نشریه پژوهش‌های زبان‌شناسی تطبیقی
سال سیزدهم- شماره ۲۶- پاییز و زمستان ۱۴۰۲

استانداردسازی داده‌ها و اطلاعات گویشی: ضرورت و راهکار

مسعود قیومی[۱]

مقاله پژوهشی

چکیده

گویش‌شناسی به مطالعهٔ علمی یک گویش و توزیع جغرافیایی آن می‌پردازد. هر گویش یک زبان است و مطالعهٔ یک گویش به تحلیل‌های بسیار زیادی نیاز دارد که همین امر به طولانی‌شدن مدت انجام مطالعات مربوط به یک گویش می‌انجامد. گردآوری داده‌های گویشی بسیار پرزحمت و زمان‌بر است. از این‌رو، نیاز است این داده به‌گونه‌ای تهیه شود تا قابلیت استفادهٔ مجدد در بررسی‌های آتی را داشته باشد. دادهٔ خام کارایی چندانی در مطالعات گویش‌شناسی ندارد و نیاز است در چارچوب روش تحلیل زبان‌شناسی ساختگرایی، تحلیل‌های زبان‌شناختی به آن اضافه گردد. باتوجه‌به حجم زیاد داده گویشی در طرح‌هایی مانند تهیه اطلس گویش‌های یک کشور و تحلیل‌های زبان‌شناختی اضافه‌شده به آن، به ساماندهی دادهٔ گویشی نیاز است. استفاده از رایانه به‌عنوان یک ابزار موجب می‌شود داده براساس ساختار مشخصی انتظام یابد. هدف اصلی این پژوهش، معرفی یک استاندار برای سازماندهی داده‌ها و اطلاعات گویشی است. این استاندارد حاوی دادهٔ گویشی، فراداده‌های مربوط به آن گویش و همچنین اطلاعات زبان‌شناختی حاصل از تحلیل دادهٔ گویشی است. این اطلاعات براساس ساختار دادهٔ درختی و زبان نشانه‌گذاری گسترش‌پذیر سازماندهی می‌گردد. این ساختار داده، قابلیت جابه‌جایی داشته و می‌تواند به‌سادگی به درون یک پایگاه داده خوانده شود.

کلیدواژه‌ها: گویش‌شناسی، استانداردسازی، انتظام داده، گویش‌شناسی رایانشی، زبان نشانه‌گذاری.

۱- دانشیار پژوهشگاه علوم انسانی و مطالعات فرهنگی
m.ghayoomi@ihcs.ac.ir |✉

تاریخ دریافت: ۱۴۰۱/۰۲/۲۱ **تاریخ پذیرش:** ۱۴۰۱/۱۰/۱۷

شناسه دیجیتال (DOI): 10.22084/RJHLL.2023.26253.2221

۱- مقدمه

گویش‌شناسی به‌عنوان یکی از زیرمجموعه‌های زبان‌شناسی اجتماعی، به مطالعهٔ علمی یک گویش و توزیع جغرافیایی آن می‌پردازد. در گویش‌شناسی با یک زبان و ویژگی‌های متعلق به آن زبان، از نظر واژه، تلفظ، ساختار، معنا و کاربرد، سروکار داریم و این‌طور نیست که گویش به‌عنوان زبان گروهی از افراد متعلق به یک طبقهٔ اجتماعی تلقی گردد یا از تبدیل لهجهٔ معیار حاصل شده باشد (لاینز[1]، ۱۹۸۱: ۲۵). مطالعات گویش‌شناختی به شناخت عمیق آن گویش منجر شده و از این شناخت می‌توان در برنامه‌ریزی زبانی، تهیهٔ اطلس گویش‌های یک کشور، پژوهش‌های گویشی برای یافتن تشابه‌ها و تفاوت‌های بین گویش‌ها و همچنین در مطالعات مربوط به تعیین «مرز همگویی» (isoglosses) بهره برد. برای رسیدن به این شناخت نیاز است داده‌های گویشی مورد بررسی و مطالعه دقیق قرار گیرد. گردآوری دادهٔ گویشی کار بسیار پرزحمتی است و به یک روش‌شناسی دقیق و مشخص نیاز دارد. علاوه‌بر این روش‌شناسی، باید شیوه‌ای برای نگهداری، سازماندهی و مطالعه داده‌ها به‌خصوص در حجم زیاد داده انتخاب گردد. واینرایش[2] (۱۹۵۴) رویکرد جدیدی را در مطالعات گویشی مطرح کرده‌است که «گویش‌شناسی ساختگرا» (structural dialectology) نامیده می‌شود. در این رویکرد، از روش تحلیل زبان‌شناسی ساختگرایی برای تحلیل داده‌های گویشی استفاده می‌شود. رایانه به‌عنوان یک ابزار پژوهش می‌تواند علاوه بر سازماندهی داده، در تحلیل آن نیز به پژوهشگران کمک نماید. بر این اساس، داده‌های گویشی باید باتوجه‌به نظر واینرایش (۱۹۵۴) براساس یک استاندارد مشخص که مورد نیاز رایانه است انتظام یابد. هدف اصلی این مقاله، ارائهٔ یک قالب استاندارد است که می‌تواند در ساختارمندسازی داده‌های گویشی به‌کار رود و امکان تحلیل داده‌ها با کمک رایانه را فراهم آورد.

ساختار این مقاله به این گونه است که در بخش ۲، پیشینهٔ مطالعاتی در مورد سه مفهوم داده، اطلاعات و دانش توضیح داده می‌شود تا مشخص گردد منظور ما از شناخت در مورد یک گویش به کدامیک از این مفاهیم مرتبط است. علاوه‌بر داده‌هایی که مربوط به خود گویش است، اطلاعات جانبی نیز وجود دارد که به فراداده معروف است که ضمن معرفی آن، تفاوت آن با داده نیز مشخص می‌گردد. در بخش ۳، چارچوب کلی که برای ساختار معیار در داده مطرح است معرفی می‌شود. در بخش ۴، نحوهٔ کاربرد این چارچوب کلی جهت سازماندهی اطلاعات زبانشناختی در لایه‌های مختلف به‌طور فشرده معرفی می‌شود. ساختار پیشنهادی دادهٔ گویشی در بخش ۵ توضیح داده شده و مقاله با جمع‌بندی و نتیجه‌گیری در بخش ۶ پایان می‌پذیرد.

۲- ارتباط داده، اطلاعات و دانش

علم نوین بر پایه آزمایش، مشاهده و طرح نظریه استوار است تا با کمک آن بتوان به تبیین رویدادهای جهان پرداخت. بنابراین تبیین، شیوه‌ای برای پرده‌گشایی از دانش جدید است. شناخت علمی از رویدادهای جهان می‌تواند به دو صورت به‌دست آید: الف) براساس مشاهدهٔ واقعیت که به علم‌الیقین معروف بوده و بسیار

1. Lyon
2. Weinreich

پیچیده است و رسیدن به آن ممکن است زمان‌بر و غیرقابل انجام باشد؛ و ب) براساس احتمالات. در نهاد شناخت علمی احتمالاتی، استدلال استقرایی نهفته است. از این‌رو، برای ارائهٔ بهترین تبیین در مورد رویدادهای جهان شایسته است از شناخت علمی مبتنی‌بر احتمالات و نه مشاهدهٔ واقعیت (علم‌الیقین) کمک گرفته شود. براساس این رویکرد چنین می‌توان نتیجه گرفت که دانش ریاضی بر تحول علوم تأثیرگذار بوده‌است (اکاشا، ۱۳۸۷). از آنجا که محاسبات ریاضی شاکلهٔ رایانه را شکل می‌دهد؛ رایانه می‌تواند در تحول علمی سهیم باشد.

