Contemporary Persian Inflectional Analyzer

Davood Heidarpour
MSc in Computational Linguistics; Faculty of New Sciences and Technologies; University of Tehran; Tehran, Iran;
Email: d.heidarpour@ut.ac.ir

Elham S. Sebt
MSc in Computational Linguistics; Faculty of New Sciences and Technologies; University of Tehran; Tehran, Iran;
Email: e.sebt@ut.ac.ir

Mahmood Bijankhan
PhD in General Linguistics; Full Professor; Faculty of Literature and Humanities; University of Tehran; Tehran, Iran;
Email: mbjkhan@ut.ac.ir

Mostafa Salehi*
PhD in Computer Engineering; Associate Professor;
Faculty of New Sciences and Technologies; University of Tehran;
Tehran, Iran Email: mostafa_salehi@ut.ac.ir

Hadi Veisi
PhD in Computer Engineering; Assistant Professor;
Faculty of New Sciences and Technologies; University of Tehran;
Tehran, Iran Email: h.veisi@ut.ac.ir

Iranian Journal of Information Processing and Management

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 36 | No. 4 | pp. 945-970
Summer 2021
https://doi.org/10.52547/jipm.36.4.945

Received: 19, Sep. 2019 **Accepted: 07, Mar. 2021**

Abstract: In recent years, the use of informal writing in Persian has grown significantly due to the increasing expansion of cyberspace and social media and platforms, and the tendency of users to bring the written language closer to colloquial speech. But on the other hand, proper tools to process this language register are not developed very much. One of the tools for low level processing of textual data is an inflectional analyzer. However, such tools are not developed for this register yet. Informal words have their own structures, stems, morphemes and clitics and they also make use of formal structures and units. Moreover, this register also consists of formal words so any analyzer for informal words should have the potential to analyze formal words, too. In this paper, it is tried to cover all inflectional structures of informal Persian language to build an inflectional analyzer. A corpus of most of its known sub-registers is constructed to extract words, morphemes and inflectional rules and morphotactics. A part of this corpus is used for testing the analyzer. After extracting 1786 unique words of the test part, inflectional analyzer f-measure is equal to 97.67%.

* Corresponding Author

This tool can be used in computational processing of Persian language and it can also be used in teaching Persian, specifically colloquial Persian to non-Persian learners.

Keywords: Computational Linguistics, Inflectional Analyzer, Informal Persian Register, Contemporary Persian, FST, Persian Teaching

تحلیلگر تصریفی فارسی معاصر

داود حیدرپور
کارشناسی ارشد زبان‌شناسی رایانشی؛ دانشکدهٔ علوم و فنون نوین؛ دانشگاه تهران؛ تهران، ایران؛
d.heidarpour@ut.ac.ir

الهام‌سادات سبط
کارشناسی ارشد زبان‌شناسی رایانشی؛ دانشکدهٔ علوم و فنون نوین؛ دانشگاه تهران؛ تهران، ایران؛
e.sebt@ut.ac.ir

محمود بی‌جن‌خان
دکتری زبان‌شناسی؛ استاد تمام؛ گروه زبان‌شناسی؛ دانشگاه تهران؛ تهران، ایران mbjkhan@ut.ac.ir

مصطفی صالحی
دکتری مهندسی کامپیوتر؛ دانشیار؛ دانشکدهٔ علوم و فنون نوین؛ دانشگاه تهران؛ تهران، ایران؛
پدیدآور رابط mostafa_salehi@ut.ac.ir

هادی ویسی
دکتری مهندسی کامپیوتر؛ استادیار؛ دانشکدهٔ علوم و فنون نوین؛ دانشگاه تهران؛ تهران، ایران؛
h.veisi@ut.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۸/۰۶/۲۸ پذیرش: ۱۳۹۹/۱۲/۱۷ مقاله برای اصلاح به مدت ۴ ماه نزد پدیدآوران بوده است.

چکیده: در سال‌های اخیر، کاربرد گونهٔ نوشتاری غیررسمی زبان فارسی به‌دلیل گسترش روزافزون فضای مجازی و شبکه‌های اجتماعی و تمایل کاربران به نزدیک کردن زبان نوشتار به گفتار رشد چشمگیری داشته است. با وجود این، ابزارهای پردازش این گونهٔ زبانی به‌میزان لازم توسعه داده نشده است. تحلیلگرهای تصریفی از جملهٔ ابزارهایی است که در پردازش زبانی کاربرد وسیعی دارد و تاکنون برای گونهٔ غیررسمی طراحی و پیاده‌سازی نشده است. با توجه به این‌که گونهٔ نوشتاری غیررسمی در کنار واژگان و قواعد صرفی و نحوی مختص به‌خود، در بخشی از واژه‌ها و ساختارها با گونهٔ رسمی مشترک است، در این پژوهش با پوشش فارسی رسمی و غیررسمی اولین ابزار تحلیل تصریفی فارسی معاصر برای همهٔ اقسام واژه توسعه داده شده و تلاش شده همهٔ ساختارهای تصریفی واژه‌های فارسی غیررسمی پوشش داده شود. این ابزار به‌صورت قاعده‌مند و مستقل از بافت و با بهره‌گیری از مبدل حالت محدود، پی‌بست‌ها و وندهای تصریفی رسمی و غیررسمی را در واژه‌های زبان شناسایی و تحلیل کرده، ستاک‌های رسمی و غیررسمی را نیز استخراج می‌کند. به‌منظور پوشش دادن تمام ساخت‌ها و حالت‌های تصریفی، با توجه به

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۶ | شمارهٔ ٤ | ص‌ص ۹۴۵-۹۷۰
تابستان ۱۴۰۰
https://doi.org/10.52547/jipm.36.4.945

رویکرد مستقل از بافت، الگوریتم برای هر واژه، تمام خوانش‌ها و معانی گوناگونی را که می‌تواند بسته به قرارگیری در بافت‌های گوناگون داشته باشد، تحلیل و ارائه می‌کند. به‌منظور استخراج و بررسی واژگان و قواعد تصریفی و نگارشی گونهٔ غیر رسمی، پیکرهٔ فارسی معاصر از سیاق‌ها و زیرسیاق‌های گوناگون این گونهٔ زبانی تهیه شده و در طراحی و آزمون تحلیلگر مورد استفاده قرار گرفت. آزمون تحلیلگر با استفاده از ۱۷۸۶ واژهٔ یکتای استخراج‌شده از پیکره، نتیجهٔ ۹۶/۶۷ درصد را در معیار اف به‌دست داده است. از این ابزار می‌توان در انواع تحلیل‌ها و کاربردهای پردازش رایانه‌ای زبان فارسی و همچنین در آموزش فارسی، به‌ویژه محاورهٔ فارسی به غیر فارسی‌زبانان استفاده کرد.

کلیدواژه‌ها: پردازش زبانی، تحلیلگر تصریفی، گونهٔ غیررسمی فارسی، فارسی معاصر، مبدل حالت محدود، آموزش فارسی

۱. مقدمه

گونه‌ای از زبان فارسی که به‌عنوان زبان رسمی کشور در اسناد و مکتوبات حکومتی، آموزشی، علمی و نیز روزنامه‌ها و نشریات به‌کار می‌رود، گونهٔ نوشتاری معیار یا رسمی است. در کنار این گونه می‌توان به گونهٔ غیررسمی اشاره کرد که به‌عنوان نوشتار ارتباطی در شبکه‌های اجتماعی، پیام‌رسان‌های تلفنی، وبلاگ‌ها و گاهی سایت‌ها و نامه‌های الکترونیکی استفاده می‌شود و کاربر تلاش می‌کند زبان نوشتار را هرچه بیشتر به گونهٔ گفتاری نزدیک کند. هرچند این گونهٔ نوشتاری، ساختاری مختص به خود دارد و از لحاظ صرفی، نحوی و واژگانی با زبان نوشتاری معیار متفاوت است، اما در کنار تفاوت‌ها وجوه اشتراک دو گونه نیز بالاست. در بررسی آماری پیکره‌ای که برای این پژوهش جمع‌آوری شده (شامل بیش از ۴۹ هزار واژه که به‌صورت دستی تقطیع[1] شده)، ۸۳/۵ درصد (بیش از ۴۱ هزار) واژه‌های متون فارسی غیررسمی با زبان فارسی گونهٔ رسمی مشترک است و تنها ۱۶/۵ درصد واژه‌ها غیررسمی است (بخش ۳). لازم به ذکر است که واژهٔ غیررسمی بر اساس این پژوهش، واژه‌ای است که حداقل یک عنصر غیررسمی در ساختار خود داشته باشد و این عنصر ریشه یا هر یک از اجزای تصریفی واژه است.

در حوزهٔ پردازش رایانه‌ای زبان، تحلیل صرفی به‌عنوان پردازش پایه در بسیاری از پردازش‌های زبانی سطح بالاتر کاربرد دارد، اما ابزارهای شناسایی فارسی غیررسمی بر خلاف گسترش بسیار این گونهٔ زبانی در رسانه‌های اجتماعی و فضای وب، توسعهٔ بسیار

1. tokenize

اندکی یافته‌اند.

