

## واژه‌های دستوری به مثابه نشانگرهای گویش فردی: رویکردی پیکرهای به شناسایی هویت نویسنده در زبان

فارسی

رامین گلشنائی\*

استادیار زبان‌شناسی، دانشگاه الزهرا، تهران، ایران.

پذیرش: ۹۶/۱۲/۲۱

دریافت: ۹۶/۸/۲۱

### چکیده

شناسایی هویت نویسنده یکی از حوزه‌های تحقیقاتی مهم در حیطه زبان‌شناسی حقوقی است که موضوع پژوهش‌های زبان‌شناختی و رایانشی گستردۀ در زبان‌های مختلف بوده است. با این حال شواهد محدودی از پژوهش‌های صورت‌گرفته با موضوع شناسایی نویسنده در زبان فارسی وجود دارد. در این پژوهش امکان شناسایی نویسنده یک متن با تکیه بر مفهوم گویش فردی و با استفاده از واژه‌های دستوری زبان فارسی بررسی شده است. واژه‌های دستوری از آن جهت که بهطور ناخودآگاه در تولید زبان به کار گرفته می‌شوند، مستقل از موضوع متن به کار می‌روند و بسامد بالایی در متون کوتاه دارند، می‌توانند نشانگرهای مؤثری برای کنگذاری گویش فردی و ممیز سبک نویسنده‌گان باشند. در این پژوهش، ابتدا پیکرهایی متنی از نوشتۀ‌های پنج محقق و نویسنده معاصر جمع‌آوری و سپس با استفاده از نرم‌افزار خطایاب وفا، استانداردسازی شدند. با استفاده از بسته سبک‌سنجی *stylo* نرم‌افزار آماری *R* واژه‌های پربسامد دستوری با توالی‌های یک تا سه‌گاشتشی از متون استخراج شدند و سپس قابلیت تفکیک متون بر اساس این واژه‌ها و با استفاده از تحلیل مؤلفه‌های اصلی و همچنین تحلیل خوش‌های بر اساس مقیاس فاصله‌ای دلتا بررسی شد. نتایج نشان داد که واژه‌های تکنگاشتشی بهتر از دونگاشتشی و سه‌گاشتشی‌ها در متون کم‌حجم است. نویسنده را دارند و عملکرد واژه‌های تکنگاشتشی بهتر از دونگاشتشی و سه‌گاشتشی‌ها در متون کم‌حجم است. همچنین نتایج پژوهش نشان داد که حجم کمینه متن برای شناسایی موقفيت‌آمیز نویسنده در متون فارسی حدود ۴۰۰۰ واژه بر اساس ۲۰ واژه دستوری پربسامد است.

واژه‌های کلیدی: گویش فردی، شناسایی نویسنده، تحلیل پیکرهای، زبان‌شناسی حقوقی، روش دلتا.

## ۱. مقدمه

شناسایی اصالت نوشته‌ها و کشف هویت نویسنده متون از دیرباز مورد توجه انسان بوده است. یکی از قیمت‌ترین نمونه‌های آن، اثبات جعلی بودن سند «بخشش کنستانتین» توسط لورنزو والا در سال ۱۴۴۰ میلادی است که بر اساس این سند ادعا شده بود کنستانتین (امپراطور روم) حکومت روم و قسمت غربی امپراطوری روم را به پاپ سپرده است (Frontini, Lynch, & Vogel, 2008). پژوهش‌هایی که امروزه در این حوزه انجام می‌شوند، عمدتاً با عنوان سبک‌سنجی<sup>۲</sup> شناخته می‌شوند. در عصر کنونی با تحولی که رایانه در پردازش متون حجیم ایجاد کرده و کاربردهایی که پژوهش‌های سبک‌سنجی در زبان‌شناسی حقوقی<sup>۳</sup> (Totty & Hardcastle, 1987) و کشف سرقت علمی و ادبی<sup>۴</sup> (Stein & Meyer zu Eissen, 2007) پیدا کرده است، بررسی نظاممند و دقیق متون نوشتاری در زبان‌های مختلف می‌تواند به افزایش درک کنونی محققان نسبت به جوانب مختلف سبک‌های زبانی منجر شود.

پژوهش‌های نوین سبک‌سنجی که بر تحلیلهای رایانشی مبتنی هستند، بر مؤلفه‌های مختلفی از متون تمرکز می‌کنند. این مؤلفه‌ها می‌توانند در سطح واژگانی (غنای واژه‌ها، بسامد واژه‌ها، غیره)، نحوی (ساخت جمله، بسامد قواعد نحوی، خطاهای، اجزای کلام)، و معنایی (هممعناها، وابستگی‌های معنایی) باشند (Stamatatos, 2009). تحقیقات انجام‌گرفته در باب مؤلفه‌های واژگانی نشان می‌دهد که نوع به خصوصی از واژه‌ها، یعنی واژه‌های دستوری، ویژگی‌های منحصر‌به‌فردی برای شناسایی سبک نویسنده دارد. برخی از این ویژگی‌ها عبارت‌اند از: بسامد بالای آن‌ها در متن که دست پژوهشگر را در تحلیل متون کوتاه‌تر باز می‌گذارد، عدم تأثیرپذیری واژه‌های نقشی از موضوع یا ژانر نوشته، و عدم کنترل خودآگاه نویسنده بر کاربرد واژه‌های دستوری (Binongo, 2003)، با اینکه در زبان انگلیسی و زبان‌های دیگر تحقیقات متعددی درباره نقش واژه‌های دستوری در شناسایی سبک نویسنده انجام شده (Argamon & Levitan, 2005; Burrows, 1987; Holmes, Gordon, & Wilson, 2001; Mosteller & Wallace, 1964; Segarra, Eisen, & Ribeiro, 2015; Zhao & Zobel, 2005)؛ اما در زبان فارسی پژوهش‌هایی چندانی در باب این موضوع صورت نگرفته است. از سوی دیگر، از آن‌جا که واژه‌های دستوری در زبان‌های مختلف به شکل‌های مختلفی نمود می‌یابند، پژوهش‌هایی که نقش مؤثر واژه‌های دستوری در تفکیک سبک‌های نوشتاری در یک زبان خاص را تأیید می‌کنند، لزوماً به معنای تأیید آن در زبان‌های دیگر از جمله زبان فارسی نمی‌تواند باشد. در این پژوهش، کارایی واژه‌های دستوری زبان فارسی در تفکیک سبک زبانی نویسنده‌های

