



## A Study on the Persian Corpus developed from Cyberspace

Ghayoomi, Masood<sup>1</sup> 

Institute for Humanities and Cultural Studies; and  
Academy of Persian Language and Literature, Tehran,  
Iran

Mesgarkhoiy, Maryam<sup>2</sup> 

Researcher of the Language and Computer Group at the  
Persian Academy of Persian Language and Letters,  
Tehran, Iran

### Abstract

Nowadays, the existence of emerging communication tools has made communication between speakers possible through writing. The electronic, global and interactive nature of such emerging technologies has facilitated and increased the speed of communication. The linguistic interaction by using these tools and the relationship between speech and writing have causes a type of writing to be created by the users of a language, a writing type known as “neography”. The main aim of this research is to investigate the properties of neography in Persian and classify them into the categories based on a linguistic corpus developed from the data published in social media. To this end, the corpus is semi-automatically annotated based on the standard writing style, and the linguistic properties at phonetic, morphological and syntax levels. Then, the words whose written forms are different from the standard form and have a type of neography are studied based on the orthography properties and also linguistic features. The results of the analyzing the data and the assigned labels show that neography in Persian in the virtual space occurs at two levels of orthography and morpho-phonetic, and the content words bear the highest amount of neography in the words.

**Keywords:** Neography, Broken writing, Colloquial writing, Virtual space, Persian orthography grammar.

1. [m.ghayoomi@ihcs.ac.ir](mailto:m.ghayoomi@ihcs.ac.ir) (Corresponding Author)

2. [m.mesgar62@gmail.com](mailto:m.mesgar62@gmail.com)

**How to cite:** Ghayoomi, M., & Mesgarkhoiy, M. (2024). A Study on the Persian Corpus developed from Cyberspace. *Language and Linguistics*, 19(37), 117 - 136. doi: [10.30465/lsi.2024.47171.1715](https://doi.org/10.30465/lsi.2024.47171.1715)



## تحلیلی بر پیکره حاصل از داده‌های زبانی فارسی در فضای مجازی

پژوهشگاه علوم انسانی و مطالعات فرهنگی و فرهنگستان زبان و ادب فارسی،  
تهران، ایران

قیومی، مسعود <sup>ID</sup>

پژوهشگر گروه زبان و رایانه فرهنگستان زبان و ادب فارسی

مسگرخویی، مریم <sup>ID</sup>

### چکیده

امروزه وجود ابزارهای ارتباطی نوظهور سبب شده‌است که ارتباط میان گویشوران از طریق نوشتن میسر شود. ماهیت الکترونیکی، جهانی و تعاملی این دسته از فناوری‌های نوظهور سبب تسهیل و افزایش سرعت در ارتباطات شده‌است. تعامل زبانی با به‌کارگیری این ابزارها و رابطه میان گفتار و نوشتار سبب می‌شود که گونه‌ای از نوشتار توسط کاربران یک زبان خلق شود، گونه‌ای که به نونویسی معروف است. هدف از انجام این پژوهش بررسی ویژگی‌های نونویسی در فارسی و طبقه‌بندی انواع آن براساس پیکره زبانی تهیه‌شده از داده‌های منتشرشده در شبکه‌های اجتماعی است. برای این هدف، داده‌های گردآوری‌شده در این پیکره براساس شیوه نگارش معیارشان، در سطوح آوایی، بن‌واژه‌ای و مقوله دستوری به‌صورت نیمه‌خودکار برچسب‌گذاری شده‌است. سپس واژه‌هایی که صورت نوشتاری‌شان متفاوت از صورت معیار است و دارای نوعی نونویسی است از جنبه ویژگی‌های خط و همچنین ویژگی‌های زبانشناختی مورد بررسی قرار گرفت. نتایج حاصل از تحلیل این داده‌ها و برچسب‌هایشان نشان می‌دهد که نونویسی در فارسی در فضای مجازی در دو سطح نگارشی و آوایی-ساخت‌واژی اتفاق می‌افتد و واژه‌های محتوایی بیشترین میزان نونویسی واژه‌ها را متحمل می‌شود.

**کلیدواژه:** نونویسی، شکسته‌نویسی، محاوره‌نویسی، فضای مجازی، دستور خط فارسی

## ۱ مقدمه

بشر امروز انواع ابزارهای ارتباطی را در اختیار دارد و می‌تواند از آن‌ها برای تبادل نظر و اطلاعات استفاده کند. یک نمونه از این دسته از ابزارها، شبکه‌های اجتماعی در محیط اینترنت هستند. پیش‌تر کار نوشتن به افراد فرهیخته و طبقه اجتماعی فرهنگی خاصی اختصاص داشت، اما چنین فناوری‌هایی سبب شده‌است که تمامی افراد، فارغ از سن، طبقه اجتماعی و سطح سواد، به خلق متن از طریق نوشتار بپردازند. متون خلق‌شده در چنین فضایی با متون معیار تفاوت و فاصله دارد و ابعاد گوناگون این تفاوت می‌تواند موضوع پژوهش زبان‌شناسان باشد. دسترسی به حجم زیادی از داده‌هایی که در این فضا تولید می‌شوند پژوهشگر را با انبوهی از داده‌ها مواجه می‌کند که می‌تواند در قالب یک پیکره زبانی مورد بررسی قرار گیرند. ظهور ابزارهای نوین ارتباطی، مانند پیامک یا شبکه‌های اجتماعی، سبب شده‌است نوشتار در این محیط‌ها به گفتاری‌نویسی تغییر یابد. بنابراین در این نوع داده، دو حالت<sup>۱</sup> گفتار و نوشتار (مکانری<sup>۲</sup> و ویلسون<sup>۳</sup>، ۲۰۰۱) همزمان با هم حضور دارند. در این نوع ارتباط، تولید محتوا و ارسال آن به محض فشردن دکمه «ارسال» یا «ایتر»<sup>۴</sup>، سبب می‌شود تعامل زبانی در لحظه و تقریباً همزمان با فشردن کلید اتفاق بیفتد (انیس<sup>۵</sup>، ۱۹۹۹)؛ همانگونه که با خروج آواهای زبانی در کانال ارتباطی، تعامل میان دو نفر برقرار می‌شود.

وجود ویژگی کم‌کوشی در فناوری‌های نوین (دژوند<sup>۶</sup> و مرسیر<sup>۷</sup>، ۲۰۰۲) سبب شده‌است نگارش واژه‌ها دستخوش تغییر شود و پدیده‌ای با عنوان نونویسی<sup>۸</sup> در نگارش واژه‌ها اتفاق بیفتد. نونویسی به دو صورت در خط تجلی پیدا می‌کند: یکی شکسته‌نویسی و دیگری محاوره‌نویسی (گفتارنویسی، عامیانه‌نویسی، خودمانی‌نویسی). باید توجه داشت که میان شکسته‌نویسی و محاوره‌نویسی تفاوت وجود دارد. در شکسته‌نویسی، واژه‌ها و عبارات باتوجه‌به شیوه بیان‌شان در گفتار نگارش می‌شوند؛ درحالی‌که در محاوره‌نویسی، واژه‌ها در قالبی صمیمانه و خودمانی، همان‌طور که بیان می‌شوند، نگارش می‌یابند. برای مثال، نگارش «ماس» به جای «ماست» (نوعی لبنیات) به گونه محاوره‌نویسی تعلق دارد. درحالی‌که، «ماست» به جای «ما است» صورت شکست تلقی می‌گردد. گاهی عدول از قواعد نگارش به گونه‌ای است که با اصطلاح «شلخته‌نویسی» (غفاری، ۱۳۹۴) شناخته می‌شود. عوامل متعددی بر شکسته‌نویسی در فضای مجازی تأثیرگذار

---

1. mode  
 2. McEnery, T.  
 3. Wilson, A.  
 4. Enter  
 5. Anis, J.  
 6. Dejongd, A.  
 7. Mercier, J.  
 8. neography

است. از سوی دیگر پدیده‌های مرتبط با خط در فضای اینترنت و وب تنها منحصر به پدیده شکسته‌نویسی نیست، به‌گونه‌ای که حتی ممکن است اینترنت بر ماهیت زبان و ادبیات نیز تأثیر بگذارد. نمونه این قبیل مطالعات، پژوهش شوهانی و حسینی (۱۳۹۷) است که به بررسی اثرگذاری فضای مجازی بر زبان و ادبیات فارسی معاصر پرداخته‌است.

