

Comparative Analysis of Editorial Tools Using a Developed Benchmark Corpus Based on the Approved Persian Orthography Rules

Masood Ghayoomi¹ 

Faculty of Linguistics, Institute for
Humanities and Cultural Studies,
Tehran, Iran

Laily Nezami² 

Research Institute for Terminology
Studies, Academy of Persian Language
and Literature, Tehran, Iran

Abstract

The development of information technology and its integration with natural language have made it possible to understand and generate linguistic content in an algorithmic manner. The existence of different research groups and their usage of different data and processing algorithms have made the results obtained from the tools not comparable. To address this shortcoming, in this article an attempt has been made to develop a benchmark corpus that is based on the approved Persian orthography rules to be used for evaluating the performance of Persian editing tools. In the current research, after examining the approved Persian orthography rules, the proposed rules are divided into eight main categories. Then, a dataset of about 98,000 words is collected and compared with the rules in two genres, namely scientific and news. A subset of the words of this data is marked based on the eight categories; and the evaluation of the models is limited to these marked words only. In the current study, the performance of six editing tools, namely VirāVirāst, FārsiYār, Virāstman, Pāknevis, Virāstyār and Gāgool, is compared. Comparing the performance of the tools with the gold standard data, it is concluded that Paknevis has the highest performance with an average of 74.07% in the two scientific and news genres, and Gagool has the lowest performance with an average of 24.72% based on the approved Persian orthography rules. Finally, the performance of the tools is compared

1. m.ghayoomi@ihcs.ac.ir (Corresponding Author)

2. lilinezami14@gmail.com

with each other in the eight categories to determine the strengths and weaknesses of the tools.

Keywords: Benchmark corpus, Persian language, approved Persian orthography rules, editing tool, linguistic corpus



تحلیل مقایسه‌ای ابزارهای ویراستاری با استفاده از پیکره سنج تهیه شده منطبق با دستور مصوب خط فارسی

مسعود قیومی^۱ ID پژوهشکده زبان‌شناسی، پژوهشگاه علوم انسانی و مطالعات فرهنگی، تهران، ایران

لیلی نظامی^۲ ID پژوهشکده مطالعات واژه‌گزینی، فرهنگستان زبان و ادب فارسی، تهران، ایران

چکیده

توسعه فناوری اطلاعات و عجز شدن آن با زبان طبیعی موجب شده است درک و تولید محتوای زبانی به صورت الگوریتمی میسر گردد. وجود گروه‌های پژوهشی مختلف و بهره‌گیری آنان از داده‌ها و الگوریتم‌های پردازشی مختلف، سبب مقایسه‌پذیر نبودن نتایج حاصل از عملکرد ابزارها شده است. برای رفع این کاستی، در مقاله حاضر تلاش شده است پیکره‌ای سنج که منطبق با دستور مصوب خط فارسی است تهیه شود و از این پیکره برای ارزیابی توانایی ابزارهای ویراستاری فارسی استفاده گردد. در پژوهش حاضر، پس از بررسی دستور خط فارسی مصوب فرهنگستان زبان و ادب فارسی، قواعد تعریف شده در دستور خط به هشت مقوله اصلی تقسیم شد. سپس در دو ژانر علمی و خبری، مجموعه داده‌ای با حدود ۹۸/۰۰۰ واژه، گردآوری و با دستور مصوب خط انطباق داده شد. تعدادی از واژه‌های این پیکره براساس مقولات هشت‌گانه نشان‌دار شده و ارزیابی مدل‌ها فقط به واژه‌های نشان‌دار شده محدود شد. در پژوهش حاضر عملکرد شش ابزار ویراویراست، فارسی‌یار، ویراست‌من، پاک‌نویس، ویراستیار و گاکول مقایسه شد. از مقایسه عملکرد ابزارها با داده، معیار طلایی این نتیجه به دست آمده است که ابزار پاک‌نویس با میانگین ۷۴/۰۷ درصد در دو ژانر علمی و خبری، بالاترین کارایی و گاکول با میانگین ۲۴/۷۲ درصد، پایین‌ترین

۴ | نام مجله | سال ؟ | شماره ؟ | فصل سال (دولت پژوهی | سال اول | شماره ۴ | زمستان ۱۳۹۵)

میزان عملکرد را در ویراستاری فارسی منطبق با دستور مصوب خط فارسی به دست آورد. در انتها، عملکرد ابزارها در مقولات هشت گانه با یکدیگر مقایسه شد تا نقاط قوت و ضعف ابزارها مشخص شود.

کلیدواژه‌ها: پیکره سنج، زبان فارسی، دستور مصوب خط فارسی، ابزار ویراستاری، پیکره زبانی

۱. مقدمه

امروزه توسعه فناوری اطلاعات و به کارگیری آن در حوزه‌های مختلف سبب شده امکانات جدید و متنوعی در اختیار بشر قرار گیرد و استفاده از این دسته از فناوری‌ها، زمینه را برای رشد و توسعه آنها فراهم آورد. یکی از این فناوری‌ها که با هوش مصنوعی نیز عجین شده و تحولات عظیمی را به همراه داشته فناوری‌های زبانی است. در این دسته از روش‌های رایانشی برای درک و تولید الگوریتمی محتوای زبانی استفاده می‌شود. این نوع فناوری، طیف بسیار متنوعی از فعالیت‌ها را شامل می‌شود. یکی از این رویکردها مربوط به حوزه‌های مختلف زبان‌شناسی است که به بررسی مسائل صرفاً زبانی در حوزه‌های آواشناسی، صرف، نحو، معناشناسی و کاربردشناسی می‌پردازد و رویکرد دیگر، بر جنبه‌های کاربردی فناوریانه دانش زبان‌شناسی، مانند ترجمه ماشینی^۱، سامانه‌های پرسش و پاسخ، غلطیاب خودکار املائی و دستوری، ویراستاری و غیره، متمرکز است. بهره‌گیری از الگوریتم‌های متنوع پردازش زبانی سبب می‌شود ضمن ایجاد و توسعه ابزارهای مبتنی بر فناوری‌های زبان، نظارتی بر عملکرد این ابزارها صورت گیرد و چالش‌های هر کدام مطرح گردد.

مقایسه عملکرد ابزارهای متنوع زبانی با استفاده از مجموعه داده یکسان سبب می‌شود زمینه برای همکاری و رقابت بین گروه‌های پژوهشی فراهم بیاید و با استفاده از معیاری مشخص، نقاط ضعف و قوت هر یک شناسایی و به توسعه آنها پرداخته شود. بنابراین، امروزه علاوه بر تهیه پیکره‌های معیار مرسوم که معمولاً برای ارزیابی کارایی الگوریتم‌ها و مدل‌های زبانی استفاده می‌شود، نتایج عملکرد آنها نیز اطلاعات مفیدی در اختیار پژوهشگران قرار می‌دهد. پوشیده نیست که زبان فارسی در مقایسه با زبان‌هایی مانند انگلیسی، منابع زبانی محدودی دارد (Shamsfard, 2019) و تنوع پیکره‌های زبانی آن برای آموزش و ارزیابی داده، چندان زیاد و متنوع نیست. نبود پیکره‌های متنوع سنجه^۲ به زبان فارسی در حوزه‌های مختلف فناوری زبانی مسأله‌ای است که مقایسه عملکرد الگوریتم‌های پیشنهادی یا ابزارها را با چالش مواجه می‌کند، زیرا در هر پژوهش، از پیکره‌ای خودساخته استفاده می‌شود که برای زبان فارسی چنین پیکره‌ای در دسترس نیست. این امر سبب می‌شود نتوان ارزیابی دقیقی از عملکرد مدل‌ها داشت. قیومی (۱۴۰۱)

1. machine translation

2. benchmark

در مقایسه پژوهش‌های مختلف درباره ابزارهای پایه پردازش فارسی به این نکات اشاره کرده است.

در پژوهش حاضر، دو مسأله مطرح و بررسی می‌شود: الف) معرفی و بررسی ویژگی‌های پیکره سنج‌ای که براساس دستور خط مصوب فرهنگستان زبان و ادب فارسی تهیه شده است. ب) ارزیابی مقایسه‌ای شش ابزار ویراستاری خط فارسی با این پیکره سنج که گروه‌های پژوهشی شرکت‌ها و سازمان‌های مختلف آن را تهیه کرده‌اند.

۲. پیشینه پژوهش

پژوهش حاضر، با سه موضوع تهیه دستور خط، ابزار پردازشی و تهیه پیکره سنج در ارتباط است. در ادامه این بخش، به صورت گذرا به مطالعات انجام‌شده در مورد دستور خط و همچنین، پردازش خط و چالش‌های ممکن پرداخته شده است. موضوع پیکره سنج بسیار متنوع است و مطالعات انجام‌شده در خصوص معیارسازی خط، در زیربخش مربوط به دستور خط توضیح داده خواهد شد.

۲-۱. دستور خط

تهیه دستور خط، فعالیت است که از دیرباز برای زبان‌های مختلف، از جمله فرانسوی و کره‌ای، انجام شده است. برای نمونه، از قرن هفدهم تا نوزدهم، دانشمندان تلاش‌هایی برای اصلاح خط فرانسوی پیشنهاد کردند که همگی تلاش‌ها به دلیل بی‌تفاوتی یا تمسخر خود دانشمندان و دستورنویسان با شکست مواجه شد. نمونه این تلاش‌ها توسط رگنیه - دمارائیس^۱ (1706) انجام پذیرفت که به انتشار دستور زبان معروف او منجر شد. تیمونیه^۲ (1967) برای اولین بار در اثری به نام «نظام گرافیکی زبان فرانسوی»، روش اصلاح هوشمندانه و منسجمی ارائه داد تا براساس آن، دانشگاهیان و معلمان قادر باشند املاي واژه‌های فرانسوی را اصلاح کنند و آموزش دهند (Rat, 1968). زبان کره‌ای تا قرن ۱۵ به خط چینی نوشته می‌شد، اما به مرور الفبای کره‌ای به نام هنگول^۳ در نوشتار به دستور شاه سجان^۴؛ چهارمین پادشاه سلسله جوسان^۵، اختراع و جایگزین شد (صفارمقدم، ۱۳۷۶).