پژوهش‌های انجام‌شده در حوزهٔ ساخت‌واژهٔ زبان فارسی بیشتر به گونهٔ رسمی زبان پرداخته‌اند، از جمله، «اسلامی» و همکاران (۱۳۸۳)، «اسلامی و علی‌زاده لجمیری» (۱۳۸۸)، Shamsfard, Jafari and Ilbeygi (2010) و «مواجی، اسلامی و وزیرنژاد» (۱۳۹۰). تحلیلگرهای تصریفی که در این پژوهش‌ها پیاده‌سازی شده، مختص زبان فارسی رسمی است و قادر به تصریف فارسی غیررسمی نیست. این پژوهش‌ها بر اساس ساختار تصریفی ارائه‌شده توسط «اسلامی» و همکاران (۱۳۸۳) طراحی شده‌اند. تنها پژوهش Megerdoomian (2000) بر اساس ساختار تصریفی‌ای که خود او استخراج و تبیین کرده، پیاده‌سازی شده است. تعداد اندکی از پژوهش‌ها نیز به فارسی غیررسمی پرداخته‌اند. «مگردومیان» به‌صورت پراکنده ساختارهای غیررسمی را شناسایی کرده است (Megerdoomian 2008). این ساختارها می‌بایست به تحلیلگر رسمی او -ساختارهایی که خود او استخراج و تبیین کرده- افزوده می‌شد تا تحلیلگر تصریفی بتواند در کنار تصریف واژگان رسمی، واژگان غیررسمی را نیز پوشش دهد، اما این طرح ناتمام باقی مانده است. گرچه این پژوهش قبل از فراگیر شدن پیام‌رسان‌ها و شبکه‌های اجتماعی انجام شده، اما با توجه به رشد وبلاگ‌های فارسی و گسترش کاربرد فارسی غیررسمی در آن اولین پژوهشی است که به فارسی غیررسمی در فضای مجازی پرداخته است. «تازه‌جانی و بحرانی» با انجام پژوهشی محدود، تنها بخش کوچکی از افعال غیررسمی را پوشش داده و تحلیلگری تصریفی برای آن ساخته‌اند (۱۳۹۲). «طبیب‌زاده» نیز مانند «مگردومیان» با مطالعه‌ای پیکره‌بنیاد، اما بر خلاف آن محدود به داستان‌ها و نمایشنامه‌های صد سال اخیر فارسی، رخدادهای کلمات غیررسمی را بررسی کرده و ساختار مشاهده‌شده در کلمات را گزارش کرده است (۱۳۹۸). این پژوهش صرفاً نظری است و فاقد ابزار تحلیل رایانه‌ای این‌گونه کلمات است.

استفاده از زبان غیررسمی در اینترنت در میان سخنوران سایر زبان‌ها نیز رایج بوده و تلاش برای تحلیل رایانه‌ای آن نیز در جریان است. برای مثال، Torjmen and Haddar (2018) برای تحلیل گونهٔ بیانی تونسی و Alshargi and Rambow (2016) برای تحلیل گونهٔ بیانی یمنی (صنعایی) و مراکشی، که همگی متفاوت از عربی استانداردِ مدرن (Ryding 2005) یا عربی نوشتاری مدرن[1] (Badawi, Carter and Gully 2013) هستند، با شناسایی

1. Modern Written Arabic (MWA)

ساختار چنین کلماتی (با استخراج از پیکره‌هایی که برای این منظور ساخته‌اند) و افزودن آن‌ها به ساختار مبدل‌های[1] عربی استاندارد یا نوشتاری مدرن، تحلیلگرهای تصریفی برای گونه‌های بیانی خود طراحی کرده‌اند.

با توجه به کاستی‌های این حوزه، در پژوهش حاضر تلاش شده است که ساختارهای تصریفی فارسی غیررسمی به‌طور کامل بررسی و شناسایی شده، و تحلیلگر تصریفی فارسی غیررسمی بر مبنای قواعد جامع و مانع در این حوزه طراحی و پیاده‌سازی شود. هدف در طراحی این تحلیلگر شناسایی و تحلیل ستاک، وندهای تصریفی و پی‌بست‌های واژه است که در سامانه‌های پردازش زبان طبیعی[2] کاربرد دارد.

از آنجا که در گونهٔ غیررسمی از گونهٔ رسمی نیز استفاده می‌شود، بنابراین، هر تحلیلگر واژه برای فارسیِ غیررسمی می‌بایست قادر به تحلیل واژه‌های فارسی رسمی هم باشد. فارسی‌ای که در این پژوهش مد نظر است، فارسی امروزین است که در نوشتار و گفتار فارسی‌زبانان در گونهٔ رسمی و غیررسمی استفاده می‌شود و شامل گونهٔ ادبی و تاریخی نمی‌شود. بنا بر نظر «لازار» فارسی معاصر کاربرد امروزین فارسی را در هر شکلش در نظر دارد و نه ساختاری که لزوماً در متون کلاسیک فارسی وجود دارد و یا تعریف شده است (Lazard 1992).

۲. چارچوب پیشنهادی

رویکرد رایج در تحلیل زبانی قاعده‌مند و در سطح واژه، استفاده از مبدل حالت محدود[3] است که می‌تواند دو رشته را به یکدیگر نگاشت کند. این نگاشت بین واژه و تمام قطعات سازندهٔ قاعدهٔ کلی واژه صورت می‌گیرد. بنابراین، ورودی مبدل می‌تواند واژه و خروجی آن قاعدهٔ سازندهٔ واژه باشد. به‌منظور پیاده‌سازی یک تحلیلگر با استفاده از مبدل حالت محدود به سه منبع زبانی نیاز داریم: واژگان، قواعد تصریفی و قواعد نگارشی (Jurafsky and Martin 2008).

واژگان[4] فهرستی از ستاک‌ها و وندهای زبان همراه با اطلاعاتی مثل مقولهٔ دستوری واژه است. قواعد هم‌نشینی[5] تک‌واژها، مدلی از هم‌نشینی و باهم‌آیی تک‌واژ در یک واژه

---

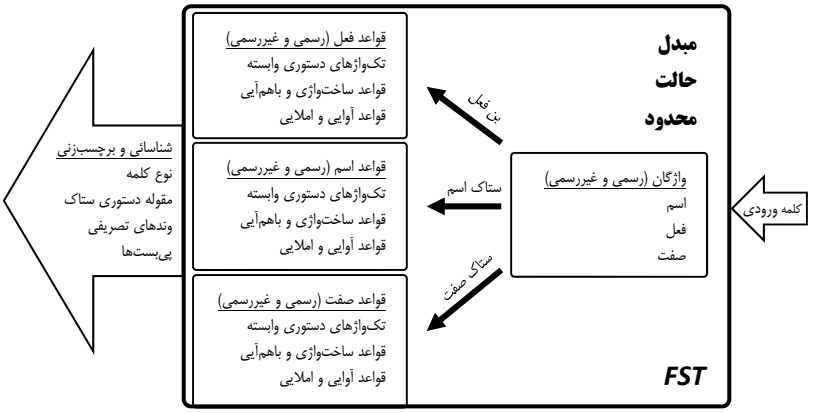
1. transducer

۲. نظیر تحلیل معنایی، نظرکاوی، بازشناسی گفتار و غیره.

3. finite state transducer (FST)
4. lexicon
5. morphotactic rules

ارائه می‌کند و قواعد املایی، تغییرات املایی ناشی از تغییرات واژ-واجی[1] را هنگام ترکیب دو تک‌واژ توصیف می‌کند.

در این پژوهش به‌منظور پردازش تصریفی قاعده‌مند از ابزار «فوما»[2] (Hulden 2009)، که یک ابزار مبدل حالت محدود متن آزاد و رایگان است، استفاده شده است. «فوما» این امکان را فراهم می‌آورد که با قرار دادن چندین مبدل در سیستم، واژهٔ ورودی را از یکی یا از همهٔ آن‌ها عبور داد و قاعده تولید کرد. هر مبدل می‌تواند برای تغییر و یا شناسایی بخشی از قواعد ساختِ واژه به‌کار رود. همچنین، کاربر می‌تواند با استفاده از کتابخانه‌های تعریف‌شده یا به‌صورت شخصی‌شده در مبدل‌ها جست‌وجو انجام دهد و یا ترتیب و نحوهٔ اجرا شدن مبدل‌ها را شخصی‌سازی کند. شکل ۱، ساختار مفهومی یک مبدل را نشان می‌دهد. کلمهٔ ورودی پس از شناسایی ستاک و ساختار تصریفی و برچسب‌زنی همهٔ اجزای سازندهٔ آن در خروجی نمایش داده می‌شود.



