوری اطلاعات ایران که به عنوان پیکره زبانی انتخاب گردیده است و با توجه به نظریه ساختوازی و معناشناسی واژگانی (لیبر ۲۰۰۴) این واژگان مورد تجزیه و تحلیل آماری قرار گرفته‌اند و علاوه بر تعیین فراوانی هر یک از عناصر واژگان مرکب، مشتق و مرکب مشتق، الگوهای ساختوازی اصطلاحات علمی استخراج شده است. سپس، با تعیین روابط معنایی آنها، زمینه طراحی هستی‌شناسی و شبکه واژگانی این حوزه علمی برای برنامه‌ریزی آموزشی و پژوهشی و در نهایت، طراحی نظام مفاهیم شبکه علمی کشور فراهم گردیده است.

سامانه استانداردساز و خطایاب متون علمی فارسی

روزانه هزاران مستند متنی متنوع در حوزه‌های مختلف علمی بر روی وب جهان‌گستر قرار می‌گیرد. این مستندات می‌توانند شامل پایان‌نامه‌ها، مقاله‌ها، گزارش‌های علمی و مواردی از این قبیل باشد. نگارش متن این مستندات علمی جهت حفظ یکنواختی باید بر اساس اصول ثابت انجام گیرد، اما همواره به طور غیر عمدی دستخوش سلیقه‌های مختلفی در طول تاریخ می‌شود. اگرچه این تغییرات

ناشی از پویا بودن زبان و خلاقیت ذهنی بشر است، اما این پویایی و خلاقیت پردازش ماشینی متن را با چالش‌های متعددی روبرو می‌کند و دقت پردازش داده‌ها را به میزان چشمگیری پایین می‌آورد. علاوه بر تنوع نگارشی، غلط‌های سهوی املائی نیز وجود دارد که فحوای گفتمانی متن را منحرف کرده و درک آن را با مشکل مواجه می‌کند. بنابراین کلیه نویسه‌های متن باید به حالت استاندارد تبدیل شوند و عاری از هرگونه خطاهای املائی گردند. پژوهشگران طرح حاضر سامانه‌ای برای استانداردسازی و خطایابی متون علمی فارسی طراحی کرده‌اند که این سامانه متون نوشتاری علمی و تخصصی فارسی را به لحاظ صحت نگارشی و املائی بررسی می‌کند و متن را به شکل استاندارد در می‌آورد. این سامانه با معیار اف ۹۰,۰۰۸ و معیار بلوی ۹۱,۶۶ به استانداردسازی می‌پردازد و با معیار اف ۹۲ درصد و با رتبه‌بندی ۸۲ درصد به اصلاح خطای املائی می‌پردازد.

اصطلاح‌شناسی و واژه‌سازی در زبان آلمانی

زبان مهم‌ترین و کامل‌ترین وسیله نه تنها برای ایجاد ارتباط همزمان میان افراد، بلکه به منظور ارتباط بین نسل‌ها و فرهنگ‌های مختلف در گذشته، حال و آینده نیز به کار می‌رود. بنابراین نقش عمده زبان، ایجاد ارتباط و مراوده اجتماعی، چه در سطوح تاریخی و چه در سطوح جغرافیایی است، و انسان برای انتقال تجربیات و اندیشه‌های خود به دیگران به ابزاری که همان زبان است، نیازمند است. چنانچه رشد ملت‌های گوناگون را با گستردگی گنجینه لغات زبان آن‌ها مقایسه کنیم، درمی‌یابیم که بین نهادهای اجتماعی زبان و فرهنگ رابطه‌ای مستقیم وجود دارد. توسعه و گسترش افکار و اطلاعات مستلزم رشد متناسب گنجینه لغات است، و همچنین تاثیر زبان و تفکر را بر یکدیگر می‌رساند.

ایرانداک برای گسترش گفتمان اصطلاح‌شناسی و زبان‌شناسی رایانشی، در میزگردها و سخنرانی‌های علمی به زمینه‌های گوناگون این حوزه می‌پردازد و کارشناسان و استادان را برای گفت‌وگوهای انتقادی در این زمینه گرد می‌آورد و دستاوردهای آنها را در اختیار همگان می‌گذارد.

