نرم‌افزار تشخیص فعل در زبان فارسی*

پویان دانش کار آراسته
کارشناس‌ارشد زبان‌شناسی

درآمد

زبان‌شناسی رایانه‌ای[1] از پویاترین شاخه‌های علم زبان‌شناسی در جهان است که ناشناخته‌تر از دیگر حوزه‌های این علم در ایران است. این مقاله، گزارشی است از تلاشی عملی که در این زمینه به‌انجام رسیده است. در این مقاله، روشی در طراحی نرم‌افزاری ارائه می‌شود که می‌تواند با کمک مشخصات صوری **فعل** در زبان فارسی، این مقوله را در متن تشخیص دهد، تقطیع تکواژی کند، و مشخصات صرفی و نحوی آن از قبیل زمان، شخص، عدد، مثبت و منفی، معلوم و مجهول، سببی، وجهی، امر، و مرکب را اعلام کند.

---

* این مقاله براساس رسالهٔ کارشناسی‌ارشد نگارنده ــ با همین عنوان ــ به راهنمایی و مشاورهٔ استادان ارجمند جناب‌آقای دکتر سید مصطفی عاصی و جناب‌آقای دکتر محمد دبیرمقدم که در بهمن ۱۳۸۱ در دانشگاه علامه طباطبایی دفاع شد، به نگارش درآمده است. علاقه‌مندان به مشاهده و تحقیق دربارهٔ کُد برنامه که به زبان Visual Basic نوشته شده است، می‌توانند به این رساله مراجعه فرمایند. در ضمن، از آقای ایمان شاکریان که در کُدنویسی این برنامه مرا بسیار یاری کردند، صمیمانه تشکر می‌کنم.

1. Computational Linguistics

فرهنگ، ۵۰-۴۹؛ بهار و تابستان ۱۳۸۳؛ ص‌ص ۴۶-۳۱

در این مقاله فقط جنبهٔ زبان‌شناختی مسئله بررسی می‌شود. جنبه‌های تخصصیِ رایانه‌ایِ کار، مستلزم مجالی دیگر و مقاله‌ای مستقل است. این طرح می‌تواند برای نرم‌افزارهای مشابه، به‌منزلهٔ یک تجربهٔ انجام‌شده، راهگشا باشد.

۱. مقدمه

بی‌شک یکی از اهداف زبان‌شناسی رایانه‌ای که از ابتدای تکوین و گسترش این علم مد نظر بوده و پیوسته دنبال شده، ارائهٔ نرم‌افزارهایی است که به کمک آنها بتوان تحلیل‌هایی برروی متن و یا پاره‌های زبانی انجام داد. تولید این‌گونه نرم‌افزارها در بلندپروازانه‌ترین شکل خود، برنامه‌های جامع ترجمهٔ ماشینی و برنامه‌های تحلیل متن است.

روشن است که یک برنامهٔ جامع، مثل ترجمهٔ ماشینی یک زبان به زبان دیگر، برنامه‌ای فوق‌العاده گسترده است که خود از اجتماع چندین برنامهٔ کوچک‌تر که هر یک وظیفه‌ای خاص برعهده دارد، تشکیل می‌شود. از عمده‌ترین اجزای چنین برنامه‌هایی، نرم‌افزارهایی است که بتواند مقولات زبانی را تشخیص و مشخصات آنها را به‌طور دقیق ارائه دهد تا برنامهٔ اصلی، بنابه هدفی که دارد، آنها را تحلیل و یا ترجمه کند.

یکی از اصلی‌ترین و درعین‌حال پیچیده‌ترین مقولات زبان فارسی، **فعل** است؛ که در مقالهٔ حاضر دربارهٔ آن بحث می‌شود.

قطعاً اجرای چنین تحقیقاتی، علاوه برطبیعت کاربردی، به دستاوردهای نظری و احیاناً شبیه‌سازی و درک فرایندهای زبانیِ مقولات مورد نظر نیز می‌انجامد.

در این مقاله ابتدا به صورتی بسیار گذرا، از مقولهٔ فعل و آنچه این نرم‌افزار درمورد آن باید تشخیص دهد سخن به‌میان می‌آید و بحثی کوتاه دربارهٔ افعال مرکب ارائه می‌شود؛ سپس ساختار کلی برنامه معرفی و تک‌تک مراحل آن بررسی می‌شود؛ در پایان نیز ویژگی‌ها و مشکلات کار به اجمال مطرح و نتیجه‌گیری می‌شود.

۲. فعل در زبان فارسی

مقولهٔ فعل در زبان فارسی بیش از هر مقولهٔ دیگر، مشخصات صوری دارد: وندهای متعدد، ازقبیل پیشوندها، شناسه، تکواژ سببی‌ساز، و مانند آن. این امر از یک‌سو نوشتن یک برنامهٔ رایانه‌ای را که طبعاً به عوامل صوری و قابل تعریف برای رایانه نیاز دارد، آسان و از سوی دیگر، به‌دلیل تعدد این مشخصات و گستردگی و حالات بسیار زیاد فعل در زبان فارسی، فوق‌العاده پیچیده می‌کند.

