مدیریت داده‌های پژوهشی، مطالعه موردی داده‌های زبانی

مرتضی رضائی شریف‌آبادی

کارشناس ارشد زبان‌شناسی رایانشی، مرکز تحقیقات کامپیوتری علوم اسلامی (نور)

mrezaeis@alum.sharif.edu

چکیده

جمع‌آوری و تحلیل داده یکی از گام‌های اساسی پژوهش در بسیاری از زمینه‌های تحقیقاتی است. مدیریت صحیح داده‌های پژوهشی موجب می‌شود که پژوهشگران هم خود بتوانند بعدها به داده‌ها مراجعه کنند و در پژوهش‌های دیگر خود از آن‌ها استفاده نمایند، و هم با اشتراک‌گذاری داده‌ها این امکان را برای سایر پژوهشگران فراهم نمایند که بتوانند آن داده‌ها را برای اهداف دیگر مورد استفاده قرار دهند. این مسئله در خصوص داده‌های زبانی، یعنی داده‌هایی که برای مطالعه یا پردازش زبان از آن استفاده می‌شوند، نیز صادق است و زبان‌شناسان و متخصصان پردازش رایانه‌ای زبان می‌توانند با رعایت نکات مربوط به مدیریت داده‌های زبانی، گام مثبتی در پژوهش‌های حوزهٔ تخصصی خود بردارند. در این پژوهش ضمن تعریف و ارائه توضیحاتی پیرامون داده‌های زبانی، ابتدا به تجربیات جهانی از موضوع مدیریت داده‌های پژوهشی به طور عام پرداخته می‌شود و سپس تجربیات به دست آمده از تولید داده‌های زبانی و راه‌اندازی و مدیریت «مرجع دادگان زبان فارسی» با خوانندگان مقاله به اشتراک گذاشته می‌شود. نکاتی چون اشتراک‌گذاری داده‌های زبانی در پایگاه‌های اشتراک دادهٔ معتبر، مشخص کردن وضعیت مالکیت معنوی و شرایط استفادهٔ داده، تهیهٔ مستندات مناسب برای توصیف داده و مشخص کردن مستندی که کاربران باید هنگام استفاده از داده به آن ارجاع دهند، استفاده از منابع اولیه مناسب برای تولید داده، ساختاربندی داده بر مبنای قالب‌های استاندارد و مشترک، کسب اطمینان از کیفیت داده با استفاده از روش‌های مختلف ارزیابی، ایجاد ابزارهای مناسب برای پردازش و نمایش داده، و استفاده از ساختارهای مناسب برای فایل داده از جمله پیشنهاداتی هستند که برای مدیریت داده‌های زبانی در این مقاله ارائه شده است.

کلیدواژگان: داده‌های زبانی، مدیریت داده‌های پژوهشی، زبان‌شناسی پیکره‌ای، پردازش زبان طبیعی، زبان‌شناسی رایانشی.

۱. مقدمه

مقصود از دادهٔ زبانی[۱] داده‌ای است که برای مطالعه یا پردازش زبان از آن استفاده می‌شود. بهره‌گیری از داده‌های حاصل از جمع‌آوری نمونه‌های واقعی کاربرد زبان سابقه‌ای طولانی در علم زبان‌شناسی دارد. علی‌رغم انتقاداتی که چامسکی[۲] از دههٔ ۱۹۵۰ متوجه زبانشناسانی کرد که در مطالعات خود به داده‌های زبانی اتکاء می‌کردند، استفاده از چنین داده‌هایی کنار گذاشته نشد. بکارگیری رایانه برای جمع‌آوری و تحلیل داده‌های زبانی کمک شایانی به روش‌های مبتنی بر داده کرد و با توسعهٔ داده‌ها و شیوه‌های تحلیل جدید از اوایل دههٔ ۱۹۸۰، «زبان‌شناسی پیکره‌ای[۳]» که با تعریفی ساده همان «مطالعهٔ زبان بر مبنای نمونه‌هایی از کاربرد واقعی زبان» است، رشد قابل توجهی داشت (مک‌انری، ۲۰۰۱).

از سوی دیگر تولید انبوه داده‌های زبانی کمک شایانی به حوزهٔ پردازش زبان طبیعی[٤] کرده است. پردازش زبان طبیعی یک زمینهٔ پژوهشی میان‌رشته‌ای است که هدف آن استفاده از رایانه‌ها برای انجام فعالیت‌های سودمندی است که با زبان انسان سر و کار دارد. از جملهٔ این فعالیت‌ها می‌توان به ترجمهٔ ماشینی، پرسش و پاسخ خودکار و . . . اشاره کرد (جورافسکی[۵]، ۲۰۰۸). روش‌های موجود برای پردازش زبان را می‌توان از جهتی به دو دستهٔ کلی «مبتنی بر قاعده[٦]» و «آماری[۷]» تقسیم کرد. در رویکرد مبتنی بر قاعده پژوهشگران مجموعهٔ محدودی از قواعد را برای انجام فعالیت مورد نظر تدوین می‌کنند. تدوین چنین قواعدی زمان‌بر است و مجموعهٔ قواعد هیچ‌گاه فهرست کاملی را تشکیل نمی‌دهند. در رویکرد آماری به دنبال پاسخ برای

---

۱ linguistic data
۲ Noam Chomsky
۳ corpus linguistics
٤ natural language processing
۵ Jurafsky, D.
٦ rule-based
۷ statistical

این پرسش هستیم که «الگوهای متداولی که در کاربرد زبان دیده می‌شوند کدام‌اند؟» (مانینگ[۱]، ۱۹۹۹). شناسایی چنین الگوهایی توسط رایانه و با روش‌های آماری مستلزم دسترسی به مجموعهٔ بزرگی از داده‌هاست که می‌تواند به صورت مجموعه‌ای از متون و یا مجموعه‌ای از مکالمات ضبط‌شده باشد.