در پژوهش حاضر به تحلیل و معرفی پیکره داده‌های زبان فارسی در فضای مجازی و پدیده نونویسی در فضای مجازی می‌پردازیم و تلاش می‌کنیم براساس رفتار کاربرانی که از شبکه‌ها و رسانه‌های اجتماعی مانند تلگرام، اینستاگرام، واتس‌آپ و توئیتر استفاده می‌کنند، شیوه نگارش فارسی شکسته و انواع نونویسی را در فضای مجازی بررسی کنیم.

## ۲. پیشینه پژوهش

تفاوت در گونه گفتاری و نوشتاری در بیشتر زبان‌های دنیا دیده می‌شود؛ از این رو، نوعی نونویسی در بسیاری از این زبان‌ها رواج دارد. پژوهش‌های گوناگونی در ایران و خارج از ایران به بررسی این موضوع پرداخته و ابعاد آن را روشن کرده‌اند. در این مطالعات تلاش شده‌است راهکارهایی برای مواجهه با این پدیده یافت شود. در این بخش، ابتدا به بررسی کوتاه برخی از مهمترین آثار مرتبط با پدیده نونویسی در زبان‌های گوناگون پرداخته و در ادامه، پژوهش‌هایی که با این نگاه به بررسی زبان فارسی در فضای مجازی پرداخته‌است را معرفی می‌کنیم.

لازار<sup>۱</sup> (۲۰۱۲) به موضوع نونویسی در گپ‌های الکترونیکی در زبان فرانسه پرداخته و تمامی تغییرات ایجادشده در نگارش واژه‌ها در محیط‌های الکترونیکی را در قالب نونویسی بررسی کرده‌است. براساس مطالعه انجام‌شده بر روی پیکره حاصل از رخدادهای گفتگو، عامل اول پدیده نونویسی، جایگزینی یک واژه با یک حرف است مانند جایگزینی *k* به جای *que*. عامل دیگر نونویسی، جایگزینی یک حرف به جای حرف دیگر است. حذف واکه یا همخوان آخر که تلفظ نمی‌شود نیز عامل دیگر نونویسی در زبان فرانسه معرفی شده‌است.

لیندمن<sup>۲</sup> (۲۰۰۵) عبارت «انگلیسی شکسته»<sup>۳</sup> را معرفی کرده و آن را گونه نوشتاری غیرمعیار و غیرمرسوم زبان انگلیسی می‌داند. وی این ویژگی زبان انگلیسی را از بُعد اجتماعی بررسی کرده و نشان داده‌است که زبان انگلیسی با چنین ویژگی‌هایی نوعی پیچین است که در نتیجه کاربرد انگلیسی در میان دانش‌آموزان، دانشجویان و فارغ‌التحصیلان غیرانگلیسی‌زبان پدید آمده‌است. به عقیده لیندمن، چنین موضوعی می‌تواند در پژوهش‌هایی که به کنش‌گری زبانی

1. Lazar, J.

2. Lindemann, S.

3. broken English

مبتنی بر فناوری و شیوه نگارش غیرمعیار شبکه‌های اجتماعی می‌پردازد دقیق‌تر مطالعه و بررسی گردد.

منزفیلد<sup>۱</sup> و همکاران (۲۰۱۹) در پژوهشی با بررسی ابعاد گوناگون متون انگلیسی محاوره‌ای، ابزاری برای تبدیل این دسته از متون به متن معیار معرفی کرده‌اند. ابزار آن‌ها یک ماشین ترجمه است که با بهره‌گیری از یک الگوریتم یادگیری عمیق دنباله-به-دنباله<sup>۲</sup>، یک رشته انگلیسی محاوره‌ای را به عنوان ورودی می‌پذیرد و آن را به یک رشته معیار تبدیل می‌کند. کوزیربایف<sup>۳</sup> و یسنبایی<sup>۴</sup> (۲۰۲۰) از همین شیوه برای تبدیل متون قزاقی محاوره‌ای به گونه معیار آن استفاده کرده‌اند. آن‌ها برای بهبود این مدل، از شبکه عصبی کدگذار-کدگشا<sup>۵</sup> و حافظه کوتاه مدت طولانی<sup>۶</sup> بهره برده‌اند.

آرفین<sup>۷</sup> و تیون<sup>۸</sup> (۲۰۲۰) نیز متن‌های مالایی محاوره‌ای را که در شبکه‌های اجتماعی تولید شده است را با ابزار مشابهی به متن معیار تبدیل کرده و در ادامه به برچسب‌گذاری مقولات دستوری این متن‌ها پرداخته‌اند.

آرمین و شمس‌فرد (۲۰۱۱) مدلی قاعده‌مند معرفی کرده‌اند که متن فارسی محاوره‌ای را به متن رسمی تبدیل می‌کند و با در نظر گرفتن احتمال کاربرد واژه‌ها در محور هم‌نشینی، پیشنهادات ارائه شده را رتبه‌بندی می‌نماید. در این پژوهش، پیکره‌ای زبانی حاوی ۴۴ هزار واژه از ترکیب زیرنویس فیلم‌ها، متون ادبی محاوره‌ای و محتوای وبلاگ تهیه شده و از پیکره بی‌جن‌خان (۱۳۸۴) برای طراحی مدلی احتمالاتی برای رتبه‌بندی استفاده شده است.

رسولی و همکاران (۲۰۲۰) نیز با استفاده از مدل زبانی مبتنی بر برت<sup>۹</sup> (دولین<sup>۱۰</sup> و همکاران، ۲۰۱۹)، مدلی دنباله-به-دنباله برای ترجمه ماشینی ارائه کرده‌اند که متن فارسی محاوره‌ای را به فارسی رسمی تبدیل می‌کند. داده‌هایی که در این پژوهش به کار رفته است صفحات ویکی‌پدیا و پیکره میزان (کاشفی، ۲۰۲۰) است.

معصومی و همکاران (۲۰۲۰) ابتدا با هدف اعتبارسنجی، داده‌هایی از شبکه اجتماعی توییتر، رسانه اجتماعی تلگرام و قسمت نظرات وبگاه دیجی‌کالا را خزش کرده و در ادامه، آن‌ها را با

- 
1. Mansfield, C.
  2. sequence-to-sequence
  3. Kozhimbayev, Z.
  4. Yessenbayev, Z.
  5. encoder-decoder
  6. long short-term memory (LSTM)
  7. Ariffin, S. N. A. N.
  8. Tiun, S.
  9. Bidirectional Encoder Representations from Transformers (BERT)
  10. Devlin, J.

