تولید پیکره برچسب‌خورده واحدساز زبان فارسی با درنظرگرفتن ملاحظات زبان‌شناسی رایانشی آن

مژگان فرهودی*، مریم محمودی و مونا داودی شمسی
پژوهشگاه ارتباطات و فناوری اطلاعات، تهران، ایران

چکیده

متون نگاشته‌شده فارسی به‌طورمعمول دو مشکل ساده، ولی مهم دارند. مشکل نخست واژه‌های چندواحدی هستند که از اتصال یک واژه به واژه‌های بعدی حاصل می‌شوند. مشکل دیگر واحدهای چندواژه‌ای هستند که از جداشدگی واژه‌هایی که با هم یک واحد واژگانی را تشکیل می‌دهند، حاصل می‌شوند. ابزار واحدساز در زبان فارسی که به‌عنوان یکی از ابزارهای پیش‌پردازش زبان است، کاربرد فراوانی در تجزیه و تحلیل متون داشته و باید بتواند واحدهای واژگانی را تشخیص دهد. به عبارتی، این ابزار، مرکز کلمات را در متون تشخیص داده و آن را به دنباله‌ای از کلمات به‌منظور تحلیل‌های بعدی تبدیل می‌کند. تنوع در رسم‌الخط فارسی و عدم رعایت قوانین جدانویسی و پیوسته‌نویسی کلمات از یک‌سو و پیچیدگی‌های واژگانی زبان فارسی از سویی دیگر فرایندهای مختلف پردازشی زبان از جمله واحدسازی را با چالش‌های بسیاری روبه‌رو می‌کند؛ لذا برای عملکرد بهینه این ابزار، لازم است ابتدا ملاحظات زبان‌شناسی رایانشی واحدسازی در زبان فارسی مشخص و سپس بر اساس این ملاحظات مجموعه‌داده‌ای برای آموزش و آزمایش آن فراهم شد. در این مقاله سعی شد ضمن تبیین ملاحظات یادشده، به تهیه پیکره‌ای در این خصوص بپردازیم. پیکره تهیه‌شده شامل ۲۱/۱۸۳ کلمه و متوسط طول جملات ۴۰/۲۸ است.

واژگان کلیدی: پیکره واحدساز زبان فارسی، پردازش زبان فارسی، زبان‌شناسی رایانشی.

Producing a Persian Text Tokenizer Corpus Focusing on Its Computational Linguistics Considerations

Mojgan Farhoodi*, Maryam Mahmoudi & Mona Davoodi Shamsi
ICT Research Institute, Tehran, Iran

Abstract

The main task of the tokenization is to divide the sentences of the text into its constituent units and remove punctuation marks (dots, commas, etc.). Each unit is a continuous lexical or grammatical writing chain that is an independent semantic unit. Tokenization occurs at the word level and the extracted units can be used as input to other components such as stemmer. The requirement to create this tool is to identify and recognize the units that are known as independent semantic units in Persian language. This tool detects word boundaries in texts and converts the text into a sequence of words.

In the English language, many activities have been done in the field of text tokenization and many tools have been development; such as: Stanford, Ragel, ANTLR, JFLex, JLex, Flex and Quex. In recent decades, valuable researches have also been conducted in the field of tokenization in Persian language that all of them have worked on the lexical and syntactic layer. In the current research, we tried to focus on the semantic layer in addition to those two layers.

Persian texts usually have two simple but important problems. The first problem is multi-word tokens that result from connecting one word to the next. Another problem is polysyllabic units, which result from the separation of words that together form a lexical unit. Tokenizer is one of the language preprocessing tools that is widely used in text analysis. This component recognizes the center of words in texts and turns it into a sequence of words for later analysis. Variety in Persian script and non-observance of the rules of separation and spelling of words on the one hand and the lexical complexities

* نویسندهٔ عهده‌دار مکاتبات
* Corresponding author

• تاریخ ارسال مقاله: ۱۳۹۹/۵/۲۹ • تاریخ پذیرش: ۱۴۰۰/۱۰/۱۸ • تاریخ انتشار: ۱۴۰۱/۱۰/۳ • نوع مطالعه: پژوهشی

of Persian language on the other hand, different language processing such as tokenization face many challenges. Therefore, in order to obtain the optimal performance of this tool, it is necessary to first specify the computational linguistics considerations of tokenization in Persian and then, based on these considerations, provide a data set for training and testing. In this article, while explaining the mentioned considerations, we tried to prepare a data set in this regard. The prepared data set contains 21.183 tokens and the average length of sentences is 40.28.

Keywords: Persian text tokenization corpus, Natural Language Processing (NLP), cyber linguistic.

۱- مقدمه

وظیفهٔ اصلی واحدساز یا تک‌واژساز کلمه، تقطیع جملات متن به واحدهای تشکیل‌دهندهٔ آن و حذف علائم نگارشی (نقطه، ویرگول و غیره) است. هر واحد، زنجیرهٔ نوشتاری واژگانی یا دستوری پیوسته‌ای است که واحد معنایی مستقل است. واحدسازی در سطح کلمات رخ می‌دهد و واحدهای استخراج‌شده می‌توانند به‌عنوان ورودی مؤلفه‌های دیگر مانند ریشه‌یاب و برچسب‌گذار استفاده شوند. لازمه ایجاد این ابزار شناسایی و تشخیص واحدهایی است که در زبان فارسی به‌عنوان واحدهای مستقل معنایی شناخته می‌شوند. این ابزار مرز کلمات را در متون تشخیص داده و متن را به دنباله‌ای از کلمات تبدیل می‌کند.

امکان استفاده از نیم‌فاصله در کنار فاصلهٔ یک‌حرفی پدیدآورندهٔ اصلی‌ترین مشکلات در نوشتار فارسی است. تنوع در رسم‌الخط فارسی و عدم رعایت قوانین جدانویسی و پیوسته‌نویسی کلمات از یک‌سو و پیچیدگی‌های واژگانی زبان فارسی از سویی دیگر فرایندهای مختلف پردازشی زبان از جمله واحدسازی را با چالش‌های بسیاری روبه‌رو می‌کند. از جمله راه‌کارهای رویارویی با این چالش‌ها سطح‌بندی واحدسازی است؛ به این صورت که واحدسازی در دو سطح صورت پذیرد. در سطح نخست، واحدساز پیشین با توجه به ویژگی‌های نوشتاری متن (مانند فاصلهٔ یک‌حرفی) به تعیین مرز واحدها می‌پردازد که البته در این مرحله ممکن است اشتباهاتی رخ دهد؛ برای مثال، واژه‌های چندبخشی به واحدهای مجزا تقسیم شوند یا واژه‌های مجزا به‌صورت یک واحد در نظر گرفته شوند. برطرف‌کردن این کاستی‌ها بر عهدهٔ واحدسازی پسین است؛ بنابراین، خروجی فرایند واحدسازی پیشین در سطح نخست، درونداد واحدسازی پسین در سطح دوم محسوب می‌شود. در ادامه مقاله ابتدا در بخش۲ به مروری بر کارهای انجام‌شده در زبان انگلیسی و فارسی می‌پردازیم، سپس در بخش۳، شیوهٔ واحدسازی عناصر مختلف زبان فارسی در سطح ۱ و ۲ تشریح شده و به چالش‌های مهمی که انتظار می‌رود ابزارهای واحدساز پیشین و پسین به شیوه‌ای صحیح از پس آنها برآیند، اشاره می‌شود. در انتهای مقاله و در بخش۴ نیز به معرفی پیکره تهیه‌شده براساس اصول ذکر شده در مقاله برای واحدسازی زبان فارسی می‌پردازیم.

۲- پژوهش‌های پیشین

در زبان انگلیسی فعالیت‌های زیادی در زمینه توسعه ابزار واحدساز انجام شده است. از نمونه‌های انگلیسی این ابزار می‌توان به استنفورد، Ragel، ANTLR، JFLex، JLex، Flex و Quex اشاره کرد. چنان‌که می‌دانیم بخش قابل‌توجهی از مطالعات در حوزهٔ واحدسازی بر روی زبان انگلیسی و حل ابهامات واحدسازی در این زبان صورت پذیرفته است؛ مانند [10-12,16] که همگی به مسألهٔ تنوعات روش‌های واحدسازی در مواجهه با چالش‌های زبان اذعان می‌کنند و هر یک روشی نوین پیش می‌نهد.