همانطور که ملاحظه می‌شود دسترسی به داده‌های با کیفیت که بتوان از آن‌ها برای مطالعه و پردازش زبان استفاده کرد بسیار حائز اهمیت است. برای همین منظور لازم است که تولیدکنندگان و استفاده‌کنندگان داده‌های زبانی با مسائل مربوط به نحوهٔ تولید، اشتراک‌گذاری و استفادهٔ مطلوب داده‌های زبانی آشنا شوند. به همین منظور در این مقاله به تجربیات جهانی از مدیریت داده‌های پژوهشی، که داده‌های زبانی یکی از انواع آن‌هاست، پرداخته می‌شود و در ادامهٔ آن نیز تجربیات حاصل از مدیریت داده‌های زبانی مورد توجه قرار می‌گیرد.

۲. مدیریت داده‌های پژوهشی

مقصود از دادهٔ پژوهشی داده‌ای است که به منظور تحلیل برای ارائهٔ نتایج اصیل پژوهشی به صورت رقومی جمع‌آوری، مشاهده و یا تولید می‌شود (دانشگاه ادینبرو[۲]،۲۰۱۵). از داده‌های پژوهشی برای انجام تحقیقات علمی در رشته‌های مختلف علمی استفاده می‌شود. برای نمونه، یک متخصص تغذیه که می‌خواهد راجع به تاثیر مصرف یک مادهٔ خوراکی بر سلامتی افراد تحقیق کند، اطلاعاتی مانند رژیم غذایی افراد، وزن آن‌ها، سنشان و . . . را جمع‌آوری می‌کند؛ کارشناس محیط زیست که در خصوص تغییرات آب و هوایی مطالعه می‌کند نیاز به اطلاعاتی راجع به وضعیت اقلیمی طی سال‌های گذشته دارد؛ کتابداری که قصد دارد دربارهٔ رفتارهای اطلاع‌یابی دانشجویان پژوهش کند نیاز دارد که بدین منظور داده‌هایی را از طریق پرسش‌نامه‌هایی به دست آورد؛ و بالاخره زبان‌شناسی که یک فرایند زبان‌شناختی را مطالعه می‌کند می‌تواند با جمع‌آوری نمونه‌های فراوان از کاربرد فرایند مورد نظر در متون مختلف آن را بررسی کند. تمام این موارد منجر به تولید داده‌هایی می‌شود که پژوهشگران بر اساس آن‌ها دست به تحلیل و نتیجه‌گیری علمی می‌زنند.

---

[۱] Manning, C. D.
[۲] University of Edinburgh

داده‌های پژوهشی را می‌توان با توجه به اینکه محقق خود آن‌ها را به صورت دست اوّل و از طریق مشاهده، پرسش‌نامه، مصاحبه و. . . بدست می‌آورد، و یا اینکه از منابع دیگر مانند تحقیقات گذشته یا آمار و اسناد سازمانی به دست می‌آیند به ترتیب «اوّلیه» و «ثانوی» نامید (خاکی، ۱۳۷۸). در مبحث مدیریت داده‌های پژوهشی بر اهمیت بهبود فرایند تولید، مستندسازی، نگهداری، و به اشتراک‌گذاری داده‌های پژوهشی اوّلیه و ضرورت دسترسی هرچه بیشتر و بهتر پژوهشگران به داده‌های پژوهشی ثانوی تأکید می‌شود. در واقع موضوع مدیریت داده‌های پژوهشی زمانی بیشتر اهمیت پیدا می‌کند که تولیدکنندهٔ داده قصد داشته باشد آن را جهت بهره‌برداری سایر پژوهشگران به اشتراک بگذارد. مزایای متعددی را می‌توان برای به اشتراک گذاشتن چنین داده‌هایی برشمرد که از آن میان می‌توان به موارد زیر اشاره کرد (دانشگاه کمبریج[۱]،۲۰۱۵):

– امکان‌پذیر کردن اعتبارسنجی مستقل نتایج: تنها هنگامی که داده‌های پژوهش در دسترس دیگران باشد ارزیابی‌های متقن امکان‌پذیر خواهند بود و عددسازی‌ها و نمودارسازی‌های احتمالی را می‌توان تشخیص داد.

– گسترده کردن دامنهٔ تأثیر و دیده‌شدن پژوهش: اشتراک‌گذاری داده‌های پژوهشی در کنار اشتراک‌گذاری نتایج تحقیقات می‌تواند موجب تأثیرگذاری بیشتر پژوهش‌ها و ارجاع بیشتر به آن‌ها شود.

– استفادهٔ بهینه از سرمایه‌گذاری‌ها با جلوگیری از دوباره‌کاری: داده‌های پژوهشی منابع ارزشمندی هستند که اغلب با صرف زمان زیاد و هزینهٔ بالا تولید می‌شوند (ایندن و همکاران، ۲۰۱۱). با عرضهٔ داده‌های پژوهشی به جامعهٔ علمی دیگر شاهد صرف هزینه و زمان برای تولید داده‌های مشابه و تکراری نخواهیم بود.

– ایجاد فرصت برای همکاری‌ها و مشارکت‌های جدید: داده‌های پژوهشی انتشاریافته می‌توانند محملی برای همکاری‌های علمی میان محققان باشند.

– پیشبرد پژوهش هنگامی که داده‌ها به اشکال جدید و نوآورانه‌ای ترکیب می‌شوند: تجمیع داده‌های پژوهشی مختلف می‌تواند موجب هم‌افزایی و دستیابی به نتایج علمی نوآورانه شود. همچنین بسیاری

[۱] University of Cambridge

از داده‌های پژوهشی را می‌توان برای کاربردهایی غیر از کاربرد اصلی که برای آن تولید شده‌اند مورد استفاده قرار داد.

