هنری کوچرا

مترجم: بهروز صفرزاده

ریاضیات زبان

زبان چنان با انسان عجین شده و وسیله‌ای بالقوه بی‌همتا برای بیان تخیلات شاعرانه گشته که بسیاری از ما به دشواری می‌توانیم آن را دارای جنبه‌های غیرتخیلی، کلیشه‌ای و قابل‌پیش‌بینی، یا حتی ریاضی بدانیم. با این حال، در طی چند دههٔ گذشته شاهد بوده‌ایم که حجم وسیعی از گفتارها و نوشتارهایمان به وسیلهٔ کامپیوترها ضبط، پردازش، و منتقل شده است. کامپیوترها قادرند واژه‌های مکتوب ما را به نشانه‌های الکترونیکی تبدیل کنند، و بدین‌ترتیب این امکان را برای ما فراهم سازند که در میان واژه‌ها به جست‌وجو بپردازیم و آنها را سازمان‌دهی کنیم، و آن‌گاه، در صورت نیاز، آنها را در قالب نشانه‌های الکترونیکی بی‌درنگ به فواصل دور بفرستیم. این کار کامپیوتر ــ که با کم‌ذوقی آن را «واژه‌پردازی» و «بازیابی اطلاعات» نامیده‌ایم ــ نه تنها صنعتی چند میلیارد دلاری، بلکه ابزار اطلاع‌رسانی نوینی است که همانند اختراع چاپ در حدود پانصد سال قبل، اطلاع‌رسانی را از پایه دگرگون می‌سازد.

هنوز هم مردم بخشی از وقتِ خود را به تدوینِ اندیشه‌ها از نوشتن رمان، و سرودن شعر اختصاص می‌دهند. اما دیگر مجبور نیستیم که برای اعمالِ چند تصحیح در متن با زحمت بسیار تمام صفحات آن را از نو حروف‌چینی کنیم. امروزه می‌توانیم مترادفات و تعاریف را در واژه‌نامه‌های الکترونیکی جست‌وجو کنیم، یا حتی از کامپیوتر بخواهیم که غلط‌های املایی ما را بیابد و تصحیح کند، و یا دستورِ زبان و سبکِ نگارشمان را محک بزند.

کامپیوترها ــ که هنوز هم برخی از ما آنها را صرفاً ماشین‌حساب‌هایی با سرعتِ برق‌آسا می‌دانیم ــ می‌توانند زبان را رمزگذاری، پردازش، و در سطحی محدود درک کنند. این امر مدیونِ ویژگی‌های چشم‌گیری است که دستگاه‌های زبانِ طبیعی براثر تحول و تکامل در طول تاریخ طولانی خود یافته‌اند. زبان ماهیتاً ساختاری سازمان‌یافته از نشانه‌های دارایِ سلسله مراتب است، که ما را قادر می‌سازد تا تعدادِ بالقوه نامحدودی مفهوم را با استفاده از ابزاری محدود بیان کنیم. اجزای اصلی این ساختار بسیار معدودند: مجموعه‌ای از آواهای منقابل ــ واج‌های زبان ــ که می‌توانیم آنها را به شکل حروف و ترکیباتِ حروف در نظام‌هایِ نوشتاری الفبایی خود نشان دهیم. این اجزای سازندهٔ اولیه به شیوه‌هایی بسیار محدود با یکدیگر ترکیب می‌شوند و واحدهای معنادار اصلی ــ واژه‌های زبان ــ را می‌سازند. واژه‌ها نیز به نوبهٔ خود بر طبق اصول روش‌مند ترکیب‌سازی زنجیره‌های معناداری را می‌سازند که بر پایهٔ قواعدِ نحو قرار دارند و ما آنها را «جملات» می‌نامیم. بدین‌ترتیب کل این نظام عبارت است از ساختن مجموعه‌ای از گفتارهای بالقوه نامحدود با استفاده از فهرست محدودی از واحدهای ناپیوسته، درست همان‌طور که از ترکیب ده عددِ نظامِ اعدادمان با یکدیگر می‌توان مجموعه‌ای نامحدود از ارزش‌ها و عبارات ریاضی متفاوت ساخت.

این سازمان‌دهی همانند نظام زبان و اعداد به‌راحتی امکان این را فراهم می‌سازد که واحدهای زبانی را به صورت اعداد نمایش دهیم و به گونه‌ای آنها را به کار ببریم که گویی واحدهای ریاضی هستند. جادو نیست که کامپیوتر واژه‌ای را که از نظر املایی غلط است برایمان پیدا کند: اگر کامپیوتر واژه‌نامه‌ای در اختیار داشته باشد که در آن برای هر واژه یک کُدِ عددی تعیین شده باشد، می‌تواند کد عددی واژه‌ای را که می‌نویسیم با ارزش عددی واژه در فهرست واژه‌هایش از طریق یک عملیات حسابی ساده مقابله و مقایسه کند. چنان‌چه مابازایی برای آن یافته نشود، واژهٔ موردنظر در واژه‌نامهٔ کامپیوتر وجود ندارد و بدین‌ترتیب می‌فهمیم که املای واژه غلط است.