در دهه‌های اخیر پژوهش‌های ارزنده‌ای نیز در حوزهٔ واحدسازی زبان فارسی انجام شده است که می‌توان آنها را به دو دسته تقسیم‌بندی کرد:

- توسعه ابزار واحدسازی متون فارسی؛ در این خصوص فعالیت‌های به‌نسبه خوب، ولی پراکنده‌ای در کشور انجام شده است و نمونه‌های متعددی از آنها در قالب سرویس‌های وب در دسترس هستند. در این خصوص می‌توان به ابزارهایی از قبیل هضم[1]، استپ‌وان[2]، ستپر[3] و آیپا[4]، پارسی‌پرداز[5] و ... اشاره کرد. تمامی این ابزارها از دادگان آموزشی خود استفاده می‌کنند که متأسفانه در دسترس نیست. اما با بررسی عملکرد این ابزارها می‌توان تا حد زیادی به دادگان آنها و نحوه برچسب‌زنی آنها پی برد. بررسی‌های انجام‌شده حاکی از آن است که ابزارهای موجود برای واحدسازی به مقوله لغوی و یا

[1] https://www.sobhe.ir/hazm
[2] http://nlp.sbu.ac.ir/Projects.aspx#section1
[3] https://www.peykaregan.ir/dataset
[4] https://aipaa.ir/demo/text-analysis?tag=syntax
[5] http://nlp.sbu.ac.ir/Projects.aspx#section1

نحوی کلمات توجه می‌کنند و براساس آن واحدسازی را انجام می‌دهند ولی در دادگان تهیه شده در پژوهش جاری، سعی شده است تا کلمات با توجه به قرار گرفتن آنها در بافت جمله تقطیع شوند. به‌عنوان مثال در جمله «پویا پسر با دقت و منظمی است»، «با دقت» یک واحد محسوب می‌شود، ولی در جمله «علی با دقت فراوان نامه را خواند»، «با دقت» دو واحد محسوب می‌شود؛ درصورتی‌که در ابزارهای موجود در هر دو جمله عبارت «با دقت» را دو واحد درنظر می‌گیرند؛ مگرآن‌که دوکلمه بدون فاصله نوشته شده باشد.

- ایجاد و توسعه دادگان؛ در این راستا می‌توان به پروژه‌ای موسوم به «پروژهٔ شیراز» اشاره داشت [17, 18]. هدف نهایی این پروژهٔ تحقیقاتی دست‌یابی به ماشین ترجمهٔ فارسی-انگلیسی بوده و دستوری با عنوان «سمبا» ارائه شده است که مسئول پیاده‌سازی یک تحلیل‌گر تک‌واژی است. در پژوهش نام‌برده واحدسازی متون در دو مرحلهٔ مستقل اما مرتبط صورت می‌پذیرد. به این گونه که واژه‌های چندبخشی در ابتدا به‌صورت یک‌دست جداسازی شده و سپس یک پساواحدساز که مجهز به اطلاعات زبان ویژه است، عناصر جداشده را به یکدیگر متصل می‌کند. پژوهش مهم دیگری که باید به آن اشاره شود، [5] است که به واحدسازی واژه‌های چندجزئیِ ایستا و پویا می‌پردازد. برای دستهٔ نخست از فهرست واژه و روش انطباق و برای دستهٔ دوم از الگوهای قاعده‌مند بهره می‌جوید. در پژوهشی مرتبط، [23] به چالش‌های پیش روی تعیین مرز واحدهای چندواژه‌ای و واژه‌های چندواحدی در زبان فارسی می‌پردازند. واحدهای چندواژه‌ای جدا نوشته می‌شوند اما باید یک واحد درنظر گرفته شوند. در مقابل، واژه‌های چندواحدی به‌اشتباه به‌صورت پیوسته نوشته می‌شوند؛ اما چند واژهٔ مستقل‌اند. گفتنی است در این پژوهش به این دو دسته واژهٔ مشکل‌ساز پرداخته خواهد شد.

پژوهش‌های دیگری نیز بر روی زبان فارسی صورت پذیرفته است که به‌طور مستقیم یا غیرمستقیم به حوزهٔ واحدسازی کلمات و جملات مرتبط‌اند. در زمینهٔ تحلیل ساخت‌واژی و نحوی متن و تهیهٔ واژگان زبان فارسی با الهام از وردنت‌های زبان انگلیسی می‌توان به [4,9,15,19-22] در میان دیگر پژوهش‌گران اشاره داشت.

با بررسی فعالیت‌های انجام‌شده در این خصوص باید گفت که در پژوهش جاری تمام تلاش این بوده است، علی‌رغم نمونه‌های موجود که تنها بر روی لایه لغوی و نحوی تمرکز کرده‌اند، لایه معنا[1] نیز در تقطیع کلمات مدنظر قرار گیرد که در ادامه مقاله با ذکر مثال به شرح آن پرداخته شده است.

۳- واحدسازی

۳-۱- واحدسازی واژه‌های ساده

واژه‌های ساده (یا بسیط) از یک تکواژ تشکیل می‌شوند (مانند «ایران») و واژه‌های غیرساده (یا غیربسیط) از بیش از یک تک‌واژ تشکیل می‌شوند (مانند «ایران‌خودرو»). واژه‌های غیرساده به دو دستهٔ مرکب و نیمه‌مرکب قابل‌تقسیم‌اند و هر دسته قواعد واحدسازی خاص خود را دارند که در بخش‌های بعدی به این قواعد اشاره می‌شود. واژه‌های ساده به‌وسیلهٔ واحدسازهای پیشین و پسین به یک واحد مجزا تقطیع می‌شوند و به‌طورمعمول مشکلی برای فرایند واحدسازی پیش نمی‌آورند. این واحدها به مقوله‌های دستوری مختلف مانند اسم، صفت، قید، حرف اضافه، فعل و غیره تعلق دارند. یک نمونه از واحدسازی واژه‌های ساده در زیر آمده است:

(۱) «تعیین و تقطیع هر واحد متن نخستین مرحله در پردازش متن محسوب می‌شود.»

واحدسازی سطح ۱ و ۲: «تعیین»، «و» «تقطیع»، «هر»، «واحد»، «نخستین»، «مرحله»، «در»، «پردازش»، «متن»، «محسوب»، «می‌شود»

جدول زیر خلاصه‌ای از واحدسازی واژه‌های ساده است:

(جدول-۱): واحدسازی واژه‌های ساده
(Table-1): Tokenization of simple words

نوع سازه	سطح ۱	سطح ۲
ساده	یک واحد	یک واحد

۳-۲- واحدسازی انواع سازه‌های غیرساده: واژه‌های مرکب، نیمه‌مرکب و گروه‌های نحوی

سازه‌های غیرساده از کنار هم قرارگرفتن عناصر زبانی مختلف مانند واژه‌ها، وندها و گروه‌ها پدید می‌آیند. گاهی اوقات از کنار هم قرارگرفتن واژه‌ها یا وندها، واژه‌ای جدید

[1] semantic

پدید می‌آید که می‌تواند از نوع اسم، صفت، قید، حرف‌اضافه، حرف ربط یا فعل مرکب باشد. گاهی نیز از کنار هم قرارگرفتن واژه‌های ساده یا مرکب یگ گروه نحوی تشکیل می‌شود. این سازه‌ها بر مبنای معیارهای نحوی و معنایی در طبقات زیر جای می‌گیرند:

الف- واژه‌های مرکب

ب- واژه‌های نیمه‌مرکب

ج- گروه‌های نحوی

از مهم‌ترین چالش‌های پیش روی فرایند واحدسازی می‌توان به تقطیع کلمات غیرساده مرکب، نیمه‌مرکب و گروه‌های نحوی و بازشناسی آنها از یکدیگر اشاره داشت؛ بنابراین، در بخش‌های بعدی ابتدا به معیارهای بازشناسی واژه‌های مرکب از گروه‌های نحوی و سپس به بازشناسی واژه‌های مرکب و نیمه‌مرکب می‌پردازیم. در این راستا از تحلیل‌های ارائه‌شده در پژوهش‌های مطرح زبان‌شناسی بهره جسته‌ایم؛ از جمله: [1,2,7,13,14,].

۳-۲-۱- بازشناسی واژه‌های مرکب از گروه‌های نحوی

واژه‌های مرکب از بیش از یک تک‌واژ– اعم از آزاد یا وابسته (وندها)- تشکیل می‌شوند و در فرایند واحدسازی همچون واژه‌های ساده به‌عنوان یک واحد در نظر گرفته می‌شوند. چنانچه سازه از نوع گروه نحوی باشد، عناصر به‌کاررفته در آن باید به واحدهای مجزا تقسیم شوند. گاه واحدهای مرکب در نوشتار فارسی با فاصلهٔ یک‌حرفی از هم جدا شده و مشکلاتی را پیش روی فرایند واحدسازی قرار می‌دهند. با استفاده از معیارهای زیر می‌توان کلمات مرکب را از گروه‌های نحوی بازشناخت.

الف- انسجام واژگانی: کلمات مرکب از انسجام برخوردارند و فرایندهای نحوی یا صرفی به ساختار درونی آنها دسترسی ندارند. حال آنکه این قواعد بر عناصر سازندهٔ گروه‌های نحوی به‌راحتی اعمال می‌شوند. به‌طور مثال، نشانه‌های جمع یا تفضیلی و عالی و انواع وابسته‌ها نمی‌توانند میان عناصر سازندهٔ کلمهٔ مرکب واقع شوند و تنها در پایان این کلمات به‌کار می‌روند (مانند «*کتابهاخانه» و «*خوشتررو» در مقایسه با «کتابخانه‌ها» و «خوش‌روترین») یا («کتابخانه» در مقایسه با «*کتاب‌این‌خانه») اما همین عناصر آزادانه میان کلمات یک گروه نحوی قرار می‌گیرند («کتابهای این خانه»، «خوش‌ترین رو از آن اوست»، «کتابهای این خانه خواندنی‌اند»).