مدیریت داده‌های پژوهشی امروزه در بسیاری از کشورها مورد توجه دولت‌ها، مؤسسات پژوهشی و پژوهشگران قرار گرفته است. نهادهای سیاست‌گذار علم و فناوری در برخی کشورها با راهکارهایی پژوهشگران را ملزم به ارائه «برنامهٔ مدیریت داده[۱]» می‌کنند؛ برای مثال بنیاد ملی علوم[۲] آمریکا، شوراهای پژوهشی انگلستان[۳]، بنیاد پژوهش آلمان[٤]، و ... که از اصلی‌ترین نهادهای تأمین‌کنندهٔ بودجه‌های پژوهشی در این کشورها هستند، دریافت‌کنندگان پژوهانه را ملزم به اشتراک‌گذاری داده‌های پژوهشی و ارائهٔ برنامه مدیریت داده می‌کنند. معمولاً در یک برنامهٔ مدیریت داده به پرسش‌های زیر پاسخ داده می‌شود (ایندن و همکاران، ۲۰۱۱):

چه داده‌ای حین پژوهش تولید خواهد شد؟ چه اقداماتی برای تضمین کیفیت داده و فرادادها انجام خواهد گرفت؟ چه برنامه‌ای برای اشتراک‌گذاری داده وجود دارد؟ پژوهشگر برای اشتراک‌گذاری داده با چه مسائل یا محدودیت‌های اخلاقی و قانونی مواجه است؟ مالکیت معنوی داده متعلق به کیست (و کاربران با چه نوع مجوزی می‌توانند از آن استفاده کنند)؟ ذخیره‌سازی و پشتیبان‌گیری از داده به چه شکل خواهد بود؟ چه افرادی در فرایند تولید داده دخیل‌اند و هر کدام چه وظایفی دارند؟ تولید داده چه هزینه‌هایی دارد و چه منابعی مورد نیاز است؟

به خاطر جدید بودن موضوع مدیریت داده‌های پژوهشی، بسیاری از کتابخانه‌های دانشگاهی در این زمینه خدمات ترویجی، آموزشی و مشاوره‌ای ارائه می‌کنند. برای مثال مؤسسه فناوری ماساچوست، دانشگاه کمبریج، کالج سلطنتی لندن، دانشگاه هاروارد، دانشگاه آکسفورد، کالج دانشگاهی لندن، دانشگاه استنفورد،

---

[۱] Data Management Plan

[۲] National Science Foundation (http://www.nsf.gov/bfa/dias/policy/dmp.jsp)

[۳] Research Councils UK (http://www.rcuk.ac.uk/research/datapolicy)

[٤] German Research Foundation (http://dfg.de/en/research_funding/proposal_review_decision/applicants/submitting_proposal/reusing_research_data)

مؤسسه فناوری کالیفرنیا، دانشگاه پرینستون و دانشگاه ییل که بر اساس رتبه‌بندی ۲۰۱۵/۲۰۱٤ کیو. اس.[۱]، ده دانشگاه برتر جهان هستند، همگی در کتابخانه و یا دفاتر پژوهش خود خدماتی را در زمینه مدیریت داده‌های پژوهشی ارائه می‌کنند. علاوه بر کتابخانه‌های دانشگاه‌ها، مراکز مستقلی نیز هستند که ترویج و آموزش مدیریت داده‌های پژوهشی را در دستور کار خود دارند که شناخته‌شده‌ترین آن‌ها عبارتند از مرکز سازماندهی دیجیتال[۲] و اتحادیهٔ داده‌های پژوهشی[۳].

از سوی دیگر، سازمان‌های انتشاراتی معتبر به انتشار مجلات داده‌ای[٤] روی آورده‌اند. از نمونه این مجلات می‌توان به «داده‌های خلاصه[٥]» از انتشارات الزویر[٦] و «داده علمی[۷]» از گروه انتشاراتی نیچر[۸] اشاره کرد. چنین نشریاتی بر توصیف داده‌های علمی در حوزه‌های مختلف تمرکز دارند و یکی از شرایط اصلی پذیرش مقاله در آن‌ها ارائهٔ شدن داده‌های مرتبط با مقاله در یک پایگاه معتبر اشتراک داده است. وبگاه «فهرست مخازن داده‌های پژوهشی[۹]» بیش از ۱۳۰۰ پایگاه اشتراک داده را از ٤۸ کشور و ۱۵۲ حوزهٔ تخصصی معرفی کرده است. این فهرست به محققان کمک می‌کند که پایگاه مناسبی را برای اشتراک‌گذاری داده پژوهشی خود انتخاب کنند. در خصوص داده‌های زبانی نیز مراکز مختلفی در سراسر جهان هست که پایگاه‌های تخصصی برای اشتراک‌گذاری چنین داده‌های ایجاد کرده‌اند. دو مورد از مهم‌ترین این مراکز عبارتند از کنسرسیوم داده‌های زبانی[۱۰] (تأسیس ۱۹۹۲) و انجمن اروپایی منابع زبانی[۱۱] (تأسیس: ۱۹۹۵).

---

[۱] QS World University Rankings (http://www.topuniversities.com/university-rankings/world-university-rankings/۲۰۱٤)
[۲] Digital Curation Centre (http://www.dcc.ac.uk)
[۳] Research Data Alliance (https://rd-alliance.org)
[٤] data journals
[٥] Data in Brief (http://www.journals.elsevier.com/data-in-brief)
[٦] Elsevier
[۷] Scientific Data (http://www.nature.com/sdata)
[۸] Nature
[۹] Registry of Research Data Repositories (http://www.re۳data.org)
[۱۰] Linguistic Data Consortium (http://www.ldc.upenn.edu)
[۱۱] European Language Resources Association (http://www.elra.info)

۳. مرجع دادگان زبان فارسی

مرجع دادگان زبان فارسی در اسفندماه سال ۱۳۹۱ به عنوان بستری برای اشتراک‌گذاری داده‌های زبان فارسی راه‌اندازی شد. در این پایگاه اینترنتی که با آدرس www.dadegan.ir در دسترس عموم کاربران وب است، هم‌اکنون تعداد ٤۳ داده معرفی شده است و تقریباً به همین تعداد داده‌های دیگر نیز شناسایی شده است که مراحل آماده‌سازی برای انتشار در مرجع را طی می‌کنند. جدول ۱ آمار داده‌های منتشرشده در مرجع دادگان را ارائه می‌کند:

جدول ۱ – آمار مربوط به داده‌های منتشرشده در مرجع دادگان

داده‌های متنی	داده‌های صوتی	داده‌های تصویری	تعداد کل داده‌ها
۳۳	۸	۲	٤۳

همانطور که در جدول فوق ملاحظه می‌شود اکثر داده‌های معرفی‌شده در مرجع دادگان داده‌های متنی هستند. از این میان ۲۳ مورد پیکره[۱] و ۱۰ مورد واژگان[۲] هستند. پیکره متنی عبارت است از مجموعهٔ بزرگی از متون که معمولاً به نحوی جمع‌آوری می‌شود که توازن دقیقی از متون یک یا چند ژانر را شامل باشد. واژگان یا منبع واژگانی[۳] مجموعه‌ای از کلمات و /یا عبارات است که به همراه اطلاعات مربوط به آن‌ها (مانند برچسب اجزای سخن[٤]) فهرست می‌شوند (برد[٥]، ۲۰۰۹).