شکل ۱. مدل مفهومی مبدل حالت محدود تحلیلگر تصریفی

۲-۱. چالش‌های پیاده‌سازی تحلیلگر تصریفی فارسی معاصر

چنان‌که ذکر شد، گونه‌ای از زبان که به‌عنوان فارسی معاصر در این پژوهش مورد توجه است، گونه‌ای ارتباطی است که به شکل گسترده در شبکه‌های اجتماعی، پیام‌رسان‌ها و وبلاگ‌ها مورد استفاده قرار می‌گیرد و دارای ویژگی‌های ارتباطی نزدیک

1. morphophonemic 2. Foma

به گفتار است. با توجه به محدودیت‌های فنی رسانهٔ[1] نوشتاری فارسی و فقدان استاندارد، کاربران از شیوه‌های گوناگونی برای نگارش استفاده می‌کنند. اما به‌رغم تغییرات آوایی و یکدست نبودن قوانین نگارشی، فارسی غیررسمی در بسیاری از موارد در ساختار تصریفی خود از قواعدی تبعیت می‌کند که در معدود پژوهش‌های این حوزه به برخی از آن‌ها اشاره شده است. اما لازم است برای تحلیل دقیق و درست، با بررسی داده‌های استاندارد این گونهٔ زبانی، ساختار جامع و مانعی برای آن تعریف شود. با توجه به در دسترس نبودن دادهٔ استاندارد برای این گونهٔ نوشتاری، نخستین قدم فراهم‌آوردن دادهٔ مناسب به‌منظور بررسی و استخراج الگوهای نوشتاری، تک‌واژهای تغییریافته و قواعد ساخت‌واژی است. با توجه به این‌که در این گونهٔ نوشتاری به‌طور معمول واژه‌ها چه در شکل رسمی و چه در شکل غیررسمی، تابع الگوی نگارشی یکسان و ثابتی نیستند و پوشش و ذخیرهٔ همهٔ این واژه‌ها در پیاده‌سازی تحلیلگر امکان‌پذیر نیست، بنابراین، بخشی از واژه‌ها، به‌ویژه آن‌هایی که تحت تأثیر تغییرات آوایی قرار می‌گیرند، تبدیل به واژه‌های خارج از واژگانِ تحلیلگر می‌شوند و شناسایی آن‌ها به‌راحتی امکان‌پذیر نیست. برای به حداقل رساندن تأثیر چنین مواردی در دقت نهایی تحلیلگر لازم است به نکات زیر توجه شود:

- ◇ حذف برخی حروف: همزه در پایان واژه‌ها یا بر روی پایه‌های «و»، «ا»، و «ی» وجودی غیر قطعی دارد و ممکن است حذف شود؛ مثلاً «أعضاء: اعضا». حرف «ه» در نقش واکه در صورتی که در پایان یک تک‌واژ قرار بگیرد، در بسیاری از موارد، در هنگام اتصال به تک‌واژی دیگر حذف می‌شود؛ مثلاً «به‌طوریکه: بطوریکه، همه شون: همشون».
- ◇ فاصله‌گذاری: فاصله، نیم‌فاصله و اتصال، سه حالتِ در کنار هم قرار گرفتن واژه‌ها و یا اجزای واژه‌هاست. گرچه «فرهنگستان زبان و ادب فارسی» استانداردی برای آن در نظر گرفته (صادقی و زندی‌مقدم ۱۳۸۵)، اما به‌طور معمول، در نگارش فارسی رعایت نمی‌شود؛ مثل «کتاب‌ها، کتابها و کتاب ها» و در نگارش غیررسمی نیز تنوع بیشتری تولید می‌کند.
- ◇ خطای املایی: خطای املایی یکی از مشکلات اجتناب‌ناپذیر در همهٔ پردازش‌های متنی به‌ویژه فارسی غیررسمی است. بخشی از این خطاها که منشأ نگرشی دارد، با

1. medium

برطرف‌کردن سایر چالش‌های پردازش فارسی غیررسمی قابل کنترل است؛ مانند جایگزینی حروف هم‌صدا با یکدیگر. سایر خطاهای ناشی از حروف‌چینی نیز با روش‌های شناسایی و رفع خطای املایی می‌بایست برطرف شود.

۳. دادگان

در تهیهٔ دادگان برای این پژوهش تلاش شد تمام زیرسیاق‌های فارسی غیررسمی شناسایی شده، برای هر کدام متونی تهیه، بررسی و برچسب‌گذاری شود. مجموعهٔ پیکرهٔ جمع‌آوری‌شده نزدیک به پنجاه‌هزار واژه است. واژگان این پیکره به شکل دستی تقطیع شده و حدود ۳۰ درصد آن برای ارزیابی کنار گذاشته شد. از بقیه برای بررسی ساخت‌های تصریفی، آمارگیری از میزان توزیع ساخت‌های مختلف، بررسی و استخراج واژگان و ساخت‌های غیررسمی استفاده شد. تمام واژه‌های غیررسمی این پیکره به شکل دستی معادل‌نویسی شده و بن‌واژهٔ[1] آن‌ها نیز استخراج شده است. ساختار پیکره‌ای که به‌طور متوازن از این زیرسیاق‌ها (که همگی مربوط به زبان فارسی غیررسمی است) استفاده می‌کند، از نوع پیکرهٔ نمونه‌گیری‌شده[2] است. برای هر سیاق[3] زبان فارسی غیررسمی نمونه‌ها، هم در مرحلهٔ جمع‌آوریِ اولیه و هم در مرحلهٔ گزینش به شکل تصادفی[4] انتخاب شده است (McEnry and Hardie 2011). حجم دادهٔ هر بخش هم به نسبت کاربرد آن زیرسیاق انتخاب شده است. جدول ۱، داده‌های جمع‌آوری‌شده در تهیهٔ پیکرهٔ فارسی معاصر را به تفکیک سیاق‌ها و زیرسیاق‌ها نشان می‌دهد.

٤. قواعد و واژگان فعلی

در این بخش واژگان فعلی تحلیلگر که شامل بن افعال و تک‌واژهای وابستهٔ متصل به فعل است و همچنین، ساخت تصریفی افعال در گونهٔ غیررسمی توضیح داده می‌شود.

1. lemma
2. sampling corpus
3. register
4. random sampling

جدول ۱. سیاق‌ها و زیرسیاق‌های موجود در پیکرهٔ فارسی معاصر

ردیف	سیاق	زیرسیاق و حجم کلی
۱	گفت‌وگو و بیان واقعی[1]	مصاحبه، سخنرانی از پیش آماده‌شده، اجرای رادیو/ تلوزیونی از روی متن؛ ۱۲۴۸ واژه
۲	گفت‌وگو و بیان خیالی[2]	رمان، نمایش‌نامه، فیلم‌نامه، زیرنویس فیلم، شعر محاوره؛ ۷۵۶۷ واژه
۳	نظر کاربران فضای مجازی	مصرف‌کنندهٔ کالا یا خدمات، مقالات خبری و متفرقه؛ ۹۰۵۰ واژه
۴	پیام‌رسان تلفنی	مکالمهٔ شخصی، مکالمهٔ گروهی، اطلاع‌رسانی؛ ۱۱۹۴۲ واژه
۵	شبکهٔ اجتماعی	پست‌ها، پاسخ‌ها؛ ۸۴۳۱ واژه
۶	وبلاگ	پست‌ها، پاسخ‌ها؛ ۸۹۰۳ واژه
۷	نامهٔ شخصی و خاطرات روزانه؛ ۲۱۴۶ واژه	

٤-۱. واژگان فعلی

در ساختمان فعل‌های رسمی ساده و پیشوندی از بن‌های فعل ماضی و مضارع استفاده می‌شود. بن‌های رسمی از واژگان «زایا» (اسلامی و همکاران ۱۳۸۳)، «ویراستیار» (کاشفی، نصری و کنعانی ۱۳۸۹)، واژگان «پرلکس» (Sagot and Walther 2010)، و پیکرهٔ جمع‌آوری‌شدهٔ فارسی غیررسمی به‌دست آمده است. بن فعل‌های حال و گذشتهٔ غیررسمی نیز از پیکرهٔ جمع‌آوری‌شده و هم با معادل‌سازیِ بن فعل‌های رسمی به‌دست آمده است که در مجموع، شامل ۵۱۶ بن ماضی سادهٔ رسمی و ۸۷ بن غیررسمی، ۴۹ بن ماضی پیشوندی رسمی و ۲۳ بن غیررسمی، ۳۷۹ بن مضارع سادهٔ رسمی و ۸۷ بن غیررسمی و نیز ۴۹ بن مضارع پیشوندی رسمی و ۲۲ بن غیررسمی است. شناسه‌ها و پی‌بست‌ها با توجه به پژوهش‌های انجام‌شده، مانند Megerdoomian (2008)، «اسلامی» و همکاران (۱۳۸۳)، «انوری و گیوی» (۱۳۹۱)، «بی‌جن‌خان» (۱۳۸۶) و از پیکرهٔ جمع‌آوری‌شده به‌دست آمده است (جدول ۲).

در ساخت افعال گونهٔ غیررسمی استفاده از پی‌بست‌های فاعلی و مفعولی رایج است. پی‌بست‌های مفعولی «م، ت، ش، مان، تان، شان» در متون رسمی نیز به‌ندرت استفاده می‌شود، اما در متون غیررسمی صیغه‌های جمع بیشتر به‌صورت «مون، تون، شون» استفاده می‌شود. پی‌بست فاعلی «ش» مختص گونهٔ محاوره است و در فعل‌های سوم‌شخص به‌کار

1. non-fictional 2. fictional

می‌رود. بیشتر نمودهای فعل ماضی مانند ماضی ساده، استمراری و بعید، صرف نظر از گذرا یا ناگذر بودن فعل پی‌بست فاعلی را می‌پذیرند. پی‌بست فاعلی می‌تواند به فعل‌های مضارع اخباری و التزامی لازم نیز افزوده شود.