در چارچوب ساختگرایی و تعریف سوسور[1](۱۹۱۶) از زبان، دو سطح صورت (form) و معنی (meaning) مطرح شده‌است. این دو سطح یک واقعیت روانشناختی است که پیوندشان با یکدیگر، تمامِ جملاتِ زبانیِ بیان‌شده توسط گویشور یک زبان را در بر می‌گیرد. تجلی صورت می‌تواند به‌شکل یک زنجیرهٔ آوایی و یا نوشتاری دارای مفهوم باشد. صورت و معنی در یک نظام مشخص، واحدهای زبانی را ایجاد می‌کند که این واحدهای زبانی واژه نامیده می‌شود، براساس روابط همنشینی (associative relation) در چارچوب قواعد آن نظام مشخص، قابلیت ترکیب با یکدیگر وجود دارد تا واحدهای بزرگتر که جمله گفته می‌شود ساخته شود (ون مارلی[2]، ۲۰۰۸). در چارچوب این نظر و رویکرد نوین علم، به تعریف جایگاه سه مفهوم داده، اطلاعات و دانش می‌پردازیم.

منظور از دانش، آگاهی فرد از حقایق یا همان اطلاعات است که به‌صورت تجربی و یا ازطریق شناخت به‌دست می‌آید. بورگین[3] نظر ارسطو در مورد دانش و آگاهی افراد را به این صورت بیان کرده‌است که تمام انسان‌ها از طریق طبیعت دانش کسب می‌کنند (بورگین، ۲۰۱۷: ۱). فعالیت‌های مربوط به دانش در دو حوزهٔ اصلی تعریف می‌شود: الف) مطالعات نظری و عملی در مورد خود دانش؛ ب) فناوری دانش که به مهندسی دانش و کاربرد و مدیریت دانش می‌پردازد. در بخش مهندسی دانش با موضوعاتی مانند فناوری تولید دانش، سازماندهی آن، تبدیل، مدیریت، انتقال، به‌دست‌آوردن و اکتساب دانش سروکار داریم. در حالی‌که در بخش کاربرد دانش، چگونگی کاربرد دانش توسط افراد و یا سازمان‌ها و همچنین توسعهٔ روش‌های جدید کاربرد دانش مورد توجه است (بورگین، ۲۰۱۷: ۳)؛ به عقیده پوستر[4] (۱۹۹۰) داده‌هایی که در حوزهٔ فناوری دانش به‌کار برده می‌شود صورت‌های مختلف دارد که شامل داده‌های متنی، صوتی، تصویری و عددی است.

دانش که به تبیین علم می‌پردازد ذاتاً به داده و اطلاعات وابسته است. فراگیرشدن کاربرد رایانه در دنیای امروز و وجود اینترنت سبب شده‌است انواع داده‌های متنی، صوتی و تصویری به‌صورت الکترونیکی ارائه گردد. اگرچه ماهیت محتوای ارائه‌شده در دنیای رقمی امروز با دنیای غیررقمی دیروز یکسان است، نحوهٔ بازنمایی داده با گذشته متفاوت شده‌است. تفاوت در نحوهٔ ارائهٔ داده موجب شده‌است روش‌شناسی پژوهش دستخوش

1. de Saussure
2. Van Marle
3. Burgin
4. Poster

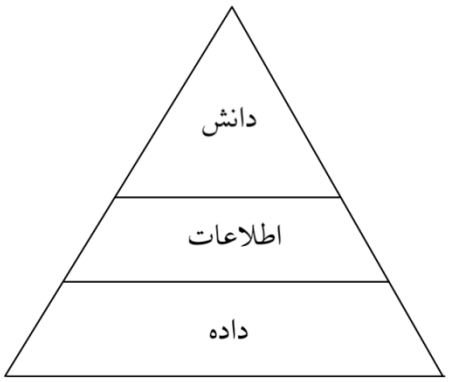
تغییر شود و تعریف افراد از مفاهیمی چون داده، اطلاعات و دانش متفاوت گردد. زینس[1] (۲۰۰۷) از بررسی آثار ۴۵ نویسنده، تعداد ۱۳۰ تعریف برای سه اصطلاح داده، اطلاعات و دانش را گردآوری کرده است که در ادامه مهم‌ترین تعاریف مربوط به توصیف این سه مفهوم مورد بررسی قرار می‌گیرد.

وقتی فردی در مورد اطلاعات صحبت می‌کند منظورش اطلاعات شناختی است (بورگین، ۲۰۰۱). مفهومی که از «اطلاعات» از قرن ۱۶ تاکنون در زبان‌های فرانسه، انکلیسی، اسپانیایی و ایتالیایی وجود دارد همان مفهوم خاصی است که ما امروزه از آن استفاده می‌کنیم (کاپورو[2]، ۱۹۹۱). با این حال، کاربرد علمیِ مفهومِ اطلاعات بیانگر این است که تعریف کلی‌تری از این معنای خاص مدنظر است، مانند هایدگر و فینک[3] (۱۹۷۰) که در حوزهٔ زیست‌شناسی، مفهوم کلی اطلاعات را در قالب اطلاعات خاص ارائه کرده‌اند. برای مثال، تداخل مفهوم کلی و خاص اطلاعات در شرایطی ظهور پیدا می‌کند که زیست‌شناس به‌طور کلی در مورد اطلاعات DNA بحث می‌کند و یا به‌طور خاص در مورد ژنتیک انسان. جملگی همه اطلاعات است.

تعریف نوتا[4](۱۹۷۰) از اطلاعات، هر چیز جدید است. اطلاعات قدیمی دیگر اطلاعات نیست؛ بنابراین «اطلاعات مقداری ناشناختگی، غیرمنتظره بودن، تعجب‌آور بودن یا اهمیت را [در خود] داراست». اُبرین[5] (۱۹۹۵) تفاوتی بین داده و اطلاعات قائل است به این صورت که منظور از داده، مواد خام است؛ و منظور از اطلاعات داده‌هایی است که به داده‌های پرمعنا و کاربردی در بافت تبدیل شده‌است. بنابراین، اطلاعات از نگاه وی مجموعه‌ای از داده‌های انتظام‌یافتهٔ قابل فهم است (لادون[6]، ۱۹۹۶). کوگلی و دبونز[7] (۱۹۹۹) تعریف دیگری را از داده ارائه کرده‌اند به این صورت که داده مجموعهٔ متونی است که به هیچ پرسشی پاسخ نمی‌هد؛ و این اطلاعات است که به پرسش‌های مرتبط با چه کسی، چه زمانی، چه چیزی و کجا پاسخ می‌دهد. دانش مجموعهٔ متونی است که به سؤالات مرتبط با چرایی و چگونگی پاسخ می‌دهد. درتسکه[8] (۲۰۰۰) اطلاعات را با «صورت زبانی» سوسور (۱۹۱۶) پیوند زده است به این مفهوم که «صورت» ساختار عینی یک شیء است؛ و اطلاعات، ویژگی «صورت» را شکل می‌دهد. در چارچوب این رویکرد، دانش همان «معنا» در رویکرد سوسور به زبان است که یک مفهوم انتزاعی است. دالکیر[9] (۲۰۰۵) داده را به‌عنوان محتوا در نظر می‌گیرد که قابل دیدن یا تغییر است. اطلاعات محتوایی است که دادهٔ تحلیل شده را بازنمایی می‌کند؛ و دانش به اطلاعات نظری و مفید اطلاق می‌شود.

