

پارس مورف: تحلیلگر صرفی زبان فارسی *

وحید مواجی^۱، محرم اسلامی^۲ و بهرام وزیرنژاد^۱

^۱ دانشگاه صنعتی شریف

^۲ دانشگاه زنجان

۲۷ اردیبهشت ۱۳۹۱

چکیده

در مقاله حاضر می‌کوشیم مبنای نظری، نحوه طراحی و عملکرد سامانه 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 مقوله دستوری و نقش هر کدام از اجزای سازنده 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 نسخه 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^۱ 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^۲ و واژه‌سازی^۳ تشکیل می‌شود و در تصریف صورت-کلمه‌ها^۴ مورد بحث قرار می‌گیرد. هر کدام از واژه‌ها^۵ با توجه به جایگاهشان در جمله می‌توانند صورت-کلمه‌های متفاوت داشته

*دوفصل‌نامه علمی-پژوهشی پردازش علائم و داده‌ها، ۱۳۹۰، شماره ۱، پیاپی ۱۵

^۱morphology

^۲inflection

^۳word-formation

^۴word-forms

^۵lexemes

باشند که در اصطلاح غیردقیق به آنها کلمه^۶ 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) (Bauer (1983). واژه بسیط (الف) واژه‌ای است که از یک ریشه تشکیل شده است. صورت-کلمه (ب) صورت‌های تصریف شده یک واژه را می‌گویند. از ج. ۱ تا ۰.۱ واژه‌های پیچیده را می‌بینیم که در چارچوب امکانات واژه‌سازی زبان درست شده‌اند. واژه مشتق (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 (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 ۰.۱ ب) که به آنها تکواژهای غیراشتقاقی و تسامحاً کل آنها را تکواژهای تصریفی می‌گوییم. بخشی دیگر (در ۰.۱ ج و ۰.۱ 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 علیزاده‌لمجیری، ۱۳۸۸) به صورت ساختمان 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 وندهای 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 Megerdooonian

⁶word

(2000) در طراحی ترجمه سامانه ماشینی فارسی-انگلیسی اشاره کرد که در میانه راه متوقف شد. دومین مورد از طراحی تحلیل خودکار فارسی مربوط به تحلیل گر 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 زبانی 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 نیست. 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 "کفشهایم" 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 حالیکه پارس مورف با دقت کامل موارد فوق را تجزیه و تحلیل می‌کند.

پارس مورف، تحلیل گر 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 زبانی 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.

۲ نظام تصریف زبان فارسی

چنانچه در مقدمه گفتیم، در طراحی پارس مورف رویکرد ساخت‌بنیاد اتخاذ شده است که به موجب آن عناصر صرفی تشکیل‌دهنده کلمه در یک ارتباط ساختاری و نظام‌مند با همدیگر قرار دارند و در ساخت کلمه، اجزا هر کدام عنوان و نقش منحصر به فردی دارند. پارس مورف صورت واژگانی کلمه را از واژگان زبانی فارسی (اسلامی و دیگران، ۱۳۸۳) می‌گیرد و در چارچوب نظام تصریف کلمه در زبان فارسی (اسلامی و عزیززاده‌لمجیری، ۱۳۸۸) به تحلیل تصریفی کلمه می‌پردازد. در واژگان زبانی مقوله دستوری هر کدام از واژه‌ها مشخص شده است و بعد از شناسایی واژه با توجه به مقوله دستوری آن ساخت تصریفی را برای آن در نظر می‌گیرد. اگر در ساختمان کلمه از وابسته‌های تصریفی پیش‌بینی شده در جایگاه‌های خاص مورد یا مواردی وجود داشته باشد، در آن صورت عنوان هر کدام را مشخص می‌کند. واژه ممکن است بدون هیچ گونه وند تصریفی در جمله به کار رود که در چنین شرایطی واژه و کلمه صورت یکسانی خواهند داشت. وابسته‌های تصریفی عناصر اختیاری در ساختمان کلمه هستند و پارس مورف قادر است حتی با استفاده از قواعد تصریف که در اختیار دارد کلمات خارج از واژگان