نرم‌افزار مورد نظر باید صورت‌های مختلف فعلی در زبان فارسی را که در زیر فهرست‌وار بیان شده است، تشخیص دهد[۱]:

الف- تمامی گروه فعلی زبان فارسی، نه فقط فعل واژگانی؛ برای مثال: **می‌خواهند بروند، گفته نشده بود، داشتیم می‌خوردیم، نپوشانده بودی، فرو افتاده بود، و...**

ب- صورت خاص هر یک از زمان‌های فعل فارسی ازقبیل حال ساده، حال استمراری، ماضی ساده، ماضی بعید، آینده، ...

بنابراین تمامی زمان‌های فعل فارسی باید تجزیه و تحلیل شود و مشخصات صوری هر زمان به‌شکل فرمول‌های قابل تعریف برای رایانه درآید. برای نمونه، فرمول زمان حال ساده و صورت سببی مجهول زمان گذشتهٔ استمراری به ترتیب چنین بیان می‌شود:

می + بن حال + شناسه (م، ی، د، یم، ید، ند)

داشت + شناسه (م، ی، ϕ یم ، ید ، ند) + بن حال + انده/انیده + می + شد + شناسه (م، ی، ϕ یم ، ید ، ند)

یادآور می‌شود که چون تحلیل فعل فارسی در متن مورد نظر است، باید صورت نوشتاری هر کلمه و یا تکواژ درنظرگرفته شود.

پ- صورت مثبت و منفی هر صورت فعلی؛ یعنی تکواژ منفی‌ساز و تمامی جایگاه‌های قرارگرفتن آن باید برای رایانه تعریف شود. برای مثال: **نرفتیم، نخواهد خواند، گفته نخواهد شد، گفته نشد، و...** (چنان‌که ملاحظه می‌شود، تکواژ منفی‌ساز در زبان فارسی جایگاه‌های مختلفی دارد که هر یک باید مشخص شود).

---

۱. برای تحلیل جامع هر یک از مواردی که بیان می‌شود، نگاه کنید به پایان‌نامهٔ نگارنده که مشخصات آن در پاورقی صفحهٔ اوّل مقاله داده شده است.

ت- صورت مجهول افعال؛ که باید به‌صورت یکپارچه تشخیص داده شود، یعنی مثلاً فعل «خورده خواهد شد» به‌طور کلی تشخیص داده شود، نه به‌صورت جدا جدا، مثلاً «خواهد» و «شد».

ث- فعل‌های سببی، که تکواژ سببی‌ساز دارند؛ مانند: **بخورانیم، خورانیده شد**.... همان‌گونه که در قسمت «ب» عنوان شد، بنابر ملاحظات رایانه‌ای، صورت نوشتاری عنصر سببی‌ساز به سه صورت **ان/انده/انیده** تعریف می‌شود.

ج- صورت‌های امر و نهی افعال؛ مثل: **برو، نرو، بسوزان**....

چ- صورت‌های نوشتاری مثبت و منفی در صیغهٔ سوم شخص مفرد زمان‌های زیر در افعالی که بن ماضی آنها در نوشتار، از اضافه کردن یک «د» به بن حال به‌دست می‌آید، مثل فعل «خوردن»، یکسان است، که باید هر دو زمان منظور شود:

حال ساده (/سببی) = ماضی استمراری (/سببی)

حال التزامی (/سببی) = ماضی ساده (/سببی)

در ضمن، ظاهر مصدر کوتاه با صیغهٔ سوم شخص مفرد فعل ماضی ساده یکی است.

ح- افعال وجهی مانند **باید (می‌باید، بایستی، می‌بایستی، می‌بایست)، می‌شود، بشود، می‌شد، می‌توان، بتوان**، همچنین صورت‌های ماضی ساده، ماضی استمراری، ماضی نقلی، ماضی التزامی، و نیز حال ساده از فعل **توانستن** (مشکوةالدینی، ۱۳۷۴: ۱۵۶).

چنان که اشاره شد، هر فعل صورت‌های زمانی مختلفی دارد. فعل را به هشت صورت مثبت، منفی، مجهول، منفی مجهول، سببی، منفی سببی، سببی مجهول، و منفی سببی می‌توان بیان کرد؛ که این صورت‌ها نیز می‌تواند برای شش شخص دستوری متفاوت باشد. در ضمن، صورت‌های امر و نهی هم وجود دارد. همچنین در بعضی از زمان‌ها ممکن است برای هر یک از موارد دو صورت وجود داشته باشد، مثل: **خوانده نباشم** یا **نخوانده باشم**.