1. Zins
2. Capurro
3. Heidegger and Fink
4. Nauta
5. O’Brien
6. Laudon
7. Quigley and Debons
8. Dretske
9. Dalkir

اختراع رایانه موجب شد بین مفاهیم داده، اطلاعات و دانش تفاوت ایجاد شود. لانداور[1](۱۹۹۸)، بویسوت و کانالز[2] (۲۰۰۴)، شارما[3] (۲۰۰۵) و رولی[4] (۲۰۰۷) نوعی ارتباط سلسه‌مراتبی هرمی بین این سه مفهوم قائل شده‌اند به این صورت که داده در پایین هرم، اطلاعات در میانهٔ هرم و دانش در رأس هرم وجود دارد که رابطهٔ بین این سه مفهوم در شکل (۱) نمایش داده شده است. این هرم در حوزهٔ مدیریت دانش با اصطلاحاتی چون «سلسله مراتب دانش» یا «هرم دانش» معرفی می‌شود؛ در حالی‌که در حوزهٔ علم اطلاعات، این هرم به «سلسله مراتب اطلاعات» یا «هرم اطلاعات» معروف است.



شکل ۱: ارتباط بین سه مفهوم داده، اطلاعات و دانش

۱-۲- از داده به فراداده

در داده‌های الکترونیکی برای هر داده امکان تعریف فراداده (metadata) وجود دارد و این اطلاعات فراداده‌ای می‌تواند در کنار داده نگه داشته شود. فراداده اطلاعات یا توصیفی در مورد جنبه یا جنبه‌هایی از داده فراهم می‌نماید (بورگین، ۲۰۱۷: ۴۰). فراداده توسط باگلی[5] (۱۹۶۸) معرفی شد و سپس در حوزه‌های بسیاری چون مدیریت اطلاعات، دانش اطلاعات، فناوری اطلاعات، کتابداری و پایگاه داده به‌صورت فراگیر مورد استفاده قرار گرفت. فراداده در اصل داده‌ای در مورد داده است. به‌عبارتی دیگر، فراداده توصیفاتی است که جنبه‌های داده را توضیح می‌هد و این توصیف‌ها بافت داده را بیان می‌کند (اینمون و همکاران[6]، ۲۰۰۸).

در منابع موجود در وب یا نرم‌افزارها، از فراداده برای توصیف‌های ساختارمند و ایجاد بافت برای داده استفاده می‌شود که از نظر کاربردی در ارائهٔ خدمات پیشرفته حوزه‌های مختلف به‌کار برده می‌شود. به عبارتی

1. Landauer
2. Boisot and Canals
3. Sharma
4. Rowley
5. Bagley
6. Inmon & et al

دیگر، فراداده برای نشانه‌گذاری استفاده می‌شود که می‌تواند طرح‌واره ساده یا پیچیده داشته باشد (نوسهدال و همکاران[1]، ۲۰۱۱). فراداده اغلب به‌صورت مقولهٔ برچسب و ارزش برچسب به داده ملحق می‌شود. ویژگی کاربردی فراداده این است که می‌تواند در سامانه‌های جستجو مورد استفاده قرار گیرد و آنچه به کاربر ارائه می‌شود دانشی است که از جستجوی فراداده‌ها به‌دست می‌آید.

به‌طورکلی سه نوع فراداده وجود دارد:[2] الف) فرادادهٔ توصیفی: این نوع فراداده برای توصیف منابعی با هدف کشف یا تشخیص به‌کار می‌رود. این نوع فراداده‌ها شامل عناصری مانند عنوان، چکیده، نام نویسنده و واژه‌های کلیدی است؛ ب) فرادادهٔ ساختارمند: این نوع فراداده‌ها مشخص می‌کند چگونه عناصر مرکب کنار یکدیگر قرار بگیرد، مانند این‌که چگونه صفحات کتاب کنار یکدیگر مرتب می‌شود تا یک فصل شکل بگیرد؛ ج) فراداده اجرایی (سیاستی): این فراداده‌ها اطلاعاتی را فراهم می‌آورد تا به مدیریت یک منبع کمک نماید، مانند زمان و چگونگی ایجاد منبع، نوع فایل و سایر اطلاعات فنی و این‌که چه کسی اجازهٔ دسترسی به آن منبع را دارد.

۳- ساختار معیار داده

«زبان نشانه‌گذاری تعمیم‌یافتهٔ معیار» (Standard Generalized Markup Language (SGML)) زبانی در حوزهٔ رایانه است که امکان نشانه‌گذاری داده را فراهم می‌آورد. این زبان دارای استاندارد ۱۹۸۶ از «سازمان بین‌المللی استاندارد» (International Standard Organization) است[3]. در این نشانه‌گذاری، ساختار داده به‌صورت یک درخت ترسیم می‌شود؛ بنابراین این درخت از یک ریشه، تعدادی شاخه و تعدادی برگ تشکیل شده است. این ساختار به‌عنوان یک راهکار می‌تواند برای ساختارمندسازی داده و کاربردهای مرتبط با پردازش متن مورد استفاده قرار گیرد. «زبان نشانه‌گذاری ابرمتنی» (HyperText Markup Language (HTML)) و «زبان نشانه‌گذاری گسترش‌پذیر» (eXtensible Markup Language (XML)) دو ساختاری است که با دو کاربرد متفاوت از زبان نشانه‌گذاری تعمیم‌یافتهٔ معیار ساخته شده است. زبان نشانه‌گذاری ابرمتنی برای ساماندهی و نمایش داده در صفحات وب به‌کار می‌رود؛ در حالی‌که زبان نشانه‌گذاری گسترش‌پذیر برای توصیف داده و ارائهٔ اطلاعات مرتبط با آن داده کاربرد دارد. در این دو زبان نشانه‌گذاری، نوع اطلاعات با استفاده از برچسب مشخص می‌شود؛ با این تفاوت که در زبان نشانه‌گذاری ابرمتنی این برچسب‌ها محدود و معنای آن از قبل مشخص‌شده است و در زبان نشانه‌گذاری گسترش‌پذیر این برچسب‌ها از پیش‌تعریف شده نبوده و با توجه به نیاز، قابل تعریف و گسترش است. در شکل (۲) می‌توان نحوهٔ کاربرد زبان نشانه‌گذاری ابرمتنی و زبان نشانه‌گذاری گسترش‌پذیر را مشاهده نمود. شایان ذکر است که به‌دلیل تفاوت در کارکرد، جایگزینی این دو زبان با یکدیگر غیرممکن است.