مختلف بررسی شده است. مسئله این پژوهش در قالب سه پرسش قابل تعریف است که عبارت‌اند از: ۱) آیا واژه‌های دستوری در نثر فارسی قابلیت تفکیک سبک نویسنده‌ها را دارند؟ ۲) کدامیک از توالی‌های یک تا سه واژه‌ای (تکنگاشتی<sup>۱</sup>، دونگاشتی<sup>۲</sup> و سه‌نگاشتی<sup>۳</sup>) از واژه‌های دستوری فارسی در تفکیک سبک‌های نویسنده‌ها موفق‌تر عمل می‌کنند؟ ۳) حداقل کلمات مورد نیاز برای تفکیک موفقیت‌آمیز سبک‌های نوشتاری بر اساس واژه‌های دستوری چقدر است؟ برای پاسخ به این پرسش‌ها، پیکرهایی متنی از نوشته‌های پنج محقق و نویسنده معاصر جمع‌آوری و سپس با استفاده از نرم‌افزار خطایاب وفا، استانداردسازی شدند. با استفاده از بسته تحلیل آماری style در نرم‌افزار<sup>۴</sup>، واژه‌های پربسامد دستوری با توالی‌های یک تا سه‌نگاشتی از متون استخراج شدند و سپس قابلیت تفکیک متون بر اساس این واژه‌ها و با استفاده از تحلیل مؤلفه‌های اصلی و همچنین تحلیل خوش‌های بر اساس مقیاس فاصله‌ای دلتا بررسی شد. در بخش بعد پیشینه برخی پژوهش‌های انجام‌شده در حوزه سبک‌سنجی و موضوعات مرتبط معرفی خواهد شد. در بخش ۳، مبانی نظری و روش پژوهش شرح داده خواهد شد. بخش ۴ به ارائه نتایج تحلیل داده‌ها می‌پردازد و در نهایت بخش ۵ به بحث و جمع‌بندی نتایج پژوهش اختصاص دارد.

## ۲. پیشینه پژوهش

در بسیاری از پژوهش‌های انجام‌شده در حوزه تشخیص هویت نویسنده، محققان از نشانگرهای سبکی<sup>۵</sup> متعددی با عنوان نشانگرهای واژگانی، نحوی و معنایی به عنوان شاخص‌های ممیز سبک نویسنده استفاده کرده‌اند. نمونه معروف مطالعاتی که از نشانگرهای واژگانی در شناسایی نویسنده استفاده کرده، مربوط به شناسایی نویسنده ۱۲ مقاله بدون نویسنده از میان «مقالات فدرالیست» است که طی سال‌های ۱۷۸۷ تا ۱۷۸۸ منتشر شدند تا شهروندان ایالت نیویورک را ترغیب کنند به قانون اساسی رأی مثبت دهند. این مقالات توسط سه نفر یعنی الکساندر همیلتون، جان جی و جیمز مدیسون نوشته شده بودند؛ اما به طور ناشناس منتشر شدند. در مطالعه‌ای که با استفاده از روش‌های آماری و به کارگیری بسامد واژه‌های دستوری به مثابه نشانگرهای واژگانی صورت گرفت (Mosteller & Wallace, 1964)، مشخص شد که ۱۲ مقاله توسط مدیسون، یکی از سه نویسنده مقالات، نوشته شده است. در حوزه زبان فارسی می‌توان به پژوهشی اشاره کرد که بر

اساس واژه‌های دستوری پرسامد و با استفاده از روش‌های آماری به تفکیک سبک نگارشی نظامی گنجوی/ شهریار و عبدالحسین زرین‌کوب/ سیمین دانشور پرداخته است (Modaber Dabagh, 2007). نتایج این پژوهش که بر روی نظم و نثر فارسی انجام شده است، نشان می‌دهد واژه‌های دستوری قادر به تفکیک سبک نویسنده‌ها هستند. نشانگر واژگانی دیگری که مبتنی بر بسامد واژگانی عمل می‌کند، نشانگر غنای واژگانی<sup>۱</sup> است. این شاخص که تنوع واژه‌های مورداستفاده توسط فرد را نشان می‌دهد معمولاً با تقسیم تعداد واژه‌های منحصر به فرد (گونه) متن بر تعداد کل کلمات (نمونه) به کار رفته به دست می‌آید (Johansson, 2008). در مورد غنای واژگانی نیز، تحقیقات صورت گرفته در حوزهٔ روان‌شناسی زبان نشان می‌دهد افراد مختلف غنای واژگانی متفاوتی دارند (Carroll, 2008). شاخص غنای واژگانی معمولاً در کتاب نشانگرهای دیگر متنی برای تفکیک سبک نگارشی نویسنده‌ها مورد استفاده قرار گرفته است (Farahmandpour & Nikmehr, 2015). اگرچه شاخص غنای واژگانی می‌تواند تصویری از تنوع واژگانی افراد به دست دهد؛ اما به دلیل تحت‌تأثیر قرار گرفتن این شاخص از اندازهٔ متن، در به کارگیری آن باید جانب احتیاط نگه داشته شود و حجم متون مورد مقایسه حدالامکان باید یکسان باشد (*ibid*).