جدول ۲. شناسه‌های فارسی رسمی و غیررسمی[1]

شخص و شمار	بعد از همخوان‌ها و واکه ی		بعد از واکه آ		بعد از واکه و	
	رسمی	غ.رسمی	رسمی	غ.رسمی	رسمی	غ.رسمی
۱مفرد	!م می‌خندم، می‌زیم، می‌گم		!(ی)م می‌آیم	!م میام	!(ی)م می‌رویم	
۲مفرد	!ی می‌خندی، می‌زیی، می‌گی		!(ی)ی می‌آیی	!ی میای	!(ی)ی می‌رویی	
۳مفرد	مضارع: !د ماضی: Ø می‌خندد، زیست	مضارع: !ه ماضی: Ø می‌گه، می‌ذاشت	!(ی)د می‌آید	!د میاد	!(ی)د می‌روید	!(ی)ه می‌رویه
۱جمع	!یم می‌خندیم، می‌زییم، می‌گیم		!(ی)یم می‌آییم	!یم میایم	!(ی)یم می‌روییم	
۲جمع	!ید می‌خندید، می‌زیید	!ین می‌گین	!(ی)ید می‌آیید	!ید/ !ین/ !(ی)ین میاین میایین، میاید	!(ی)ید می‌روید	!(ی)ین می‌رویین
۳جمع	!ند می‌خندند	!ن می‌گن	!(ی)ند می‌آیند	!ن میان	!(ی)ند می‌روند	!ین می‌روین

(۱). اومدش و بچه‌ها رو به صف کرد، بعدشم بردشون بیرون.

(۲). بچه‌های ما همش سراغ دختر کوچولوتونو می‌گیرن، فردا که میاییـن با خودتون بیارینـش.

(۳). خودش گفت می‌خواسته ماجرا رو بگه ولی اونا نذاشتن.

افزون بر پی‌بست‌های فاعلی و مفعولی پی‌بست‌های «و»، «م» نیز که صورت کوتاه‌شدهٔ حرف عطف «و» و حرف ربط «هم» است و نیز پی‌بست تأکید «ها» می‌توانند به ساخت افعال گونهٔ غیررسمی اضافه شوند (Megerdoomian 2008؛ شقاقی ۱۳۹۴) که در توزیع تکمیلی یکدیگر هستند.

(۴). فایده نداشت اگه می گفتیَم گوش نمی‌کرد.

۱. !: بدون هیچ فاصله‌ای به تک‌واژ قبل متصل می‌شود. ×: با نیم‌فاصله به تک‌واژ قبل متصل می‌شود.

(۵). اونم رفتو با یه مامور برگشت.

همچنین، در نوشتار گونهٔ غیررسمی از پی‌بست تأکیدی «ا» نیز استفاده می‌شود که با بررسی پیکرهٔ جمع‌آوری‌شده شناسایی و استخراج شد.

(۶). اگه حرف گوش نکنی نمی‌برمتا!

٤-۲. ساخت تصریف فعلی

به‌منظور طراحی و پیاده‌سازی تحلیلگر تصریفی فارسی معاصر قواعد مربوط به ساخت تصریفی افعال گونهٔ رسمی از منابع و پژوهش‌های انجام‌شده در این حوزه از جمله از «اسلامی» و همکاران (۱۳۸۳) و «انوری و گیوی» (۱۳۹۱) استخراج شد و مورد استفاده قرار گرفت. ساخت تصریفی فعل‌ها در گونهٔ غیررسمی نیز از الگوی ساخت فعل در گونهٔ رسمی تبعیت می‌کند، با این تفاوت که افعال غیررسمی امکان پذیرش طیف گسترده‌تری از پی‌بست‌ها را دارند که به‌عنوان قواعد افزوده به تحلیلگر ساخت‌واژی معرفی شده است. در ساخت فعل غیررسمی حداقل یک عنصر غیررسمی باید وجود داشته باشد که عبارت است از: بن غیررسمی فعل، شناسهٔ غیررسمی، پی‌بست مفعولی، پی‌بست فاعلی، تک‌واژهای محاوره (جدول ۶). به‌عنوان مثال، فعل «گفتیشان» گرچه از بن رسمی «گفت»، شناسهٔ رسمی «ی» و پی‌بست ضمیری رسمی «شان» تشکیل شده است، اما دارای ساختی غیررسمی است؛ زیرا فعل‌های رسمی پی‌بست مفعولی نمی‌پذیرند (Megerdoomian 2000, 2008). فعل «رفتش» نیز که از بن رسمی و شناسهٔ سوم شخص مفرد (Ø) تشکیل شده، به این دلیل که پی‌بست فاعلی «ش» پذیرفته است، یک فعل غیررسمی به حساب می‌آید (اسلامی و علی‌زاده لجمیری ۱۳۸۸). در این پژوهش با بررسی تحقیقی که به‌صورت ناتمام در این زمینه توسط Megerdoomian (2008) انجام شده، همین‌طور، منابع مشترک با فعل رسمی ودیگری پیکرهٔ جمع‌آوری‌شده، الگوی تصریفی افعال غیررسمی مطابق شکل ۲، استخراج و مورد استفاده قرار گرفت.

$$\text{ساختمان فعل رسمی} + \begin{bmatrix}(\text{پی‌بست مفعولی})\\(\text{پی‌بست فاعلی})\end{bmatrix} + (\text{اجزای ساختمان غیررسمی})$$

شکل ۲. تصریف افعال غیررسمی

فعل‌های گذرا می‌توانند پی‌بست مفعولی و یا پی‌بست فاعلی بپذیرند. در مقابل، فعل‌های ناگذر تنها می‌توانند پی‌بست فاعلی بپذیرند. در فعل‌های گذرا پذیرفتن پی‌بست فاعلی یا پی‌بست مفعولی سوم‌شخص مفرد هر دو منجر به افزوده شدن «ش» به انتهای بن فعل می‌شود که ابهام ایجاد می‌کند.

٤-٢-١. موارد استثنا در ساخت فعل امر

در چارچوب قاعدهٔ تصریفی فعل امر غیررسمی می‌توان چند مورد استثنا شناسایی و معرفی کرد:

- ◇ در ساخت فعل‌های غیررسمی این امکان وجود دارد که بن‌های رسمی یا غیررسمی با شناسه‌های رسمی یا غیررسمی ترکیب شوند؛ به‌عنوان مثال: «بگذرانید، بگذرانین، بگذرونید، بگذرونین». در ساخت فعل امر مفرد در صورتی که بن مضارع مختوم به واکه (رو، گو، شو، ده) باشد، از بن غیررسمی استفاده نمی‌شود.
(۷) مصدر گفتن- بن رسمی: گو «بگویید»، بن غیررسمی: گ «بگید» ساخت امر مفرد: «بگو»
- ◇ مصدر پیشوندی «وایسادن» نیز برای ساخت امری مفرد و جمع دو بن‌واژه متفاوت دارد، با این تفاوت که هر دو بن‌واژهٔ آن غیررسمی است. امر مفرد: «وایسا»، امرجمع: «وایسید».
- ◇ از بن مضارع گریستن (گری) فعل امر مفرد ساخته نمی‌شود و در حالت امر صرفاً از فعل مرکب استفاده می‌شود؛ به‌عنوان مثال: «گریه کن»

٤-٢-٢. فعل‌های ناقص

برخی بن فعل‌های غیررسمی مانند فعل‌های ناقص که همهٔ ساخت‌ها و زمان‌هایشان متداول نیست (انوری و گیوی ۱۳۹۱)، در همهٔ ساخت‌ها استفاده نمی‌شود و در صورتی می‌تواند در ساخت فعل مورد استفاده قرار گیرد که همراه با پیشوندهای تصریفی (می، ن) باشد، در غیر این صورت، بیشتر، از بن رسمی استفاده می‌شود. فعل‌های پیشوندی که با این بن‌ها ساخته می‌شوند نیز الگویی مشابه دارند (جدول ۳).

جدول ۳. فعل‌های ناقص ساخت‌های ماضی

	بن ناقص فعل ماضی	ذاشت (گذاشت)	شوند[1] (نشاند)	شست (نِشست)	واذاشت (واگذاشت)	فروشست (فرونشست)
ساخت مجاز	ساده منفی	نذاشتم	نَشوندی	نَشستم	وانَذاشت	فرونَشست
	استمراری	(ن)میذاشتم	(ن)میشوندی	(ن)میشستم	وا(ن)میذاشت	فرو(ن)میشِست
	نقلی مستمر	(ن)میذاشته	(ن)میشونده	(ن)میشسته	وا(ن)میذاشته	فرو(ن)میشِسته
	نقلی بعید التزامی-منفی	نذاشته	نَشونده	نَشسته	وانذاشته	فرونَشسته
ساخت غیرمجاز	ساده	ذاشتم*	شوندی*	شِستم*	واذاشتم*	فروشست*
	نقلی بعید التزامی-مثبت	ذاشته*	شونده*	شِسته*	واذاشته*	فروشسته*

٤-۲-۳. اتصال پی‌بست‌های ضمیری فعل مختوم به واکه

«مگردومیان» تغییرات آوایی و نگارشی را هنگام اتصال پی‌بست‌های ضمیری به افعال غیررسمی مضارع اخباری و ماضی نقلی سوم‌شخص مفرد بررسی کرده است (Megerdoomian 2008). با توجه به این‌که این افعال به واکه «ِ» (از لحاظ نگارشی «ه») ختم می‌شوند، در هنگام تلفظ، واکه به همخوان «ت» تبدیل می‌شود. این تغییر آوایی در صورت نگارشی، با جایگزینی «ه» با «ت» و یا درج «ت» در واژه نمایش داده می‌شود.
(۸). همه بچه‌ها رو صحیح و سالم با اتوبوس برده و اُوردِتشون/ اُورده‌تشون.

