پژوهشگاه علوم و فناوری اطلاعات ایران

(ایرانداک)

نقش ایرانداک

در گسترش خط و زبان فارسی

پاییز ۱۳۹۷

به نام خدا

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 دهۀ 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 ماموریت‌های ملی ایرانداک می‌باشد. اسناد بالا دستی همچون نقشه جامع علمی کشور، سند چشم انداز، 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.

مقدمه

فناوری‌های نوین بر بازنمایی و بازیابی دانش در زمینه مستندسازی تأثیر بسیاری داشته‌اند. این فناوری‌ها تغییرات بزرگی را در بازنمایی اطلاعات دیجیتالی و بازیابی آنها پدید آورده و بر رفتار کاربران نیز تأثیر بسیاری گذاشته‌اند. در سایه گسترش روزافزون متن‌های دیجیتالی، 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 به‌کار می‌روند. هستان‌شناسی در بازنمایی دانش از کلیدی‌ترین ابزارهای سازمان‌دهی اطلاعات در وب به‌شمار 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، برابرنهادهای درست 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 بازنشر 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 مرجح با سرمدخل‌هایی برای مترادف‌ها یا شبه‌مترادف‌ها همراه باشند. اصطلاح‌نامه، راهنمای نمایه‌ساز و کاربر برای گزینش اصطلاح یکسان یا ترکیب یکسانی از اصطلاح‌های مرجح برای بیان یک

موضوع است. با ساخت واژه‌های استاندارد برای یک موضوع، یک‌دستی مدخل‌ها در نمایه‌سازی فراهم می‌شود. با تفکیک هم‌نویسه‌ها/هم‌نگاره‌ها به گونه‌ای که هر یک تنها به یک معنی رهنمون شود و گزینش اصطلاح مرجح 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 به‌شکل خودکار می‌انجامد؛ این امر با کاربست الگوریتم‌ها و ساختارهای داده در علوم رایانه، به‌ویژه 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، نیاز به مجموعه‌ای از ابزارها برای پیش‌پردازش و آماده‌سازی متن دارد. این ابزارها در دو دسته‌اند که گاهی آمیخته آنها نیز به‌کار 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 گفت‌وشنود 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 گفت‌وشنود، 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 به‌جای پیکره‌های متنی عمومی در بازار، پیکره‌هایی بر پایه داده‌های علمی و فنی پایگاه‌های داده ایرانداک ساخته شود. از آنجایی‌که این داده‌ها از پایان‌نامه‌ها و رساله‌های بروز به‌دست می‌آیند، می‌توانند در بسیاری از کارهای پژوهشی به‌کار روند و دستاوردهای بهتری را به‌بار آورند.

سامانه‌ها

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

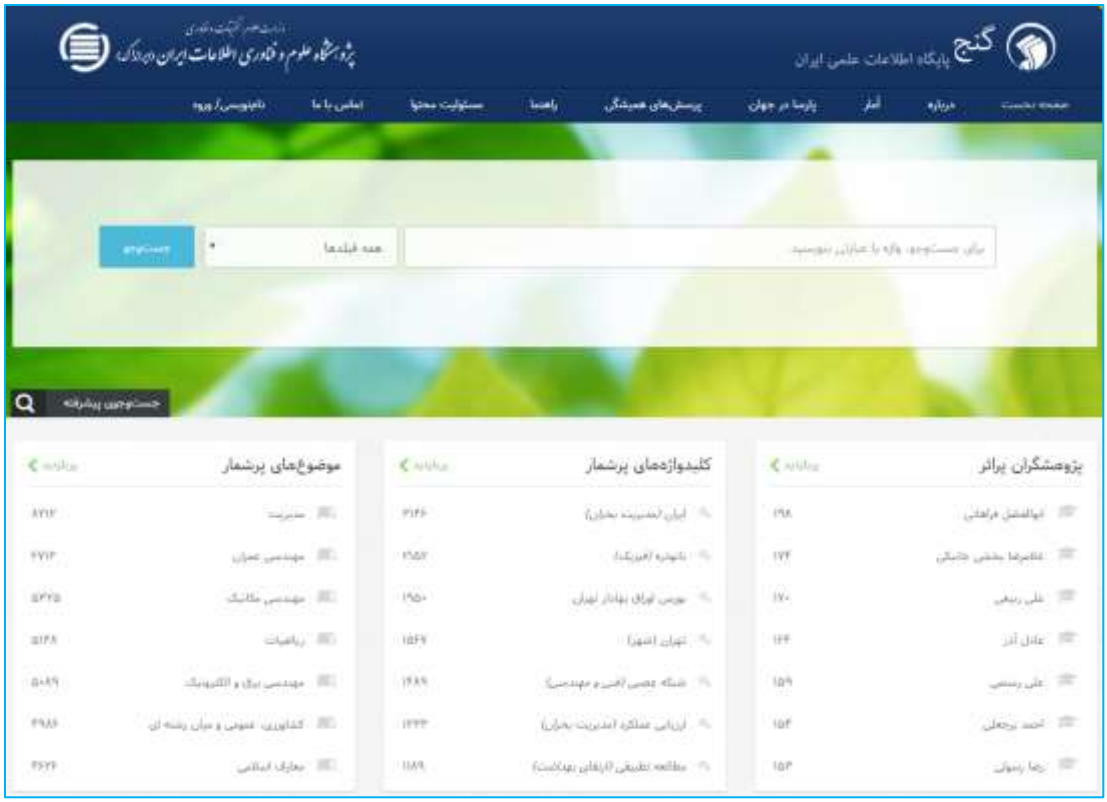
سامانه ملی ثبت پایان‌نامه، رساله، و پیشنهاده