ب- ترکیب‌پذیری معنایی: معنای تمامی گروه‌های نحوی همواره ترکیب‌پذیر است، بدین معنا که معنای گروه برآیندی است از معنای تک‌تک عناصر حاضر در گروه. بار اصلی معنای گروه نیز بر دوش هستهٔ معنایی آن است. برای مثال، در گروه «گل‌های زیبای این خانه» از رابطهٔ هسته («گلها») و توصیف‌گرها به معنای کل گروه پی می‌بریم. این درحالی است که معنای بسیاری از کلمات مرکب، از اجزای سازندهٔ آنها، مانند ترکیب‌های «بهارنارنج» یا «خارپشت». قابل‌پیش‌بینی نیست.

ج- مقولهٔ نحوی: گروه‌های نحوی همواره واحد هسته‌ای نحوی هستند که تعیین‌کنندهٔ مقولهٔ نحوی کل گروه است. برای مثال، مقولهٔ نحوی ساخت‌هایی چون «زیباتر از همیشه» و «منظرهٔ بسیار زیبا» همان مقولهٔ نحوی هسته‌های گروه‌ها یعنی «زیباتر» و «منظره» است؛ ساخت نخست، گروه صفتی و ساخت دیگر گروه اسمی است. این مسأله در رابطه با بسیاری از کلمات مرکب صادق نیست و مقولهٔ نحوی آنها از عناصر سازنده‌شان قابل‌پیش‌بینی نیست مانند ترکیب‌های صفتی «خداترس» و «انسان‌ستیز».

د- همپایگی: همپایگی دو گروه اسمی در فارسی امکان‌پذیر است («کتابخانه و گلخانه») اما بخش‌های مختلف واژه‌های ساده یا مرکب را نمی‌توان با یکدیگر هم‌پایه کرد («*کتاب و گلخانه»).

ه- جابه‌جایی: عناصر حاضر در کلمات مرکب قابلیت جابه‌جایی ندارند و چنانچه دو عنصر جابه‌جا شوند یا ساختی نادستوری حاصل می‌آید (مانند «خوش‌آب‌وهوا» در مقایسه با «*خوش‌هواوآب») یا معنا دگرگون می‌شود (مانند «تلویزیون سیاه‌وسفید» در مقایسه با گروه نحوی «تلویزیون سفید و سیاه»).

۳-۲-۲- بازشناسی واژه‌های مرکب از نیمه‌مرکب

در این مقاله، عناوین مرکب و نیمه‌مرکب به تمامی واژه‌های غیربسیطی اطلاق می‌شود که از بیش از یک تکواژ تشکیل شده باشند، خواه تکواژها آزاد باشند یا وابسته. معیار اصلی بازشناسی کلمات مرکب از نیمه‌مرکب انسجام درونی عناصر آنها است. عناصر حاضر در کلمه‌های مرکب جداناپذیرند و روی‌هم‌رفته یک واحد محسوب می‌شوند؛ اما کلمات نیمه‌مرکب رفتاری بینابین از خود

نشان می‌دهند. به این صورت که از برخی جهات شبیه کلمات مرکب و از جهاتی دیگر مانند گروه‌های نحوی عمل می‌کنند. طبقات مختلف واژه‌ها از جمله اسم‌ها، صفت‌ها، حروف اضافه و ربط و فعل‌ها می‌توانند نیمه‌مرکب باشند. در ادامه، به‌ترتیب برای هریک از طبقات دستوری کلمات نیمه‌مرکب مثالی می‌آوریم.

الف- اسم/صفت

اسمی چون «مهدکودک» طبق بیشتر معیارهای بخش پیشین، مرکب محسوب می‌شود؛ اما وند تصریفی جمع می‌تواند انسجام درونی واژه را بر هم زند. این وند از یک سو مانند واژه‌های مرکب به پایان واژه متصل («مهدکودک‌ها») و از سویی دیگر مانند گروه‌های نحوی میان عناصر حاضر می‌شود («مهدهای کودک»). چنین زنجیره‌ایی را نیمه‌مرکب می‌خوانیم و در سطح یک واحدسازی عناصر سازندهٔ آن را واحدهای مجزا در نظر می‌گیریم؛ اما به‌سبب ویژگی‌های ترکیبی آن در سطح دو آن را یک واحد قلمداد می‌کنیم:

(۲) مهدکودک

به‌عنوان مثال‌هایی دیگر می‌توان اسم‌ها و صفت‌های نیمه‌مرکبی چون «آذربایجان شرقی» یا «دوزبانه» را در نظر گرفت. این قبیل واژه‌ها از بیشتر آزمون‌های کلمات مرکب سربلند بیرون می‌آیند؛ اما عناصر سازنده‌شان را به‌صورت محدود می‌توان همپایه کرد (مانند «آذربایجان شرقی و غربی» و «دو یا سه‌زبانه»). این قبیل واژه‌های نیمه‌مرکب نیز در سطح یک واحدسازی دو واحد و در سطح ۲ یک واحد محسوب می‌شوند.

نکته: عناصر حاضر در اسم‌ها و صفت‌های نیمه‌ترکیبی در حالت معطوف در هر دو سطح واحدسازی به واحدهای مجزا تقسیم می‌شوند:

(۳) آذربایجان شرقی و غربی

واحدسازی سطح ۱: «آذربایجان»، «شرقی»، «و»، «غربی»
واحدسازی سطح ۲: «آذربایجان»، «شرقی»، «و»، «غربی»

ب- حرف اضافه

زنجیره‌های شامل «حرف‌اضافه+اسم» همچون «درمورد»، «براساس»، «باوجود»، «دربارهٔ» را با گروه‌هایی چون «بر بام»، «در خانهٔ»، «بر فرش» مقایسه کنیم. حروف اضافه در دستهٔ نخست حرف‌اضافه نیمه‌مرکب‌اند و در سطح دو واحدسازی یک واحد در نظر گرفته می‌شوند. در مقابل، دستهٔ دوم گروه حرف‌اضافه‌ای هستند و عناصر حاضر در آن در هر دو سطح دو واحد مجزا محسوب می‌شوند:

(۴) براساس

واحدسازی سطح ۱: «بر» «اساس»
واحدسازی سطح ۲: «براساس»

(۵) بر بام

واحدسازی سطح ۱: «بر» «بام»
واحدسازی سطح ۲: «بر» «بام»

با استفاده از آزمون گسترش‌پذیری اسم، استدلال‌هایی برای بازشناسی این دو طبقه ارائه داده است. به‌طور کلی، انواع مختلف وابسته‌های پیشین یا پسین قادر به توصیف اسم در گروه حرف‌اضافه‌ای هستند (مثال ۶) اما در حروف‌اضافهٔ نیمه‌مرکب این امکان وجود ندارد (مثال‌های ۷-الف و ب). نکتهٔ شایان ذکر آن‌که حروف اضافهٔ نیمه‌مرکب تنها در برخی بافت‌های محدود وابستهٔ صفت اشاره می‌پذیرند (مثال ۷-ج).

(۶) الف. بر زیباترین بام خانه پا گذاشتم.

ب. بر بامهای خانه پا گذاشتم.

ج. بر این بام (خانه) پا گذاشتم.

(۷) الف. *بر جدیدترین اساس مطالعات، هوای زمین

واحدسازی سطح ۱: «مهد»، «کودک»
واحدسازی سطح ۲: «مهدکودک»

گرم‌تر شده است.

ب. *بر اساس‌های مطالعات دانشمندان، هوای زمین گرم‌تر شده است.

ج. بر این اساس (*مطلب)، هوای زمین گرم‌تر شده است.