را شناسایی و عناصر آن را مشخص کند و اگر ستاک ناشناخته پیچیده باشد، اجزای آن را هم با توجه به قواعد واژه‌سازی روشن می‌کند. در خصوص کلمه در کلمه‌ها (words within word) نیز در پارس مورف تمهیداتی اندیشیده شده است که در گزینه‌های بعدی که در اختیار کاربر قرار می‌دهد به آنها اشاره می‌کند. کلمه در کلمه‌ها واژه‌هایی هستند که در درون آنها می‌توان کلمات دیگری نیز پیدا کرد مانند "مادر" که در عین حال واژه بسیط و دو کلمه مجزا ("ما + در") است. البته در شناسایی کلمه در کلمه‌ها در پارس مورف از قواعد صرفی استفاده کرده‌ایم. در ادامه به ساخت تصریفی اسم و فعل اشاره می‌کنیم. شایان ذکر است که در زبان فارسی، صفت‌ها بالقوه اسم هستند و با اشتقاق صفر به اسم تبدیل می‌شوند و صورت‌های تصریفی اسم را می‌پذیرند. این موارد در طراحی پارس مورف لحاظ شده است. در ساخت تصریفی اسم، هسته صورت واژگانی اسم و عنصر اجباری است که می‌تواند وابسته‌های تصریفی اختیاری داشته باشد که به همین دلیل در ساخت تصریفی اسم در درون () قرار دارند.

$$[اسم] + [(تکواژ جمع)] + \left[\begin{array}{c} (یای نکره) \\ (یای بند موصولی) \\ (واژه‌بست‌های شخصی / ضمائر متصل) \\ (کسره اضافه) \end{array} \right] + [(واژه‌بست‌های ربطی)]$$

شکل ۱: ساختار تصریفی اسم (اسلامی و علیزاده‌لمجیری، ۱۳۸۸)

چنانچه در ساخت تصریفی اسم می‌بینیم، اسم بالقوه می‌تواند در سه جایگاه، وابسته‌هایی بپذیرد که این جایگاه‌ها معتبر هستند و هرگونه جابجایی در نظم وابسته‌ها منجر به کلمه بدساخت می‌شود. در هر کدام از جایگاه‌ها نیز صورت‌های مختلف تکواژ (مثلاً تکواژهای مختلف جمع در جایگاه اول) یا طبقات گوناگونی از وابسته‌ها به کار می‌رود که حضور یکی از آنها مانع حضور بقیه گونه‌های یک تکواژ و یا دیگر طبقات مربوط به آن جایگاه می‌شود. بنابراین پارس مورف هیچگاه دو تکواژ جمع یا دو تکواژ مربوط به جایگاه دوم یا سوم در ساختمان اسم شناسایی نمی‌کند. به این محدودیت اصطلاحاً توزیع تکمیلی^۷ می‌گویند که پارس مورف در تحلیل ساخت تصریفی کلمه محدودیت بالا را از نظر تصریف و واژه‌سازی لحاظ می‌کند. در واژه‌سازی نیز پارس مورف تکرار تکواژ مثلاً "گر" را به طور همزمان در ساختمان کلمه مجاز نمی‌داند. تحلیل ساخت تصریفی فعل به دلیل پذیرفتن پیشوند و پسوند تصریفی دشوارتر است. در واژگان زایا، بن مضارع و بن ماضی صورت‌های واژگانی فعل در نظر گرفته شده است و پارس مورف با توجه به این موضوع ساخت تصریفی فعل را تحلیل می‌کند. در ادامه ساخت تصریفی فعل فارسی را می‌بینیم که محدودیت‌های ذکر شده در خصوص اسم در فعل نیز مشاهده می‌شود.

$$[(تکواژ نفی)] + [(تکواژ وجه امری و التزامی) / (تکواژ استمراری)] + [فعل] + [(تکواژ نمود کامل)] + [(شناسه‌ها)] + [(واژه‌بست‌های ربطی)]$$

شکل ۲: ساختار تصریفی فعل (اسلامی و علیزاده‌لمجیری، ۱۳۸۸)

دیگر اقسام کلمه مانند صفت‌ها، قیدها، اعداد، حروف اضافه نیز ساخت تصریفی خاص خود را دارند که در طراحی پارس مورف به طور کامل مد نظر بوده‌اند.

⁷complementary distribution

۳ نظام واژه‌سازی زبان فارسی

نظام واژه‌سازی زبان فارسی نیز همانند نظام تصریف این زبان در طراحی پارس مورف با دقت صورت‌بندی شده است. دو شیوه عمده واژه‌سازی یعنی اشتقاق^۸ و ترکیب^۹ در زبان فارسی عمده‌ترین شیوه‌های واژه‌سازی هستند. پارس مورف با مراجعه به واژگان زبانی زبان فارسی (اسلامی و دیگران، ۱۳۸۳) و نیز با توجه به امکانات اشتقاق و ترکیب به تجزیه کلمات مشتق و مرکب می‌پردازد.