برای مثال، صورت‌های زمان حال ساده از مصدر «خوردن» برای صیغهٔ سوم شخص مفرد چنین است:

۱. صورت مثبت: **می‌خورد**

۲. صورت منفی: **نمی‌خورد**

۳. صورت مجهول: **خورده می‌شود**
۴. صورت منفی مجهول: **خورده نمی‌شود**
۵. صورت سببی: **می‌خوراند**
۶. صورت منفی سببی: **نمی‌خوراند**
۷. صورت سببی مجهول: **خورانده/خورانیده می‌شود**
۸. صورت منفی سببی مجهول: **خورانده/خورانیده نمی‌شود**

۲ـ۱. افعال مرکب

دربارهٔ افعال مرکب ـ که به‌دلیل ویژگی خاصی که دارند؛ در اینجا جداگانه دربارهٔ آنها بحث می‌شود ـ بسیاری از زبان‌شناسان بررسی‌های دقیقی انجام داده‌اند؛ از جمله: باطنی (۱۳۴۸)، عاصی (۱۳۷۱)، صادقی (۱۳۷۲)، دبیرمقدم (۱۳۷۴). در این زمینه، اختلافاتی وجود دارد: گروهی همچون خانلری (۱۳۶۵: ۱۸۱-۱۱۳) و مشکوةالدینی (۱۳۷۴: ۱۶۳-۱۵۸) فعل‌های فارسی را به سه دستهٔ ساده، پیشوندی، و مرکب تقسیم کرده‌اند؛ ولی دبیرمقدم (۱۳۷۴) فعل‌های پیشوندی و مرکب ـ هر دو ـ را فعل مرکب می‌شمارد و به تقسیم‌بندی فعل ساده و فعل مرکب قائل است. همچنین، خانلری (۱۳۶۵: ۱۸۱-۱۷۶) عباراتی مثل «خوشم آمد»، «خوابشان می‌آید» و مانند آن را تحت عنوان فعل‌های ناگذر تبیین می‌کند؛ امّا به‌اعتقاد دبیرمقدم (۱۳۷۴)، این عبارات اساساً نه فعل مرکب بلکه جملات تمام و کمالی هستند که در آنها جزء اسمی، فاعل است.

در اینجا یادآوری یک نکته ضروری است. اصولاً در یک برنامهٔ رایانه‌ای، آنچه ارزش دارد، عوامل صوری قابل تعریف برای رایانه است. از این دیدگاه، مسائل نظری تا هنگامی دخیل هستند که به کشف و تبیین این مشخصه‌های صوری بینجامند. بنابراین، صرف‌نظر از اینکه کدام‌یک از نظریه‌ها و تقسیم‌بندی‌های ذکرشده قابل قبول‌تر و از منظر زبان‌شناسی نظری توجیه‌پذیرتر است، در این مقاله به ترتیب زیر عمل خواهد شد:

۱. افعال مرکبی که اصطلاحاً پیشوندی نامیده می‌شوند، مثل «بازآمدن»، «فروافتادن»، و مانند آن، به‌دلیل محدودبودن پیشوندهای فعلی، قابل تعریف و

تشخیص برای رایانه هستند و در طرح تشخیص برنامه به ترتیبی که گفته خواهد شد (بخش ۴-۳-۴)، گنجانده می‌شوند.

۲. در عباراتی که اصطلاحاً ناگذر خوانده می‌شوند، تنها کلمهٔ فعل درنظر گرفته می‌شود؛ مثال: (۱) خوابش خواهد گرفت. (خواهد گرفت: فعل آیندهٔ ساده، مثبت، صیغهٔ سوم شخص مفرد). که البته این کار باعث افزایش درصد خطای برنامه خواهد شد؛ زیرا فعل‌های جملاتی نظیر «خوابشان خواهد گرفت.»، «خوابت خواهد گرفت.»، و غیره نیز همانند فعل جملهٔ (۱) اعلام خواهد شد. چنین مواردی را به‌ناچار باید به پس‌ویرایش[1] سپرد (نگاه کنید به بخش ۵ در همین مقاله).

۳. با اینکه افعالی مثل «بودن»، «کردن»، «زدن»، «آمدن»، «فرمودن»، «افتادن»، و بسیاری دیگر، در بیشتر مواقع به‌صورت مرکب به‌کار می‌روند ــ با توجه به بسامد بسیار بالای افعال مرکب در فارسی ــ مثل: «دلخور بودن»، «صبر کردن»، و مانند آن، موارد بسیاری هم وجود دارد که این افعال به‌صورت مرکب به‌کار نمی‌روند؛ ضمن اینکه بسیاری از آنها با گروه حرف اضافه‌ای که مشخص نیست از چند کلمه تشکیل می‌شود، به کار می‌روند، مانند «به‌دست‌آوردن»، «از کارافتادن». بنابراین، اگر بخواهیم کلمه یا کلمات قبل از این افعال را ــ که به هرحال نمی‌توان برای رایانه قابل تشخیص کرد و یا قابل تشخیص کردن آنها اقتصادی نیست ــ به این افعال منضم کنیم، درصد خطای برنامه بیش از حد بالا خواهد رفت. پس با توجه به اینکه برنامهٔ مورد نظر جزء فعلی این افعال را تشخیص خواهد داد و همهٔ اطلاعات صرفی و نحوی این افعال مانند شخص، عدد، و غیره در جزء فعلی آنها وجود دارد، بهتر است بدون آنکه درصد خطای برنامه را افزایش دهیم، تشخیص جزء غیرفعلی این افعال را به پس‌ویرایش بسپاریم.