واژه‌های متناظر سالمشان برچسب‌گذاری کرده‌اند. آنها با استفاده از این داده‌ها مدلی طراحی کرده‌اند که داده‌های ورودی را از تلگرام می‌پذیرد و آن را به فارسی رسمی تبدیل می‌کند. ادیبیان و ممتازی (۱۴۰۱) به مدلی دست یافته‌اند که عبارات حاوی شکسته‌نویسی در محتوای شبکه‌های اجتماعی را به‌عنوان داده ورودی می‌پذیرد و شکل سالم واژه‌ها را به‌عنوان خروجی ارائه می‌دهد. مدلی که در این مبدل استفاده شده است مبتنی بر شبکه عصبی عمیق است و از الگوریتم کدگذار-کدگشا و یادگیری دنباله-به-دنباله بهره برده است. دقت گزارش شده این مدل ۷/۷۰ درصد است.

شکسته‌نویسی در متون ادبی فارسی نیز از جمله موضوعات مطرح در محافل علمی متمرکز بر زبان فارسی است. به همین منظور مجموعه جلساتی با عنوان «هم‌اندیشی و ویرایش» در شرکت انتشارات فنی ایران برگزار شده است و بزرگانی چون ابوالحسن نجفی، محمدرضا باطنی، علی صلح‌جو، امید طیب‌زاده و دیگران به موضوع شکسته‌نویسی در فارسی پرداخته‌اند. ویدئوهای این جلسات در وب در دسترس است.<sup>۱</sup> چند اثر منتشرشده ایرانی نیز به موضوع شکسته‌نویسی در فارسی پرداخته‌اند که در ادامه بررسی می‌گردد.

صلح‌جو (۱۳۸۶) از پدیده شکسته‌نویسی در فارسی دفاع کرده و برای آن حد و حدودی تعیین کرده است. وی گفتار و نوشتار را در قالب حافظه صوتی و حافظه تصویری مورد توجه قرار داده و به تغییرات آوایی و شیوه نگارش واژه‌ها و همچنین تفاوت در نحو گفتار و نوشتار پرداخته است. به عقیده وی، بیان شیوه گفتار در قالب نوشتار در متون ادبی موجب می‌شود خواننده ارتباط بیشتری با متن برقرار کند و مشکلی در شکسته‌خواندن واژه‌ها نداشته باشد. وی این حد از شکسته‌نویسی را مجاز دانسته و نویسندگان را از افراط در شکسته‌نویسی برحذر داشته است؛ چراکه به اعتقاد وی افراط بر شکسته‌نویسی، خوانش و درک متن را با اختلال مواجه می‌کند.

طیب‌زاده (۱۳۹۸ الف) در پژوهش خود به معرفی مبانی و دستور خط فارسی شکسته پرداخته است. وی در این پژوهش با استناد به آثار داستانی و نمایشی صد سال اخیر، دستور خطی توصیفی را ارائه کرده است. وی معتقد است صورت‌های گویشی باید جدا از صورت‌های سبکی شکسته در نظر گرفته شود؛ از این رو وی در پژوهش خود، گونه‌های گویشی را در بررسی آثار داستانی و نمایشی فارسی استخراج و آمار آن را جداگانه گزارش کرده است. علاوه بر گونه‌های گویشی، وی واژه‌های گفتاری، مانند واسه، واژه‌های عامیانه، مانند خفن، یا زبان مخفی، مانند پرپا (کبوتر)، را جزء شکسته‌نویسی نمی‌داند؛ چراکه این موارد متعلق به فارسی معیار نیست و

1. <https://www.aparat.com/v/EyTgo>

شکل نوشتاری سالم در زبان فارسی ندارد. همچنین از نظر وی، مواردی از محاوره‌گرایی مانند درج پسوند معرفه‌ساز e- در «دختره» یا پسوند تصغیرساز ke- در «مردکه»، ki- در «راستکی» یا پسوند القایی u- در «چاقالو»، جملگی ربطی به شکسته‌نویسی ندارد.

طیب‌زاده (۱۳۹۸ب) در تکمیل پژوهش پیشین خود، در کتاب فارسی شکسته ۲۵ سؤال را مطرح کرده و به‌صورت مستدل به آن‌ها پاسخ گفته است. در ادامه وی دستوری برای خط شکسته ارائه کرده و فرهنگی املائی حاوی صورت‌های شکسته و معادل سالم آن را تهیه نموده است.

بی‌جن‌خان (۱۳۹۱) در پژوهش خود به بررسی خط و زبان فارسی در فضای مجازی پرداخته است و ضمن ارائه یک طبقه‌بندی جامع از زبان گفتاری و نوشتاری، به بررسی هر یک از اعضای این دسته‌بندی پرداخته است. وی در این اثر، گونه خط و زبان فارسی در فضای مجازی را با اصطلاح «گونه اینترنتی» معرفی کرده است.

هدایت مفیدی و همکاران (۱۳۹۶) محتوای نوشتاری دو کانال خبری در تلگرام را از نقطه‌نظر فنی، زبانی و بلاغی مورد بررسی قرار داده‌اند. در این بررسی، ۷۵۳ جمله‌ای که در آن‌ها الگوهای معیار رعایت نشده بود به‌عنوان پیکره پژوهش بررسی شده است. براساس داده‌های تحلیل شده، اشکالات زبانی، فنی و بلاغی به‌ترتیب حاوی بیشترین کاربرد غیرمعیار در فضای مجازی بوده است.

مسگرخویی (۱۳۹۹) به آسیب‌شناسی پژوهش‌های معطوف به خط و زبان فارسی در فضای مجازی و ظهور پدیده نونویسی و گفتاری‌نویسی در متن‌های تولیدشده در فضای مجازی پرداخته است. از نظر وی، چهار عامل اصلی مسئول عدول از قواعد زبانی و اعمال تغییرات به‌ظاهر خودسرانه در متون فضای مجازی است: الف) ماهیت الکترونیکی، جهانی و تعاملی اینترنت؛ ب) «نت‌گفتار» و رابطه آن با نوشتار و گفتار؛ ج) ماهیت کاربران از نظر جنسیت، سن، طبقه اجتماعی و میزان تحصیلات؛ د) وجود تحولات اجتماعی و مسئله چندفرهنگی و سنت‌گریزی. مسگرخویی همچنین عواملی چون بستر ارتباط، هدف از ایجاد ارتباط در این فضا، میزان آشنایی با فعالیت در چنین محیط‌هایی، پهنه جغرافیایی کاربران و دارا بودن گویش منطقه‌ای یا قومی، استفاده از زبان نشاندار و سرعت تحولات در دنیای امروز را جزء دلایل تغییرات به‌وجودآمده در نگارش متن در فضای مجازی برشمارده است.

طیب‌زاده (۱۳۹۹) در مقاله خود با عنوان «تغییرات آوایی و صورت‌های شکسته در فضای مجازی و استفاده از آن‌ها در فرهنگ جامع زبان فارسی» به بررسی صورت‌های شکسته و سالم در فضای مجازی پرداخته است. وی کوتاه‌نوشت‌ها، مانند «ک» به جای «که»، و «ه» به‌عنوان

کسره اضافه، مانند «تابه بلند» به جای «تاب بلند» را به‌عنوان قواعد شکسته‌نویسی در فضای مجازی معرفی کرده‌است.

### ۳. چارچوب نظری

در مکتب ساختگرایی، سوسور<sup>۱</sup> (۱۹۱۶) معتقد است که زبان دارای دو سطح «صورت» و «معنا» است. تجلی صورت در زبان می‌تواند صورت آوایی داشته و در کانال ارتباطی از طریق امواج صوتی از گوینده به شنونده منتقل شود. همچنین صورت می‌تواند از طریق خط و نوشتار مکتوب، تجلی یابد و یک کانال ارتباطی بین نویسنده و خواننده را فراهم آورد. چنین می‌توان نتیجه گرفت که خط شکل مکتوب زبان است و معنا در تجلی صورت، چه صورت آوایی و چه صورت نوشتاری، مستتر است. از آنجاکه زبان متشکل از تعدادی قاعده است که به‌صورت تکرارپذیر استفاده می‌شود تا محتوا را شکل دهد، خط نیز باید از اصول و قواعد مشخصی پیروی نماید که به مجموعه این اصول و قواعد دستور گفته می‌شود.