هر چند سامانهٔ ملی ثبت پایان‌نامه، رساله، و پیشنهاده از سال ۱۳۸۶ و بر پایهٔ بخش‌نامهٔ وزارت عتف (شمارهٔ ۴۳۸۹/۱۲۲۳۸ تاریخ ۱۳۸۶/۸/۲۰) به‌راه افتاد، ولی این سامانه در آبان ۱۳۹۵ برای پشتیبانی از آیین‌نامهٔ پیش‌گفته در نشانی sabt.irandoc.ac.ir درست شد. دانشجویان پس از تصویب پیشنهادهٔ خود، به این سامانه مراجعه و اطلاعات پیشنهادهٔ خود را ثبت و شناسهٔ ره‌گیری دریافت می‌کنند. «پارسا»ها نیز پس از ثبت و بارگذاری فایل تمام‌متن و تأیید ایرانداک، آماده پذیرش دانشگاه می‌شوند. دریافت شناسهٔ ره‌گیری به‌منزلهٔ 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. تا پایان آذر ۱۳۹۶ بیش از ۲۹۳ هزار «پارسا» و بیش از ۱۱۶ هزار پیشنهاده در این سامانه ثبت شده‌اند.

پایگاه اطلاعات علمی ایران (گنج)

پایگاه اطلاعات علمی ایران (گنج)، دستاورد کاربرد فناوری اطلاعات در خدمات مدیریت اطلاعات علم و فناوری در ایرانداک است. پیشینه این بخش از خدمات ایرانداک به دهه چهل خورشیدی بازمی‌گردد که کار گردآوری، سازمان‌دهی، و اشاعه اسناد علمی و فناورانه در مرکز اسناد علمی آغاز شد. بیشترین داده‌های روزآمد گنج از پایان‌نامه‌ها و رساله‌ها و پیشنهاده‌های آنهاست، ولی 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 «مین‌فریم» 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.

در پاییز ۱۳۹۶، روزانه بیش از ۹هزار کاربر یکتا به گنج مراجعه کرده و در نه‌ماه نخست ۱۳۹۶ میانگین روزانه نزدیک به ۲۶۱هزار جست‌وجو انجام داده‌اند.



سامانهٔ پیشینهٔ پژوهش

در پاسخ به خواست جامعهٔ علمی کشور برای هدفمندسازی و افزایش کارایی و اثربخشی پژوهش‌ها، ایرانداک سامانهٔ پیشینهٔ پژوهش را با پشتوانهٔ در حال افزایش صدها هزار «پارسا» از آبان ۱۳۹۳ راه‌اندازی کرده و در نشانی pishineh.irandoc.ac.ir در اختیار همهٔ پژوهشگران کشور قرار داده است. راه‌اندازی سامانهٔ پیشینهٔ 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. این سامانه تاکنون به بیش از ۳۸۱هزار درخواست پاسخ داده و در پاییز ۱۳۹۶ میانگین روزانهٔ درخواست‌ها، ۴۵۵ بوده است.

سامانه همانندجو

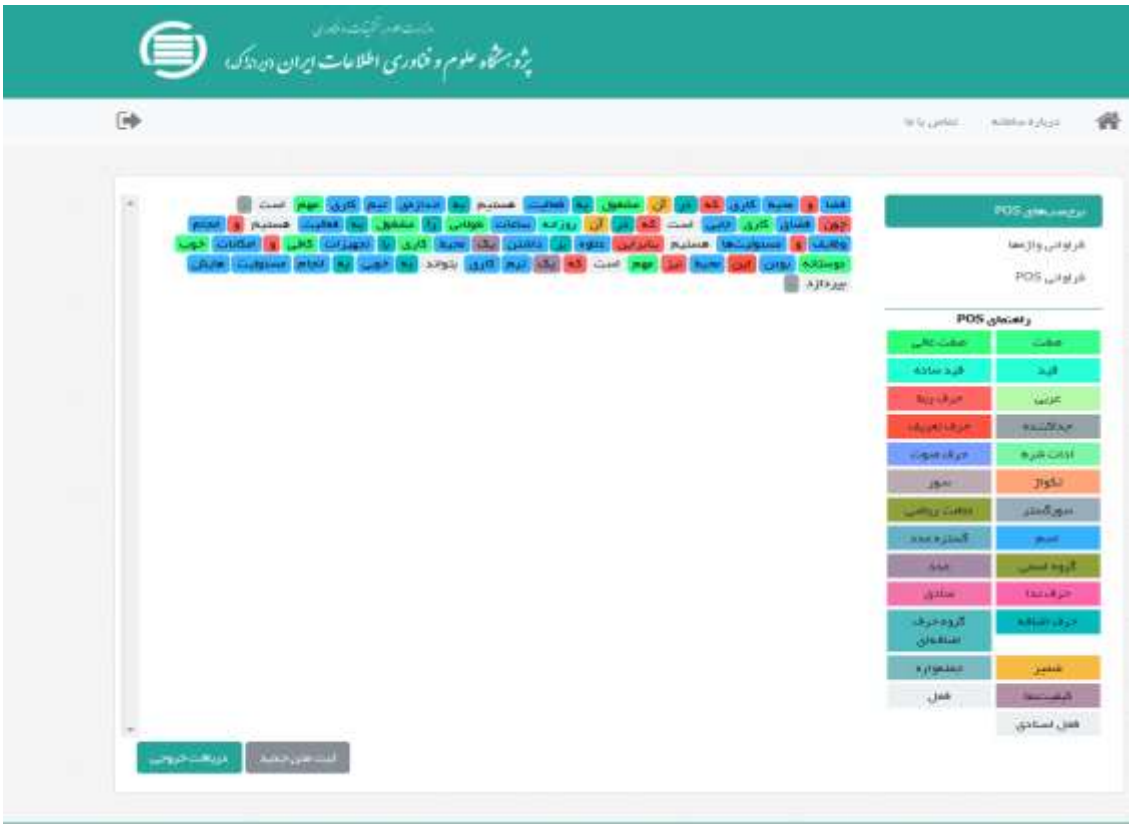
برای اجرای سیاست‌های کلی علم و فناوری و در پاسخ به خواست جامعهٔ علمی کشور برای گسترش اخلاق پژوهش و پشتیبانی از مالکیت فکری و معنوی و هم‌چنین پیش‌گیری از بدرفتاری‌های علمی، به‌ویژه سایه‌نویسی و دستبرد علمی در نگارش پایان‌نامه‌ها و رساله‌ها، ایرانداک سامانهٔ «همانندجو» را در نشانی tik.irandoc.ac.ir با پشتوانهٔ تمام‌متن و فزایندهٔ صدها هزار «پارسا» ۲۰ آذر