کارآیی و حشو

زبان‌ها ــ از آنجا که در طی قرون و اعصار به طور خودجوش در جوامع رشد و تکامل یافته‌اند ــ دارای ویژگی‌های ریاضی‌ای به مراتب پیچیده‌تر از ویژگی‌های ریاضی ناشی از ساختار سلسله مراتبی واحدهای ناپیوسته هستند. زبان‌ها در آن واحد هم کارآیی دارند و هم حشو. این دو خصیصهٔ متناقض به

گونه‌ای با یکدیگر توازن یافته‌اند که ارتباطات را سودمند و قابل اعتماد می‌سازند.

نخست حشو را در نظر بگیریم: از میان حدود ۳۳ واج زبان انگلیسی، تنها زیرمجموعهٔ کوچکی از جای‌گشت‌های بالقوهٔ آنها در ساخت واژه‌های بالفعل زبان به کار می‌رود. مثلاً انگلیسی‌زبانانِ بزرگ‌سال می‌دانند که trip واژه‌ای انگلیسی است. اما آنها این را هم می‌دانند که tlip واژه‌ای انگلیسی نیست، و لزومی هم ندارد که برای اطمینان از این مسئله به واژه‌نامه رجوع کنند. شمّ زبانی به آنها می‌گوید که هیچ واژهٔ انگلیسی‌ای نمی‌تواند با tl آغاز شود. اما همین انگلیسی‌زبانان اگر با trin روبه‌رو شوند ــ گرچه این واژه را نمی‌شناسند ــ ممکن است ناچار شوند از واژه‌نامه یاری بجویند. لفظِ trin دست‌کم از لحاظ نظری بالقوه می‌تواند واژه‌ای انگلیسی باشد، چرا که هیچ‌یک از محدوده‌هایِ عامی را که در زنجیره‌های آوایی مجاز زبان انگلیسی هست، نقض نمی‌کند.

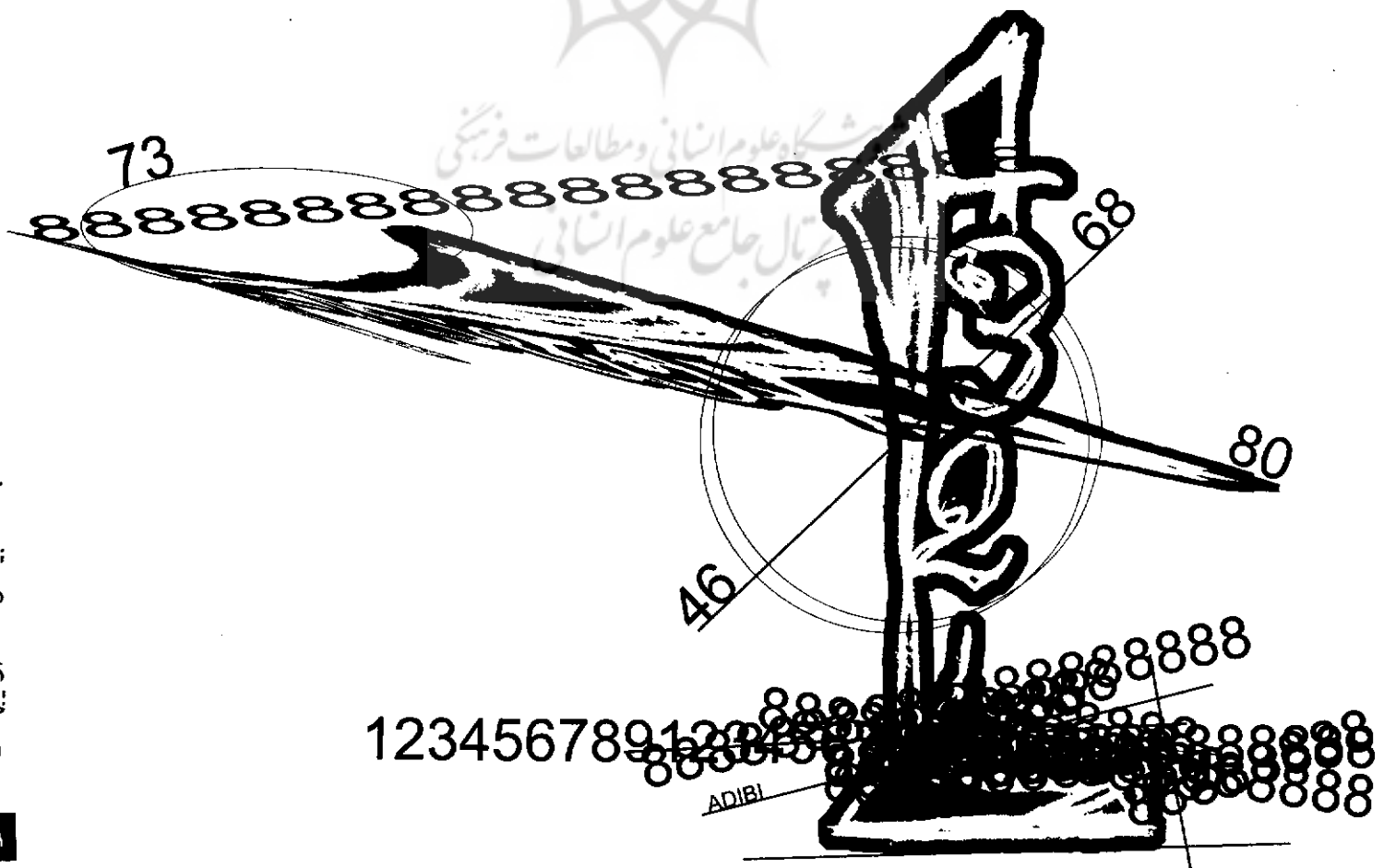
می‌بینیم که حتی در همین سطح ابتدایی واج‌شناسی نیز حشوِ قابل توجهی وجود دارد: تحمیلِ محدودیت‌هایی به زنجیره‌های بالقوه، و در نتیجه ورود اطلاعات زائد به نظام زبان.