سخنرانی علمی: فعالیت‌های بین‌المللی اصطلاح‌شناسی و کاربرد آن در بازنمایی دانش

در این نشست که با حضور رئیس اینفوترم (International Information Centre for Terminology) که نهاد بین‌المللی فعالی در حوزه اصطلاح‌شناسی است برگزار شد، در ابتدا فعالیت‌های اصطلاح‌شناسی و اصطلاح‌نامه‌های علوم پایه تدوین یافته در گروه اصطلاح‌شناسی پژوهشگاه علوم و فناوری اطلاعات و آنتولوژی حاصل از آنها توسط اساتید ایرانداک معرفی شد. در ادامه، رئیس اینفوترم به تشریح نقش اصطلاح‌شناسی در سازماندهی و مدیریت دانش نتایج فعالیت‌های تحقیقاتی، نقش کلیدواژه‌ها و اصطلاحات در بازنمایی دانش و انتقال آن به متخصصین و پژوهشگران، کاربرد زبان و برنامه‌ریزی اصطلاح‌شناسی در سطح ملی و بین‌المللی، استانداردها، اصول و روش‌های اصطلاح‌شناسی، معرفی سازمان اینفوترم و نقش آن در هماهنگی فعالیت‌های اصطلاح‌شناسی در سطح بین‌الملل پرداختند.



در این سخنرانی، ابتدا درباره سه معنای اصطلاح‌شناسی (ترمینولوژی) صحبت شد. این سه معنا شامل: ۱- اصطلاح‌شناسی به معنی مجموعه اصطلاحات علمی، ۲- اصطلاح‌شناسی به معنی فعالیت واژه‌گزینی و معادل‌یابی و ۳- اصطلاح‌شناسی به معنی رشته‌ای علمی. سپس، برخی از مکاتب اصطلاح‌شناسی مانند مکتب وین، مکتب مسکو، مکتب پراگ، مکتب کبک، مکتب شناختی و مکتب اجتماعی معرفی شدند و در ادامه به تأثیرپذیری مکاتب اصطلاح‌شناسی از مکاتب زبان‌شناسی پرداخته شد. در نهایت به تفصیل درباره مکتب وین و تأثیرپذیری آن از برخی از مکاتب زبان‌شناسی توضیح داده شد و در این راستا به تقابل مکتب زبان‌شناسی با مکتب وین پرداخته شد.



کاربران به صورت مداوم پرسش جست‌وجوی قبلی خود را به امید بازیابی بهتر اصلاح می‌کنند. این اصلاحات فرمول‌بندی مجدد یا اصلاح پرس‌وجو نامیده می‌شود. بیتس (۱۹۹۰) اشاره می‌کند، گاهی کاربر ترجیح می‌دهد که فرایند جست‌وجو را تا حدی، مطابق با استراتژی‌ها و تاکتیک‌های مشخصی که بتواند به طور موثر در رابط کاربر حمایت شود، کنترل کند. بیتس تاکتیک‌ها را به عنوان «یک یا تعداد انگشت‌شماری از حرکات که جست‌وجو را پیش می‌برند» تعریف می‌کند و به عنوان مثال، آنها را شامل تغییر یک اصطلاح به اصطلاحات خاص‌تر یا اضافه کردن اصطلاحات مشابه می‌داند. در این میان، هستی‌شناسی‌ها می‌توانند برای حمایت مستقیم تاکتیک‌ها استفاده شوند. (García and Sicilia)

۲۰۰۳). در این سخنرانی، به تاکتیک‌هایی که از سوی هستی‌شناسی‌ها حمایت می‌شوند پرداخته شده و وضعیت کاربران گنج در استفاده از این تاکتیک‌ها به تصویر کشیده می‌شود.



سخنرانی علمی: رتبه‌بندی حوزه‌های پژوهشی زبان‌شناسی رایانشی با روش تحلیل سلسله‌مراتبی (AHP)

در این سخنرانی به مساله تعیین اولویت‌های پژوهشی گروه زبان‌شناسی رایانشی ایرانداک پرداخته شده و نتایج به دست آمده با استفاده از روش تحلیل سلسله‌مراتبی (AHP) برای این مساله ارائه شد. با توجه به مطالب ذکر شده، روش تحلیل سلسله‌مراتبی دارای دو سطح «گزینه‌ها و معیارها» است که در مساله موردنظر، گزینه‌های پژوهش حوزه‌های زبان‌شناسی رایانشی بوده که برای تعیین آنها از روش اصطلاح‌نامه‌ای استفاده شده، و معیارهای پژوهش بر اساس اسناد بالادستی پژوهشگاه و با نظر خبرگان بدست آمده است. علاوه‌براین، مشخص گردید که اولویت‌بندی معیارها و گزینه‌ها توسط خبرگان زبان‌شناسی صورت گرفته و در نهایت با تعیین وزن معیارها و گزینه‌ها، اولویت‌های پژوهشی بر اساس نظر خبرگان به دست آمده است.