۱۳۹۵ پس از یک دورهٔ آزمایشی 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. هم‌اکنون در این سامانه ۲۵۵ مؤسسه و بیش از ۲۰ هزار استاد و دانشجو عضویت دارند.

سامانهٔ برچسب‌گذاری اجزای کلام برای متون فارسی

برچسب‌گذاری اجزای کلام (Part Of Speech tagging)، عمل انتساب برچسب به کلمات تشکیل‌دهندهٔ یک متن یا یک پیکره است. این برچسب‌گذاری براساس مقولهٔ آن کلمه در متن (مانند « اسم»، «فعل»، «قید»، «صفت»، و غیره) صورت می‌پذیرد. برچسب‌گذاری اجزای واژگانی کلام، عملی کاربردی در بسیاری از حوزه‌های پیشرفته‌تر پردازش زبان طبیعی از جمله ترجمهٔ ماشینی، خطایاب، تبدیل متن به گفتار، بازیابی اطلاعات، موتورهای جستجو و کمک به مدل‌های آماری است. سامانهٔ برچسب‌گذاری اجزای کلام (که دسترسی به آن در نشانی http://saba.irandoc.ac.ir فراهم شده است)، امکان وارد کردن متون گوناگون فارسی، برچسب‌گذاری مقوله‌ای کلمات تشکیل‌دهندهٔ متون، مشاهدهٔ فهرست کلمات برچسب‌خورده همراه با فراوانی آن کلمات در متن، مشاهدهٔ فراوانی برچسب‌ها در متن، مشاهدهٔ فهرست اسم‌ها به ترتیب فراوانی آنها در متن، رفع ابهام از برچسب برخی از هم‌نگاره‌های اسمی و صفتی فارسی و دریافت خروجی هر کدام از فهرست‌ها را

فراهم می‌آورد. طراحی سامانهٔ برچسب‌گذاری اجزای کلام، مبتنی بر کدهای ابزار برچسب‌گذاری «هضم» است. اما سعی شده است از طریق به کار بردن نرم‌افزار مجهز به رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «ی» در فارسی، که فراوانی بالایی در پیکره‌های متنی فارسی دارند، عملکرد این ابزار بهبود یابد. نرم‌افزار مذکور مبتنی بر الگوهای حساس به بافت نحوی مستخرج از پیکرهٔ «بی‌جن‌خان» است. پس از ارزیابی‌های انجام شده، تنها الگوهایی که تأثیر مثبت در برچسب‌زنی داشته‌اند به برچسب‌زن «هضم» (که مجددا با «پیکرهٔ بی‌جن‌خان» آموزش دیده است)، اضافه شده است.



سامانهٔ هوشمند استخراج پژوهشگران مشابه

در پایگاه داده‌های ثبت وگنج 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

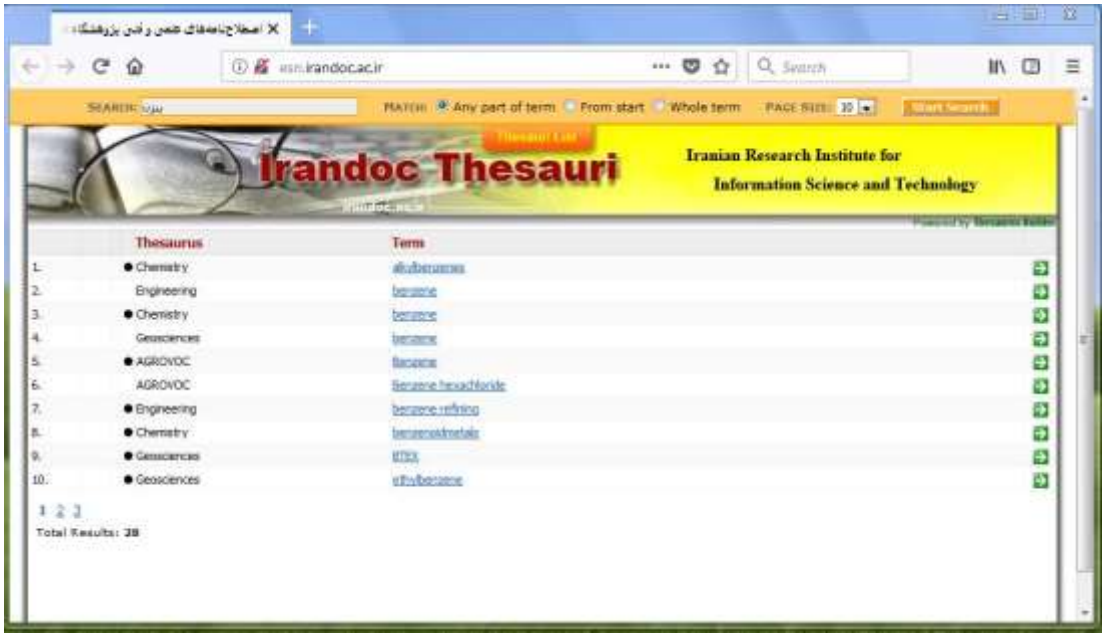
استفاده از الگوریتم‌های زبان‌شناسی رایانشی و تطبیق هوشمند رشته، پژوهشگران یکتایی که به اشکال گوناگون ثبت شده‌اند را استخراج می‌کند.