وجود این اطلاعات زائد لازم است، تا ما را قادر سازد که بدون انبوهی از اشتباهات و سوءتفاهمات، با دیگران ارتباط برقرار کنیم. اگر هر جای‌گشت بالقوهٔ آواها در انگلیسی واژه‌ای بالفعل می‌بود، نظام ارتباطی ما حقیقتاً بسیار کارآمد می‌شد، و تمام واژه‌های ما بسیار کوتاه می‌بودند. در آن صورت هیچ نیازی به واژه‌های بیش از چهارآوایی نداشتیم، چرا که از لحاظ ریاضی بیش از یک میلیون واژهٔ چهارآوایی می‌توانستیم داشته باشیم. اما ارتباط برقرارکردن در چنین نظامی فوق‌العاده دشوار است. محدودیت‌های فیزیولوژیکی ما در زمینهٔ تولید و درک آواها و زنجیره‌های آوایی، و هم‌چنین ویژگی‌های حافظهٔ انسان، آموزش و به کارگیری چنین نظامی را عملاً غیرممکن می‌سازد. حتی اگر به فرض بتوانیم چنین زبان کارآمدی را بیاموزیم، هر صدایی، هر عیب و نقص یا اشتباهی که به دریافت ولو تنها یک صدا لطمه بزند، درک ما را مخدوش خواهد ساخت. علت این امر آن است که فقدان حشو در نظام به ما اجازه نخواهد داد که حدس بزنیم چه چیزی را احتمالاً از دست داده‌ایم. از این هم بدتر آن که اگر صدایی را به جای صدای دیگری بشنویم، واژه‌ای معقول ــ اگرچه برخلاف مراد گوینده ــ را شنیده‌ایم. بنابراین حشو، که ویژگی تمام زبان‌های جهان است، یکی از ارزش‌مندترین دوستان ارتباطی ماست.

بدیهی است که محدودیت‌های زنجیره‌های آوایی مجاز که برای دست‌یابی به این حشو به کار می‌روند، ممکن است در زبان‌های گوناگون تفاوت‌های چشم‌گیری داشته باشند. مثلاً زبان‌های بسیاری هست که در آنها واژه‌ای می‌تواند با خوشهٔ آوایی ــ tl آغاز شود، همان خوشهٔ آوایی‌ای که در انگلیسی جایی ندارد. در سطوح بالاتر نیز این حکم صادق است. زبانِ ترکیبیِ انگلیسی برای نشان دادن بسیاری از کارکردهای دستوری خود متکی به ترتیب واژه‌هاست. در این زبان زنجیره‌های بالقوهٔ واژه‌ها در جمله به‌شدت محدود است. اما در زبان‌هایی مانند لاتین یا روسی، که ترتیب واژه‌های آنها «آزاد» است، ظاهراً جای‌گشت‌های واژه‌ها نامحدود است. آن‌چه نامحدود بودن جای‌گشت‌ها را در این زبان‌ها ممکن می‌سازد، دستگاه و صورت‌های صرفی دقیق آنهاست که در آن تعداد اندکی پسوند به چند شیوهٔ بسیار محدود با هستهٔ واژه‌ها ترکیب می‌شوند تا روابط نحوی‌ای را که در انگلیسی از طریق ترتیب واژه‌ها بیان می‌شود، نشان دهند.