۱.۳ اشتقاق در زبان فارسی

فهرست تکواژهای اشتقاقی با استفاده از ساخت اشتقاقی واژه در فارسی امروز (کلباسی، ۱۳۷۱) و اشتقاق پسوندی در زبان فارسی امروز (کشانی، ۱۳۷۱) و نیز ملاحظات نویسندگان این مقاله تهیه شد و پارس مورف با در اختیار داشتن این فهرست و الگوهای اشتقاق به تجزیه کلمات مشتق فارسی می‌پردازد. در الگوهای اشتقاق مشخص شده است که هر کدام از تکواژهای اشتقاقی به چه نوع ستاکی از نظر مقوله افزوده می‌شود و واژه حاصل از رهگذر اشتقاق چه مقوله دستوری خواهد داشت. به عنوان مثال کلمه "دانشگاهی" به صورت "دان(V1) + ش(اسم‌ساز=N) + گاه(N=اسم‌ساز) + ی(N=اسم‌ساز)" تجزیه می‌شود. پارس مورف با در اختیار داشتن امکانات ذکر شده می‌تواند کلمات خارج از واژگان را نیز تجزیه و تحلیل کند و نیز ترکیب‌های بالقوه را برای کلمه مورد نظر در گزینه‌های بعدی پیشنهاد کند. نظر به عدم قطعیت در استفاده از فاصله در متون فارسی به عنوان مرز کلمه، پارس مورف این قابلیت را دارد که علاوه بر تجزیه و تحلیل کلمه در سطح اشتقاق برای کلمه مورد نظر صورت‌های ترکیبی نیز در نظر بگیرد. با افزودن اطلاعات نحوی به پارس مورف در موارد مشکوک به اشتقاق یا ترکیب و یا دو کلمه مستقل بدون فاصله می‌توان از گزینه‌هایی دیگری استفاده کرد که پارس مورف ارائه می‌کند. مسأله مهم دیگر در تجزیه کلمات مشتق، موضوع هم‌نویسه‌ها^{۱۰} است که کار تجزیه خودکار کلمه را دچار مشکل می‌کند. به عنوان مثال "ی" با ساخت واجی "i/" می‌تواند در نقش‌های زیر به کار رود: الف) تکواژ اشتقاقی (یای نسبت) که معمولاً اسم را به صفت و برعکس صفت را به اسم تبدیل کند. مثلاً در کلمات "دانشگاهی، زیبایی"، ب) در نقش "یای" نکره مثلاً در "کتابی" در جمله "کتابی خریدم" که اشاره به کتاب نامعین می‌کند، ج) در نقش "یای" بند موصولی ساز، مثلاً در "پسری" در جمله "پسری که دیدی" و د) در نقش واژه‌بست ربطی، مثلاً در "تو پسری" به معنی "تو پسر هستی". علاوه بر صورت‌های یکسان تکواژهای اشتقاقی و غیراشتقاقی نوع دیگری از هم‌نویسه هست که در آنها بین صورت واژگانی و صورت تصریفی ابهام وجود دارد. مثلاً "مردم" که به صورت‌های زیر تجزیه می‌شود:

الف) مردم(N1)

ب) مرد(N1) + am (واژه بست ربطی)

ج) مرد(N1) + am (واژه بست شخصی یعنی مردِ من)

د) مُرد(V2) + am (واژه بست شخصی)

البته در ساخت‌های مبهم از اطلاعات واژه‌سازی و تصریفی برای رفع ابهام استفاده می‌شود. مثلاً در "دانشگاهی‌ها"، پارس مورف به هیچ وجه "ی" را "یای" غیر اشتقاقی قلمداد نمی‌کند چرا که قبل از تکواژ جمع آمده است.

۲.۳ ترکیب در زبان فارسی

با توجه به تعریفی که از ترکیب در مقدمه مقاله حاضر ارائه کردیم، اصول و الگوهای ترکیب صورت‌بندی و پارس مورف با مراجعه به واژگان زبانی زبان فارسی (اسلامی و دیگران، ۱۳۸۳) و با در دست داشتن الگوهای ترکیب

⁸derivation

⁹compounding

¹⁰homographs

(طباطبائی، ۱۳۸۲) کلمات مرکب را شناسایی و اجزای ترکیب را با ذکر مقوله دستوری آنها مشخص می‌کند. واژه‌های بسیط و واحدهای واژگانی غیر پیچیده (مانند بن‌های مضارع) در چارچوب الگوهای ترکیب (همان) می‌توانند با هم ترکیب شوند و واژه مرکب درست کنند. در اغلب مواقع اجزای واژه مرکب خود مدخل‌های واژگانی در واژگان برنامه هستند که در آن صورت مقوله‌های دستوری اجزای ترکیب و نیز مقوله دستوری واژه مرکب مشخص است. کلمات مرکب خارج از واژگان نیز با مراجعه به واژگان برنامه و با در نظر گرفتن الگوهای ترکیب تجزیه می‌شوند.