1. Nocedal & et al
2. http://www.niso.org/publications/press/UnderstandingMetadata.pdf
3. https://www.w3.org/TR/NOTE-rdfarch

```
<!DOCTYPE html>
<html>
<body>

<h1>This tag is used for First
Heading.</h1>
<p> This tag is used for
determining a paragraph.</p>

</body>
</html>
```

(الف): کاربرد زبان نشانه‌گذاری ابرمتنی

```
<?xml version="1.0"?>
<catalog>
  <book id="bk101">
    <author>Gambardella,
Matthew</author>
    <title>XML Developer's Guide
</title>
    <genre>Computer</genre>
    <price>44.95</price>
<publish_date>2000-10-01
</publish_date>
    <description>An in-depth look at
creating applications with XML.
</description>
  </book>
</catalog>
```

(ب): کاربرد زبان نشانه‌گذاری گسترش‌پذیر

شکل ۲: کاربرد زبان نشانه‌گذاری ابرمتنی و گسترش‌پذیر

همان‌گونه که در شکل (۲) مشخص است، در سطر اول نوع زبان نشانه‌گذاری یک سند، خواه زبان نشانه‌گذاری ابرمتنی خواه زبان نشانه‌گذاری گسترش‌پذیر، تعریف شده‌است. در شکل (۲–الف) برچسب‌های محدود با مفهوم مشخص تعریف شده‌است، مانند <html> برای تعیین شروع یک سند و </html> برای تعیین انتهای آن سند، <body> برای تعیین شروع محتوای سند مورد نظر و </body> برای تعیین انتهای محتوای آن سند، <h1> برای تعیین شروع عنوان با اندازه قلم بسیار درشت و </h1> برای تعیین انتهای آن عنوان، برای تعیین شروع یک پاراگراف و برای تعیین انتها آن پاراگراف. کاربرد این برچسب‌های از پیش تعریف‌شده سبب می‌شود چینش اطلاعات در یک صفحهٔ وب براساس ساختاری که تعریف شده است انجام پذیرد. در شکل (۲–ب) ساختار داده برای یک کتاب‌فروشی به‌گونه‌ای تنظیم شده است که شامل تمام اطلاعات مربوط به یک کتاب اعم از نام نویسنده، نام ناشر، عنوان کتاب، تاریخ نشر، ژانر کتاب و توصیفی در مورد آن کتاب است. تغییر این کالا به ماشین سبب می‌شود برچسب‌های جدیدی با توجه به نیاز تعریف گردد. این ویژگی بیانگر انعطاف‌پذیری بالای زبان نشانه‌گذاری گسترش‌پذیر است.

سادگی در ساختارمندسازی داده در زبان نشانه‌گذاری گسترش‌پذیر و انعطاف‌پذیری در تعریف برچسب‌ها برای ساماندهی اطلاعات سبب شده است این زبان نشانه‌گذاری به‌طور فراگیر به‌خصوص در حوزهٔ وب مورد استفاده قرار گیرد. برای جستجوی درخت زبان نشانه‌گذاری گسترش‌پذیر از زبان جستجوی مسیر زبان نشانه‌گذاری گسترش یافته (XML Path Language (XPath)) استفاده می‌شود. با کمک این زبان

می‌توان درخت حاصل از زبان نشانه‌گذاری گسترش‌پذیر را پیمود و با دسترسی به گره‌های مادر، فرزند و خواهر، به جستجوی اطلاعات پرداخت.[۱]

ویژگی دیگر زبان نشانه‌گذاری گسترش‌پذیر این است که امکان تعریف فراداده در آن وجود دارد. این فراداده به‌صورت ویژگی (attribute) و ارزش (value) در هر گرهٔ درخت مربوط به زبان نشانه‌گذاری گسترش‌پذیر قابل تعریف است. امروزه، اهمیت فراداده به‌اندازه خودِ داده است و از اطلاعات آن در حوزهٔ درکِ عمیقِ معنایی استفاده می‌شود.

٤- ساختارمندسازی لایه‌های تحلیل زبان‌شناختی

در بخش پیشین، به‌صورت مختصر، انواع ساختار معیار داده در رایانه معرفی شد. در این بخش به ساختارهای معیار دادهٔ زبانی که در حوزهٔ تهیهٔ پیکرهٔ زبانی مطرح است خواهیم پرداخت. در حوزهٔ پردازش زبان طبیعی، دادهٔ ورودی باید براساس یک چارچوب مشخص ساختارمند گردد که در این خصوص «چارچوب نشانه‌گذاری واژگانی» (Lexicon Markup Framework) که توسط سازمان بین‌المللی استاندارد معرفی شده و به نحوهٔ مدیریت منابع زبانی می‌پردازد مورد پذیرش قرار گرفته‌است. ویژگی این ساختارمندسازی، قابلیت استفادهٔ مجدد از داده‌های گردآوری شده در پژوهش‌های بعدی است که برای این هدف معمولاً از ساختار زبان نشانه‌گذاری گسترش‌پذیر استفاده می‌شود. قیومی (۱۳۹۸) نحوهٔ ساماندهی تحلیل‌های زبان‌شناختی در لایه‌های مختلف را بیان کرده است که در این بخش به‌طور فشرده شیوهٔ ساماندهی اطلاعات مربوط به لایه‌های تحلیل زبان‌شناسی بیان می‌شود. وی شیوهٔ ساماندهی داده‌ها را با بهره‌گیری از زبان نشانه‌گذاری گسترش‌پذیر و ساختار «همایش یادگیری رایانشی زبان طبیعی» (Conference on Computational Natural Language Learning (CoNLL)) (بوچهولز و مارسی[۲]، ۲۰۰۶) معرفی کرده است. ساختار داده در این همایش که به ساختار CoNLL معروف است به این صورت است که اطلاعات در قالب سطر و ستون در یک فایل متنی سازماندهی می‌شود. ناگفته نماند امکان تبدیل داده از هرکدام از این دو ساختار به یکدیگر وجود دارد.

٤-۱- حوزهٔ آوا

به دو صورت می‌توان دادهٔ آوایی را ساماندهی نمود: الف) ذخیرهٔ داده به‌صورت امواج صوتی حاصل از مکالمهٔ دو یا چند انسان که برای ساماندهی این نوع داده از استاندارد زبان نشانه‌گذاری گسترش‌پذیر صوتی (VoiceXML) استفاده می‌شود.[۳] ب) ذخیرهٔ داده به‌صورت امواج صوتی محصول گفتگوی میان انسان و

---

۱. در زبان نشانه‌گذاری گسترش‌پذیر، رابطهٔ بین گره‌ها با اصطلاحاتی مانند گرهٔ مادر (mother) یا والد (parent)، فرزند (child) و خواهر یا برادر (sibling) معرفی می‌گردد که این اصطلاحات با رابطهٔ گره‌ها در حوزهٔ زبان‌شناسی متفاوت است.

2. Buchholz and Marsi

3. https://www.w3.org/TR/voicexml20/

ماشین که برای ساماندهی این نوع داده، از زبان نشانه‌گذاری تحلیل گفتار ((Speech Synthesis (SSML) Markup Language) استفاده می‌شود.[1] ساختار دو زبان نشانه‌گذاری گسترش‌پذیر صوتی و تحلیل گفتار براساس زبان نشانه‌گذاری گسترش‌پذیر ایجاد شده و مورد تأیید «کنسرسیوم جهانی وب» (World Wide Web Consortium (W3C)) است.

٤-۲- حوزهٔ صرف و نحو

در حوزه صرف دو سطح تحلیل وجود دارد: الف) ارائهٔ بن‌واژهٔ هر یک از صورت‌واژه‌ها؛ ب) تحلیل صرفی و نمایش پیشوندها و پسوندهای به‌کاررفته در ساختار یک واژه. این اطلاعات براساس ساختار CoNLL قابل ساماندهی است.