شاخه‌ای از ریاضیات به نام «نظریهٔ اطلاعات» ابزاری صوری برای سنجش حشو در یک نظام ارتباطی به دست می‌دهد. این سنجش‌ها برای زبان‌های طبیعی پیچیده و دشوارند؛ با این حال، برآوردهایی اجمالی را با آنها می‌توان انجام داد. در سطح واج‌شناسی برای چند زبان (انگلیسی، آلمانی، روسی و چکی) محاسبه کرده‌ایم که میزان حشو به حدود ۵۰٪ می‌رسد، البته در این محاسبه تنها محدودیت‌های زنجیره‌های آوایی درون هجاها را درنظر گرفته‌ایم. بدیهی است که ترتیب قرار گرفتن هجاها نیز در تمام زبان‌ها محدودیت‌هایی دارد. بدین ترتیب در مجموع میزان حشو را باید دست‌کم حدود ۸۰٪ برآورد کرد.

روی دیگر سکه در زبان «کارآیی» است. از دیرباز



دانسته شده است که واژه‌های بسیار پربسامد معمولاً کوتاه هستند. وقتی واژه‌ها رایج‌تر می‌شوند، ما آنها را کوتاه یا مخفف می‌کنیم: telephone را به صورتِ phone, airplane و مانندِ آنها.

تجزیه و تحلیلِ کامپیوتری نمونه‌های فراوانی از متون زبان، اطلاعات دقیقی در تأیید این نتیجه‌گیری کلی به دست می‌دهد. پیکره‌ای یک‌میلیون‌واژه‌ای از انگلیسی آمریکایی امروز تهیه شده که به نام «پیکرهٔ براون» نیز معروف است. این پیکره با استفاده از نمونه‌هایی برگرفته از ۵۰۰ منبع گوناگون از پانزده‌گونه و سبک نوشتاری مختلف تألیف شده است. در این پیکره واژه‌هایی که ۵۷٪ متن‌های مورد مطالعه (به عبارت دیگر، ۵۷٪ از یک میلیون نشانه) را تشکیل می‌دهند، دارای چهار حرف یا کمترند. چند نمونه از این واژه‌ها عبارت‌اند از: have, that, of, but, and, the.

اما چنان‌چه واژه‌نامه‌ای از این پیکره تهیه کنیم، مسئله شکلی کاملاً متفاوت خواهد داشت. یعنی مجموعه‌ای از واژه‌های گوناگون ــ که در زبان‌شناسی صوری «گونه‌ها» نامیده می‌شوند ــ فراهم کنیم، به‌طوری که هر واژه فقط یک بار در فهرست بیاید، بدون در نظر گرفتن دفعاتی که آن واژه ممکن است در متن تکرار شده باشد. در این صورت، واژه‌هایِ دارایِ چهار حرف یا کمتر نزدیک به ۹٪ واژه‌نامه را تشکیل می‌دهند.

اختلاف میان دو رقم مذکور نشان‌دهندهٔ کارآیی ارتباطی زبان است: نظامِ زبان به گونه‌ای طراحی شده که واژه‌های کوتاه بارها در یک متن معمولی تکرار می‌شوند و بدین ترتیب بسامد بالایی را به خود اختصاص می‌دهند. واژه‌های بلندتر کم به کار می‌روند، و بسامد واژه‌های واقعاً بلند نیز ناچیز است. درازای هر بار کاربرد یک واژهٔ ده‌حرفی هشت بار کاربرد یک واژهٔ سه‌حرفی، و درازای هر بار کاربرد یک واژهٔ بیست‌حرفی ۳۵۲۴ بار کاربرد یک واژهٔ سه‌حرفی وجود دارد.

این اصل در زبان‌های انسان مشابه طراحی نظام‌های ارتباطی مصنوعی است. در الفبای بین‌المللی مرس، پراستعمال‌ترین حرف انگلیسی، یعنی e، کوتاه‌ترین نشانه را دارد ــ یک سیگنال کوتاه که مستلزم کمترین زمان انتقال است. برعکس، کم‌استعمال‌ترین حروف، مانند j، q و y، بلندترین نشانه‌ها را دارند ــ زنجیره‌های گوناگونی از سه سیگنال بلند و یک سیگنال کوتاه. ساموئل مرس چیزی را طراحی کرد که زبان‌ها در جریان تحول و تکامل طبیعی خود به آن دست یافته‌اند.