از سوی دیگر، فرگوسن<sup>۲</sup> (۱۹۵۹) مفهوم دوزبانگونی<sup>۳</sup> را برای اولین بار مطرح کرده است. دوزبانگونی به این معناست که دو یا چند گونه زبانی معیار برای یک زبان مشخص وجود دارد که باتوجه به شرایط و موقعیت‌های اجتماعی، یکی از آن گونه توسط گویشور استفاده گردد. نمونه‌ای از شرایط دوزبانگونی هنگامی است که فرد در منزل با یک گونه زبانی صحبت می‌کند و در اجتماع، با گونه زبانی دیگری که رسمی‌تر است. هادسون<sup>۴</sup> (۲۰۰۲) گونه زبانی محاوره‌ای و غیررسمی را فروگونه<sup>۵</sup> و گونه رسمی‌تر است فراگونه<sup>۶</sup> نامیده‌است. این گونه‌های زبان می‌تواند بازنمایی نوشتاری متفاوتی داشته باشد؛ بنابراین ما با شرایطی مواجه می‌شویم که دو گونه نوشتاری برای یک زبان به کار می‌رود. وجود فناوری‌های نوظهوری چون وبلاگ‌ها و شبکه‌های اجتماعی به چنین شرایطی دامن می‌زند و صورت زبانی که به‌واسطه خط در یک محیط الکترونیک تولید می‌شود توسط گویشور دچار تحول شده و سبب می‌شود پدیده دوزبانگونی اتفاق بیفتد. چراکه براساس نظر انیس (۱۹۹۹)، نوشتار در محیط الکترونیکی نوعی تعامل زبانی است که در آن فشردن کلید ارسال همچون بازکردن دهان در گفتار است. این پدیده در تمامی زبان‌ها، از جمله زبان فارسی، اجتناب‌ناپذیر است.

---

1. de Saussure, F.  
2. Ferguson, C. A.  
3. Diglossia  
4. Hudson, A.  
5. Low variety  
6. High variety



طیب‌زاده (۱۳۹۸ الف: ۱۳) وجود دوزبان‌گونه‌گی در زبان فارسی را رد نکرده است اما معتقد است که تفاوت فارسی رسمی و غیررسمی به حدی زیاد نیست که بتوان وجود دوزبان‌گونه‌گی را در فارسی قائل شد؛ ازسویی دیگر نمی‌توان وجود آن را انکار کرد. نمونه این وضعیت را می‌توان در پیام‌های ردوبدل‌شده در پیام‌رسان‌ها در فضای مجازی دید. مکاتبات الکترونیکی در این فضا سبب شده‌است غالباً یک متن نوشتاری غیررسمی تولید و به مخاطب ارسال شود و پاسخی با همین ویژگی دریافت گردد. می‌توان چنین پنداشت که فضای مجازی سبب شده است فرایند دوزبان‌گونه‌گی در زبان فارسی به‌واسطه به‌کارگیری دو گونه نوشتاری رسمی و غیررسمی تسریع شود. از آنجاکه حجم فراوانی از این پیام‌ها، متن نوشتاری است، خط جایگاه مهمی پیدا می‌کند. بنا به تعریف طیب‌زاده (۱۳۹۸ الف: ۳)، منظور از «صورت شکسته» و «شکسته‌نویسی» آن است که «صورت شکسته به شکل نوشتاری کلمه یا وند یا پی‌بستی اطلاق می‌شود که اولاً مبین تلفظ سبک گفتاری فارسی باشد، و ثانیاً دارای معادل آوایی سالمی در زبان فارسی باشد. ... در این معنا، شکسته‌نویسی به انعکاس تلفظ فارسی گفتاری در نثر فارسی نامیده می‌شود». وی علاوه بر صورت شکسته به «فارسی سالم» نیز اشاره دارد و آن را این گونه تعریف می‌کند: «متنی که صورت کلمات آن مبین تلفظ نوشتاری یا رسمی آن‌ها است» (۱۳۹۸ الف: ۵). همچنین وی به این نکته اشاره دارد که محاوره‌گرایی (گفتاری‌نویسی) نباید با شکسته‌نویسی خلط گردد. طیب‌زاده (۱۳۹۸ الف: ۸-۹) بر این باور است که صورت‌های شکسته متون داستانی و نمایشی را می‌توان نماینده تام و تمام صورت‌های شکسته در پیام‌ها و ارتباطات اینترنتی دانست. در این پژوهش تلاش می‌شود تجلی صورت زبانی فارسی در ارتباطات الکترونیکی که دارای نونویسی در نوشتار واژه‌ها است بررسی گردد و طبقه‌بندی از آن ارائه نماید.

#### ۴. گردآوری داده‌های پژوهش

زبان‌شناسی پیکره‌ای یکی از رویکردهای نوین پژوهش در حوزه زبان‌شناسی نظری است که با هدف یافتن شواهد زبانی و همچنین ارزیابی فرضیه به کار می‌رود. منظور از پیکره زبانی، مجموعه‌ای از داده‌های زبانی است که از منابع مختلف گردآوری شده و به‌صورت الکترونیکی ضبط می‌شود و می‌توان از آن اطلاعات آماری و الگوهای زبانی را استخراج نمود. بنابراین پیکره زبانی دربرگیرنده توانش زبانی گویشوران مختلفی است که می‌تواند در مطالعات زبانی مورد استفاده قرار گیرد.

مکانری و ویلسون (۲۰۰۱) ویژگی‌های پیکره زبانی را این‌گونه معرفی می‌کنند: (۱) نماینده بودن داده‌های گردآوری‌شده. داده‌های گردآوری‌شده می‌بایست متنوع و متوازن بوده و سوگیری خاصی نداشته باشند تا بتوانند نماینده واقعی زبان تلقی شوند. (۲) حجم محدود داده‌های

گردآوری شده. داده‌های گردآوری شده با توجه به هدف پژوهش باید حجم محدودی داشته باشند. حالت<sup>۱</sup> داده‌های گردآوری شده. داده‌ها ممکن است نوشتاری و یا گفتاری باشند. داده‌های نوشتاری می‌توانند از کتاب، مجله، و یا روزنامه گردآوری شود و داده‌های گفتاری ممکن است رسمی و غیررسمی باشد. داده‌های وب و فضای مجازی نیز می‌تواند یکی از منابع تهیه تلقی گردد (کیلگاریف<sup>۲</sup> و گرفن‌اشتت<sup>۳</sup>، ۲۰۰۳). داده‌ها، چه به صورت نوشتاری و چه به صورت گفتاری می‌بایست الکترونیکی شود تا امکان تجزیه و تحلیل و پردازش آن‌ها به کمک رایانه میسر شود. ۴) معیار بودن داده‌های گردآوری شده به عنوان نماینده زبان که حاوی تنوع زبانی از نظر ژانر، سبک و سیاق باشد. این تنوع زبانی می‌تواند به صورت زبان معیار، فرامعیار یا زیرمعیار نمود پیدا کند (داگلاس<sup>۴</sup>، ۲۰۰۳). براساس معیارهای آتکینز<sup>۵</sup> و همکاران (۱۹۹۲) دو گونه نوشتاری معیار (رسمی) و غیرمعیار (غیررسمی) به واسطه خط در زبان تجلی پیدا می‌کند. گونه غیرمعیار (غیررسمی) خط را «خط شکسته» می‌نامیم<sup>۶</sup>. آنچه در زیرگونه نوشتاری شکسته تجلی پیدا می‌کند دراصل گونه گفتاری زبان است که به واسطه خط تجلی پیدا کرده است. این نوع نوشتار معمولاً در رمان‌ها و گفتگوهای بین افراد در شبکه‌های اجتماعی قابل مشاهده است. به نوشتار درآوردن گونه گفتاری و محاوره‌ای زبان اصول نگارش واژه‌ها با استفاده از خط و قواعد دستور خط را تحت تأثیر قرار می‌دهد.

