نقش پیکره‌های زبانی در نوشتن دستورزبان معرفی یک نرم‌افزار رایانه‌ای

محمود بی‌جن‌خان

چکیده

در رویکرد تاریخ‌گرایی و ساخت‌گرایی به زبان از پیکره‌های زبانی برای حل بعضی از مسائل زبانی استفاده می‌شده است. همچنین برای دست‌یابی به انگارهٔ مارکفی زبان و ضرورت استخراج توزیع احتمال مشروط کلمات زبان از پیکره‌های زبانی استفاده شده و می‌شود. اگرچه با تأکید زبان‌شناسان زایشی بر داده‌های بالقوه و نه بالفعل زبانی برای کشف پیچیدگیهای ذهن انسان، جایگاه پیکره‌ها در تجزیه و تحلیل دستوری تضعیف شده، با رشد سریع فن‌شناسی اطلاعات ضرورت تهیهٔ پیکره‌ها و دادگانهای زبانی با حجم بسیار بالا از اولویت بالایی برخوردار شده است. در این مقاله بر ضرورت تحلیل یافته‌های زبان‌شناسی نظری دربارهٔ دستورزبان با استفاده از پیکره‌های زبانی تأکید شده و برای این منظور یک نرم‌افزار رایانه‌ای معرفی شده است.

مقدمه

اگر زبان را یک مجموعهٔ نامتناهی از رشته‌های خوش‌ساخت واجی، صرفی، نحوی و معنایی فرض کنیم، به دو روش می‌توان دستورزبان را توصیف کرد.

روش اول: این روش مبتنی بر استدلال استقرایی است و شامل مراحل زیر:

۱. جمع‌آوری داده‌های یک زبان خاص (در یک فاصلهٔ زمانی تعریف‌شده) اعم از مکتوب و ملفوظ،

۲. طبقه‌بندی نظام‌مند داده‌ها،

۳. توصیف و صورت‌بندی آنها با استفاده از یک انگارهٔ دلخواه، مانند انگارهٔ احتمالی یا ساختاری.

روش دوم: این روش مبتنی بر استدلال قیاسی است و شامل مراحل زیر:

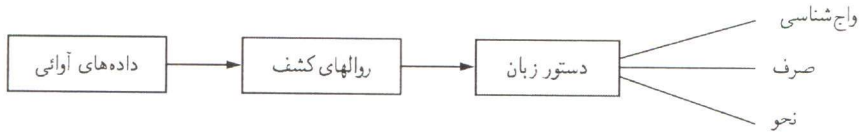
۱. جمع‌آوری داده‌های یک زبان خاص (در فواصل زمانی دلخواه) اعم از مکتوب و ملفوظ،

۲. استخراج اصول و پارامترهای زبانی از داده‌ها،

۳. توصیف داده‌های هر زبان ممکن در چارچوب اصول و مقادیر پارامترها.

با استفاده از روش اول می‌توان دستور هر زبان را مجزا از زبانهای دیگر بررسی کرد و نتایج حاصل از آن صرفاً برای داده‌های آن زبان صادق است، اما با استفاده از روش دوم می‌توان دستورزبان هر جامعهٔ زبانی را بررسی کرد، و در صورت لزوم در اصول زبانی تجدیدنظر کرد تا برای ساخت داده‌های زبانی بیشتری صادق باشند. با این روش می‌توان از دستور یک زبان خاص، مانند زبان فارسی، شروع کرد و به دستور جهانی زبان نوع بشر رسید.

نودستوریان در اواخر قرن نوزدهم مطالعات زبانی خود را در چارچوب روش اول آغاز کردند. لسکین، اُشتهوف و بروگمان (۱۸۷۸) متأثر از فلسفهٔ پوزیتیویسم منطقی تحقیقات خود را در خصوص کشف روابط خویشاوندی بین زبانها و استخراج خانواده‌های زبانی با استفاده از همین روش جزء به کل انجام دادند، یعنی جمع‌آوری کلمات در زبانهای خواهر، طبقه‌بندی و مطابقت دادن آنها با یکدیگر و بازسازی صورت آوائی آنها در زبان مادر. فرآنتس بوآس (۱۹۴۲ـ۱۸۵۸) و لئونارد بلومفیلد (۱۹۴۹ـ۱۸۸۷) نیز با همین روش به مطالعهٔ زبان سرخ‌پوستان امریکایی پرداختند (کنستوویچ ۱۹۹۴). بلومفیلد (۱۹۳۳) در کتاب خود زبان (ترجمهٔ حق‌شناس ۱۳۷۹) با پیروی از همین روش استقرایی، ساخت زبان را با تحلیل داده‌های گفتاری در حوزه‌های واج‌شناسی، صرف و نحو به دست داده و از این رهگذر پایه‌های زبان‌شناسی ساختگرا را در امریکا بنا نهاده است. وی به زبان از دیدگاه روان‌شناسی رفتارگرای اسکینر می‌نگریست. بی. اف. اسکینر (۱۹۰۴ـ۱۹۹۰) در کتاب مشهور خود رفتار شفاهی زبان را به عنوان مجموعه‌ای از رفتارهای یادگیری شده از طریق فرایند انگیزه و پاسخ معرفی کرد (چامسکی ۱۹۵۹). بنابراین در این رویکرد، زیر بنای تحلیل زبانی داده‌های آوائیِ حاصل از ارتباط گفتاری گویشوران زبان با یکدیگر است. اوج ساخت‌گرایی بلومفیلد، در روالهای کشف زلیگ هریس (۱۹۵۱) ظاهر شد. وی در کتاب خود روشهای زبان‌شناسی ساختاری نشان داده است که چگونه می‌توان با استفاده از یک مجموعهٔ متناهی از دستورالعملهای صوری که بر داده‌های آوائی یک زبان خاص اعمال می‌شود، دستور آن زبان را کشف کرد. دستورالعملهای صوری همان روالهای کشف هستند که به ترتیب ساخت واجی، صرفی و نحوی زبان را به دست می‌دهند (شکل ۱).



شکل ۱. استخراج دستورزبان با استفاده از روالهای کشف

درونداد روالهای کشف، داده‌های آوائی و برونداد آن دستورزبان است. یعنی اگر حجم قابل ملاحظه‌ای از داده‌های آوائی یک زبان خاص را در یک دورهٔ زمانی معین جمع‌آوری کنیم و تعدادی دستورالعمل صوری را که به جستجوی الگوهای زبانی، از قبیل توزیع آواها، واجها، تکواژها و کلمات در داده‌ها می‌پردازند، به کار گیریم، در نهایت ساخت واجی، صرفی و نحوی زبان مورد مطالعه را به دست خواهیم آورد. بنابراین، روالهای کشف در ساختارگرایی هریس دقیقاً یک روش استقرایی داده‌بنیاد برای کشف دستورزبان است. بر همین اساس بود که در دههٔ ۵۰ بعضی بر این اعتقاد بودند که می‌توان با آوانویسی حجم انبوهی از داده‌های گفتاری و نوشتن برنامه‌های رایانه‌ای مبتنی بر روالهای کشف، از رایانه‌ها برای استخراج دستورزبان استفاده کرد. پیتر ای. بنش (۱۹۹۱) معتقد است که گسترش انقلاب چامسکی بعد از ساختارگرایی هریس تا حدّی ناشی از کاستیهای سخت‌افزاری و نرم‌افزاری رایانه‌ها برای اجرای برنامهٔ موجود در نمودار بالاست.

در واقع نوشتن دستورزبان با استفاده از روالهای کشف مستلزم داشتن ابزارهایی است که اولاً به کمک آنها بتوان حجم بسیار زیادی از داده‌های زبانی را به عنوان نمونهٔ آماری زبان مورد مطالعه جمع‌آوری و به صورت رقمی در حافظهٔ رایانه ذخیره کرد و ثانیاً سرعت پردازشی رایانه‌ها برای استخراج الگوهای زبانی قابل قبول باشد. فن‌شناسی اطلاعات در دورهٔ ساخت‌گرایی هریس در مقایسه با امروز در دوران طفولیت خود به سر می‌برد و چنین ابزارهایی در اختیار محققان نبود. یعنی به عقیدهٔ بنش، اگر فن‌شناسی رایانه در دههٔ ۵۰ و ۶۰ در حد و اندازهٔ فن‌شناسی اواخر قرن بیستم بود چه بسا روش تحقیق در مطالعات زبانی به نفع مکتب زایشی تغییر جهت نمی‌داد و در نتیجه به جای آنکه داده‌های تحقیق برای توصیف ساخت زبان، جملات یا ساختهای بالقوه در ذهن زبان‌شناس باشند، همانا شامل جملات بالفعلی می‌شد که گویشوران زبان عملاً می‌گویند یا می‌شنوند، یعنی همان چیزی که تحت عنوان «پیکرهٔ زبانی» (corpus) مطرح بوده و هست.

