پارس مورف: تحلیلگر صرفی زبان فارسی

وحید مواجی[۱]، محرم اسلامی[۲] و بهرام وزیرنژاد[۱]

[۱]دانشگاه صنعتی شریف
[۲]دانشگاه زنجان

چکیده

در مقالهٔ حاضر می‌کوشیم مبنای نظری، نحوهٔ طراحی و عملکرد سامانهٔ تحلیل‌گر صرفی زبان فارسی با عنوان اختصاری "پارس مورف" را معرفی کنیم. پارس مورف سامانه‌ای مبتنی بر قواعد صرفی زبان فارسی است که ساخت درونی کلمات فارسی را با توجه به نظام تصریف و نظام واژه‌سازی زبان تجزیه و تحلیل می‌کند و مقولهٔ دستوری و نقش هر کدام از اجزای سازندهٔ کلمه را مشخص می‌کند. پارس مورف با استفاده از یک واژگان حدوداً ۴۵۰۰۰ واژه‌ای و نیز در چارچوب قواعد صرفی زبان فارسی که بر یک تحقیق جامع زبان‌شناختی استوار است می‌تواند واژه‌های پیچیده و نیز صورت‌های ممکن تصریفی و حتی واژه‌های خارج از واژگان را تحلیل کند. دقت نسخهٔ اول پارس مورف حدود ۹۵% است که افزودن اطلاعات نحوی و مسائل مربوط به هم‌نویسه‌ها و نیز لحاظ کردن ویژگی‌های خط فارسی می‌تواند این دقت را به ۱۰۰% نزدیک کند. از پارس مورف می‌توان در مطالعات محض زبانی و نیز پردازش ماشینی زبان فارسی استفاده کرد.

کلیدواژه‌ها: پارس مورف، تحلیل‌گر صرفی، زبان فارسی، واژه‌سازی، تصریف، اشتقاق، ترکیب.

۱ مقدمه

صرف یا ساختواژه (morphology) به عنوان بخشی از زبانشناسی ناظر بر مطالعه ساخت درونی کلمات و روابط نظام‌مند صورت و معنا در کلمات است Booij (2007) . وقتی صحبت از ساخت (کلمه) می‌شود، تلویحاً می‌پذیریم که کلمه دارای اجزایی است و در نتیجه هرکدام از اجزا نقش و جایگاه مشخصی در ساخت کلمه دارد. در این تحقیق پس از صورت‌بندی ساخت درونی کلمه، سامانه‌ای را طراحی کرده‌ایم که قادر است به طور خودکار ساخت درونی کلمات فارسی را تحلیل و اجزای سازنده کلمه را مشخص کند. پیش از ورود به بحث اصلی، در ادامه به طور مختصر به مبانی صرف اشاره می‌کنیم تا در جریان بحث پیرامون عملکرد پارس مورف ابهامی وجود نداشته باشد.

ساختواژه از دو نظام مستقل تصریف (inflection) و واژه‌سازی (word-formation) تشکیل می‌شود و در تصریف صورت-کلمه‌ها (word-forms) مورد بحث قرار می‌گیرد. هر کدام از واژه‌ها (lexemes) با توجه به جایگاهشان در جمله می‌توانند صورت-کلمه‌های متفاوت داشته باشند که در اصطلاح غیردقیق به آنها کلمه (word) نیز می‌گویند. صورت-کلمه‌ها واژه‌های جدید نیستند، بلکه صورت تصریفی واژه به حساب می‌آیند. در نظام واژه‌سازی، واژه‌ها یا بسیط‌اند و یا پیچیده. واژه‌های بسیط می‌توانند با هم ترکیب شوند و واژه مرکب بسازند و از طرف دیگر واژه‌های بسیط و یا مرکب می‌توانند با عناصر واژه‌ساز (تکواژهای اشتقاقی) زبان ترکیب شوند و به ترتیب واژه‌های پیچیده مشتق و یا مرکب-مشتق بسازند. به نمونه‌های زیر توجه کنید:

الف) واژه بسیط: /کار/
ب) صورت کلمه: کار، کارها، کارش، کارشه
ج) واژه مشتق: /کارگر، کارمند، کارایی، کاری(کارکن)/
د) واژه مرکب: /کارخانه‌ساز، کاربلد، کارآگاه، کارنامه/
ه) واژه مرکب-مشتق: /کاردرمانی، کارخانه‌چی، کارمندمحور، کاردرمانگری/