برای گردآوری داده‌های این پژوهش، آن دسته از نوشته‌های کاربران شبکه‌های اجتماعی تلگرام، اینستاگرام، واتس‌آپ و توئیتر که حاوی نوعی نونویسی بود و با نوشتار معیار مواژه‌ها و دستور خط مصوب فرهنگستان زبان و ادب فارسی منطبق نبود را گردآوری کرده‌ایم. تلاش کرده‌ایم داده‌های گردآوری شده در سه موضوع متنوع اجتماعی، سیاسی و اقتصادی متوازن باشد. در نهایت، بعد از بررسی حدود ۲۰ هزار نظر، پیکره‌ای حاوی ۲۰۱۷ پاره‌گفتار که متشکل از ۱۴۵۷۱ صورت‌واژه است به دست آمد.

به منظور کاربردی تر شدن داده‌های پژوهش، چند سطح تحلیل به شکل خام و اولیه داده اضافه شده است. یکی از این سطوح، درج دستی و غیرخودکار صورت معیار برای صورت‌های دارای نونویسی است. برای درج صورت‌های معیار دستور خط مصوب فرهنگستان (۱۴۰۲) ملاک قرار گرفته است. سطح دوم، درج خودکار بن‌واژه صورت‌واژه‌ها با ابزار لمینگ است که توسط

- 
1. mode
  2. Kilgarriff, A.
  3. Grefenstette, G.
  4. Douglas, F.
  5. Atkins, S.

۶. لازم به ذکر است که اصطلاح «خط شکسته» با شیوه تحریر خط در خوشنویسی، مانند خط شکسته نستعلیق که به عنوان یک قلم در نظر گرفته می‌شود، متفاوت است

مولر<sup>۱</sup> و همکاران (۲۰۱۵) تهیه شده و توسط قیومی (۱۳۹۸) برای فارسی سازگار شده است. لایه سوم برچسب مقولات دستوری واژه‌های شکسته با ابزار مارموت مولر و همکاران (۲۰۱۳) است. برای آموزش این ابزار، از پیکره بی‌جن‌خان (۱۳۸۴) استفاده شده است که براساس ساختار برچسب‌های معرفی شده توسط قیومی (۲۰۱۲) توسعه یافته است. مقولات دستوری واژه‌ها به دو گروه واژه‌های محتوایی و نقشی قابل تقسیم است. ۴۵/۷۸٪ از واژه‌های این پیکره واژه‌های محتوایی، شامل اسم، فعل، صفت، قید، ضمیر، شاخص و شبه‌جمله، و بقیه واژه‌های دستوری، شامل حرف ربط، حرف تعریف، حرف اضافه، حرف نشانه، عدد و متفرقه است. در ادامه، پس از برچسب‌زنی‌های خودکار، همه برچسب‌ها به صورت دستی کنترل شده است. برخی آمار استخراج شده از پیکره هدف براساس تحلیل‌های فوق، در جدول ۱ گزارش شده است.

جدول ۱: آمار استخراج شده از پیکره هدف

تعداد واژه	تعداد جمله یا عبارت	تعداد واژه یکتا	تعداد بن‌واژه	متوسط جمله یا عبارت (واژه)	طول		تعداد واژگانی	تنوع واژگانی
					تعداد	تنوع		
14571	2017	5169	3055	7/22	11431	3140	0/36	0/21

در جدول ۲ توزیع آماری مقولات دستوری واژه‌ها گزارش شده است. همان‌طور که مشاهده می‌شود، مقولات دستوری اسم و فعل به ترتیب پربسامدترین و شبه‌جمله، متفرقه و شاخص کم‌بسامدترین مقوله‌های دستوری است که جمعاً حدود یک درصد از داده‌های پیکره را شامل می‌شود.

جدول ۲: توزیع آماری مقولات دستوری واژه‌ها در پیکره

مقوله دستوری	بسامد نسبی (%)	مقوله دستوری	بسامد نسبی (%)
اسم	۸۰/۳۳	حرف تعریف	۲۳/۴
فعل	۸۹/۲۲	حرف نشانه «را»	۳۰/۲
حرف اضافه	۵۹/۹	عدد	۱۴/۲
صفت	۷۸/۸	شبه‌جمله	۶۹/۰
قید	۲۶/۷	متفرقه	۳۰/۰
حرف ربط	۴۰/۶	شاخص	۰۳/۰
ضمیر	۶۷/۵		

1. Müller, T.

در جدول ۳ چند نمونه از پیکره داده‌های فضای مجازی به همراه صورت معیار آنها آورده شده است.

جدول ۳. نمونه داده‌های فارسی در فضای مجازی

ردیف	داده اصلی	صورت سالم
1	اتفاقاً منم تو بد شرایطیم	اتفاقاً من هم توی بد شرایطی هستم
2	عه اون سودابه بود لفت داد	ا آن سودابه بود لفت داد
3	جاتووووون خالیااااا	جاتون خالی ها
4	حقوقش اونقد نیس ک کفاف زندگیمونو بده	حقوقش آنقدر نیست که کفاف زندگیمان را بدهد
5	اومده تحقیق دربارت	آمده است تحقیق درباره‌ات
6	خونت کوچیک	خانه‌ات کوچک است
7	دو روز کات کردم	دو روز است [که] کات کرده‌ام
8	من و مقصر میدونن	من را مقصر می‌دانند
9	دگ بریدم	دیگر بریده‌ام
10	تو رو دوس دارم	تو را دوست دارم
11	نزاشت من خواهرمو ببینم	نگذاشت من خواهرم را ببینم
12	کاره پسره تو مبله	کار پسر توی مبل است

## ۵. تحلیل داده‌ها

تحلیل ارائه‌شده در دو سطح است: یکی در سطح خط و ویژگی‌های نگارشی و دیگری براساس ویژگی‌های آوایی-ساختوازی. در جدول ۴ گزارشی از اعمال قواعد براساس پیکره حاصل از فضای مجازی گزارش شده است.

جدول ۴: نتایج کلی از اعمال قواعد

داده‌های حاوی تغییر		داده‌های بدون تغییر
با تکرار قواعد	عدم تکرار قواعد	
10/08	33/02	56/90

اگر بخواهیم نمایی کلی از قواعد اعمال‌شده بر این پیکره ارائه دهیم، مشاهده می‌کنیم که در ۹۰/۵۶ درصد از داده‌ها هیچ‌گونه تغییری اعمال نشده است و صورت نوشتاری گویشوران فارسی‌زبان در فضای مجازی با صورت نوشتاری معیار مصوب فرهنگستان زبان و ادب فارسی یکسان است. این عدد به این مفهوم است ۱۰/۴۳ درصد از واژه‌های نوشته‌شده در متون

منتشر شده در فضای مجازی دارای شکسته‌نویسی یا نونویسی است که عدد قابل توجهی است و اهمیت بررسی آن را مشخص می‌کند. ۰۲/۳۳ درصد از واژه‌ها گاهی اوقات ممکن است در یک واژه چندین بار نونویسی رخ دهد که براساس داده‌های تحلیل شده در این پژوهش، حدود ۱۰ درصد از قاعده‌های شکسته‌نویسی، امکان تکرار یک قاعده در یک واژه را دارد.