دیوید کریستال (۱۹۹۴) می‌نویسد: «پیکرهٔ زبانی، مجموعه‌ای از داده‌های زبانی است که به صورت متون نوشتاری یا آوانوشته‌های گفتاری گردآوری شده‌اند. هدف اصلی از گردآوری یک

پیکرهٔ زبانی عبارت است از تأیید یک فرضیهٔ زبانی. به عنوان مثال، فرضیه‌ای دربارهٔ چگونگی تغییر در استفاده از یک آوا، کلمه یا ساخت نحوی خاص. زبان‌شناسی پیکره‌ای به مطالعهٔ اصول و روش به‌کارگیری پیکره‌های زبانی در مطالعهٔ زبان می‌پردازد. یک پیکرهٔ زبانی رایانه‌ای شامل متون زیادی است که توسط رایانه قابل خواندن باشد.»

تام مک‌آرتور نیز می‌نویسد: «پیکرهٔ زبانی، در زبان‌شناسی و واژگان‌نگاری (lexicography) دلالت بر مجموعه‌ای از متون، گفته‌ها (utterances) یا سایر داده‌های زبانی دارد که اولاً بتواند کم و بیش نمونهٔ آماری جامعهٔ داده‌های یک زبان خاص بوده و معمولاً به صورت دادگان[1] (database) الکترونیکی ذخیره شده باشد.» در حال حاضر، پیکره‌های زبانی رایانه‌ای میلیونها کلمه را در خود جای می‌دهند، به طوری که مشخصات آنها را می‌توان با استفاده از برچسب‌دهی (تعیین مرز کلمات یا ساختهای بزرگتر از کلمه و طبقه‌بندی آنها با استفاده از مجموعه‌ای از برچسبها) و همچنین به کارگیری برنامه‌های متنوع آماری تحلیل کرد. زبان‌شناسی پیکره‌ای به مطالعهٔ داده‌ها در هر پیکرهٔ زبانی با چنین مشخصاتی می پردازد.

از دههٔ پنجاه به بعد زبان‌شناسان برای مطالعهٔ ساخت زبان به تهیهٔ پیکره‌های زبانی پرداختند، مانند پیکرهٔ زبانی «بررسی جامع به‌کارگیری زبان انگلیسی» (Survey of English Usage=SEU) توسط راندلف کرک (Randolph Quirk) در سال ۱۹۵۹ در دانشگاه لندن (UCL) و پیکرهٔ زبانی «بررسی جامع زبان انگلیسی گفتاری (Survey of Spoken English=SSE) توسط یان سوراتویک (Jan Svartvik) در سال ۱۹۷۵ در دانشگاه لوند (Lund) و پیکرهٔ زبانی براون توسط دبلیو. ان. فرانسیس و اچ. کوکرا (W. N. Francis & H. Kucera) در سال ۱۹۶۴ در گروه زبان‌شناسی دانشگاه براون امریکا. این روند از دههٔ ۹۰ تاکنون شتاب بیشتری پیدا کرده است، به طوری که دو مرکز تولید و توزیع پیکره‌های زبانی در امریکا و اروپا فعالیت دارند: کنسرسیوم داده‌های زبانی (LDC) در امریکا و انجمن اروپایی منابع زبانی (ELRA) در فرانسه. در همین دوران پیکرهٔ ملی زبان انگلیسی (BNC) در سال ۱۹۹۱ برای زبان انگلیسی بریتانیایی با حجم ۱۰۰ میلیون کلمه توسط شورای تحقیقات علوم مهندسی انگلستان و پیکرهٔ زبانی بانک درختی پنسیلوانیا (Penn Treebank) در سال ۱۹۹۶ برای زبان انگلیسی امریکایی با حجم ۴٫۵ میلیون کلمه، و البته همراه با تقطیع نحوی و معنایی توسط دانشگاه پنسیلوانیای امریکا تولید و توزیع شده است و همچنان برای درج اطلاعات زبان‌شناختی در حال توسعه و تکمیل هستند.

۱. اگر داده‌های موجود در پیکرهٔ زبانی را برحسب متغیرهای تعریف‌شده‌ای که مانند: متغیرهای واجی، صرفی، نحوی و یا معنایی و متغیرهای غیر زبانی مانند: سن، جنس و میزان تحصیلات گویشوران زبان، چنان سازماندهی کنیم که کاربر بتواند از طریق یک برنامهٔ رایانه‌ای به جستجوی اطلاعات مورد نیاز بپردازد و به آنها دسترسی داشته باشد، پیکرهٔ زبانی را دادگان گوییم.

نقدهایی بر پیکره‌های زبانی

با ظهور چامسکی از میزان توجه بسیاری از زبان‌شناسان به پیکرهٔ زبانی به عنوان یک منبع اصلی برای توصیف ساخت دستوری زبان کاسته شد.

نقد اول: دستورزبان، دستور زبان درونی است نه زبان بیرونی. وی معتقد بود که پیکرهٔ زبانی در اصل گردایه‌ای از گفته‌های فیزیکی است، یا به عبارت دیگر داده‌های کنشی زبان را شامل می‌شود و بنابراین راهنمای ضعیفی برای انگارهٔ توانش زبانی است (منبع اینترنتی در فهرست منابع).

نقد دوم: علاوه بر آن، زبان مجموعه‌ای نامتناهی از جملاتی است که طول آنها محدود است (چامسکی ۱۹۵۷، ترجمهٔ سمیعی، ۱۱)، در حالی‌که نه‌تنها هر پیکرهٔ زبانی یک مجموعهٔ متناهی از جملات زبان است، و لذا فاقد بسیاری از جملات است که اهل زبان در هر لحظه ممکن است تولید کنند (از جمله جملاتی که در همین مقاله آورده شده‌اند)، بلکه طول جملات تابع عوامل غیرزبانی مانند سبک نویسنده یا گویندهٔ اثر، روش ویراستاران، موضوع اثر و عوامل دیگر است.

نقد سوم: فرض کنیم زبان یک مجموعهٔ متناهی از جملات باشد، در آن صورت اقتصادی است که زبان‌شناس برای تصمیم‌گیری دربارهٔ خوش‌ساخت بودن یک جملهٔ مفروض به شم زبانی خود مراجعه کند تا یک پیکرهٔ زبانی که شامل همهٔ جملات خوش‌ساخت باشد.

اما نکتهٔ حائز اهمیت این است که چامسکی کفایت انگاره‌های آماری را برای توصیف و تبیین ساخت دستوری زبان زیر سؤال برد (همان: ۳۰ـ۲۹) و این به معنی نفی به‌کارگیری پیکره‌های زبانی برای استخراج توزیع آماری یا احتمالی الگوهای زبانی به‌منظور مطالعهٔ یادگیری زبان، حافظهٔ زبانی، جنبهٔ شناختی ساختهای زبانی، مهارتهای زبانی، زبان‌پریشی و بسیاری مسائل دیگر نیست که امروزه در حوزه‌های روان‌شناسی زبان، جامعه‌شناسی زبان، عصب‌شناسی زبان مطرح‌اند.

آنچه چامسکی در فصل سوم کتاب ساختهای نحوی بدان پرداخته است ناظر بر این واقعیت است که با استفاده از یک انگارهٔ مارکفی، یعنی یک فرایند مارکفی با حالات محدود، نمی‌توان دستور زبانهای طبیعی مانند دستورزبان انگلیسی را مدل‌سازی کرد. به عبارت دیگر، اگر دستورزبان انگلیسی را هم‌ارز با یک فرایند مارکفی n مرحله‌ای در نظر بگیریم، خوش‌ساخت بودن بسیاری از جمله‌های بالقوهٔ انگلیسی را با استفاده از این دستور نمی‌توان توضیح داد. این بحث جای تأمل دارد، بنابراین به جزئیات آن می‌پردازیم.

کلود ای. شانن (۱۹۴۸:۴۵ـ۴۲) در مقالهٔ مشهور خود «نظریهٔ ریاضی ارتباط»، هر متن انگلیسی را به مثابهٔ یک نظام اطلاعاتی گسسته در نظر می‌گیرد، به طوری که رشتهٔ کلماتی که اهل

زبان در نوشته‌های خود می‌نویسند تابع یک نظام احتمالاتی مشخص است. وی برای توصیف ریاضی زبان نوشتاری از دانش آماری موجود در ساخت زبان استفاده کرد. برای این منظور ماشینی را در نظر گرفت که در اولین گام برای نزدیک شدن به ساخت زبان انگلیسی، رشتهٔ کلمات را صرفاً بر اساس فراوانی هر کلمه و استقلال کلمات از یکدیگر تولید کند و آن را تقریب مرتبهٔ اول زبان انگلیسی نامید. ماشین رشتهٔ کلمات زیر را تولید کرد:

Representing and speedily is an good apt or come can different natural here he the a in came the to of to expert gray come to furnishes the line message had be these.

در دومین گام برای نزدیک شدن به ساخت زبان انگلیسی، رشتهٔ کلمات را با در نظر گرفتن احتمال وقوع یک کلمه به شرط دانستن کلمهٔ قبل از آن تولید کرد و آن را تقریب مرتبهٔ دوم زبان انگلیسی نامید. ماشین در این گام رشتهٔ کلمات زیر را تولید کرد:

The head and in frontal attack on an English writer that the character of this point is therefore another method ...