در برخی دیگر از پژوهش‌ها از نشانگرهای نحوی نظری اجزای کلام<sup>۱</sup> به عنوان ممیز سبک نویسنده استفاده کرده‌اند (Gamon, 2004; Koppel & Schler, 2003). نشانگرهای نحوی همانطور که از نامش پیداست، بر مبنای ساختار نحوی و وابستگی‌های دستوری بین عبارات زبانی عمل می‌کند. پیش‌فرض استفاده از نشانگرهای نحوی در تشخیص هویت نویسنده این است که افراد در تولید زبان از الگوهای نحوی کمابیش مشابه استفاده می‌کنند. نکته‌ای که استفاده از رویکردی نحوی‌بینیاد را با چالش مواجه می‌کند، استفاده از ابزارهای خودکار تحلیل نحوی است. در واقع برای اینکه بتوان در تحلیل متنی از ساختار نحوی بهره جست، نیاز است که با استفاده از نرم‌افزارهای رایانشی خاص (نظری تقطیع‌گر نحوی یا برچسب‌زن نحوی) بازنمایی نحوی جمله را به دست دهیم. این پیش‌پردازش در بسیاری از زبان‌ها به دلیل محدودیت در دسترسی به نرم‌افزارهای لازم امکان‌پذیر نیست.

مشخصه‌های معنایی متن، یکی دیگر از مؤلفه‌هایی هستند که این پتانسیل را دارند تا برخی مشخصه‌های سبکی نویسنده یک متن را فاش کنند. این نشانگرهای معنایی می‌توانند طیف وسیعی از مشخصه‌ها نظری شمار دستوری اسمی، زمان و نمود دستوری فعل و مشخصه‌های

زیرمقوله‌ای فعل را شامل شوند (Gamone, 2004). برخی پژوهش‌ها همچنین از رویکردهای نظاممندتر و دستوربینان‌تری به این مسئله استفاده کرده‌اند. مثلاً در برخی پژوهش‌ها (Whitelaw & Argamon, 2004; Whitelaw & Patrick, 2004) مجموعه‌ای از مشخصه‌های معنایی بر اساس دستور نظاممند نقشگرا<sup>۱۱</sup> تعریف شده‌اند که کلمات یا عبارات زبانی بهخصوصی را با اطلاعات معنایی مرتبط می‌کنند. برخی مطالعات اخیر از معناشناسی قاب‌بینان نظریه‌ای جدید در معناشناسی بهویژه در آثار ترجمه‌شده، استفاده کرده‌اند. معناشناسی قاب‌بینان نظریه‌ای جدید در معناشناسی است که معنای یک واژه را در دایره دانش دایرة المعرفی آن واژه و روابط معنایی با واژه‌های دیگر تعریف می‌کند. برای نمونه، طبق این نظریه، معنای واژه‌ای نظیر «فروختن» بدون در نظرگرفتن داشت دایرة المعرفی قاب<sup>۱۲</sup> داده‌شده و ارتباط آن با دیگر واژه‌ها نظیر «خریدن»، «خریدار»، «کالای مورد معامله» و غیره قابل درک نخواهد بود. در پژوهشی اخیر (Hedegaard & Simonsen, 2011) که برای شناسایی نویسنده یک متن ترجمه‌شده انجام شده، پژوهشگران با استفاده از روابط معنایی تعریف شده بر اساس شبکه قاب‌ها دقت شناسایی نویسنده اصلی متن را افزایش دادند. فرض بنیادین آن‌ها در استفاده از معناشناسی قاب‌بینان این بوده که ترجمة یک متن بسته به سبک مترجم می‌تواند نشانگرهای نحوی و واژگانی نویسنده اصلی را دکرگون کند؛ اما قاب‌های معنایی به کاررفته در متن اصلی در متن ترجمه‌شده دکرگون نمی‌شوند. تحلیل متن بر اساس نشانگرهای معنایی نیاز به تحلیل عمیق متون و استفاده از ابزارهای رایانشی خودکار دارد. با توجه به اینکه ابزارهای خودکار پردازش معنایی زبان در ابتدای مسیر هستند، چالش‌های استفاده از نشانگرهای معنایی بیشتر از نشانگرهای صوری واژگانی و نحوی است.