نکته: چنانچه حروف اضافهٔ نیمه‌ترکیبی وابسته بپذیرند، عناصر سازندهٔ آنها در هر دو سطح واحدسازی به واحدهای مجزا تقسیم می‌شوند:

(۸) بر این اساس

واحدسازی سطح ۱: «بر»، «این»، «اساس»
واحدسازی سطح ۲: «بر»، «این»، «اساس»

ج- فعل

از عمده‌ترین چالش‌های مرتبط با افعال در زبان فارسی، بازشناسی افعال موسوم به مرکب از گروه‌های فعلی است. به این تعبیر که کاربرد هر عنصر غیرفعلی در کنار فعل را نباید فعل مرکب قلمداد کرد؛ گاه سروکار ما با گروه فعلی است. در اینجا مراد از گروه فعلی گروهی است که متشکل از یک موضوع و فعلی ساده باشد. برای مثال، به‌رغم شباهت ظاهری دو زنجیرهٔ «بستنی خوردن» و «ضربه خوردن» ساخت نخست، گروه فعلی متشکل از موضوع فعل (مفعول) و فعل ساده است؛ اما ساخت دوم، فعلی

مرکب است که در واژگان ذهنی گویشوران و فرهنگ‌های لغت فهرست می‌شود. پیداست که باید فعل مرکب را از گروه فعلی بازشناخت تا واحدسازی به شیوه‌ای صحیح انجام شود. عناصر حاضر در فعل مرکب، روی‌هم‌رفته یک واحد و عناصر سازندهٔ گروه‌های فعلی، واحدهای مستقل قلمداد می‌شوند. برمبنای معیارهایی که در بخش ۳-۱ آمد، افعال مرکب از بسیاری از جهات مرکب محسوب می‌شوند؛ معنایشان ترکیب‌پذیر نیست، عنصر پیش‌فعلی و فعل سبک روی‌هم‌رفته یک واحد معنایی محسوب می‌شوند و قیدها نمی‌توانند درون بسیاری از آنها واقع شوند («کاملاً گوش کرد» در مقایسه با «*گوش کاملاً کرد»)؛ بنابراین، شاید بتوان گفت افعال مرکب نه یک گروه فعلی که یک واحد فعلی‌اند؛ اما رفتار این سازه‌ها از جهاتی شبیه گروه‌های فعلی است. به‌طور مثال، واحدهای سازندهٔ این افعال به‌طورمعمول می‌توانند از یکدیگر فاصله گیرند و عناصری مانند افعال وجهی یا وندهای نفی میان عناصر پیش‌فعلی و افعال سبک واقع می‌شوند (مانند «پایان خواهم داد» یا «پایان نخواهد داد»)؛ بنابراین، این عناصر را نیمه‌مرکب در نظر می‌گیریم و در سطوح یک و دو آنها را به‌ترتیب به دو و یک واحد تقطیع می‌کنیم:

(۹) پایان داد

واحدسازی سطح ۱: «پایان»، «داد»
واحدسازی سطح ۲: «پایان داد»

نکته: چنانچه عنصری میان سازه‌های فعل مرکب قرار گیرد، عناصر سازندهٔ آن در هر دو سطح به واحدهای مجزا تقسیم می‌شوند:

(۱۰) پایان خواهم داد

واحدسازی سطح ۱: «پایان»، «خواهم»، «داد»
واحدسازی سطح ۲: «پایان»، «خواهم»، «داد»

جدول (۲) خلاصه‌ای از مطالب مطرح شده در این بخش است:

(جدول-۲): واحدسازی انواع سازه‌ها
(Table-2): Tokenization of various structures

نوع سازه	سطح ۱	سطح ۲
مرکب (مانند «کتابخانه»)	۱ واحد	۱ واحد
نیمه‌مرکب (مانند «مهدکودک»، «دوزبانه»، «براساس»، «پایان داد»)	۲ واحد	۱ واحد
گروه (مانند «خانهٔ کتاب»، «بر بام»، «غذا خورد»)	۲ واحد	۲ واحد

۳-۳- فرایندهای ساختِ واژه‌های غیرساده

فرایندهای ترکیب، اشتقاق و تصریف در ساخت واژه‌های غیرساده (مرکب و نیمه‌مرکب) دخیل‌اند. در ادامه فرایندهای نامبرده با مثال توضیح داده می‌شود.

۳-۳-۱- ترکیب

فرایند ترکیب با کنار هم گذاردن دو یا چند تکواژ آزاد واژگانی یا دستوری کلمهٔ مرکب یا نیمه‌مرکب با انواع مقوله‌های دستوری می‌سازد. در این ارتباط، به قواعد زیر که تنها چند نمونه از قواعد ترکیب است، می‌توان اشاره کرد. واژه‌های ساخته‌شده به‌وسیلهٔ فرایند ترکیب از نوع مرکب یا نیمه‌مرکب‌اند. قواعد ترکیب در دستهٔ (الف) واژه‌های مرکب می‌سازند، به این معنا که عناصر تشکیل‌دهندهٔ کلمات به هیچ روی از یکدیگر جدا نشده و مشمول فرایند عطف نمی‌شوند. در مقابل، عناصر سازندهٔ کلمات در دستهٔ (ب) در شرایط خاص جدا می‌شوند یا مشمول فرایند عطف قرار می‌گیرند.

الف- واژه‌های مرکب

صفت + اسم ← صفت مرکب (مانند «کم‌آب»، «بدنام»، «لاغراندام»)

اسم/ضمیر + اسم ← صفت مرکب (مانند «دل‌رحم»، «راه‌راه»، «خودرأی»)

اسم + بن فعل ← اسم/صفت مرکب (مانند «آب‌پاش»، «گوشت‌کوب»، «مردم‌فریب»، «خداشناس»)

صفت/اسم + اسم ← اسم مرکب (مانند «پیرمرد»، «بزرگراه»، «هنرپیشه»، «پسرخاله»)

صفت اشاره/حرف ربط + ضمیر موصولی ← کلمهٔ مرکب (مانند «همین‌که»، «اینکه»، «چراکه»، «چونکه»)

عدد + اسم ← اسم مرکب (مانند «دوشنبه»، «پنج‌شنبه»)

اسم + حرف عطف + اسم ← اسم/صفت مرکب پیوندی (مانند «سیاه‌وسفید»، «سروصدا»، «صداوسیما»)

ب- واژه‌های نیمه‌مرکب

اسم + صفت ← اسم نیمه‌مرکب (مانند «کرهٔ شمالی»، «آذربایجان شرقی»)

عدد + اسم ← صفت نیمه‌مرکب (مانند «دوزبانه»، «دوطبقه»، «دوساعته»، «دومتری»)

اسم + فعل سبک ← فعل نیمه‌مرکب (مانند «کتک زدن»، «دوست داشتن»، «لطمه زدن»)

صفت + فعل سبک ← فعل نیمه‌مرکب (مانند «دراز کشیدن»، «سبک کردن»، «پهن کردن»)
گروه حرف اضافه‌ای + فعل سبک ← فعل نیمه‌مرکب (مانند «از یاد بردن»، «به باد دادن»)
قید + فعل سبک ← فعل نیمه‌مرکب (مانند «بالا بردن»، «بیرون کردن»)

نکته: واژه‌های به‌دست‌آمده از فرایند ترکیب چنانچه مرکب باشند، در هر دو سطح واحدسازی یک واحد در نظر گرفته می‌شوند و چنانچه نیمه‌مرکب باشند در سطح ۱ دو واحد و در سطح ۲ یک واحد قلمداد می‌شوند. چنانچه عنصری میان سازه‌های واژه‌های نیمه‌مرکب قرار گیرد، عناصر سازندهٔ آن در هر دو سطح به واحدهای مجزا تقسیم می‌شوند (مثال ۱۱):

(۱۱) دوشنبه

واحدسازی سطح ۱: «دوشنبه»
واحدسازی سطح ۲: «دوشنبه»

(۱۲) الف- دوطبقه

واحدسازی سطح ۱: «دو»، «طبقه»
واحدسازی سطح ۲: «دوطبقه»

(۱۳) دو یا سه طبقه

واحدسازی سطح ۱: «دو»، «یا»، «سه»، «طبقه»
واحدسازی سطح ۲: «دو»، «یا»، «سه»، «طبقه»

نکته: «واو» عطف در برخی از کلمات مرکب پیوندی به واکه تبدیل می‌شود، مانند «زناشویی»، «کمابیش» و «تکاپو». در برخی از دیگر ترکیب‌ها «واو» عطف حذف می‌شود مانند «بسازبفروش» و «بشوربپوش» و غیره. در تمامی این موارد نیز با واژهٔ مرکب سروکار داریم که در هر دو سطح واحدسازی یک واحد قلمداد می‌شوند.

۳-۳-۲- اشتقاق

فرایند اشتقاق به‌طورمعمول با افزودن یک پیشوند یا پسوند اشتقاقی به تکواژی آزاد، کلمهٔ مرکب یا نیمه‌مرکب با انواع مقوله‌های دستوری می‌سازد. وندهای اشتقاقی واجد معنااند و با اضافه‌شدن به کلمه، واژه‌ایی جدید می‌سازند که به‌طورمعمول به‌طور جداگانه در فرهنگ‌های لغت فهرست می‌شود. این وندها اغلب مقولهٔ نحوی کلمه را تغییر داده (مانند «شاد»← «شادی»، «گفت»← «گفتار») یا معنای کلمه را دستخوش تغییر می‌کنند (مانند «روا»← «ناروا»). در زیر به چند نمونه از فرایندهای واژه‌سازی اشتقاقی اشاره می‌شود:

