پایگاه داده‌های زبان فارسی در اینترنت[1]

مصطفی عاصی
مدیر گروه زبان‌شناسی پژوهشگاه علوم انسانی و مطالعات فرهنگی
S_M_ASSI@IHCS.AC.IR

۱- پایگاه داده‌های زبانی [2]

امروزه دیگر کسی درباره لزوم بنیاد نهادن بررسی‌های زبانی و زبان‌شناختی بر داده‌های واقعی و مستند تردیدی ندارد. برای هر نوع پژوهش، به پیکره زبانی[3] ویژه‌ای که دربردارندهٔ نمونه‌های مناسب و کافی باشد، نیاز است. پیکره هرچه گسترده تر و متنوع تر باشد معتبرتر و سودمندتر خواهد بود. گستردگی و تنوع پیکره در شکل‌های سنتی محدودیت‌های بسیاری را بهمراه دارد. هنگامی که حجم پیکره از مرزی می گذرد، ساماندهی و بهره گیری از آن مشکل و سپس ناممکن میگردد. گوناگونی داده‌ها گرچه در بیشتر بررسی‌ها اهمیت بسیار و نقش تعیین کننده می یابد، اما بازهم مشکل را پیچیده تر می سازد.

از سوی دیگر، بسیاری از فعالیت‌های علمی در حوزه زبان، ادبیات و زبان‌شناسی به داده‌های مشابهی نیاز دارند که هریک برای خود به گوشه ای از گستره زبان می‌پردازند. چه بسا پیکره‌های مشابه یا دارای همپوشی فراگیر که بدون آگاهی از وجود دیگری و با صرف وقت و هزینه زیاد به وجودآمده و پس از بهره برداری به کناری نهاده شده اند. ایراد دیگری که اغلب بر این داده‌های پراکنده وارد است، داشتن ناراستی‌های فراوان بدلیل یکبارمصرف بودن آنهاست. چراکه کمتر فرصتی برای آزمودن، ویراستن و پیراستن آنها فراهم می شود. بالاخره باتوجه به ماهیت ایستای اینگونه پیکره‌ها، حتی اگر بخواهیم از آنها در طرح‌های دیگری بهره بگیریم، پس از گذشت مدتی کهنه و شاید بی اعتبار به شمار آیند.

هدف از ایجاد پایگاه داده‌های زبان فارسی (دادگان زبان فارسی)، فراهم کردن مجموعه‌ای از پیکره‌های مطلوب، مناسب و دور از نارسایی‌های یادشده است.پیکره ای که باوجود حجم عظیمی از داده‌های زبانی با گستردگی و گوناگونی‌های بسیار، دارای ساختاری بسامان و منطقی باشد تا امکان هرگونه جستجو و دستیابی سریع به آگاهی‌های موردنیاز را در هرزمان فراهم نماید. چنین پیکره ای می تواند همواره روزآیند گردد و پاسخگوی نیاز همه پژوهندگان زبان فارسی و کاربران گوناگون در همه زمینه‌های نظری وکاربردی باشد.[4]

۲- ویژگیهای پایگاه

از اوایل سال ۱۳۷۲ کار ایجاد پایگاه داده‌هایی برای زبان فارسی با طراحی و سرپرستی نگارنده در پژوهشگاه علوم انسانی آغاز شد و تاسال۱۳۷۸ دو مرحله (فاز) آن به اجرا درآمد و مرحله سوم که مهمترین فاز یعنی گسترش و افزایش حجم داده‌ها و دگرگونی اساسی در نرم افزار و ایجاد امکانات نوین شبکه ای برای ارائه خدمات و اطلاعات آن در شبکه جهانی اینترنت بود بدلیل نبود منابع مالی چندسالی از اجرا بازماند تا اینکه با کمک مالی وزارت ارتباطات و فناوری اطلاعات از سال ۱۳۸۱ اجرای فاز سوم طرح آغاز گردید ودوسال بعد به پایان رسید.

پایگاه داده‌های زبان فارسی فراگیر و متنوع است. به‌سخن دیگر فراتر از یک یا چند پیکرهٔ خاص است و کاربران برپایهٔ نیاز و هدف پژوهشی خود می‌توانند پیکرهٔ مناسب را از آن برگزینند. حتی پژوهندگان می‌توانند پیکره‌های اختصاصی خود را وارد پایگاه‌کنند و تحلیل‌ها و فهرست‌گیری‌های موردنظر خود را انجام دهند.

پایگاه داده‌های زبان فارسی همگانی است و از راه اینترنت در دسترس همهٔ دانش‌پژوهان و علاقه‌مندان به زبان فارسی است.