در حوزهٔ نحو، به دو صورت از ساختار CoNLL بهره برده می‌شود: الف) تعیین مقولهٔ دستوری هر واژه براساس بافت جایگاهی واژه؛ و ب) تعیین رابطهٔ نحوی میان واژه‌های تشکیل‌دهندهٔ یک جمله و برچسب‌زنیِ نوع این رابطه. سادگی ساختارمندسازی داده براساس CoNLL سبب شده است این ساختار به‌طور فراگیر مورد استفاده ابزارهای پردازش زبان طبیعی قرار گیرد. تعداد ستون‌هایی که حاوی اطلاعات صرفی و نحوی است براساس ساختار CoNLL در سال ۲۰۰۶ و ۲۰۰۹ بین ۱۰ تا ۱۴ ستون متغیر است که در هر ستون اطلاعات مشخصی تعریف شده است. علاوه بر ساختار CoNLL می‌توان از ساختار زبان نشانه‌گذاری گسترش‌پذیر در سطح جمله برای نمایش نمودار سلسله‌مراتبی درختی نیز استفاده کرد؛ چراکه ساختار زبان نشانه‌گذاری گسترش‌پذیر با ساخت سلسله‌مراتبی تحلیل جمله منطبق است.

٤-۳- حوزهٔ معنی

از سال ۱۹۹۸ سلسله‌کارگاه‌هایی به‌نام «ارزیابی مفهوم» (SENSEEVAL) برای پژوهش در حوزهٔ معناشناسی رایانشی برگزار شده است که عمدهٔ این پژوهش‌ها برروی ابهام‌زدایی معنایی متمرکز بوده است[2]. از سال ۲۰۰۷ این کارگاه به «ارزیابی معنایی» (SEMEVAL) تغییر نام داده است. ساختار داده‌ای که در این سلسله‌کارگاه‌ها مورد استفاده قرار گرفته است براساس زبان نشانه‌گذاری گسترش‌پذیر است به این صورت که معنای واژهٔ هدف به‌صورت ویژگی و ارزش در هر گرهٔ درخت مبتنی بر زبان نشانه‌گذاری گسترش‌پذیر تعریف شده است.

۵- ساختار پیشنهادی دادهٔ گویشی

در ابتدای بخش ۲ مقالهٔ حاضر مطرح شد که در هر طرح پژوهشی، دو بُعد مطالعات نظری و فناوری دانش مورد توجه است؛ و داده نیز باید به یکی از صورت‌های عنوان شده ارائه گردد. چنانچه بخواهیم با رویکرد

1. https://www.w3.org/TR/speech-synthesis11/
2 http://www.itri.brighton.ac.uk/events/senseval/ARCHIVE/index.html

گویش‌شناسی ساختگرای واینرایش (۱۹۵۴) به مطالعات گویشی بپردازیم، دادهٔ صوتی باید به‌صورت ساختارمند مورد بررسی قرار گیرد. روش مرسوم برای تحلیل این نوع داده، بازنمایی دادهٔ صوتی به‌صورت نگارش آوایی است که در این حالت دادهٔ صوتی به دادهٔ متنی تبدیل می‌شود. ویژگی دادهٔ متنی این است که قابل مشاهده است و می‌تواند به‌صورت دستی و یا ماشینی با کمک الگوریتم‌های پردازش زبان طبیعی تحلیل گردد. ناگفته نماند این امکان وجود دارد موقعیت جغرافیایی گویش مورد مطالعه به یک نقشه جغرافیایی متصل گردد تا ضمن امکان تحلیل داده‌های خامِ گویشیِ گردآوری شده بتوان «مرز همگویی» را نیز روی نقشه تعیین نمود. دستیابی به این اطلاعات به بررسی‌های موشکافانه نیاز دارد تا به دانش مربوط به گویش تحت مطالعه منجر گردد.

در این مقاله، یک ساختار مشخص در قالب یک استاندارد برای سازماندهی فراداده‌ها و تحلیل‌های زبان‌شناختی داده‌های گویشی در لایه‌های مختلف، اعم از آوا، صرف، نحو، معنی و تحلیل کلام، ارائه می‌گردد. این ساختار که براساس زبان نشانه‌گذاری گسترش‌پذیر پایه‌ریزی شده است، ضمن انعطاف‌پذیری برای نگهداری صورت‌های مختلف اطلاعات و سادگی جابه‌جایی داده‌ها، امکان تبدیل به ساختارهای دیگر و خوانش اطلاعات به یک پایگاه داده را فراهم می‌آورد.

۵-۱- ساختار کلی داده

در ساختار دادهٔ پیشنهادی، محتوای هر فایل صوتی در یک فایل مبتنی‌بر زبان نشانه‌گذاری گسترش‌پذیر ذخیره می‌گردد. ساختار زبان نشانه‌گذاری گسترش‌پذیر به این صورت خواهد بود که محتوای فایل در یک گرهٔ ریشه به‌نام <DOCUMENT> ذخیره می‌گردد. این گره دو فرزند دارد که یکی از فرزندها به‌نام <MetaData> بوده و حاوی فراداده مربوط به دادهٔ گویشی است و فرزند دیگر <Data> است که حاوی دادهٔ گویشی و تحلیل زبانشناختی آن می‌باشد.

۵-۲- ساختار فراداده

گرهٔ <MetaData> شامل چهار گره فرزند است که در این چهار گره، اطلاعاتی کلی در مورد فایل صوتی، اطلاعاتی کلی در مورد گویش تحت مطالعه، اطلاعاتی در مورد فردی که کار گردآوری داده را برعهده داشته و همچنین اطلاعاتی در مورد گویشور وجود دارد که به تفصیل توضیح داده می‌شود.

۵-۲-۱- اطلاعات کلی در مورد فایل صوتی

در گرهٔ <GeneralInfo> اطلاعات کلی در مورد فایل گویشی ذخیره شده است. این اطلاعات عبارت است از شماره یکتای فایل مبتنی بر زبان نشانه‌گذاری گسترش‌پذیر که در گرهٔ <DocumentID> ذخیره می‌شود. نام فایل صوتی دادهٔ گویشی در گرهٔ <VoiceFileName> ذخیره می‌شود. مسیر دسترسی به این فایل در رایانه مقصد در گرهٔ <PathToVoiceFile> موجود است. از آنجا که ممکن است وسایل ضبط صدا، فایل‌های صوتی را با قالب‌بندی‌های متفاوت تولید کند و برای کاربرد این فایل صوتی از قالب‌بندی

هماهنگ با نرم‌افزار تجزیهٔ و تحلیل استفاده نماید، باید نوع فایل صوتی در گرهٔ <FileType> مشخص گردد. اطلاعاتی در مورد طول مدت فایل صوتی و همچنین زمان تهیهٔ داده به‌ترتیب در دو گرهٔ <VoiceDuration> و <CollectionDate> تعریف می‌گردد.

۵-۲-۲- اطلاعات مربوط به گویش

اطلاعات گرهٔ <DialectInfo> حاوی نام گویش، نام روستا، نام شهر و استان محل گردآوری داده‌های گویشی است که اطلاعات آنها به‌ترتیب در گره‌های <Dialect>، <Village>، <City> و <Province> ذخیره می‌گردد. علاوه‌بر این موارد، اطلاعاتی در مورد موقعیت جغرافیایی منطقهٔ گویشیِ مورد نظر در گرهٔ <Geographical _location> ذخیره می‌گردد. اطلاعات مربوط به طول و عرض جغرافیایی به ترتیب در دو گرهٔ <NS> و <EW> موجود است. در این دو گره، اطلاعات مربوط به زاویه، ساعت، دقیقه و جهت نقطهٔ جغرافیایی موجود است. اطلاعات دیگری که در گرهٔ <DialectInfo> موجود است عبارت است از <Google_map_link>، <DialectPopulation> و <DialectFamily> است که به‌ترتیب حاوی اطلاعات مربوط به پیوند این موقعیت جغرافیایی در نقشهٔ گوگل، جمعیت منطقهٔ گویشی و نام خانواده گویش مورد نظر است. پیوند به نقشهٔ جغرافیایی گوگل سبب می‌شود امکان جانمایی موقعیت گویش براساس این نقشه فراهم آید.