1. Régnier-Desmarais, F. S.

2. Thimonnier, R.

3. Hangul

4. Sejong the Great

5. Joseon

آلبرتس^۱ (2013) دستورالعملی برای شیوه نگارش زبان‌های آفریقایی جنوبی ارائه کرده است. کمیسیون ملی املای اوکراین در سال ۲۰۱۶ به تهیه سندی برای شیوه نگارش واژه‌های اوکراینی اقدام کرده است.^۲ درمورد فارسی، فرهنگستان زبان و ادب فارسی در سال ۱۳۸۱ اقدام به تهیه دستور خط کرد و سپس آن را بازبینی کرد (فرهنگستان زبان و ادب فارسی، ۱۳۸۱؛ ۱۴۰۲). یکی دیگر از تلاش‌های انجام‌شده برای معیارسازی شیوه نگارش واژه‌ها، تدوین فرهنگ املایی برای فارسی است که صادقی و زندی‌مقدم (۱۳۸۵) به رشته تحریر درآوردند و حاوی نگارش معیار بیش از ۳۶ هزار بن‌واژه است.

آلبرتس (2006) به مواردی اشاره کرده که لازم است هنگام تدوین دستور خط به آنها توجه شود. این موارد عبارت‌اند از: قواعد املایی مربوط به واژه‌های مرکب، درج خط‌تیره میان واژه‌ها، قواعد املایی برای واژه‌ها و همخوان‌ها، قواعد املایی برای اسامی خاص، هنجارهای درج علائم سجاوندی، هنجارهای اختصارسازی، تهیه فهرست اختصارات و واژه‌های اختصاری و فهرست واژه‌های دشوار. نگارش تمامی موارد ذکرشده، به دستورالعملی واحد نیاز دارد تا ضمن جلوگیری از تشتت نگارش واژه‌ها، سادگی خواندن و نوشتن را به همراه داشته باشد.

۲-۲. مسائل خط فارسی و ابزارهای معیارسازی

اگرچه دستور خط فارسی سبب ایجاد یکدستی در نظام نوشتار می‌شود، اما همچنان در خط فارسی مشکلاتی به چشم می‌آید. پژوهشگرانی چون قیومی و ممتازی^۳ (2009)، قیومی و همکاران (2010)، شمس‌فرد^۴ (2011)، اسلامی (۱۳۸۱)، ستوده و هنرجویان (۱۳۹۱؛ ۱۳۹۳) و قیومی (2025) به شماری از دشواری‌هایی اشاره کرده‌اند که هنگام تهیه پیکره زبان فارسی با آن مواجه می‌شویم و گاه راهکارهایی را برای معضلات پیشنهاد داده‌اند. اهم این مسائل عبارت‌اند از: کدگذاری حروف، تنوع نگارشی واژه‌ها، واحدسازی صحیح واژه‌های فارسی، تحلیل ساختوازی و واژه‌سازی مربوط به نواژه‌ها و واژه‌های

1. Alberts, M.

2.

<https://mon.gov.ua/storage/app/media/news/%D0%9D%D0%BE%D0%B2%D0%B8%D0%BD%D0%B8/2018/08/22/UkrainianSpelling.pdf>

3. Ghayoomi, M., & Momtazi, S.

4. Shamsfard, M.

مرکب و نگارش فارسی واژه‌های لاتین. امروزه، ابزارها یا افزونه‌هایی تهیه شده که می‌توانند در راه نیل به اهداف موردنظر در تعریف دستور خط مفید باشند.

کاشفی (۱۳۹۰) ابزاری به نام ویراستیار^۱ تهیه کرده که در قالب افزونه^۲ نرم‌افزار «مایکروسافت ورد»^۳، برای اصلاح خطاهای املائی، اشتباهات ویرایشی و نشانه‌گذاری، و همچنین معیارسازی متون فارسی به کار می‌رود. امکانات این ابزار در بخش ۵-۱ توضیح داده خواهد شد.

بگ‌جانی (۱۳۹۹) در پژوهش خود ضمن معرفی ابزار ویراستاری ویراستیار به توضیح درمورد امکانات این نرم‌افزار براساس مصوبات فرهنگستان زبان و ادب فارسی پرداخته است. در این پژوهش، وی ابتدا بخش‌های تشکیل‌دهنده و امکانات این ابزار را معرفی کرده و در ادامه با جزئیات بیشتری به بررسی هر امکان پرداخته است.

کاشفی (۱۴۰۱) تلاش کرده است تا ابعاد، پیچیدگی‌ها و چالش‌های پردازی و راهکارهای موجود و موردنیاز برای مواجهه و مرتفع ساختن این مسائل، به‌منظور توسعه سامانه‌های خودکار پردازش و خطایابی زبان فارسی را از نظر بگذرانند. وی چالش‌های عمده خط فارسی را در ساخت‌واژه، فاصله‌گذاری، واژه‌های هم‌آوا^۴ و حروف هم‌شکل دیده است. راه‌حل برطرف ساختن این معضله‌ها معیارسازی است که بهتر است نهاد مرجعی مانند فرهنگستان زبان و ادب فارسی، عهده‌دار این مسئولیت شود. بر این اساس، می‌توان معیارسازی را به بسترهای دیگری تسری داد: بستر نگارش مبتنی بر تبادل و شیوه نمایش اطلاعات، چیدمان حروف و علائم فارسی بر صفحه کلید رایانه، قلم، خطایابی املائی و تولید پیشنهادها، جایگزین، رتبه‌بندی پیشنهادها، مدل‌سازی خطایابی، خطایابی مبتنی بر یادگیری ماشین و پژوهش‌ها و ابزارهای موجود پیرامون ویرایش و خطایابی املائی فارسی. پژوهش‌های وی به تهیه ابزاری به نام ویراستیار منجر شده است.

قیومی (۱۴۰۱) در پژوهش خود گزارش جامعی از مطالعات انجام‌شده و همچنین ابزارهای موجود در پردازش‌های پایه خط و زبان فارسی ارائه کرده است. یکی از نکاتی که او در جمع‌بندی نتایج پژوهش‌های مختلف به آن پرداخته یکسان‌نبودن داده‌های آموزش و ارزیابی مدل‌ها و ممکن‌نبودن مقایسه عملکرد الگوریتم‌های پردازشی است.

1. <https://virastyar.ir/>

2. plug-in

3. Microsoft Word

4. homophone

نام خانوادگی نویسنده اول و دوم (بیش از دو نویسنده نام خانوادگی نویسنده اول و همکاران | ۹

بنابراین، تعیین «وضعیت روز»^۱ در پدیده‌های فارسی، چندان واضح و مشخص نیست. در مقاله حاضر تلاش می‌شود ابزارهای ویراستاری مختلف تهیه‌شده برای زبان فارسی با پیکره سنجۀ یکسان ارزیابی شوند تا توانمندی ابزارها مقایسه‌پذیر گردد.

۳. روش

دو موضوع اساس چارچوب نظری مقاله حاضر را شکل می‌دهند: یکی دستور خط و دیگری پیکره سنجۀ که در ادامه توضیح داده خواهند شد.

۳-۱. دستور خط

توجه به قواعد املائی و نگارشی زبان و پابندی به این قواعد هنگام تدوین فرهنگ‌های لغت عمومی و تخصصی توسط فرهنگ‌نویسان و اصطلاح‌شناسان، بسیار اهمیت دارد. اصول واژه‌سازی هر زبان، بخشی از این قواعد را تشکیل می‌دهد و اصطلاحات جدید تنها در صورتی می‌توانند در یک زبان خاص ابداع شوند که قواعد املائی و نگارشی آن زبان را رعایت کرده باشند. بدون نظام املائی و نگارشی مناسب، شیوۀ نگارش کاربران زبان، نامنظم خواهد بود. بنابراین، تدوین این قواعد از اهمیت بسزایی برخوردار است. معمولاً برنامه‌ریزی‌ها، سیاست‌ها و اصول اولیه در املا و نگارش واژه‌ها بدون تغییر باقی می‌ماند، اما توسعه زبان معمولاً بر املا و نگارش تأثیر می‌گذارد و آنها را دستخوش تغییر می‌کند (Alberts, 2013). فینلایسون^۲ (2004؛ نقل از آلبرتس (2013)) معتقد است «اگر قواعد املائی و نگارشی استاندارد نباشند، نمی‌توان از زبان در ارتباطات نوشتاری استفاده کرد و آن را در فرهنگ لغت ثبت نمود. واژه‌ها و اصطلاحات جدیدی که می‌توانند ابداع شوند به پایۀ زبانی مستحکمی با اصول صحیح واژه‌سازی به‌عنوان هنجار نیاز دارند».

زبان طبیعی در مقایسه با زبان‌های برنامه‌نویسی و صوری، مشکلات و پیچیدگی‌های پردازشی بسیاری دارد. یکی از عوامل بروز این مشکلات، مسائل فناورانه است. از جمله عوامل دیگر می‌توان به عامل انسانی اشاره کرد که سبب تنوع نگارش واژه‌ها شده و پردازش داده‌های زبانی را با معضلات متعددی مواجه می‌کند. شایان ذکر است گردآوری داده‌های زبانی از گویشوران متفاوت یک زبان نیز مسأله دیگری است که

1. state-of-the-art

2. Finlayson, R.

به این پیچیدگی می‌افزاید. یکی از مسائلی که جمع‌آوری داده‌های زبانی را پیچیده می‌کند مسئله خط است. همانطور که می‌دانیم خط نیز باید پیرو اصول و ضوابطی باشد که مجموع این اصول و ضوابط دستور خط نامیده می‌شود.