سخنرانی علمی: رابطه خط و شخصیت

در این سخنرانی به رابطه بین دست خط افراد و شخصیت آنان پرداخته خواهد شد. از روی ویژگی‌های خط هر فرد و نحوه آرایش نوشتار و شیوه استفاده از قلم می‌توان به برخی ویژگی‌های درونی افراد پی برد. شناسایی خلق و خوی نویسنده از روی دست خط به خصوصیات بومی خط نیز مربوط می‌شود.



سخنرانی علمی: از خواب‌نامه تا لغت‌نامه (ملاحظات در باب انتخاب واژگان بیگانه)

مشابهت غربی میان فرهنگ لغت و خواب‌نامه وجود دارد: هر لغت هم در میان لغت‌نامه وجود دارد و هم در خواب‌نامه. اما برای لغت‌های مدرن و نوظهور معادلی وجود ندارد. این نکته به خودی خود دیرهمضمی این لغات را در انبانه ناخودآگاه ذهن ایرانی نشان می‌دهد. هدف از این سخنرانی توجه‌دادن مخاطبان به رابطه خواب به عنوان تصویر (رویا=رویت) و کلمه و پرداختن به اهمیت انتخاب لغات جدید و تأثیر عمیقی است که می‌توانند بر ناخودآگاه بگذارند.



در ابتدا ایجاد طرحواره مدیریت دانش در پژوهشگاه علوم و فناوری اطلاعات مطرح شد؛ طبق این طرحواره اسناد و مدارک به صورت دیداری و شنیداری، کتاب‌ها، پایان‌نامه‌ها و گزارش‌های دولتی جمع‌آوری، سپس نمایه و وارد پایگاه‌های اطلاعاتی می‌شوند و نتایج تحلیل این اطلاعات علمی-فنی در اختیار سیاستگذاران کشور قرار می‌گیرند. مأموریت اصلی مرکز اطلاعات و مدارک علمی ایران در سالهای ۱۳۷۰ به عنوان یک مرکز اصلی و مادر در کشور، پشتیبانی مراکز تحقیقاتی و دانشگاهی و مراکز اجرایی و سیاست‌گذاری در زمینه اطلاعات علمی بود که این سازماندهی و مدیریت اطلاعات بدون کاربرد زبان‌های تخصصی و اصطلاح‌شناسی ممکن نبود. در این بخش اصطلاح‌شناسی و ارتباط آن با نشانه‌شناسی نمادهای کلامی و غیرکلامی، داده‌پردازی و انفورماتیک و مستندسازی، اطلاع‌رسانی و نظریه‌های علوم شناختی در این حوزه مطرح می‌شد. بر این اساس اصطلاح‌نامه به عنوان ابزار سازماندهی و بازیابی اطلاعات ایجاد گردید. در این سخنرانی فرایند شکل‌گیری اصطلاح‌نامه جامع، مراحل تشکیل هر اصطلاح‌نامه و فرایند تولید آن، مراحل ایجاد هستی‌شناسی‌ها با استفاده از اصطلاح‌نامه‌ها و مراحل ساخت ابزار مناسب برای انتقال اطلاعات یکسان شده از واژگان علوم پایه به نرم‌افزار آنتولوژی مورد بحث قرار گرفته است.