با توجه به نمونه‌های ارائه‌شده در مثال ۱ دستگاه صرف دو نقش اساسی برعهده دارد: تصریف و واژه‌سازی (طباطبائی، ۱۳۸۲). واژه بسیط (۱.الف) واژه‌ای است که از یک ریشه تشکیل شده است. صورت-کلمه (۱.ب) صورت‌های تصریف شده یک واژه را می‌گویند. از ۱.ج تا ۱.ه واژه‌های پیچیده را می‌بینیم که در چارچوب امکانات واژه‌سازی زبان درست شده‌اند. واژه مشتق (۱.ج) واژه‌ای است که در ساختمان آن یک واژه بسیط یا واحد واژگانی بسیط و یک یا چند تکواژ اشتقاقی به کار رفته است. واژه مرکب واژه‌ای است که از دو یا چند واژه بسیط یا واحد واژگانی بسیط درست شده باشد. واژه مرکب-مشتق (۱.ه) واژه‌ای است که در ساختمان آن دو یا چند واژه بسیط یا واحد واژگانی بسیط و یک یا چند تکواژ اشتقاقی به کار رفته باشد. با توجه به مثال‌های بالا می‌بینیم که بخشی از تکواژهای مقید واژه جدید نمی‌سازند (مانند ۱.ب) که به آنها تکواژهای غیراشتقاقی و تسامحاً کل آنها را تکواژهای تصریفی می‌گوییم. بخشی دیگر (در ۱.ج و ۱.ه) واژه جدید می‌سازند که به آنها تکواژهای اشتقاقی می‌گوییم. به عبارت دیگر در تصریف با واژه‌پردازی سروکار داریم، اما در اشتقاق با واژه‌سازی (طباطبائی، ۱۳۷۶). ستاک (stem) به صورت-کلمه‌ای اطلاق می‌شود که وندهای تصریفی از آن حذف شده باشند. ستاک می‌تواند بسیط یا پیچیده باشد و ستاک بسیط همان ریشه است Booij (2007).

در طراحی پارس مورف، ساخت تصریفی کلمه در زبان فارسی (اسلامی و علیزاده، ۱۳۸۸) به صورت ساختمند در نظر گرفته شده است و جایگاه طبقات مختلف وندهای تصریفی در ساختمان انواع کلمات مشخص شده است. پارس مورف در تحلیل تصریفی کلمه ابتدا ستاک را شناسایی می‌کند و متناسب با ساخت تصریفی پیش‌بینی شده برای آن نوع ستاک به دنبال انواع وندهای تصریفی در جایگاه‌های خاص، با در نظر گرفتن صورت‌های مختلف نوشتاری هرکدام از طبقات، می‌گردد. یافتن وندی مثلاً در جایگاه دوم پس از ستاک اسمی نشان می‌دهد که جایگاه اول خالی مانده است. در بخش ۲ در مقاله حاضر به طور کامل به موضوع ساخت تصریفی کلمه در زبان فارسی پرداخته‌ایم.

اگر ستاک پیچیده باشد، پارس مورف با اعمال قواعد واژه‌سازی زبان فارسی، ستاک مورد نظر را از حیث مشتق یا مرکب بودن تجزیه و تحلیل و اجزای سازنده آن را با ذکر مقوله دستوری و نقش آنها مشخص می‌کند. به دلیل معتبر نبودن فاصله به عنوان مرز کلمه در متون فارسی، پارس مورف در تجزیه ستاک‌های پیچیده ترکیب‌های بالقوه را نیز به عنوان گزینه‌های بعدی در اختیار ما می‌گذارد. مثلاً ستاک پیچیده "کارگر" در بخش اشتقاق به عنوان کلمه مشتق شناسایی می‌شود و یا در ترکیب گفته می‌شود که در فارسی ممکن است آن ترکیب "کار (N۱) + گر "(Adv) یعنی اسم + قید باشد که این دو کلمه بی‌فاصله در کنار هم آمده‌اند. در بخش ۳ به عملکرد پارس مورف در ریشه‌یابی و تعیین اجزای واژه‌های پیچیده خواهیم پرداخت.

در تمامی تحقیقات مربوط به پردازش‌های خودکار زبانی در زبان فارسی، به خصوص در پردازش متن فارسی برای مقاصد مختلف اعم از ترجمه ماشینی، تبدیل متن به گفتار و غیره از نوعی تحلیل‌گر صرفی استفاده می‌شود؛ اگر چه در اغلب مواقع تحلیل‌گرهای صرفی در تحقیقات پیشین محدود، هدف‌محور و فاقد پشتوانه جامع زبان‌شناختی است. به عنوان مرور پیشینه پژوهش در ادامه تنها به مواردی اشاره می‌کنیم که در تحلیل ساخت درونی کلمه فارسی نگاه ساخت‌مند داشته‌اند و با یک رویکرد زبان‌شناختی- مهندسی به تحلیل صرفی کلمه فارسی پرداخته‌اند. در این خصوص ابتدا می‌توان به مطالعات دقیقی اشاره کرد که در پروژه شیراز Megerdoomian (2000) در طراحی ترجمه سامانهٔ ماشینی فارسی-انگلیسی اشاره کرد که در میانه راه متوقف شد. دومین مورد از طراحی تحلیل خودکار فارسی مربوط به تحلیل‌گر تصریفی زبان فارسی است که در سامانه تبدیل متن به گفتار فارسی "گویا" به کار گرفته شد (اسلامی و دیگران، ۱۳۸۳) که تنها به تحلیل تصریفی کلمه محدود می‌شد. آماده‌سازی متن معیار برای زبان فارسی ۱ با عنوان اختصاری STeP۱ سامانه دیگری است که قادر است کلمات فارسی را از نظر صرفی تجزیه و تحلیل کند (شمس‌فرد، ۱۳۸۸). STeP۱ با استفاده از واژگان زایای زبان فارسی (اسلامی و دیگران، ۱۳۸۳) و قواعد صرفی که طراحان آن در نظر گرفته‌اند کار می‌کند.