در برخی دیگر از پژوهش‌های اخیر محققان سعی کرده‌اند تمایز سبکی نویسنده‌ها/ گویشوران را با استفاده از مفهوم نظری و زبان‌شناختی گوییش فردی<sup>۱۳</sup> تبیین کنند که به شیوه منحصر به فرد افراد در به کارگیری عناصر زبانی نظیر آواه، واژدها، ساختهای دستوری و غیره گفته می‌شود (Wardhaugh & Fuller, 2015: 9). اهمیت این پژوهش‌ها آن‌جا مشخص می‌شود که برخی از زبان‌شناسان در گذشته همواره نسبت به واقعیت گوییش فردی تشکیک کرده‌اند از مطالعه اخیری که بر روی گفتار پنج سخنگوی خبری کاخ سفید انجام شده (Barlow, 2010) شواهد محکمی از رد پای گوییش فردی در

ساختهای دستوری به دست می‌دهد. فرضیه مورد آزمون این پژوهش، که بعداً تأیید می‌شود، این است که تنوع در گویش فردی گویشوران در قالب مؤلفه‌های اصلی واژی دستوری نمود می‌یابد و به استعمال برخی واژه‌ها و عبارات خاص و منحصر به فرد محدود نمی‌شود.

در پژوهشی دیگر (Johnson & Wright, 2014) که بر روی شناسایی نویسنده رایانامه‌های شرکت انرژی انرون انجام گرفته است، محققان با استفاده از توالی‌های واژگانی آنگاشتی، سعی می‌کنند مصادیقی از گویش فردی را به دست دهند. در این پژوهش پیکرهای ۲/۵ میلیون کلمه‌ای متشكل از ۶۳۰۰۰ ایمیل که توسط ۱۷۶ کارمند شرکت امریکایی انرون نوشته شده است، آماده و به لحاظ سبکی مورد تحلیل واقع می‌شود. نتایج تحلیل نشان می‌دهد که هر یک از کارمندان شرکت الگوهای سبکی متمایز از الگوی دیگران را در نگارش خود به کار می‌گیرند. در ادامه تحلیل‌ها پژوهشگران همچنین با استفاده از یک آزمایش آماری نشان می‌دهند که با استفاده از آنگاشتها می‌توان نویسنده‌گان رایانامه‌های مجھول الهویت را با دقت ۱۰۰ درصد شناسایی کرد.

در این پژوهش قصد بر این است که با رویکردی مبتنی بر گویش فردی، تحلیلی پیکرهای از سبک نویسنده‌گان مختلف در زبان فارسی به دست داده شود و قابلیت تفکیک سبک نویسنده‌ها بر اساس واژه‌های دستوری برسی شود. بر اساس شواهد موجود، واژه‌های دستوری به دلیل اینکه مجموعه‌ای بسته از عناصر زبانی را تشکیل می‌دهند، مستقل از محتوا و موضوع متن به کار گرفته می‌شوند، و سرنخ‌هایی از ساختهای کلان نحوی به دست می‌دهند، نشانگرهای مؤثری برای شناسایی هویت نویسنده محسوب می‌شوند. در فصل بعد، توضیحات مربوط به روش‌شناسی پژوهش، پیکره و الگوریتم‌های مورد استفاده ارائه شده است.

### ۳. مبانی نظری و روش پژوهش

شناسایی نویسنده را می‌توان فرایندی تعریف کرد که طی آن با اندازه‌گیری برخی مؤلفه‌های متنی می‌توان متنی را که توسط نویسنده‌های مختلف نوشته شده، از یکیگر متمایز کرد (Stamatatos, 2009). پیش‌فرض مهمی که در تحقیقات مربوط به شناسایی نویسنده وجود دارد این است که نویسنده یک متن فاقد نویسنده، از بین چند نویسنده مظنون انتخاب می‌شود و بنابراین بدیهی است که فرایند «شناسایی» به جستجوی فرد از میان نویسنده‌گان ناشناخته و بی‌شمار دلالت ندارد. برای مقایسه متنی که هویت نویسنده آن نامعلوم است با متنی که نویسنده‌گان

مشخصی دارند، نیاز به شاخصهایی است که ویژگی‌های سبکی متون را بتوان بر اساس آن شاخص‌ها استخراج و با یکی‌گر مقایسه کرد. این پژوهش براساس نشانگرهای واژگانی انجام شده است.

این پژوهش، همسو با تحقیقات مروارشده، مبتنی بر این فرض زبان‌شناختی است که هر فردی دارای گویش فردی منحصر به خود است که می‌تواند در نشانگرهای مختلف متى ازجمله واژه‌های دستوری تظاهر پیدا کند. در تعریف دقیق مفهوم گویش فردی می‌توان گفت «گویش فردی شیوه منحصر به فرد [هر گویشور] در صحبت کردن، شامل به کارگیری صدای، کلمات، سنتور زبان و سبک زبانی است» (Wardhaugh & Fuller, 2015: 9). اگر بخواهیم از استعاره اثر انگشت استفاده کنیم، می‌توان این فرضیه را مطرح کرد که گویش فردی هر کس از آنجا که منحصر به خود است، حکم اثر انگشت زبانی‌اش را خواهد داشت. پیامد این فرضیه در حوزه شناسایی نویسنده یک متن این خواهد بود که اگر بتوان مشخصه‌های سبکی و مبتنی بر گویش فردی متون نگاشته شده توسط یک نویسنده را استخراج کرد، می‌توان با استخراج مشخصه‌های سبکی یک متن فاقد هویت و مقایسه آن با مؤلفه‌های سبکی استخراج شده از متون دارای هویت، مشخص کرد که سبک متن فاقد هویت به سبک کدام نویسنده نزدیکتر است.