با توجه به افزایش حجم اسناد و نیاز به پردازش ماشینی، عدم کارایی سیستم‌های پردازشی مبتنی بر نحو و کلیدواژه، نیاز روزافزون به سطوح بالاتر پردازش زبان (پردازش‌های معنایی)، نیاز به تهیه منابع زبانی و حرکت به سمت رویکردهای معنایی، ایجاد شبکه‌های واژگانی الزامی غیر قابل انکار است. در این سخنرانی، «واژه‌ستان‌شناسی» به عنوان یکی از گلوگاه‌ها در مسیر پردازش زبان فارسی بیان شده و برخی از کاربردهای آن مورد بررسی قرار گرفت. سپس، از «فارس‌نت» به عنوان پروژه‌ای عظیم برای ساخت واژه‌ستان‌شناسی در زبان فارسی نام برده شد که شبکه واژگانی (وردنت) فارسی، افزودن اطلاعات معنایی تکمیلی مانند قاب افعال، ساختارهای آرگومانی (نقش‌های موضوعی) و محدودیت‌های گزینشی و تبدیل واژگان معنایی به واژه‌ستان‌شناسی با افزودن روابط و ویژگی‌های متنوع‌تر، فازهای اول تا سوم آن را تشکیل می‌دهند. در ادامه، عناصر پایه‌ای وردنت فارسی معرفی شده و ساختار و ویژگی‌ها، گستره تحت پوشش، آمار کنونی دادگان، مفاهیم بنیادی، روابط میان مقوله‌ای و

درون مقوله‌ای، روابط مشترک میان مقوله‌های نحوی، روابط خاص هر مقوله، روش‌های ایجاد شبکه‌های واژگانی، نحوه ساخت وردنت فارسی و ویژگی‌های آن، ابزارهای ساخت نیمه خودکار، نگاشت نیمه خودکار وردنت فارسی به انگلیسی، مشکلات نگاشت، و اشاره به برخی الگوها مانند الگوی هیرست و سایر الگوها مورد بحث واقع شد. در خاتمه این سخنرانی، به فعالیت‌های آتی در مسیر تکمیل فارس‌نت پرداخته شد.

سخنرانی علمی: ترجمه ماشینی مبتنی بر پیکره

یکی از کاربردهای مهم پیکره، استفاده از پیکره در ترجمه ماشینی است که نوعاً مبتنی بر استفاده از پیکره‌های موازی است. هدف از برگزاری این سخنرانی، معرفی نقش پیکره زبانی در توسعه و بهبود سامانه‌های ترجمه خودکار/ماشینی بود. در این راستا، ابتدا به تعریف مفهوم پیکره و ذکر برخی از انواع پیکره پرداخته شد. سپس، تاریخچه کوتاهی از صنعت ترجمه ارائه گردید. در ادامه به طور مفصل درباره معماری ترجمه و رویکردها و مدل‌های مطرح در آن مورد بررسی قرار گرفت و در نهایت، آخرین نوآوری‌ها و ویژگی‌های آن‌ها در ترجمه معرفی شد. نتایج بررسی انواع رویکردها با تکیه بر تاریخچه به کارگیری آن‌ها، نشان می‌دهد، مدل شبکه عصبی مبتنی بر پیکره، در مقایسه با سایر مدل‌ها (مدل مبتنی بر قاعده و مدل آماری مبتنی بر پیکره) تقریباً در همه جفت زبان‌هایی که بررسی شده است، عملکرد بهتری داشته است و خطاهای کمتری دیده شده است.



در راستای مشخص نمودن وجه تمایز و تشابه «فرهنگ‌نگاری» با «اصطلاح‌نامه‌نگاری» در این رویداد، پس از بررسی پیکره‌های مهم‌ترین فرهنگ‌های فارسی به فارسی، از کهن‌ترین آن‌ها تا امروز، به معرفی طرح تدوین فرهنگ جامع زبان فارسی که زیر نظر دکتر علی‌اشرف صادقی در فرهنگستان زبان و ادب فارسی در دست اجراست، پرداخته شد و به بررسی مهم‌ترین رکن این فرهنگ یعنی پیکره رایانه‌ای آن، متشکل از شواهد به‌کارگیری واحدهای واژگانی در متون زبان فارسی پرداخته و ویژگی‌های برجسته آن شرح داده شد. همچنین در این جستار، به برخی از مهم‌ترین پیکره‌های زبانی رایانه‌ای که در سال‌های اخیر در ایران فراهم آمده نیز به اختصار اشاره گردید. سپس در سخنرانی دوم و سوم به تعریف اصطلاح‌نامه و روش تدوین و کاربردهای آن، پرداخته شد. یکی از جدیدترین کاربردهای اصطلاح‌نامه در امر آموزش است. در این راستا درباره کاربرد اصطلاح‌نامه در آموزش و تاثیر آن بر افزایش یادگیری فراگیران توضیح داده شد. در پایان، به تعریف مفهوم «فرااصطلاح‌نامه» و تفاوت آن با اصطلاح‌نامه پرداخته شد. سپس ویژگی‌ها و کاربرد فرااصطلاح‌نامه مطرح و در نهایت، نقش آن در ساخت آنتولوژی بررسی گردید.