۳. پایگاه داده

برای این برنامه، یک پایگاه داده[2] متشکل از بن‌های ماضی و بن‌های مضارع فعل‌های زبان فارسی امروز، به‌صورت الفبایی، تعبیه شده است. در آینده هرگاه

---

1. Post Editing
2. Data Base

کلمهٔ جست‌وجو[1] به کار رود، منظور جست‌وجو در این بانک است. نمونه‌ای از این پایگاه داده که با نرم‌افزار Access نوشته شده است ــ و می‌توان آن را با هر نرم‌افزار پایگاه دادهٔ دیگر هم نوشت ــ به شکل زیر است:

Presentroots	Pastroots	Number
آز	آخت	۱
آرام	آرمید	۲
آزار	آزرد	۳

۴. ساختار برنامه

چون هدف در طرح مورد نظر، تشخیص فعل در متن است، مراحل کلی زیر را می‌توان برای گام‌های مختلف عملکرد نرم‌افزار متصور بود:

۱. گرفتن متن به‌عنوان درونداد
۲. تشخیص اولیهٔ کلمه به‌طور مستقل
۳. بررسی فعل بودن یا نبودن کلمه
۴. اعلام مشخصات درصورت فعل بودن کلمه
۵. تشخیص و بررسی کلمهٔ بعد

۴ـ۱. گرفتن متن به‌عنوان درونداد

در این مرحله، کاربر متنی را که ممکن است شامل یک کلمه، یک عبارت، یک جمله، و یا متنی متشکل از جمله‌های بسیار باشد، در محل مخصوص وارد یا کپی می‌کند. خوشبختانه، زبان‌های برنامه‌نویسی جدید، به‌ویژه برنامه‌هایی که از ابزار ویندوز[2] استفاده می‌کنند و اصطلاحاً Visual نامیده می‌شوند، ابزار مورد نیاز در این قسمت مانند کادر متن[3] را دراختیار برنامه‌نویس قرار می‌دهند.

---

1. Search
2. Windows
3. Text Box

۴ـ۲. تشخیص کلمه به‌صورت مستقل

ابزار تشخیص کلمه در این برنامه، فاصله[1] است. مجموعهٔ حروفی که بین دو فاصله قرار داشته باشد، یک کلمه درنظر گرفته می‌شود. مانند:

علی دیروز به مدرسه رفت.

گاهی در نوشتار فارسی، برخی از تکواژهای واژهٔ فعل به‌صورت جدا از هم نوشته می‌شود، مانند «می‌روم»، «گفته‌اید»، و... مواردی از این دست برای برنامه تعریف شده است و برنامه به‌طور خودکار فاصلهٔ بین این تکواژها را حذف[2] می‌کند و این کلمات را به‌صورت یک کلمهٔ یکپارچه درنظر می‌گیرد.

۴ـ۳. بررسی "فعل" بودن یا نبودن کلمه

این مرحله که درحقیقت بدنهٔ اصلی برنامه را تشکیل می‌دهد، بیشترین حجم کد برنامه را دربر می‌گیرد، و متشکل از بخش‌های[3] متعددی است. پایه و اساس این قسمت بر این است که افعال زبان فارسی یا هیچ وندی ـ پیشوند یا پسوند ـ در ظاهر ندارند یا یک و یا هر دو این وندها را دارند. بنابراین اگر بتوانیم وندهای یک فعل را جدا کنیم، آنچه باقی می‌ماند، بن مضارع یا بن ماضی خواهد بود. در این صورت می‌توانیم با شناختن وندهای به‌کاررفته در ساختمان فعل و ترتیب قرارگرفتن آنها و شناختن بن فعل، به مشخصات صرفی و نحوی فعل پی ببریم و آنها را اعلام کنیم. یک نمای بسیار کلی از این مرحله ـ که حدود ۳۹۰۰ خط از برنامه را دربر می‌گیرد ـ چنین است:

۱. بررسی ابتدایی کل کلمه
۲. جداکردن تک‌تک حروف از ابتدای کلمه و بررسی پیشوندها
۳. جداکردن تک‌تک حروف از انتهای کلمه و بررسی پسوندها
۴. عملیات میانی برای تشخیص فعل‌های چندکلمه‌ای

---

1. Space
2. Trim
3. Modules

۴ـ۳ـ۱. بررسی ابتدایی کل کلمه

در این مرحله، کل کلمه یک بار در بانک جست‌وجو می‌شود؛ زیرا افعالی مثل «خورد» یا «رفت»، هیچ پسوند یا پیشوندی ندارند. اگر کلمه در بانک پیدا شود، قسمت اعلام مشخصات اجرا می‌شود.