در سال‌های اخیر که کاربرد رایانه در عرصه زبان و خط روزافزون شده و دایره شاغلان به امر چاپ و تکثیر، از حلقه متخصصان سنتی این حرفه فراتر رفته، توجه به مسئله خط ضرورت و اهمیت بیشتری پیدا کرده است. به همان نسبت، نگرانی از خطر بروز تشتت و اعمال سلیقه‌های مختلف و متضاد نیز نسبت به گذشته افزایش یافته است. بعضی از تغییرات در شیوه نگارش واژه‌ها می‌تواند در قالب تنوع نگارشی (قیومی و همکاران، ۱۳۹۴)، شکسته‌نویسی (صلح‌جو، ۱۳۸۶؛ طیب‌زاده، ۱۳۹۸) و گفتاری‌نویسی (قیومی و مسگرخویی، ۱۴۰۲) مطرح گردد.

فرهنگستان زبان و ادب فارسی به‌عنوان نهادی دولتی که طبق اساس‌نامه خود، عهده‌دار پاسداری از خط و زبان فارسی است در پی آن بوده که تا سرحد امکان خط فارسی را قانونمند و قاعده‌پذیر سازد تا از آشفتگی و هرج‌ومرج جلوگیری کند. نمونه‌ای از این آشفتگی، تنوع نگارشی در واژه‌ها است که شریفی و همکاران (۱۳۹۶) بررسی کرده‌اند. دستور خط فارسی در سال ۱۳۸۱ پس از جلسات متعدد با اعضای پیوسته فرهنگستان و با ملاحظه همه اظهارنظرهای صاحب‌نظران تصویب شد و در نسخه ۱۴۰۲ مورد بازبینی کلی قرار گرفت. در تدوین دستور خط فارسی و با هدف معیارسازی آن، به ملاک‌های زیر توجه شده است: ۱) حفظ چهره خط فارسی، ۲) حفظ استقلال خط، ۳) تطابق مکتوب و ملفوظ، ۴) فراگیر بودن قاعده، ۵) سهولت نوشتن و خواندن، ۶) سهولت آموزش قواعد و ۷) فاصله‌گذاری در خط فارسی (فرهنگستان زبان و ادب فارسی، ۱۳۸۱؛ ۱۴۰۲).

۲-۳. تهیه پیکره سنجه

پیکره سنجه، مجموعه داده‌ای است که برای سنجش عملکرد مدل‌ها و الگوریتم‌ها به کار می‌رود. این اصطلاح، در نتیجه تلفیق رویکرد زبان‌شناسانه به پیکره و رویکرد فناوریانه مقایسه‌پذیر به منظور سنجش نرم‌افزارها حاصل شده است. تهیه پیکره سنجه دو پیامد دارد: یکی ارزیابی کیفی سامانه‌های نرم‌افزاری است که سنجش عملکرد مدل یا الگوریتم را براساس خروجی گرفته‌شده ممکن می‌سازد. پیامد دیگر، امکان مقایسه نتایج حاصل از

عملکرد ارزیابی مدل‌ها و الگوریتم‌ها و مقایسه آنها با یکدیگر است. پیکره سنجه دو ویژگی اصلی دارد: یکی از ویژگی‌ها گستردگی پیکره سنجه است تا بتواند به مجموعه‌ای از فعالیت‌ها و وضعیت آنها در دنیای واقعی نزدیک باشد. ویژگی دیگر، دسترسی به حجم و تعداد مناسب نمونه‌ها برای انجام آزمایش‌های مقیاس‌پذیر است؛ بنابراین، پیکره سنجه باید دربرگیرنده نمونه‌های زیادی باشد (Neumayer et al., 2007).

با توجه به این دو ویژگی که در بالا ذکر شد، پیکره سنجه به سه دسته کلی تقسیم می‌شود:

الف) پیکره کامل از نظر محتوا: این پیکره انواع محتوای مرتبط با یک موضوع را پوشش می‌دهد تا بتواند ابعاد مختلف آن موضوع موردنظر را در برگیرد؛ ب) پیکره کامل از نظر ویژگی: این پیکره حاوی انواع ویژگی‌هایی است که به یک مسأله مربوط می‌شود و برای آزمایش کامل‌بودن یک مدل یا الگوریتمی که مختص یک مسأله است به کار می‌رود؛ ج) پیکره تعریف‌کننده عملکرد: مجموعه داده بزرگی که می‌توان از آن برای مقایسه دو یا چند مدل یا الگوریتم استفاده کرد. ترکیب پیکره کامل از نظر محتوا و ویژگی به خوبی می‌تواند بیانگر عملکرد باشد (Neumayer et al., 2007).

معضلاتی که در «حفاظت دیجیتال پیکره»^۱ مطرح است عبارت است از: الف) اهمیت تعریف دقیق فعالیت موردنظر برای کاربردی‌شدن و موفقیت پیکره، ب) اندازه مناسب پیکره که بتواند تمامی نیاز را برآورده سازد، ج) لایه‌بندی و توزیع عناصر در یک پیکره براساس معیارهای مختلف، د) بازنمایی داده به گونه‌ای که قابلیت خواندن و پردازش توسط رایانه را داشته باشد، ه) تعریف دقیق معیارهای ارزیابی که معمولاً توسط نیروی انسانی برچسب‌زن تعریف می‌شود (Neumayer et al., 2007).

در پژوهش حاضر تلاش می‌شود با توجه به ملاک‌های معرفی‌شده، در گام نخست، پیکره سنجه برای مقایسه ابزارهای ویراستاری فارسی تهیه گردد و در گام دوم به ارزیابی مقایسه‌ای ابزارها پرداخته شود.

۳-۳. روش

مقایسه‌پذیری نتایج مدل‌های زبانی و الگوریتم‌های پردازش زبان طبیعی، مسأله کلیدی و بسیار مهمی است که ضمن یافتن نقاط ضعف و قوت آن مدل یا الگوریتم، برای همکاری

و رقابت بین گروه‌های پژوهشی، بسیار مفید واقع می‌شود. در این خصوص، می‌توان پیکره‌ای را براساس معیارهای مشخص موردنظر، نشانه‌گذاری کرد و از آن در آزمایش‌ها و ارزیابی‌ها بهره برد. این مجموعه داده باید به‌طور مداوم، طراحی و توسعه یابد و حاوی ژانرهای مختلف و ویژگی‌های فنی متفاوت در زمینه‌های گوناگون باشد. برای حصول اطمینان از بهره حداکثری از چنین مجموعه‌ای، تعیین میزان پیچیدگی و هدف از ساخت این مجموعه، دشواری‌های اصلی تهیه این نوع پیکره هستند.

همان‌گونه که در مقدمه به آن اشاره شد، هدف از پژوهش حاضر، مقایسه ابزارهای ویراستاری فارسی براساس یک پیکره سنج بر مبنای دستور خط زبان فارسی مصوب فرهنگستان زبان و ادب فارسی است. در این بخش به معرفی پیکره سنج می‌پردازیم. برای رسیدن به این هدف، با دو چالش روبه‌رو هستیم: یکی ژانر داده و دیگری مسأله هدف در خط که باید در هنگام تهیه داده سنج مدنظر قرار گیرند.

برای مسأله اول، دو ژانر را انتخاب کردیم که نوع داده‌ها را مشخص می‌کند: یکی ژانر علمی است و دیگری ژانر خبری. بین این دو ژانر تفاوت فاحشی وجود دارد؛ چه از نظر واژگان و چه از نظر ساخت زبانی. معمولاً در ژانر علمی جملات کوتاه و واژه‌های تخصصی به کار می‌روند؛ درحالی‌که در ژانر خبری جملات بلندتر هستند و به‌طور عمده از واژه‌های عمومی استفاده می‌شود. در ژانر علمی ۱۰ مقاله کامل را به‌صورت تصادفی از مجموعه مقالات علمی - پژوهشی نامه فرهنگستان انتخاب کردیم.^۱ سپس، مقالات را به فایل متنی ساده تبدیل و پس از انطباق با آخرین نسخه دستور مصوب خط فارسی، داده‌ها را براساس ساختار یک واژه در هر ردیف فایل اکسل، سازماندهی کردیم. در ژانر خبری نیز، اخبار کامل و متنوعی را از وبگاه‌های خبری آرمان ملی^۲، خبر ورزشی^۳، دنیای اقتصاد^۴، رسالت^۵، شرق^۶، هم‌میهن^۷ و همشهری^۸ به‌صورت تصادفی و دستی گردآوری کردیم. داده

-
1. <https://nf.apll.ir/>
 2. <https://www.armanmeli.ir/>
 3. <https://www.khabarvarzeshi.com/>
 4. <https://donya-e-eqtasad.com/>
 5. <https://resalat-news.com/>
 6. <https://www.sharghdaily.com/>
 7. <https://hammihanonline.ir/>
 8. <https://www.hamshahrionline.ir/>

گردآوری شده حاصل از متون علمی شامل ۴۲,۹۸۶ واژه و داده خبری شامل ۵۵,۳۳۲ واژه است. به عبارتی دیگر، پیکره گردآوری شده شامل بیش از ۹۸,۰۰۰ واژه است.

در چالش دوم، دستور مصوب خط فارسی را بررسی و موارد مطرح شده در آن را به ۸ مقوله اصلی تقسیم کردیم. این مقولات عبارت است از: (۱) پیشوندها و پسوندها، (۲) تنوین و تشدید، (۳) حروف اضافه مرکب و ربط، (۴) ترکیبها، (۵) واژهها و ترکیبهای عربی، (۶) نویسهها و فاصلهها، (۷) یای نکره و (۸) کسره اضافه و همزه. در این مرحله، دادههایی که براساس دستور مصوب خط فارسی معیار شده براساس هر مقوله موردنظر نشاندار می شود.

