

## تبدیل متن فارسی به زنجیره واجی با استفاده از تحلیلگر صرفی

وحید مواجی<sup>۱</sup>، محرم اسلامی<sup>۲</sup>

<sup>۱</sup>مرکز زبان‌ها و زبان‌شناسی، دانشگاه صنعتی شریف [mavaij@alum.sharif.edu](mailto:mavaij@alum.sharif.edu)

<sup>۲</sup>دانشگاه زنجان [melsami@znu.ac.ir](mailto:melsami@znu.ac.ir)

### چکیده

در مقاله حاضر می‌کوشیم روشی خودکار برای تبدیل متون فارسی به زنجیره واجی ارائه دهیم. خط فارسی به دلیل دشواری‌های پردازشی که دارد ورودی مناسبی برای برنامه‌های پردازش متن به حساب نمی‌آید. از ویژگی‌های خط فارسی می‌توان به عدم نمایش واژه‌های کوتاه و به دنبال آن موضوع هم‌نویسه‌گی، مسأله کسره اضافه، فاصله بین اجزای کلمه واحد، فقدان فاصله بین کلمه‌های مستقل، موضوع جدانویسی و پیوسته‌نویسی و غیره اشاره کرد. برخورداری خط فارسی از ویژگی‌های که برشمردیم موجب می‌شود قبل از انجام هرگونه پردازشی، متون فارسی را به زنجیره واجی تبدیل کنیم. خروجی برنامه تبدیل متن به زنجیره واجی کاربردهای متعددی منجمله در تبدیل خودکار متن به گفتار، واج‌نویسی صحیح متون، آموزش زبان فارسی به غیرفارسی‌زبانان، فرهنگ نویسی و غیره دارد. در این مقاله با استفاده از تحلیلگر صرفی پارس-مورف که توسط نگارندگان طراحی و پیاده‌سازی شده است، متن ورودی از لحاظ صرفی تحلیل شده و اجزای صرفی آن از قبیل پیشوندها، پسوندها، اشتقاق و ترکیب بدست آمده و سپس با استفاده از واژگان زبانی فارسی، صورت واجی آنها با هم ترکیب شده و در نهایت صورت واجی متن ورودی به دست می‌آید.

### کلمات کلیدی

متن فارسی، زنجیره واجی، تحلیلگر صرفی، تبدیل متن به گفتار

### ۱- مقدمه

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 در پروژه شیراز [4] در طراحی سامانه ترجمه 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 [5] که تنها به تحلیل تصریفی کلمه محدود می‌شد. آماده‌سازی متن معیار برای زبان فارسی ۱ با عنوان اختصاری STePI سامانه دیگری است که قادر است کلمات فارسی را از نظر صرفی تجزیه و تحلیل کند [6]. STePI با استفاده از واژگان زبانی فارسی [7] و قواعد صرفی که طراحان آن در نظر گرفته‌اند کار می‌کند.

در این مقاله روشی جدید برای تبدیل خودکار متن فارسی به زنجیره فارسی ارائه می‌شود. در این روش با استفاده از تحلیلگر صرفی که طراحی کرده‌ایم عبارت ورودی را به اجزای

یک سامانه تبدیل متن به گفتار از دو قسمت تبدیل متن به زنجیره واج‌های تشکیل‌دهنده آن و نیز قسمت تبدیل زنجیره واج‌ها به گفتار تشکیل می‌گردد. در این مقاله روی قسمت اول یعنی تبدیل متن به زنجیره واجی تمرکز داریم. روش‌های متعددی برای تبدیل متن به زنجیره واجی مورد استفاده قرار گرفته است. در [1] از قواعد تبدیل نویسه به صورت واجی استفاده شده و استثنائات نیز از یک فرهنگ استخراج می‌شود. استفاده از یک درخت تصمیم چندسطحی که هر نویسه را نسبت به حروف مجاور آن به صورت یک درخت نمایش می‌دهد در [2] مورد مطالعه قرار گرفته است. استفاده از روش‌های زبان طبیعی نیز مورد بررسی قرار گرفته است. در این روش، هر تکواژ به همراه اطلاعات مربوط به صورت‌های صرفی مختلف آن مانند صورت جمع، گذشته، حال، و غیره و کلیه اطلاعات صرفی مربوطه در یک دادگان ذخیره می‌گردد. در این حالت نیز اگر کلمه در فهرست تکواژها موجود نبود از قواعد نویسه به صورت واجی یا فرهنگ استثناءها استفاده می‌شود [3].

