مجموعه مقالات چهارمین همایش ملی زبان‌شناسی رایانشی

به کوشش

مسعود قیومی

آذین شهریاری‌فرد

سرشناسه	همایش ملی زبان‌شناسی رایانشی (چهارمین : ۱۳۹۶ : تهران)
عنوان و نام پدیدآور	مجموعه مقالات چهارمین همایش ملی زبانشناسی رایانشی/به کوشش مسعود قیومی، آذین شهریاری‌فرد ؛ برگزارکننده انجمن زبانشناسی ایران؛ با همکاری پژوهشگاه علوم انسانی و مطالعات فرهنگی، پایگاه استنادی علوم جهان اسلام(ISC)، نشر نویسه پارسی.
مشخصات نشر	تهران : نشر نویسه پارسی، ۱۳۹۶.
مشخصات ظاهری	۲۰۹ ص.: مصور، جدول، نمودار.
شابک	۹۷۸-۶۰۰-۷۰۳۰-۵۹-۲
وضعیت فهرست نویسی	فیپا
موضوع	زبان‌شناسی کامپیوتری -- کنگره‌ها
موضوع	Computational (Linguistics) -- Congresses
موضوع	زبان‌شناسی -- ایران -- کنگره‌ها
موضوع	Linguistics -- Iran -- Congresses
شناسه افزوده	قیومی، مسعود، ۱۳۵۸ - ، گردآورنده.
شناسه افزوده	شهریاری‌فرد، آذین، ۱۳۶۶ - ، گردآورنده.
شناسه افزوده	منظوری، ریحانه، ۱۳۶۳ - ، ویراستار.
شناسه افزوده	انجمن زبان‌شناسی ایران.
رده بندی کنگره	PIR۲۶۵۳/ ۸۸۸ت ۱۳۹۶
رده بندی دیویی	۴فا۰/۲۸
شماره کتابشناسی ملی	۵۰۶۶۰۶۶

انجمن زبان شناسی ایران

تهران، بزرگراه چمران، پل مدیریت، خیابان علامه طباطبایی جنوبی، دانشکده ادبیات فارسی و زبان‌های خارجی دانشگاه علامه طباطبایی. طبقه اول، اتاق ۱۱۷. صندوق پستی: ۱۵۹۷۶۳۳۱۱۱
تلفن:۸۸۶۹۰۰۲۲ نمابر: ۸۸۶۹۰۰۲۲
www.lsi.ir

نشر نویسه پارسی

تهران، صندوق پستی ۱۳۷۹-۱۶۷۶۵
تلفن:۷۷۰۵۳۲۴۶
نمابر:۷۷۰۵۳۲۴۶
سامانه پیام کوتاه:۳۰۰۰۴۵۵۴۵۵۴۱۴۲
وبگاه نشر نویسه پارسی:
www.neveeseh.com

ISC
Islamic World Science Citation Center

دارای درجه علمی ـ پژوهشی

نمایه شده در پایگاه استنادی علوم جهان اسلام (ISC)

دارای درجه علمی ـ پژوهشی، بر اساس بند ۱ ماده ۳ آیین‌نامه نحوه برگزاری و ساماندهی همایش‌های علمی ابلاغیه شماره ۳/۱۶۸۲۱۶ مورخ ۱۳۹۳/۹/۱۱ وزارت علوم، تحقیقات و فن‌آوری.

شابک: ۹۷۸-۶۰۰-۷۰۳۰ -۵۹ -۲
ISBN: 978-600-7030-59-2

مجموعه مقالات چهارمین همایش ملی زبان‌شناسی رایانشی

به کوشش	مسعود قیومی- آذین شهریاری‌فرد
ویراستار چکیده‌های انگلیسی	ریحانه منظوری
طرح جلد و یونیفورم	محمد محرابی www.mehraabi.com
صفحه‌آرایی و آماده‌سازی چاپ	محمد محرابی www.mehraabi.com
چاپ و صحافی	روز
شمارگان	۲۰۰ نسخه
نوبت چاپ	اول، ۱۳۹۶
قیمت	۲۰۰۰۰ تومان

مجموعه مقالات چهارمین همایش ملی زبان‌شناسی رایانشی

برگزارکننده

انجمن زبان‌شناسی ایران

با همکاری

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پایگاه استنادی علوم جهان اسلام (ISC)
نشر نویسه پارسی