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

سامانه جامع اصطلاح‌نامه‌های علمی و فنی

در این سامانه؛ esn.irandoc.ac.ir ؛ تمامی اصطلاح‌نامه‌هایی که تا کنون 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 در سال ۱۳۸۶، اصطلاح‌نامه‌های ریاضی، شیمی، فیزیک، زمین‌شناسی، زیست‌شناسی، فنی و مهندسی، کشاورزی، مدیریت بحران، ارتقای بهداشت، و جامعه‌شناسی را ترجمه و تدوین و منتشر کرده است.

در سال ۱۳۹۷ ویرایش جدید و روزآمد شده تمامی این اصطلاح‌نامه‌ها در وبگاه آزمایشی بارگذاری شده و به زودی جایگزین ویرایش قدیمی خواهند شد. 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.



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 اعضا 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 پژوهشکدهٔ «علوم اطلاعات» قرار گرفته‌اند. این دو گروه هم‌اکنون شش عضو هیات علمی دارد.

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

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

تقی رجبی
دانش‌آموخته کارشناسی‌ارشد شیمی فیزیک از دانشگاه تربیت مدرس
پژوهشکده علوم اطلاعات
گروه پژوهشی اصطلاح‌شناسی و هستان‌شناسی

دکتر الهام علایی ابوذر
دانش‌آموخته دکتری تخصصی زبان‌شناسی همگانی از دانشگاه تهران
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

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

پژوهش

اصطلاح‌نامه‌نگاری

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 معناییِ گوناگون میان واژه‌ها را نشان می‌دهد. اصطلاح‌نامه‌ها در سازماندهی اطلاعات برای ذخیره و نمایه‌سازی، بازیابی اطلاعات به کمک شبکهٔ 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 اصطلاح‌نامهٔ خود را با نام اصطلاح‌نامهٔ 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. این اصطلاح‌نامه‌ها هم‌اکنون مورد روزآمدسازی قرارگرفته و بر روی وبگاه ایرانداک پیاده‌سازی شده‌اند. ایرانداک با سازمان‌های جهانی اصطلاح‌شناسی مانند «ترم نت» و «اینفوترم» نیز همکاری دارد و

توانسته است با کمک یونسکو برخی از اصطلاح‌نامه‌های خود را در سایت‌های جهانی نیز بگذارد.

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 ساختواژی-نحوی- معنایی و لغوی متن است که تاثیر 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 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 می‌شود. سپس در مرحله بعد راهکارها یا راه‌حل‌هایی برای رفع یا بهبود آنها بر پایه مطالعات و روش‌های بکار گرفته شده در پژوهش‌های موجود در این زمینه ارائه می‌شود.

طراحی یک روش هوشمند تجزیه رشته‌های مرجع در زبان فارسی

یک رشته مرجع مجموعه‌ای از مولفه‌ها مانند نام نویسندگان، عنوان، محل نشر، سال نشر، شماره صفحات و ... بوده و بسته به ترتیب قرارگیری این مولفه‌های می‌تواند به فرمت‌های گوناگونی نوشته شود. تجزیه رشته‌های مرجع گامی ابتدایی و ضروری برای کارهای پیشرفته‌تری مانند تطبیق رشته‌های مرجع و تحلیل مراجع و مآخذ است. در حالیکه تجزیه رشته‌های مرجع توسط کاربر انسانی به راحتی انجام‌پذیر است، تنوع موجود در فرمت‌های نگارش رشته‌های مرجع در کنار اشتباهات رخ داده توسط نویسندگان در نگارش این رشته‌ها، خودکارسازی این عملیات را سخت کرده است. روش‌های زیادی برای خودکارسازی تجزیه رشته‌های مرجع ارائه شده است اما این روش‌ها وابسته به زبان بوده و استفاده از روشی ارائه شده برای یک زبان در زبانی دیگر منجر به نتایج اشتباه می‌شود. تحقیقات صورت گرفته بیان‌گر این است که تاکنون هیچ روشی برای خودکارسازی تجزیه رشته‌های مرجع در زبان فارسی ارائه نشده است. با توجه به این مهم و نقش گسترده این مساله در ساخت خودکار شبکه‌های استنادی مدارک علمی و فرایندهای بازیابی اطلاعات، در این پژوهش به این مساله پرداخته و با استفاده از قواعد موجود در زمینه استناددهی و تکنیک‌های یادگیری ماشین روشی جهت تجزیه خودکار رشته‌های مرجع و استخراج مولفه‌های آن در زبان فارسی ارائه خواهد شد.