سخنرانی علمی: نقش پیکره‌های زبانی در تحلیل گفتمان (تحلیل گفتمان پیکره‌یار)

هدف از برگزاری این رویداد معرفی مفهوم «پیکره»، «زبان‌شناسی پیکره‌ای»، کاربردهای پیکره با

تمرکز روی یکی از کاربردهای مهم پیکره (تحلیل گفتمان) بود. بر این پایه، در این سخنرانی پس از تعریف «تحلیل گفتمان پیکره‌یار» ابتدا به بررسی وجه تمایز این حوزه را با پژوهش‌های پیکره‌بنیاد و پژوهش‌های برگرفته از پیکره، پرداخته شد. سپس، روش کار در تحلیل گفتمان پیکره‌یار توضیح داده شد. در ادامه یکی از نرم‌افزارهای پرکاربرد در تحلیل واژگانی متن معرفی شد. انواع کارایی این نرم افزار، «ورداسمیت»، شامل: قابلیت پردازش متون فارسی و انگلیسی، کاربرد در زمینه سبک‌شناسی، امکان جستجوی واژه‌ها و عبارت‌ها در متن، امکان انتخاب پیکره مورد بررسی برای جستجوی واژه‌ها و عبارت‌ها، مشخص کردن باهم‌آیندهای یک واژه و مشخص کردن الگوی توزیع یک واژه در متن است. با انتخاب گزینه‌های گوناگون این نرم‌افزار نتایج مختلفی از به کار رفتن واژه‌ها در بافت‌های گوناگون به دست می‌آید که می‌تواند تحلیل‌گر را در انواع تحلیل‌های گفتمانی از جمله مشخص نمودن سبک یک متن، نگرش و سوگیری خاص یک متن، تحلیل گفتمان متون سیاسی، تاریخی، مذهبی و.... یاری رساند.



سخنرانی علمی: زبان‌شناسی و کاربرد آن در حوزه بالینی (گفتاردرمانی)

هدف از برگزاری این رویداد، معرفی زبان‌شناسی بالینی به عنوان حوزه‌ای میان‌رشته‌ای و پرداختن به نقش دانش زبان‌شناسی در تشخیص و درمان بسیاری از اختلال‌های رشدی و اکتسابی زبان بود. محورهای این نشست شامل: پرداختن به مفهوم ارتباط و مولفه‌های آن، تعریف اختلال‌های رشدی زبان و عوامل پاتولوژیکی و غیرپاتولوژیکی مطرح، طبقه‌بندی انواع اختلال‌های رشدی زبان، نقش زبان‌شناسی در پژوهش‌های گفتاردرمانی و اتخاذ رویکردهای درمانی مناسب، بررسی آسیب‌های زبان‌شناختی در اختلال‌های زبانی و ارتباطی بزرگسالان، بررسی پارامترهای روان‌شناسی زبان، انواع

درمان‌ها در حوزه اختلال‌های اکتسابی زبانی و نقش زبان‌شناسی در انتخاب مولفه‌های گوناگون زبانی برای تحریک‌های زبانی مناسب است. جمع‌بندی مطالب ارائه شده در این نشست را می‌توان اینگونه خلاصه کرد: زبان‌شناسی در ارایه چارچوبی محکم و قابل استناد برای روند رشد طبیعی زبان در کودکان فارسی زبان، بررسی فرآیندهای گوناگون دخیل در اختلال‌های رشدی زبان، ساخت آزمون‌های گوناگون، بررسی ویژگی‌های زبان مربوط به هر کدام از اختلال‌ها و تهیه داده اولیه مربوط به روند رشد طبیعی زبان در کودک، می‌تواند به گفتاردرمانی کمک کند. در حوزه اختلال‌های زبانی مربوط به بزرگسالی نیز، انتخاب پارامترهای روان‌شناسی زبان و مشخصه‌های زبانی خاص به منظور تحریک‌های موثر زبانی، با همکاری زبان‌شناس صورت می‌پذیرد. همچنین در ساخت ابزارها و آزمون‌های خاص تشخیصی برای آلزایمر، دمانس و سایر اختلال‌های زبانی مربوط به بزرگسالان می‌توان به صورت تیمی (متشکل از گفتار درمان، زبان‌شناس، روان‌شناس، عصب‌شناس، متخصص هوش مصنوعی و....) عمل کرد.