نتایج تحلیل داده‌ها از نظر تکرارپذیری یا اعمال تکی قواعد در هر دو سطح تحلیل در جدول ۵ گزارش شده است. با احتساب ویژگی تکرارپذیری قاعده‌ها، به تحلیل پیکره فضای مجازی می‌پردازیم. چنانچه بخواهیم تصویر واضحی از اعمال این قواعد بر داده‌های فضای مجازی داشته باشیم نیاز به تفکیک این دو سطح تحلیل داریم. به طور کلی، ۲۴/۱۳ درصد از مجموع تغییرات حاوی شکسته‌نویسی یا نونویسی در داده‌های فضای مجازی در سطح خط و ویژگی‌های نگارشی است؛ و ۰۴/۳۶ درصد از مجموع تغییرات براساس ویژگی‌های آوایی-ساختوازی است. ۵۴/۱۲ درصد از قواعد تکی، به تغییرات حاوی شکسته‌نویسی واژه‌ها براساس ویژگی‌های نگارشی تعلق دارد و ۸۵/۳۱ درصد از قواعد تکی به ویژگی‌های آوایی-ساختوازی متعلق است.

جدول ۵: نتایج تحلیل داده‌ها از نظر تکرارپذیری یا اعمال تکی قواعد هر دو سطح تحلیل

ویژگی‌های آوایی-ساختوازی		ویژگی‌های نگارشی	
عدم تکرار قواعد	با تکرار قواعد	عدم تکرار قواعد	با تکرار قواعد
12/54	0/70	31/85	4/19
13/24		36/04	

نکته قابل تأمل این است که براساس داده‌های تحلیل شده، بیشتر گزارش بود که به‌طورکلی، ۱۰/۴۳ درصد از محتوای گردآوری شده از فضای مجازی دارای نونویسی است. درحالی‌که با تجمیع تحلیل داده‌ها از نظر تکرارپذیری یا اعمال تکی قواعد در سطح نگارشی و همچنین آوایی-ساختوازی در جدول ۴، به عدد ۲۸/۴۹ درصد می‌رسیم. اختلاف این دو عدد بیانگر این نکته است که در ۲۹/۱ درصد از داده‌ها، چند قاعده متفاوت، ولی بدون تکرار، در یک واژه اتفاق افتاده است و در سایر موارد، ۸۹/۴ درصد از داده‌ها، چند قاعده متفاوت ولی تکرارپذیر بر روی یک واژه اعمال شده است. مثال‌های ۱۳ تا ۱۶ شامل شواهدی است که دارای این ویژگی است. شایان ذکر است معیار اعمال قواعد در بافت زبانی و نه یک واژه منفک از بافت است. در مثالی مانند «این وضعیت خیلی اسفناک» اگرچه فعل نمود نگارشی ندارد ولی در /asafnAke/? نمود آوایی دارد؛ بنابراین، این مثال یک جمله است و نگارش فعل «است» حذف شده است.

۱۳: بینوه (شکل سالم: بی.ام.و.)

تغییرات نگارشی اعمال‌شده در این واژه حذف، درج و تغییر در مقایسه با شکل معیار نگارش «بی‌ام‌و» است. همچنین به نظر می‌رسد نوعی گونه‌زبانی یا گویشی خاصی در شیوه نگارش این واژه نهفته است.

۱۴: «ایندت» (شکل سالم: آینده‌ات را)

تغییرات نگارشی اعمال‌شده در این واژه عبارت است از جایگزینی نگارش «ا» به جای «آ» و دو تغییر ساختوازی که شامل نگارش گونه‌های مختلف ضمائر متصل پس از واژه‌های منتهی به «های» غیرملفوظ و حذف تکواژ "را" در گفتار است.

۱۵: «عصفاک» (شکل سالم: اسفناک)

تغییرات نگارشی اعمال‌شده در این واژه عبارت است از حذف نویسه «است»، دو تغییر نگارشی که شامل تغییر کاربرد «ع» به جای «ا» و دیگری «ص» به جای «س» است.

۱۶: «اونور ابشون» (شکل سالم: آن‌ورآبشان)

تغییرات نگارشی اعمال‌شده در این واژه عبارت است از دو تغییر نگارشی در نگارش «ا» به جای «آ»، یک تغییر ساختوازی با تبدیل /an/ به /un/ در واژه «آن» و یک تغییر آوایی با تبدیل /an/ به /un/ در واژه‌بست «-شان».

در جدول ۶ توزیع آماری مقولات دستوری واژه‌ها در دو سطح تحلیل گزارش شده‌است. همان‌طور که مشاهده می‌گردد، نونویسی در ۸۱/۴۶ درصد از واژه‌های محتوایی و ۷۲/۸ درصد از واژه‌های دستوری اتفاق افتاده‌است. سه مقوله دستوری «فعل» و «اسم» به ترتیب پرمسامدترین مقوله‌های دستوری است که نونویسی واژه‌ها بر روی آنها اعمال شده‌است.

جدول ۶: توزیع آماری مقولات دستوری واژه‌ها در دو سطح تحلیل

مقوله دستوری	بسامد نسبی (%)	مقوله دستوری	بسامد نسبی (%)
فعل	12/90	حرف ربط	1/15
اسم	11/26	عدد	1/15
صفت	3/79	حرف نشانه «را»	1/06
حرف اضافه	3/64	شبه‌جمله	0/42
قید	3/02	متفرقه	0/28
ضمیر	2/83	شاخص	0/01
حرف تعریف	1/45		

## ۵-۱. فراتحلیل داده‌ها براساس ویژگی‌های نگارشی

تحلیل نگارشی سطحی از تحلیل است که در آن شیوه نگارش داده اصلی و معادل معیارشده آن واژه را در پیکره مقایسه می‌کنیم تا مشخص گردد هنگام نگارش واژه‌ها در فضای مجازی (داده اصلی) نسبت به داده معیار چه نوع تغییری به لحاظ خط اعمال شده است. برای این هدف، پنج عاملی که سبب تغییر در شیوه نگارش معیار واژه‌ها در فضای مجازی می‌شود را شناسایی کرده‌ایم که در جدول ۷، نتایج حاصل از تحلیل داده‌ها براساس ویژگی‌های نگارشی گزارش شده است. عوامل شناسایی شده مربوط به شکسته‌نویسی یا نونویسی واژه‌ها عبارت است از: الف) حذف حروف، مانند «راجب» به جای شکل معیار نوشتاری «راجع به»، ب) تغییر حروف، مانند «اصطلاک» به جای شکل معیار نوشتاری «استهلاک»، پ) درج حروف، مانند «فوهش» به جای شکل معیار نوشتاری «فحش»، ت) کاربردشناختی حروف، مانند «عامریکا» به جای شکل معیار نوشتاری «آمریکا» و ث) تکرار حروف، مانند «خبرررر» به جای شکل معیار نوشتاری «خبر».