طراحی سامانه برچسب‌دهی به اجزای کلام برای متون فارسی

در پژوهش حاضر به این مسئله پرداخته شد که آیا با رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «-ی»، که فراوانی بالایی در پیکره‌های متنی فارسی دارند، کارایی یک سیستم برچسب‌زنی خودکار، افزایش می‌یابد و در نهایت می‌توان سامانه‌ای طراحی کرد که عمل برچسب‌دهی خودکار را با در نظر گرفتن رفع ابهام از برچسب هم‌نگاره‌های اسمی و صفتی مختوم به «-ی» در فارسی، با کارایی بهتری انجام دهد؟ سیستم مورد مطالعه در پژوهش حاضر، سیستم «هضم» بود. در پژوهش حاضر، نرم‌افزاری جهت رفع ابهام از برچسب نحوی هم‌نگاره‌های 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 می‌توان برچسب درست را به هم‌نگاره‌های مذکور اختصاص داد. ارزیابی کلی نرم‌افزار تهیه شده جهت رفع ابهام از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «-ی» در فارسی، نشان می‌دهد اگر تنها الگوهای حساس به بافت نحوی که تأثیر مثبت در برچسب‌زنی داشته‌اند را به برچسب‌زن «هضم» اضافه کنیم، صحت (Accuracy) کلی برچسب‌زن ۹۵,۶۹۱ درصد می‌شود که ۱,۳۴ درصد نسبت به حالتی که از تمام الگوهای حساس به بافت نحوی استفاده شود، بالاتر است.

ارائه روشی هوشمند برای استخراج کلیدواژه از مستندات علمی زبان فارسی بر اساس سیستم‌های پیشنهاددهنده

با توجه به افزایش حجم تولید و ثبت مستندات علمی، به خصوص در پایگاه گنج، نیاز است که فرایند نمایه‌سازی با سرعت بیشتری انجام شود. هم‌اکنون علاوه بر متوسط نرخ ثبت روزانه ۲۰۰ عنوان پایان‌نامه در روز در گنج، حدود ۱۴۰هزار سند موجود است که هنوز نمایه نشده‌اند. در حال حاضر فرایند نمایه‌سازی در ایرانداک براساس دانش و تخصص نمایه‌سازان انجام می‌شود و روش‌های ماشینی برای پردازش خودکار متون و نمایه‌سازی آنها استفاده نمی‌شود. در بسیاری از پایگاه‌های 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 همین راستا، در این پژوهش، با هدف تسریع فرایند نمایه‌سازی، با استفاده از روش‌های یادگیری ماشین و تحلیل آماری متون روشی ارائه می‌گردد 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 در نهایت انتخاب کند. پیشنهاد کلیدواژه به نمایه‌ساز، براساس رویکرد سیستم‌های پیشنهاددهنده و با اتکا به دانش ضمنی موجود در مجموعه مستندات نمایه شده قبلی صورت می‌گیرد.

طراحی و ساخت سامانه‌ی استانداردساز و خطایاب متون فارسی

روزانه هزاران مستند متنی متنوع در حوزه‌های مختلف علمی بر روی وب جهان‌گستر 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 انجام‌گیرد، 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 املایی نیز وجود دارد که فحوای گفتمانی متن را منحرف کرده و درک آن را با مشکل مواجه می‌کند. بنابراین، کلیه‌ی نویسه‌های متن باید به حالت استاندارد تبدیل شوند و عاری از هرگونه خطاهای املایی 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 املایی 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 موثراند. 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 مقدمه‌ی 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 گروهای 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. علاوه‌براین، درحالیکه 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 تقریبا 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 با جمع‌آوری و پردازش تمام پایان‌نامه‌ها و رساله‌های وزارت علوم، تحقیقات و فناوری از جایگاه ممتاز و رو به رشدی برای دانشجویان، اساتید و مدیران برخوردار شده 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 پژوهشگرمی‌باشد. 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 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 علمیِ 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 ۶ 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 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 به‌کارگیری عناصر زبانی از جمله واژه‌ها، ساختار دستوری و آواها گفته می‌شود که اخیرا 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 به‌کار 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 برچسب نحوی هم‌نگاره‌های اسمی و صفتی فارسی

بررسی هم‌نگاره‌ها در پیکره‌های متنی فارسی نشان می‌دهد اکثر این هم‌نگاره‌ها، در اثر یکسان بودن نمود نوشتاری تکواژ یای نکره، یای اسم‌ساز، شناسهٔ دوم شخص مفرد، یای صفت‌ساز و یای متصل به گروه اسمی ایجاد شده‌اند. در صورت رفع ابهام از برچسب نحوی این هم‌نگاره‌ها (هم‌نگاره‌های اسمی و صفتی مختوم به «-ی»)، رفع ابهام معنایی از کلمات با سهولت بیشتری انجام خواهد شد. در پژوهش حاضر ابتدا فهرست مبسوطی از هم‌نگاره‌های اسمی و صفتی مختوم به «-ی» تهیه شد؛ قواعد حساس به بافت نحوی جهت رفع ابهام از برچسب نحوی هم‌نگاره‌های مختوم به «ی» استخراج شد. سپس جهت بررسی صحت قواعد مذکور، برنامهٔ ماشینی تهیه شد که صحت قواعد مسخرج 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 هم‌نگاره‌های اسمی و صفتی به منظور کمک به برچسب‌دهی «اسم» به کلیدواژه‌ها در پیکره‌های علمی

در تحقیق حاضر ساخت‌واژهٔ 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 هم‌نگاره‌ها در زبان فارسی پرداخت. سپس، انواع هم‌نگاره‌ها در زبان فارسی مورد مطالعهٔ دقیق قرار گرفت و در نهایت از طریق جستجو به دو روش ماشینی و دستی، فهرست مبسوطی از هم‌نگاره‌ها از پیکره‌های«پیکرهٔ متنی زبان فارسی»، «پایگاه دادگان زبان فارسی» و «پیکرهٔ وابستگی نحوی زبان فارسی» تهیه شد. بررسی کلی هم‌نگاره‌ها در پیکره‌های مورد مطالعه نشان می‌دهد که بیشتر هم‌نگاره‌ها، فراوانی بالایی در پیکره‌های متنی فارسی دارند و اکثر آنها در اثر یکسان بودن نمود نوشتاری تکواژ یاء نکره، یاء اسم‌ساز، شناسهٔ دوم شخص مفرد، یاء صفت‌ساز و یاء متصل به گروه اسمی، ایجاد شده‌اند.