تعداد واژه‌ها

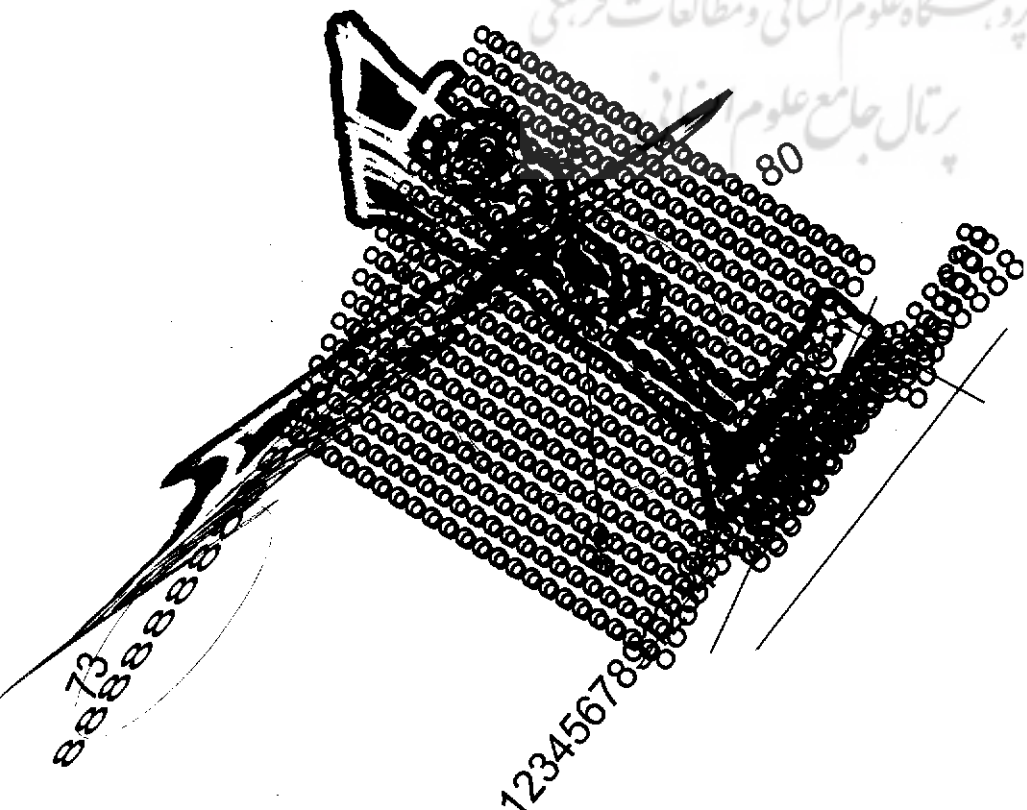
هر زبان‌شناسی که به جنبه‌های کمّی زبان علاقه‌مند باشد، گه‌گاه به نوعی این پرسش‌های نهایتاً عددی به ذهنش خطور می‌کند که: «شکسپیر چند واژه به کار برده است؟»، «هر شخص چند واژه بلد است؟»، «واژه‌نامه چند واژه باید داشته باشد؟».

دست‌کم پرسش نخست پاسخی مشخص ــ هر چند نه ساده ــ دارد: کلیات شکسپیر جمعاً ۸۸۴۶۴۷ واژهٔ متن دارد، که از ۲۹۰۶۶ واژهٔ متمایز ــ ازجمله نام‌های خاص ــ تشکیل یافته است. برای درک معنای این ارقام، باید کاملاً مطمئن باشیم که منظورمان از واژه چیست. چه چیزی را شمارش می‌کنیم؟ حتی اگر تنها بر روی زبان نوشتار تمرکز کنیم و واژه را صرفاً زنجیره‌ای از حروف بدانیم که از دو سو با فاصله محصور شده است، مشکلمان کاملاً برطرف نخواهد شد. باید تعیین کنیم که صورت‌های صرفی ــ مثلاً صورت‌های صرفی فعل play (played, playing, plays) ــ را واژه‌هایی مستقل می‌دانیم، یا صرفاً اعضای یک طبقه که با شکل هستهٔ play نمایش داده می‌شوند. اگر شیوهٔ نخست را برگزینیم، چهار واژهٔ متمایز خواهیم داشت. اگر شیوهٔ دوم را برگزینیم، تنها یک واژه خواهیم داشت، که متشکل است از مجموعه‌ای از صورت‌های دستوری که فقط از لحاظ صرفی با هم تفاوت دارند. زبان‌شناسان چنین واژه‌ای را «لِما» می‌نامند. حتی در انگلیسی، که دستگاهِ صرفیِ کاملاً محدودی دارد، هنگام دادن هر نوع آماری دربارهٔ تعداد واژه‌هایی که نویسنده‌ای به کار برده یا در اثری به کار رفته است، باید بسیار دقیق و سنجیده از اصطلاحات استفاده کنیم. در زبانی بسیار تصریفی چون روسی، که بسیاری از اسم‌ها دارای ده صورت و حالت صرفی و شمارشی متفاوت هستند، اختلاف میان تعداد صورت‌های واژه و تعداد لماها حقیقتاً چشم‌گیر است.

واژه‌نامه‌های متعارف عموماً مجموعه‌هایی هستند از صورت‌های هستهٔ لماها که به شکل مدخل‌هایی با حروف سیاه و به ترتیب الفبا ارائه می‌شوند. صورت‌های صرفی آنها هنگامی ذکر می‌شود که بی‌قاعده یا دارای تغییر املایی باشند؛ اما تصریفات باقاعده عموماً ذکر نمی‌شود. از این رو، تصور ما از واژه، به دلیل آن که تحت تأثیر ساختار واژه‌نامه شکل گرفته است، بیشتر به لما نزدیک است تا صورت‌های صرفی.