همان‌طور که ملاحظه می‌شود رشتهٔ کلماتی که ماشین در گام دوم تولید کرده، به مراتب به یک متن انگلیسی نزدیکتر است تا رشتهٔ کلماتی که در گام اول تولید نموده است. شانن سپس می‌گوید هرچه وقوع یک کلمه مشروط به دانستن تعداد بیشتری از کلمات قبلی باشد، یا به عبارت دیگر از تقریب مرتبه‌های بالاتر زبان استفاده شود، به یک متن طبیعی زبان انگلیسی نزدیکتر می‌شویم. برای این منظور کافی است توزیع احتمال وقوع یک کلمه به شرط آنکه n کلمه قبل از آن را از یک پیکرهٔ متنی زبان انگلیسی به دست آوریم تا به تقریب مرتبهٔ n زبان انگلیسی دست یابیم. وی در ادامهٔ بحث خود اضافه می‌کند که ماشین تولیدکنندهٔ عبارات زبان انگلیسی از فرایندهای مارکف تبعیت می‌کند. هر فرایند مارکف شامل n وضعیت (state)، یعنی S_n، ...، S_2، S_1 و همچنین مجموعهٔ احتمالات $P_{i(j)}$ است، به طوری که هر عضو مجموعه احتمال گذار از وضعیت i را به وضعیت j نشان می‌دهد. بنابراین ماشین مولّد زبان یک فرایند مارکف است که به هنگام گذار از وضعیت i به وضعیت j، یک کلمهٔ معین (W) از واژگان زبان را تولید کند. حال اگر در وضعیت i یک کلمه مانند W_1 حاضر باشد، یعنی احتمال وقوع کلمه W در وضعیت j مشروط به داشتن کلمهٔ W_1 در وضعیت i باشد، آنگاه فرایند مارکف از نوع مرتبهٔ اول، و اگر در وضعیت i دو کلمه مانند W_1W_2 حاضر باشد، یعنی احتمال وقوع کلمهٔ W در وضعیت j مشروط به دانستن رشتهٔ W_1W_2 در وضعیت i باشد، آنگاه فرایند مارکف از نوع مرتبهٔ دوم، و اگر در وضعیت i، n کلمه مانند

$W_1 W_2 \ldots W_n$ حاضر باشد، یعنی احتمال وقوع کلمهٔ W در وضعیت j مشروط به دانستن رشتهٔ n کلمه‌ای $W_1 W_2 \ldots W_n$ در وضعیت i باشد، آنگاه فرایند مارکف از نوع مرتبهٔ n است. چامسکی (۱۹۵۷، ترجمهٔ سمیعی، ۲۱) چنین ماشینی را که با ساخت و کار احتمالات شرطی عبارتهای زبانی را تولید می‌کند، یک دستورزبان مرحله به مرحله یا یک انگارهٔ مارکفی برای زبان نامید.

حال فرض کنیم یک دستورزبان مرحله به مرحله برای زبان فارسی داشته باشیم. در این صورت فارسی‌زبانان برای تولید یا درک عبارتهای زبانی مجهز به یک ماشین (یا فرایند) مارکفی مرتبهٔ n هستند، به‌طوری که اگر وضعیتهایی که ماشین در آن قرار می‌گیرد، مقوله‌های نحوی ـ معنایی زبان باشد، آنگاه به‌هنگام عبور از وضعیت S_i به S_j با احتمال $P_{i(j)}$ کلمهٔ W را تولید یا درک می‌کنند. به‌عنوان مثال جملهٔ زیر را در نظر می‌گیریم:

(۱) دوست علی امروز صبح به دانشگاه رفت.

فارسی‌زبانان برای تولید یا درک جملهٔ (۱) باید مجهز به مجموعهٔ وضعیتها (states) و مجموعهٔ احتمالات گذار (TPROB) باشند. مجموعهٔ وضعیتها برای جملهٔ (۱) را به‌صورت زیر در نظر می‌گیریم:

$$\text{STATES}=\left\{\begin{array}{c}\text{وضعیت پایان، فعل لازم، اسم مکان، حرف اضافهٔ مکانی،}\\ \text{قید زمان، اسم خاص، اسم عام، وضعیت شروع}\end{array}\right\}$$

با این فرض که دستورزبان یک ماشین مارکفی مرتبهٔ ۲ باشد، فارسی‌زبانان باید مجهز به مجموعهٔ احتمالات گذار از وضعیت i به j مشروط به دانستن (حداکثر) دو کلمه در وضعیت i باشند. می‌توان چنین فرض کرد که فارسی‌زبانان در طول یادگیری زبان مادری، این مجموعهٔ احتمالات را از داده‌های گفتاری و نوشتاری به‌دست می‌آورند. بنابراین مجموعهٔ احتمالات برای تولید یا درک جملهٔ (۱) به‌صورت زیر است:

۱. احتمال اینکه به‌هنگام گذار از وضعیت شروع به وضعیت اسم عام کلمهٔ «دوست» تولید یا درک شود.

۲. احتمال اینکه به‌هنگام گذار از وضعیتهای شروع و اسم عام به وضعیت اسم خاص، پس از کلمهٔ «دوست» کلمهٔ «علی» تولید یا درک شود.

۳. احتمال اینکه به‌هنگام گذار از وضعیتهای اسم عام و اسم خاص به وضعیت قید زمان، پس از رشتهٔ دو کلمه‌ای «دوست علی» کلمهٔ «امروز» تولید یا درک شود.

۴. احتمال اینکه به‌هنگام گذار از وضعیتهای اسم خاص و قید زمان به وضعیت قید زمان، پس از رشتهٔ دو کلمه‌ای «علی امروز» کلمهٔ «صبح» تولید یا درک شود.

۵. احتمال اینکه به‌هنگام گذار از وضعیتهای قید زمان و قید زمان به وضعیت حرف اضافهٔ مکانی، پس از رشتهٔ دو کلمه‌ای «امروز صبح» کلمهٔ «به» تولید یا درک شود.

۶. احتمال اینکه به‌هنگام گذار از وضعیتهای قید زمان و حرف اضافهٔ مکانی به وضعیت اسم مکان، پس از رشتهٔ دو کلمه‌ای «صبح به» کلمهٔ «دانشگاه» تولید یا درک شود.
۷. احتمال اینکه به‌هنگام گذار از وضعیتهای حرف اضافهٔ مکانی و اسم مکان به وضعیت فعل لازم، پس از رشتهٔ دو کلمه‌ای «به دانشگاه» کلمهٔ «رفت» تولید یا درک شود.
۸. احتمال اینکه به‌هنگام گذار از وضعیتهای اسم مکان و فعل لازم، پس از رشتهٔ دو کلمه‌ای «دانشگاه رفت»، پایان جمله تولید یا درک شود.[۱]
در اینجا ذکر دو نکته حائز اهمیت فراوان است:

نکتهٔ اول: در یک دستورزبان مبتنی بر ماشینِ مارکفیِ مرتبهٔ n، تولید و درک یک کلمه به n کلمهٔ قبلی وابسته است. به‌عبارت دیگر اطلاعات مربوط به مرتبه‌های بالاتر از n در حافظهٔ ماشین وجود ندارد (استرگوس ۲۰۰۱:۲۴). به‌عنوان مثال در جملهٔ (۱) احتمال تولید کلمهٔ «رفت» صرفاً تابع رشتهٔ دو کلمه‌ایِ «به دانشگاه» و مستقل از کلماتِ «دوست علی امروز صبح» است، زیرا دستورزبان یک ماشین مارکف دومرحله‌ای تعریف شده است. در حالی‌که اگر یک فارسی‌زبان بعد از «به دانشگاه» بخواهد از فعل لازم «رفتن» در پایان جمله استفاده کند، احتمال تولید شش صورت تصریفی فعل بر حسب مطابقت با فاعل به‌طور مساوی وجود دارد. در واقع آنچه باعث می‌شود صورت سوم شخص مفرد فعل تولید شود، عبارت «دوست علی» است که حداکثر فاصله را با فعل «رفت» دارد. بنابراین اگر بخواهیم به دستورزبان مرحله به مرحله در توصیف ساخت زبان وفادار بمانیم، برای تولید یا درک جملاتی چون (۱) نیاز به یک ماشین مارکفی شش‌مرحله‌ای خواهیم داشت. این بدان معنی است که اهل زبان باید در ذهن خود احتمال وقوع هر کلمه را به شرط هر شش کلمهٔ قبل از آن در حافظه ذخیره کرده باشند. به این ترتیب اگر فرض کنیم حجم واژگان زبان فارسی ۱۰۰۰ مدخل باشد، تعداد عبارتهای به‌طول ۶ فارسی برابر با $۱۰۰۰^{۶}$ خواهد بود. اگرچه احتمال وقوع بسیاری از این عبارتها صفر است ولی فرضیهٔ مجهز بودن حافظهٔ اهل زبان به احتمالاتی از این دست قابل دفاع نیست. علاوه‌بر آن اگر فاصلهٔ بین دو عنصر زبانی که به‌لحاظ ساخت وابسته هستند، مانند فاعل و فعل، زیادتر شود، باید به عبارتهای با طول بیشتر از شش کلمه قائل شد. به‌عنوان مثال اگر طول جملهٔ (۱) با نشاندن دو بند موصولی بیشتر شود، «دوست علی که او را می‌شناسی، صبح امروز به دانشگاهی که برادرم در آن تحصیل می‌کند، رفت»، در این صورت برای تولید یا درک جملهٔ اخیر نیاز به یک ماشین مارکف شانزده‌مرحله‌ای

۱. باید توجه داشت که در این مقاله به شیوهٔ محاسبهٔ احتمال گذار که نیاز به مباحث آماری و نظریهٔ احتمال بیزی (Bayesian) دارد، نمی‌پردازیم.