سازمان همایش

دکتر مسعود قیومی دبیر علمی
آذین شهریاری‌فرد دبیر اجرایی

کمیته علمی

دکتر سیدمصطفی عاصی عضو هیئت علمی پژوهشگاه علوم انسانی و مطالعات فرهنگی
دکتر محمود بی‌جن‌خان عضو هیئت علمی دانشگاه تهران
دکتر اکبر حسابی عضو هیئت علمی دانشگاه اصفهان
دکتر علی رضاقلی‌فامیان عضو هیئت علمی دانشگاه پیام‌نور
دکتر مرتضی طاهری اردلی عضو هیئت علمی دانشگاه شهرکرد
دکتر مسعود قیومی عضو هیئت علمی پژوهشگاه علوم انسانی و مطالعات فرهنگی
دکتر علی رضاقلی‌فامیان عضو هیئت علمی دانشگاه پیام‌نور
دکتر مرتضی طاهری اردلی عضو هیئت علمی دانشگاه شهرکرد
دکتر سعیده ممتازی عضو هیئت علمی دانشگاه امیرکبیر
سید ابوالقاسم میرروشن‌دل عضو هیئت علمی دانشگاه گیلان

هیئت داوران

دکتر سیدمصطفی عاصی	عضو هیئت علمی پژوهشگاه علوم انسانی و مطالعات فرهنگی
دکتر کامبیز بدیع	عضو هیئت علمی موسسه تحقیقات ارتباطات و فناوری اطلاعات
دکتر محمد بحرانی	عضو هیئت علمی دانشگاه صنعتی شریف
دکتر محمود بی‌جن‌خان	عضو هیئت علمی دانشگاه تهران
دکتر حسین صامتی	عضو هیئت علمی دانشگاه صنعتی شریف
دکتر هشام فیلی	عضو هیئت علمی دانشگاه تهران
دکتر اکبر حسابی	عضو هیئت علمی دانشگاه اصفهان
دکتر مسعود قیومی	عضو هیئت علمی پژوهشگاه علوم انسانی و مطالعات فرهنگی
دکتر سعیده ممتازی	عضو هیئت علمی دانشگاه امیرکبیر
دکتر علی رضاقلی‌فامیان	عضو هیئت علمی دانشگاه پیام‌نور
دکتر بهروز مینایی	عضو هیئت علمی دانشگاه علم و صنعت ایران
دکتر محمدمهدی همایون‌پور	عضو هیئت علمی دانشگاه صنعتی امیرکبیر

کمیته اجرایی

زهرا ابراهیم‌بانکی، سهیلا ایزدی، محمدحسن ترابی، بیتا قوچانی، مریم محمدی، ریحانه منظوری، طاهره همتی

فهرست مطالب

پیشگفتار

چهارمین همایش ملی زبان‌شناسی رایانشی در ۲۶ بهمن ۱۳۹۶ در پژوهشگاه علوم انسانی و مطالعات فرهنگی به همت انجمن زبان‌شناسی ایران و با همکاری نشر نویسه پارسی و پایگاه استنادی علوم جهان اسلام برگزار شد. تعداد ۲۲ مقاله از مراکز آموزشی و پژوهشی مختلف کشور به دبیرخانه همایش ارسال شد که مقالات ارسالی در چارچوب محورهای ذکرشدهٔ همایش بود. پس از داوری و برگزاری شورای کمیتهٔ داوران، با میزان پذیرش ۴۱ درصد، تعداد ۹ مقاله برای چاپ در مجموعه‌مقالات برگزیده شد که از این تعداد، ۸ مقاله به‌صورت سخنرانی ارائه شد.

زبان‌شناسی رایانشی میان‌رشته‌ای است که به بررسی مسائل زبانی با رویکرد رایانشی می‌پردازد. هدف اصلی این حوزه، درک و تولید زبان طبیعی در قالب گفتار و نوشتار است. برای رسیدن به این هدف، دانش زبانی باید به‌گونه‌ای الگوریتمی که قابلیت فهم و استفاده توسط رایانه را دارد، ارائه گردد. زبان‌شناسی رایانشی بخشی از علوم شناختی نیز محسوب می‌شود که با حوزهٔ هوش مصنوعی به‌عنوان زیرمجموعه‌ای از علم رایانه، هم‌پوشی بسیار زیادی دارد. زبان‌شناسی رایانشی دو جنبهٔ نظری و کاربردی دارد. جنبهٔ نظری آن با زبان‌شناسی نظری و علوم شناختی در ارتباط است تا دانش زبانی برای درک و تولید به‌صورت صوری ارائه گردد. این‌گونه صوری‌سازی به‌شکلی دانش زبانی را شبیه‌سازی می‌کند که قابلیت پیاده‌سازی در نرم‌افزارهای رایانشی را داشته باشد. نرم‌افزارهای تهیه‌شده بیانگر جنبهٔ کاربردی زبان‌شناسی رایانشی است.