۴ـ۳ـ۲. جداکردن تک‌تک حروف از ابتدای کلمه و بررسی پیشوندها

در این مرحله، اولین حرف سمت راست ـ ابتدای ـ کلمه جدا و سپس بررسی می‌شود که آیا این حرف می‌تواند جزء اولین حروف پیشوندهای فعل فارسی باشد یا خیر؛ برای مثال، حرف «م» می‌تواند اولین حرف پیشوند باشد، ولی مثلاً «چ» اولین حرف هیچ پیشوند فعلی در زبان فارسی نیست. به همین ترتیب، انتخاب‌های بعدی نیز محدودتر می شود؛ مثلاً بعد از «م» می‌تواند «ی» بیاید ولی «ل» نمی‌تواند قرار گیرد. به این ترتیب، اگر پیشوندی در کلمهٔ فعل وجود داشته باشد، جدا و سپس ذخیره می‌شود. درضمن، در هر یک از مراحل جداسازی، هرگاه لازم باشد، جست‌وجویی در بانک بن‌ها انجام می‌گیرد.

۴ـ۳ـ۳. جداکردن تک‌تک حروف از انتهای کلمه و بررسی پسوندها

در این مرحله، اولین حرف سمت چپ ـ انتهای ـ کلمه جدا و سپس بررسی می‌شود که آیا این حرف می‌تواند جزء آخرین حروف پسوندهای فعل فارسی باشد یا خیر؛ مثلاً حرف «م» می‌تواند آخرین حرف پسوند (مثل ـ یم) باشد، ولی مثلاً «ر» آخرین حرف هیچ پسوند فعلی در زبان فارسی نیست. به همین ترتیب، انتخاب‌های بعدی محدودتر می‌شود؛ مثلاً قبل از «م» می‌تواند «ی» بیاید ولی «پ» نمی‌تواند قرار گیرد. به این ترتیب، اگر پسوندی در کلمهٔ فعل وجود داشته باشد، جدا و سپس ذخیره می‌شود. در هر یک از مراحل جداسازی، در مواقع لزوم، جست‌وجویی در بانک بن‌ها انجام می‌گیرد.

۴ـ۳ـ۴. عملیات میانی برای تشخیص گروه‌های فعلی

برخی از فعل‌ها در فارسی از بیش از یک کلمه تشکیل می‌شود، مانند «رفته است»، «گفته خواهد شد»، «داشتید می‌رفتید»، «فرو خورده شد»، ...

در این گونه افعال ــ که تعداد زیادی از حالت‌های فعل را دربر می‌گیرد ــ تمامی حالات بررسی و برای رایانه تعریف شده است؛ به این معنی که با استفاده از مشخصات صوری، همهٔ حالات، حروف، فعل‌های کمکی، پیشوندها و پسوندها، در قبل و بعد از افعال مشخص و کدنویسی شده است، در نتیجه همهٔ افعال چندکلمه‌ای مثل موارد بالا و یا افعال پیشوندی تشخیص داده می‌شود. حتی در مواردی که بین کلمات فعل جدایی می‌افتد، مانند «داشتیم به تهران می‌رفتیم»، با تعریف متغیرهای گوناگون که عمر معینی دارد (نگاه کنید به: کد برنامه)، این موارد برای برنامه قابل تشخیص شده است.

نموداری که در پایان مقاله آمده، تنها نمونهٔ بسیار کوچک و ساده‌ای است از موارد مختلفی که در فعل فارسی وجود دارد، آن هم فقط در افعال ساده. بدیهی است که برای هر یک از افعال گروهی، چندین نمودار از این دست لازم است. چنان‌که ملاحظه می‌شود، در این نمودار فقط حالات مختلف افعالی که به «د» ختم می‌شود و پسوندهای مختلف و ساختمان آنها ــ بدون پیشوندها ــ نمایش داده شده است. برای نمونه، اگر اولین مورد را دنبال کنیم، مثلاً برای فعل «خوردن»، خواهیم داشت: **خور** + د که می‌شود **خورد**، و الی آخر.

برای گروه‌های فعلی به‌ازای هر فعلی که در ساختمان این افعال به‌کار می‌رود مانند «شدن» برای مجهول و یا «خواستن» برای فعل‌های آینده و یا اسم مفعول در فعل‌های مجهول یا ماضی نقلی و بعید و...، یک مرحلهٔ جداگانه نوشته شده است. برای مثال، در مورد فعل «داشت خورده می‌شد»، برنامه ابتدا فعل «داشت» را تشخیص می‌دهد و اعلام می‌کند. از آنجا که فعل «داشتن» ازجمله فعل‌های شرکت‌کننده درگروه‌های فعلی است، متغیری فعال می‌شود و آن را ذخیره می‌کند؛ سپس کلمهٔ «خورده» را به‌سبب اینکه فعل نیست، تشخیص نمی‌دهد و از آن می‌گذرد. آنگاه با رسیدن به کلمهٔ «می‌شد»، آن را به‌عنوان فعل تشخیص می‌دهد و اعلام می‌کند. در اینجا برنامه به مرحلهٔ مخصوص فعل «شدن» می‌رود. طبق