همانطور که در نتایج مشخص است، حذف حروف پرکاربردترین و رایج‌ترین ویژگی نونویسی است که در داده‌های فضای مجازی اتفاق افتاده است. به نظر می‌رسد دلیل حذف حروف بیشتر برای تسهیل نگارش الکترونیکی و سرعت در تحریر حروف باشد. کم‌کاربردترین ویژگی نگارشی تکرار حروف است که ممکن است برای تأکید یا بیان رابطه صمیمانه بین نویسنده و مخاطب به کار می‌رود. تغییر حروف نیز ویژگی دیگر نگارشی است که با هدف تسهیل در نگارش الکترونیکی اتفاق می‌افتد. درج حروف و تغییر کاربردشناختی از دیگر ویژگی‌های نگارشی است.

جدول ۷: نتایج تحلیل داده‌ها براساس ویژگی‌های نگارشی

تکرار حروف	تغییر کاربردشناختی	درج حروف	تغییر حروف	حذف حروف
46	143	246	463	1142
0/32	0/98	1/69	3/18	7/84

مثال‌هایی که برای حذف حروف در پیکره گردآوری شده یافت شد در جدول ۸ همراه با صورت معیار، بسامد مطلق و بسامد نسبی گزارش شده است.

جدول ۸: نمونه‌های پربسامد دارای حذف حروف

واژه شکسته	صورت معیار	بسامد مطلق	بسامد نسبی	واژه شکسته	صورت معیار	بسامد مطلق	بسامد نسبی
ب	به	105	0/0072	واقعا	واقعاً	10	0/0007
ک	که	89	0/0061	فعلا	فعالاً	10	0/0007
ی	یک	46	0/0032	کلا	کلاً	9	0/0006
چ	چه	25	0/0017	دگ	دیگر	9	0/0006
ن	نه	23	0/0016	خودت	خودت است	8	0/0005
خب	خوب	18	0/0012	حتما	حتماً	8	0/0005
اصلا	اصلاً	15	0/0010	بهرتر	بهرتر است	8	0/0005
دیگ	دیگر	13	0/0009	واس	واسه	7	0/0005
مردم	مردم است	11	0/0008	راجب	راجع‌به	7	0/0005
بخاطر	به‌خاطر	11	0/0008	بجای	به‌جای	7	0/0005

## ۲-۵. فراتحلیل داده‌ها براساس ویژگی‌های آوایی-ساختواژی

تحلیل آوایی-ساختواژی سطحی از تحلیل است که در آن برچسب آوایی داده اصلی و معادل معیارشده آن را در پیکره مقایسه می‌کنیم تا مشخص گردد هنگام کاربرد واژه‌ها در فضای مجازی (داده اصلی) نسبت به داده معیار چه نوع تغییر آوایی و ساختواژی اعمال شده است. برای پاسخ به این پرسش ۷ تغییر مد نظر قرار گرفته شده است که در ادامه توضیح داده می‌شود:

۱. تکرار آوا: تکرار واکه در یک واژه (فقط واکه و نه همخوان)، مانند «جااااااان»؛
۲. حذف تکواژ: حذف تکواژ از یک پاره‌گفتار، مانند «آمده» به جای «آمده است»؛
۳. درج آوا: اضافه شدن یک آوا به ساختمان واژه، مانند «کتابه من» به جای «کتاب من»؛
۴. تغییر آوایی: تبدیل شدن یک آوا به آوایی دیگر، مانند «اومد» به جای «آمد»؛
۵. کاربرد گویشی: کاربرد واژه‌ای از گویش یا زبانی غیر از فارسی، مانند «فیک»؛
۶. کاهش آوایی: حذف آوایی از ساختمان واژه، مانند «یه» به جای «یک»؛
۷. تغییر ساختواژی: استفاده از یک تکواژگونه به جای دیگری، مانند «فعالیت» به جای «فعالیت است».

روش تحلیل داده‌ها براساس ویژگی‌های آوایی-ساختواژی به این صورت است ابتدا هر داده‌ای که آوانویسی صورت معیار آن با آوانویسی صورت اصلی اش متفاوت باشد باید یکی از فرایندهای بالا بر روی آن اعمال شود. در این روند، آوانویسی داده معیار به‌عنوان داده پایه قرار



می‌گیرد و آوانویسی داده اصلی با توجه به آن تحلیل می‌شود. اگر در یک داده، یک تحلیل دوبار یا سه بار تکرار شود، به ازای تعداد دفعات تکرار، به آن تحلیل، وزن داده می‌شود. چنانچه در یک داده، دو یا چند نوع تحلیل مختلف ارائه شده باشد، هر یک از تحلیل‌ها جداگانه لحاظ می‌گردد.

خلاصه نتایج حاصل از تحلیل پیکره تهیه شده براساس ویژگی‌های آوایی-ساختواژی در جدول ۹ گزارش شده است. ۶۰/۴۰ درصد از تغییرات اعمال شده در پیکره منتخب در حوزه تغییرات آوایی-ساختواژی است. همانطور که مشاهده می‌گردد، تغییر ساختواژی بیشترین ویژگی ساختواژی-آوایی است که در حدود ۲۸ درصد از داده‌های پیکره اتفاق افتاده است. ویژگی‌های آوایی اعم از کاهش و تغییر آوایی و همچنین تکرار و درج آوا ۷۶/۹ درصد از تحلیل داده‌ها را شکل داده است. حذف تکواژ کمتر از ۲ درصد از واژه‌های پیکره را شامل می‌شود و کمتر از یک درصد از داده‌ها حاوی ویژگی گویشی است.

جدول ۹: نتایج تحلیل داده‌ها براساس ویژگی‌های آوایی-ساختواژی

تغییر ساختواژی	کاهش آوا	تغییر آوا	حذف تکواژ	گویش	تکرار آوا	درج آوا	
4086	717	662	280	128	33	10	بسامد مطلق
28/04	4/92	4/54	1/92	0/88	0/23	0/07	بسامد نسبی (%)

## ۶. نتیجه گیری

در این پژوهش به موضوع نونویسی واژه‌ها و عبارات فارسی و تحلیل این گونه از داده‌ها که نمونه‌های آن را به‌وفور می‌توان در شبکه‌های اجتماعی ملاحظه کرد پرداختیم. برای این هدف، یک پیکره زبانی از این گونه داده گردآوری شد و به‌صورت دستی شکل معیار داده را به‌دست آوریم و به‌صورت الگوریتمی به برچسب‌گذاری داده‌ها و تحلیل آن پرداختیم.

براساس داده‌های تحلیل شده، در بیش از نیمی از واژه‌های پیکره، نونویسی واژه‌ها رخ نداده و دستور خط معیار در نگارش واژه‌ها لحاظ شده است. رعایت دستور خط معیار در این حجم از داده بیانگر این نکته است که همه واژه‌ها امکان نونویسی و انعطاف‌پذیری لازم برای این هدف را ندارد. به‌طورکلی دو ویژگی عمده نگارشی و آوایی-ساختواژی در شکسته‌نویسی واژه‌ها مشاهده شد و برای زیرشاخه‌های هریک از ویژگی‌ها مشخص شد. دو نتیجه‌ای که از تحلیل داده‌ها به‌دست آمد عبارت است از:

۱- نونویسی واژه‌ها براساس ویژگی‌های آوایی-ساختواژی در پیکره متداول‌تر از ویژگی نگارشی بود؛ درحالی‌که به نظر می‌رسد به دلیل تبادل اطلاعات از طریق خط در فضای مجازی، نونویسی واژه‌ها از نوع نگارشی باید متداول‌تر باشد که اینگونه نبود. این نکته به این مفهوم است که تهیه ابزار در پژوهش‌های کاربردی نباید در حد پردازش خط باقی بماند و لحاظ کردن اطلاعات زبان‌شناسی آوایی و ساختواژی از اهمیت شایانی برخوردار است.