غنی‌سازی روابط معنایی هستان‌شناسی علوم پایه (شیمی)

هستان‌شناسی در واقع ابزار بازنمودن 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 اولویت‌بعدی قرار دارد. در این پژوهش به منظور توسعه روابط معنایی هستان‌شناسی علوم‌پایه، 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 این 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 جز فعلی «خوردن»پرداخته شده است. در فعل مرکب، جز فعلی نظیر «خوردن« که به آن فعل سبک نیز می‌گویند با عناصر غیرفعلی متعددی نظیر «ضربه«، «غذا»، «زمین« و غیره هم‌نشین می‌شود و این هم‌نشینی در افعال مرکب مختلف سبب می‌شود که فعل سبک «خوردن« معانی مختلفی در کنار اجزا غیرفعلی داشته باشد. آنچه مد نظر این پژوهش بوده، استفاده از روشی داده‌بنیاد برای طبقه‌بندی معانی مختلف فعل سبک «خوردن» بوده است برای انجام پژوهش، در پیکره‌ای ۵۰ میلیون کلمه‌ای که به‌صورت تصادفی از پیکره همشهری نمونه‌گیری شده است مورد استفاده قرار گرفت. با استفاده از نرم‌افزار تحلیل پیکره‌ای AntConc عناصر غیرفعلی فعل سبک «شدن» استخراج شد. سپس ۳۵ فعل به‌ترتیب بسامد پیکره انتخاب و برای هر فعل ۲۰ نمونه کاربردی از پیکره استخراج و با استفاده از روش تحلیل نمای رفتاری، مشخصه‌های ساخت‌واژی، نحوی و معنایی آن برچسب‌دهی شد. سپس جدول نمای رفتاری افعال با استفاده از تحلیل خوشه‌ای سلسه‌مراتبی در نرم‌افزار R تجزیه و تحلیل شد. خوشه‌های معنایی به‌دست آمده از تحلیل خوشه‌ای نشان می‌دهد که معانی بر اساس بافت کاربردی افعال به خوشه‌های متمایز تقسیم می‌شوند و افعال مرکبی که اجزا غیر فعلی آنها به لحاظ حوزه معنایی مشترک هستند، در خوشه‌های مشابه و نزدیک به هم جای می‌گیرند.

تعیین قاعده وابسته‌های پسین گروه اسمی

منظور از پردازش زبان فراهم‌کردن تمهیداتی رایانه‌بنیاد است که براساس آن بتوان زبان را به شیوه‌ای ماشینی و در حد امکان بدون دخالت انسان فهم و درک کرد. در این پژوهش ابتدا پیشینه بررسی گروه اسمی و مشخصا وابسته‌های پسین آن بررسی شده و سپس قواعد زبانی وابسته‌های پسین با ذکر مثال عرضه شده است. در نهایت این قواعد به صورت فرمول‌های قراردادی ارائه شده است. لازم به ذکر است که هدف از انجام این پژوهش استخراج قواعد صوری وابسته‌های پسین گروه اسمی زبان فارسی است که زیربنای پردازش گروه اسمی زبان فارسی خواهد بود.

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

در برنامه‌های تبدیل متن به گفتار معمولا پردازش نوایی با تغییر فرکانس پایه و دیرش واحدهای بازسازی صورت می‌گیرد، که گفتار حاصل تا حدی ویژگی‌های نوایی جمله را دارد اما با توجه به آزمون‌های ادراکی و پژوهش‌های جدیدی که در زمینه بازسازی واحدهای گفتاری صورت گرفته به نظر می‌رسد بهتر است واحدهای بازسازی از بافت‌های نوایی مختلف و حساس به ساختار نوایی انتخاب شده و سپس برای تولد گفتار هم‌گذاری 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 پیچیدگی‌ها و دشواری‌های موجود، هجاهای مورد نظر تنها از چهاربافت نوایی ۱) هجای بی‌تکیه 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 عمدتا از الگوهای معناشناختی بهره می‌برند وجوه مشترکی با واژگان دارند. یک مجموعه واژگان اصطلاحات را براساس مفاهیم مرتبط سازماندهی می‌کند، در حالی که هستی‌شناسی‌ها مفهوم‌سازی کرده و روابط مفاهیم را تعیین می‌نماید. یک واژگان مشترک پیش‌زمینه و ابزاری است در به اشتراک‌گذاشتن دانش جمعی و یک هستی‌شناسی مشترک علاوه بر این ابزاری است در به اشتراک‌گذاشتن 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

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 OWL2 و بر اساس استاندارد ۲۵۹۶۴ در پایگاه اطلاعات مفهوم- بنیاد به عنوان ابزاری قوی جهت ذخیره و بازیابی اطلاعات به‌کار 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 متمادی 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 نرم‌افزار 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 ابهام‌آوایی می‌شوند. اما این تمایزات توسط برخی از پارامترهای اکوستیکی باقیمانده در سیگنال گفتار، انتقال پیدا می‌کنند. تحلیل ۱۸۱ نمونه آوایی که توسط ۱۳ زن و ۱۴ مرد در بافت‌های مختلف آوایی تولید شده‌اند، نشان می‌دهد که پارامترهای نوفه سایش شامل دیرش سایش و مرکز ثقل طیف و نیز اطلاعات آوایی واکه شامل دیرش واکه، بسامد پایانه F2 تمایز واجی [seI] -[se] را حتی در نمونه‌های کاهش‌یافته‌ای مانند [seF] در سیگنال گفتار انتقال می‌دهند. از آنجا که پارامترهای واکه، سرنخ‌های دقیق‌تر و بهتری برای انتقال تمایز واجی [seI] -[se] در سیگنال تلفنی هستند، سه پارامتر دیرش واکه، بسامد پایانه F2 و شیب گذر F2 در ۸۱۱ نمونه آوایی (۴ زن و ۹ مرد) مورد اندازه‌گیری قرار گرفت و به عنوان داده‌های 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 مل‌کپستروم 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 روش‌هایی که می‌توان برای سیستم مبتنی بر پارامتر، معیار اطمینان تعریف کرد، استفاده از درجه شباهت نرمال‌شده‌ای است که با استفاده از توزیع گوسی به دست می‌آید. دقت حدود ۱ درصد بیشتر از سیستم بازشناسی مبتنی بر HMM۳۳/۹۳درصد است و به ۳۱/۹۴ درصد