۵. قواعد و واژگان غیرفعلی

در این بخش به واژگان غیرفعلی تحلیلگر، شامل ستاک‌های غیرفعلی و وندهای تصریفی و پی‌بست‌های متصل به آن‌ها پرداخته و قواعد تصریفی آن بررسی می‌شود.

۵-۱. واژگان

واژگان در این بخش شامل ستاک‌های غیرفعلی، وندهای تصریفی و پی‌بست‌های رسمی و غیررسمی است که می‌توانند به هر ستاک متصل شوند. ستاک‌ها بر اساس دستور زبان فارسی «انوری و گیوی» (۱۳۹۱) دسته‌بندی و مقولات واژگانی اسم، قید،

۱. با توجه به داده‌های پیکره برای بن «نشاند» دو تک‌واژگونه غیررسمی «نشوند» و «شوند» شناسایی و به تحلیلگر معرفی شده است.

عـدد، حـرف ربـط، حـرف اضافـه، شـاخص، ضمیـر (مشـترک، شـخصی)، ضمیـر/ صفـت (اشـاره، مبهـم، پرسشـی، تعجبـی) متمایـز شـده‌اند. در بخـش ضمیـر/ صفـت در صورتـی کـه سـتاک به‌تنهایـی اسـتفاده شـود، صفـت و یـا ضمیـر اسـت (مسـتقل از بافـت قابـل تشـخیص نیسـت) و در غیـر ایـن صـورت، ضمیـر اسـت. ایـن نکتـه نیـز قابـل ذکـر اسـت کـه برچسـب‌دهی بـه مقـولات واژگانـی مسـتقل از بافـت تعریـف شـده اسـت. جـدول ۵، تعـداد سـتاک‌هایی را کـه ذیـل هـر مقولـهٔ واژگانـی بـرای تحلیلگـر تعریـف شـده، نشـان می‌دهـد.

ماننـد بخـش افعـال، در ایـن بخـش نیـز در صورتـی کـه واژه سـاختار یـا حداقـل یـک عنصـر غیررسـمی داشـته باشـد، به‌عنـوان واژهٔ غیررسـمی شناسـایی و پـردازش می‌شـود؛ به‌عنـوان مثـال: «کتابـا، جعبه‌مـون». همچنیـن، بـرای تمـام اسـامی، اطلاعاتـی در مـورد مفـرد یـا جمـع بـودن هـر اسـم، به‌عنـوان مثـال اسـامی جمـع مکسـر عربـی و نیـز وندهـای جمعـی کـه هـر اسـم می‌توانـد بپذیـرد، مثـلاً «ان، هـا» بـرای اسـامی جانـدار، وارد شـده اسـت.

جدول ٤. مقولات واژگانی غیرفعلی تعریف‌شده برای تحلیلگر تصریفی

مقوله واژگانی	اسم	اسم خاص	قید	صفت	عدد
تعداد رسمی	۲۷۹۴۰	۶۲۴۲	۱۳۷۴	۱۶۶۵۰	۴۸
غیررسمی	۸۰	۱۸	۲۲۷	۴۰	۱۷
مقوله واژگانی	صفت ضمیر	شاخص	ضمیر	حرف ربط	حرف اضافه
تعداد رسمی	۷۷	۲۹	۱۰	۳۰۲	۱۶۴
غیررسمی	۲۲	۳	۳	۲۶	۴

۵-۱-۱. تک‌واژهای جمع

در زبـان فارسـی اسـم‌ها بـا وندهـای تصریفـی «هـا»، «ان» و «گان (تک‌واژگونـه ان)» جمـع بسـته می‌شـود. افـزون بـر آن، وندهـای جمـع عربـی ماننـد «یـن، ون، ات (جـات)» در زبـان فارسـی رایـج و مـورد اسـتفاده اسـت. در سـاختار غیررسـمی نیـز افزون‌بـر وندهـای فـوق، از «ا» به‌تنهایـی اسـتفاده می‌شـود.

(۹). سلام دوستان! طرحای جدیدمون آماده‌س.

۵-۱-۲. تک‌واژهای نکره/ موصول، کسرهٔ اضافه و معرفه‌ساز

جدول۵. الگوهای نگارشی تک‌واژهای نکره/ موصول، کسرهٔ اضافه و معرفه‌ساز

ستاک	مختوم به همخوان	مختوم به واکه الف/ واو	مختوم به واکه ه	مختوم به واکه ی
تک‌واژ کسرهٔ اضافه	–	!ی بدگوی، هیولای	Ø / ×ی برده، برده‌ی	–
تک‌واژ نکره	!ی *کتابی*	!یی / !ئی / !عی *کتابایی، کدوئی، طلاعی، (یه) جوراعی، ایتالیاعی*	×ای / ×ئی / ×یی پرده‌ای، خنده‌ئی، گربه‌یی	×ای / ×ی / ×عی دوگانگی‌ای، اپیدمی‌ی، آگاهی‌عی
تک‌واژ معرفه‌ساز	!ه *کتابه*	!ئه / !عه *آب‌نمائه (عه)، تابلوئه (عه)*	!ه / ×هه / ×ئه / ×عه *گیرهه، پنبه‌هه، گشنه‌ئه، قراضه‌ئه*	!ه نظامیه

تک‌واژ معرفه‌ساز «ه» با اتصال به پایهٔ اسم یا صفتِ جایگزین اسم، آن را شناسه می‌کند و منحصر به گونهٔ غیررسمی است. جدول ۶، الگوهای نگارشی را برای تک‌واژهای نکره/ موصول، کسرهٔ اضافه و معرفه‌ساز در گونهٔ غیررسمی، با توجه به بافت حرفی که در آن قرار می‌گیرند، نشان می‌دهد.

۵-۱-۳. پی‌بست‌های ربطی

این پی‌بست ساخت اسنادی را به اسم‌ها و غالبِ اقسام دیگر واژه که در این بخش آمده، اضافه می‌کند. جدول ۶، الگوهای مختلف نگارشی را در بافت حرفی نشان می‌دهد.

جدول ٦. پی‌بست‌های ربطی[1]

ش.ش.	رسمی‌بودن	همخوان	واکه واو	واکه ه	واکه الف	واکه ی
مفرد ۱، ۲، ۳	رسمی	س !(م، ی، ست)	س !ی(م، ی) س !ست	س ×(ام، ای، ست)	س !ی(م،ی) س !ست	س [×ام، !م] س ×[ای، ی] س !ست
	غیررسمی	س !(م، ی، ه) *مریضم، مریضی، مریضه*	س![ام، م] س!ای س!ئه *بلـگوام، بلـگوم، بلـگوای، بلـگوئه،*	س ×م س ×یی س ×س *برده‌م، برده‌یی، برده‌س*	س![ام، م] س!ای س!س *هیولام، بالام، بالای، بالاس*	س [×ام، !م] س [×ای، ×ی] س !ست *نظامی‌ام، نظامیم، نظامی‌ای، نظامی‌ی نظامیست*
جمع ۱، ۲، ۳	رسمی	س ! (یم، ید، ند)	س !ی(یم، ید،ند)	س ×(ایم، اید، اند)	س !ی(یم، ید،ند)	س [×ایم، ×یم] س [×اید، !ید] س [×اند، !ند]
	غیررسمی	س ! (یم، ین، ن) *مریضیم، مریضین، مریضن*	س!ایم س![اید، این، یین] س![اند، ند، ن] *پرروایم، بلـگواید، بلـگواین، پرروییین، پرروانـد، پررونـد، پررون*	س ×ییم س ×[ییـد، این، یین] س ×[ند، ان، ن] *برده‌ییم، برده‌یید، برده‌این، برده‌یین، برده‌نـد، برده‌ان، برده‌ن*	س!ایم س![اید، یین] س![اند، ن] *تنهاایم، تنهاایـد، تنهایین، تنهاانـد، تنهان*	س ×ییم س ×[ییـد، این، یین] س [× ان، !ن] *نظامی‌ییم، نظامی‌ییـد، نظامی‌این، نظامی‌یین، نظامی‌ان، نظامین*

۵-۱-٤. پی‌بست‌های ضمیری (شخصی)

این پی‌بست‌ها می‌توانند در نقش ملکی یا مفعولی ظاهر شوند و نیز با اتصال به جزء غیرفعلی فعل مرکب نقش مطابقهٔ مفعولی یا فاعلی را به عهده گیرند (شقاقی ۱۳۹۴). از آنجا که این تمایز بیشتر در سطح واژه و تصریف قابل تمیز نیست، همهٔ آن‌ها در بخش واژه‌های غیرفعلی با عنوان پی‌بست‌های ضمیری (شخصی) نام‌گذاری شده‌اند.

۱. الگوهای نگارشی متفاوت برای یک تک‌واژ داخل کروشه آورده شده است.