خواهیم داشت، زیرا وقوع فعل «رفت» تابع فاعل «دوست علی» است که به‌اندازهٔ ۱۶ کلمه از آن فاصله دارد، یعنی اهل زبان برای تولید یا درک جملاتی از این نوع، باید احتمال وقوع هر کلمه را به شرط هر شانزده کلمهٔ قبل از آن در ذهن خود ذخیره کرده باشد.

چامسکی (۱۹۵۷، ترجمهٔ سمیعی، ۲۶) به ساختهایی از زبان انگلیسی اشاره می‌کند که فاصلهٔ بین عناصر وابسته در اثر قرارگرفتن یک جمله بین آن دو عنصر، می‌تواند بالقوه زیاد شود و بنابراین برای توصیف چنین ساختهایی با استفاده از ماشین مارکفی n مرحله‌ای، باید قائل به وجود مجموعهٔ احتمالات وقوع یک کلمه به شرط n کلمهٔ قبل از آن (طول جملهٔ درج‌شده بین آن دو عنصر) شد، به‌طوری که n می‌تواند بالقوه عدد بزرگی باشد. ساختهای مزبور عبارت‌اند از:

(2) If S_1, then S_2.

(3) Either S_3, or S_4.

(4) The man who said that S_5, is arriving today.

در جملهٔ (۲) وقوع «then» مشروط به وقوع «if» است، در حالی‌که جملهٔ S_1 بین این دو قرار دارد. در جملهٔ (۳) وقوع «or» مشروط به وقوع «either» است، در حالی‌که جملهٔ S_3 بین این دو قرار دارد. در جملهٔ (۴) وقوع «is» مشروط به وقوع «the man» یا هر اسم جاندار مفرد است، در حالی‌که جملهٔ S_5 بین این دو قرار دارد. حال اگر طول جملات S_1، S_3 و S_5 زیاد باشد (در مقام نظر می‌تواند بی‌نهایت باشد)، فاصلهٔ بسیار زیادی بین این عناصر وابسته به‌وجود خواهد آمد و در آن صورت، طرح مجموعهٔ احتمالات وقوع کلمات then، or و is به شرط بی‌نهایت کلمهٔ قبل بیهوده خواهد بود. در نهایت چامسکی (۱۹۵۷، ترجمه سمیعی، ۲۴) نتیجه می‌گیرد که ساختن یک ماشین مارکفی که ساختهای نحوی فوق‌الذکر را توصیف کند، نه‌تنها دشوار بلکه محال است. زیرا اولاً هیچ پیکرهٔ زبانی وجود ندارد که در آن به تعداد کافی جملاتی از نوع (۲)، (۳) و (۴) با طولهای متفاوت S_1، S_3 و S_5 وجود داشته باشد، و ثانیاً با فرض وجود چنین پیکره‌ای، حتی با در اختیار داشتن پیشرفته‌ترین رایانه‌ها محاسبهٔ مجموعهٔ احتمالات مشروط زمان‌بر و پرهزینه خواهد بود.

نکتهٔ دوم: محدودیتهای گزینشی کلمات بیشتر درون‌ساختی هستند تا بین‌ساختی. به‌عبارت دیگر، این گرایش در زبانها وجود دارد که کلماتی که درون یک گروه نحوی هستند، همدیگر را بیشتر محدود می‌کنند تا کلماتی که در مرز بین دو گروه نحوی قرار دارند. به‌عنوان مثال در جملهٔ (۱)، احتمال تولید «علی» بعد از کلمهٔ «دوست» به‌مراتب بیشتر از احتمال تولید «امروز» یا «امروز صبح» بعد از «دوست علی» است، زیرا «دوست علی» یک گروه اسمی است که در آن پس از تولید

«دوست» احتمال وقوع بسیاری از کلمات به‌عنوان متمم هستهٔ گروه اسمی صفر یا کم می‌شود و با احتمال زیادی به یک اسم یا ضمیر جاندار محدود می‌گردد. در حالی‌که پس از «دوست علی» احتمال وقوع هر گروه نحوی و در نتیجه هر کلمهٔ فارسی تقریباً به یک اندازه وجود دارد و حتی می‌توان گروه قیدی «امروز صبح» را قبل از «دوست علی» تولید کرد. به‌همین دلیل است که گفته می‌شود آرایش کلمات در جملات فارسی آزاد است، یا زبان فارسی یک زبان دارای scrambling است.

چون ماشینهای مارکفی در تولید عبارتهای زبانی از یک انگارهٔ خطی تبعیت می‌کنند، به‌طوری که مرزهای نحوی را دخالت نمی‌دهند و بالمآل در محاسبهٔ توزیع احتمالات مشروط، رشته‌های به‌طول n کلمه را با یک چشم نگاه می‌کنند، نمی‌توانند انگارهٔ درستی از محدودیتهای گزینشی کلمات در ساختهای نحوی به‌دست دهند. بنابراین توصیفی که یک دستورزبان مبتنی بر ماشین مارکفی مرتبهٔ n از تولید یا درک عبارتهای زبانی به‌دست می‌دهد، کفایت توضیحی ندارد.

چامسکی، در نهایت، دستورزبان در چارچوب انگارهٔ مارکفی را کنار گذاشت و برای توصیف ساخت زبان به دستور ساخت گروهی و گشتارها متوسل شد (چامسکی ۱۹۵۷، ترجمهٔ سمیعی). اما این به مفهوم عدم کارآیی انگارهٔ مارکفی در توصیف ساختهای زبانی نیست. به‌عنوان مثال، همایندها (collocations) ساختهایی هستند که از ترکیب دو یا چند کلمه به‌دست می‌آیند، به‌طوری که معمولاً دلالت بر معنی غیرترکیبی (non-compositionl) دارند، یعنی معنی ترکیب تابعی از معنی کلمات سازنده نیست. چامسکی (۱۹۱:۱۹۶۴) از همایندها با عنوان «ساختهای بسته» یاد می‌کند. این ساختها را نمی‌توان با آزمونهای نحوی از قبیل همپایگی، جانشینی یا گسترش مطالعه و توصیف کرد، زیرا اساساً یک ترکیب نحوی نیستند، بلکه یک ترکیب ثابت از کلماتی هستند که برای افادهٔ یک معنای مشخص به یکدیگر جوش خورده‌اند و می‌توانند بخشی از واژگان ذهنی اهل زبان محسوب شوند (بن‌سون و دیگران، ۱۹۸۶). بعضی از همایندهای فارسی عبارت‌اند از: در این باره، از آنجا که، در صورتی که، در این صورت، راجع به، نسبت به، بر روی، در روی، بر اساس، خروج از، علاقه به، خواب رفتن، از درس/پا افتادن و ...» (صفوی ۱۳۸۰). یکی از مسائل جالب زبان‌شناختی می‌تواند این باشد که کدام ترکیب از کلمات فارسی را می‌توان همایند نامید. برای پاسخ به این سؤال چاره‌ای جز بررسی آماری یک پیکرهٔ جامع زبانی نیست. در مرحلهٔ اول با استفاده از انگارهٔ مارکفی می‌توان ترکیبهای n کلمه‌ای را که احتمال وقوع بیشینه دارند، به‌عنوان گزینه‌هایی برای همایندهای زبانی در نظر گرفت و در مرحلهٔ دوم ترکیبهایی که به آزمونهای نحوی پاسخ مثبت ندهند، به‌عنوان همایندهای فارسی در نظر گرفت.