تعریف، برنامه حرف آخر کلمهٔ ماقبل را ـ که همیشه در این برنامه ذخیره می‌شود ـ جدا می‌کند. اگر این حرف «ه» نباشد، عملیات پایان می‌پذیرد، متغیرها خاموش می‌شوند، و برنامه به کلمهٔ بعد می‌رود؛ ولی اگر «ه» باشد، بقیهٔ آن کلمه در بانک بن‌های ماضی جست‌وجو می‌شود، اگر پیدا نشود، عملیات مثل مرحلهٔ قبل تمام می‌شود، ولی اگر پیدا شود، آنگاه سه کلمهٔ «داشت»، «خورده»، و «می‌شد» به هم می‌پیوندد و به‌صورت یکجا در برنامه اعلام می‌شود.

۴ـ۴. اعلام مشخصات در صورت "فعل" بودن کلمه

با مشخص‌شدن پیشوندها و پسوندها ـ اگر کلمه فعل باشد و پیشوند یا پسوند هم داشته باشد ـ و همچنین بن ماضی یا مضارع فعل، متغیرهای گوناگونی فعال می‌شود. مثلاً با پیداشدن بن ماضی، متغیری به‌نام Past و یا با پیداشدن پیشوند «می»، متغیری به‌نام Mi فعال می‌شود. در این هنگام، قسمتی از برنامه فعال می‌شود که وظیفهٔ آن اعلام مشخصات فعل است؛ به این ترتیب که ابتدا تمامی پیشوندها، پسوندها، و بن‌های یافت‌شده که ذخیره شده‌اند، به‌ترتیب و با علامت «+» در یک کادر فهرست[1] اعلام می‌شوند، مانند:

«خورد + ه + خواه + د + شد» برای فعل «خورده خواهد شد»؛

سپس برمبنای متغیرهای فعال‌شده، مشخصات فعلِ یافت‌شده درکادر فهرست دیگری اعلام می‌شود. یعنی مثلاً فعال‌شدن Mi باعث چاپ‌شدن استمراری، فعال‌شدن همزمان صفت مفعولی به‌علاوهٔ فعل «شدن» نمایندهٔ مجهول، و فعال‌شدن عنصر سببی‌ساز نشانگر فعل سببی می‌شود، و الی آخر. بنابراین، برای فعل نمونهٔ بالا چنین پیامی چاپ می‌شود:

فعل آیندهٔ سادهٔ مجهول صیغهٔ سوم شخص مفرد از مصدر خوردن

1. List Box

۴ـ۵. تشخیص و بررسی کلمهٔ بعد

در این قسمت، برنامه به مرحلهٔ ۴ـ۱ بازمی‌گردد و عملیات جدید را ازسر می‌گیرد.

۵. پس‌ویرایش

ساختن نرم‌افزارهایی که به‌صورت کامل و بدون خطا متن‌ها یا پاره‌های زبانی را تحلیل و یا ترجمه کند، از دههٔ ۴۰ میلادی همواره آرزوی بشر بوده است. امّا این آرزو حتی در زمینهٔ متن‌های تخصصی با واژه‌ها و ساختارهای محدود نیز هیچ‌گاه محقق نشده است؛ و همواره وجود درصدی از خطا در برنامه‌های رایانه‌ای زبانی، طبیعی تلقی می‌شود. به همین جهت، نرم‌افزارهای طراحی‌شده یا به پیش‌ویرایش[1] و یا به پس‌ویرایش نیاز دارد؛ به این معنی که مواردِ موجب خطا یا غیر قابل پردازش صحیح را کاربر پیش از دادن متن به رایانه به‌طور خاص ـ قابل فهم برای رایانه ـ علامتگذاری یا پس از پردازش ویرایش می‌کند.

در برنامهٔ مورد بحث ما نیز مواردی مانند منضم‌کردن جزء غیرفعلی افعال مرکب، فعل‌های ناگذر و موارد مشکل‌ساز دیگر به پس‌ویرایش سپرده می‌شود.

۶. ویژگی‌های برنامه

الف ـ سرعت بالا: به‌دلیل کوچک‌بودن پایگاه داده‌ها ـ که از فقط حدود ۴۵۰ عضو تشکیل شده است ـ امکان جست‌وجوهای بسیار سریعی فراهم می‌شود که در نهایت به سرعت زیاد برنامه می‌انجامد. طراحی این نرم‌افزار برمبنای داده‌های زبان‌شناختی و ویژگی‌های ساختاری زبان فارسی نیز سبب نوشتن کمترین کد با پوشش‌دادن بیشترین حالات ممکن شده است که خود این امر هم سرعت برنامه را بالا می‌برد.