بهره‌گیری از این پایگاه برای کاربردهای پژوهشی رایگان است.

پایگاه داده‌های زبان فارسی تنها مجموعه‌ای از مواد خام زبانی نیست بلکه دارای متن‌های نشانه‌گذاری‌شده (از جمله شناسنامهٔ متن، برچسب‌های دستوری، آوایی، ریشه‌ای و معنایی) است که همواره افزایش می‌یابند.

این دادگان مجهز به نرم‌افزارهای اختصاصی جستجو، تقطیع و تحلیل متن است که می‌تواند انواع فهرست‌های واژگانی، بسامدی و آماری را ارائه کند.

دادگان زبان فارسی در دوره‌های زمانی منظم روزایند می‌شود و امکانات نرم‌افزاری و متن‌های تازه به آن افزوده میگردد.

در زیر برخی از ویژگی‌های یادشده تشریح می‌شود:

۲-۱ گستره زبانی

زبان فارسی مفهومی بسیار وسیع دارد و میتواند در برگیرنده همه گونه‌های گفتاری، نوشتاری، سبکی و کاربردی این زبان در همه دوران‌های تحول آن باشد. برای نزدیک شدن به این دریای داده‌ها لازم است آن را به محدوده‌هایی بخش کنیم و در مراحل منظم و بتدریج آنها را پوشش دهیم. در نخستین مرحله با توجه به نیازهای گوناگون پژوهشی و کاربردی، از طیف دوران‌های تاریخی زبان فارسی، برش فارسی معاصر برگزیده شد.

همین برش نیز که به‌طور قراردادی از آغاز قرن چهاردهم خورشیدی تا امروز را دربرمی‌گیرد، خود دارای گونه‌های بسیاری است، از جمله گونهٔ رسمی نوشتاری یا به‌اصطلاح، فارسیِ معیار و گونهٔ گفتاری آن، گونه‌های ادبی، سبکی و حرفه‌ای فارسی، گونه‌های محاوره ای و عامیانهٔ آن، و گونه‌هایی که متغیرهای زبانی و اجتماعی دیگری مانند سن، جنس، سواد و تحصیل، طبقهٔ اجتماعی و محیط‌های مختلفِ ارتباطی، عامل تمایز آنها به‌شمار می‌روند.

۲-۲ منابع گردآوری داده‌ها

با توجه به گونه‌های یادشده، بایسته است که با روش‌های متفاوت و مناسب داده‌های موردنیاز فراهم شده، در درون حافظه رایانه سازماندهی گردد. برای مثال، از گونه‌هایی که به نوشتار وابسته اند با استفاده از متن‌های معتبر و با رعایت معیارهای مختلف نمونه‌گیری شده و هیچگونه محدودیت و امساکی در مورد آثار مهم ادبی و نویسندگان سرشناس و بویژه صاحب سبک و تأثیرگذار اعمال نمی شود. تاکنون گردآوری، درونداد و سازماندهی داده‌ها در چند مرحله انجام گردیده است و بازهم ادامه خواهد یافت :

۱- نخست فهرستهای مفصلی از همهٔ منابع مهم نظم و نثر فارسی فراهم شد. این فهرست‌ها بطور جداگانه برای آثار شعری، داستانی، غیرداستانی، نمایشنامه و فیلمنامه، ادبیات کودکان، نشریه‌های ادواری و مجلات علمی، تخصصی و ادبی فراهم گردید. عناوین آثاری که در این فهرستها قرار گرفتند بیش از یک هزار و پانصد متن گردید که پس از بررسی و کنار نهادن موارد مشابه، بیش از پانصد عنوان برای درونداد به پایگاه داده‌ها برگزیده شد. می‌توان ادعا کرد که نمونه‌های برگزیده، نماینده‌ای واقعی از زبان فارسی معاصر به‌شمار می‌رود.

فهرست کامل ۱۵۰۰ متن مهم نظم و نثر ادبیات معاصر ایران و انواع دیگر متون زبانی شامل :

- ۴۵۲ اثر داستانی و غیرداستانی نثر
- ۲۴۹ اثر شعری از شاعران معاصر
- ۸۴ عنوان مجله و نشریهٔ علمی، ادبی و تخصصی
- ۳۱۱ عنوان نمایشنامه
- ۸۰ عنوان فیلمنامه
- ۲۰۰ عنوان ادبیات کودکان
- چندین عنوان روزنامه و نشریه خبری، همه پسند[۵] و متنوع
- برخی از کتاب‌های درسی دانشگاهی و دبیرستانی
- برخی از کتاب‌های دبستانی
- نامه‌های اداری و بخشنامه‌ها
- مجموعه‌ای از قوانین و مقررات
- نشریه‌ها و جزوه‌های پراکنده، پوسترها، دیوارنوشته‌ها و مانند اینها

۲ - فهرستی با بیش از ۵۰۰ اثر از میان آثار بالا برای تایپ دستی برگزیده شد.