داده‌های صوتی عبارتند از مکالمات ضبط‌شده به همراه پیاده‌سازی متن آن‌ها که عمدتاً در سامانه‌های تبدیل متن به گفتار و گفتار به متن مورد استفاده قرار می‌گیرند و منظور از داده‌های تصویری داده‌هایی است که

[۱] corpus
[۲] lexicon
[۳] lexical resource
[٤] part of speech tag
[٥] Bird, S.

برای فعالیت‌هایی چون نویسه‌خوانی نوری[1] مورد استفاده قرار می‌گیرند و مشتمل بر تصاویری از متون چاپی و یا دست‌نویس هستند.

برای معرفی هر یک از داده‌ها در مرجع دادگان زبان فارسی، اعم از متنی، صوتی و تصویری، یک صفحه مجزا با یک آدرس وب منحصر به فرد اختصاص داده می‌شود. در ادامه ضمن معرفی اطلاعات موجود در صفحات معرفی داده، به نکاتی پرداخته می‌شود که حین تهیهٔ دادهٔ زبانی باید به آن‌ها توجه شود. این نکات می‌تواند علاوه بر داده‌های زبانی به مدیریت انواع دادهٔ پژوهشی تعمیم داده شود. هر صفحهٔ معرفی داده شامل بخش‌های زیر است:

عنوان داده: هر داده دارای عنوانی است که در ابتدای صفحه درج می‌شود. خوب است که برای داده‌های زبانی تولیدشده از ابتدای انتشار نامی مناسب انتخاب شود که کاربران هنگام ارجاع به داده به صورت یک‌دست از آن استفاده کنند. در انتخاب نام پیکره باید به نکاتی از قبیل گویا بودن، کوتاه بون، بیش از اندازه عام نبودن و . . . توجه شود.

کد داده: هر داده‌ای که در مرجع دادگان زبان فارسی ثبت می‌شود دارای کدی منحصر به فرد است که در صفحه معرفی داده نمایش داده می‌شود. کاربرد اصلی این کدها در مرجع دادگان، آرشیو کردن داده‌ها و اطلاعات و مدارک مربوط به آن‌ها است. همچنین در مورد برخی صفحات که عنوان پیکره برای استفاده در آدرس صفحه مناسب نیست، از کد داده برای این منظور استفاده می‌شود. در کنار این کد تاریخ ثبت داده در مرجع و تعداد بازدید از صفحهٔ داده نیز نمایش داده می‌شود.

معرفی داده: برای معرفی هر داده توصیف مختصری به اندازهٔ یک پاراگراف کوتاه در بالای صفحهٔ معرفی داده ارائه می‌شود.

مالکیت معنوی و شرایط استفاده: یکی دیگر از اطلاعاتی که در صفحهٔ معرفی داده‌ها ارائه می‌شود این است که مالکیت معنوی دادهٔ زبانی تولیدشده متعلق به چه شخص و یا سازمانی است و چه محدودیت‌هایی برای استفاده از داده وجود دارد. داده‌ها را می‌توان با محدودیت‌هایی چون استفاده تنها برای اهداف پژوهشی، غیرتجاری یا غیرنظامی و استفاده تنها توسط فرد یا واحد سازمانی دریافت‌کنندهٔ داده و عدم امکان بازنشر

[1] optical character recognition

داده عرضه کرد و یا آزادی بیشتری به کاربران داد و اجازهٔ هرگونه بهره‌برداری از داده‌ها را به آن‌ها داد. یک نکتهٔ قابل توجه این است که هر چه محدودیت‌های کمتری برای استفادهٔ از داده‌ها وضع شود، دامنهٔ کاربرد داده‌ها وسیع‌تر خواهد بود و این همان موضوعی است که طرفداران ایدهٔ «داده‌های باز»[۱] بر آن تأکید دارند. بسیاری از تولیدکنندگان داده‌های زبانی از اجازه‌نامه‌های[۲] آماده‌ای استفاده می‌کنند که درجات مختلفی از آزادی را به کاربران می‌دهند، از آن میان می‌توان به اجازه‌نامه‌های گنو[۳]، کریتیو کامانز[٤]، ام‌آی‌تی[٥]، ال‌جی‌پی‌ال‌آر[٦] و ... اشاره کرد.

اطلاعات ارجاع: هنگامی که داده‌های زبانی به عنوان یک منبع علمی مورد استفاده قرار می‌گیرند باید همچون سایر منابع به شکلی شایسته به آنها ارجاع داد. امروزه رایج‌ترین شیوهٔ ارجاع به داده‌های زبانی در جامعه علمی زبان‌شناسی و پردازش زبان، ارجاع به مقاله، کتاب و یا گزارشی است که دادهٔ مورد نظر را توصیف کرده باشد. خوب است به آدرس اینترنتی و زمانی که داده دریافت شده است نیز اشاره شود. همچنین داده‌های ثبت‌شده در برخی پایگاه‌های دارای شناساگر شیء دیجیتال[۷]هستند که می‌توان هنگام ارجاع از این کد نیز استفاده کرد.

مستندات: همانطور که در بخش اطلاعات ارجاع توضیح داده شد، معمولاً داده‌های زبانی در یک مقاله، کتاب و یا گزارش توصیف می‌شوند. مستندات توصیف‌کنندهٔ داده‌های زبانی عمدتاً شامل اطلاعاتی چون منابع تولید، روش‌های تولید، ساختار دقیق داده‌ها، ارزیابی داده‌ها و آمار داده‌ها است. در زیر به هر یک از این موارد می‌پردازیم:

– **منابع تولید:** انتخاب منابع اولیهٔ مناسب برای تولید داده‌های زبانی از مسائل مهمی است که پژوهشگر، مخصوصاً اگر قصد اشتراک‌گذاری داده‌ها را دارد، باید از ابتدا به آن توجه نماید. منابعی چون کتاب‌ها و