در فرایند نشاندار کردن واژهها، دو مسأله مدنظر قرار گرفته است: (۱) توانایی ابزار در تشخیص مورد غلط و اصلاح آن: در این موارد واژه معیار، آگاهانه براساس مقوله موردنظر تغییر پیدا می کند. برای مثال اگر مقوله موردنظر «تنوین و تشدید» است، نمادهای تنوین و تشدید از واژه هدف حذف می گردد. از ابزار موردنظر انتظار می رود در موارد نشاندار شده این موارد را اصلاح و به شکل معیار تبدیل کند. (۲) توانایی مدل در تشخیص مورد صحیح و صرف نظر از اعمال تغییر: در این موارد، واژه معیار موردنظر، بدون هرگونه تغییری براساس مقوله موردنظر نشانه گذاری می گردد. برای مثال، اگر مقوله موردنظر «تنوین و تشدید» باشد، نمادهای تنوین و تشدید از واژه هدف حذف نمی شود و عیناً در داده آزمون وجود خواهد داشت. انتظار می رود که ابزار، این موارد نشاندار شده را بدون هرگونه اصلاح بپذیرد و در داده ورودی، تغییری اعمال نکند. تعداد کل واژههای نشاندار شده در ژانر علمی ۱۲۴۶ واژه و در ژانر خبری ۹۳۷ واژه است. به عبارتی دیگر، این پیکره سنج، حاوی ۲۱۸۳ واژه نشانه گذاری شده است. آمار تعداد واژههای داده خام و نشاندار شده در دو ژانر علمی و خبری برای مقولات هشت گانه موردنظر در جدول ۱ گزارش شده است.

جدول ۱. دادههای پیکره سنج

داده علمی		داده خبری		مقوله
تعداد واژههای داده خام	تعداد واژههای نشاندار شده	تعداد واژههای داده خام	تعداد واژههای نشاندار شده	

۳۵	۳۳۶۲	۵۵۲	۱۳۶۹۰	پیشوندها و پسوندها
۴۸	۵۵۳۲	۱۰۷	۵۰۴۱	تنوین و تشدید
۵۷	۵۰۱۱	۵۰	۲۷۵۶	حروف اضافه مرکب و ربط
۳۲۱	۱۸۲۴۱	۳۴	۸۴۳	ترکیبها
۹	۱۷۵۰	۷۵	۷۳۹۰	واژه‌ها و ترکیبهای عربی
۲۷۸	۱۰۶۰۷	۱۸۲	۲۰۵۴	نویسه‌ها و فاصله‌ها
۲۶	۳۵۹۴	۱۲۶	۴۸۷۹	یای نکره
۱۶۳	۱۰۵۹۷	۱۲۰	۶۳۳۳	کسره اضافه و همزه
۹۳۷	۵۸۶۹۴	۱۲۴۶	۴۲۹۸۶	جمع

شیوهٔ برچسب‌گذاری پیکرهٔ سنجه در نمونهٔ مثال جدول ۲ قابل‌مشاهده است. همان‌طور که مشخص است، دو واژهٔ «تازه» و «بین‌المللی» ازجمله واژه‌های هدف نشان‌دار در پیکرهٔ سنجه هستند که برای دو مقولهٔ «کسره اضافه و همزه» و «نویسه‌ها و فاصله‌ها» نشانه‌گذاری شده‌اند.

جدول ۲. نمونهٔ پیکرهٔ سنجه

دادهٔ اصلی	دادهٔ اصلی معیارشده	دادهٔ نشان‌دارشده	مشخص‌شدن واژهٔ نشان‌دارشده	تعیین نوع مقولهٔ واژهٔ نشان‌دارشده
چارچوب	چارچوب	چارچوب		
تازه	تازه	تازه	*	کسره اضافه و همزه
همکاری	همکاری	همکاری		
ایران	ایران	ایران		
و	و	و		
آژانس	آژانس	آژانس		
بین‌المللی	بین‌المللی	بین‌المللی	*	نویسه‌ها و

فصلها				
انرژی	انرژی	انرژی		
اتمی	اتمی	اتمی		
نهایی	نهایی	نهایی		
شد	شد	شد		
.	.	.		

انتظار می‌رود که ابزار ویراستاری، واژه «تازه» را بدون هرگونه تغییری در خروجی خود داشته باشد، اما در واژه «بین‌المللی» قاعده دستور خط را اعمال کند و آن را به «بین‌المللی» تغییر دهد. چنانچه این دو مورد به صورت‌های گفته‌شده در ستون داده‌های معیارشده تغییر یابند، به ابزار دو امتیاز مثبت داده می‌شود و در صورت داشتن خروجی‌ای به‌جز موارد فهرست‌شده در ستون داده‌های معیار، از امتیاز مثبت کسر می‌شود و به امتیاز منفی اضافه می‌گردد. در مورد شیوه ارزیابی، توضیحات بیشتر در بخش ۵ ارائه خواهد شد.

۴. یافته‌های پژوهش

در انجام پژوهش حاضر، شش ابزار ویراستاری را در نظر گرفته‌ایم که داده‌های سنجۀ تهیه‌شده به این ابزارها داده شد و عملکرد خروجی ابزارها پس از مقایسه واژه‌های نشان‌دار با داده‌های معیار ارزیابی گردید.

۴-۱. معرفی ابزارها

ابزارهایی که در پژوهش حاضر برای ارزیابی مقایسه‌ای عملکرد استفاده شد عبارت است از: ویراویراست^۱، فارسی‌یار^۲، ویراست‌من^۳، پاک‌نویس^۴، ویراستیار^۵ و گاکول. بعضی از ابزارها و توانمندی‌هایشان در بخش دوم مقاله معرفی شده است. توانمندی‌های ابزارها یکسان نیست و بعضی ابزارها امکانات بیشتری برای کاربر فراهم می‌آورد.

1. <https://viravirast.com/>
2. <https://text-mining.ir/>
3. <https://virastman.ir/>
4. <https://paknevis.ir/>
5. <https://virastyar.ir/>

فارسی‌یار، مجموعه ابزار ویراستاری فارسی و حاوی امکانات زیر است: هنجارسازی متن^۱، تقطیع جملات و واژه‌ها، تشخیص کسره اضافه، حذف ایست‌واژه‌ها^۲، ریشه‌یابی کلمات یا بُن‌واژه‌یاب، تبدیل عبارات محاوره‌ای به رسمی، تصحیح‌کننده خطاهای املائی یا خطا در حروفچینی واژه‌ها، برچسب‌زنی مقولات دستوری واژه‌ها در جمله، قطعه‌بند^۳ یا تجزیه نحوی سطحی^۴، تجزیه نحوی^۵ جملات، برچسب‌زن نقش معنایی، تشخیص موجودیت‌های نامدار^۶، مرجع‌یابی ضمائر، ایجاد هستان‌شناس از سند ورودی، شبکه معنایی واژگان و محاسبه شباهت میان دو عبارت، جمله یا متن.

ویراستیار ابزاری است که در قالب افزونه به نرم‌افزار «مایکروسافت ورد»، برای اصلاح خطاهای املائی، اشتباهات ویرایشی و نشانه‌گذاری، همچنین معیارسازی متون فارسی به کار می‌رود. امکانات غلط‌یاب این ابزار عبارت‌اند از: اصلاح املائی واژه‌ها، ارائه فهرستی از واژه‌های صحیح پیشنهادی، اصلاح انواع اشتباهات در فاصله‌گذاری، چسباندن واژه‌های متوالی مرکب به هم، درج فاصله اشتباه میان واژه‌های مختوم به حروف «آ - ا»، «د»، «ذ»، «ر»، «ز»، «ژ» و «و»، اصلاح کاربرد نابه‌جای فاصله به جای نیم‌فاصله، تلفیق درج اشتباه فاصله و چسبیدن واژه‌های متوالی، تشخیص و اصلاح پسوندها در واژه‌ها، تشخیص و اصلاح تکرار متوالی واژه‌ها، اصلاح غلط‌های ناشی از هم‌آوایی، امکان افزودن واژه‌های جدید به واژه‌نامه، امکان اصلاح یک مورد غلط به‌طور یکباره در کل متن، امکان نادیده گرفتن یک مورد غلط و اصلاح نشدن آن و امکان نادیده گرفتن یک مورد غلط و اصلاح نشدن آن در کل متن. علاوه بر این موارد، این ابزار، امکان اصلاح انواع حروف «ی» و «ک»، تبدیل ارقام عربی به معادل فارسی، اصلاح اعراب، حذف نیم‌فاصله‌های تکراری و اصلاح درج علائم سجاوندی را دارد. ویراستیار به سفارش دبیرخانه شورای عالی اطلاع‌رسانی تولید شده است و به‌صورت رایگان و متن‌باز عرضه می‌شود.

ویراست‌من ابزار دیگری است که قابلیت‌های متعددی برای معیارسازی و اصلاح متن دارد که به‌صورت افزونه به نرم‌افزار ورد^۷ اضافه شده و توانایی خطایابی معنایی

-
1. text normalizer
 2. stop word
 3. chunker
 4. shallow parser
 5. parsing
 6. named entity recognition
 7. Microsoft Word

نام خانوادگی نویسنده اول و دوم (بیش از دو نویسنده نام خانوادگی نویسنده اول و همکاران | ۱۷

باتوجه به بافت و پیشنهاد واژه‌های جایگزین، اصلاح شیوه نگارش براساس دستور خط فارسی در اصلاح نیم‌فاصله‌ها و فاصله بین علائم و حروف و همچنین، هنجارسازی متن با جایگزینی حروف عربی با فارسی را دارد.

پاکنویس ابزار ویراستاری است که به صورت افزونه در نرم‌افزار ورد، توانایی غلط‌یابی املایی، رعایت نشانه‌گذاری، ارائه پیشنهادها و ویرایشی، رعایت دستور خط مصوب فرهنگستان و رفع اشکالات مرتبط با اعداد را دارد.