دستورزبان و پیکرهٔ زبانی

چامسکی (۱۹۵۷، ترجمه سمیعی، ۱۴) از یک‌سو مجموعهٔ جمله‌های دستوری را با هیچ پیکره‌ای از گفته‌ها، که زبان‌شناس برای تحقیقات زبانی گردآوری می‌کند، همسان نمی‌داند، و از سوی دیگر معتقد است دستورزبان از رهگذر یک پیکرهٔ زبانی محدود و تصادفی از گفته‌ها، توانایی توصیف بی‌نهایت گفته یا جملهٔ دستوری را پیدا می‌کند. با فراگیر شدن آراء چامسکی در امریکا و اروپا توجه بیشتر زبان‌شناسان به توصیف الگوها یا ساختهای زبانی که بالقوه می‌توان تولید کرد معطوف شد و نه به ساختهایی که عملاً تولید می‌شوند و در پیکره‌های زبانی وجود دارند. به این ترتیب، جمع‌آوری پیکره‌های زبانی از دههٔ ۶۰ به بعد به‌تدریج اهمیت خود را از دست دادند. اما با شروع دههٔ ۹۰ و برگزاری دوسالانهٔ همایشهای بین‌المللیِ پردازشِ زبانهای گفتاری[1] و یوروسپیچ[2] در امریکا، کانادا، اروپا و شرق آسیا و اهمیت روزافزون بازشناسی و بازسازی رایانه‌ای زبان و گفتار و ترجمهٔ ماشینی توجه زبان‌شناسان مجدداً به پیکره‌های زبانی معطوف گشت (بی‌جن‌خان ۱۳۷۱: ۱۱۴ـ۱۰۸). پیترلدفوگد (۱۹۹۲، ترجمهٔ بی‌جن‌خان، ۹۴ـ۸۴) در سخنرانی افتتاحیهٔ دومین همایش بین‌المللی پردازش زبانهای گفتاری که در بانف (Banff) کانادا برگزار شد، چنین گفت:

«من جداً معتقدم که استفاده از بانکهای داده‌ای عظیم (پیکره‌های زبانی) ممکن است به دگرگونی کامل آواشناسی، واج‌شناسی و حتی احتمالاً کل زبان‌شناسی در سالهای آتی منجر شود . . . از زمان ظهور نوام چامسکی، زبان‌شناسی بر توصیف توانش گوینده، یعنی ساخت ذهنی زبان و نه تظاهر زبان در گفتار، تأکید داشته است . . . می‌توان از یک رویکرد متفاوت با رویکرد ذهنی سود جست. عاقلانه نیست به زبان‌شناسی امریکایی در آغاز دههٔ ۱۹۵۰ برگردیم، زمانی‌که توصیف زبان صرفاً شامل داده‌هایی بود که در پیکره به‌حساب می‌آمدند. اما به همین اندازه نیز غیرعاقلانه است که همچنان روی صندلی راحتی بنشینیم و در خصوص زبان گفتاری که در ذهن یک گویندهٔ فرضی قرار دارد، حکم قطعی بدهیم.»

در دههٔ ۹۰ اوج توجه زبان‌شناسان و مهندسان فن‌شناسی اطلاعات و ارتباطات به جمع‌آوری پیکره‌های زبانی، اعم از گفتاری و نوشتاری، به‌منظور آماده‌سازی مواد خام زبانی برای مدل‌سازی مهندسی تواناییها و مهارتهای زبانی انسان در سامانه‌های رایانه‌ای بوده است، از قبیل بازشناسی و تولید گفتار و بازشناسی هویت گویشوران زبان از طریق گفتار. با توسعهٔ فن‌شناسی اطلاعات و جمع‌آوری بیشتر پیکره‌های زبانی، توجه زبان‌شناسان حتی زبان‌شناسان زایشی به نحوهٔ استفاده از اطلاعات موجود در پیکره‌ها برای بهینه‌سازی یافته‌های نظری خود جلب شد. در اینجا به ذکر دو مثال از زبان آلمانی به نقل از دبلیو. دتمار مویررس (W. Detmar Meurers) و دو مثال از زبان فارسی می‌پردازیم.

1. International Conference on Spoken Language Processing (ICSLP)
2. Eurospeech

مثال اول: افعال AcI

مویررس (۲۰۰۲) آورده است که پتر سوخسلند (Peter Suchsland) زبان‌شناس آلمانی با مطالعهٔ ساخت گروه فعلی زبان آلمانی به این نتیجه رسیده است که در ساختهای با نمود کامل، افعالی چون sehen (دیدن) یا hören (شنیدن) (افعال AcI) همیشه به‌صورت مصدری و نه به‌صورت وجه وصفی ماضی ظاهر می‌شوند. به‌عنوان مثال، جملهٔ زیر را در نظر می‌گیریم.

Er hat ihn über die Straße gehen **sehen**/***gesehen**
He has him over the street go see inf/seen past-part

همان‌طور که مشاهده می‌شود فعل «دیدن» با نمود کامل به‌صورت مصدری «sehen» ظاهر شده است نه به‌صورت صفت مفعولی «gesehen». اما مویررس با استفاده از پیکرهٔ متون دو روزنامهٔ آلمانی با حجم ۴۰۹، ۰۳۹، ۴۸۰ کلمه به مثالهای نقض برای یافته‌های نظری سوخسلند دست یافت، که در آنها افعال «دیدن» و «شنیدن» با نمود کامل به‌صورت صفت مفعولی «gesehen» و «gehört» وجود داشتند.

مثال دوم: افعال وجه‌نما

مویررس (۲۰۰۲) یکی از مسائل نظری در حوزهٔ گروه فعلی زبان آلمانی را ناشی از این می‌داند که چون متمم یک فعل وجه‌نما می‌تواند فعل وجه‌نمای دیگری باشد، بنابراین اینکه چه محدودیتهایی در توالی دو فعل وجه‌نما وجود دارد و هر توالی چه تعبیر معنایی دارد، از مباحث جالب نحوی و معنایی است. وی معتقد است برای پاسخ به این‌گونه مسائل می‌توان از پیکره‌های متنی با حجم خیلی زیاد استفاده کرد.

مثال سوم: حرف اضافهٔ مرکب

در زبان فارسی حروف اضافهٔ بسیط «به، از، با، در، بر» با سایر حروف اضافه مانند «روی، بالای، پایین، زیر، ...» ترکیب می‌شوند و حرف اضافهٔ مرکب می‌سازند (ابوالحسنی چیمه ۱۳۸۱)، مانند: از روی، در روی، بر روی، در بالای، از پایین، بر بالای، از بالای، ...

در این خصوص تحقیق دربارهٔ دو سؤال اهمیت دارد. اولاً چه محدودیتهایی در ترکیب حروف اضافهٔ بسیط با یکدیگر وجود دارد؛ و مهمتر از آن تعبیر معنایی هر ترکیب ممکن چیست. ساده‌ترین و مطمئن‌ترین روش تحقیق، جستجو در یک پیکرهٔ زبانی با حجم بسیار زیاد است تا از رهگذر مطالعهٔ بافت زبانی تعبیر معنایی هر حرف اضافهٔ مرکب را به‌دست آورد.

مثال چهارم: حرف اضافهٔ مرکب یا گروه حرف اضافه‌ای

در دستورزبان فارسی این بحث مطرح است که ترکیب حاصل از باهمایی حروف اضافهٔ بسیط

و کلماتی که نوعاً دارای مقولهٔ اسم با مشخصه‌های معنایی {انتزاعی، غیرارجاعی، غیرمشخص} هستند، یک حرف اضافهٔ مرکب است یا گروه حرف اضافه‌ای (ابوالحسنی چیمه ۱۳۸۱: فصل پنجم). بعضی از این ترکیبها عبارت‌اند از: دربارهٔ براساس، برحسب، باوجود، به‌صرف و در مورد، که در آنها «در، بر، با، به» حرف اضافهٔ بسیط و «باره، اساس، حسب، وجود، صرف و مورد» اسم بامشخصه‌های معنایی فوق‌الذکر هستند. بر طبق نظر فرشیدورد (۱۳۵۱) چون این ترکیبها از بسامد وقوع بالایی برخوردارند، پس یک کلمهٔ (حرف اضافهٔ) مرکب محسوب می‌شوند، در حالی‌که اگر چنین باشد بسیاری از گروههای حرف اضافه‌ای مانند: از من، به خانه، با همه و . . . که چه‌بسا بسامد بالاتری در پیکره‌های زبانی داشته باشند، باید کلمهٔ مرکب محسوب شوند (ابوالحسنی چیمه ۱۳۸۱، فصل پنجم). اما با استفاده از انگارهٔ مارکفی می‌توان این فرضیه را مطرح کرد که اگر ترکیبهای موردنظر حرف اضافهٔ مرکب باشند، انتظار می‌رود احتمال گذار از این‌گونه ترکیبها به گروه اسمی با احتمال گذار از حرف اضافهٔ بسیط به گروههای اسمی تفاوت معنی‌داری نداشته باشد. برای آزمون این فرضیه نیز باید به یک پیکرهٔ زبانی با حجم بسیار زیاد متوسل شد و معنی‌دار بودن این دو احتمال را با استفاده از یک آزمون آماری مناسب به‌دست آورد.