مقالات پذیرفته‌شده در این همایش هر دو جنبهٔ زبان‌شناسی رایانشی را پوشش داده‌است. تقریباً در همهٔ مقالات، **زبان فارسی** به‌عنوان زبان هدف برای بررسی و پردازش انتخاب شده‌است. ویژگی رویکرد کاربردی‌سازی در حوزهٔ زبان‌شناسی رایانشی و پردازش زبان فارسی سبب ارتقای جایگاه زبان فارسی در دنیای دیجیتال امروز می‌شود. واکاوی مقالات انتخاب‌شده مبین این نکته است که مقالات عمدتاً به معرفی یک الگوریتم یا روش با کاربرد مشخص برای بررسی

یک مسئله پرداخته‌است. خروجی این الگوریتم‌ها می‌تواند مستقیماً به تهیه نرم‌افزاری خاص برای پردازش زبان فارسی بیانجامد یا این که به‌عنوان بخشی از یک نرم‌افزار به‌کار رود. وجود این نرم‌افزارها می‌تواند به مطرح‌شدن سؤالات جدید و تلاش برای یافتن پاسخ به آنها در پژوهش‌های آینده منجر شود.

مقالات پذیرفته‌شده این همایش موضوعات زیر را در بر می‌گیرد: استانداردسازی دادهٔ زبانی، تهیهٔ درخت‌بانک سازه‌ای و تجزیهٔ نحوی خودکار، خلاصه‌سازی خودکار متن، تشخیص موجودیت‌های نامدار، نظرکاوی، و تجزیه معنایی خودکار.

بر خود لازم می‌دانم از تمامی بزرگانی که در شکل‌گیری این همایش تلاش کردند، تشکر و قدردانی نمایم. نخست از رئیس انجمن زبان‌شناسی ایران، سرکار خانم دکتر بلقیس روشن؛ از مسئولان محترم پژوهشگاه علوم انسانی و مطالعات فرهنگی؛ رئیس پژوهشکدهٔ زبان‌شناسی پژوهشگاه، جناب آقای دکتر مصطفی عاصی؛ دبیر اجرایی، سرکار خانم آذین شهریاری‌فرد و همچنین اعضای شورای اجرایی همایش که بزرگوارانه در اجرای همایش بسیار تلاش کردند؛ از سرکار خانم ریحانه منظوری که در ویرایش چکیده‌های انکلیسی کمک شایانی نمودند؛ و از جناب آقای امیر احمدی، مدیر محترم نشر نویسه پارسی، برای همراهی‌هایشان در اطلاع‌رسانی‌های همایش و همچنین چاپ مجموعه‌مقالات.

در پایان شایسته است از تمامی اعضای محترم کمیتهٔ علمی و هیأت داوران همایش صمیمانه تشکر و قدردانی نمایم. همچنین از استادان، دانشجویان و پژوهشگران گرانقدری که با ارسال مقاله‌های پُر بار خود به غنای علمی این همایش افزودند و دستاوردهای پژوهشی خود را با علاقمندان به این حوزه در میان گذاشتند، صمیمانه سپاسگزاری می‌کنم.

مسعود قیومی

بهمن ۱۳۹۶

مسئلهٔ چندواژگی در پردازش نحو رایانشی زبان فارسی

مسعود قیومی[۱]