در تمامی تحقیقات مربوط به پردازش‌های خودکار زبانی در زبان فارسی، به خصوص در پردازش متن فارسی برای مقاصد



آخرین ویرایش واژگان زبانی فارسی [۷] که حدود ۴۵۰۰۰ واژه در آن قرار دارد و در چارچوب قواعد صرفی 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.

تشکیل دهنده آن تقطیع می‌کنیم و سپس هر قطعه را به عنوان ورودی به تحلیلگر صرفی می‌دهیم تا به اجزای صرفی تشکیل دهنده خود از قبیل پسوندها و پیشوندهای تصریفی و اشتقاقی، پایه‌های ترکیب و اشتقاق و غیره تجزیه شود و سپس صورت واجی اجزا با هم ترکیب شده و زنجیره واجی کل متن به دست می‌آید.

## ۲- تحلیلگر صرفی

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. در طراحی پارس مورف، ساخت تصریفی کلمه در زبان فارسی [۸] 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 وندهای مختلف در جایگاه دوم پس از ستاک اسمی نشان می‌دهد که جایگاه اول خالی مانده است.

## ۳- تبدیل متن به زنجیره واجی

در برنامه تحلیلگر صرفی پارس-مورف گزینه‌ای به نام Phonology وجود دارد که از طریق آن متن فارسی به عنوان ورودی برنامه داده شده و نتیجه به صورت زنجیره واجی نشان داده می‌شود. برای علائم واجی زبان فارسی از علائم موجود در [۹] استفاده کرده‌ایم. مثلاً واج‌های  $\text{?}$ ،  $\text{š}$ ،  $\text{â}$ ،  $\text{j}$ ،  $\text{č}$  و  $\text{ž}$  به ترتیب نمایانگر نویسه‌های /همزه/، /ش/، /آ/، /ج/، /چ/ و /ژ/ می‌باشند.

خروجی برنامه برای جمله "آن یکی نحوی به کشتی درنشت" در شکل (۱) آمده است. همانطور که مشاهده می‌گردد، دو صورت واجی برای این عبارت نشان داده شده است که در یکی کلمه "کشتی" به صورت /kašti/ و در دیگری به صورت /kešti/ واج‌نویسی شده است.

?ân yeki nahvi beh kašti dar nešast
?ân yeki nahvi beh kešti dar nešast

شکل (۱): زنجیره واجی عبارت "آن یکی نحوی به کشتی درنشت"

در شکل (۲)، زنجیره واجی عبارت "مردم حضور دارند" آمده است. در این حالت هم دو نوع واج‌نویسی برای کلمه "مردم" آمده است: اولی دارای مقوله دستوری اسم است که به صورت /mardom/ نوشته شده و دیگری نتیجه تحلیل صرفی به صورت فعل "مردن" در حالت اول شخص مفرد است یعنی  $\text{mord+am}$ . از آنجا که تحلیل نحوی روی عبارت ورودی انجام نمی‌شود، این زنجیره واجی نیز از لحاظ برنامه معتبر است. ولی با افزودن اطلاعات نحوی به برنامه می‌توان دقت آن را بالاتر برد.

mardom hozur dârand
mordam hozur dârand

شکل (۲): زنجیره واجی عبارت "مردم حضور دارند"

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 (NI) + گر (Adv)" یعنی اسم + قید باشد که این دو کلمه بی‌فاصله در کنار هم آمده‌اند.