در خصوص کلمات مرکب- مشتق نیز پارس مورف قواعد اشتقاق و ترکیب را همزمان لحاظ می‌کند و این واژه‌ها را به اجزای سازنده‌اش تجزیه می‌کند که در آن نقش و مقوله دستوری هر کدام از اجزا مشخص شده است. به عنوان مثال کلماتی مانند "کارخانه‌چی" و "یا دانشجو" را به ترتیب به صورت "کار(N1) + خانه(N1) + چی تکواژ(N=) اشتقاقی اسم‌ساز) و "دان(V1) + ش تکواژ(N=) اشتقاقی اسم‌ساز) + جو(V1) تجزیه می‌کند.

۴ نتیجه‌گیری

در این تحقیق، سامانه تحلیل صرفی کلمه در زبان فارسی با عنوان پارس مورف را معرفی کرده‌ایم که بر پایه یک مطالعه دقیق زبان‌شناختی از نظام صرفی زبان فارسی استوار است. پارس مورف قادر است ساخت درونی کلمه فارسی را از حیث نظام تصریف و واژه‌سازی تجزیه و تحلیل کند و برای هر کدام از اجزا در درون کلمه عنوان و نقش زبانی خاصی اختصاص دهد. در اجرای این تحقیق پس از صورت‌بندی دقیق نظام تصریف و واژه‌سازی، الگوریتم عملیاتی برنامه تحلیل‌گر فارسی پی‌ریزی و تکمیل شد. پارس مورف به لحاظ نظری بر پایه علمی استوار است و نسخه اول آن قادر است با دقت بالای ۹۵٪ ساخت درونی کلمات فارسی را تجزیه و تحلیل کند و با افزودن اطلاعات مربوطه به هم‌نویسه‌ها و افزودن اطلاعات نحوی و نیز رفع خطای برنامه که جزء لاینفک فرایند توسعه نرم‌افزار است این دقت به نزدیک ۱۰۰٪ خواهد رسید.

پارس مورف می‌تواند در دو زمینه کاملاً مجزا به کار گرفته شود و یاریگر محققان در این دو زمینه باشد. در مطالعات زبان‌شناختی محض؛ مثلاً در تهیه منابع زبانی برای زبان فارسی و نیز در زمینه انواع پردازش خودکار زبان فارسی به طور مشخص در تبدیل متن به گفتار، ترجمه ماشینی، استخراج اطلاعات، خلاصه‌سازی متون و غیره به کار گرفته شود.

مراجع

- Bauer, L. (1983), English Word-Formation, Cambridge University Press. 2
- Booij, G. (2007), The Grammar of Words, Oxford University Press, 2nd ed. . 1, 2
- Megerdooian, K. (2000), "Persian Computational Morphology: A unification-based approach," in NMSU, CLR, Memoranda in Computer and Cognitive Science Report. 2
- اسلامی، محرم، شریفی آتشگاه، مسعود، علیزاده لمجیری، صدیقه، و زندی، طاهره (۱۳۸۳)، "واژگان زبانی فارسی"، در اولین کارگاه پژوهشی زبان فارسی و رایانه، دانشگاه تهران. ۳، ۵
- اسلامی، محرم و علیزاده لمجیری، صدیقه (۱۳۸۸)، "ساختار تصریفی کلمه در زبان فارسی"، زبان و ادب فارسی، ۱۸-۱. ۲، ۳، ۴

شمس فرد، مهنوش (۱۳۸۸)، "STeP۱: تهیه متن معیار برای زبان فارسی"، گزارش طرح تحقیقی، آزمایشگاه پردازش زبان طبیعی، دانشگاه شهید بهشتی. ۳

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. ۶، ۲

کشانی، خسرو (۱۳۷۱)، اشتقاق پسوندی در زبان فارسی امروز. مرکز نشر دانشگاهی. ۵

کلباسی، ایران (۱۳۷۱)، ساخت اشتقاقی واژه در فارسی امروز. موسسه مطالعات و تحقیقات فرهنگی. ۵