چکیده

این مقاله به بررسی چالش چندواژگی در پردازش نحو رایانشی زبان فارسی و ارائه راهکار برای رفع آن می‌پردازد. این چالش به دو دستهٔ عمده تقسیم می‌شود: واحدهای واژگانی چند قطعه‌ای و چندقطعه‌ای‌های واژگانی یک واحدی. چالش دستهٔ اول زمانی ظاهر می‌گردد که در یک زنجیره، چند واژه به‌اشتباه به یکدیگر جوش‌خورده‌اند، و یا واژه‌بست به میزبان ملحق شده‌است. در دستهٔ دوم، چند زنجیره باید با هم ترکیب شوند تا یک واژه حاصل گردد. این دو چالش در پردازش نحوی و بن‌واژه‌سازی واژه‌ها متبلور است. برای رفع این دو دسته چالش، سه الگوریتم معرفی می‌شود که به‌ترتیب بر روی پیکرهٔ بی‌جن‌خان اجرا می‌گردند. ویژگی الگوریتم‌ها این است که در آنها از روش‌های قاعده‌مند و روش‌های مبتنی‌بر آمار استفاده شده‌است تا به‌طور منسجم بتوانند بر مشکلات حاصل از چندواژگی در پردازش نحو رایانشی زبان فارسی فائق آیند. پس از اعمال الگوریتم‌ها، کار ارزیابی با استفاده از دادهٔ آزمون نشان می‌دهد که با اجرای الگوریتم اول که در آن واژه‌بست از میزبان جدا می‌گردد، دقت ۸۰/۵۲ به‌دست می‌آید. با اجرای الگوریتم دوم برای رفع مشکل جوش‌خوردگی عناصر، دقت ۷۵/۴۳ به‌دست می‌آید. دقت ۸۶/۳۸ درصد با اجرای الگوریتم سوم برای ترکیب چند زنجیره و ساخت یک واژه به‌دست می‌آید.

کلیدواژه‌ها: چندواژگی، پردازش زبان طبیعی، نحو، پیکره زبانی.

[۱] استادیار گروه زبان‌شناسی، پژوهشگاه علوم انسانی و مطالعات فرهنگی.
m.ghayoomi@ihcs.ac.ir

۱. مقدمه

ازجمله کاربردهای زبان‌شناسی پیکره‌ای یافتن مصادیق از داخل پیکره و استخراج آمار است. بنابراین دادهٔ زبانی باید به‌گونه‌ای معیار شده باشد تا امکان جستجوی دقیق و استخراج آمار دقیق فراهم گردد. کاربرد دیگر پیکرهٔ زبانی آن است که با نشانه‌گذاری دادهٔ زبانی امکان جستجوهای پیشرفته میسر گردد. این نشانه‌گذاری در سطوح مختلف زبانی ازجمله نحو انجام می‌پذیرد. در انجام این امور، مؤلفه اصلی خط است، بنابراین هرگونه اشکالی در آن، بر داده، نتیجهٔ جستجو، و آمار استخراج‌شده تأثیرگذار است. خط فارسی به‌دلیل ویژگی‌هایی که دارد سبب می‌شود پردازش رایانشی دادهٔ زبانی را با اشکال مواجه کند. آنچه در این مقاله به آن پرداخته می‌شود چندواژگی به‌عنوان یک چالش اصلی در پردازش رایانه‌ای زبان فارسی به‌خصوص در قسمت نحو است. علی‌رغم پژوهش‌های انجام‌شده در این زمینه مانند شریفی‌آتشگاه و بی‌جن‌خان (۲۰۰۹) و طباطبایی و صراف (۱۳۹۶)، نیاز است با نگاهی عمیق‌تر و اشراف بیشتر این مسئله مورد بررسی و ارزیابی قرار گیرد.

این مقاله در ۷ بخش تهیه شده‌است. در بخش ۲، چالش‌هایی که برای تهیه یک پیکره برای زبان فارسی با آن مواجه می‌شویم، بیان می‌شود. بخش ۳ به معرفی داده و ابزار به‌کاررفته در این پژوهش می‌پردازد. در بخش ۴، چندواژگی به‌عنوان یکی از چالش‌ها مطرح شده و توضیح داده می‌شود. بخش ۵ به معرفی سه الگوریتم و ارائه راهکارهایی برای حل چالش چندواژگی می‌پردازد و توضیح داده می‌شود. در بخش ۶، نتایج به‌دست‌آمده بحث و بررسی خواهد شد. در انتها در بخش ۷، به نتیجه‌گیری راهکارهای ارائه‌شده خواهیم پرداخت.