با این حال، در بررسی واژگان و بسامد واژگان یک نویسنده یا مجموعه‌ای از متون، معمولاً صورت‌های صرفی را شمارش می‌کنند، نه لماها را. این کار چند دلیل دارد: کامپیوترها زنجیره‌ای از حروف را که عین زنجیره‌ای دیگر نباشد، به عنوان واژهٔ متمایز تشخیص می‌دهند. آنها نمی‌توانند فقرات متعلق به یک لما را به آسانی تشخیص و در کنار یکدیگر قرار دهند. لمابندی مستلزم چند تصمیم‌گیری روش‌شناختی زبانی است، که برخی از آنها بسیار دشوارند. بنابراین رقم ۲۹۰۶۶ واژهٔ متمایز کلیات شکسپیر ناظر بر صورت‌های صرفی است، نه لماها. اگر فرض کنیم که نتایج تحقیقات انجام شده بر روی انگلیسی جدید ــ دست‌کم به طور تقریبی ــ درمورد آثار شکسپیر نیز صادق است، می‌توانیم نتیجه بگیریم که مجموع لماهای آثار این شاعر در حدود ۱۸۰۰۰ است.

با این حساب، غنای واژگان شکسپیر را چگونه می‌توان با انگلیسی امروز مقایسه کرد؟ در تجزیه و تحلیل پیکرهٔ براون، صورت‌های صرفی بر اساس اصول دقیقاً تعریف‌شده‌ای در ضمن گروه‌های لمای خود آورده شده‌اند؛ اما آمار صورت‌های واژهٔ منفرد نیز تهیه شده است. دو دسته نتایج نشان می‌دهد که این پایگاه داده‌های یک‌میلیون‌واژه‌ای دارای ۶۱۸۰۵ صورت واژه است، که به ۳۷۸۵۱ لما تعلق دارند. بنابراین تعداد واژه‌های متمایز در پیکرهٔ براون بیش از دو برابر کلیات شکسپیر است، گرچه این دو پایگاه داده‌ها از نظر حجم با هم قابل مقایسه‌اند. آیا این بدان معناست که واژگان شکسپیر کم بوده است، یا این که در طی چندصد سال اخیر زبان انگلیسی از لحاظ واژگان به مراتب غنی‌تر شده است؟ ضرورتاً چنین نیست. بی‌تردید واژه‌های بسیاری به مخزن واژگانی ما افزوده شده است، اما تقریباً با اطمینان می‌توان گفت که علت اصلی اختلاف مذکور این است که پیکرهٔ نوشته‌های شکسپیر از نظر محتوا و سبک کاملاً منجانس است. حال آن که پیکرهٔ براون تعمداً نامتجانس تهیه شده، تا نمایانگر زبان معاصر باشد. پیکرهٔ براون از میان ۵۰۰ منبع گوناگون



برگزیده شده، از روزنامه‌ها گرفته تا نوشته‌های علمی و ادبیات داستانی.

اگر تمام واژه‌ها به دفعات یکسان به کار رفته باشند، آن‌گاه هر صورت واژه تقریباً ۱۶ بار و هر لما تقریباً ۲۶ بار در متن یک‌میلیون‌واژه‌ای به چشم می‌خورد. در عمل، میزان تکرار واژه‌های منفرد، و بنابراین بسامد آنها، فوق‌العاده ناهمسان است. آمار کلی بسیار قابل توجه است: میزان کاربرد صد واژهٔ نخست پربسامد چنان زیاد است که این صد واژه ۴۷/۴ ٪کل متن را تشکیل می‌دهند. از میان تمام واژه‌هایِ (نشانه‌هایِ)، متن یک‌میلیون‌واژه‌ای، صد لمای نخست پربسامد ۴۹/۶ ٪کل متن را تشکیل می‌دهند. ۲۸۵۴ صورت واژهٔ متمایز، که به ۲۱۲۴ لمای متمایز تعلق دارند، به تنهایی ۸۰٪کل متن یک‌میلیون واژه‌ای را به خود اختصاص داده‌اند.