۲-۴. ارزیابی ابزارها

ورودی ابزارهای ویراستاری، داده‌های معیاری است که براساس دستور مصوب خط فارسی تهیه شده است و بعضی از واژه‌های این داده، آگاهانه نشان‌دار شده است. در مرحله ارزیابی، خروجی مدل، صرف‌نظر از امکانات ابزار، با داده معیار مقایسه می‌شود. اطلاعات استخراج‌شده از متن‌های موردنظر، چه در ژانر علمی و چه در ژانر خبری، براساس ماتریس درهم‌ریختگی قیومی (۱۳۹۷)، در جدول ۳، ذکر شده است:

جدول ۳. ماتریس درهم‌ریختگی

داده پردازش شده			
غلط		درست	
داده ورودی	غلط	درست ← درست	درست ← غلط
	غلط	غلط ← درست	غلط ← غلط

منظور از «داده ورودی» در جدول ۳ داده‌ای است که نیروی انسانی، بعضی از واژه‌های آن را براساس مقولات هشت‌گانه موردنظر در دستور مصوب خط فارسی به‌طور آگاهانه نشان‌دار کرده تا توسط ابزار به شکل صحیح تبدیل شود. این تبدیل ممکن است به صورت صحیح یا غلط انجام پذیرد. «داده پردازش‌شده» مربوط به خروجی و عملکرد ابزار بر روی داده است که نتیجه به دست آمده از کار پردازشی ابزار موردنظر ممکن است صحیح یا غلط باشد. درهم‌ریختگی این دو، چهار وضعیت را فراهم می‌آورد:

(الف) صحیح ← صحیح: در این وضعیت، یک واژه صحیح نشان‌دار شده در داده ورودی بدون تغییر و به همان صورت در خروجی ابزار دیده می‌شود.

(ب) غلط ← صحیح: در این وضعیت، یک واژه غلط نشان‌دار شده در داده ورودی، پس از اعمال عملکرد ابزار بر روی آن و انجام کار پردازشی، به یک واژه صحیح براساس دستور مصوب خط فارسی تبدیل می‌شود.

(ج) صحیح ← غلط: در این وضعیت، یک واژه صحیح در داده نشان‌دار شده ورودی، پس از عملکرد ابزار بر روی آن و انجام کار پردازشی، به یک واژه غلط تبدیل می‌شود.

(د) غلط ← غلط: در این وضعیت، یک واژه غلط نشان‌دار شده در داده ورودی، بدون تغییر و به همان صورت غلط در خروجی دیده می‌شود و ابزار، هیچ‌گونه کار پردازشی بر روی آن انجام نمی‌دهد.

براساس تعاریف بالا، تساوی (۱) میزان دقت ابزار موردنظر را محاسبه می‌کند. در این تساوی، نسبت تمام صحیح‌ها در خروجی ابزار به تمام حالات در داده‌ها سنجیده می‌شود. بالابودن این عدد، بیانگر میزان دقت در داده خروجی است. فاصله میزان درصد دقت تا ۱۰۰ درصد، بیانگر میزان غلط در داده خروجی و نیاز ابزار به اصلاح و بهبود الگوریتم پردازشی آن برای دستیابی به دقت ۱۰۰ درصد است.

(۱)

$$\text{دقت} = \frac{\text{صحیح} \leftarrow \text{صحیح} + \text{غلط} \leftarrow \text{صحیح}}{\text{غلط} \leftarrow \text{غلط} + \text{غلط} \leftarrow \text{صحیح} + \text{صحیح} \leftarrow \text{غلط} + \text{صحیح} \leftarrow \text{صحیح}}$$

۱-۲-۵. ابزارهای تک‌حالتی

دو ابزار ویراویراست و فارسی‌یار، امکانات متنوعی ندارند و بدون انجام هرگونه تنظیماتی، داده ورودی را به داده خروجی تبدیل می‌کنند. عملکرد ابزارها برای مقولات هشت‌گانه منطبق با دستور مصوب خط فارسی در جدول‌های ۴ و ۵ گزارش شده است.

نام خانوادگی نویسنده اول و دوم (بیش از دو نویسنده نام خانوادگی نویسنده اول و همکاران | ۱۹

جدول ۴. عملکرد ابزار ویراویراست

مقوله	علمی				خبری			
	درست <	درست < غلط	غلط <	غلط < غلط	درست <	درست < غلط	غلط <	غلط < غلط
پیشوندها و پسوندها	۶۷	۴	۳۲۲	۱۵۹	۰	۰	۱۵	۲۰
تنوین و تشدید	۳۸	۱	۴۱	۲۷	۰	۳	۸	۳۷
حروف اضافهٔ مرکب و ربط	۱	۱	۲۹	۱۹	۰	۳	۲۱	۳۳
ترکیب‌ها	۰	۰	۲۲	۱۲	۴	۲۶	۱۲۲	۱۶۹
واژه‌ها و ترکیب‌های عربی	۱۵	۳	۲۸	۲۹	۰	۰	۱	۸
نویسه‌ها و فاصله‌ها	۵	۰	۱۲۶	۵۱	۱	۲۵	۱۱۱	۱۴۱
یای نکره	۱۳	۰	۹۹	۱۴	۰	۷	۵	۱۴
کسرهٔ اضافه و همزه	۱۲	۰	۱۰۸	۰	۵	۱۹	۱۲	۱۲۷
جمع	۱۵۱	۹	۷۷۵	۳۱۱	۱۰	۸۳	۲۹۵	۵۴۹
عملکرد نسبی	۱۲/۱۲	۰/۷۲	۶۲/۲۰	۲۴/۹۶	۶/۶۶	۰/۸۰	۲۳/۶۸	۴۴/۰۶

جدول ۵. عملکرد ابزار فارسی‌یار

مقوله	علمی				خبری			
	درست <	درست < غلط	غلط <	غلط < غلط	درست <	درست < غلط	غلط <	غلط < غلط

پیشوندها و پسوندها	۵۵	۱۴	۲۴۸	۲۳۵	۱	۰	۲۸	۶
تنوین و تشدید	۶	۳	۷	۹۱	۴	۰	۲۲	۲۲
حروف اضافه مرکب و ربط	۰	۱	۲۳	۲۶	۴	۰	۲۳	۳۰
ترکیبها	۰	۰	۱۶	۱۸	۲۴	۶	۱۵۴	۱۳۷
واژه‌ها و ترکیب‌های عربی	۱۵	۳	۲۰	۳۷	۰	۰	۵	۴
نویسه‌ها و فاصله‌ها	۲	۰	۱۳۰	۵۰	۲۳	۱	۲۰۱	۵۳
یای نکره	۱۶	۰	۵۰	۶۰	۷	۰	۱۷	۲
کسره اضافه و همزه	۲	۹	۲۱	۸۸	۱۰	۱۹	۳۸	۹۶
جمع	۹۶	۳۰	۵۱۵	۶۰۵	۷۳	۲۶	۴۸۸	۳۵۰
عملکرد نسبی	۷/۷۰	۲/۴۱	۴۱/۳۳	۴۸/۵۶	۵/۸۶	۲/۰۹	۳۹/۱۷	۲۸/۰۹

براساس نتایج گزارش شده درمورد عملکرد ویراویراست در جدول ۴ مشخص است که عملکرد این ابزار در ژانر علمی، بهتر از ژانر خبری است و براساس داده معیار، اشتباهات بسیاری را به درست تبدیل کرده است؛ درحالی که میزان عمل نکردن مدل در تبدیل داده‌های غلط به درست در ژانر خبری، بسیار بالا است که سبب کاهش کارایی کلی ابزار می‌گردد. یکی از ویژگی‌های ویراویراست این است که در کمتر از یک درصد از داده‌ها، واژه درست را به غلط تبدیل کرده است. در جدول ۵ عملکرد فارسی‌یار در مقایسه با ویراویراست کاملاً برعکس است، زیرا حجم داده‌های غلط در خروجی ژانر علمی بیشتر از ژانر خبری است.

۲-۲-۵. ابزارهای چندحالتی

ابزارهای ویراست‌من، پاک‌نویس، ویراستیار و گاگول امکانات متنوعی دارند که با انجام تنظیمات مختلف آنها می‌توان داده ورودی را به داده خروجی تبدیل کرد. عملکرد کلی ابزارها برای تنظیمات مختلف برای دو ژانر علمی و خبری در جدول ۶ گزارش شده است.

جدول ۶. عملکرد کلی ابزارهای ویراست‌من، پاک‌نویس، ویراستیار و گاگول با تنظیمات مختلف

ابزار	حالت ۱		حالت ۲		حالت ۳		حالت ۴		حالت ۵		حالت ۶		حالت ۷	
	ویراست‌من	پاک‌نویس	ویراستیار	گاگول	ویراست‌من	پاک‌نویس	ویراستیار	گاگول	ویراست‌من	پاک‌نویس	ویراستیار	گاگول	ویراست‌من	پاک‌نویس
ویراست‌من	۴۳/۹۸	۴۳/۶۳	۴۲/۹۴	۳۲/۴۲	۴۳/۲۶	۳۰/۴۲	۴۰/۴۵	۳۲/۱۸						
پاک‌نویس	۷۲/۴۷	۵۶/۸۲	۷۲/۳۱	۵۶/۰۲	۷۲/۴۷	۷۵/۶۷	۷۲/۳۱	۵۶/۹۸						
ویراستیار	۱۲/۱۲	۹/۳۱	۱۳/۴۸	۹/۶۳	۶۸/۲۲	۵۳/۹۳	۱۳/۶۴	۱۰/۶۷	۶۹/۹۰	۵۴/۱۷	۶۸/۲۲	۵۳/۹۳	۷۰/۵۹	۷۳/۰۱
گاگول	۲۱/۴۳	۱۳/۰۰	۱۳/۰۸	۱۲/۱۲	۱۵/۴۹	۱۰/۱۱	۲۳/۳۵	۱۶/۶۱	۲۵/۲۸	۱۵/۱۷	۱۷/۵۸	۱۳/۱۶	۲۶/۱۷	۲۳/۲۶

مفهوم حالت‌های ۱ تا ۷ در ابزارهای مختلف در زیر تعریف شده است:

الف) ابزار ویراست‌من:

حالت ۱: مدل پایه با فعال‌نبودن گزینه‌های «پیشنهاد‌های فرهنگستان» و «پیشنهاد‌های

عمومی»؛

حالت ۲: فعال‌بودن گزینه «پیشنهاد‌های فرهنگستان»؛

حالت ۳: فعال‌بودن گزینه «پیشنهاد‌های عمومی»؛

حالت ۴: فعال‌بودن هر دو گزینه «پیشنهاد‌های فرهنگستان» و «پیشنهاد‌های عمومی».