سخنرانی علمی: علوم انسانی در رایاسپهر

دوران کنونی را عصر انفجار اطلاعات، دوران موج سوم تحولات جهانی، انقلاب انفورماتیک و روزگار فناوری اطلاعات نامیده‌اند. ما در فضائی رایانشی و الکترونیکی (یعنی رایاسپهر) شناوریم. برای اینکه بتوانیم در چنین فضائی جایگاه شایسته خود را به دست آوریم و همدوش دیگران پیش رویم، چه باید بکنیم؟ وضعیت و جایگاه علوم انسانی بویژه زبان‌شناسی و زبان و فرهنگ ما چگونه است؟ نقش برنامه‌ریزان زبان، زبان‌شناسان و دست‌اندرکاران این حوزه چیست؟ برای پاسخ به این

سوال‌ها، در این رویداد، ضمن ارائه توضیحات مفصل درباره مفهوم «رایاسپهر»، به بررسی جایگاه علوم و فنون گوناگون، به طور خاص علوم انسانی، در رایاسپهر پرداخته شد تا به این وسیله اهمیت امکانات موجود در رایاسپهر در پیشرفت‌های علوم انسانی مشخص گردد. سپس، به بررسی جایگاه زبان‌شناسی و زبان فارسی در رایاسپهر پرداخته شد و در نهایت گام‌های اساسی برای تثبیت بیشتر جایگاه خط و زبان فارسی در رایاسپهر معرفی گردید که این گام‌ها شامل: زمینه‌سازی، بسترسازی، ابزارسازی و محتواسازی است. اکنون سه گام اول در کشور به خوبی برداشته شده است و تنها چیزی که لازم داریم تولید محتوای مناسب صوری، فکری، علمی و ... است.



سخنرانی علمی: پیکره و برجسب‌گذاری آن با کمک رایانه: ویژگی‌ها، کاستی‌ها و چالش‌ها

یکی از چالش‌های مطرح در زبان‌شناسی پیکره‌ای و رایانشی، نشانه‌گذاری داده‌ها (اضافه کردن اطلاعات زبانی توضیحی-تفسیری به پیکره) است. در این رویداد سعی شده است انواع نشانه‌گذاری (دستی و ماشینی) و ویژگی‌های مثبت و کم و کاستی‌های هر کدام از رویکردها معرفی شود. در این راستا برخی از مدل‌های ماشینی نشانه‌گذاری داده برای زبان فارسی، نوعاً درخت نحوی/تری‌بنک توضیح داده شده است. بررسی نقاط ضعف و قوت نحوه نشانه‌گذاری داده (دستی و ماشینی) نشان می‌دهد، یادگیری ماشینی فعال، روش مناسبی برای نشانه‌گذاری داده در حداقل حجم و کاهش زمان و هزینه است. همچنین در این رویداد به بررسی چالش‌های مربوط به داده‌های زبان فارسی،

برچسب‌زنی و مدل‌یادگیرنده نیز پرداخته شده است. نتایج بررسی‌ها نشان می‌دهد علیرغم چالش‌ها، برچسب‌گذاری ماشینی بسیار بهتر از برچسب‌گذاری دستی است و این چالش‌ها در مقایسه با صرف زمان و هزینه برچسب‌گذاری دستی، ناچیز است.



سخنرانی علمی: اختصارسازی در زبان فارسی

در زبان فارسی کلمات و اصطلاحات تخصصی از رشته‌های مختلف وارد زبان شده است. این مسئله به خصوص در رشته‌هایی مانند پزشکی و رایانه که در زندگی مردم عادی بیشتر وارد شده‌اند، پررنگ‌تر است. بنابراین مردم عادی مجبورند که واژه‌ها و اصطلاحات تخصصی همچنین سرنام‌های این حوزه‌های تخصصی را بشناسند و با آنها آشنا باشند تا مشکل تفهیم و تفاهم پیش نیاید. در این سخنرانی، پس از بیان مقدمات، دلایل پیدایش اختصارات بیان شد. سپس، صورت‌های مخمر، صورت‌های ادغام‌شده و سرواژه‌ها به عنوان صورت‌های مختلف اختصارسازی بررسی شدند. در انتها نکات موردنیاز به در نظر گرفته شدن در ایجاد صورت‌های سرواژه‌سازی ذکر شدند.