الف- واژه‌های مرکب

پیشوند اشتقاقی+ صفت/اسم ← صفت مشتق (مانند «ناراضی»، «باهوش»، «بی‌رحم»، «به‌روز»)
پیشوند اشتقاقی + اسم/صفت + پسوند اشتقاقی ← قید مشتق (مانند «ناآگاهانه»، «ناامیدانه»)
پیشوند اشتقاقی + بن فعل ← اسم/صفت مشتق (مانند «پیشامد»، «نشکن»، «هم‌نشین»)
بن فعل + پسوند اشتقاقی ← اسم/صفت مشتق (مانند «ریزش»، «خندان»، «کوشا»)
اسم + پسوند ← اسم/صفت (مانند «شهرک»، «مردانه»، «تعمیرگاه»)
صفت + پسوند ← اسم/صفت مشتق (مانند «سفیدک»، «گرما»، «آگاهانه»)
اسم + پسوند ← حرف اضافهٔ مرکب (مانند «بنابر»، «علاوه‌بر»، «بنابر»)
پیشوند + اسم ← قید مرکب (مانند «به‌استثنا»، «به‌تدریج»، «به‌مرور»)

ب- واژه‌های نیمه‌مرکب

پیشوند + بن فعل ← فعل مشتق (مانند «درگذر»، «بازآ»، «برگرد»)
پیشوند + اسم ← حرف اضافهٔ نیمه‌مرکب (مانند «برمبنای»، «براساس»، «به سختی»)
حرف اضافهٔ مکان/زمان + پسوند ← قید مرکب (مانند «پشت به»، «پس از»)

نکته: واژه‌های به‌دست‌آمده از فرایند اشتقاق چنانچه مرکب باشند، در هر دو سطح واحدسازی یک واحد در نظر گرفته می‌شوند و چنانچه نیمه‌مرکب باشند، در سطح ۱ دو واحد و در سطح ۲ یک واحد قلمداد می‌شوند. چنانچه عنصری میان سازه‌های واژه‌های نیمه‌مرکب قرار گیرد، عناصر سازندهٔ آن در هر دو سطح به واحدهای مجزا تقسیم می‌شوند:

(۱۴) به‌استثنا

واحدسازی سطح ۱ و ۲: « به‌استثنا»

(۱۵) پس از

واحدسازی سطح ۱: «پس»، «از»
واحدسازی سطح ۲: «پس‌از»

(۱۶) پیش یا پس از

واحدسازی سطح ۱: «پیش»، «یا»، «پس»، «از»
واحدسازی سطح ۲: «پیش»، «یا»، «پس»، «از»

نکته: گاه هر دو فرایند اشتقاق و ترکیب در ساخت یک واحد مرکب دخیل‌اند. واژه‌هایی چون «غیرقابل‌قبول»، «خودداری»، «تلافی‌جویانه» مشتق‌مرکب محسوب می‌شوند و در هر دو سطح واحدسازی یک واحد در نظر گرفته می‌شوند.

(۱۷) غیرقابل‌قبول

واحدسازی سطح ۱: «غیرقابل‌قبول»
واحدسازی سطح ۲: «غیرقابل‌قبول»

۳-۳-۳- تصریف

فرایند تصریف به‌طورمعمول با افزودن یک پیشوند یا پسوند تصریفی به تکواژی آزاد، کلمهٔ مرکب می‌سازد. وندهای تصریفی فقط معنای دستوری دارند و تغییری در مقولهٔ نحوی کلمه یا معنای واژگانی آن ایجاد نمی‌کنند. این وندها به‌طورمعمول پس از وندهای اشتقاقی قرار می‌گیرند. نشانهٔ جمع (مانند «- ها»، «- ان»، «- ات» و غیره) و نشانه‌های تفضیلی و عالی (مانند «- تر» و «- ترین») از جمله وندهای تصریفی اسم‌ها و صفت‌ها هستند. افعال نیز به لحاظ زمان، شخص، شمار و نمود صرف شده و صورت‌های مختلفی می‌پذیرند. وندهای تصریفی فعلی شامل نشانه‌های زمان (مانند «رفتم»، «بروم»)، شخص (مانند «می‌روم» «می‌روی»)، شمار (مانند «رفتم»، «رفتیم»)، نفی (مانند «نرفتم») و نمود می‌شوند. نمود دستوری می‌تواند با حضور فعل کمکی (نوشته است)، یا پیشوند (می‌نوشت) مشخص شود.

قواعد زیر نمونه‌هایی از قواعد تصریف به‌شمار می‌روند:

اسم/صفت + پسوند تصریفی ⟵ اسم (مانند «رساله‌ها»، «دانشمندان»، «اشتباهات»)

صفت/قید + پسوند تصریفی ⟵ صفت/قید (مانند «باهوش‌تر/ترین»، «دلیرانه‌تر/ترین»

نشانهٔ امر/حال التزامی + بن فعل ⟵ فعل صرف‌شده (مانند «بخوان»، «بخوانم»)

نشانهٔ نمود ناقص + فعل ⟵ فعل صرف‌شده (مانند «می‌خوری»، «می‌خورد»)

صفت مفعولی + نشانهٔ نمود کامل ⟵ فعل صرف‌شده (مانند «نوشتی»، «نوشته‌اند»، «نوشته‌است»)

بن فعل + شناسه‌های فعلی ⟵ فعل صرف‌شده (مانند «سرودم»، «سرودی»، «سرود»)

نشانه‌های نفی + فعل ⟵ فعل صرف‌شده (مانند «نخروشید»، «میازار»)

فعل + پی‌بست ⟵ گروه واژه‌بست (مانند «خواندمش»، «خوردی‌اش»)

نکته: برخلاف فرایندهای ترکیب و اشتقاق، فرایندهای تصریفی کلمهٔ نیمه‌مرکب نمی‌سازند؛ بنابراین وندهای تصریفی به‌همراه پایه‌های واژگانی‌شان در هر دو سطح از واحدسازی یک واحد در نظر گرفته می‌شوند:

(۱۸) نوشته است

واحدسازی سطح ۱: «نوشته‌است»
واحدسازی سطح ۲: «نوشته‌است»

در این مقاله، واحد مرکب به تمامی واژه‌های غیربسیط اعم از مرکب، مشتق، مشتق‌مرکب و صرف‌شده اطلاق می‌شود که از بیش از یک تکواژ تشکیل شده باشد؛ مانند انواع واژه‌های «ایران‌خودرو»، «ایرانی»، «ایرانی‌ها».

۳-۴- تأثیر بافت بر فرایند واحدسازی

گاه برخی عناصر را نمی‌توان به‌صورت منفرد واحدسازی کرد. به این معنا که واحدسازی عناصر وابسته به بافت زبانی است که آن عنصر در آن به‌کار رفته است. به این تعبیر که زنجیره‌ای از واحدهای زبانی در یک بافت خاص روی‌هم‌رفته یک واحد مرکب می‌سازند و در بافتی دیگر تشکیل‌دهندهٔ گروه نحوی هستند و به واحدهای مجزا تقسیم می‌شوند.

۳-۴-۱- بازشناسی واژه‌های مشتق از گروه‌های نحوی

زنجیره‌هایی چون «با دقت»، «با ادب»، «بی کار» و امثال آنها می‌توانند متعلق به سطح واژه (مثال ۱۹) یا گروه نحوی (مثال ۲۰) باشند؛ اما این مسأله تنها با توجه به بافت مشخص می‌شود:

(۱۹) پویا پسر **بادقتی** است.

واحدسازی سطح ۱: «بادقت»
واحدسازی سطح ۲: «بادقت»

(۲۰) علی **با دقت** فراوان نامه را خواند.

واحدسازی سطح ۱: «با»، «دقت»
واحدسازی سطح ۲: «با»، «دقت»

۳-۴-۲- بازشناسی طبقات مختلف افعال از گروه‌های فعلی

در فرایندهای واحدسازی، به‌طورمعمول بیشتر صورت‌های تصریفی فعل به‌درستی یک واحد در نظر گرفته می‌شوند. آنچه این فرایند را با چالش روبه‌رو می‌کند تفاوت‌های

افعال ربطی، کمکی، سبک و وجهی است. افعال سبک که ذکرشان در بخش‌های پیشین رفت و نحوهٔ واحدسازی آنها مشخص شد. افعال وجهی پیش از فعل واقع شده و معنایی وجهی (نگرش گوینده به رخداد) به ساخت اضافه می‌کنند (مانند «**می‌خواهم** بروم»، «**می‌توانم** بخوانم»). این افعال ساختاری مستقل از گروه فعلی بعدشان دارند و باید واحدهای مستقل درنظر گرفته شوند.

(۲۱) می‌خواهم بروم.

واحدسازی سطح ۱ و ۲: «می‌خواهم»، «بروم»

افعال ربطی یا اسنادی معنای مستقلی ندارند و صرفاً ارتباط‌دهندهٔ مسند و مسندٌالیه هستند (جملهٔ ۲۲). افعال کمکی عناصری هستند که به صرف افعال کمک می‌کنند (جملهٔ ۲۳).

(۲۲) غذا نپخته/خام است.

(۲۳) او غذا را نپخته است.

از جملهٔ زیر دو خوانش متفاوت قابل‌برداشت است که متأثر از دو ساختار متمایز بالا است و درک این‌که کدام ساخت مدنظر است، تنها با استناد به بافت امکان‌پذیر است.

(۲۴) غذا نپخته است.