۵-۲-۳- اطلاعات مربوط گردآوری‌کننده دادهٔ گویشی

اطلاعات گرهٔ <DataCollector> به فردی تعلق دارد که کار جمع‌آوری داده را انجام داده است. برای پیگیری فعالیت‌های این فرد نیاز است هویت وی با یک شماره یکتا مشخص گردد که گرهٔ <DataCollectorID> حاوی این اطلاعات است. اطلاعاتی در مورد نام، نام خانوادگی، سن و جنسیت فردی که کار گردآوری داده را انجام داده است به‌ترتیب در گره‌های <FirstName>، <LastName>، <Age> و <Gender> ذخیره می‌شود. اطلاعاتی در مورد رشته تحصیلی، مقطع تحصیلی و نام دانشگاه این فرد نیز مفید است که گره‌های <EducationField>، <EducationLevel> و <SchoolName> حاوی این اطلاعات است. اطلاعات تماس این فرد در گره‌های <ContactInfoTel> و <ContactInfoEmail> ذخیره می‌گردد.

۵-۲-٤- اطلاعات مربوط به تحلیل‌گر دادهٔ گویشی

اطلاعات گرهٔ <DataAnalyzer> به فردی تعلق دارد که کار تحلیل زبان‌شناختی دادهٔ گویشی را انجام می‌دهد. اطلاعات مورد نیاز این گره همانند اطلاعات معرفی‌شده مربوط به فرد گردآوری‌کنندهٔ دادهٔ گویشی است که از تکرار این اطلاعات پرهیز می‌شود.

۵-۲-۵- اطلاعات مربوط به گویشور

اطلاعات گرهٔ <Speaker> مربوط به یک یا چند گویشور است که حاوی اطلاعاتی مانند شناسه یکتای گویشور، نام، نام خانوادگی، سن، جنسیت، سطح تحصیلات و شغل وی است که این اطلاعات به‌ترتیب در

گـــــــــره‌هـــــــــای <SpeakerID>، <FirstName>، <LastName>، <Age>، <Gender>، <EducationLevel> و <Job> ذخیره می‌گردد. چنانچه بیش از یک نفر گویشور در فایـل صـوتی مـورد بررسی مشارکت داشته باشد، گرهٔ <Speaker> به تعداد نفرات با شناسه‌های یکتای متفاوت تکرار می‌شود و ویژگی‌های مرتبط ذخیره می‌گردد.

۵-۳- ساختار داده

گرهٔ <Data> شامل داده‌های گویشی است که داده‌های متعلق به هر گویشور مشارکت‌کننده در فایل صوتی مورد مطالعه، به‌طور جداگانه در گرهٔ <Speaker> ذخیره می‌گردد. این گره حـاوی دو نـوع اطلاعـات اسـت. یکی شناسهٔ گویشور است که بـه‌صـورت ویژگـی id و ارزش مشـخص شـده اسـت. ارزش ایـن ویژگـی بـا شناسه‌های تعریف شده در گرهٔ <Speaker> که در بخش ۵-۲-۵ معرفی شد منطبق است. اطلاعات دیگـر گرهٔ <Data> یک یا چند گرهٔ <S> است که تمام جملات بیان شده توسط گویشور را شـامل مـی‌شـود. در این گره، شمارهٔ یکتا جمله به‌عنوان شناسـهٔ آن و بـه‌صـورت ویژگـی و ارزش بیـان مـی‌شـود. ویژگـی بـا id مشخص شده و ارزش آن براساس ترتیب شمارهٔ متن در گرهٔ <Document ID> و ترتیب جمـلات تنظـیم می‌گردد.

گرهٔ <S> چندیدن شاخه دارد که این شاخه‌ها واژه‌های مربوط به آن جمله است؛ چراکه یک جمله از یک یا چند گرهٔ <w> تشکیل شده است. برگ‌های گرهٔ <w> آوانویسـی واژهٔ مـورد نظـر در جملـه اسـت. انـواع اطلاعات زبان‌شناختی با ساختار ویژگی و ارزش در هر گرهٔ <w> تعریف می‌گردد کـه در ادامـه توضـیح داده می‌شود. در شکل (۳) ویژگی‌های معرفی شده برای تعریف اطلاعات زبان‌شناختی نمایش داده شده است:

```
<w
id=“”
phonological_transcription=“”
pos=“”
lemma=“”
morphAnalysis=“”
ergative=“”
meaning=“”
case=“”
voiceFrag_Start=“”
voiceFrag_End=“”
anaphorResolusion=“”
cons=“”
dep=“”
depType=“”
>
 </w>
```

شکل ۳: ویژگی‌های معرفی شده برای تعریف اطلاعات زبان‌شناختی

در ادامه در مورد فراداده‌های گرهٔ <w> توضیح داده می‌شود:

- ویژگی id بیانگر شمارهٔ اندیس واژه در جمله است. این شماره برای هر واژه در متن یکتا است به این صورت که شمارهٔ متن، شمارهٔ جمله و شمارهٔ واژه در ارزش این ویژگی تعریف می‌گردد.
- ویژگی phonological_transcription حاوی نگارش واجی واژهٔ هدف است.
- ویژگی pos حاوی اطلاعات مقوله دستوری واژهٔ هدف براساس یک جدول مشخص است. این اطلاعات می‌تواند دانه‌درشت[1] یا دانه‌ریز[2] باشد. در اطلاعات دانه‌درشت فقط مقوله دستوری اصلی واژه مشخص می‌شود؛ در حالی‌که در اطلاعات دانه‌ریز اطلاعات صرفی- نحوی و حتی معنایی می‌تواند کدگذاری گردد. مجموعهٔ برچسب معرفی شده توسط بی‌جن‌خان و همکاران (۲۰۱۱) که برای فارسی معاصر تعریف شده است می‌تواند توسعه یابد و با توجه به نیاز برای داده‌های گویشی تغییر یابد.
- ویژگی lemma حاوی بن‌واژهٔ واژهٔ هدف است.
- ویژگی morphAnalysis حاوی تحلیل صرفی واژهٔ هدف است. در این تحلیل، تکواژهای تصریفی و اشتقاقی از پایه جدا می‌شود و تحلیل ساختار واژه را ارائه می‌کند.
- ویژگی ergative بیانگر ویژگی ارگتیو در گویشِ موردِ مطالعه است. این ویژگی حاوی ارزشِ دوگانه[3] است به این صورت که اگر واژه‌ای ارگتیو باشد ارزش «صحیح» و در غیر این‌صورت ارزش «غلط» برای آن واژه تعریف می‌شود.
- ویژگی meaning حاوی معنای واژهٔ هدف است.
- ویژگی case بیانگر حالت نحوی واژهٔ هدف در جمله است.
- ویژگی voiceFrag_Start حاوی زمان شروع برش فایل صوتی برای بیان واژهٔ هدف است.
- ویژگی voiceFrag_End حاوی زمان انتهای برش فایل صوتی برای بیان واژهٔ هدف است.
- ویژگی anaphorResolusion بیانگر مرجع ضمیر در جمله است. ارزش این ویژگی، اندیس واژهٔ ارجاع داده شده است.
- ویژگی cons حاوی تجزیهٔ نحوی واژهٔ هدف در تجزیهٔ سلسله‌مراتبی جمله است. برای تهیه این ساختار از استاندارد معرفی شده توسط CoNLL در سال ۲۰۱۱ استفاده می‌شود.
- ویژگی dep حاوی تجزیهٔ نحوی واژهٔ هدف در تجزیهٔ وابستگی جمله است. در این ساختار از شیوهٔ برچسب‌گذاری داده در CoNLL متعلق به سال ۲۰۰۶ استفاده می‌گردد. بر این اساس، ارزش این ویژگی، اندیس واژهٔ هسته است.
- ویژگی depType بیانگر نوع وابستگی بین عنصر هسته و وابسته است که ارتباط آنها در ویژگی dep مشخص شده است.