برای درک ۸۰٪ یک متن نمونهٔ انگلیسی جدید دانستن کمتر از ۳۰۰۰ واژه کفایت می‌کند. اما این حقیقت بدان معنا نیست که این نوع واژگان می‌توانند بقای فرهنگی هیچ‌یک از ما را در جامعه‌ای مدرن تضمین کند. پیش از هر چیز باید دانست که بسیاری از پر بسامدترین واژه‌های زبان انگلیسی نقش‌واژه‌ها هستند: حروف تعریف، حروف اضافه، و صورت‌های افعال معین، مثلاً صورت‌های have, be و یا do. حرف معرفهٔ the با فاصلهٔ بسیار، پربسامدترین واژهٔ زبان انگلیسی است. این واژه ۶۹۹۷۵ بار در متن یک‌میلیون‌واژه‌ای به کار رفته است. گرچه در کل متن، اسم‌ها و افعال مقوله‌های دستوری غالبند، که به ترتیب در حدود ۲۶٪ و ۱۸٪کل نشانه‌های واژه‌ای را تشکیل می‌دهند، اما در پربسامدترین طبقه، از میان مقوله‌های دستوری، کفهٔ نقش‌واژه‌ها کاملاً سنگین‌تر است. لماها که رتبهٔ ۱ تا ۳۲ بسامد را دارند، همگی نقش‌واژه یا ضمیرند. نخستین معناواژه، یعنی فعل say، رتبهٔ ۳۳، و نخستین اسم، یعنی man، رتبهٔ ۴۴ را به خود اختصاص داده است.

نقش‌واژه‌ها، که وجودشان برای نشان دادن نقش دقیق معناواژه‌ها و رابطهٔ نحوی آنها در جمله ضروری است، دقیقاً همان واژه‌هایی نیز هستند که به دلیل بسیار کلیشه‌ای و قابل پیش‌بینی بودن، اگر در متنی حذف شوند، بسیار راحت‌تر از واژه‌های دیگر می‌توان حدس زد که حذف شده‌اند. چنین حذف‌هایی دقیقاً مشخصهٔ سبکی تیترهای روزنامه‌هاست.

Actor Found in Critical Condition after Explosion

تیتر فوق فاقد هرگونه حرف تعریف و فعل معینی است. در یک متن کامل، جملهٔ مذکور به صورت زیر خواهد بود:

An actor has been found in a critical condition after an explosion.

بدین ترتیب جمله‌ای دوازده‌واژه‌ای به تیتری هفت‌واژه‌ای تقلیل یافته، بی آن‌که چیزی از اطلاع‌رسانی آن کاسته شود. برعکس، هرچه در متنی واژه‌ای کمتر کلیشه‌ای باشد، وجودش به شگفتی‌آفرینی و نقش اطلاع‌رسانی جمله بیشتر کمک می‌کند. به این اعتبار، هرچه واژه‌ای کم‌بسامدتر باشد، اهمیتش ــ دست‌کم از لحاظ آماری ــ برای درک خبر بیشتر خواهد بود.

اگر درک ۸۰٪ هر متن برایمان کافی باشد تا باسواد قلمداد شویم، می‌توانیم با واژگانی کمتر از ۳۰۰۰ گذران کنیم، و از واژه‌نامه‌ها بی‌نیاز گردیم. اما از آن‌جا که چنین نیست، واژه‌نامه‌ها که املا، تلفظ، و تعاریف واژه‌ها (اعم از پربسامد و نادر) را در اختیارمان می‌گذارند، یاران جدایی‌ناپذیر ما هستند.

کامپیوتر و دستورزبان

در چند سال گذشته کامپیوترهایی ساخته‌ایم که از توانایی‌های زبانی چشم‌گیرتری برخوردارند: اکنون برنامه‌هایی در اختیار داریم که دست‌کم برخی از اشتباهات دستوری را می‌توانند بیابند و تصحیح کنند. در نگاه نخست، این نوع توانایی کامپیوتر ممکن است شگفت‌انگیز بنماید. تسلط یافتن بر ظرائف املای انگلیسی شاید کار دشواری باشد، اما درک طرز کار غلط‌یاب کامپیوتری از آن دشوارتر نیست. واژه‌نامهٔ املایی با فهرست مشخصی از واژه‌ها سروکار دارد؛ اما چنین نیست که دستور زبان هم با فهرست‌های مشخصی از جملات سروکار داشته باشد. سروکار دستور زبان با انواع بالقوه بی‌شماری از عبارات انسان است. جملاتی که به آنها برمی‌خوریم، اغلب برایمان کاملاً تازگی دارند؛ با این حال، می‌توانیم آنها را بفهمیم، هرچند که پیش از این هرگز آنها را ندیده یا نشنیده‌ایم. بی‌گمان هیچ‌کس زبان بومی‌اش را از راه حفظ کردن جملات نمی‌آموزد، بلکه هر کسی دانشی انتزاعی از ساخت‌های پذیرفته‌شدهٔ زبان خود دارد که از آن برای ساختن و درک زنجیره‌های جدید واژه‌ها استفاده می‌کند. از طریق ملکهٔ ذهن کردن این دانش است که فرد زبان بومی‌اش را فرامی‌گیرد.

همهٔ انگلیسی‌زبانان، حتی آنهایی که هرگز دستور زبان نخوانده‌اند، دارای شمِ زبانی اساسی و قابل‌اعتمادی هستند که به آنها می‌گوید کدام جمله می‌تواند انگلیسی باشد و کدام نمی‌تواند، درست همان‌طوری که می‌دانند کدام واژه می‌تواند انگلیسی باشد و کدام نمی‌تواند. اما این شم دستوری به مراتب انتزاعی‌تر و پیچیده‌تر از شمّ واج‌شناختی گوینده است. این مثال ساده را در نظر بگیرید: هر انگلیسی‌زبانی می‌داند که جملهٔ She told him to behave himself درست است، اما جملهٔ She told him to behave him و She told himself to behave himself کاملاً مردودند.