ب) ابزار پاک‌نویس:

حالت ۱: مدل پایه با فعال‌نبودن گزینه‌های «پیشنهاد ویرایشی» و «مصوب فرهنگستان»؛

حالت ۲: فعال‌بودن گزینه «پیشنهاد ویرایشی»؛

حالت ۳: فعال بودن گزینه «مصوب فرهنگستان»؛

حالت ۴: فعال بودن هر دو گزینه «پیشنهاد ویرایشی» و «مصوب فرهنگستان».

ج) ابزار ویراستیار:

حالت ۱: فعال بودن گزینه «نویسه»؛

حالت ۲: فعال بودن گزینه «دستور خط»؛

حالت ۳: فعال بودن گزینه «ویراستاری»؛

حالت ۴: فعال بودن گزینه‌های «نویسه» و «دستور خط»؛

حالت ۵: فعال بودن گزینه‌های «نویسه» و «ویراستاری»؛

حالت ۶: فعال بودن گزینه‌های «دستور خط» و «ویراستاری»؛

حالت ۷: فعال بودن گزینه‌های «نویسه»، «دستور خط» و «ویراستاری».

د) ابزار گاکول:

حالت ۱: فعال بودن گزینه «واژه‌ها»؛

حالت ۲: فعال بودن گزینه «وندها»؛

حالت ۳: فعال بودن گزینه «نویسه‌ها»؛

حالت ۴: فعال بودن گزینه‌های «واژه‌ها» و «وندها»؛

حالت ۵: فعال بودن گزینه‌های «واژه‌ها» و «نویسه‌ها»؛

حالت ۶: فعال بودن گزینه‌های «وندها» و «نویسه‌ها»؛

حالت ۷: فعال بودن گزینه‌های «واژه‌ها»، «وندها» و «نویسه‌ها».

براساس نتایج میانگین دو ژانر علمی و خبری در حالت‌های مختلف تنظیمات ابزارها، این نتیجه به دست آمد که ویراست‌من در حالت ۱، پاک‌نویس در حالت ۳ و ویراستیار و گاکول در حالت ۷ بهترین عملکرد را دارد. به عبارتی دیگر، فعال کردن امکانات در ابزار ویراست‌من با توجه به داده سنجۀ تهیه شده بر عملکرد کلی نرم‌افزار، تأثیر منفی می‌گذارد. فعال بودن گزینه «مصوب فرهنگستان» در پاک‌نویس و فعال بودن تمامی امکانات ابزار در ویراستیار و گاکول با توجه به داده سنجۀ تهیه شده بر عملکرد ابزارها تأثیر

نام خانوادگی نویسنده اول و دوم (بیش از دو نویسنده نام خانوادگی نویسنده اول و همکاران | ۲۳

مثبت دارد. در جدول‌های ۷ تا ۱۰ عملکرد این دسته از ابزارها در مقولات هشت گانه گزارش شده است.

جدول ۷. عملکرد ابزار ویراست‌من

مقوله	علمی				خبری			
	درست <	غلط <	غلط <	غلط <	درست <	غلط <	غلط <	غلط <
پیشوندها و پسوندها	۷۲	۶	۲۳۱	۲۴۳	۱	۰	۱۷	۱۷
تنوین و تشدید	۹	۱	۵	۹۲	۵	۰	۳۶	۲۶
حروف اضافه مرکب و ربط	۲	۰	۱۸	۳۰	۵	۰	۱۷	۳۵
ترکیب‌ها	۰	۰	۱۵	۱۹	۲۸	۳	۷۵	۲۱۵
واژه‌ها و ترکیب‌های عربی	۱۵	۲	۱۳	۴۵	۰	۰	۳	۶
نویسه‌ها و فاصله‌ها	۱	۰	۸۷	۹۴	۲۶	۰	۱۳۹	۱۱۳
یای نکره	۱۵	۲	۲۶	۸۳	۷	۰	۱۰	۹
کسره اضافه و همزه	۵	۶	۳۴	۷۵	۱۱	۱۲	۳۷	۱۰۳
جمع	۱۱۹	۱۷	۴۲۹	۶۸۱	۸۳	۱۵	۳۳۴	۵۲۴
عملکرد نسبی	۹/۵۵	۱/۳۶	۳۴/۴۳	۵۴/۶۵	۶/۶۶	۱/۲۰	۲۶/۸۱	۴۲/۰۵

جدول ۸. عملکرد ابزار پاک‌نویس

مقوله	علمی				خبری			
	درست <	غلط <	غلط <	غلط <	درست <	غلط <	غلط <	غلط <
پیشوندها و	۷۱	۵	۳۳۸	۱۳۸	۲	۰	۳۰	۳

پسوندها								
تنوین و تشدید	۱۰	۰	۱۲	۸۵	۱	۰	۳۸	۹
حروف اضافهٔ مرکب و ربط	۲	۰	۳۳	۱۵	۳	۱	۳۱	۲۲
ترکیبها	۰	۰	۳۰	۴	۲۸	۰	۱۹۸	۹۵
واژه‌ها و ترکیبهای عربی	۱۷	۳	۳۶	۱۹	۰	۰	۶	۳
نویسه‌ها و فاصله‌ها	۴	۰	۱۴۱	۳۷	۲۳	۱	۲۰۴	۵۰
یای نکره	۱۶	۰	۸۸	۲۲	۷	۰	۱۸	۱
کسرهٔ اضافه و همزه	۱۲	۰	۹۳	۱۵	۲۸	۰	۹۲	۴۳
جمع	۱۳۲	۸	۷۷۱	۳۳۵	۹۲	۲	۶۱۷	۲۲۶
عملکرد نسبی	۱۰/۵۹	۰/۶۴	۶۱/۸۸	۲۶/۸۹	۷/۳۸	۰/۱۶	۴۹/۵۲	۱۸/۱۴

جدول ۹. عملکرد ویراستیار

مقوله	علمی				خبری			
	درست <	درست <	غلط <	غلط <	درست <	درست <	غلط <	غلط <
پیشوندها و پسوندها	۷۶	۰	۳۳۶	۱۳۳	۱	۰	۳۲	۲
تنوین و تشدید	۸	۲	۴۰	۵۷	۴	۰	۳۴	۱۰
حروف اضافهٔ مرکب و ربط	۲	۰	۲۱	۲۷	۳	۲	۲۱	۳۱
ترکیبها	۰	۰	۱۶	۱۸	۳۱	۰	۲۱۸	۷۲
واژه‌ها و ترکیبهای عربی	۱۹	۰	۲۹	۲۷	۰	۰	۵	۴
نویسه‌ها و فاصله‌ها	۱	۰	۱۴۲	۳۶	۲۶	۰	۲۰۰	۵۲

نام خانوادگی نویسنده اول و دوم (بیش از دو نویسنده نام خانوادگی نویسنده اول و همکاران | ۲۵

۱۷	۰	۸۵	۲۲	۷	۰	۱۶	۳	یای نکره
۱۲	۰	۶۷	۴۱	۲۹	۰	۵۷	۷۷	کسره اضافه و همزه
۱۳۵	۲	۷۳۶	۳۶۱	۱۰۱	۲	۵۸۳	۲۵۱	جمع
۱۰/۸۳	۰/۱۶	۵۹/۰۷	۲۸/۹۷	۸/۱۱	۰/۱۶	۴۶/۷۹	۲۰/۱۴	عملکرد نسبی

جدول ۱۰. عملکرد گاگول

مقوله	علمی				خبری			
	درست <	درست <	غلط <	غلط <	درست <	درست <	غلط <	غلط <
پیشوندها و پسوندها	۷۴	۰	۱۳۵	۳۴۳	۱	۰	۱۰	۲۴
تنوین و تشدید	۱۱	۰	۸	۸۸	۴	۰	۲	۴۲
حروف اضافه مرکب و ربط	۲	۰	۱۵	۳۳	۳	۰	۱۸	۳۶
ترکیبها	۰	۰	۱۰	۲۴	۲۹	۰	۶۳	۲۲۹
واژه‌ها و ترکیبهای عربی	۲۰	۰	۴	۵۱	۰	۰	۲	۷
نویسه‌ها و فاصله‌ها	۲	۰	۱۶	۱۶۴	۲۴	۰	۱۶	۲۳۸
یای نکره	۱۶	۰	۱	۱۰۹	۷	۰	۷	۱۲
کسره اضافه و همزه	۱۲	۰	۰	۱۰۸	۲۹	۰	۳	۱۳۱
جمع	۱۳۷	۰	۱۸۹	۹۲۰	۹۷	۰	۱۲۱	۷۱۹
عملکرد نسبی	۱۱/۰۰	۰	۱۵/۱۷	۷۳/۸۴	۷/۷۸	۰	۹/۷۱	۵۷/۷۰

براساس نتایج گزارش شده عملکرد ویراست‌من و گاگول در جدول‌های ۷ و ۱۰، مشخص است که عملکرد این دو ابزار در تبدیل موارد غلط به صحیح در دو ژانر علمی و

خبری تقریباً یکسان است و هردو نسبتاً غلط‌های زیادی را به صحیح تبدیل نکرده است. در جدول‌های ۸ و ۹ عملکرد پاکنویس و ویراستیار در مقایسه با ویراست‌من و گاگول کاملاً برعکس است؛ زیرا حجم داده‌های غلط در هر دو ژانر بسیار کمتر است و مدل‌ها عملکرد بهتری را در تبدیل داده‌های غلط به درست داشته است. یکی از ویژگی‌های گاگول در مقایسه با تمامی ابزارها این است که هیچ واژه درستی را به غلط تبدیل نکرده؛ درحالی‌که سایر ابزارها، حتی در حدود یک درصد، این تبدیل را انجام داده است.

برای جمع‌بندی عملکرد مدل‌ها در یک نگاه و مقایسه آنها با یکدیگر، در جدول ۱۱ کارایی مدل‌ها در دو ژانر علمی و خبری و همچنین، امتیاز میانگین حاصل از عملکرد ابزارها در این دو ژانر گزارش شده است. همان‌گونه که از امتیاز میانگین ابزارها مشخص است، پاکنویس با ۷۴/۰۷ درصد، بالاترین و گاگول با ۲۴/۷۲ درصد، پایینترین امتیاز میانگین را به‌دست آورده‌اند.