---

[۱] open data
[۲] license
[۳] GNU (http://www.gnu.org/licenses/)
[٤] Creative Commons (http://creativecommons.org/)
[٥] MIT (https://opensource.org/licenses/MIT)
[٦] Lesser General Public License For Linguistic Resources (http://www.iran-inde.cnrs.fr/IMG/html/LGPLLR.html)
[۷] Digital Object Identifier (DOI)

مقالات، صفحات اینترنتی، زیرنویس فیلم‌ها، سخنرانی‌ها، مکالمات تلفنی، و ... می‌توانند برای تهیه داده‌های زبانی استفاده شوند. برای نمونه در مستندات پیکرهٔ استاندارد سامانه‌های خلاصه‌ساز (پاسخ)[۱] آمده است که منابع اولیه مورد استفاده برای جمع‌آوری متون این پیکره عبارتند از اخبار با موضوعات مختلف از ۷ خبرگزاری داخلی و یا برای ایجاد مجموعهٔ فارس‌دات[۲] ۳۰۰ فارسی‌زبان از مناطق مختلف جملاتی را ادا کرده‌اند و صدای آن‌ها ضبط شده است. یکی از نکاتی که هنگام انتخاب منابع باید به آن توجه داشت مسئلهٔ حقوق مالکیت معنوی است. به‌کارگیری و انتشار اسناد موجود در بعضی منابع مستلزم کسب مجوز از صاحبان آن‌هاست. از سوی دیگر برخی منابع حاوی اطلاعات خصوصی اشخاص و یا مطالبی است که دارای ملاحظات امنیتی هستند. داده‌هایی که با استفاده از چنین منابعی تهیه شده باشند هنگام انتشار با مشکل مواجه خواهند شد. در برخی مواقع می‌توان با پیشنهاداتی چون تقطیع کردن سندهای موجود در منابع اولیه و استفاده از آن‌ها در داده زبانی بدون رعایت ترتیب اولیه و یا حذف اطلاعات شخصی از داخل سندها، رضایت صاحبان آثار منابع اولیه را جهت استفاده از آثارشان در منابع زبانی به دست آورد. همچنین در مواردی که داده‌های زبانی با روش‌هایی چون مصاحبه جمع‌آوری می‌شود باید رضایت افرد را برای انتشار مکالمات حاصل از مصاحبه به صورت داده‌های زبانی جلب کرد.

– **روش تولید:** لازم است پژوهشگران روش خود را در تولید داده‌های زبانی به صورت کامل تشریح کنند. امروزه برای تولید انواع داده‌های زبانی استانداردها و رویه‌های جاافتاده‌ای وجود دارد که می‌توان با مرور پژوهش‌های پیشین در حوزهٔ مورد نظر با این استانداردها و رویه‌ها آشنا شد. محققی که قصد دارد یک دادهٔ زبانی جدید تولید کند باید ضمن مشخص کردن اهداف و کاربردهای مدنظر خود، دلایل استفاده و یا عدم استفاده از روش‌های مرسوم پیشین و ابتکارات به کار رفته در ایجاد دادهٔ جدید را تشریح کند. مواردی که می‌توان در این بخش به آن‌ها پرداخت عبارتند از: نظریه زبانی مد نظر، اقدامات انجام گرفته برای آنکه دادهٔ تولیدشده به خوبی نمایندهٔ گونهٔ زبانی مورد مطالعه باشد، نرم‌افزارها و سخت‌افزارهای مورد استفاده برای تهیهٔ داده، نیروی انسانی دخیل در پروژهٔ تولید داده و . . . . برای نمونه برای تولید دادگان درختی فارسی

[۱] http://dadegan.ir/catalog/pasokh
[۲] http://dadegan.ir/catalog/farsdat

در چارچوب دستور ساخت سازه‌ای هسته‌بنیان[۱]، همانطور که در مقالهٔ مربوط به این پیکره توضیح داده شده است، از روش بوت‌استرپینگ[۲] استفاده می‌شود. در این روش ابتدا روابط نحوی درون تعدادی از جملات پیکره با تعداد محدودی قاعده تا حد امکان تعیین می‌شود، سپس پژوهشگر روابط نحوی را تکمیل می‌کند و قواعد پرتکرار حاصل از تکمیل روابط نحوی توسط پژوهشگر به مجموعهٔ قواعد اولیه اضافه شده و بخش دیگری از داده با مجموعهٔ جدید قواعد تعیین رابطه می‌شود و این کار ادامه پیدا می‌کند تا روابط نحوی در کل جملات دادهٔ مورد نظر تعیین شود.

– **ساختار داده‌ها:** داده‌های زبانی باید ساختارمند باشند. استفاده از ساختارهای استاندارد و مشترک باعث می‌شود که داده‌ها را بتوان با ابزار مشترک پردازش کرد و مورد تحلیل قرار داد. برای مثال پیکره وابستگی نحوی زبان فارسی[۳] بر اساس ساختار همایش یادگیری زبان طبیعی[٤]در سال ۲۰۰٦ و یا پیکره متنی زبان فارسی[٥] بر مبنای استاندارد ایگلز[٦] تولید شده است. جهت روشن شدن موضوع، قالب داده‌های وابستگی که در همایش یادگیری زبان طبیعی سال ۲۰۰٦ ارائه شده است در ادامه شرح داده می‌شود. در این قالب، داده شامل جملاتی است که هر کدام با یک سطر خالی از هم جدا شده‌اند. هر جمله شامل یکیا چند واژه است که هر یک در سطری مجزا قرار می‌گیرند. هر سطر شامل ۱۰ فیلد است که با یک کارکتر تب[۷] از هم جدا شده‌اند و استفاده از فاصله درون فیلدها مجاز نیست. فیلدهای هر سطر در قالب همایش یادگیری زبان طبیعی به ترتیب از چپ به راست حاوی اطلاعات زیراست:

۱. شماره واژه که در هر جمله از ۱ شروع می‌شود
۲. صورت واژه (مثال: حاضران، اجرای، برخاستند)
۳. ریشه واژه (مثال: حاضر، اجرا، بر#خاست#خیز)

---

[۱] http://dadegan.ir/catalog/pertreebank
[۲] bootstrapping
[۳] http://dadegan.ir/catalog/perdt
[٤] Conference on Natural Language Learning (CONLL)
[٥] http://dadegan.ir/catalog/matni
[٦] EAGLES (Expert Advisory Group on Language Engineering Standards)
[۷] tab

٤. برچسب اجزای سخن درشت[۱] (مثال: اسم، اسم، فعل)

۵. برچسب اجزای سخن ریز[۲] (مثال: جاندار، بی‌جان، معلوم)

٦. مجموعه‌ای از ویژگی‌های نحوی، صرفی و . . . که با کاراکتر | از هم جدا شده‌اند (ویژگی‌های همچون شمار، شخص، زمان/وجه/نمود افعال، . . .)