### ۱-۲. نشانگرهای سبکی

نشانگر سبکی مورد استفاده در این پژوهش، نشانگرهای واژگانی و از نوع واژه‌های دستوری است. واژه‌ها طبق تقسیم‌بندی‌های زبان‌شناختی به دو دسته عمده واژه‌های محتوایی و واژه‌های دستوری تقسیم می‌شوند. واژه‌های محتوایی تمام واژه‌هایی را در بر می‌گیرند که محتوای معنایی پررنگ‌تری دارند، تعدادشان نامحدود است و حجم عمدۀ فرهنگ‌های لغت را به خود اختصاص می‌دهند. واژه‌های دستوری (شامل حروف اضافه، حروف ربط، غیره)، در مقابل، محتوای معنایی پررنگی ندارند، تعدادشان در هر زبان محدود است و نقشی دستوری در جملات ایفا می‌کنند. در این پژوهش، به دلایل زیر نشانگرهای واژگانی از مقوله واژه‌های دستوری انتخاب شده‌اند (Kestemont, 2014):

- واژه‌های دستوری بسامد بالای نسبت به واژه‌های محتوایی در متون دارند. این ویژگی امتیاز بزرگی در حوزه تشخیص سبک نویسنده محسوب می‌شود؛ زیرا در دنیای واقعی و در

موقعیت‌های کاربردی، حجم متون ایده‌آل نیست و گاهی محقق ناچار است با داده‌های اندکی که در اختیار دارد، سبک نویسنده متن را مشخص یا نویسنده را شناسایی کند.

• واژه‌های دستوری اساساً مستقل از عنوان و موضوع متن به کار برده می‌شوند. از آنجا

که واژه‌های دستوری نقش دستوری در جمله ایفا می‌کنند، مورد استعمال آن‌ها متکی به موضوع متن نیست. به همین دلیل مزیت مهم این واژه این است که متون مورد تحلیل در شرایط آزمایشگاهی نیازی به داشتن موضوع یکسان ندارند که بتوان آن‌ها را با یکدیگر مقایسه کرد.

• واژه‌های دستوری به طور ناخودآگاه به کار گرفته می‌شوند و نویسنده خلاف واژه‌های محتوایی کنترل خودآگاه بر استفاده یا استفاده نکردن از آن‌ها ندارد.

تعداد نشانگرهای دستوری مورد استفاده در این پژوهش که بر اساس آن‌ها سبک نویسنده‌ها با یکدیگر مقایسه و تمایز می‌شدند، ۲۰ واژه دستوری پرسامد تعیین شد.

## ۲-۳. پیکره پژوهش

پیکره مورد استفاده در این پژوهش، از مجموعه مقالات نوشتاری پنج نویسنده که به صورت الکترونیکی در دسترس بوده‌اند، گردآوری شده است. این متون از مقالات قابل دسترس ناصر فکوهی و نیوشان صدر در وبگاه اینترنتی «انسان‌شناسی و فرهنگ»<sup>۱۴</sup>، و بگاه شخصی صادق زیباکلام<sup>۱۵</sup>، و بگاه شخصی رضا داوری اردکانی<sup>۱۶</sup>، و بگاه شخصی حمید پارسانیا<sup>۱۷</sup> جمع‌آوری شده‌اند. در جمع‌آوری پیکره، ملاحظات زیر مدنظر قرار گرفت:

- سعی بر آن بود موضوعاتی که نویسنده‌گان به آن‌ها پرداخته‌اند تا حد امکان مشابه انتخاب شوند یا حداقل در یک حوزه (در اینجا علوم انسانی) باشند. هرچند استفاده از نشانگرهای واژه‌های دستوری در این پژوهش که مستقل از موضوع به کار گرفته می‌شوند، ضرورت تشابه متون را منتفی می‌سازد؛ اما این نکته برای بالا بردن روایی پژوهش مدنظر بوده است.

- در جمع‌آوری متون همواره این موضوع مورد تأکید بود که اطیبان حاصل شود متن به قلم خود نویسنده نوشته شده است. از آنجا که این پژوهش، به صورت آزمایشگاهی انجام می‌شود و هدف از آن آزمودن کارایی واژه‌های دستوری در تمایز کردن سبک نویسنده‌های فارسی است، اطیبان از اصالت نویسنده‌ها بسیار حائز اهمیت است. این دغدغه موجب شد به جای استفاده از مقالات یا نویسنده‌جات علمی معمول که حاوی ارجاع به منابع دیگر هستند، از مقالات و

خودنگاشتهای موجود در وبلاگ‌های شخصی برخی محققان و اندیشمندان که پیش‌تر اشاره شد استفاده شود. مقالات پس از گردآوری به‌طور مجزا برای هر نویسنده در فایلی متنی ذخیره شدند. قبل از اینکه مطالب گردآوری شده در آزمایش‌های سبک‌سنگی استفاده شوند، لازم بود که برخی عملیات پیش‌پردازشی بر روی آن‌ها انجام شود. در ابتدا هر یک از فایل‌ها به‌صورت دستی مرور شدند تا برخی مطالب غیرمرتبط با پژوهش پالایش شوند. مثلاً نام نویسندهان مقالات از ابتدای آن‌ها حذف شد. همچنین در مواردی مشاهده شد که متن مقاله به روایت شخص دیگری غیر از خود نویسنده است که این مقالات کنار گذاشته شدند. همچنین سعی شد مقالاتی که به‌طور فراوان از منابع دیگر نقل قول مستقیم کرده بودند، از پیکرهٔ پژوهش کنار گذاشته شوند.