نرم‌افزار پیکرهٔ زبانی

نرم‌افزار هر پیکرهٔ زبانی تسهیلات موردنیاز برای انواع جستجو در پیکره و همچنین تجزیه و تحلیل آماری داده‌ها را به همراه گزارشهای مربوطه فراهم می‌کند. نرم‌افزاری که نگارنده برای طراحی و تهیهٔ یک پیکرهٔ زبانی با حجم حدود ۸٫۵میلیون کلمه از آن استفاده کرده، دارای مشخصات زیر است:

۱ جمع‌آوری داده‌ها

در جمع‌آوری داده‌های الکترونیکی چند متغیر اساسی را باید در نظر گرفت، از جمله موضوع و نوع داده. از آنجا که یک استاندارد واحد برای نویسه‌های خط فارسی وجود ندارد (عاصی ۱۳۷۹) و مراکز یا منابع تولید اطلاعات متنی از قلمهای رایانه‌ای متعددی استفاده می‌کنند، جمع‌آوری نظام‌مند متون فارسی کار ساده‌ای نیست. بنابراین در تهیهٔ هر پیکرهٔ متنی زبان فارسی چاره‌ای جز انتخاب یک استاندارد واحد و تعریف یک نگاشت چند به یک از نویسه‌های یک حرف در قلمهای متعدد رایانه‌ای به یک نویسهٔ مشخص برای آن حرف در استاندارد انتخاب‌شده مانند استاندارد میکروسافت، وجود ندارد. همچنین به مجموعهٔ نویسه‌های فارسی باید بعضی از نویسه‌های حروف عربی را که در متون فارسی به‌کار می‌روند، مانند «ي، ؤ» نیز افزود. علاوه‌بر آن، در جمع‌آوری داده‌ها توزیع فراوانی متون بر حسب موضوع باید یکنواخت و عنوان موضوع نیز کلی باشد.

اما آنچه که در تهیهٔ پیکرهٔ زبانی اهمیت زیادی دارد، نوع داده است. براساس طبقه‌بندی

اسوارتویک (۱۹۹۵)، داده‌های زبانی یا گفتاری هستند یا نوشتاری. داده‌های گفتاری یا دوطرفه (dialogue) هستند یا یک‌طرفه (monologue). داده‌های گفتاری دوطرفه یا به‌صورت گفتگوی خصوصی (conversation) دو نفر با همدیگر است، یا به‌صورت گفتگوی دو نفر در ملاء عام، مانند مصاحبه یا مناظره. داده‌های گفتاری دوطرفهٔ خصوصی یا به‌صورت رودررو است یا به‌صورت تلفنی. در دو حالت اخیر، داده‌های گفتاری محرمانه (surreptitous) است، اگر حداقل یکی از شرکت‌کنندگان در گفتگو نداند مکالمه‌شان ضبط می‌شود و در غیراین‌صورت داده‌های گفتاری غیرمحرمانه (non-surreptitous) محسوب می‌شود. اما دادهٔ گفتاری یک‌طرفه یا فی‌البداهه (spontaneous) است، مانند گفتار کسی که بدون برنامه‌ریزی قبلی دربارهٔ موضوعی سخن بگوید، یا آماده (prepared)، مانند سخنرانی شخصیتهای سیاسی از روی یک نوشته، یا سخنان یک وکیل در صحنهٔ دادگاه براساس آمادگی قبلی به‌منظور اقامهٔ شواهد به نفع موکل خود.

از سوی دیگر، داده‌های نوشتاری بر سه نوع‌اند: چاپی، غیرچاپی و خواندنی. داده‌های نوشتاری چاپی یا جنبهٔ اطلاعاتی دارند، مانند متون علمی، روزنامه‌ها، جملات، کتاب، مقالات و نامه‌های اداری و یا جنبهٔ آموزشی دارند مانند کتب درسی و کتابهای راهنما و یا جنبهٔ انگیزشی، مانند متون مذهبی و بعضی متون سیاسی. داده‌های نوشتاری غیرچاپی شامل دست‌نوشته‌ها، یادداشتها و انواع نامه‌های شخصی است و بالاخره داده‌های نوشتاری خواندنی شامل متونی است که با قصد خواندن یا روخوانی از آنها نوشته می‌شوند، مانند نمایشنامه‌ها، خطابه‌ها، داستانها و متن اخبار.

درونداد نرم‌افزار یک پروندهٔ متنی با پسوند txt.، متعلق به یکی از انواع داده‌های نوشتاری است که نویسه‌های آن منطبق با نویسه‌های سینا[1] است. داده‌هایی که تاکنون در پیکره گنجانده شده‌اند از روزنامه، مجله و کتاب جمع‌آوری شده‌اند.

۲ آماده‌سازی داده‌ها

آماده‌سازی داده‌ها شامل ویرایشهای زبانی است که به‌صورت نیمه‌خودکار بر پرونده‌های متنی اعمال می‌شوند. ضرورت انجام این فعالیت از آنجا ناشی می‌شود که متون فارسی با سلیقه‌های متفاوت تایپ یا نوشته می‌شوند. مرز کلمه نیز در بسیاری از موارد مشخص نیست که خود باعث تنوع و عدم یکدستی در نگارش متون می‌شود. علاوه‌برآن، خطای املایی در تایپ یا نوشتن متن محتمل‌الوقوع است. بنابراین، آماده‌سازی داده‌ها دلالت بر تقطیع متون به کلمات زبان‌شناختی با استفاده از ویرایشهای نیمه‌خودکار دارد.

۱. چون طرح پژوهشی پیکرهٔ زبانی را نگارنده از سال ۱۳۷۷ شروع کرده، مجبور به استفاده از نویسه‌های سینا بوده است. در نسخهٔ جدید نرم‌افزار از نویسه‌های سیستم عامل ویندوز ۲۰۰۰ (مایکروسافت) استفاده می‌شود.

بعضی از ویرایشها عبارت‌اند از:

۱. چسباندن واحدهای نوشتاری که در واقع یک کلمهٔ تصریفی هستند، مانند چسباندن «می» به «رفت» برای ساختن کلمهٔ تصریفی «می‌رفت».

۲. چسباندن واحدهای نوشتاری که در واقع یک کلمهٔ اشتقاقی هستند. مانند چسباندن «ناک» به «خطر» برای ساختن کلمهٔ اشتقاقی «خطرناک».

۳. چسباندن واحدهای نوشتاری که در واقع یک کلمهٔ مرکب هستند. مانند چسباندن «کنندگان» به کلمهٔ قبل از آن برای ساختن ترکیبهایی چون «برگزارکنندگان» و «توزیع‌کنندگان».

۴. جداسازی واحدهای سازندهٔ یک ترکیب نوشتاری و بالمال تجزیهٔ آن به دو کلمهٔ جدید. مانند تجزیهٔ «آنمرحوم» به کلمات «آن» و «مرحوم»، یا تجزیهٔ «بعلت» به کلمات «به» و «علت».

برای این قبیل ویرایشها که در متون الکترونیکی فارسی به‌وفور می‌توان یافت، وندهای تصریفی و اشتقاقی و همچنین کلمات سازندهٔ بعضی از ترکیبها در نرم‌افزار تعریف شده‌اند و ویرایش زبانی به‌صورت خودکار انجام می‌شود.

اما حجم بسیار زیادی از متون را باید به‌طور دستی تقطیع کرد. به‌عنوان مثال، واحد نوشتاری «وبا» می‌تواند اسم یک بیماری باشد، که در آن‌صورت یک کلمه محسوب می‌شود، اما اگر ترکیبی از حرف ربط «و» و حرف اضافهٔ «با» باشد، باید به کلمات سازنده‌اش تجزیه شود. همچنین در عبارت «بنابراین گزارش» اگر «این» وابستهٔ اشاره در گروه اسمی «این گزارش» باشد، یا بخشی از قید یا حرف ربط «بنابراین» باشد، به‌ترتیب به دو صورت «بنا # بر # {این # گزارش}» و «بنابراین # گزارش»[1] تجزیه می‌شود.

۵. یکی دیگر از ویرایشهای زبانی ناظر بر واحدهای نوشتاری متفاوت است که دلالت بر یک کلمهٔ نوشتاری دارد، و بالعکس یک واحد نوشتاری که می‌تواند دلالت بر دو کلمهٔ نوشتاری متفاوت داشته باشد. به‌عنوان مثال، سه واحد نوشتاری «همه»، «همهٔ» و «همه‌ی» در ترکیبهای «همه/همهٔ/همه‌ی کتابها» دلالت بر سور عمومی دارند که دارای کسرهٔ اضافه است. مثال دیگر، اینکه در استاندارد یونی‌کد (unicode) به نویسه‌های «ه، ﻫ، ﻩ» یک کد اختصاص داده می‌شود، بنابراین اگر در تایپ کلمهٔ «یگانه‌ای» بین «یگانه» و «ای» به‌جای استفاده از نویسهٔ نیم‌فاصله[2] (pseudo-space) از نویسهٔ فاصله استفاده شود،چون نرم‌افزار «ای» را به‌عنوان پسوند به «یگانه» می‌چسباند، در آن صورت پس از ویرایش کلمهٔ «یگانهای» به‌دست می‌آید. به این ترتیب، واحد «یگانهای» می‌تواند به

۱. البته در این حالت انتظار می‌رود که بین دو کلمه یک ویرگول وجود داشته باشد، اما علائم نقطه‌گذاری در بسیاری از متون به‌درستی به‌کار نمی‌رود.