می‌رسد. بنابراین می‌توان 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 نمایه‌سازی ماشینی منابع فارسی

در این پژوهش، در ابتدا به بررسی رابطه بین نمایه‌سازی و بازیابی اطلاعاتی (جست‌وجو)، مباحث نظری و سیر تحول نمایه‌سازی ماشینی از دیدگاه تئوریک، نمایه‌سازی ماشینی به صورت عملیاتی و از نگاه فرایندی، پروژه‌های موفق انجام شده در دنیا و برای زبان‌های مختلف و نیازمندی‌های نمایه‌ساز ماشینی در زبان فارسی پرداخته شده است. در ادامه، اجزای یک نمایه‌ساز زبان فارسی تشریح گردیده و مدل مفهومی نمایه‌ساز ماشینی به صورت دقیق ترسیم شده و روابط بین اجزا، جزئیات زیر سیستم‌ها و مدل منطقی اولیه سیستم طراحی شده است. در انتها نیازمندی‌های توسعه نمایه‌ساز ماشینی برای پژوهشگاه علوم و فناوری اطلاعات ایران مورد بحث واقع شده و وضعیت پروژه‌هایی 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 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 بوده پرداخته شده است. برنامه خروجی

که فعل‌شکن فارسی نامیده شده برنامه‌ای رایانه‌ای است که صیغگان فعل‌های زبان فارسی را تجزبه 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. نرم‌افزار شکل مکتوب را می‌خواند و آن را با قواعدی که در حافظه دارد منطبق می‌کند و سپس آن را به صورت اعداد عرضه می‌کند

واژه‌شکن فارسی

واژه‌شکن فارسی برنامه‌ای رایانه‌ای است که کلمات پیچیده را به اجزاء سازنده‌اش تجزیه و مقوله دستوری و معنایی آن را مشخص می‌کند. این نرم‌افزار حافظه‌ای دارد که کلمات بسیط و تکواژها و قواعد ساخت کلمات در آن قرار دارد. نرم‌افزار ابتداء کلمه‌ای را که به آن عرضه شده در حافظه جستجو و مقوله دستوری و معنایی آن را می‌یابد. چنانچه صورت کامل کلمه موردنظر در حافظه وجود نداشته باشد نرم‌افزار در صدد تجزیه آن برآمده و در صورتی که اجزاء کلمه مفروض قاعده‌مند باشد، مقوله آن را مشخص می‌کند. چنانچه کلمه غیر دستوری باشد، نرم‌افزار تنها به تجزیه کلمه اکتفاء می‌کند.

طراحی الگوی نمایه استنادی نشریات فارسی ایران

نمایه استنادی مهمترین و اصلی‌ترین ابزار در تحلیل استنادی می‌باشد، از این طریق می‌توان ضمن آنکه امکان بازیاب اطلاعات را فراهم‌آورد، به ارزیابی کمی و کیفی تولیدات علمی، مطالعه تاریخ و ساختار علم، شناخت رابطه بین استنادها و ترسیم ساختار موضوعی رشته‌های 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 می‌تواند به عرضه کتاب‌شناختی مقالات کلیدی و برجسته علمی دست بزند و علاوه بر آن ارزیابی پژوهش و پژوهشگران را در میان انبوه اطلاعات امکان‌پذیر سازد. در این پژوهش به مطالعه و بررسی تعاریف و مفاهیم نمایه استنادی، شاخص‌های انتخاب نشریات و سوابق سایر کشورها و ایران پرداخته شده است. علاوه بر آن ساختار پایگاه‌های استنادی موسسه اطلاعات علمی آمریکا مورد بررسی قرار گرفته و فرایند تولید و انجام کار در این نمایه‌های استناد ارائه گردیده است. نهایتا بر اساس مجموعه اطلاعات جمع‌آوری شده الگوی ایجاد یک نمایه استنادی برای نشریات فارسی در ایران ارائه شده است.

فرایند نمایه‌سازی

استفاده موثر از اطلاعات به سازماندهی آن نیاز دارد و سازماندهی به فرایندهای متنوع بازنمایی محتوای مدرک برای بهبود جست‌وجو و بازیابی اطلاعات همراه است. یکی از مهمترین شیوه‌های بازنمایی محتوا، نمایه‌سازی است که با تخصیص توصیفگرها به مدارک، محتوای موضوعی آنها را نشان می‌دهد. 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 کمک 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 استاندارد کردن، یکسان‌سازی، ارجاع دهی و معادل‌یابی قرار گرفتند و در نهایت در یک مجلد به چاپ رسیدند، تا مورد استفاده دانش‌پژوهان و محققان در بخش صنعت و دانشگاه جهت تعیین اولویت‌های تحقیقاتی و جلوگیری از دوباره‌کاری 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 کاربرگه‌های 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 باز پردازش 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 قراردهند. 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 نهایتا ایجاد پایگاه تخصصی شیمی از امهات 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.