۳- تاکنون بیش از ۳۰۰ متن در رویهم بیش از ۲۴۰۰۰ صفحه که به بیش از ۵ میلیون واژه می‌رسد تایپ شده است.

۴- متن‌های دیگری شامل کتاب و مقاله‌های تخصصی با نزدیک به ۱۰،۰۰۰،۰۰۰ واژه گردآوری شده که که بخشی از آنها وارد پایگاه شده و بقیه در دست تبدیل، ویرایش و درون‌داد است.

۵- بیش از ۶۰ ساعت گفتار پیوسته مربوط به محاورهٔ عادی افراد، برنامه‌های رادیویی و تلویزیونی برروی نوار و یا به‌صورت فایل‌های دیجیتالی ضبط شده

۶- متن‌های گفتاری از نوار بر روی کاغذ پیاده سازی شده

۷- متن‌های یادشده با بیش از ۲۰۰۰،۰۰۰ واژه تایپ شده
۸- بخشهای مشخصی از متن‌های نوشتاری و گفتاری با بیش ازسه میلیون واژه ویرایش دوباره گردیده است.
۹- بخش‌های برگزیده از متون ویرایش شده برچسب دهی دستوری و معنایی شده و پیوسته ادامه دارد.
۱۰- متن‌های زیر با روش‌های گوناگون به صورت الکترونیکی فراهم گردید :
- متن ۱۲ واژه نامه مختلف
- متن مجموعه قوانین کشوری و مقاله‌ها و رویه‌های حقوقی وقضایی
- همهٔ متن‌های روزنامهٔ همشهری از سال ۱۳۷۵ تا آغاز سال ۱۳۸۲ (رویهم ۳۴۵ مگابایت با فرمت HTML و شامل تعداد ۱۹۰۲۰۶ مقاله و ۶۳ میلیون واژه)
- همهٔ متن‌های روزنامهٔ همشهری شش ماهه آغازین سال ۱۳۸۲ (رویهم ۶۳ مگابایت و بیش از ۶،۲۵۰،۰۰۰ واژه) این بخش به‌شکل موضوعی جداسازی ودسته‌بندی شده‌است.
- متن دورهٔ شش‌ماههٔ روزنامهٔ ایران نیز به همین روش درون‌داد شده‌است.
- متن‌های نمونه‌ای از روزنامه‌ها و نشریات دیگر نیز با این روش گردآوری گردیده است.
البته از متن‌های روزنامه‌ای تنها نمونه‌هایی با نزدیک به ۳۰ میلیون واژه برگزیده شده و به پایگاه داده‌ها درونداد شده است
مجموع متن‌های گردآوری شده نزدیک به **یک‌صد میلیون** واژه می گردد که تاکنون تنها **۶۰ میلیون** واژه از آن به درون پایگاه وارد شده است.
این کار به‌صورت فعالیتی همیشگی و با افزودن منابع تازه دنبال خواهدشد.

۲-۳ ساختار زبانی پیکره

داده‌ها به شکل‌ها و قالب‌های[۶] گوناگون در این پایگاه ذخیره می شوند: به‌صورت متن‌های پیوسته کامل و یا گزیده آثار ادبی یا نوشته‌های مهم، به‌صورت فهرست‌های واژه نما و بسامدی از همین متن‌ها و متن‌های دیگر، یعنی فهرست همه واژه‌های آنها به همراه چند سطر از بافت زبانی آنها و بسامدشان، و نیز به‌صورت واژه نامه‌های تک زبانه و دوزبا نه، همچنین، متن‌های بازنویسی شده داده‌های گفتاری چه به‌صورت متن پیوسته و چه به‌صورت فهرست‌های بسامدی در پیکره جای‌دارند و پیش‌بینی شده با بکارگیری امکانات چند رسانه ای[۷]، تلفظ آوایی داده‌ها نیز ارائه گردد. همچنین متن‌های مشخصی از پیکره با روش‌های خودکار ونرم افزاری و دستی نشانه گذاری شده اند. نشانه گذاری متن برای افزودن اطلاعات گوناگون دستوری، واژگانی، معنایی، ریشه شناختی تلفظی و کاربردی به آن صورت می گیرد.