۷. شماره والد (واژه‌ای که واژه مورد نظر وابسته به آن است)

۸. نوع رابطه وابستگی با والد (فاعل، مفعول، مسند، . . .)

۹ و ۱۰. نوع خاصی از رابطه وابستگی با والد که در بسیاری از پیکره‌ها از جمله پیکره وابستگی زبان فارسی مورد استفاده قرار نگرفته است و لذا این فیلدها در چنین پیکره‌هایی با خط تیره (–) پر شده‌اند.

– **آمار داده‌ها:** اطلاعاتی چون تعداد کلمات یا جملات موجود در دادهٔ زبانی، میزان وقوع ساخت‌های مشخص و . . . معمولاً در مستندات مربوط به داده ارائه می‌شود. برای مثال در مقالهٔمربوط به پیکره موازی انگلیسی–فارسی تهران[۳]، آمار مربوط به تعداد کلمات، تعداد حروف، میانگین طول جملات، تعداد واژه‌های منحصر به فرد و تعداد خطوط پیکره در قالب جدول و اطلاعات مربوط به توزیع جملات انگلیسی و فارسی بر اساس طول آن‌ها به صورت نمودار ارائه شده است. همچنین در داده‌های مورد استفاده برای پردازش زبان که معمولاً قسمتی از داده به عنوان مجموعهٔ آموزش[٤] و بخشی به عنوان مجموعهٔ آزمایش[۵] ارائه می‌شود، آمار مربوط به این تقسیم‌بندی نیز ثبت می‌شود.

– **ارزیابی داده:** معمولا برای نشان دادن کارایی داده، در پایان معرفی داده، به نتایج به دست آمده از پردازش یا تحلیل بر اساس دادهٔ تولیدشده پرداخته می‌شود. برای مثال اگر یک درخت‌بانک[٦] نحوی تهیه شود، یک تجزیه‌گر[۷] نحوی بر اساس دادهٔ تهیه‌شده آموزش داده می‌شود و دقت تجزیه‌گر گزارش می‌شود. از روش‌های دیگری نیز می‌توان برای ارزیابی داده استفاده کرد. برای مثال برای ارزیابی پایگاه داده گفتار احساسی زبان

---

[۱] coarse-grained part-of-speech tag
[۲] fine-grained part-of-speech tag
[۳] http://dadegan.ir/catalog/tep
[٤] training set
[۵] test set
[٦] treebank
[۷] parser

فارسی[۱] که مجموعه‌ای است از جملات بیان‌شده با احساسات مختلف (عصبانیت، شادی، غم، ترس، چندش)، جملات صوتی ضبط‌شده در اختیار ۳٤ فارسی‌زبان قرار گرفته است تا در خصوص احساسات موجود در آن‌ها اظهار نظر کنند و نهایتاً جملاتی که با درصد بالایی درست تشخیص داده شده‌اند در مجموعه قرار گرفته‌اند. یکی از روش‌هایی که برای ارزیابی کیفیت داده‌ها استفاده می‌شود بررسی توافق میان برچسب‌زن‌ها[۲] است. «توافق میان برچسب‌زن‌ها» معیاری است که در طی تهیهٔ داده‌های بزرگ که نیاز به چند نفر برچسب‌زن دارد مورد استفاده قرار می‌گیرد. برای اندازه‌گیری توافق میان برچسب‌زن‌ها، بخشی از داده به بیش از یک برچسب‌زن داده می‌شود و میزان هماهنگی میان برچسب‌زن‌ها بررسی می‌شود. پایین بودن توافق میان برچسب‌زن‌ها ممکن است معانی متفاوتی داشته باشد از جمله اینکه این که برچسب‌های تعریف‌شده باید بازبینی شوند و یا اینکه دستورالعمل‌ها باید بازنویسی شوند. مدیر پروژه پس از انجام اصلاحات لازم، پس از مدتی مجدداً توافق میان برچسب‌زن‌ها را مقایسه کرده و در صورت نیاز باز هم تغییراتی در روال کار ایجاد می‌کند. در پایان پروژه نیز توافق نهایی میان برچسب‌زن‌ها بررسی شده و گزارش می‌شود.

اطلاعات تکمیلی: در این بخش اطلاعات تکمیلی که لازم است کاربر راجع به داده بداند ارائه می‌شود. برای مثال اگر توضیحات تفصیلی مربوط به داده در یک وبگاه دیگر موجود است، آدرس آن وبگاه در این قسمت نمایش داده می‌شود و یا اگر مرورگر برخطی برای داده ایجاد شده است پیوند مربوط به آن ارائه می‌شود. برای نمونه «سامانهٔ جستجوی دادگان»[۳] نام یک مرورگر برخط است که بازنمایی گرافیکی بسیار خوبی از اطلاعات موجود در فرهنگ ظرفیت نحوی افعال فارسی[٤] و پیکرهٔ وابستگی نحوی زبان فارسی ارائه می‌کند. این سامانه در بخش اطلاعات تکمیلی صفحات معرفی دو دادهٔ نامبرده معرفی می‌شود.