۲. با استفاده از این نویسه می‌توان نشان داد هنوز کلمهٔ نوشتاری کامل نشده است، در حالی‌که نویسهٔ فاصله (space) برای تعیین مرز کلمه استفاده می‌شود.

دو کلمهٔ نوشتاری متفاوت «یگانهای» و «یگانه‌ای» دلالت داشته باشد. در این‌گونه موارد با استفاده از برچسب نحوی-معنایی ابهام‌زدایی می‌شود. به این ترتیب که «همه»، «همهٔ» و «همه‌ی» برچسب «سور-کسرهٔ اضافه» دارند و لذا یک کلمهٔ نوشتاری واحد محسوب می‌شوند، و «یگانهای» در ترکیب «خداوند # یگانهای» اگر دارای برچسب «صفت ساده-یاء نکره» باشد، به‌صورت کلمهٔ «یگانه‌ای» در نظر گرفته می‌شود.

۳ برچسب‌دهی نحوی-معنایی

برچسب نحوی-معنایی عبارت است از یک کد چندحرفی که نشانگر مقولهٔ نحوی زیرمقوله‌های نحوی و معنایی (در صورت نیاز) کلمه است (لیچ و دیگران ۱۹۹۴). به‌عنوان مثال، کلمهٔ «علوی» دارای، برچسب «SPN» است که نشانگر «اسم مفرد خاص» و کلمهٔ «آبادان» دارای برچسب «SPLN» است که نشانگر «اسم مفرد مکان خاص» است. در نرم‌افزار مقوله‌های نحوی و زیرمقوله‌های نحوی و معنایی در قالب بازنمایی درختی در شش پنجرهٔ متوالی با عناوین فارسی آنها تعریف شده‌اند. کاربر می‌تواند به دلخواه آنها را تغییر دهد و همزمان با تقطیع نیمه‌خودکار پروندهٔ متنی در مرحلهٔ آماده‌سازی متون به هر کلمه پربسامدترین برچسب نحوی-معنایی را، با استفاده از یک فرهنگ لغت که در جریان تولید پیکره ساخته و تکمیل می‌شود، اختصاص دهد. این فعالیت را می‌توان برچسب‌دهی خودکار نامید. اما چون برچسب کلمات با توجه به بافت زبانی تغییر می‌کند، (عاصی و حاج عبدالحسینی ۲۰۰۱) به ناچار برچسب بعضی از کلمات باید به‌طور دستی اصلاح شود. امکانات لازم برای اصلاح دستی و سریع برچسبها در نرم‌افزار وجود دارد.

پس از تقطیع برچسب‌دهی نحوی-معنایی، یک پروندهٔ جدید با همان نام پروندهٔ متنی ولی با پسوند lmf. ساخته می‌شود که حاوی شناسنامهٔ متن و کلیهٔ اطلاعات مربوط به برچسب کلمات است.

۴ آمارگان

منظور از آمارگان مجموعهٔ برنامه‌های رایانه‌ای است که برحسب نیاز کاربر بر روی پرونده‌های با پسوند lmf. اجرا شده و نتایج آن به‌صورت فهرستهای آماری قابل نمایش و چاپ است. بعضی از عملیات زبانی که در برنامه‌ها پیش‌بینی شده، عبارت‌اند از: باهمایی کلمات در چارچوب انگارهٔ مارکفی، شامل بسامد نسبی هر کلمه، ترکیبهای دوکلمه‌ای، سه‌کلمه‌ای و چهارکلمه‌ای به‌منظور استخراج ساختهای همایند، باهمایی مقوله‌های دستوری (برچسبهای نحوی-معنایی)، شامل بسامد نسبی هر مقوله، ترکیبهای دومقوله‌ای و چهارمقوله‌ای، توزیع احتمال وقوع هر کلمه به شرط مقولهٔ دستوری حداکثر سه کلمهٔ قبل از آن، واژگان بسامدی، شناسایی هم‌نویسه‌ها و

بافتهای زبانی آنها، استخراج مطابقه‌ها (concordances) (بارن‌بروک ۱۹۹۶) و گزارش‌گیری برحسب فراوانی و ترتیب حروف الفبای فارسی.

بحث دربارهٔ مباحث فوق در حوصلهٔ این مقاله نمی‌گنجد، اما یکی از مهمترین فعالیتهای این نرم‌افزار که در دست ایجاد و توسعه است، فراهم نمودن جستجوهای هوشمند در پیکرهٔ زبانی برای آزمون یافته‌های زبان‌شناسی نظری است.

جستجوی هوشمند در پیکرهٔ زبانی

در یک پیکره دو نوع جستجوی رایانه‌ای برای عناصر زبانی می‌توان در نظر گرفت. نوع اول، جستجوی بخشی از کلمه یا کلمه یا چند کلمهٔ متوالی است، صرف‌نظر از اطلاعات ساختاری. این قابلیت در بسیاری از ویراستارهای رایانه‌ای وجود دارد. در این حالت، کاربر باید مورد جستجو را عیناً وارد کند تا رایانه عملیات جستجو را انجام دهد. به‌عنوان مثال، اگر کاربر بخواهد کلمات اشتقاقی با پسوند «مند» را جستجو کند، مورد جستجو را باید به‌صورت «مند< blank >»[۱] وارد کند. اگرچه نتیجهٔ عملیات جستجو شامل کلماتی چون «سمند، کمند، می‌نامند، ...» خواهد شد، اما می‌توان آنها را در تحلیل خود به‌حساب نیاورد. نوع دوم، جستجوی هر رشته عناصر زبانی با توجه به ساخت نحوی است. به‌عنوان مثال، اگر کاربر یا محقق بخواهد مطابقت فاعل و فعل فارسی را در پیکره برحسب مؤلفه‌های معنایی شمار و جانداری تجزیه و تحلیل آماری کند، نمی‌تواند از جستجوی نوع اول استفاده کند، زیرا نیاز به گنجاندن اطلاعات ساختاری مانند مفرد/جمع و جاندار/بی‌جان در «مورد جستجو» را دارد. برای حل این مسئله و بسیاری از مسائل زبان‌شناسی نظری، طراحان دادگان یک زبان، جستجو در پیکرهٔ زبانی را تعریف می‌کنند. درونداد زبان جستجو، سؤال (query) جستجو است که کاربر بر طبق یک دستورالعمل می‌پرسد و برونداد آن پاسخی است که رایانه پس از پردازش در پیکره در اختیار محقق می‌گذارد. در این صورت، رایانه هوشمندانه جستجو می‌کند و میزان پیچیدگی آن تابعی است از میزان اطلاعات ساختاری که در قالب برچسبهای نحوی-معنایی برای داده‌های پیکره تعریف شده‌اند. بنابراین زبان‌شناس برای دسترسی به داده‌های موردنظر خود در پیکره باید بتواند مسئلهٔ زبان‌شناختی خود را به زبان جستجو ترجمه کند.

زبان جستجو دارای سه بخش اصلی است: واژگان، نحو و معنی‌شناسی

واژگان شامل نویسه‌های حرفی، عددی و علامتی (مانند پرانتز، کروشه، آکولاد و ...)، متغیرها، ثوابت (constants) و عملگرهای منطقی است. متغیر بر دو نوع است: WORD که برای تعریف کلمه و POS که برای تعریف مقولهٔ نحوی-معنایی به‌کار می‌رود. ثوابت نیز بر دو نوع‌اند:

۱. منظور از < blank > فاصله‌ای است که می‌توان با فشردن کلید جای خالی روی صفحه کلید، آن‌را وارد کرد.

صورت نوشتاری یا واجی کلمه و علائم اختصاری برای مقولهٔ نحوی-معنایی کلمات. عملگرهای منطقی شامل باهمایی (concatenation)، ترکیب عطفی (conjunction)، ترکیب فصلی (disjunction) و نفی (negation) است. نحو زبان جستجو شامل قواعد ترکیب عناصر واژگانی با یکدیگر برای ساختن سؤال جستجو است. قواعد ترکیب ناظر به الگوهای خوش‌ساخت در نحو زبان مورد مطالعه است. و بالاخره معنی‌شناسی زبان جستجو که مبتنی بر روش انطباق الگو (pattern matching) است. بر این اساس، الگوی نحوی موجود در سؤال جستجو با داده‌های موجود در پیکره مطابقت داده می‌شود و مواردی (tokens) از داده‌ها که منطبق با سؤال جستجو است به‌عنوان پاسخ جستجو در اختیار کاربر قرار می‌گیرد.