پس از پیش‌پردازش دستی اولیه، کل پیکره به‌طور خودکار توسط نرم‌افزار خطایاب وفا (Faili, Ehsan, Montazery, & Pilehvar, 2016) استاندارسازی شد. این خطایاب به‌صورت افزونه‌ای در نرم‌افزار مایکروسافت وورد نصب می‌شود و متن را به‌لحاظ یکست‌سازی حروف (جاگزینی حروف عربی با فارسی)، رعایت نیمفاصله‌ها، نحوی و معنایی خطایابی می‌کند. مهم‌ترین انگیزه از به‌کارگیری نرم‌افزار خطایاب در این پژوهش، یکست‌سازی نگارش و کشف خطاهای املایی در متن بوده است.

پس از مرحلهٔ پیش‌پردازش، پیکرهٔ آماده استفاده شد. تعداد کلمات پیکرهٔ اصلی پژوهش به‌تفکیک نویسندهان آن به‌صورت جدول ۱ است.

جدول ۱: تعداد کلمات موجود در پیکره به‌تفکیک نویسندهان

Table 1. Number of words per author in the corpus

نویسنده	تعداد کلمات پیکره
صادق زیباکلام	۱۰۸۶۹۰
حیدر پارسانیا	۱۰۷۴۹۹
ناصر فکوهی	۱۰۷۱۰۷
رضا داوری اردکانی	۴۴۸۵۴
نیوشاصدر	۲۴۲۴۶

### ۳-۲. روش تحلیل

برای تحلیل سبک نویسنده‌گان از بسته تخصصی استایلو (Eder et al., 2013) که در زبان برنامه‌نویسی آماری R (Core Team, 2015) ایجاد شده است، استفاده شد. این بسته قابلیت دریافت فایل‌های پیکره، نمونه‌گیری تصادفی از آن‌ها، تهیه لیست واژه‌های پربسامد<sup>۱۰</sup>-نگاشتی، و انجام تحلیل‌های مورد نظر بر اساس برخی از الگوریتم‌های ازپیش‌موجود را دارد. در این پژوهش از دو روش بینظارت تحلیل مؤلفه‌های اصلی<sup>۱۱</sup> و تحلیل خوش‌های<sup>۱۲</sup> استفاده شده است. تحلیل مؤلفه‌های اصلی برای بررسی نقش هر یک از واژه‌های دستوری در تمایز سبک نویسنده‌های مختلف و تحلیل خوش‌های برای خوشبندی آثار نویسنده‌گان مورد استفاده قرار گرفت. مقیاس فاصله‌ای بهکار رفته برای محاسبه فاصله آثار نویسنده‌ها با یکدیگر بر اساس مقیاس دلتا (Burrows, 2002) معین شد.

### ۴-۲. روال انجام پژوهش

این پژوهش در قالب ۶ آزمایش انجام شده و هر آزمایش با هدف بررسی کارایی واژه‌های دستوری تکنگاشتی، دونگاشتی و سه‌نگاشتی انجام شده است. همچنین در هر آزمایش مقدار متغیری از تعداد واژه‌ها به عنوان پیکره از کل پیکره پژوهش نمونه‌گیری شده‌اند. جدول ۲ میزان حجم پیکره برای هر آزمایش را به همراه نویسنده‌گان نمونه متن‌ها نشان می‌دهد.

جدول ۲: حجم تقریبی پیکره مورد استفاده در هر آزمایش

Table 2. Size of the corpus used in each experiment

آزمایش	حجم پیکره برای هر نمونه متن	نویسنده متن
۱	۵....	فکوهی، پارسانیا، زیبایکلام
۲	۱....	فکوهی، پارسانیا، زیبایکلام، صدر، اردکانی
۳	۱....	فکوهی، پارسانیا، زیبایکلام، صدر، اردکانی
۴	۲...	فکوهی، پارسانیا، زیبایکلام، صدر، اردکانی
۵	۳...	فکوهی، پارسانیا، زیبایکلام، صدر، اردکانی
۶	۴...	فکوهی، پارسانیا، زیبایکلام، صدر، اردکانی

روال انجام آزمایش‌ها به این صورت بوده که ابتدا در آزمایش اول متون ۵۰۰۰ واژه‌ای از نویسندهان تحلیل شدند. برای این منظور، متون سه نویسنده دارای تعداد واژه کافی یعنی ۱۰۰۰۰ واژه بودند که برای هر نویسنده دو پیکره ۵۰۰۰ واژه‌ای به طور تصادفی نمونه‌گیری شدند. در آزمایش دوم، حجم نمونه‌منتها به ۱۰۰۰ کاهش داده شد و متون همه پنج نویسنده در تحلیل قرار گرفت. در آزمایش سوم، حجم نمونه‌منتها به هزار کاهش یافت. هدف از اعمال این کاهش‌ها ارزیابی عملکرد واژه‌های دستوری در تمایز سبک نویسنده‌ها در حجم داده‌های مختلف است. در سطح ۱۰۰۰ واژه چون عملکرد تشخیص نویسنده بچار افت دقت شد، در هر یک از آزمایش‌های بعدی هزار واژه به حجم متون اضافه شد تا دقت شناسایی نویسنده‌ها حداقل در سطح واژه‌های دستوری تکنگاشتی به صد درصد برسد.