چنانچه «است» را فعل کمکی در نظر گیریم، فعل واژگانی و کمکی روی‌هم‌رفته یک واحد محسوب می‌شوند:

واحدسازی سطح ۱ و ۲: «نپخته‌است»

چنانچه ساخت اسنادی مدنظر باشد مسند و فعل ربطی دو واحد مجزا قلمداد می‌شوند:

واحدسازی سطح ۱ و ۲: «نپخته»، «است»

نکته: تمامی ساخت‌های از نوع صفت+شدن (مانند «خنک شدن»، «گرم شدن») را به پیروی از [6] ساخت اسنادی قلمداد می‌کنیم؛ ساخت‌هایی که معنای استعاری دارند فعل مرکب محسوب می‌شوند و طبق قواعد کلمات نیمه‌مرکب واحدسازی می‌شوند (مانند «پهن کردن (سفره)» یا «داغ کردن» (عصبانی شدن)).

در رابطه با فعل «شدن» به‌کاررفته در ساخت‌های مجهول (مانند «دیده شد») نیز به شیوه‌ای مشابه با افعال اسنادی عمل می‌کنیم و صورت صرف‌شدهٔ فعل و «شدن» را واحدهای مجزا در نظر می‌گیریم:

(۲۵) او دیده شده است.

واحدسازی سطح ۱ و ۲: «دیده»، «شده‌است»

۳-۵- واحدسازی کوته‌نوشت‌ها

اختصارسازی از فرایندهای رایج واژه‌سازی در زبان است که در آن یک یا چند حرف جایگزین یک یا چند کلمه می‌شود. کلمات اختصاری یا کوته‌نوشت‌ها بر حسب این‌که به‌عنوان یک واژهٔ قابل‌تلفظ به‌کار روند یا نه زیر دو دسته قرار می‌گیرند:

(۱) سرنام‌ها که از ترکیب حروف آغازین واژه‌ها ساخته می‌شوند و بعنوان واژه‌های مستقل قابل‌تلفظ به‌کار می‌روند (مانند «ناجا» یا «ساواک»).

(۲) سرواژه‌ها که به شیوه‌ای مشابه با سرنام‌ها ساخته می‌شوند؛ اما یک واژهٔ واحدِ قابل‌تلفظ نیستند (مانند «ای.اف.سی» یا «ه» «ش»).

در فرایند واحدسازی هر دو دستهٔ بالا یک واحد در نظر گرفته می‌شوند:

واحدسازی سطح ۱ و ۲: «ای.اف.سی»
واحدسازی سطح ۱ و ۲: «ناجا»

۳-۵-۱ واحدسازی اعداد

اعداد به‌کار رفته در تاریخ‌ها به‌صورت زیر واحدسازی می‌شوند:

واحدسازی سطح ۱ و ۲: « ۱۳۹۸/۲/۲۳»
واحدسازی سطح ۱ و ۲: «۲۳»، «اردیبهشت»، «۱۳۹۸»
واحدسازی سطح ۱ و ۲: «بیست و سوم»، «اردیبهشت»، «هزار و سیصد و نود و هشت»

اعداد به‌کار رفته در شماره تلفن‌ها به‌صورت زیر واحدسازی می‌شوند:

واحدسازی سطح ۱ و ۲: «۰۹۱۲۱۱۱۱۱۱»
واحدسازی سطح ۱ و ۲: «صفر»، «نهصد و دوازده»، «یازده»، «صد و یازده»

اعداد در گروه‌های نحوی به‌صورت زیر واحدسازی می‌شوند:

واحدسازی سطح ۱ و ۲: «۵۰»، «درصد»
واحدسازی سطح ۱ و ۲: «دو»، «متر»
واحدسازی سطح ۱ و ۲: «چهار»، «سال»، «و»، «نیم»

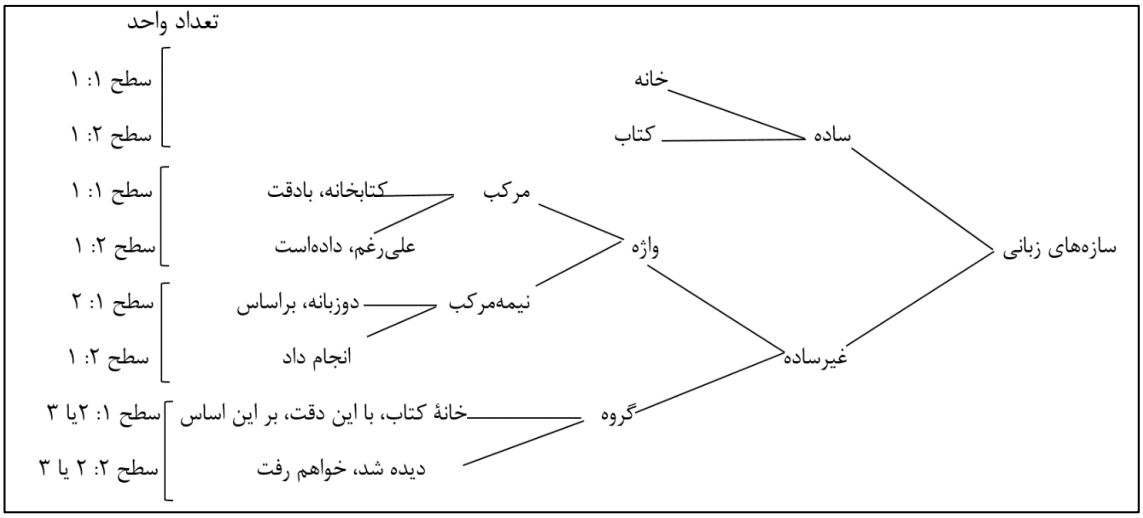
۳-۶- نشانه‌های سجاوندی

واحدسازی نشانه‌ها و علائم نقطه‌گذاری مانند نقطه، خط فاصله، خط مورب و پرانتز حذف می‌شوند مگر آنکه این علائم درون کلمات به‌کار رفته باشند (مانند «(فرا)نظریهٔ صورت‌گرایی»، «پیش‌نویس»، «۲۰:۳۰»، «۵/۵» و غیره).

در اینگونه موارد علائم در درون واژه‌ها باقی مانده و کل واژه یک واحد در نظر گرفته می‌شود:

واحدسازی سطح ۱ و ۲: « (فرا)نظریه»
واحدسازی سطح ۱ و ۲: «۲۰:۳۰»

شکل (۱) خلاصه‌ای از موضوعات مطرح‌شده را نشان می‌دهد.



(شکل-۱): واحدسازی کلمه در زبان فارسی

(Figure-1): Tokenization of Persian words

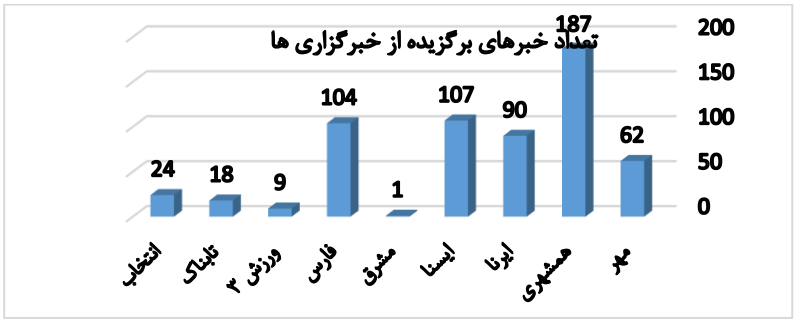
۴- ایجاد داده‌های محک مرتبط با واحدساز

در این بخش به تشرح فعالیت‌های انجام‌شده در راستای تهیه مجموعه‌داده محک مربوط به ابزار واحدساز می‌پردازیم. برای این منظور ابتدا جملات موردنظر انتخاب و سپس براساس موارد مطرح در این مقاله برچسب‌گذاری شد. در ادامه پس از تشریح روال انتخاب جملات، به تشریح ویژگی‌های مجموعه‌دادگان مربوطه می‌پردازیم.

برای ساخت مجموعه‌دادگان مربوطه، امکان تمرکز بر روی انواع مختلف گونه‌های متنی وجود داشت؛ ولی با درنظرگرفتن ملاحظات مختلف در [6]، گونه خبری انتخاب شد. با توجه به آن‌که نیاز است مجموعه‌داده دارای پراکندگی باشند و سبک و سیاق‌های مختلفی را دربرگیرند، فهرستی از خبرگزاری‌ها و سایت‌های خبری انتخاب شد تا منبع دریافت اسناد قرار گیرند. این فهرست به شرح زیر است: خبرگزاری مهر، خبرگزاری همشهری، خبرگزاری ایرنا (خبرگزاری جمهوری اسلامی)، خبرگزاری ایسنا (خبرگزاری دانشجویی ایران)، خبرگزاری فارس، خبرگزاری ورزش۳، خبرگزاری انتخاب، خبرگزاری تابناک. انتخاب اسناد خبری از فهرست خبرگزاری‌ها و سایت‌های خبری با رعایت موازین زیر صورت گرفت:

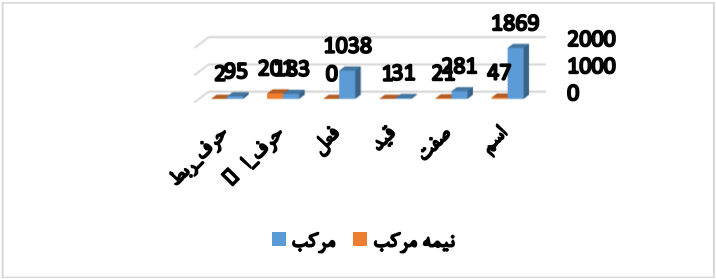
- توزیع موضوعی پیکره باید شامل توزیع متنوعی از موضوعات خبری باشد.
- انتخاب اسناد باید در زمان‌های مختلف صورت گیرد تا بتوان پوشش زمانی نسبتا خوبی را نیز در دادگان درنظر گرفت.
- متوسط طول جملات باید خیلی کوتاه یا بلند نباشد.