۲- از میان واژه‌های محتوایی و دستوری، اگرچه واژه‌های دستوری پرکاربردتر است، نونویسی در واژه‌های محتوایی بسیار متداول‌تر است و واژه‌های دارای مقولات دستوری «فعل» و «اسم» بیشترین درصد از نونویسی را شامل می‌شود.

## منابع

- ادیبان، مجید، و ممتازی، سعیده (۱۴۰۱) «تبدیل متن محاوره به رسمی فارسی با استفاده از شبکه‌های عصبی مبتنی بر مبدل»، مجله زبان و زبان‌شناسی، ۱۸ (۳۵): ۴۵-۶۷.
- بی‌جن‌خان، محمود (۱۳۸۴) "نقش پیکره زبانی در نوشتن دستور زبانک معرفی یک نرم‌افزار رایانه‌ای"، مجله زبان‌شناسی، ۱۹ (۳۷): ۴۸-۶۷.
- بی‌جن‌خان، محمود (۱۳۹۱) "خط و زبان فارسی در فضای مجازی"، همایش محتوای ملی در فضای مجازی. کنسرسیوم محتوای ملی، کتابخانه ملی ایران، تهران.
- دستور خط مصوب فرهنگستان (۱۴۰۲) دستور خط مصوب فرهنگستان. تهران: فرهنگستان زبان و ادب فارسی.
- سمیعی‌گیلانی، احمد (۱۳۹۱) نگارش و ویرایش. تهران: سازمان مطالعه و تدوین کتب علوم انسانی دانشگاه‌ها (سمت).
- شوهانی، علیرضا، و حسینی، سارا (۱۳۹۷) «بررسی وجود تأثیر فضای مجازی بر زبان و ادبیات فارسی معاصر»، نشریه زبان و ادب فارسی، ۷۱ (۲۳۸): ۷۵-۱۰۱.
- صلح‌جو، علی (۱۳۸۶) "بشکنیم یا نشکنیم". فصلنامه مترجم، ۴۵ (۱۷): ۹-۲۲.
- صلح‌جو، علی (۱۳۹۱) اصول شکسته‌نویسی: راهنمای شکستن واژه‌ها در گفت‌وگوهای داستان. تهران: نشر مرکز.
- طیب‌زاده، امید (۱۳۹۸الف) مبانی و دستور خط فارسی شکسته براساس صد سال آثار داستانی و نمایشی. تهران: پژوهشگاه علوم انسانی و مطالعات فرهنگی.
- طیب‌زاده، امید (۱۳۹۸ب) فارسی شکسته: دستور خط و فرهنگ املائی. تهران: کتاب بهار.
- طیب‌زاده، امید (۱۳۹۹) «تغییرات آوایی و صورت‌های شکسته در فضای مجازی و استفاده از آنها در فرهنگ‌های جامع زبان فارسی»، دستور ویژه‌نامه نامه فرهنگستان، ۱۶: ۱۷۵-۱۹۴.
- غفاری، مهسا (۱۳۹۴) «واکوی تأثیر شلخته‌نویسی فضای مجازی بر زبان فارسی (از قافیه‌نویسی رودکی تا شلخته‌نویسی امروز)»، روزنامه عطر یاس، شماره ۶۷۵، تاریخ ۱۸/۱۰/۱۳۹۴، ص ۴.

- قیومی، مسعود (۱۳۹۸) «گذار از بن‌واژه‌سازی قاعده‌مند به آماری در فارسی» در مجموعه مقالات پنجمین همایش زبان‌شناسی رایانشی. صص: ۵۷-۸۶، تهران: نشر نویسه پارسی.
- مسگرخویی، مریم (۱۳۹۹) «پیش‌نیازهای بررسی آسیب‌شناختی خط و زبان فارسی در فضای مجازی»، دستور ویژه‌نامه نامه فرهنگستان، ۱۶: ۱۴۹-۱۷۴.
- هدایت مفیدی، مسحه، کامیابی گل، عطیه، و علیزاده، علی (۱۳۹۶) «فضای مجازی و زبان فارسی: غیرمعیارهای نوشتاری در شبکه‌ی اجتماعی تلگرام»، مطالعات رسانه‌ای، ۱۲: ۶۵-۸۲.
- Anis, J. (1999) *Internet, Communication and French Language*. Paris: Hermes.
- Ariffin, S. N. A. N., and Tiun, S. (2020) "Rule-based text normalization for Malay social media texts," *International Journal of Advanced Computer Science and Applications*. 11(10), 156-162
- Armin, N., & Shamsfard, M. (2011) "Converting Persian colloquium text to formal by n-grams," *Computer Society of Iran. for statistical machine translation*. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp: 1724-1734.
- Atkins, S., Clear, J., and Ostler, N. (1992) "Corpus design criteria," *Literary and Linguistic Computing*, 7(1): 1-16.
- Dejong, A. and Mercier, J. (2002) *French cyber@ngue*. Brussels, The Renaissance of the Book.
- Douglas, F. (2003) "The Scottish corpus of texts and speech: Problems of corpus design," *Literary and Linguistic Computing*, 18, 23-37.
- de Saussure, F. (1916) *Cours de linguistique générale*. Lausanne, Paris: Payot.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019), "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171-4186.
- Ferguson, C. A. (1959) *Diglossia*, *WORD*, 15:2, 325-340
- Ghayoomi, M. (2012) "Bootstrapping the development of an HPSG-based treebank for Persian," *Linguistic Issues in Language Technology*, 7(1).
- Hudson, A. (2002) "Outline of a theory of diglossia," *International Journal of the Sociology of Language*, 157: 1-48.
- Kashefi, O. (2020) "MIZAN: A large Persian-English parallel corpus," <https://arxiv.org/abs/1801.02107>
- Kilgarriff, A. and Grefenstette, G. (2003) "Introduction to the special issue on the Web as Corpus," *Computational Linguistics*, 29, 333-348.
- Kozhimbayev, Z. and Yessenbayev, Z. (2020) "Kazakh text normalization using machine translation approaches," *CEUR Workshop Proceedings*, Vol. 2780, CEUR-WS, 115-122.
- Lazar, J. (2012) "Quelques observations sur les néographies photisantes en français tchaté," *Linguistica Pragensia*, 22(1): 18-28.
- Lindemann, S. (2005) "Who speaks 'Broken English'? US undergraduates' perception of non-native English." *International Journal of Applied Linguistics*, vol. 15(2): 187-212.

- Mansfield, C., Sun, M., Sun, M., Liu, Y., Gandhe, A., & Hoffmeister, B. (2019) "Neural text normalization with subword units," In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Volume 2, pp: 190-196.
- Masoumi, V., Salehi, M., Veisi, H., Haddadian, G., Ranjbar, V., & Sahebdel, M. (2020) "TeleCrowd: A Crowdsourcing Approach to Create Informal to Formal Text Corpora," arXiv preprint arXiv:2004.11771.
- McEnery, T. and Wilson, A. (2001), *Corpus Linguistics: An Introduction*, Edinburgh University Press.
- Müller, T., Cotterell, R., Fraser, A., & Schütze, H. (2015). Joint lemmatization and morphological tagging with lemming. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 2268-2274). Lisbon, Portugal. Association for Computational Linguistics.
- Müller, T., Schmid, H., & Schütze, H. (2013). Efficient higher-order CRFs for morphological tagging. In Proceedings of the 2013 Conference on Empirical Methods in Natural language Processing (pp. 322-332). Seattle, Washington, USA. Association for Computational Linguistics.
- Rasooli, M. S., et al. (2020). Automatic Standardization of Colloquial Persian. arXiv preprint arXiv:2012.05879.