اگر به جزئیات فرآیند ذهنی‌ای که ما را قادر می‌سازد تا این قضاوت‌های به ظاهر ساده را انجام دهیم بیندیشیم، عمق پیچیدگی آنها بی‌درنگ نمایان خواهد شد. می‌دانیم که در مثال نخست، آخرین واژه باید himself باشد، نه him، زیرا شخصی که از او خواسته شده تا مؤدب باشد، در آن واحد هم فاعل زیرساختی فعل behave است و هم مفعول آن. زبان‌شناسان این را «ساختار انعکاسی» می‌نامند. از سوی دیگر، سومین واژهٔ همان مثال، یعنی ضمیر him، مفعول فعل told است که به‌وضوح از فاعل، یعنی she، متمایز است، و از همین رو هرگز نمی‌توان آن را با ضمیر انعکاسی بیان کرد.

برنامه‌ریزی کردن کامپیوتر به طوری که اشتباهات دستوری را در ساختار جمله تشخیص دهد و تصحیح کند، کاری است بس دشوار؛ اما اگر به کمتر از کمال مطلوب راضی باشیم، یقیناً کاری است شدنی. اساس این کار آن است که به کامپیوتر تجزیهٔ جمله را بیاموزیم؛ یعنی این که چگونه جمله را به مقوله‌های دستوری‌اش تجزیه کند و روابط نحوی آنها را آشکار سازد، اجزای مهم جمله (از قبیل گروه‌های منفرد، فاعل‌ها، و مفعول‌ها) را بیابد، تعیین کند که آیا فاعل و فعل با هم مطابقت می‌کنند یا نه، و مانند آنها.

آنچه باید به خاطر داشت، این است که تجزیه و تحلیل دستوری تنها از آن رو میسر است که ما می‌توانیم قواعد ساخت‌های زبانی را تدوین کنیم. کامپیوترها فقط در صورتی می‌توانند کارهای شگفت‌انگیز انجام دهند که ما به شیوه‌های صوری دقیق و بسیار روشن طرزکار را به آنها بیاموزیم. توصیف قواعد انتزاعی دستور زبان ــ ویژگی‌های ریاضی زبان در عام‌ترین معنای اصطلاح ــ این کار را ممکن می‌سازد. بدین ترتیب می‌توانیم کاملاً مطمئن باشیم که چه کسی رابطهٔ متقابل انسان ــ کامپیوتر را کنترل می‌کند: عملکرد نهایی کامپیوترها هرقدر هم که عظیم و تحسین‌برانگیز باشد، هنوز هم برعهدهٔ انسان است که قواعد زبانی را کشف کند و به کامپیوتر طرز استفاده از آنها را بیاموزد.

منبع

Kucera, Henry. "The Mathematics of Language". in *The American Heritage Dictionary of the English Language*, Third Edition, Boston, USA: Houghton Mifflin Company, 1992, pp. xxxi - xxxiii.

اصطلاحات

انگلیسی جدید (Modern English) زبان انگلیسی از حدود سال ۱۵۰۰ میلادی به بعد.

پیکره (corpus) مجموعهٔ بزرگی از نوشته‌های دارای سبک، موضوع، یا مؤلف مشخص.

جای‌گشت (permutation) هر یک از آرایش‌های اعضای یک مجموعه. مثلاً سه حرف «د»، «ر» و «س» حداکثر دارای شش جای‌گشت هستند: درس، دس‌ر، ردس، رس‌د، س‌در، س‌رد.

لما (lemma) مجموعه‌ای از صورت‌های دستوری دارای هستهٔ مشترک. مثلاً صورت‌های «رفتم»، «رفته بودم»، «می‌روم»، «بروم»، «برو» و «رفتن» همگی یک لما را تشکیل می‌دهند؛ و یا صورت‌های «دست»، «دستی» (یک دست)، «دست‌ها»، و «دستان».

لمابندی (lemmatization) دسته‌بندی واژه‌ها بر اساس لماها.

معناواژه (content word) واژهٔ معنادار که در تقابل با نقش‌واژه قرار می‌گیرد.

نقش‌واژه (function word) واژهٔ فاقد معنا یا دارای بار معنایی ضعیف که برای نشان دادن روابط دستوری واژه‌های دیگر به کار می‌رود. برخی از نقش‌واژه‌ها عبارت‌اند از: حروف اضافه، حروف ربط، حروف تعریف.

واژگان (vocabulary) ۱. مجموعهٔ واژه‌های یک زبان. ۲. مجموعهٔ واژه‌هایی که فرد می‌داند یا به کار می‌برد.