1. coarse-grained
2. fine-grained
3. boolean

برای مثال، دو جملهٔ (۱) و (۲) متعلق به گویش انارکی در زیر آورده شده است:

(۱)

on	yome
he/she	came+3SG

او آمد

(۲)

on	on	o	evine
he/she	him/her	DOM	seeing+3SG

او او را می‌بیند

ساختار دادهٔ پیشنهاد شده برای این دو مثال در زیر ارائه می‌گردد که قسمت‌های مربوط به این ساختار پیشتر در بخش‌های ۵-۱ و ۵-۲ توضیح داده شده است و در اینجا به اجمال توضیح داده می‌شود. گفته شد هر فایل مبتنی بر زبان نشانه‌گذاری گسترش‌پذیر متشکل از یک گرهٔ ریشه (<DOCUMENT>) است. این گره حاوی فراداده و داده‌های گویشی است. فراداده‌ها شامل اطلاعات کلی در مورد فایل صوتی، اطلاعاتی در مورد گویش و اطلاعاتی در مورد گردآورنده داده، تحلیل‌گر داده و گویشور است. قسمت داده نیز دربرگیرنده جملات گویش تحت بررسی است که توسط یک گویشور بیان می‌شود که تمام این اطلاعات در گرهٔ <Speaker> مشخص می‌گردد. این گره متشکل از تعدادی گرهٔ جمله (<S>) است. از آنجا که هر جمله از واژه تشکیل شده و واژه نیز حجم زیادی از اطلاعات را در خود جای می‌دهد، هر گرهٔ واژه (<w>) حاوی اطلاعات زبان‌شناسی است که در قالب فراداده معرفی می‌گردد.

```
<DOCUMENT>
<MetaData>
    <GeneralInfo>
      <DocumentID>1</DocumentID>
       <VoiceFileName>anarak1.wav</VoiceFileName>

      <PathToVoiceFile>C:\Dialects\Anaraki\</PathToVoiceFile>
      <FileType>Wave</FileType>
      <VoiceDuration>00:06:34:68</VoiceDuration>
      <CollectionDate>21.12.2006</CollectionDate>
    </GeneralInfo>
    <DialectInfo>
      <Dialect>انارکی</Dialect>
      <Village>انارک</Village>
      <City>نایین</City>
      <Province>اصفهان</Province>
```

```
<Geographical _location>
  <NS>
    <Angle>30</Angle>
    <Hour>18</Hour>
    <Second>40</Second>
    <Direction>N</Direction>
  </NS>
  <EW>
    <Angle>53</Angle>
    <Hour>41</Hour>
    <Second>54</Second>
    <Direction>N</Direction>
  </EW>
</Geographical _ locationEW>
<Google_map_link>https://www.google.com/maps/place/33%C2%B
018'40.0%22N+53%C2%B041'54.0%22E/@33.311111,53.698333,12z/data=!4m
5!3m4!1s0x0:0x0!8m2!3d33.311111!4d53.698333?hl=en”</Google_map_link>
<DialectPopulation>1903</DialectPopulation>
<DialectFamily></ DialectFamily >
</DialectInfo>
<DataCollector>
<DataCollectorID>100000</DataCollectorID>
<FirstName>[1]***</FirstName>
<LastName>انارکی ***</LastName>
<Age>21</Age>
<Gender>مؤنث</Gender>
<EducationField>زبانشناسی</EducationField>
<EducationLevel>کارشناسی ارشد</EducationLevel>
<SchoolName>دانشگاه آزاد تهران مرکزی</SchoolName>
<ContactInfoTel>0912*******</ContactInfoTel>
<ContactInfoEmail>***@gmil.com</ContactInfoEmail>
</DataCollector>
<DataAnalyzer>
  <DataAnalyzerID>100000</DataAnalyzerID>
<FirstName>*** </FirstName>
<LastName>انارکی ***</LastName>
<Age>21</Age>
```


۱. برای حفظ اطلاعات شخصی افراد در مقاله، از علامت * استفاده شده است.

<Gender>مؤنث</Gender>
<EducationField>زبانشناسی</EducationField>
<EducationLevel>کارشناسی ارشد</EducationLevel>
<SchoolName>دانشگاه آزاد تهران مرکزی</SchoolName>
<ContactInfoTel>0912*******</ContactInfoTel>
<ContactInfoEmail>***@gmil.com</ContactInfoEmail>
<AnnotationType>آوانگاری</AnnotationType>
</DataAnalyzer>
<Speaker>
<SpeakerID>200000</SpeakerID>
<FirstName>***</FirstName>
<LastName>انارکی ***</LastName>
<Age>20</Age>
<Gender>مذکر</Gender>
<EducationLevel>کارشناسی</EducationLevel>
<Job>دانشجو</Job>
</Speaker>
</MetaData>
<Data>
<Speaker id="200000">
<S id="1-1">
<w id="1-1-1" phonological_transcription="on" pos="PRON,3,SG" lemma="on" morphAnalysis="on" ergative="False" meaning="او" case="nominative" voiceFrag_Start="00:14:62" voiceFrag_End="00:14:64" anaphorResolusion="X" cons=" (ROOT(S(NP(PRON*))" dep="1-1-2" depType="Subject">on</w>

<w id="1-1-2" phonological_transcription="yome" pos="V,POS,PAST,3,SG" lemma="yom" morphAnalysis="yom+e" ergative="False" meaning="آمد" case="" voiceFrag_Start="00:14:64" voiceFrag_End= "00:14:70" anaphorResolusion="1-1-1" cons="(VP(V*))))" dep="1-1-0" depType="ROOT"> yome</w>
</S>
<S id="1-2">

<w id="1-2-1" phonological_transcription="on" pos="PRON,3,SG" lemma="on" morphAnalysis="on" ergative="False" meaning="او" case="nominative" voiceFrag_Start="00:14:72" voiceFrag_End="00:14:74"

anaphorResolusion="1-1-1" cons="(ROOT(S(NP(PRON*))" dep="1-2-4" depType="subject">on</w>

<w id="1-2-2" phonological_transcription="on" pos="PRON,3,SG" lemma="on" morphAnalysis="on" ergative="False" meaning="او" case="accusative" voiceFrag_Start="00:14:74" voiceFrag_End="00:14:76" anaphorResolusion="X" cons="(VP(NP(PRON*)" dep="1-2-4" depType="object">on</w>