۲. چالش‌های تهیهٔ پیکرهٔ زبان فارسی

قیومی و ممتازی (۲۰۰۹)، قیومی و همکاران (۲۰۱۰)، و شمس‌فرد (۲۰۱۱) چالش‌هایی را که به‌هنگام تهیهٔ پیکرهٔ زبان فارسی با آنها مواجه می‌شویم، بررسی کرده‌اند که در این بخش به‌طور خلاصه به آنها اشاره می‌شود. کدگذاری حروف فارسی و عربی و تداخل بین کد حروف «ی» و «ک» فارسی با کد عربی حروف «ي» و «ك» یکی از این چالش‌هاست. این تداخل در تهیهٔ فهرست الفبایی نیز تأثیرگذار است. چالش دیگر چندواژگی است که در ادامه این بخش توضیح داده خواهد شد. سبک نوشتاری از نظر نوشتار معیار، فرامعیار، یا فرومعیار چالش دیگر است که باید در تهیهٔ پیکرهٔ معیار که به‌عنوان نمایندهٔ زبان طلقی می‌شود در نظر گرفته شود. خلاقیت زبانی متجلی‌شده در خط نیز چالش دیگر است. این ویژگی در پیامک‌ها و نظرات کاربران و وبلاگ‌ها کاملاً مشهود است. ابهام به‌دلیل هم‌نویسگی واژه‌ها چالشی دیگر است. ممکن است چندواژگی سبب هم‌نویسگی گردد و سپس چالش ایجاد گردد. عامل دیگر هم‌نویسگی مانند عدم‌درج یا تغییر حروف و علایم قرضی از خط عربی مانند همزه یا تنوین سبب ایجاد هم‌نویسگی شده و پردازش خط را با چالش مواجه کند. متأسفانه علی‌رغم وجود دستور املایی مصوب فرهنگستان زبان و ادب فارسی (۱۳۸۹) فهرستی از استثناها مطرح می‌گردد که این خود منجر به آشفتگی و عدم‌قطعیت در املاء و پراگندگی آمار واژه می‌گردد. عدم وجود هرگونه معیار برای نگارش واژه‌های لاتین با حروف فارسی، مانند اسم لاتین نویسنده، در یک متن فارسی منجر به خلق واژه‌های جدید با توزیع آماری مستقل می‌گردد، بنابراین این خود یک چالش محسوب می‌شود.

تعدادی از این چالش‌ها را قیومی و همکاران (۱۳۹۴) تحت عنوان تنوع نگارشی بررسی کرده و با معرفی یک الگوریتم تلاش کرده‌اند، تنوع‌های نگارشی را به‌طور خودکار بیابند. در این مقاله چندواژگی و رفع مشکل آن

مطرح می‌گردد تا دادهٔ حاصل را بتوان برای نشانه‌گذاری برای سطوح دیگر زبان‌شناسی استفاده کرد.

شریفی‌آتشگاه و بی‌جن‌خان (۲۰۰۹) به‌صورت پیکره‌بنیان ویژگی چندواژگی را در زبان فارسی مطرح کرده‌اند. در بررسی آنها چندواژگی به دو دسته تقسیم می‌گردد. این دو دسته، محصول عدم‌رعایت فاصله‌گذاری صحیح بین واحدهای چندواژه‌ای است. دستهٔ اول، واحدهای واژگانی چند قطعه‌ای هستند که چندواژه با هم یک واحد واژگانی می‌سازند. در این دسته، فاصلهٔ کامل[۲] به‌جای فاصلهٔ مجازی[۳] درج شده‌است. بنابراین بین عناصر واژگانی فاصلهٔ کامل قرار گرفته‌است و با توجه‌به این که فاصلهٔ کامل به‌عنوان مرز واژه تلقی می‌گردد، سبب می‌شود یک واحد واژگانی به‌صورت چند واژه توسط رایانه تشخیص داده شود. برای مثال، عدم فاصله‌گذاری صحیح منجر می‌شود واژهٔ «درغیراین‌صورت» به‌صورت «درغیراین صورت» نگارش گردد و دو واژهٔ «درغیراین» و «صورت» توسط رایانه تشخیص داده شود. واژه‌های مرکب، مانند «برای اینکه»، و یا مواردی که بین وند تصریفی یا اشتقاقی و شکل پایه‌ای واژه، فاصلهٔ کامل به‌جای فاصلهٔ مجازی درج شده است، مانند «دانش آموز»، «بی هدف»، «فروشگاه ها» و غیره را می‌توان در این دسته قرار داد.

دستهٔ دوم، چندقطعه‌ای‌های واژگانی یک واحدی هستند که چند واحد واژگانی با هم یک واژه را تشکیل می‌دهد. در این دسته، عدم‌درج فاصلهٔ کامل سبب می‌شود چندین واژه به‌عنوان یک واژه تلقی گردد. منشاء اصلی این مشکل این است که تعدادی از حروف فارسی، مانند «آ»، «ا»، «د»، «ذ»، «ر»، «ز»، «ژ»، «و»، شکل متصل به حرف بعدی ندارند؛ بنابراین اتصال این حروف به حرف بعدی اختلالی را در امر خواندن ایجاد نمی‌کند. برای مثال عدم‌درج فاصله سبب می‌شود واحد «و یا در برابر او ایستادم» به‌صورت یک