پارس مورف، تحلیل‌گر صرفی زبان فارسی، که با زبان برنامه‌نویسی جاوا (Java) نوشته شده است 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



[۵] اسلامی، محرم، مسعود شریفی آتشگاه، زهرا احمدی نیا، علی بهرامی راد و طاهره زندی، "تبدیل رایانه‌ای متن به گفتار فارسی (گویا)"، مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه دانشگاه تهران، ایران ۱۳۸۳.

[۶] شمس‌فرد، مهنوش، "STePI: تهیه متن معیار برای زبان فارسی". آزمایشگاه پردازش زبان طبیعی، دانشگاه شهید بهشتی، ایران، ۱۳۸۸.

[۷] اسلامی، محرم، مسعود شریفی آتشگاه، صدیقه علیزاده لمجیری و طاهره زندی، "واژگان زایای زبان فارسی"، مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه، دانشگاه تهران، تهران، ۱۳۸۳.

[۸] اسلامی، محرم و صدیقه علیزاده، "ساخت تصریفی کلمه در زبان فارسی"، زبان و ادب فارسی، نشریه دانشکده ادبیات و علوم انسانی دانشگاه تبریز، سال ۵۲، شماره مسلسل ۲۱۱، تبریز، ۱۳۸۸.

[۹] ثمره، یدالله، آواشناسی زبان فارسی، ویرایش دوم، مرکز نشر دانشگاهی، تهران، ۱۳۷۸.

[۱۰] اسلامی، محرم، دشواری‌های پردازش رایانه‌ای خط فارسی، نشر دانش، سال نوزدهم، شماره سوم، پاییز ۱۳۸۱، مرکز نشر دانشگاهی، تهران، ۱۳۸۱.

مسئله بعدی، مسأله کسره اضافه است که باید تمهیدی برای آن اندیشیده شود. چون کسره اضافه در متن فارسی نمایش داده نمی‌شود، ولی در زنجیره واجی حضور دارد؛ لازم است اطلاعات نحوی به برنامه تبدیل متن به زنجیره واجی اضافه شود تا دقت برنامه افزایش یابد. برای مثال در شکل (۳) خروجی برنامه به ازای عبارت "کتاب من کو؟" آورده شده که در آن کسره اضافه را نمی‌بینیم.

ketâb man ku

شکل (۳): زنجیره واجی عبارت "کتاب من کو؟"

#### ۴- نتیجه

در این مقاله، با استفاده از سامانه تحلیل‌گر صرفی کلمه در زبان فارسی با عنوان پارس مورف، که بر پایه یک مطالعه دقیق زبان‌شناختی از نظام صرفی زبان فارسی استوار است، سعی کردیم تا فرایند خودکار تبدیل متن فارسی به زنجیره واجی را توسعه دهیم. 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.

از چالش‌های پیش رو که در کارهای آتی باید به آنها پرداخته شود، در نظر گرفتن دشواری‌های موجود در پردازش متن فارسی [۱۰] است که باید راهکارهایی برای حل آن اندیشیده شود. از نتایج این کار می‌توان در زمینه انواع پردازش خودکار زبان فارسی و به طور مشخص در تبدیل متن به گفتار استفاده نمود.

#### مراجع

- [1] Allen, J., "Overview of Text-To-Speech Systems", Advances in Speech Signal Processing, Edited by S. Furui, 1991.
- [2] Torkkola, K., "An Efficient Way to Learn English Grapheme-to-Phoneme Rules Automatically", IEEE, pp. 199-202, 1993.
- [۳] صیادیان، ابوالقاسم و مجید نصیرزاده، "تجربه‌ای در مدل‌سازی زبان فارسی برای یک سیستم تبدیل متن به گفتار"، مجموعه مقالات دومین کنفرانس سالانه انجمن کامپیوتر ایران، ص ۱۰۵-۱۱۱، ۱۳۷۵.
- [4] Megerdooian, Karine., "Persian Computational Morphology, A Unification-Based Approach", NMSU, CRL, Memoranda in Computer and Cognitive Science, (MCCS-00-321), 2000.