برای روشن‌شدن مسئله، مصدر سادهٔ «خوردن» را درنظر می‌گیریم. این مصدر در زمان‌های مختلف با احتساب صورت‌های مجهول، سببی، امر، و حالات مثبت و

1. Pre-Editing

منفی در شش صیغه ـ بدون احتساب اینکه فعل‌های «گشتن» و «گردیدن» هم می‌توانند در صورت‌های مجهول به‌کار روند، و البته این امر در برنامه گنجانده شده است ـ بیش از ۵۴۰ حالت مختلف به خود می‌گیرد. حال با توجه به اینکه مصدرهای رایج سادهٔ فارسی بیش از ۴۵۰ عدد است، تنها برای افعال ساده ـ غیرمرکب، غیرپیشوندی،... ـ به پایگاه داده‌ای شامل بیش از ۲۵۰ هزار عضو نیاز است! درصورتی‌که این برنامه با یک پایگاه دادهٔ ۴۵۰ عضوی قادر به تشخیص تمامی افعال گروهی نیز است؛ و این خود نشانگر برتری نرم‌افزارهای زبانی است که برمبنای دستاوردهای زبان‌شناختی طراحی می‌شوند.

ب ـ هوشمندبودن: همان‌طور که در قسمت الف توضیح داده شد، این برنامه برمبنای جست‌وجو قرار ندارد؛ بلکه با تحلیلی ساختواژی، مرحله‌به‌مرحله، به شناسایی هوشمندانهٔ اجزای کلمه اقدام و با جداسازی و تجزیهٔ این اجزا، نوع کلمه را مشخص می‌کند. حتی در بخش اعلام مشخصات فعل، این مشخصات برای افعال گوناگون ازپیش نوشته نشده است؛ بلکه با فعال‌شدن متغیرهای گوناگون، توضیحات بخش اعلام مشخصات فعل مرحله به مرحله، کامل می‌شود، که این امر را می‌توان به‌عنوان یک نمونه از شبیه‌سازی عملکرد مغز برای تشخیص مقولات زبانی درنظر گرفت.

۷. مشکلات کار

این بخش را نگارنده از آن جهت در مقاله گنجاند که درصورتی‌که پژوهشگران دیگری بخواهند نقایص برنامهٔ موجود را رفع و آن را بهینه کنند و یا نرم‌افزارهای مشابهی برای دیگر مقولات زبانی طراحی کنند، پیش‌تر از برخی مشکلات کار آگاه باشند.

الف ـ هزینهٔ زیاد: چنین پروژه‌هایی درحد خود بسیار پرهزینه‌تر از تحقیق در زمینه‌های دیگر است. هزینه‌های تهیهٔ رایانه و نرم‌افزارهای گوناگون، و آموزش را می‌توان به‌عنوان هزینه‌های اصلی این گونه تحقیقات نام برد.

ب ـ وقت‌گیربودن: چنین تحقیقاتی بسیار وقت‌گیر است و به شکیبایی فراوان دربرابر ناکامی‌های متعدد، جواب‌نگرفتن از یک شیوه و عوض‌کردن آن در میانهٔ راه،

و... نیاز دارد.

پ ـ کار گروهی: در این‌گونه تحقیقات، منطقی‌ترین شیوه، کارکردن گروهی زبان‌شناس مطلع از دانش رایانه‌ای و افراد متخصص رایانه و آشنا با مسائل زبان‌شناسی است؛ که یافتن، ترغیب‌کردن، گردآوردن، و به‌پایان‌رساندن کارِ این جمع، بس دشوار ــ و در عین حال بسیار ارزشمند ــ است.