ضبط چندگانه صورت‌هایی واحد با خط فارسی باعث ایجاد مشکلاتی در کار تعلیم و تربیت می‌شود و وقت بسیار 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 رسم‌الخط، 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 علومی 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 الگوهایی کلی برای یکسان‌نویسی آنها پیشنهاد شود. این پیشنهادات منطبق با اصول و ضوابط رسم‌الخط پیشنهادی فرهنگستان زبان و ادب فارسی است.

تاریخچهٔ آواشناسی و سهم ایرانیان

آواشناسی عمدتاً در چارچوب مطالعهٔ زبان بررسی شده است. به این سبب دستاوردهای آواشناسی در این کتاب از میان آثاری که دربارهٔ مسائل کلی زبان است، استخراج و با تسلسلی تاریخی عرضه شده است. محدودهٔ زمانی آثار از حدود شش قرن قبل از میلاد تا قرن پانزدهم میلادی است. نویسنده سعی کرده داور عادلی باشد و سهم واقعی دانشمندان ایرانی را در آواشناسی مشخص کند، نه آنکه دانشمندان هموطنش را از سر احساس در بهترین جای یک قصهٔ 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 به‌عنوان 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 به‌همراه 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 به‌کاربردن مفاهیم مطرح‌شده 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 «هم-زمانی»، زبان محاوره‌ای فارسی تهرانی افراد تحصیل‌کرده 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 شبکهٔ 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 رشتهٔ 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 اصطلاح‌نامهٔ کلان است که موجب تلفیق اصطلاح‌نامه‌های متعدد موجود در سطح ملی و بین‌المللی می‌گردد و در بازیابی، 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 پیکرهٔ زبانی انتخاب گردیده است و با توجه به نظریهٔ ساختواژی 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 مرکب‌مشتق، الگوهای ساختواژی اصطلاحات علمی استخراج شده است. سپس، با تعیین روابط معنایی آنها، زمینهٔ طراحی هستی‌شناسی و شبکهٔ واژگانی این حوزهٔ علمی برای برنامه-ریزی آموزشی و پژوهشی و در نهایت، طراحی نظام مفاهیم شبکهٔ علمی کشور فراهم گردیده است.

میزگرد و سخنرانی علمی

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.

سخنرانی علمی: مکتب اصطلاح‌شناسی

در این سخنرانی، ابتدا دربارهٔ سه معنای اصطلاح‌شناسی (ترمینولوژی) صحبت بحث 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 تاثیرپذیری مکاتب اصطلاح‌شناسی از مکاتب زبان‌شناسی پرداخته شد. در نهایت به تفصیل دربارهٔ مکتب وین و تاثیرپذیری آن از برخی از مکاتب زبان‌شناسی توضیح داده شد و در این راستا به تقابل مکتب زبان‌شناسی با مکتب وین پرداخته شد.

سخنرانی علمی: تاکتیک‌های جست‌وجو و هستی‌شناسی‌ها

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 2003). 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 دست‌خط 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 دست‌خط به خصوصیات بومی خط نیز مربوط می‌شود.

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

مشابهت غریبی 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 دیرهضمی 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.

سخنرانی علمی: ایجاد هستی‌شناسی‌ها از اصطلاح‌نامه‌های موجود

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 اصطلاح نامه 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.

سخنرانی علمی: مسائل زبان‌شناختی بازشناسی گفتار فارسی

در هر ارتباط گفتاری، شنونده قادر است معنا و منظور گفتار گوینده را درک و در مورد آن اظهارنظر کند، به پرسش‌های مطرح شده پاسخ دهد و فراتر از آن در مورد گفتار بیان شده طرح پرسش کند. بدیهی است این توانایی در افراد ناشنوا به دلیل نقص در ساختارهای مغزی و شنوایی وجود ندارد. بر همین اساس بازشناسی خودکار گفتار را می‌توان «توانایی درک گفتار انسان به وسیلهٔ رایانه» یا به بیان دیگر «مدلسازی رفتار زبانی انسان به وسیلهٔ رایانه» تعریف کرد. در واقع بازشناسی خودکار گفتار بخشی از فعالیت‌های هوش مصنوعی است که در آن فعالیت‌ها و واکنش‌های طبیعی یک انسان در برخورد با گفتار، در قالب یک نظام رقومی بازسازی می‌شود. در این سخنرانی، پس از معرفی بازشناسی گفتار، به بخش‌های مختلف بازشناسی گفتار و چگونگی فرایند آن، ارتباط بازشناسی گفتار با دانش زبان‌شناسی، سیستم‌های مختلف بازشناسی گفتار و موانع رایج در سیستم‌های بازشناسی گفتار پرداخته شد.

سخنرانی علمی: معرفی فارس‌نت: نخستین شبکه واژگانی زبان فارسی

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،گسترهٔ 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، نحوهٔ 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 اسن سخنرانی، به فعالیت‌های آتی در مسیر تکمیل فارس‌نت بیان شد.

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، پر رنگ‌تر 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، ارتقا روش‌های 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 علوم‌شناختی و پردازش زبان طبیعی و علوم شناختی برگزار گردیده 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 با حدود ۳۰ نفر شرکت‌کننده 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 دورهٔ آموزشی

ایرانداک آموزش‌های گوناگونی را برای گسترش چگونگی تدوین اصطلاح‌نامه‌ها و پردازش زبان طبیعی ارائه می‌دهد که شرکت در آنها برای همگان آزاد است. این آموزش‌ها تاکنون بارها برگزار شده‌اند که نام آنها در جدول زیر آمده است.

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