پس از مشخص شدن پیکره‌های تحلیل، اولین گام در تحلیل سبک‌های نویسنده‌ها، استخراج واژه‌های دستوری پربسامد بود که نقش نشانگر واژگانی را در تمایز کردن سبک نویسنده‌گان ایفا می‌کردند. واژه‌های دستوری پربسامد با استفاده از فرامین موجود در بسته استایلو استخراج و سپس به صورت دستی بررسی شدند تا واژه‌های غیردستوری از آن پالایش شوند. واژه‌های غیردستوری که از فهرست واژه‌های پربسامد حذف شدند عمدها شامل مواردی می‌شدند که در جدول ۳ آمده است.

جدول ۳. فهرست واژه‌های مذوف از جدول واژه‌های پربسامد

Table 3. List of words removed from the high-frequency words list

مثال	واژه‌های مذوف
	اعداد
یک، دو، سه، ...	علامت جمع بدون نیمه‌فاصله
ها، های	وندهای فاق نیمه‌فاصله
می، ای	افعال (دیطی و غیردیطی)
است، بود، شد، ...	ضمایر شخصی
او، آن‌ها، ...	واژه‌های محترابی
فرهنگ، ایران، علم، ...	

طی فرایند شمارش خودکار واژه‌های پربسامد و قبل از حذف واژه‌های غیردستوری، گزینش درصدی بر روی واژه‌های پربسامد اعمال شد. طبق این گزینش، واژه‌هایی که تا درصد خاصی

در همه متون به کار نرفته بودند، کنار گذاشته می‌شدند. حد کمینه این گزینش ۲۰ درصد و حد بیشینه آن ۸۰ درصد در نظر گرفته شد. گزینش واژه‌های پربسامد با حد ۲۰ درصد بین معناست که واژه‌هایی که در حافظ ۲۰ درصد از متن‌ها به کار رفته باشند، در تحلیل لحاظ خواهند شد.

پس از محاسبه تعداد واژه‌های پربسامد و حذف واژه‌های غیردستوری، جدول حاصله آماده تحلیل سبک‌سنجی است. در مرحله بعد، نحوه تمایز نویسندهان و سهم هر کدام واژه‌ها در تمایز سبک‌ها به واسطه تحلیل مؤلفه‌های اصلی بررسی شدند. سپس با استفاده از تحلیل خوش‌بندی و بر اساس مقیاس فاصله‌ای دلتا، خوش‌بندی متون مشابه به لحاظ سبکی به دست داده شد.

برای ارزیابی دقت خوش‌بندی، از شاخصی به نام شاخص تعديل شده رند<sup>۲۰</sup> (Hubert & Arabie, 1985) استفاده شد. این شاخص، نسخه بهبودیافتدۀ شاخص رند (Rand, 1971) است که خوش‌بندی تولید شده توسط الگوریتم را با خوش‌بندی ایده‌آل (که آثار هر نویسنده باید در خوش‌های مجزا قرار گیرند) مقایسه می‌کند و میزان مطابقت آن را به دست می‌دهد. اگر U و V را دو طبقه‌بندی متفاوت از اشیای مشخص در نظر بگیریم، شاخص رند را می‌توان طبق فرمول زیر برای مقایسه این دو طبقه‌بندی محاسبه کرد:

$$RI = \frac{a+d}{a+b+c+d}$$

a. جفت‌هایی که هم در طبقه‌بندی U و هم در طبقه‌بندی V همگروه هستند؛

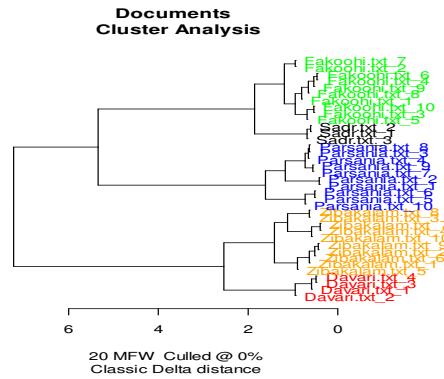
b. جفت‌هایی که در طبقه‌بندی U همگروه ولی در طبقه‌بندی V غیرهمگروه هستند؛

c. جفت‌هایی که در طبقه‌بندی V همگروه ولی در طبقه‌بندی U غیرهمگروه هستند؛

d. جفت‌هایی که هم در طبقه‌بندی U و هم در طبقه‌بندی V غیرهمگروه هستند.

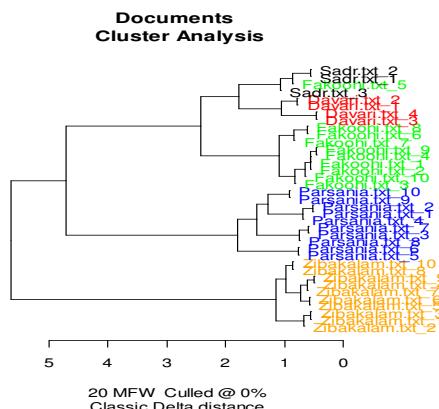
#### ۴. نتایج تحلیل داده‌ها

در هر یک از شش آزمایش که بر روی متونی با تعداد کلمات متفاوت صورت گرفته‌اند، عملکرد واژه‌های دستوری تکنگاشتی، دونگاشتی و سنهنگاشتی با استفاده از مقیاس فاصله‌ای دلتا و تحلیل خوش‌بندی به دست آمد. نمودارهای ۱ و ۲ و ۳ تحلیل خوش‌بندی واژه‌های دستوری تکنگاشتی، دونگاشتی و سنهنگاشتی را در سطح ۱۰۰۰۰ واژه نشان می‌دهد.