جدول ۷. پی‌بست‌های ضمیری

شخص شمار	رسمی‌بودن	ستاک مختوم به همخوان	ستاک مختوم به واکه واو- واکه الف	ستاک مختوم به واکه ه	ستاک مختوم به واکه ی
۱،۲،۳ مفرد	رسمی	س!(م،ت،ش) *کتابم*	س!ی(م،ت،ش)	س ×ا(ام،ات،اش)	س!(م،ت،ش) *سبزیم*
	غ.رسمی		س!(م،ت،ش) *لبوش، کارفرمات*	س ×(م،ت،ش) *جعبه‌م*	
۱،۲،۳ جمع	رسمی	س!(مان،تان،شان)	س!ی(مان،تان،شان)	س ×ی(مان،تان،شان	س!(مان،تان،شان)
	غ.رسمی	س!(مون،تون،شون) *کتابشون*	س!(مان،تان،شان) س!(مون،تون،شون) *لبوشان، لبوتون، کتیراتان، کتیراشون*	س ×(مان،تان،شان) س ×(مون،تون،شون) *جعبه‌تان، جعبه‌شون*	س!(مون،تون،شون) *سبزیشون*

۵-۱-۵. سایر تک‌واژهای وابسته مختص گونهٔ غیررسمی

«شقاقی» پی‌بست‌های سادهٔ زبان فارسی را که صورت کوتاه‌شدهٔ تک‌واژهای مستقل «را، و، هم» است و بیشتر در محاوره و گفتار سریع و خودمانی به‌کار می‌رود، معرفی می‌کند (۱۳۹۴). این پی‌بست‌ها در گونهٔ نوشتاری غیررسمی نیز مطابق جدول ۸، به‌عنوان آخرین تک‌واژ به ساختار واژه افزوده می‌شود. افزون بر موارد ذکرشده، با توجه به بررسی داده‌ها در گونهٔ غیررسمی، پی‌بست‌های تأکیدی «ها، ا» که «شقاقی» تنها قابل اتصال به گروه فعلی می‌داند، می‌توانند به مقوله‌های غیرفعلی نیز افزوده شوند که در واژگان غیرفعلی تحلیلگر تعریف شده است.

(۱۰). آدم خیلی تَنبلی بود، تَنبَلا!، باورت نمیشه!، دست به سیاه سفید نمی‌زد.

جدول ۸. برخی تک‌واژهای وابسته مختص گونهٔ غیررسمی[۴۱]

ستاک مختوم به ...	ها (تاکید)	ا (تاکید)	و (عطف)	رو (را)	و (را)	رم (را+هم)	م (هم)
همخوان	س ×ها *کتاب‌ها*	س !ا *کتابا*	س !و *کتابو*	س !رو *کتابرو*	س !و *کتابو*	س !رم *کتابرم*	س !م *کتابم*
واکه واو	س !ها *لبوها*	س!ا *لبوا*	س !و *لبوو*	س !رو *لبورو*	س !و *لبوو*	س !رم *لبورم*	–
واکه ی	س !ها *نظامیها*	س !ا *نظامیا*	س !و نظامیو	س !رو *نظامیرو*	س !و *نظامیو*	س !رم *نظامیرم*	س !م *نظامیم*

۱. س: ستاک

ستاک مختوم به ...	ها (تاکید)	ا (تاکید)	و (عطف)	رو (را)	و (را)	رم (را+هم)	م (هم)
واکه الف	س !ها *کتاباها*	–	س !و *کتاباو*	س !رو *کتاباررو*	س !و *کتاباو*	س !رم *کتاباررم*	س !م *کتاباام*
واکه ه	س ×ها *پنبه‌ها*	–	س ×و *پنبه‌و*	س×رو *پنبه‌رو*	س×و *پنبه‌و*	س!رم *پنبه‌رم*	–

۵-۲. قواعد غیرفعلی

ساختار تصریفی مقوله‌های واژگانی یا به‌عبارتی نوع و ترتیب پیوستن وندهای تصریفی و پی‌بست‌ها به ستاک‌های غیرفعلی در گونهٔ رسمی زبان با توجه به پژوهش‌های پیشین، از جمله «اسلامی» و همکاران (۱۳۸۳) و «انوری و گیوی» (۱۳۹۱) تعیین شده و در تحلیلگر مورد استفاده قرار گرفته است. برای گونهٔ غیررسمی نیز تلاش شد با توجه به پیکرهٔ جمع‌آوری‌شده و سایر پژوهش‌های این حوزه از جمله Megerdoomian (2008) برای هر یک از مقوله‌های واژگانی قواعد جامع و مانع به‌گونه‌ای تعریف شود که هم تمام ساخت‌های ممکن برای هر مقوله را شامل شود و هم از تولید ساخت‌های نادرست و زاید جلوگیری کند. این قواعد به‌طور خلاصه در جدول‌های ۹ و ۱۰ آورده شده است. جدول ۹، ترتیب و امکان باهم‌آیی را برای وندها و پی‌بست‌های گوناگون نشان می‌دهد که در ۹ گروه تعریف شده است. جدول ۱۰، ساخت‌های ممکن برای هر مقولهٔ واژگانی را با استفاده از گروه‌های تک‌واژی معرفی‌شده در جدول ۹، توصیف می‌کند. در هر دو جدول تک‌واژ یا تک‌واژهایی که داخل پرانتز قرار گرفته‌اند، اختیاری‌اند و امکان حذف از ساختار واژه را دارند. در جدول ۹، تک‌واژ یا تک‌واژهایی که نسبت به هم در توزیع تکمیلی هستند، در ردیف‌های مجزا، تعریف شده و در جدول ۱۰ با نشانه «/» از هم جدا شده‌اند. به‌عنوان مثال، جدول ۱۰ در مقولهٔ «حرف اضافه» نشان می‌دهد که پی‌بست ضمیری می‌تواند به حرف اضافه متصل شود و بعد از آن پی‌بست ربطی و تأکیدی (ساخت ۲) قرار گیرد؛ «حواست بهشونه‌ها!». همچنین، این امکان وجود دارد که بعد از پی‌بست ضمیری، پی‌بست کوتاه‌شده «هم: م» قرارگیرد که با پی‌بست ربطی و تأکیدی در توزیع تکمیلی است؛ «حواست بهشونم (بهشون هم) باشه». همین‌طور در «تو خودت جیره‌خورشونی!» اسم بعد از همراه شدن با ساختار پی‌بست ضمیری (شون) با یک پی‌بست ربطی دوم شخص مفرد (ی) همراه می‌شود (جدول ۹، شماره ۹).

جدول ۹. الگوهای باهم‌آیی وندهای تصریفی و پی‌بست‌ها در گونهٔ غیررسمی

شماره	باهم‌آیی تک‌واژها	شماره	باهم‌آیی تک‌واژها
۱	را + (تاکید)	۲	(ربطی) + (تاکید)
۳	(را) + (هم)*	۴	نکره + (ربطی)
۵	را + (هم * / تاکید)	۶	ضمیری + (ربطی / را / هم)
۷	(تکواژ عالی‌ساز / تکواژ تفضیلی‌ساز)	۸	موصولی + (ربطی / را + (هم)) ... + که
۹	(جمع) + (((ربطی) + (تاکید) / را + (هم * / تاکید) / نکره + (ربطی) + (عطف / تاکید) / موصولی + ((ربطی / هم) + (ضمیری) + (را + (هم))) ..+که / ضمیری + ((ربطی) + (تاکید) / را + (تاکید) / هم / عطف) / عطف / اضافه) / معرفه + ((ربطی) + (تاکید) / (هم * / را) + تاکید))		

جدول ۱۰. الگوهای باهم‌آیی وندهای تصریفی و پی‌بست‌ها در ساختار مقوله‌های واژگانی غیرفعلی در گونهٔ غیررسمی

مقوله واژگانی	ساختار	مقوله واژگانی	ساختار
شخصی	شخصی + (۸/ ۵/ ۲/ عطف/ اضافه)	اسم	اسم + ۹
پرسش	پرسش + جمع + (۶/ ۴/ ۳/ ۲)	صفت	صفت + ۷ + ۹
مبهم	مبهم + ضمیری + (۲ /۱ / هم)	عدد	عدد + تا + (جمع)
	مبهم + (۵/ ۲/ عطف/ اضافه)		عدد+ (ترتیبی۱+ (ترتیبی۲))+ (جمع)
مشترک	مشترک + (ضمیری) + (۲/ ۱/ هم/ عطف)	مصدر	(پیشوند)+ (منفی)+بن ماضی+ن+۹
حرف اضافه	حرف اضافه + (ضمیری) + (۲/ هم)	صفت مفعولی	(پیشوند)+(منفی)+صفت مفعولی+۹
اشاره[1]	اشاره + ۹		