<w id="1-2-3" phonological_transcription="o" pos="PAR" lemma="o" morphAnalysis="o" ergative="False" meaning="را" case="accusative" voiceFrag_Start= "00:14:76" voiceFrag_End="00:14:77" anaphorResolusion="" cons="(POSP*))" dep="1-2-2" depType="accMarker">o</w>

<w id="1-2-4" phonological_transcription="evine" pos="V,POS,PRES,3,SG" lemma="diy" morphAnalysis="evin+e" ergative="False" meaning="می‌بیند" case="" voiceFrag_Start="00:14:77" voiceFrag_End="00:14:90" anaphorResolusion="1-1-1" cons="(VP(V*)))))" dep="1-2-0" depType="ROOT">evine</w>
</S>
</Speaker>
</Data>
</DOCUMENT>

٦- جمع‌بندی و نتیجه‌گیری

در این مقاله به نحوهٔ ساختارمندسازی دادهٔ گویشی و چگونگی ساماندهی اطلاعات اضافه‌شده زبان‌شناختی به داده‌های گویشی پرداختیم تا در قالب یک استاندارد بتواند در گویش‌شناسی رایانشی مورد استفاده قرار گیرد. برای رسیدن به هدف، سه مفهموم داده، اطلاعات و دانش بیان شد. بنابر تعاریف ارائه‌شده، منظور از دادهٔ گویشی، دادهٔ خام ضبط‌شده از یک گویشور بوده و منظور از اطلاعات، تحلیل‌های زبان‌شناسی اضافه شده به دادهٔ گویشی است. دانش نیز توصیفی از گویش است که توسط پژوهشگر بیان شده و از تحلیل اطلاعات و نه دادهٔ خام به‌دست می‌آید. اگر بخواهیم این سه مفهوم را با سه سطح کفایت مطرح شده توسط چامسکی[1] (۱۹۶۵) برای مقایسه نظریه‌های زبانی مقایسه کنیم، به این نتیجه می‌رسیم که اطلاعات زبان‌شناسی با کفایت مشاهده‌ای و دانش با کفایت توصیفی همتراز است. اضافه‌شدن اطلاعات زبان‌شناسی به داده‌های گویشی در قالب فراداده موجب شود تا مختصات زبان برای کفایت مشاهده‌ای تعیین گردد. این امر به انبوهی از اطلاعات منجر می‌شود که برای استفادهٔ عملی از آن به ساختارمندسازی داده نیاز است. ساختارمندسازی

1. Chomsky

دادهٔ گردآوری شده موجب می‌شود این داده در محیط رایانه الکترونیکی شود و قابلیت جستجو در داده و استفاده مجدد از آن فراهم گردد. علاوه بر این موارد، قابلیت پردازش الگوریتمی این داده به‌صورت خودکار به‌وجود آمده و به پژوهشگر کمک شایانی در استخراج دانش از یک گویش و دستیابی به کفایت توصیفی می‌نماید.

منابع

– اکاشا، سمیر (۱۳۸۷). *فلسفه علم*، ترجمه هومن پناهنده. تهران: فرهنگ معاصر.

– قیومی، مسعود (۱۳۹۸). «ساماندهی تحلیل های چندلایه‌ای زبان شناختی در پیکره‌های زبانی،» در *واژه واژه زندگی: جشن‌نامه استاد ویدا شقاقی*، ویراستاران قطره، فریبا و شهرام مدرس خیابانی؛ تهران، ایران: نشر نویسه، ۲۸۷-۳۱۲.

– Bijankhan, M., J. Sheykhzadegan, M. Bahrani, and M. Ghayoomi (2011). “Lessons from building a Persian written corpus: Peykare”. *Language Resources and Evaluation*, 45(2): 143–164.

– Boisot, M., and Canals, A. (2004). “Data, information, and knowledge: Have we got it right,” *Journal of Evolutionary Economics*, 14: 43–67.

– Bagley, P. (1968). *Extension of Programming Language Concepts*, Philadelphia: University City Science Center.

– Buchholz, S. and Marsi, E. (2006). “CoNLL-X shared task on multilingual dependency parsing,” In *Proceedings of the 10th Conference on Computational Natural Language Learning*, Stroudsburg, PA, USA: Association for Computational Linguistics, 149–164.

– Burgin, M. (2001). “Information in the context of education,” *The Journal of Interdisciplinary Studies*, 14: 155-166.

– Burgin, M. (2017). *Theory of Knowledge: Structures and Processes*. Singapore: World Scientific Publishing Co. Pte. Ltd.

– Capurro, R. (1991). “Foundations of information science: Review and perspectives”, In *Proceedings of the International Conference on Conceptions of Library and Information Science*, University of Tampere, Tampere, Finland, pp. 26–28.

– Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, Massachusetts MIT Press.

– de Saussure, F. (1916). *Cours de linguistiquegénérale*, Lausanne, Paris: Payot.

– Dalkir, K. (2005). *Knowledge Management in Theory and Practice*. Amsterdam: Elsevier Science Ltd.

– Dretske, F. (2000). *Perception, Knowledge and Belief: Selected Essays*, Cambridge: Cambridge University Press.

– Heidegger, M. and Fink, E. (1970). *Heraklit*, Frankfurt am Main: Klostermann.

– Inmon, W.H., O'Neil, B. and Fryman, L. (2008). *Business Metadata: Capturing Enterprise Knowledge*. MA: Elsevier Morgan Kaufmann Publishers.

- Landauer, C. (1998). "Data, information, knowledge, understanding: Computing up the meaning hierarchy," *Proceedings of the 1998 IEEE International Conference on Systems, Man, and Cybernetics*, San Diego, California, pp. 2255–2260.
- Laudon, K.C. (1996). *Information Technology and Society*. California: Wadsworth P.C.
- Lyons, J. (1981). *Language, Meaning and Context*. London: Fontana.
- Nauta, D. (1970). *The Meaning of Information*, Paris: Mouton.
- Nocedal, A. S., Gerrikagoitia Arrien, J. K., & Burgin, M. (2011). "A mathematical model for managing XML data," *International Journal of Metadata, Semantics and Ontologies*, 6 (1): 56–73.
- O'Brien, J.A. (1995). *The Nature of Computers, The Dryden Press*, Philadelphia/San Diego.
- Okasha S. (2008). *Philosophy of Science*. Tehran: Farhang-e Moaser.
- Poster, M. (1990). *The Mode of Information: Post-structuralism and Social Contexts*. Chicago: University of Chicago Press.
- Quigley, E. J., & Debons, A. (1999). "Interrogative theory of information and knowledge", In *Proceedings of SIGCPR '99*, ACM Press, New Orleans, 4–10.
- Rowley, J. (2007). "The wisdom hierarchy: representations of the DIKW hierarchy," *Journal of Information Science*, 33 (2): 163–180.
- Sharma, N. (2005) *The Origin of the "Data Information Knowledge Wisdom" Hierarchy* (electronic edition: http://www-personal.si.umich.edu/~nsharma/ dikw_origin.htm)
- Van Marle, J. (2008). "Paradigmatic and syntagmatic relations," *1. Halbband: Ein internationales Handbuch zur Flexion und Wortbildung*, edited by G. Booij, C. Lehmann, J. Mugdan, W. Kesselheim and S. Skopeteas, Berlin, New York: De Gruyter Mouton, 225-234.
- Weinreich, U. (1954). "Is a structural dialectology possible?" *Word*, 10: 388–400.
- Zins, C. (2007). "Conceptual approaches for defining data, information, and knowledge", *Journal of the American Society for Information Science and Technology*, 58(4): 479–493.