۲-۴ ساختار رایانه ای

مجمو عه داده‌های یادشده در یک پایگاه داده‌های پیوندی[۸] به گونه ای سازماندهی شده که هر واژه با پیوندهای گوناگون به متن اصلی یا بافت خود، به همه مشخصات شناسنامه ای متن مانند نام نویسنده، نام اثر، ناشر وسال و جای انتشار، شماره سطر و صفحه، دسته بندی‌های گوناگون مربوط به نوع، سبک، موضوع و رشته اثر ارتباط یابد. پیوندهایی نیز میان واژه و معنی‌های آن، مترادف‌هایش، مقوله دستوری و تلفظ آن وجود دارد که امکان هر گونه جستجو را فراهم می سازد. در این پایگاه فرایندهای پردازشی برای انجام انواع جستجوها و فهرست گیری و گزارش‌گیری‌ها و استخراج واژه نامه‌های بسامدی و فهرست‌های آماری بکارگرفته می شود.
از آنجا که طراحی پیشین پایگاه داده‌ها به چهارده سال پیش برمی‌گردد و دارای ساختاری ساده و محدود بود، شایسته بود که باتوجه به پیشرفتهای فناوری و امکانات تازهٔ آن، ساختار جدیدی تعریف گردد تا امکان به‌کارگیری حجم‌های بزرگی از متن‌های متنوع زبان فارسی و انواع پردازشهای پیچیده با سرعت زیاد فراهم گردد. مهمترین ویژگیهای سیستم جدید عبارتند از:
بهره‌گیری از پایگاه اطلاعاتی Oracle در محیط عامل ویندوز Server 2003 برای دستیابی به توانایی‌های بیشتر و انعطاف بهتر برای دریافت، ذخیره‌سازی، پردازش و ارائه متون فارسی
توان ذخیره‌سازی حجم زیادی از متون فارسی با چندصد میلیون واژه (به‌صورت متن‌های پیوسته)
امکان بررسی متن‌های درونادی به‌صورت گزینشی، پیاپی و صفحه به صفحه
امکان ویرایش متنهای درونادی با کمک ویراستار درونی نرم‌افزار

امکان ایجاد، تغییر یا جابجایی اطلاعات شناسنامه‌ای متن‌ها
امکان جستجوهای پیچیده و چندلایه با گزینه‌های متعدد
امکان تهیه فهرستهای آماری و بسامدی از واژه‌های یک یا چند متن
امکان اجرای عملیات برچسب‌دهی دستی یا خودکار
امکان بازبینی و گشت و گذار در متون با حالت ساده یا با نمایش برچسبها.
به‌کارگیری استانداردهای نوین میانای کاربری(user interface) در محیط‌های وب، اینترنتی و اینترانتی
امکان ارائه خدمات و اطلاعات و گزارشهای یادشده به کاربران و پژوهندگان ایرانی و جهانی به‌صورت برخط[۹] و برون‌خط[۱۰]
امکان نام‌نویسی و اشتراک اینترنتی
سخنگاه اختصاصی برای تبادل نظر، بحث و گفتگوی کاربران دادگان فارسی و مطرح کردن مسائل و کاستی‌ها و ارائهٔ راه‌حل و پیشنهادهایی برای بهبود و گسترش آن.

۳-کاربری‌های پایگاه

از اطلاعات و امکانات این پایگاه به روش‌های گوناگون می توان بهره گرفت :

- به روش برخط و وارد شدن به وبگاه[۱۱] پژوهشگاه و صفحه آغازه **دادگان زبان فارسی**.
- از راه اینترنت با نشانی : http://pldb.ihcs.ac.ir که در آن pldb اختصارِ Persian Linguistic Database است.

در حالت عادی کاربران اینترنتی به عنوان میهمان (guest) می‌توانند نمونه کوچکی از امکانات را برصفحه نمایشگر مشاهده کنند اما کسانی که با شرایطی ثبت نام کرده مشترک می شوند، به امکانات بیشتری دسترسی پیدا خواهندکرد.

- با درخواست گزارش به روش برون‌خط هرگونه جستجو (که در زیر خواهد آمد) در پیکره انجام می گیرد و نتیجه آن در گزارش‌هایی بوسیله چاپگر، چاپ می شود و یا به شکل پرونده رایانه ای ارائه می گردد.

۴- انواع جستجو

می توان برپایه هریک از اقلام اطلاعاتی و یا ویژگی‌های مربوط به آنها جستجوهای تک موردی، گروهی یا کلی انجام داد. از جمله :

- جستجوی واژگانی (برپایه یک یا چندکلیدواژه)
- جستجوی مفهومی (بر پایه مفهوم یا معنای مورد نظر)
- جستجوی تلفظی (برپایه صورت تلفظی یک واژه)
- جستجوی هم بافت (برپایه واژه‌های همایند[۱۲] و یا بافت‌های همسا یه)
- گشت و گذار[۱۳] در متن‌ها و واژه نامه‌ها

این جستجوها را می توان در محدوده‌های دلخواه (مثلا دوره زمانی معین، یا نویسنده ای مشخص، یا حجم معینی از پیکره) انجام داد.