[2]white space
[3]pseudo-space

واحد «ویادربرابراواستادم» نگارش شود که این زنجیره به‌صورت یک واحد توسط رایانه تشخیص داده می‌شود. عدم رعایت فاصله‌گذاری ممکن است به ابهام بیانجامد و مشخص نگردد آیا این زنجیره از حروف باید از منظر یک واحد چندواژه‌ای تحلیل گردد یا از منظر یک واژه چندواحدی یا خود یک واژهٔ بسیط و مستقل است. برای مثال «وبا» می‌تواند یک نوع بیماری باشد؛ بنابراین یک واژهٔ بسیط است. همچنین ممکن است یک واحد چند واژه‌ای «و» و «با» باشد. مثال‌های دیگری چون «مادر»، «توهم»، «راداری»، «بریزید»، و غیره بیانگر این ابهام است که نشان می‌دهد تصحیح داده به‌صورت خودکار نمی‌تواند کار ساده‌ای باشد. حالت کوتاه‌شدگی واژه‌ها در متون ادبی، مانند «مرا»، «زوی»، «کزو» و غیره، و یا در گفتار، مانند «بچته»، «چته» یا «بچتو» را می‌توان در دستهٔ یک واحد چندواژه‌ای قرار داد. ضمایر ملکی یا عنصر فعلی که به‌صورت واژه‌بست به پایه متصل می‌شوند، مانند «کتابش»، «درعذابند»، یا «داراست»، در این دسته قرار می‌گیرند.

به‌هنگام کار با پیکره و نشانه‌گذاری داده در سطوح مختلف زبان‌شناسی دو چالش مطرح‌شده می‌تواند مشکل‌زا باشد. در پیکرهٔ بی‌جن‌خان (بی‌جن‌خان، ۱۳۸۳) مقولهٔ دستوری واژه‌ها مشخص شده‌است که در بخش بعدی توضیح داده خواهد شد. مجموعه برچسب‌های به‌کارگرفته‌شده برای نشانه‌گذاری این داده به‌گونه‌ای تعریف شده‌است که می‌تواند مشکل دستهٔ اول را توجیه نماید ولی در مورد مشکل دستهٔ دوم، مقولهٔ دستوری واژه‌ها به‌صورت مستقل مشخص شده‌است و مقوله دستوری عناصر ترکیب‌شده با هم نامشخص است. این مشکلات در سطح بالاتر تحلیل زبان‌شناختی مانند نحو که در آن باید ترسیم نمودار درختی جمله از روابط واژه‌ها با هم و نه مستقل از هم به‌دست آید، به‌خوبی خود را نشان می‌دهد. این مشکلات در بن‌واژه‌سازی نیز اشکالاتی ایجاد می‌کند. در این مقاله تلاش می‌شود برای رفع هر یک از این دو دسته چالش‌ها، راهکاری معرفی گردد تا قابلیت استفاده در پردازش نحو

جملات فارسی و همچنین تهیهٔ دادگان درختی فارسی و بن‌واژه‌سازی داشته باشد.

۳. معرفی داده و ابزار به‌کاررفته

در پیکرهٔ بی‌جن‌خان (۱۳۸۳) مقولات دستوری به‌صورت سلسله‌مراتبی برچسب‌گذاری شده‌است (بی‌جن‌خان و همکاران، ۲۰۱۱). در این برچسب‌گذاری از استاندارد ایگلز[4] (لیچ و ویلسون، ۱۹۹۹) استفاده شده‌است. نحوهٔ سلسله‌مراتبی این برچسب‌گذاری به‌این‌صورت است که ابتدا مقولهٔ دستوری اصلی واژه مشخص می‌گردد و سپس ویژگی‌های صرفی-نحوی و تا حدی معنایی برای هر واژه مشخص می‌شود. در جدول ۱ این نحوهٔ برچسب‌گذاری نمایش داده شده‌است. ترتیب قرارگرفتن اطلاعات صرفی-نحوی منجر به ایجاد ۵۸۶ برچسب متنوع شده‌است.

در این برچسب‌گذاری از ویژگی‌های معنایی برای ابهام‌زدایی همنویسه‌ها استفاده شده‌است. بنابراین در نمونه‌های جدول ۱، اضافه‌شدن ویژگی «مکان» سبب می‌شود «دفتر» به‌معنای «محل کار» از «دفتر» به‌معنای «دفتر یادداشت» مجزا گردد.