جدول ۱۱. عملکرد کلی ابزارها با استفاده از پیکره سنج

نام ابزار	داده علمی	داده خبری	میانگین
ویراویراست	۷۴/۳۲	۴۰/۳۵	۵۷/۳۳
فارسی‌یار	۴۹/۰۳	۵۹/۸۷	۵۴/۴۵
ویراست‌من	۴۳/۹۸	۴۳/۶۳	۴۳/۸۱
پاکنویس	۷۲/۴۷	۷۵/۶۷	۷۴/۰۷
ویراستیار	۷۰/۵۹	۷۳/۰۱	۷۱/۸۰
گاگول	۲۶/۱۷	۲۳/۲۶	۲۴/۷۲

برای آنکه بتوانیم به درک عمیق‌تری از عملکرد ابزارها در مقولات هشت‌گانه برسیم، امتیاز میانگین دو ژانر علمی و خبری را برای هر مقوله و ابزار محاسبه کرده‌ایم که در جدول ۱۲ گزارش شده است.

جدول ۱۲. عملکرد کلی ابزارها در مقولات مختلف

مقوله	ویراویراست	فارسی یار	ویراست من	پاکنویس	ویراستیار	گاگول
پیشوندها و پسوندها	۵۶/۶۶	۶۸/۸۷	۵۳/۱۶	۸۲/۷۶	۸۴/۴۶	۳۴/۶۵
تنوین و تشدید	۴۸/۳۷	۳۳/۱۶	۴۹/۲۵	۵۰/۹۱	۶۲/۰۱	۱۵/۱۳
حروف اضافه مرکب و ربط	۵۱/۰۵	۴۶/۶۸	۳۹/۳۰	۶۴/۸۲	۴۴/۰۵	۳۵/۴۲
ترکیبها	۵۵/۴۱	۵۱/۲۶	۳۸/۱۰	۷۹/۳۲	۶۲/۳۱	۲۹/۰۴
واژه‌ها و ترکیبهای عربی	۳۴/۲۲	۵۱/۱۱	۳۵/۳۳	۶۸/۶۷	۵۹/۷۸	۲۷/۱۱
نویسه‌ها و فاصله‌ها	۶۰/۴۵	۷۶/۵۵	۵۳/۸۵	۸۰/۶۶	۷۹/۹۳	۱۲/۱۴
یای نکره	۶۷/۵۲	۷۲/۳۴	۴۸/۹۶	۸۹/۳۵	۸۴/۷۱	۳۳/۶۷
کسره اضافه و همزه	۵۹/۵۱	۲۴/۳۱	۳۰/۹۷	۸۰/۵۶	۵۹/۳۰	۱۴/۸۲

براساس نتایج گزارش شده، ویراستیار با ۸۴/۴۶ درصد امتیاز، بالاترین کارایی را در دو مقوله «پیشوندها و پسوندها» و همچنین «تنوین و تشدید» به دست آورده و توانسته بهترین عملکرد را در میان ابزارها داشته باشد. ابزار پاکنویس بالاترین امتیاز را در مقولات «حروف اضافه مرکب و ربط» با ۶۴/۸۲ درصد، «ترکیبها» با ۷۹/۳۲ درصد، «واژه‌ها و ترکیبهای عربی» با ۶۸/۶۷ درصد، «نویسه‌ها و فاصله‌ها» با ۸۰/۶۶ درصد، «یای نکره» با ۸۹/۳۵ درصد و «کسره اضافه و همزه» با ۸۰/۵۶ درصد به دست آورده است. همچنین، دو ابزار فارسی یار و ویراست من در مقوله «نویسه‌ها و فاصله‌ها»، به ترتیب با ۷۶/۵۵ درصد و ۵۳/۸۵ درصد بالاترین کارایی را از آن خود کرده‌اند. گاگول که پایین ترین کارایی را داشت از میان مقولات هشت گانه توانسته است در مقوله «حروف اضافه مرکب و ربط» بالاترین کارایی را با ۳۵/۴۲ درصد به دست آورد.

نقطه ضعف ابزارهای دارای کمترین امتیاز متنوع است. ویراویراست در مقوله «واژه‌ها و ترکیبهای عربی» با ۳۴/۲۲ درصد کارایی کمترین امتیاز را در میان مقولات به دست آورده است. فارسی یار و ویراست من، در مقوله «کسره اضافه و همزه»، به ترتیب، کمترین امتیاز را با ۲۴/۳۱ درصد و ۳۰/۹۷ درصد در میان مقولات به خود اختصاص داده

است. پاکنویس در مقوله «تنوین و تشدید» با ۵۰/۹۱ درصد، ویراستیار در مقوله «حروف اضافه مرکب و ربط» با ۴۴/۰۵ درصد و گاگول در مقوله «نویسه‌ها و فاصله‌ها» با ۱۲/۱۴ درصد کارایی، کمترین امتیاز را در میان مقولات کسب کرده است.

می‌توان از زاویه دیگری به ارزیابی ابزارها پرداخت و آن، عملکرد ابزارها در مقولات هشت‌گانه نسبت به امتیاز میانگین گرفته‌شده خود ابزار در دو ژانر علمی و خبری و نه در مقایسه با عملکرد سایر ابزارها است. به این صورت که اگر امتیاز یک ابزار در یک مقوله (جدول ۱۳) کمتر از امتیاز میانگین عملکرد ابزار (جدول ۱۲) باشد، آن مقوله به‌عنوان چالش ابزار معرفی شود. نتیجه این مقایسه که از جدول‌های ۱۱ و ۱۲ به‌دست آمده در جدول ۱۳ قابل مشاهده است. موارد چالش برانگیز با ستاره مشخص شده است.

جدول ۱۳. عملکرد ابزارها نسبت به خود در مقایسه با امتیاز میانگین در مقولات مختلف

مقوله	ویراویراست	فارسی‌یار	ویراست‌من	پاکنویس	ویراستیار	گاگول	جمع ستاره‌ها
پیشوندها و پسوندها	*						۱
تنوین و تشدید	*	*		*	*	*	۵
حروف اضافه مرکب و ربط	*	*	*	*	*		۵
ترکیب‌ها	*	*	*	*	*		۴
واژه‌ها و ترکیب‌های عربی	*	*	*	*	*		۵
نویسه‌ها و فاصله‌ها						*	۱
یای نکره							۰
کسره اضافه و همزه		*	*	*	*	*	۴

براساس بررسی انجام‌شده در جدول ۱۳ چنین می‌توان نتیجه گرفت که مقولات «پیشوندها و پسوندها»، «نویسه‌ها و فاصله‌ها» و «یای نکره»، مشکل چندانی برای ابزارها به‌حساب نمی‌آید و ابزارها می‌تواند بر این مسائل فائق آید. این در حالی است که مقولاتی مانند «ترکیب‌ها» و «کسره اضافه و همزه» چالش‌هایی است که بیشتر ابزارها در پردازش آنها هنوز با مشکل مواجه است تا بتواند کارایی قابل‌قبولی را به‌دست دهد. مقولات «تنوین

و تشدید»، «حروف اضافه مرکب و ربط» و «واژه‌ها و ترکیب‌های عربی» عمده چالش‌هایی است که بیشتر ابزارها در پردازش آنها با خطای زیادی دست‌وپنجه نرم می‌کند. به عبارتی دیگر، این مقولات جزء دشوارترین مقولاتی است که ابزارها برای معیارسازی خط فارسی براساس دستور مصوب خط با آنها مواجه است.

۶. نتیجه‌گیری

در پژوهش حاضر کوشیدیم ضمن تهیه یک پیکره سنجه براساس دستور مصوب خط فارسی و تقسیم‌بندی ویژگی‌های این دستور در هشت مقوله، به بررسی مقایسه‌ای شش ابزار ویراستاری فارسی، از جمله ویراویراست، فارسی‌یار، ویراست‌من، پاکنویس، ویراستیار و گاکول، در دو ژانر علمی و خبری پردازیم.

این پژوهش با هدف تهیه پیکره سنجه و مقایسه‌پذیرکردن ابزارهای ویراستاری فارسی انجام پذیرفت که ممکن است توسط افراد و سازمان‌های مختلف تهیه شده و با داده‌های مختلف، آموزش دیده باشد. براساس خروجی‌های حاصل از ویراستاری ابزارها و مقایسه آنها با داده‌های معیار در دو ژانر علمی و خبری، این نتیجه به‌دست آمد که پاکنویس، بالاترین و گاکول، پایین‌ترین کارایی ابزارها را دارد. با توجه به مقولات هشت‌گانه، می‌توان براساس امتیاز میانگین عملکرد مدل‌ها به بررسی نقاط قوت و ضعف ابزارها در مقایسه با یکدیگر در مواجهه با دستور خط مصوب فارسی و حتی عملکرد خود ابزار پرداخت. با دانستن نقاط ضعف ابزارها می‌توان چالش‌های مربوط به عملکرد آنها در مقولات هشت‌گانه را در هنگام تهیه نسخه‌های بعدی رفع کرد تا به افزایش کارایی ابزارها منجر گردد. این امر سبب می‌شود ضمن افزایش کارایی مدل‌ها، فضای رقابتی برای رفع چالش‌های پردازش خط زبان شکل بگیرد.

تعارض منافع

تعارض منافع ندارم.

ORCID

Masood Ghayoomi  <http://orcid.org/0000-0001-6685-1332>

Laily Nezami  <http://orcid.org/0009-0000-2516-1052>

منابع

اسلامی، محرم. (۱۳۸۱). دشواری‌های پردازش رایانه‌ای خط فارسی. *نشر دانش*، ۱۹ (۳)، ۲۸-۳۲.
بگ‌جانی، عباس. (۱۳۹۹). ویراستیار (برنامه ویرایش متون فارسی). *ویژه‌نامه نامه فرهنگستان*، ۱۶، ۲۵۱-۲۳۹.