اگرچه STeP۱ در تجزیه کلمه به اجزای سازنده آن تا حد زیادی موفق عمل می‌کند ولی مبنای علمی زبان‌شناختی

آن نیاز به بازنگری دارد تا از اشکالات فعلی پرهیز شود. به عنوان مثال برای کلمه "دانشگاهها" سه ریشه در نظر می‌گیرد، به ترتیب "دانشگاه، دانش، دان" که نشان می‌دهد حداقل منظور پدیدآورندگان آن از "ریشه" در مفهوم علمی آن اصطلاح نسیت. یا در کلمه "دانشگاههایمان" علاوه بر اختصاص سه ریشه فوق همزمان دو تحلیل زیر برای آن ارائه می‌شود:

الف) اسم + علامت جمع + ی + ضمیر ملکی اول شخص جمع که در آن "دانشگاه" ریشه است، به صورت "دانشگاه + ها + ی + مان".

ب) اسم + گاه + علامت جمع + ی + ضمیر ملکی اول شخص جمع که در آن "دانش" ریشه است، به صورت "دانش + گاه + ها + ی + مان".

چنانچه می‌بینیم هیچ مبنای علمی برای ریشه‌بودن "دانشگاه" یا "دانش" در مثال فوق وجود ندارد، از طرف دیگر اجزای تجزیه شده لزوماً واحدهای نظام واژه‌سازی و نظام تصریف نیستند. به عنوان مثال "ی" در مثال بالا چه عنوان و نقش زبانی دارد؟ جز اینکه به خاطر شرایط واژ-واجی در زنجیره واجی کلمه ظاهر شده است. همچنین کلماتی مانند "کفشهایمه" یا "کفشهایمند" را STeP۱ نمی‌تواند تجزیه کند و هیچ ریشه‌ای برای این قبیل تصریف‌ها و صورت-کلمه‌ها و همچنین برخی ستاک‌های پیچیده و بسیط خارج از واژگان پیدا نمی‌کند. این در حالیست که پارس مورف با دقت کامل موارد فوق را تجزیه و تحلیل می‌کند.

پارس مورف، تحلیل‌گر صرفی زبان فارسی، بر یک مبنای کاملاً علمی زبانی استوار است و در چارچوب ساختمان صرفی کلمه فارسی به تجزیه و تحلیل و تعیین نقش هر کدام از اجزا در درون کلمه می‌پردازد. پس از صورت‌بندی دقیق اطلاعات نظام تصریف و واژه‌سازی در زبان فارسی سعی کردیم در مرحله اجرا و طراحی سامانه پارس مورف تمامی آن اطلاعات را به شکل دقیق به کار بگیریم. در حال حاضر پارس مورف با استفاده از آخرین ویرایش واژگان زایای زبان فارسی (اسلامی و دیگران، ۱۳۸۳) که حدود ۴۵۰۰۰ واژه در آن قرار دارد و در چارچوب قواعد صرفی زبان فارسی که در اختیار دارد، با دقت بالای ۹۵% می‌تواند ساخت درونی کلمات فارسی را تحلیل کند. نیز می‌تواند با استفاده از امکانات و اطلاعات صرفی که در اختیار دارد کلمات خارج از واژگان را نیز از حیث تصریف و واژه‌سازی تجزیه و تحلیل کند.

مراجع

Booij, G. (2007), *The Grammar of Words*, Oxford University Press, 2nd ed. . ۱, ۲

Megerdoomian, K. (2000), "Persian Computational Morphology, A Unification-Based Approach (Memoranda in Computer and Cognitive Science)," Tech. Rep. MCCS-00-321, Computing Research Laboratory, New Mexico State University. ۲

اسلامی، محرم، شریفی آتشگاه، مسعود، علیزاده لمجیری، صدیقه، و زندی، طاهره (۱۳۸۳)، "واژگان زایای زبان فارسی،" در اولین کارگاه پژوهشی زبان فارسی و رایانه، دانشگاه تهران. ۲، ۳

اسلامی، محرم و علیزاده، صدیقه (۱۳۸۸)، "ساخت تصریفی کلمه در زبان فارسی،" زبان و ادب فارسی، ۵۲. ۲

شمس‌فرد، مهرنوش (۱۳۸۸)، "STeP۱: تهیه متن معیار برای زبان فارسی،" گزارش طرح تحقیقی، آزمایشگاه پردازش زبان طبیعی، دانشگاه شهید بهشتی. ۲

طباطبائی، علاءالدین (۱۳۷۶)، فعل بسیط فارسی و واژه‌سازی. مرکز نشر دانشگاهی. ۲

طباطبائی، علاءالدین (۱۳۸۲)، اسم و صفت مرکب در زبان فارسی. مرکز نشر دانشگاهی. ۲