٦. ارزیابی

به‌منظور ارزیابی تحلیلگر تصریفی فارسی معاصر، از حدود ۳۰ درصد پیکرهٔ فارسی معاصر که در بخش ۳، توضیح داده شد، استفاده شده است. به این منظور، جملاتی از هر سیاق زبان فارسی غیررسمی به‌صورت تصادفی به‌گونه‌ای انتخاب شد که نسبت تنوع گونه و سیاق در آن‌ها حفظ شود و خطای سوگیری و خطای تصادفی کنترل‌شده به حداقل برسد (Biber 1993a/1993b). با توجه به این‌که تحلیل واژه‌ها در تحلیلگر، مستقل از بافت انجام می‌شود، نخست، واژه‌های جملات به‌صورت دستی جداسازی و استخراج شده، ۱۷۸۶ واژهٔ یکتای به‌دست‌آمده برای آزمون تحلیلگر مورد استفاده قرار گرفت. در ارزیابی رویکرد مستقل از بافت، توانایی تحلیلگر در تولید درست تمام قواعد ممکن برای هر واژه سنجیده می‌شود. هر یک از این قواعد تعریف‌کنندهٔ تایپ[2] خاصی از واژه است که در بافت مخصوص به خود معنا پیدا کرده و استفاده می‌شود. هر تایپ قاعدهٔ سازندهٔ مختص به خود دارد. به‌طور مثال، برچسب‌هایی که در پیکرهٔ متنی زبان فارسی (Bijankhan et. Al. 2011) استفاده شده، مختص به قاعدهٔ سازندهٔ (تایپ) آن کلمه در بافتی خاص است، اما قاعده(های) سازنده‌ای که تحلیلگر در خروجی خود تولید می‌کند، مربوط به تمام بافت‌هایی است که کلمه می‌تواند در آن‌ها ظاهر شده و به نسبت آن

۱. اشاره شامل ساختار (کسره) اضافه و معرفه نمی‌شود.

2. type

معانی مختلف داشته باشد. به‌عبارت دیگر، تمام تایپ‌های کلمه پوشش داده می‌شود. جدول ۱۱، خروجی تحلیلگر را برای ورودی «مردم» نشان می‌دهد. تمام تایپ‌های ممکن این ورودی در بافت‌های مختلف در میان خروجی آن قرار دارد. از میان خروجی‌ها تایپ اول و سوم غیررسمی و بقیه رسمی‌اند.

جدول ۱۱. نمونهٔ خروجی تحلیلگر تصریفی برای ورودی «مردم»

ورودی	خروجی	مثال تایپ در بافت	برچسب اجزای سخن تایپ
مردم	<ف.م.س=مرد+ش۳+هم>	اگه مردم (مُرد هم) جوابشو نده.	فعل ماضی ساده
	<ف.م.س=مرد+ش۱+رسمی>	اگه مردم (من مُردم) سر خاکم بیا.	فعل ماضی ساده
	<اسمعام=مرد+هم>	مردم (مَرد هم) مردای قدیم.	اسم عام
	<اسمعام=مرد+وشخصی۱+رسمی>	مردم (مَرد من) غیرتیه.	اسم عام
	<اسمعام=مرد+وربطی۱+رسمی>	صدام نشون میده که مردم (مَرد هستم).	اسم عام
	<اسمعام=مردم+رسمی>	مَردُم جمع شده بودن.	اسم عام

در این رویکرد، مثبت‌های درست[1] قاعده‌هایی هستند که توسط تحلیلگر به‌درستی تولید شده‌اند. مثبت‌های اشتباه[2] قاعده‌هایی هستند که به اشتباه تولید شده و منفی اشتباه[3] قاعده‌هایی هستند که می‌بایست توسط تحلیلگر تولید می‌شده، اما تولید نشده‌اند. فراخوانی[4] نشان‌دهندهٔ میزان قاعده‌های درست شناخته‌شده برای قطعات است و از تقسیم TP بر جمع TP و FN به‌دست می‌آید. صحت[5] نشان‌دهندهٔ میزان درست بودن قواعدی است که توسط تحلیلگر تولید شده‌اند و از تقسیم TP بر جمع TP و FP به‌دست می‌آید. معیار اف[6] نیز از ترکیب دو معیار فراخوانی و صحت به شکل متوازن به‌دست می‌آید، به‌طوری که می‌توان میزان تأثیر هر یک از دو معیار را مشخص کرد. رابطهٔ (۱) توزیع مساوی (۵۰ درصد– ۵۰ درصد) هر دو معیار را به‌دست می‌دهد.

$$\text{معیار اف} = ۲ \times \frac{\text{درستی} \times \text{فراخوانی}}{\text{درستی} + \text{فراخوانی}} \quad (۱)$$

جدول ۱۲، نتایج ارزیابی تحلیلگر را با استفاده از سه معیار فراخوانی، صحت و معیار اف برای واژه‌های رسمی و غیررسمی آزمون نشان می‌دهد.

1. true positive (TP)
2. false positive (FP)
3. false negative (FN)
4. recall
5. precision
6. f-measure

جدول ۱۲. نتایج ارزیابی تحلیلگر برای گونهٔ رسمی، غیررسمی و مجموع هر دو گونه

گونهٔ زبانی	فراخوانی	صحت	معیار اف
رسمی	۹۵/۱	۹۹/۴۱	۹۷/۲۱
غیررسمی	۹۶/۲۳	۹۹/۱۶	۹۷/۶۷
معاصر (مجموع)	۹۵/۵۶	۹۹/۶۵	۹۷/۵۶

تحلیلگر تصریفی برای مجموع ۱۷۸۶ واژهٔ یکتای آزمون، ۳۷۰۴ قاعده یا تایپ را با موفقیت (TP) تولید کرده و همین‌طور، ۱۷۲ قاعده را می‌بایست تولید می‌کرده، اما از دست داده (FN) است و ۱۳ قاعده را هم به‌اشتباه (FP) تولید کرده است.

جدول ۱۳[1]، بررسی موردی خطای FN را به تفکیک برای دو گونهٔ رسمی و غیررسمی نشان می‌دهد. همان‌طور که مشاهده می‌شود، کمبود واژگان با وزن ۷۱/۵ درصد، مؤثرترین عامل در رخداد خطای FN در بین همهٔ تایپ‌هاست. رتبه‌های بعدی شامل خطای املایی (۲۵ مورد، ۱۴/۵ درصد کل عوامل) و نقص در قاعده (۱۰ مورد، ۵/۸ درصد) است.

جدول ۱۳. بررسی خطای FN برای هر دو گونهٔ زبانی رسمی و غیررسمی

گونهٔ زبانی	کمبود واژگان	نقص در قاعده	خطای هم‌صدا و کسرهٔ اضافه	به‌هم‌چسبیده	تغییر آوایی	خطای املایی
رسمی	۸۳	۶	۰	۱	۰	۱۷
غیررسمی	۴۰	۴	۳	۳	۵	۸

به‌علت نمونه‌گیری تصادفی در جمع‌آوری پیکره و دادهٔ آزمون، نسبت واژه‌های رسمی آزمون به واژه‌های غیررسمی در حدود ۶۰ به ۴۰ است. از طرف دیگر، دایرهٔ واژه‌های رسمی گسترده‌تر از واژه‌های غیررسمی است و این منجر به تولید واژه‌های خارج از واژگان بیشتر رسمی نسبت به گونهٔ غیررسمی می‌شود (طبق جدول ۱۳، دو برابر؛ ۸۳ به ۴۰). واژه‌های خارج از واژگان فراخوانی تایپ‌های رسمی را بیش از غیررسمی کاهش داده است. از طرف دیگر، ساخت‌های ساده‌تر رسمی صحت تایپ‌های رسمی را نسبت به غیررسمی افزایش داده است.

در این پژوهش جداسازی واژه‌های آزمون به‌صورت دستی انجام شده است. پیش‌بینی می‌شود اگر این جداسازی به شکل خودکار انجام شود، فراخوانی را به‌ویژه برای واژه‌های غیررسمی کاهش دهد.

۷. نتیجه‌گیری

تحلیلگر تصریفی پایین‌ترین سطح پردازش را (بعد از جداساز) بر روی متن فارسی انجام می‌دهد و اطلاعات تصریفی را به قطعه‌های متن خام اضافه می‌کند. اطلاعات به‌دست‌آمده از تحلیلگر تصریفی را می‌توان در سطوح بالاتر تحلیل نحوی و شناسایی گروه‌های نحوی استفاده کرد. با توجه به این‌که واژه‌های رسمی بخش قابل توجهی از واژه‌های فارسی امروز را تشکیل می‌دهد، تحلیل تصریفی این واژه‌ها نیز در کنار واژه‌های فارسی غیررسمی بخشی از فرایند پردازشی در تحلیلگر تصریفی فارسی معاصر است. برای تحلیل تصریفی همهٔ این نوع واژگان فارسی از مبدل حالت محدود استفاده شده است.