ستوده، هاجر، و هنرجویان، زهره. (۱۳۹۱). مروری بر دشواری‌های زبان فارسی در محیط دیجیتال و تأثیرات آنها بر اثربخشی پردازش خودکار متن و بازیابی اطلاعات. *مجله کتابداری و اطلاع‌رسانی*، ۱۵ (۴)، ۵۹-۹۲. https://lis.aqr-libjournal.ir/article_42651.html

ستوده، هاجر، و هنرجویان، زهره. (۱۳۹۳). بررسی تنوع الگوهای نگارش فارسی و تأثیر آن بر جامعیت بازیابی اطلاعات (مطالعه موردی: پیکره همشهری). *مجله کتابداری و اطلاع‌رسانی*، ۱۷ (۲)، ۳۱-۴۹. https://lis.aqr-libjournal.ir/article_42322.html

شریفی، ساغر، صنعتی، مرضیه، و قیومی، مسعود. (۱۳۹۶). تنوع نگارشی در فارسی و تهیه پیکره‌های زبانی در قالب فرهنگ. *نامه فرهنگستان: ویژه‌نامه فرهنگ‌نگاری*، ۱۲، ۹۳-۱۰۸.
صادقی، علی اشرف، و زندی‌مقدم، زهرا. (۱۳۸۵). *فرهنگ املائی خط فارسی: براساس دستور خط فارسی مصوب فرهنگستان زبان و ادب فارسی*. نشر آثار فرهنگستان زبان و ادب فارسی.

صفارمقدم، احمد. (۱۳۷۶). برنامه‌ریزی و پاک‌سازی زبان در کره جنوبی. *نامه فرهنگستان*، ۱۰، ۱۱۲-۸۴.

صلح‌جو، علی. (۱۳۸۶). «بشکنیم یا تشکنیم» (پاسخی به نظر منوچهر انور در نمایشنامه عروسک‌خانه). *فصلنامه مترجم*، ۴۵ (۱۷)، ۲۲-۹.

<https://motarjemjournal.ir/2020/07/3747>

- طیب‌زاده، امید. (۱۳۹۸). مبانی و دستور خط فارسی شکسته براساس صد سال آثار داستانی و نمایشی (از ۱۲۹۸ تا ۱۳۹۷). پژوهشگاه علوم انسانی و مطالعات فرهنگی.
- فرهنگستان زبان و ادب فارسی. (۱۳۸۱). دستور خط فارسی: مصوب فرهنگستان زبان و ادب فارسی. نشر آثار فرهنگستان زبان و ادب فارسی.
- فرهنگستان زبان و ادب فارسی. (۱۴۰۲). دستور خط فارسی: مصوب فرهنگستان زبان و ادب فارسی. نشر آثار فرهنگستان زبان و ادب فارسی.
- قیومی، مسعود. (۱۴۰۱). پیش‌پردازش و ابزارهای پایه. در مهنوش شمس‌فرد و محمود بی‌جن‌خان (ویراستاران)، پردازش متن و گفتار فارسی مروری بر مبانی نظری و آخرین یافته‌های پژوهشی (صص. ۸۶-۱۱۳). سمت.
- قیومی، مسعود، شریفی، ساغر، و صنعتی، مرضیه. (۱۳۹۴). تنوع نگارشی در زبان فارسی و تهیه خودکار دادگان املایی از پیکره زبانی مبتنی بر وب. مجموعه مقالات اولین همایش بین‌المللی پژوهش‌های وب (صص. ۷-۱۰). دانشگاه علم و فرهنگ.
- قیومی، مسعود، و مسگرخویی، مریم. (۱۴۰۲). تحلیلی بر پیکره حاصل از داده‌های زبانی فارسی در فضای مجازی. زبان و زبان‌شناسی، ۱۹ (۳۷)، ۱۱۷-۱۳۶.
- کاشفی، امید. (۱۳۹۰). ویراستار: مطالعه تطبیقی یک فعالیت پردازشی متن‌باز در زبان فارسی. ره‌آورد نور، ۳۴، ۹۶-۱۰۱.
- کاشفی، امید. (۱۴۰۱). خطایابی و ویرایش و استانداردسازی. در مهنوش شمس‌فرد و محمود بی‌جن‌خان (ویراستاران)، پردازش متن و گفتار فارسی مروری بر مبانی نظری و آخرین یافته‌های پژوهشی (صص. ۲۰۹-۲۲۵). سمت.

References

- Academy of Persian Language and Literature (2002). *Persian orthographical rules: Approved by the Persian Language and Literature Academy*. Publication of the Works of the Academy of the Persian Language and Literature. [In Persian]
- Academy of Persian Language and Literature (2023). *Persian orthographical rules: Approved by the Persian Language and Literature Academy*. Publication of the Works of the Academy of the Persian Language and Literature. [In Persian]
- Alberts, M. (2006). *Standardisation, modernisation and harmonisation — A sociolinguistic perspective*. In Y. Wang, Y. Wang, & Y. Tian (Eds.) *Proceedings of the International Conference on Terminology*,

- Standardization and Technology Transfer, (pp. 182-192). Encyclopedia of China Publishing House.
- Alberts, M. (2013). The Afrikaans orthographic rules as guide for other South African languages. *Lexikos*, 23(1), 1-28.
<https://doi.org/10.5788/23-1-1202>
- Begjani, A. (2020). Virāstyār (Persian Text Editing Software). *Special Issue of Nāme-ye Farhangestān*, 16, 239-251. [In Persian]
- Eslami, M. (2002). Difficulties in computer processing of the Persian script. *Dānesh Publishing*, 19(3), 28-32. [In Persian]
- Finlayson, R. (2004). Revision of the spelling and orthography rules of the official African languages. Presented at the *Pan South African language board consultative meeting*. Pretoria.
- Ghayoomi, M. (2022). Preprocessing and basic tools. In M. Shamsfard & M. Bijankhan (Eds), *Persian text and speech processing: A review of theoretical foundations and recent research findings* (pp. 86-113). Samt. [In Persian]
- Ghayoomi, M. (2025). Developing, compiling and annotating corpora for the Persian language. In M. V. Farahani & Z. Ghane (Eds.), *New frontiers in corpus-based studies of Persian: Challenges, innovations and applications* (pp. 25-62). Springer. https://doi.org/10.1007/978-3-031-98989-6_2
- Ghayoomi, M., & Mesgarkhoyi, M. (2023). A study on the Persian corpus developed from cyberspace. *Language and Linguistics*, 19(37), 117-136. [In Persian]
- Ghayoomi, M., & Momtazi, S. (2009). Challenges in developing Persian corpora from on-line resources. *IEEE international conference on Asian language processing* (pp. 108-113). Singapore.
- Ghayoomi, M., Momtazi, S., & Bijankhan, M. (2010). A study of corpus development for Persian. *International journal on Asian language processing*, 20(1), 17-34.
- Ghayoomi, M., Sharifi, S., & Sanaati, M. (2015). Writing diversity in Persian and automatic preparation of orthographic data from web-based linguistic corpus. In *Proceedings of the first international conference on web research* (pp. 7-10). University of Science and Culture. [In Persian]
- Kashefi, O. (2011). Virāstyār: A comparative study of an open-source processing activity in Persian. *Rahāvard Noor*, 34, 96-101. [In Persian]
- Kashefi, O. (2022). Error detection, editing and standardization. In M. Shamsfard & M. Bijankhan (Eds), *Persian text and speech processing: A review of theoretical foundations and recent research findings* (pp.209-225). Samt. [In Persian]
- Neumayer, R., Kulovits, H., Rauber, A., Thaller, M., Nicchiarelli, E., Day, M., Hofman, H., & Ross, S. (2007). On the need for benchmark corpora

- in digital preservation. In *Proceedings of the 2nd DELOS conference on digital libraries*. Tirrenia.
- Rat, M. (1968). L'orthographe française. *Revue Des Deux Mondes (1829-1971)* (pp 184–192). *Revue des Deux Mondes*. <http://www.jstor.org/stable/44599448>
- Régnier-Desmarais, F. S. (1706). *Traite de la grammaire Francoise*. Kessinger Publishing.
- Sadeghi, A. A., & Zandi-Moghaddam, Z. (2006). *Persian spelling dictionary: Based on the approved Persian orthography rules of the Academy of Persian Language and Literature*. Publication of the Works of the Academy of Persian Language and Literature. [In Persian]
- Saffar Moghaddam, A. (1997). Language planning and purification in South Korea. *Nāme-ye Farhangestān*, 10, 84-112. [In Persian]
- Shamsfard, M. (2011). Challenges and open problems in Persian text processing. *The 5th language and technology conference: Human language technologies as a challenge for computer science and linguistics* (pp. 65-69). Poznań, Poland.
- Shamsfard, M. (2019). Challenges and opportunities in processing low resource languages: A study on Persian. In *Proceedings of the language technologies for all* (pp. 291–295). Paris.
- Sharifi, S., Sanati, M., & Ghayoomi, M. (2017). Writing diversity in Persian and the preparation of language corpora in the form of a dictionary. *Nāme-ye Farhangestān: Special Issue on Philology*, 12, 93-108. [In Persian]
- Solhjoo, A. (2007). "Shall we break or not break?" (A response to Manouchehr Anvar's opinion on the play *The Dollhouse*). *Translator's Quarterly*, 45(17), 9-22. <https://motarjemjournal.ir/2020/07/3747> [In Persian]
- Sotoudeh, H., & Honarjoyan, Z. (2012). A review of the difficulties of Persian language in the digital environment and their impacts on the automatic text processing and information retrieval. *Journal of Library and Information Science*, 15(4), 59-92. https://lis.aqr-libjournal.ir/article_42651.html [In Persian]
- Sotoudeh, H., & Honarjoyan, Z. (2014). A study of the diversity of Persian writing patterns and its impact on the comprehensiveness of information retrieval (Case Study: Hamshahri's Corpus). *Journal of Library and Information*, 17(2), 31-49. https://lis.aqr-libjournal.ir/article_42322.html [In Persian]
- Tabibzadeh, O. (2019). *Basics and grammar of broken Persian script based on a hundred years of fiction and drama (from 1919 to 2018)*. Institute for Humanities and Cultural Studies. [In Persian]

Thimonnier, R. (1967). *Le système graphique du français: Introduction à une pédagogie rationnelle de l'orthographe*. Plon.