با رعایت شروط گفته‌شده انتظار می‌رود که پیکره‌های تولیدشده، نماینده خوبی از دنیای واقعی باشند؛ زیرا سه پارامتر توزیع موضوعی، زمانی و طول جملات به‌صورت هم‌زمان در نظر گرفته است. البته گفتنی است به منظور یک‌دست‌شدن این داده با دادگان محک مربوط به ابزارهای تشخیص موجودیت اسمی [6] و نیز مرجع‌گزینی [3]، دویست سند خبری از این مجموعه‌دادگان انتخاب شد و حدود صد سند خبری جدید هم با رعایت موازین بالا از خبرگزاری‌های یادشده مجددا انتخاب شد که در مجموع سیصد سند خبری را دربرمی‌گیرد. سعی شد از هر یک از این اسناد خبری به‌طور متوسط دو جمله انتخاب شود که در مجموع ۶۰۲ جمله برای برچسب‌گذاری انتخاب و آماده شد. حجم کل هر یک از مجموعه‌ها ۲۱/۱۸۳ کلمه و متوسط طول جملات نیز ۴۰/۲۸ است. شکل (۲) توزیع پراکندگی جملات را در خبرگزاری‌های مختلف نشان می‌دهد.



(شکل-۲): توزیع پراکندگی جملات مجموعه داده از خبرگزاری‌های مختلف

(Figure-2): Distribution of sentence scatter in data sets from different news agencies



(شکل-۳): توزیع کلمات مرکب و نیمه مرکب به تفکیک نوع مقوله آنها در دادگان محک واحدساز

(Figure-3): Distribution of compound and semi-compound words by their POS in tokenization dataset

برای تهیه دادگان محک واحدساز، از موارد یادشده در بخش ۳ استفاده و دادگان در سه سطح برچسب‌گذاری شد. در سطح نخست، مرکب و نیمه‌مرکب‌بودن کلمات مشخص شد. در سطح دوم، نوع مقوله یا POS آنها تعیین می‌شود و در نهایت در سطح سوم، چالش مربوطه تعیین می‌شود. شکل (۳) توزیع کلمات برچسب‌خورده را در دو سطح نخست نشان می‌دهد. جدول (۳) آمار برچسب‌ها را به تفکیک و باجزئیات بیشتر نشان می‌دهد:

(جدول-۳): جزئیات توزیع برچسب‌ها در مجموعه‌دادگان واحدساز

(Tabl-3): Details of the distribution of tags in the dataset

سطح۱	سطح۲	سطح۳	مجموع
مرکب	اسم	هیچکدام	۹
مرکب	اسم	عطفی	۳۹
مرکب	اسم	اعداد	۱۳۴
مرکب	اسم	مصدر	۶۹
مرکب	اسم	تصریفی	۷۹۳
مرکب	اسم	موجودیت اسمی	۱۷۵
مرکب	اسم	کوته نوشت	۱۱
مرکب	اسم	نشانه‌های سجاوندی	۶
مرکب	اسم	ضمیر	۱۵
مرکب	اسم	سایر	۶۱۸

نیمه_مرکب	تک	هیچکدام	۱
نیمه_مرکب	اسم	عطفی	۲
نیمه_مرکب	اسم	اعداد	۶
نیمه_مرکب	اسم	تصریفی	۲
نیمه_مرکب	اسم	موجودیت اسمی	۸
نیمه_مرکب	اسم	سایر	۲۸
مرکب	صفت	هیچکدام	۱
مرکب	صفت	اعداد	۱۵
مرکب	صفت	عطفی	۱
مرکب	صفت	تصریفی	۸۵
مرکب	صفت	سایر	۱۷۹
نیمه_مرکب	صفت	اعداد	۱۷
نیمه_مرکب	صفت	سایر	۴
مرکب	قید	اعداد	۱
مرکب	قید	تصریفی	۲
مرکب	قید	سایر	۲۸
نیمه_مرکب	قید	سایر	۱
مرکب	فعل	هیچکدام	۴
مرکب	فعل	اعداد	۲
مرکب	فعل	عطفی	۵
مرکب	فعل	تصریفی	۵۱۸

مرکب	فعل	سایر	۵۰۹
مرکب	حرف_اضافه	هیچکدام	۹
مرکب	حرف_اضافه	سایر	۱۷۴
نیمه_مرکب	حرف_اضافه	هیچکدام	۹
نیمه_مرکب	حرف_اضافه	سایر	۱۹۲
مرکب	حرف_ربط	هیچکدام	۲
مرکب	حرف_ربط	سایر	۹۳
نیمه_مرکب	حرف_ربط	سایر	۲
مجموع			۳۷۶۹

۵- جمع‌بندی و پژوهش‌های آینده

در سال‌های اخیر پردازش زبان طبیعی در زبان فارسی و تقویت آن در محیط وب و رایانه، همچون دیگر زبان‌های غیرانگلیسی، با محدودیت‌هایی از جمله کمبود پیکره‌ها و منابع داده‌ای و عدم وجود ابزارهای زیرساختی دارای کیفیت و استانداردهای لازم روبرو بوده است.

دادگان و ابزارهای پایه پردازش زبان فارسی از جمله مواردی هستند که برای بخش بزرگی از محصولاتی که نیاز به پردازش زبان دارند مورد نیاز هستند. از طرفی به‌دلیل این که ایجاد و توسعه این ابزارها به‌طورمعمول دارای سوددهی اقتصادی مستقیم نیستند، بسیاری از شرکت‌ها و نهادها نگاه جامعی به توسعه آن نداشته و اغلب برای رفع نیاز خود و در حوزه کاری خود اقدام به ایجاد آن کرده‌اند. از سوی دیگر ضعف کارایی این ابزارها و دادگان، تأثیر مستقیم بر کارایی محصولات حوزه خط و زبان فارسی دارد و به نظر می‌رسد بسیاری از محصولات فعلی به‌دلیل همین ضعف در ابزارهای پایه به سطح کیفی مطلوب نرسیده‌اند؛ بنابراین یکی از ارکان بهبود عملکرد محصولات این حوزه، تقویت ابزارهای پردازش خط و زبان در کاربردهای مختلف پرداخت.

همان‌طور که در مقاله یاد شد، ابزار واحدساز به‌عنوان یکی از ابزارهای پیش‌پردازش زبان است که کاربرد فراوانی در تجزیه و تحلیل متون دارد. این مؤلفه، مرکز کلمات را در متون تشخیص داده و آن را به دنباله‌ای از کلمات به‌منظور تحلیل‌های بعدی تبدیل می‌کند. تنوع در رسم‌الخط فارسی و پیچیدگی‌های واژگانی آن منجر به ایجاد چالش‌های زیادی در این حوزه شده است. در این مقاله سعی شد تا با درنظرگرفتن مباحث و مسائل زبان‌شناسی این حوزه، به تدوین قواعد و شیوه‌های صحیح واحدسازی در زبان فارسی بپردازیم که کاربرد فراوانی در پردازش متون موجود در وب و فضای مجازی خواهد داشت. همچنین سعی شد برای نمونه، مجموعه اسنادی از متون خبرگزاری‌های معتبر تهیه و برچسب‌گذاری آن مطابق با شیوه‌نامه تهیه شده انجام شود و می‌تواند به‌عنوان داده محک برای ارزیابی ابزارهای واحدساز مورد استفاده قرار گیرد.

شیوه‌نامه ارائه‌شده می‌تواند مبنایی برای گسترش‌های آینده پیکره‌ها و ابزارهای واحدساز باشد. از این طریق می‌توان ضمن جلوگیری از پراکنده‌کاری فعالیت‌های این حوزه، امکان ترکیب پیکره‌های تولیدشده از طریق این شیوه نامه را نیز فراهم کرد.

در ادامه در نظر است، ضمن لحاظ‌کردن پیچیدگی‌های بیشتر زبان فارسی (به‌ویژه در لایه معنا)، به توسعه این پیکره پرداخت تا بدین طریق بتوان از آن برای آموزش بهتر ابزارهای توسعه داده‌شده استفاده کرد.