میزگرد تخصصی: زبان‌شناسی رایانشی

یکی از دغدغه‌هایی همیشگی ایرانداک به عنوان آرشیو ملی پایان‌نامه‌ها و رساله‌های دانش‌آموختگان تحصیلات تکمیلی کشور، ارتقا‌روش‌های ذخیره‌سازی، جستجو و بازیابی اطلاعات

در زبان فارسی برای افزایش بهره‌وری و رضایت پژوهشگران کشور می‌باشد. با توجه به نقش زبان-شناسی رایانشی در ذخیره‌سازی، جستجو و بازیابی اطلاعات، ایرانداک زبان‌شناسی رایانشی را یکی از حوزه‌های پژوهشی اصلی خود قرار داده و علاوه بر ایجاد زیرساخت‌های فنی موردنیاز و تعریف و حل مسائل این حوزه توسط پژوهشگران پژوهشگاه، میزگردهایی تخصصی میان صاحب‌نظران این حوزه نیز برگزار کرده است.

ایرانداک سه میزگرد تخصصی با حدود ۱۰۰ نفر شرکت‌کننده در این زمینه برگزار کرده است. این میزگردها با عناوین زبان‌شناسی رایانشی، زبان‌شناسی رایانشی و علوم شناختی و پردازش زبان طبیعی و علوم شناختی برگزار شده است.



میزگرد تخصصی: اصطلاح‌شناسی در عصر حاضر

اصطلاح‌نامه گنجینه‌ای از واژه‌هاست که افزون بر نظم الفبایی فرهنگ‌ها، دارای نظامی شبکه‌ای و

مفهومی میان واژه‌های یک یا چند حوزه از دانش بشری است. این نظام مفهومی که ارتباطات گوناگون میان واژه‌ها را نشان می‌دهد، در سازمان‌دهی و بازیابی اطلاعات، تحلیل اطلاعات و علم‌سنجی، و آموزش کاربردهای بسیاری دارد. با توجه به اهمیت اصطلاح‌نامه‌ها، علاوه بر تدوین و به‌روزرسانی اصطلاح‌نامه‌های تخصصی به عنوان یک از اهداف پژوهشی اصلی در ایرانداک، این موضوع دست‌مایهٔ گفتمانی میان صاحب‌نظران این حوزه نیز قرار گرفته است.

ایرانداک میزگرد تخصصی در این زمینه برگزار کرده است که در آن سخنرانان درباره موضوعات «اصطلاح‌شناسی ابزاری برای سازمان‌دهی دانش»، «حرکت دیالکتیکی در پیدایش و خلق اصطلاح‌نامه‌ها و هستی‌شناسی‌ها»، «کاربردهای اصطلاح‌نامه در آموزش و تحلیل اطلاعات» و «کاربرد اصطلاح‌نامه در محیط دیجیتال» به ایراد سخنرانی پرداختند.



کارگاه و دوره آموزشی

ایرانداک آموزش‌های گوناگونی را برای گسترش چگونگی تدوین اصطلاح‌نامه‌ها و پردازش زبان طبیعی ارائه می‌دهد که شرکت در آنها برای همگان آزاد است. از جمله آموزش‌های برگزار شده توسط ایرانداک (و شمار برگزاری آنها) در این زمینه می‌توان به موارد ارائه شده در جدول زیر اشاره کرد.

شمار	نام
۱۸	آشنایی با نمایه‌سازی و نمایه‌سازی ماشینی
۱۲	تدوین و کاربرد اصطلاح‌نامه
۸	آشنایی با شبکه‌های عصبی و استفاده از جعبه‌ابزار شبکه‌عصبی نرم‌افزار (MATLAB)
۴	پردازش زبان طبیعی
۴	آموزش تدوین اصطلاح‌نامه
۳	آشنایی با داده‌کاوی و کاربردهای آن
۲	وب معنایی: مفهوم‌سازی و کاربرد هستان‌شناسی (آنتولوژی)
۱	معرفی واژه‌شکن فارس
۱	فن ترجمه متون تخصصی