۵- انواع گزارش‌ها

گزارش‌های پایگاه به گونه‌های صوری و محتوایی مختلفی طراحی شده اند تا پاسخگوی نیازهای گوناگون باشند :

- به شکل فهرست‌های واژگانی، آماری و بسامدی (صعودی، نزولی، الفبایی و الفبایی وارونه)
- به شکل اطلاعات موردی و شناسنامه آثار
- به شکل فهرست واژه نما concordance (واژه مورد نظر در شکل کاربردی آن همراه با اطلاعاتی در باره بافت زبانی آن مانند جمله شاهد، شماره سطر و صفحه متن، نام نویسنده و مشخصات اثر، تاریخ کاربرد، بسامد در پیکره و مانند آن)

۶- کاربران پایگاه

این پایگاه برای استفاده همگانی در نظر گرفته شده است، اما مراحل و سطوح دستیابی متفاوت است. درسطح مدیریت و آماده سازی، تنها مجریان و همکاران طرح به اطلاعات دسترسی دارند، اما همه افراد می توانند از راه اینترنت مشترک پایگاه شوند و اطلاعات مورد نیاز خود را دریافت کنند. از نظر سطح دستیابی، افراد، گروه‌های پژوهشی و سازمان‌ها، دارای دامنه‌های مجاز متفاوتی خواهند بود. مثلاً تنها مجریان طرح می توانند هرگونه تغییرات را در ساختار و محتوای داده‌ها بوجود آورند، اما برخی از گروه‌های پژوهشی ممکن است بتوانند به افزایش داده‌ها بپردازند و دیگران تنها دریافت کننده اطلاعات به‌شمارمی‌آیند.

۷- آیندهٔ پایگاه

پایگاه‌های داده‌های زبانی روز به روز اهمیت بیشتری می یابند. شمار آنها و موضوع و زمینه‌های کاربردشان گسترده‌تر می‌گردد. اکنون از پایگاه‌های معرفتی[۱۴] گفتگو می شود که بسیاری از رشته‌های دانش و فن به آنها مجهز می‌شوند و همه‌گونه آگاهی‌ها و معارف به صورت الکترونیک در آنها نگهداری می‌شود.[۱۵] در بانک‌های اطلاعاتی گوناگونی که در سراسر جهان در دسترس همه است، پایگاه‌های داده‌های زبانیِ بسیاری برای زبان‌های مهم جهان فراهم گردیده است. اما، تاپیش‌ازاین در این دریای بیکران اطلاعاتی، داده‌های قابل استناد برای زبان فارسی یافت نمی‌شد.

پایگاه داده‌های زبان فارسی (دادگان فارسی) در ایران، و در وهله نخست برای پاسخگویی به نیازهای پژوهندگان ایرانی ایجادشده است و در مرحله بعد به‌عنوان یک بانک اطلاعاتی ایرانی در دسترس همه کسانی است که در باره زبان فارسی در نقاط دیگر جهان پژوهش می‌کنند.

یادداشت‌ها و مراجع

۱- این مقاله با تفاوت‌هایی پیش‌تر در نشریهٔ ***پژوهشگران***، شماره ۲، فروردین و اردیبهشت۱۳۸۴، پژوهشگاه علوم انسانی و مطالعات فرهنگی به چاپ رسیده است.

۲. linguistic database

۳. linguistic corpus

۴. پیشینهٔ ایجاد و بهره گیری از پایگاه‌های داده‌ها و نمونه‌های مهم آن در جهان و دلایل نیاز به چنین ابزاری در مقالهٔ " طرح ایجاد پایگاه داده‌های زبان فارسی با کمک کامپیوتر" در نشریه ***اطلاع رسانی***، دوره ۱۱، شماره ۱، تهران، زمستان ۱۳۷۳ آمده است.

۵. popular

۶. format

۷. multimedia

۸. relational database

۹. on-line

۱۰. off-line

۱۱. website

۱۲. collocations

۱۳. browse

۱۴. knowledge base

۱۵. انواری، مرتضی و ملک آفاق فتحیان پور. " پایگاه‌های معرفتی در سیستم‌های اطلاع رسانی"، ***اطلاع رسانی***، دوره ۱۱، شماره ۱، تهران، زمستان ۱۳۷۳.