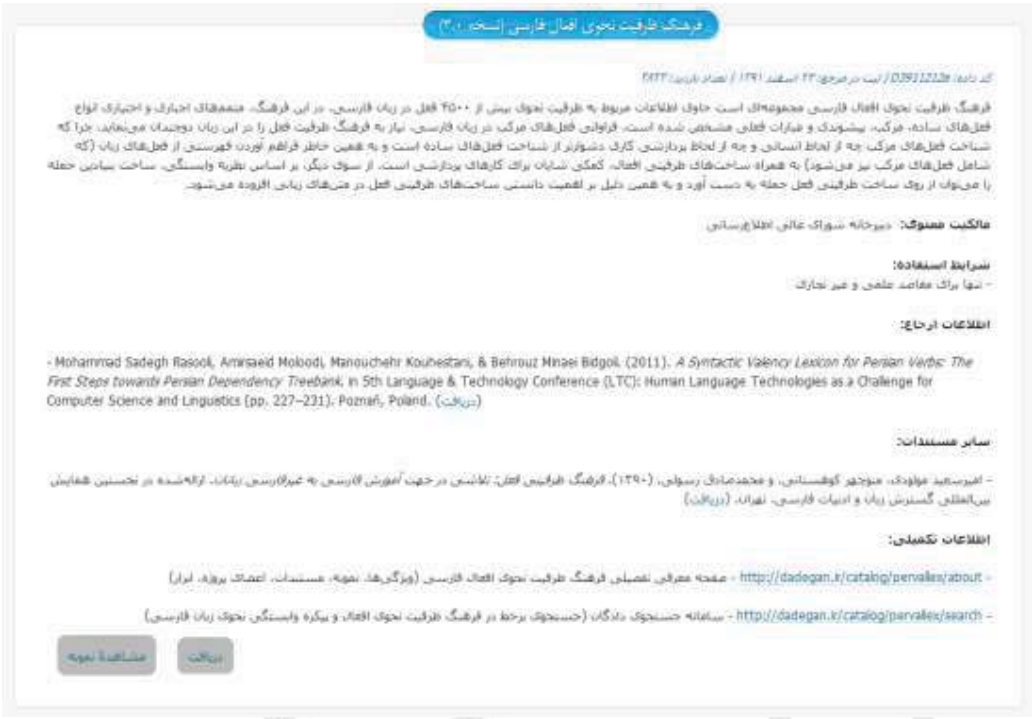
مشاهدهٔ نمونه: این گزینه به کاربران اجازه می‌دهد که پیش از دریافت فایل داده بخشی از آن را ببینند و با ساختار آن آشنا شوند و بررسی کنند که آیا با داده‌ای که مد نظر آن‌هاست انطباق دارد یا خیر.

---

[۱] http://dadegan.ir/catalog/pesd
[۲] Inter-annotator agreement
[۳] http://dadegan.ir/catalog/pervallex/search
[٤] http://dadegan.ir/catalog/pervallex

دریافت داده: این گزینه امکان دریافت داده را برای کاربران فراهم می‌کند. دریافت داده در مرجع دادگان زبان فارسی به دو صورت کلی انجام می‌گیرد. یا فایل داده توسط تولیدکنندهٔ داده در اختیار مرجع دادگان قرار گرفته است و کاربران فایل داده را مستقیماً از مرجع دریافت می‌کنند (در برخی مواقع پس از تکمیل تفاهمنامهٔ کاربری) و یا اینکه کاربران برای دریافت داده به صفحات دریافت داده در وبگاه‌های دیگر هدایت می‌شوند. یک نکته که در خصوص فایل داده‌ها وجود دارد این است که چنین فایل‌هایی باید در قالب‌های قابل قبول ارائه شوند. برای مثال داده‌های متنی بهتر است در قالب‌های ماشین‌خوان مانند txtباشد و به هیچ وجه نباید از قالب‌هایی چون pdf برای این منظور استفاده کرد. از سوی دیگر پوشه‌بندی فایل باید به نحوی باشد که دسترسی به فایل‌ها به سهولت انجام پذیرد (پوشهٔ داده، پوشهٔ مستندات و پوشهٔ ابزار). همچنین نامگذاری فایل‌ها باید طبق روال مشخصی باشد (معنادار و یکدست) و از کارکترهای نامناسب در آن‌ها استفاده نشود. عدم رعایت این نکات ساده باعث تحمیل زحمت به کاربران داده‌ها می‌شود.

در شکل ۱ یکی از صفحات معرفی داده در مرجع دادگان زبان فارسی به عنوان نمونه ارائه شده است.


فرهنگ ظرفیت نحوی افعال فارسی (نسخه ۳.۰)

کد داده: fa/0391221b / ثبت در مرجع: ۲۳ اسفند ۱۳۹۱ / تعداد بازدید: ۵۴۲۲

فرهنگ ظرفیت نحوی افعال فارسی مجموعه‌ای است حاوی اطلاعات مربوط به ظرفیت نحوی بیش از ۴۵۰۰ فعل در زبان فارسی. در این فرهنگ، متمم‌های اجباری و اختیاری انواع فعل‌های ساده، مرکب، پیشوندی و عبارات فعلی مشخص شده است. قراردادی فعل‌های مرکب در زبان فارسی، نیاز به فرهنگ ظرفیت فعل را در این زبان دوچندان می‌نماید، چرا که شناخت فعل‌های مرکب چه از لحاظ انسانی و چه از لحاظ پردازشی کاری دشوارتر از شناخت فعل‌های ساده است و به همین خاطر فراهم آوردن فهرستی از فعل‌های زبان (که شامل فعل‌های مرکب نیز می‌شود) به همراه ساخت‌های ظرفیتی افعال، کمکی شایان برای کارهای پردازشی است. از سوی دیگر، بر اساس نظریهٔ وابستگی، ساخت بنیادین جمله را می‌توان از روی ساخت ظرفیتی فعل جمله به دست آورد و به همین دلیل بر اهمیت دانستن ساخت‌های ظرفیتی فعل در متن‌های زبانی افزوده می‌شود.

مالکیت محتوی: دبیرخانه شورای عالی اطلاع‌رسانی

شرایط استفاده:
- تنها برای مقاصد علمی و غیر تجاری

اطلاعات ارجاع:

- Mohammad Sadegh Rasooli, Amirsaeid Moloodi, Manouchehr Kouhestani, & Behrouz Minaei Bidgoli. (2011). *A Syntactic Valency Lexicon for Persian Verbs: The First Steps towards Persian Dependency Treebank*. In 5th Language & Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics (pp. 227–231). Poznań, Poland. (دریافت)

سایر مستندات:

- امیرسعید مولودی، موجهر کوهستانی، و محمدصادق رسولی. (۱۳۹۰). فرهنگ ظرفیتی افعال: شناسی در جهت آموزش فارسی به غیرفارسی زبانان. ارائه‌شده در نخستین همایش بین‌المللی گسترش زبان و ادبیات فارسی. تهران. (دریافت)