به‌دلیل رویکرد مستقل از بافت تحلیلگر، برای هر واژه تمام تحلیل‌ها و قواعد احتمالی مربوط به خوانش و تحلیل معنایی، وابسته به بافت‌های گوناگون ارائه می‌شود. با افزودن یک مدل آماری برچسب‌زن صرفی-نحوی[1] می‌توان این ابزار را به یک تحلیلگر تصریفی مبتنی بر بافت تبدیل کرد. در مرحلهٔ شناسایی قاعدهٔ دقیق از بین قاعده‌های تولیدی برای هر واژه، جدا از روش‌های آماری و مبتنی بر بافت، می‌توان از اطلاعات افزوده‌ای که در تصریف غیررسمی نسبت به تصریف رسمی وجود دارد، کمک گرفت. وندها و واژه‌بست‌هایی که امکان اتصال به ستاک در ساختار غیررسمی را دارند، به‌دلیل تنوع بالاتر دربردارندهٔ اطلاعات صرفی-نحوی بیشتری هستند که می‌توانند در تجزیه و تحلیل نحوی و معنایی مورد استفاده قرار گیرد. همچنین، خروجی این تحلیلگر به‌گونه‌ای طراحی شده است که بتوان ابزارهای مکمل و ساده‌ای مانند ریشه‌یاب و ابزار تقطیع ساخت‌واژی بر اساس آن ساخت تا در تحلیل زبانی به‌کار آید. از جمله کاربردهای دیگر این ابزار می‌توان به آموزش فارسی به غیرفارسی زبانان اشاره کرد. با توجه به پوشش فارسی غیررسمی، آموزش ساختار تصریفی کلمات گفتار به زبان‌آموزان فارسی و تمرین و تکرار آن با این ابزار توسط زبان‌آموز امکان‌پذیر است.

بر اساس جدول ۱۲، بیشترین منفی‌های اشتباه در فارسی رسمی و در فارسی غیررسمی مربوط به کمبود واژگان است. با افزودن واژگان مناسب تا اندازهٔ قابل توجهی فراخوانی تحلیلگر افزایش خواهد یافت. بعد از نقص کمبود واژگان بیشترین تعداد منفی اشتباه مربوط به خطاهای املائی است. در صورت استفاده از یک خطایاب و اصلاح‌کنندهٔ خطای املایی در کنار تحلیلگر تصریفی، فراخوانی به‌ویژه برای گونهٔ رسمی افزایش خواهد یافت.

1. morpho-syntactic part-of-speech tagger

فهرست منابع

اسلامی، محرم، مسعود آتشگاه، صدیقه علی‌زاده لمجیری، و طاهره زندی. ۱۳۸۳. واژگان زایای زبان فارسی. اولین کارگاه پژوهشی زبان فارسی و رایانه. تهران.

اسلامی، محرم، و صدیقه علی‌زاده لجمیری. ۱۳۸۸. ساختار تصریفی کلمه در زبان فارسی. زبان و ادب فارسی. *نشریه دانشکدهٔ ادبیات و علوم انسانی دانشگاه تبریز* ۲۱۱: ۱-۱۸.

انوری، حسن، و حسن گیوی. ۱۳۹۱. *دستور زبان فارسی*. ویرایش چهارم. تهران: انتشارات فاطمی.

بی‌جن‌خان، محمود. ۱۳۸۶. *پیاده‌سازی استاندارد ایگلز در پیکرهٔ متنی زبان فارسی معاصر. مطالعه و تحقیق جهت تدوین پژوهشنامهٔ عملیاتی دادگان*. تهران: دبیرخانه شورای عالی اطلاع‌رسانی.

تازه‌جانی، سمیه، و محمد بحرانی. ۱۳۹۲. *بررسی روند تغییرات در تبدیل فرم رسمی افعال به فرم محاوره‌ای و ارائهٔ یک تحلیلگر صرفی برای افعال محاوره‌ای*. اولین هم‌اندیشی زبان فارسی و اینترنت. تهران.

حیدرپور، داود، مصطفی مصالحی، محمود بی‌جن‌خان، هادی ویسی، و وحید رنجبر. ۱۳۹۸. *شناسایی و پوشش واحدهای خارج از واژگان در فارسی غیررسمی*. پنجمین همایش ملی زبان‌شناسی رایانشی. انجمن زبان‌شناسی ایران، تهران.

شقاقی، ویدا. ۱۳۹۴. ونـدِ گروهی. زبان و زبان‌شناسی. *مجلهٔ انجمن زبان‌شناسی ایران، پژوهشگاه علوم انسانی و مطالعات فرهنگی* ۹ (۱۷): ۱-۲۶.

صادقی، علی‌اشرف، و زهرا زندی‌مقدم. ۱۳۸۵. *فرهنگ املایی زبان فارسی*. تهران: فرهنگستان زبان و ادب فارسی.

طیب‌زاده، امید. ۱۳۹۸. *مبانی و دستور خط فارسی شکسته*. تهران: پژوهشگاه علوم انسانی و مطالعات فرهنگی.

کاشفی، امید، میترا نصری، و کامیار کنعانی. ۱۳۸۹. *خطایابی املایی خودکار در زبان فارسی*. تهران: دبیرخانهٔ شورای عالی اطلاع‌رسانی.

مواجی، وحید، محرم اسلامی، و بهرام وزیرنژاد. ۱۳۹۰. پارس مورف: تحلیلگر صرفی زبان فارسی. *پردازش علائم و داده‌ها* ۸ (۱): ۳-۸.

References

Alshargi, Faisal, and Owen Rambow. 2016. Morphologically Annotated Corpora and Morphological Analyzers for Moroccan and Sanaani Yemeni Arabic. In 10th Language Resources and Evaluation Conference (LREC 2016). Portoroz.

Badawi, El Said, Michael Carter, and Adrian Gully. 2013. *Modern Written Arabic: A Comprehensive Grammar*. London: Routledge.

Biber, Douglas. 1993a. Representativeness in Corpus Design. *Literary and Linguistic Computing* 8 (4): 57-243.

Biber, Douglas. 1993b. Using Register-Diversified Corpora for General Language Studies. *Computational Linguistics* 19 (2): 41-219.

Bijankhan, Mahmood, Javad Sheykhzadegan, Mohammad Bahrani, and Masood Ghayoomi. 2011. Lessons from building a Persian written corpus: Peykare. *Language Resources and Evaluation* 45 (2): 143–164.

Ferguson, Charles A. 1959. Diglossia. *Word* 15 (2): 325–340.

Hulden, Mans. 2009. Foma: A Finite-State Compiler and Library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*: 29–32. Association for Computational Linguistics. Athens.

Jurafsky, Daniel, and James H Martin. 2008. *Speech and Language Processing*. 2nd Edition. New Jercsey: Prentice Hall.

Lazard, G. 1992. *A grammar of contemporary Persian*. Costa Mesa, CA: Mazda Publishers.

McEnery, Tony, and Andrew Hardie. 2011. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Megerdoomian, Karine. 2008. *Analysis of Farsi Weblogs*. Washington DC: MITRE Corporation,.

_____. 2000. *Persian Computational Morphology: A Unification-Based Approach*. New Mexico: Computing Research Laboratory, New Mexico State University.

Ryding, Karin C. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge: Cambridge university press.

Sagot, Benoit, and Geraldine Walther. 2010. A morphological lexian for the Persian language. In Proceedings of the 7th Language Resources and Evaluation Conference (LREC'10). Malta.

Shamsfard, M., H. S. Jafari, and M. Ilbeygi. 2010. *STeP-1: A Set of Fundamental Tools for Persian Text Processing*. In LREC 2010-8th Language Resources and Evaluation Conference. Malta.

Torjmen, R. and K. Haddar. 2018. *Morphological analyzer for the Tunisian dialect*. In International Conference on Text, Speech, and Dialogue (pp. 180-187). Springer, Cham. Czech Republic.

داود حیدرپور

متولد سال ۱۳۶۲، دانش‌آموختهٔ رشتهٔ نرم‌افزار و مترجمی زبان انگلیسی در مقطع کارشناسی و رشتهٔ زبان‌شناسی رایانشی در مقطع کارشناسی ارشد از دانشگاه تهران است.

از جمله موضوعات تحقیقی مورد علاقهٔ وی پردازش زبان طبیعی، متن‌کاوی و نظرکاوی است.

الهام‌سادات سبط

دارای مدرک کارشناسی فیزیک مهندسی و کارشناسی ارشد زبان‌شناسی رایانشی از دانشگاه تهران است.

از جمله موضوعات تحقیقی مورد علاقهٔ وی پردازش نحوی و ساخت‌واژی، نرمال‌سازی متن، طراحی و تهیهٔ دادهٔ زبانی و مدل‌های ماتریسی در تحلیل نحوی (دستور ماتریسی) است.

محمود بی‌جن‌خان

متولد سال ۱۳۳۷، دارای مدرک تحصیلی دکتری در رشتهٔ زبان‌شناسی از دانشگاه تهران است. ایشان هم‌اکنون استاد تمام گروه زبان‌شناسی دانشگاه تهران است.

آواشناسی، واج‌شناسی، زبان‌شناسی پیکره‌ای، تولید پیکره‌های زبانی برای آموزش و ارزیابی سامانه‌های فناوری زبان از جمله علایق پژوهشی وی است.

مصطفی صالحی

دارای مدرک دکتری در رشتهٔ مهندسی کامپیوتر از دانشگاه صنعتی شریف است. ایشان هم‌اکنون عضو هیئت علمی دانشکدهٔ علوم و فنون نوین دانشگاه تهران است.

پایش رسانه‌های اجتماعی از جمله علایق پژوهشی وی است.

هادی ویسی

دارای مدرک دکتری مهندسی کامپیوتر از دانشگاه صنعتی شریف است. وی هم‌اکنون عضو هیئت علمی دانشکدهٔ علوم و فنون نوین دانشگاه تهران است.

هوش مصنوعی و زبان‌شناسی رایانشی از جمله علایق پژوهشی وی است.