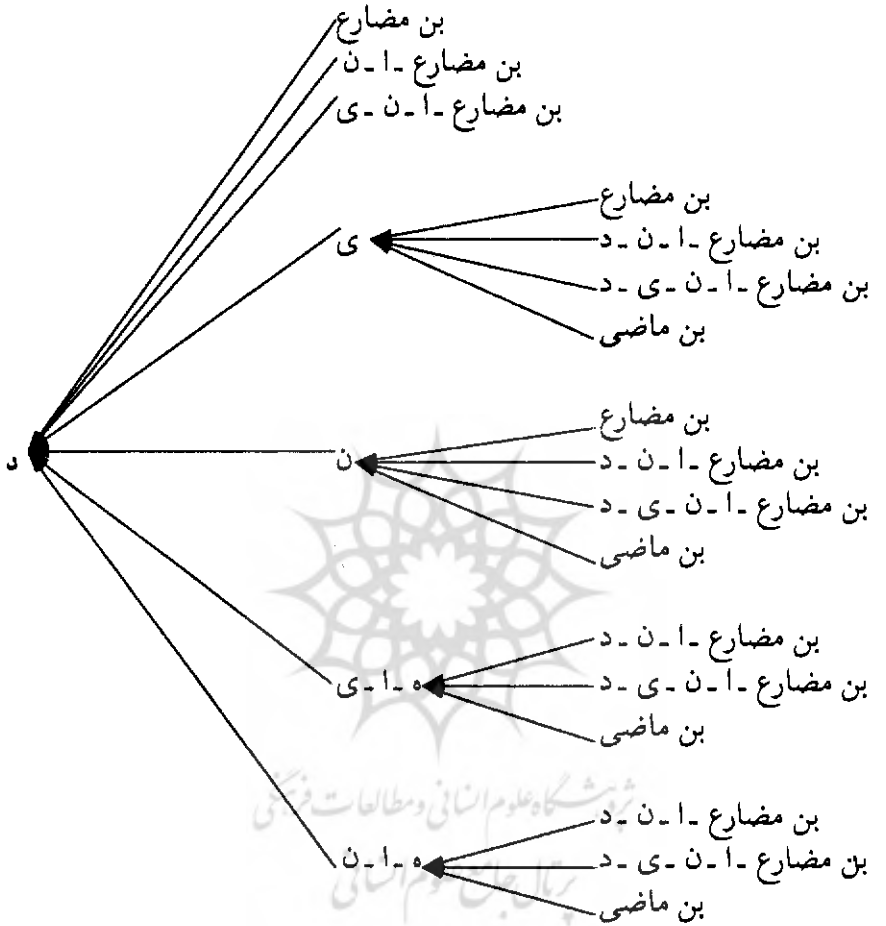
۸. نتیجه‌گیری

چنان که ملاحظه شد، ساخت نرم‌افزارهای تحلیل خودکار زبان، کاری دشوار ولی ممکن است. با پیشرفت روزافزون رایانه‌ها در ابعاد مختلف ازنظر سرعت، ظرفیت، زبان‌های برنامه‌نویسی، و مانند آن، و همچنین دستاوردهای هرچه بیشتر زبان‌شناسی، و نیز سرمایه‌گذاری‌های مالی و انسانیِ کافی در این بخش، می‌توان امیدوار بود که در آیندهٔ نزدیک، شاهد تحقق رؤیای شصت‌سالهٔ بشر در ساخت ماشین ترجمه باشیم.

مورد دیگری که اهمیت آن کمتر از مطلب فوق نیست، تأیید پیچیدگی و گستردگی قوای زبانی بشر است؛ زیرا همان‌گونه که شرح آن رفت، تنها برای تشخیص مقولهٔ فعل در زبان فارسی گونهٔ نوشتاری معیار، به حدود ۴۰۰۰ خط برنامه نیاز است؛ که خود متکی به میلیون‌ها خط برنامه‌های جانبی ازپیش‌طراحی‌شده است. اکنون می‌توان تصور کرد که اگر بخواهیم سیستمی طراحی کنیم که همهٔ مقولات دستوری یک زبان و شاید مقولات مشترک و جهانی[1] زبان‌های مختلف را بررسی کند، به چه حجمی از داده‌ها و برنامه‌ها نیاز داریم.

1. Universal

نمودار پسوندهای فعل در زبان فارسی که به حرف «د» ختم می‌شوند



کتابنامه

باطنی، محمدرضا. ۱۳۷۰. توصیف ساختمان دستوری زبان فارسی. چاپ چهارم. تهران: انتشارات امیرکبیر.

دانش کار آراسته، پویان. ۱۳۸۱. نرم‌افزار تشخیص فعل در زبان فارسی. پایان‌نامهٔ کارشناسی‌ارشد. تهران: دانشگاه علامه طباطبایی.

دبیرمقدم، محمد. ۱۳۶۷. «ساخت‌های سببی در زبان فارسی»، زبان‌شناسی، س۵، ش۱، تهران: مرکز نشر دانشگاهی.

ـــــــــــ . ۱۳۷۴. «فعل مرکب در زبان فارسی». زبان‌شناسی، س۱۲، ش۱و۲. (تاریخ انتشار: ۱۳۷۶)، تهران: مرکز نشر دانشگاهی.

عاصی، مصطفی. ۱۳۷۱. «نقش ترکیب در گسترش واژگان زبان فارسی با نگرشی بر آثار نظامی گنجوی»، فرهنگ، کتاب دهم.

ـــــــــــ . ۱۳۷۲. «کاربرد کامپیوتر در زبان‌شناسی و فرهنگ‌نگاری»، پژوهشگران، ۴، تهران: مؤسسهٔ مطالعات و تحقیقات فرهنگی.

ـــــــــــ . ۱۳۷۳. «پردازش متن فارسی با کامپیوتر»، پژوهشگران، ۱۲، تهران: مؤسسهٔ مطالعات و تحقیقات فرهنگی.

مشکوة‌الدینی، مهدی. ۱۳۷۴. دستور زبان فارسی بر پایهٔ نظریهٔ گشتاری. چاپ سوم. مشهد: دانشگاه فردوسی مشهد.

ناتل خانلری، پرویز. ۱۳۷۰. دستور زبان فارسی. چاپ دوازدهم. تهران: انتشارات توس.

ـــــــــــ . ۱۳۷۲. دستور تاریخی زبان فارسی. تهران: انتشارات توس.