اطلاعات تکمیلی:

- http://dadegan.ir/catalog/pervallex/about - صفحه معرفی تفصیلی فرهنگ ظرفیت نحوی افعال فارسی (ویژگی‌ها، نمونه، مستندات، اعضای پروژه، ابزار)

- http://dadegan.ir/catalog/pervallex/search - سامانه جستجوی داده‌گان (جستجوی پرخط در فرهنگ ظرفیت نحوی افعال و پیکره وابستگی نحوی زبان فارسی)

دریافت مشاهدهٔ نمونه

شکل ۱: نمونه‌ای از صفحات معرفی داده در مرجع دادگان زبان فارسی (صفحه معرفی فرهنگ ظرفیت نحوی افعال فارسی)

٤. نتیجه‌گیری و کارهای آینده

نکاتی چون اشتراک‌گذاری داده‌های زبانی در پایگاه‌های اشتراک دادهٔ معتبر، انتخاب عنوان مناسب برای داده، مشخص کردن وضعیت مالکیت معنوی و شرایط استفادهٔ داده، تهیهٔ مستندات مناسب برای توصیف داده و مشخص کردن مستندی که کاربران باید هنگام استفاده از داده با آن ارجاع دهند، استفاده از منابع اولیه مناسب برای تولید داده، استفاده از روش علمی و قابل دفاع برای تولید داده، ساختاربندی داده بر مبنای قالب‌های استاندارد و مشترک، کسب اطمینان از کیفیت داده با استفاده از روش‌های مختلف ارزیابی، ایجاد ابزارهای مناسب برای پردازش و نمایش داده، و استفاده از ساختارهای مناسب برای فایل داده از جمله

پیشنهاداتی هستند که برای مدیریت داده‌های زبانی در این مقاله ارائه شده است. این نکات می‌توانند در مدیریت انواع دیگر داده‌های پژوهشی نیز مفید واقع شوند. داشتن برنامه‌ریزی برای هر یک از موارد فوق از ابتدای راه، همان چیزی که از آن به عنوان برنامهٔ مدیریت داده یاد می‌شود، می‌تواند به پژوهشگر در انجام پروژهٔ تهیه داده کمک کند.

در خصوص مرجع دادگان زبان فارسی بد نیست به این نکته اشاره شود که اکنون با رونق گرفتن مباحثی چون مدیریت داده‌های پژوهشی و داده‌های باز، پلت‌فرم‌های[۱] مخصوص برای به اشتراک‌گذاری داده ایجاد شده است که از آن میان می‌توان به پلت‌فرم متن‌باز سی‌کن[۲] اشاره کرد. بازسازی مرجع دادگان زبان فارسی بر اساس چنین پلت‌فرمی از اقدامات ارزشمندی است که می‌توان برای ارتقای آن انجام داد.

موضوع مدیریت داده‌های پژوهشی، با تعریفی که ارائه شد، علی‌رغمی اهمیت فراوان از موضوعات مغفول در جامعهٔ علمی ایران است. جا دارد که متخصصان رشته‌هایی چون علم اطلاعات و دانش‌شناسی (کتابداری و اطلاع‌رسانی) و سیاست‌گذاری علم و فناوری به مسائل این حوزه بپردازند و دانشگاه‌ها، پژوهشگاه‌ها و مراکزی چون کتابخانهٔ ملی، ایرانداک و مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری بسترهای لازم برای اشتراک‌گذاری داده‌های پژوهشی را فراهم آورند.

۵. تقدیر و تشکر

در پایان جا دارد از جناب آقای دکتر مهدی بهنیافر، معاون محترم مرکز تحقیقات کامپیوتری علوم اسلامی (نور) به خاطر حمایت‌های مدیریتی در راه‌اندازی مرجع دادگان زبان فارسی تشکر کنم. ضمناً مراتب سپاس و قدردانی خود را از تمامی فراهم‌کنندگان داده و کاربران گرامی مرجع دادگان زبان فارسی که طی سال‌های اخیر با نظرات ارزشمند و ابراز محبت‌های خود به پیشرفت مرجع دادگان کمک کرده‌اند اعلام می‌دارم.

٦. فهرست منابع

[۱] platform
[۲] CKAN (Comprehensive Knowledge Archive Network)

حافظ‌نیا، محمدرضا. (۱۳۹۳). *مقدمه‌ای بر روش تحقیق در علوم انسانی*؛ تهران: سمت.

خاکی، غلامرضا. (۱۳۷۸). *روش تحقیق با رویکردی به پایان‌نامه‌نویسی*؛ تهران: مرکز تحقیقات علمی کشور با همکاری کانون فرهنگی انتشاراتی درایت.

ملک‌افضلی، حسین؛ مجدزاده، سید رضا؛ فتوحی، اکبر؛ توکلی، سامان. (۱۳۸۳). *روش‌شناسی پژوهش‌های کاربردی در علوم پزشکی*؛ تهران: دانشگاه علوم پزشکی و خدمات بهداشتی درمانی تهران، معاونت پژوهشی.

Bird, S. , Klein, E. , & Loper, E. (۲۰۰۹). *Natural language processing with Python*. Sebastopol, CA: O'Reilly Media, Inc. .

Eynden, V. V. D. , Corti, L. , Woolard, M. , Bishop, L. , & Horton, L. (۲۰۱۱). *Managing and Sharing Data*; Colchester: UK Data Archive.

Jurafsky, D. , & Martin, J. H. (۲۰۰۸). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition and Computational Linguistics,* ۲*nd edition*. Upper Saddle River, NJ: Prentice-Hall.

Manning, C. D. , & Schütze, H. (۱۹۹۹). *Foundations of statistical natural language processing*. Cambridge, MA: MIT press.

McEnery, T. , & Wilson, A. (۲۰۰۱). *Corpus linguistics: An introduction*. Edinburgh: Edinburgh University Press.

University of Cambridge. (۲۰۱۵). Research Data Management: Data sharing and Open Access. Cambridge University Library. Retrieved from http://www.lib.cam.ac.uk/dataman/pages/sharing.html

University of Edinburgh. (۲۰۱۵). DataShare repository: Our definitions. University of Edinburgh Information Services. Retrieved from http://www.ed.ac.uk/information-services/research-support/data-library/data-repository/definitions