جدول ۱: نمونه برچسب‌گذاری در پیکرهٔ بی‌جن‌خان

واژه	برچسب	تفسیر برچسب
دفتر	N,COM,SG,LOC	اسم، عام، مفرد، مکان
دفتر	N,COM,SG,LOC,EZ	اسم، عام، مفرد، مکان، اضافه
دفتر	N,COM,SG	اسم، عام، مفرد
دفتر	N,COM,SG,EZ	اسم، عام، مفرد، اضافه

[4] EAGLES

در بخش دوم مقاله حاضر دو دسته چندواژگی مطرح شد. در دستهٔ اول چندواژگی، چند واحد واژگانی به‌عنوان یک واحد تلقی می‌شود. کوتاه‌شدگی و عدم‌رعایت فاصله‌گذاری سبب می‌شود بین عناصر واژگانی جوش‌خوردگی اتفاق افتد و این عناصر به‌شکل یک واحد تلقی گردد. واژه‌بست نیز عنصر دیگری است که ذاتاً نیاز به عنصری دیگر برای جوش‌خوردگی دارد. در اینجا به نمونه‌هایی از پیکرهٔ بی‌جن‌خان و نحوهٔ برچسب‌دهی آنها اشاره خواهد شد که در جدول شماره ۲ نمایش داده شده‌است. در بخش ۴، این نحوهٔ برچسب‌دهی مقولات دستوری به‌عنوان چالش چندواژگی در تهیهٔ دادگان درختی مطرح می‌شود.

جدول ۲: برچسب‌گذاری چندواژگی به‌صورت یک واحد در پیکرهٔ بی‌جن‌خان

واژه	برچسب	تفسیر برچسب
وبا	CONJ,P N,COM,SING	حرف ربط، حرف نشانه را اسم، عام، مفرد
بچتو	N,COM,SING,2,POSTP	اسم، عام، مفرد، واژه‌بست ضمیری دوم شخص مفرد، حرف نشانه را
کتابتو	N,COM,SING,2, POSTP	اسم، عام، مفرد، واژه‌بست ضمیریدوم شخص مفرد، حرف نشانه را
کزو	CONJ,P,PRO,PERS,SING,3	حرف ربط، حرف‌اضافه،ضمیر شخصی سوم شخص مفرد
اویم	V,COP,PROC,1	فعل، ضمیر پیش فعلی، اول شخص مفرد

واژه	برچسب	تفسیر برچسب
درعذابند	V,COP,PREPNC,6	فعل، حرف‌اضافه به‌همراه گروه اسمی متممی، سوم شخص جمع
جدیدش	AJ,SIM,3	صفت، ساده، واژه‌بست ضمیری سوم شخص مفرد

در دستهٔ دوم، یک واحد واژگانی از ترکیب یک یا چند واژه حاصل می‌گردد. برای این ترکیب نیاز است واژه‌های موردنظر با هم ترکیب شوند تا یک واحد واژگانی ساخته شود. برای انجام این کار می‌توان از ابزاری مانند سامانهٔ کلارک[۵] (سیمو و اوسنوا، ۲۰۰۲) استفاده کرد که شاخص‌ترین ویژگی‌های این ابزار در ادامه توضیح داده خواهد شد.

سامانهٔ کلارک[۶] که به‌طور رایگان قابل دسترس است، براساس XML بنا نهاده شده‌است. این سامانه مجهز به زبان XPATH است تا جستجو در اسناد را میسر سازد. در این سامانه، قواعد دستوری می‌تواند با استفاده از اطلاعات واژی-نحوی به‌صورت عبارات باقاعده[۷] بیان گردد. برای جلوگیری از فرازایش[۸] این قواعد، بافت محلی قواعد مورد توجه قرار می‌گیرد تا با تعریف محدودیت‌هایی برای قواعد، از فرازایش این قواعد کاسته شود. قواعد تعریف‌شده در این سامانه به‌عنوان «ماشین‌های حالت محدود»[۹] نقش‌آفرینی می‌کند که براساس دستورِ آبشاریِ[۱۰] آبنی (آبنی، ۱۹۹۶) این قواعد به‌صورت

[5] CLaRK
[6] http://www.bultreebank.org/clark/
[7] Regular Expression
[8] Over-generate
[9] Finite State Automata
[10] Cascaded Grammar

سلسله‌مراتبی با هم در ارتباط بوده و خروجی حاصل از یک قاعده (ماشین) به‌عنوان ورودی قاعدهٔ بعدی (ماشین بعدی) تعریف می‌گردد.