سپاس و قدردانی

از همکاری صمیمانه خانم دکتر شجاعی و آقای دکتر بیجن‌خان که در تهیه و تبیین چالش‌های زبانی ابزار واحدساز صمیمانه به ما کمک کردند، نهایت سپاس‌گزاری را داریم.

۶- مراجع 6- References

[۱] انوشه، مزدک، "فرافکن‌های نمود و زمان در صفت‌های فاعلی مرکب بر پایهٔ نظریهٔ صرف توزیعی"، مجلهٔ جستارهای زبانی، شماره ۵، صفحات ۴۹-۷۲، ۱۳۹۴.

[1] A. Mazdak, "Aspect and tense projections in the complex agentive adjectives: A distributed morphology approach", *Language Related Research*, Vol.6, No. 5, pp. 49-72, 2015

[۲] انوشه، مزدک، "مسئلهٔ مجهول در زبان فارسی: رویکردی کمینه‌گرا"، مجلهٔ پژوهش‌های زبانی، شماره ۱، صفحات ۱-۲۰، بهار و تابستان ۱۳۹۴.

[2] A. Mazdak, "Passive structure in Persian: a minimalist approach", *Language Research*, Vol. 6, No. 5, pp. 1-20, 2015

[۳] رحیمی، زینب ، حسین نژاد، شادی، "هم مرجع یابی مبتنی بر پیکره در متون فارسی"، فصلنامه پردازش علائم و داده ها، شماره ۱، ۱۳۹۹.

[3] Z. Rahimi , sh. Hosseinnejad , "Corpus based coreference resolution for Farsi text",*Signal and*

Data Processing, Vol. 17, No. 1, pp. 79-98, 2020

[۴] رستم پور، سعیده، "ساخت واژگان محاسباتی برای زبان فارسی"، پایان نامه کارشناسی مهندسی نرم‌افزار، دانشگاه شهید بهشتی، تهران، ۱۳۸۵.

[4] S. Rostampour, "Computational vocabulary building for Persian language", Master's thesis in software engineering, Shahid Beheshti University, Tehran, 2006

[۵] شریفی آتشگاه، مسعود، "تولید نیمه‌خودکار درخت‌بانک گروه‌های نحوی در متون فارسی"، رسالهٔ دکتری، دانشگاه تهران، ۱۳۸۸.

[5] M. Sharifi Atashgah, "Semi-automatic production of tree bank of syntactic groups in Persian texts", Doctoral dissertation, Tehran University, 2009

[۶] شهشهانی، مهساسادات، مهدی محسنی، آزاده شاکری، هشام فیلی، "پیما: پیکره برچسب‌خورده موجودیت‌های اسمی زبان فارسی"، فصلنامه علمی پردازش علائم و داده‌ها، شماره۱، صفحات ۹۲-۱۱۰، ۱۳۹۸.

[6] M. A. Shahshahani, M. Mohseni, A. Shakery, H. Faili, "PAYMA: A Tagged Corpus of Persian Named Entities", *Signal and Data Processing*, Vol. 16, Issue 1, pp. 91-110, 2019

[۷] طباطبایی، علاء الدین، اسم و صفت مرکب در زبان فارسی. تهران، مرکز نشر دانشگاهی، ۱۳۸۲.

[7] A. Tabatabaee, "Compound nouns and adjectives in Persian language", University Publication Center, 2003

[۸] طباطبایی، علاء الدین، "فرایندهای واژه‌سازی زبان فارسی و استقلال صرف و نحو"، مجلهٔ پژوهش‌های زبان‌شناسی، صفحات ۸۷-۹۹، ۱۳۸۹.

[8] A. Tabatabaee, "Word formation processes of Persian language and the independence of grammar and syntax", *Journal of Linguistic Research*, pp. 87-99, 2010

[۹] قیومی، مسعود، "ارائهٔ یک روش مبتنی بر مدل زبانی برای واحدسازی پیکرهٔ فارسی"، مجلهٔ زبان و زبان‌شناسی، شماره ۲۷، صفحات ۲۱-۵۰، ۱۳۹۷.

[9] M. Ghayoomi, "Proposing a Method based on Language Modeling to Tokenize a Persian Corpus", *Journal of Language and Linguistic*, Vol. 14, Issue 27, pp. 21-50, 2018

[10] R. Dridan, & S. Oepen, "Tokenization: Returning to a long solved problem a survey, contrastive experiment, recommendations, and toolkit", *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers*, Vol 2. pp 378–382, 2012.

[11] K. Evang, V. Basile, G. Chrupała, J. Bos. "Elephant: Sequence Labeling for Word and Sentence Segmentation", *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1422–1426. October 2013.

[12] M. Fares, S. Oepen, , & Y. Zhang, "Machine Learning for High-Quality Tokenization Replicating Variable Tokenization Schemes", *In Alexander Gelbukh, editor, Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Proceedings*, Part I, pp. 231–244, Berlin, Heidelberg. Springer, 2013.

[13] R. Folli, H., Harley, & S. Karimi. "Determinants of the event type in Persian complex predicates", *Lingua*, vol. 115, pp.1365–1401 .2005.

[14] Gh. Karimi-Doostan, "Separability of light verb constructions in Persian", *Studia Linguistica*, pp. 70–95. 2011.

[15] S. Kiani, & M. Shamsfard, "Persian Text Segmentation and Tokenization", *Nineteenth Meeting of Computational Linguistics in the Netherlands (CLIN 19)*, Netherland. 2009.

[16] J. Mark, & O. Bojar, "TrTok: A Fast and Trainable Tokenizer for Natural Languages", *in the Prague Bulletin of Mathematical Linguistics*, vol. 98(-1), 2012.

[17] K. Megerdoomian, R. Zajac, "*Processing Persian text: Tokenization", in the Shiraz project. Computing Research Laboratory, New Mexico State University.* 2000.

[18] K. Megerdoomian, "Developing a Persian Part of Speech Tagger", *The 1th Workshop on Persian Language and Computer*, pp. 99-105. 2004.

[19] D. Riazati, "Computational Analysis of Persian Morphology", *MSc thesis, RMIT*. 1997.

[20] M. Shamsfard, "Developing FarsNet: A Lexical Ontology for Persian", *4th Global WordNet Conference (GWC'08).* 2008.

[21] M. Shamsfard, S. Kiani & Y. Shahedi, *Third International Workshop on Computational Approaches to Arabic-Script based languages (CAASL3)*, Ottawa, Canada. 2009.

[22] M. Shamsfard, A. Hesabi, H. Fadaei, N. Mansoory, A. Famian., S. Bagherbeigi, E. Fekri, M. Monshizadeh, M. Assi, "Semi-Automatic Development of FarsNet", *The Persian WordNet, Global WordNet Conference*, Mumbai, India. 2010.

[23] M. Sharifi Atashgah, M. Bijankhan. "Corpus-based analysis for multi-token units in Persian", *International Journal of Information and Communication Technology. Tehran: Iran Telecom Research Center*, vol. 1 (3), pp 15-26. 2009.

مژگان فرهودی مدرک کارشناسی خود را از دانشگاه خوارزمی در رشته مهندسی کامپیوتر (نرم‌افزار) و مدرک کارشناسی ارشد خود را در رشته مهندسی فناوری اطلاعات از دانشگاه صنعتی امیرکبیر دریافت کرده است. وی از سال ۱۳۸۰ تاکنون در پژوهشگاه ارتباطات و فناوری اطلاعات به‌عنوان پژوهشگر مشغول به فعالیت است. ایشان در حال حاضر به‌عنوان عضو هیئت علمی پژوهشگاه در حوزه‌های بازیابی اطلاعات، پردازش زبان طبیعی، کلان‌داده، داده‌کاوی و وب‌کاوی فعالیت می‌کند.

farhoodi@itrc.ac.ir

مریم محمودی دوره کارشناسی خود را در دانشگاه آزاد اسلامی واحد تهران جنوب در رشته مهندسی کامپیوتر (نرم‌افزار) و دوره کارشناسی ارشد خود را در رشته مهندسی فناوری اطلاعات در دانشگاه صنعتی امیرکبیر گذرانده است. وی از سال ۱۳۸۲ تاکنون در پژوهشگاه ارتباطات و فناوری اطلاعات به‌عنوان پژوهش‌گر پژوهشگاه در حوزه‌های بازیابی اطلاعات، پردازش زبان طبیعی، داده‌کاوی و وب‌کاوی فعالیت می‌کنند.

نشانی رایانامه ایشان عبارت است از:

mahmoudy@itrc.ac.ir

مونا داودی شمسی مدرک کارشناسی خود را از دانشگاه شهید بهشتی در رشته مهندسی کامپیوتر (نرم‌افزار) دریافت کرده است. او از سال ۱۳۸۰، تاکنون در پژوهشگاه ارتباطات و فناوری اطلاعات مشغول به فعالیت پژوهشی است. ایشان در سال‌های اخیر، به‌عنوان پژوهش‌گر در پژوهشگاه در حوزه‌های مرتبط با پردازش زبان طبیعی فعالیت می‌کنند.

نشانی رایانامه ایشان عبارت است از:

davoudi@itrc.ac.ir