به‌عنوان مثال، در دستورزبان فارسی حروف اضافهٔ بسیط (با، به، از، بر، در) با کلمات مکان‌نما ترکیب می‌شوند. ترکیب حاصل دلالت بر صراحت و وضوح بیشتر یک مکان مشخص دارد (ابوالحسنی چیمه ۱۳۸۱، فصل سوم). کلمات مکان‌نما عبارت‌اند از: «رویِ، زیرِ، داخلِ، بیرونِ، کنارِ، پهلویِ، بغلِ، سرِ، بالایِ، پایینِ، لبِ، پایِ، پشتِ، دمِ، جلویِ، نزدیکِ، پسِ، پیشِ، میانِ، بینِ، عقبِ».

جملات زیر را در نظر بگیریم:

(۱) بچه‌ها در کنار دریا شن‌بازی می‌کنند (نه در نزدیک دریا یا در روی دریا یا در ...).

(۲) علی از زیر میز کتاب را آورد (نه از کنار میز یا از روی میز یا از ...).

در جملهٔ اول می‌توان حرف اضافهٔ اول (در) را حذف کرد و جمله همچنان خوش‌ساخت باشد: بچه‌ها کنارِ/نزدیکِ دریا شن‌بازی می‌کنند. ولی در جملهٔ دوم با حذف حرف اضافهٔ اول (از) جمله بدساخت می‌شود:

* علی کنار/روی/زیر میز کتاب آورد.

در اینجا دو سؤال مطرح می‌شود:

سؤال اول: ترکیب حروف اضافهٔ بسیط با کلمات مکان‌نما یک حرف اضافهٔ مرکب می‌سازد یا یک گروه حرف اضافه‌ای که در آن کلمهٔ مکان‌نما متمم حرف اضافه است؟

سؤال دوم: ترکیب هر حرف اضافهٔ بسیط با کلمات مکان‌نما چه تعبیر معنایی دارد؟

برای پاسخ به این دو سؤال می‌توان از پیکرهٔ زبانی با حجم بسیار بالا استفاده کرد. برای این منظور کافی است با استفاده از زبان جستجوی پیکره یک سؤال متناسب با الگوهای نحوی موردنظر طرح کرد:

([POS = "prep"])[WORD = "رویِ.*"|"بالایِ.*"|"بیرونِ.*"]within S.

همان‌طور که ملاحظه می‌شود سؤال جستجو سه عبارت صوری دارد که در مجموع متناظر با الگوی نحوی ساخت مورد مطالعه است. چون بر طبق جملات ۱ و ۲ حرف اضافهٔ اول قابل

حذف است، لذا عبارت اول اختیاری است و در داخل پرانتز قرار داده شده است. عبارت دوم بیانگر یکی از کلمات «رویِ»، «بالایِ» و «بیرونِ» است که بعد از آنها می‌تواند رشته‌ای از کلمات وجود داشته باشد. این مفهوم با استفاده از نویسهٔ* بعد از هرکدام از کلمات مشخص شده است. واضح است که برای یک جستجوی کاملتر باید تمامی کلمات مکان‌نما در این عبارت آورده شوند. عبارت سوم (within S) یک عبارت مرزی است که تعداد کلمات را در عبارت دوم تا مرز جمله محدود می‌کند. بنابراین رایانه عمل انطباق الگوی موجود در سؤال جستجو را با داده‌های پیکره انجام می‌دهد تا کلیهٔ جملات یا پاره جملات را که این عمل برای آنها موفق می‌شود، به‌عنوان پاسخ جستجو بر طبق فرمان کاربر نمایش دهد یا چاپ کند.

نتیجه

آنچه تحت عنوان پیکره‌ها و دادگانهای زبانی یا گفتاری مطرح است، ریشه در رویکرد تاریخی و ساختاری به زبان دارد، که جوهرهٔ آن محوریت داده‌های بالفعل زبانی اعم از متن و گفتمان در استخراج دستورزبان است. با استفاده از این روش نه‌تنها می‌توان ساخت احتمالاتی نظامهای زبانی را مطالعه کرد، بلکه یافته‌های زبان‌شناسی نظری را در حوزهٔ دستورزبان در قالب فرضیه‌های زبانی محک زد. برای این منظور یک نرم‌افزار رایانه‌ای برای مطالعهٔ دستورزبان فارسی معرفی شد.

سپاسگزاری

از معاونت پژوهشی دانشگاه تهران و پژوهشکدهٔ پردازش هوشمند علائم که نگارنده را در اجرای طرح پژوهشی «امکان‌سنجی برای مدل‌سازی زبان فارسی» یاری کرده‌اند، قدردانی و تشکر می‌شود.

کتابنامه

Assi, S. M. and Haji Abdolhosseini, 2000. "Grammatical Tagging of a Persian Corpus". *Int. Journal of Corpus Linguistics.* Vol. 5, No. 1, pp. 69-82.

Barnbrook, G., 1996. *Language and Computers, A practical Introduction to the Computer Analysis of Language.* Edinburgh Textbooks in Empirical Linguistics, Edinburgh University Press, U. K.

Bensch, Peter, A., 1991. "Neo-Structuralism: A Commentary on the Correlations between the Work of Zelig Harris and Jeffrey Elman". Center for Research in Language Newsletter, UCSD, March, Vol. 5, No. 2.

Benson, M., et.al., 1986. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*, John Benjamins Publishing Co. Philadelphia, U.S.A.

Chomsky, N., 1959. "A Review of B. F. Skinners' *Verbal Behavior*", *Language.* 35, No: 1, 26-58.

Crystal, David, 1994. *An Encyclopedic Dictionary of Language and Languages.* Oxford, p. 85.

Harris, Zelig, 1951. *Methods in Structural Linguistics.* Chicago Univ.: University of Chicago Press.

Kenstowicz, M., 1955. *Phonology in Generative Grammar.* Blackwell Publishers, Cambridge, Mass., USA.

Leech, G., et. al., 1994. "CLAWS4: the Tagging of a British National Corpus", *Proceedings of 15th Int. Conf. on Computational Linguistics (COLING 94).* Kyoto, Japan, p. 622-628.

McArthur, Tom, 1994. *The Oxford Companion to the English Language.* Oxford, pp. 265-266.

Meurers, W. Detmar, 2002. *On the Use of Electronic Corpora for Theoretical Linguistics, Case Studies from the Syntax of German.* Dept. of Linguistics, The Ohio State Univ., USA.

Shannon, C. E., 1948. "Mathematical Theory of Communication", Bell System Technical Journal, July and October.

Stergos Afantenos D., 2001. *On Grammars, the Chomsky Hierarchy and Probabilistic Grammars.* N. C. S. R. Institute of Informatics and Telecommunications. Software and Knowledge Engineering Laboratory.

Svartvik, Jan, 1995. *The London Corpus of Spoken English: Description and Research.* Lund Studies in English 82: Lund University Press.

http://www.ling.lancs.as.uk/monkey/ihe/linguistics/corpus1/1chom.htm

ابوالحسنی چیمه، زهرا، ۱۳۸۱. بازشناسی حروف اضافهٔ مرکب از گروههای حرف اضافه‌ای در زبان فارسی معاصر بر اساس نحو ایکس تیره. پایان‌نامهٔ دکتری دانشگاه تهران، دانشکدهٔ ادبیات و علوم انسانی، گروه زبان‌شناسی همگانی.

بلومفیلد، لئونارد، ۱۹۳۳. زبان. ترجمهٔ علی محمد حق‌شناس، تهران، مرکز نشر دانشگاهی، ۱۳۷۹.

بی‌جن‌خان، محمود، ۱۳۷۱. «گزارش علمی شرکت در کنفرانس یوروسپیچ ۹۳»، مجلهٔ زبانشناسی، س ۹، ش ۲، پاییز و زمستان ۱۳۷۱.

چامسکی، نوام، ۱۹۵۷. ساختهای نحوی. ترجمهٔ احمد سمیعی، تهران، انتشارات خوارزمی، ۱۳۶۲.

صفوی، کوروش، ۱۳۸۰. «نگاهی تازه به مسئلهٔ چندمعنایی واژگانی»، نامهٔ فرهنگستان، دورهٔ پنجم، ش ۲، ص ۶۷ـ۵۰.

عاصی، مصطفی، ۱۳۷۹. «رایانه و استانداردسازی در زبان»، مجموعهٔ مقالات چهارمین کنفرانس زبان‌شناسی نظری و کاربردی. تهران، انتشارات دانشگاه علامه طباطبایی، ج ۱، ص ۶۰ـ۴۷.

فرشیدورد، خسرو، ۱۳۵۱. «کلمهٔ مرکب و معیار تشخیص آن در زبان فارسی»، مجموعهٔ سخنرانیهای دومین کنگرهٔ تحقیقات ایرانی. مشهد، دانشگاه مشهد، ج ۱، ص ۲۱۷ـ۱۶۹.

لده‌فوگد، پیتر، ۱۹۹۲. «میزان دانش کافی برای تجزیه و تحلیل زبانهای گفتاری»، ترجمهٔ محمود بی‌جن‌خان، مجلهٔ زبانشناسی، س ۹، ش ۲، پاییز و زمستان ۱۳۷۱، ص ۹۵ـ۸۴.