همان‌طور که پیشتر عنوان شده بود، ساختار داده در کلارک براساس XML است؛ بنابراین هر واژه در یک گره قرار می‌گیرد و اطلاعات زبان‌شناختی مربوط به آن واژه به‌صورت مشخصه[۱۱] و ارزش[۱۲] بیان می‌شود. دسترسی به اطلاعاتی مانند صورت واژه و مقولهٔ دستوری آن، به انعطاف‌پذیری در تعریف قواعد باقاعده و همچنین محدودکردن اعمال این قواعد می‌افزاید. در کلارک سه علامت مفهوم خاصی دارند که برای افزایش انعطاف‌پذیری قواعد می‌توان از این علایم به‌هنگام نگارش قواعد استفاده کرد. علامت @ به‌مفهوم هیچ یا یک حرف است، علامت % به‌مفهوم هیچ یا بیش‌از یک حرف است، و علامت # به‌مفهوم یک یا بیش‌از یک حرف است.

۴. چالش چندواژگی در تهیهٔ دادگان درختی

در بخش ۲، ویژگی‌های برچسب‌های دستوری واژه‌ها معرفی شد. آنچه در برچسب‌ها مشاهده می‌شود، این است که تلاش شده‌است برچسب‌ها با شکل نوشتاری منطبق گردد که در این صورت نحوهٔ تعریف برچسب‌ها در راستای خط فارسی بوده و شیوهٔ نگارش در برچسب‌ها لحاظ شده‌است. اتخاذ چنین روشی در تحلیل و نمایش اطلاعات زبان‌شناختی می‌تواند در سطح مشخص شدن نقش دستوری واژه مفید باشد، ولی برای سطوح تحلیل عمیق‌تر مانند تهیهٔ تجزیهٔ درختی جمله ناکارامد است. ازآنجاکه دادگان درختی[۱۳] جملات می‌تواند برای آموزش یک تجزیه‌گر خودکار[۱۴] توسط رایانه مورد استفاده قرار گیرد، واضح و مشخص بودن اطلاعات زبان‌شناختی باید در ساختار داده

[11] attribute
[12] value
[13] treebank
[14] parsing

لحاظ گردد تا بهترین سطح کارایی تجزیه‌گر به‌دست آید. برای درک بهتر این چالش، جمله (۱) را در نظر بگیرید.

(۱) حسین آمد وبا حسن به مدرسه رفت.

«با» هستهٔ گروهٔ حرف‌اضافه‌ای عبارت «با حسن» است. در قواعد ساخت سازه‌ای تعریف‌شده در کلاک، حرف‌اضافه و اسم با هم ترکیب می‌شوند تا یک گروه حرف‌اضافه‌ای بسازد. جوش‌خوردگی «و» و «با» و درج برچسب دستوری «P,CONJ» در پیکره سبب می‌شود امکان اتصال دو واژهٔ «با» و «حسن» میسر نگردد. دلیل آن این است که در زنجیرهٔ «وبا حسن» هستهٔ حرف‌اضافه‌ای و برچسب آن مشخص نیست تا قاعده تعریف‌شده اعمال گردد. در مثال (۲) تا (۴) پدیدهٔ واژه‌بست مطرح است.

(۲) کتابتو از حسن گرفتم.
(۳) کتاب جدیدش را از حسن گرفت.
(۴) آنها سخت درعذابند.

در «کتابتو» دو واژه‌بست وجود دارد: یکی نقش ضمیر ملکی را داراست، و دیگری نقش‌نمای مفعولی. می‌دانیم ضمیر ملکی و حرف نشانه «را» در چارچوب دستور ساخت سازه‌ای هسته‌بنیان[۱۵] متمم هستهٔ اسم (سمولیان و تسنگ، ۲۰۱۰؛ مولر و قیومی، ۲۰۱۰) تلقی می‌گردند که در مثال (۲) این دو واژه‌بست متمم «کتاب» هستند. از آنجاک ه این عناصر واژگانی (واژه‌بست‌ها) با نقش‌های دستوری متفاوت (واژه‌بست ملکی و واژه‌بست نقش‌نما) به‌صورت یک واحد آمده‌است، نیاز است از پایه منفک گردند تا نقش و روابط دستوری آنها به‌عنوان یک واحد واژگانی کاملاً مستقل مشخص گردد و تحلیل صحیح جمله به‌دست آید.

[15] Head-driven Phrase Structure Grammar