نمودار ۱: تحلیل خوشه‌ای نوشته‌های نویسندهای بر اساس واژه‌های دستوری تکنگاشتی در متنون ۱۰۰۰۰ واژه‌ای

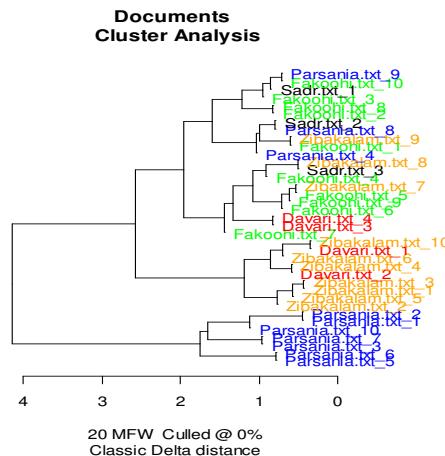
**Figure 1.** Cluster analysis of authors' writings based on monogram function words in 10000 word texts.



نمودار ۲: تحلیل خوشه‌ای نوشته‌های نویسندهای بر اساس واژه‌های دستوری دونگاشتی در متنون ۱۰۰۰۰ واژه‌ای

**Figure 2.** Cluster analysis of authors' writings based on bigram function words in 10000 word texts.

با توجه به تحلیل خوش‌های نمودارهای ۱ تا ۳ می‌توان دریافت که واژه‌های دستوری یک، دو و سه‌نگاشتی عملکرد متفاوتی در تفکیک نوشت‌های نویسنده‌ان داشته‌اند. در واقع واژه‌های دستوری تکنگاشتی (نمودار ۱) بهترین عملکرد را در مقایسه با واژه‌های دستوری دونگاشتی و سه‌نگاشتی داشته‌اند. با توجه به نمودار ۲، واژه‌های دستوری دونگاشتی در تفکیک متون مربوط به «داوری» «صدر» و «فکوهی» اندکی دچار خطا شده است. میزان خطا در واژه‌های دستوری سه‌نگاشتی (نمودار ۳) بسیار بیشتر شده است. به نظر می‌رسد با افزایش توالی واژه‌ها، بسامد توالی‌ها در متون کاهش می‌یابد و این موضوع موجب می‌شود تفکیک متون با افت دقت روبرو شود.

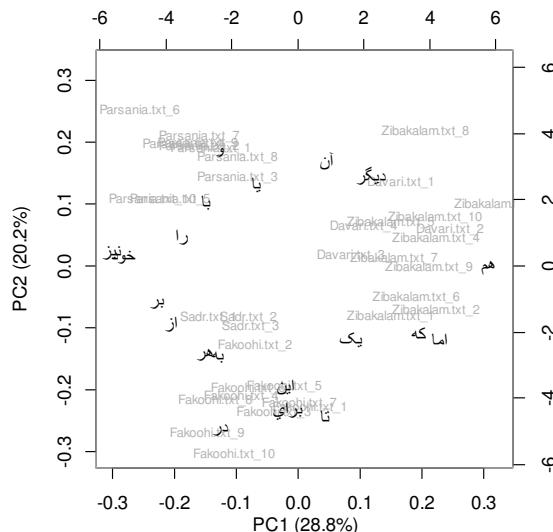


نمودار ۳: تحلیل خوش‌های نوشت‌های نویسنده‌ان بر اساس واژه‌های دستوری سه‌نگاشتی در متون ۱۰۰۰۰ واژه‌ای

**Figure 3.** Cluster analysis of authors' writings based on trigram function words in 10000 word texts.

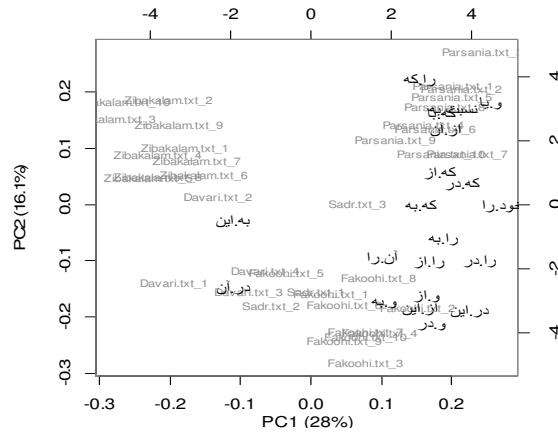
همچنین علاوه بر تحلیل خوش‌های، برای شناسایی واژه‌ها یا توالی واژه‌ایی که بیشترین تأثیر را در تفکیک متون نویسنده‌ها داشته‌اند، داده‌ها با استفاده از تحلیل مؤلفه‌های اصلی بررسی شدند. نمودار ۴ تا ۶ نتایج تحلیل مؤلفه‌های اصلی را در قالب نمودار بار عاملی<sup>۲۱</sup> برای واژه‌های دستوری

تکنگاشتی (نمودار ۴)، دونگاشتی (نمودار ۵) و سه‌نگاشتی (نمودار ۶) در سطح ۱۰۰۰۰ واژه نشان می‌دهد. در این نمودارها، متون نویسنده‌ها با رنگ خاکستری و واژه‌های دستوری (بار عاملی) با رنگ مشکی تمایز شده‌اند. نزدیکی برخی (توالی) واژه‌های دستوری با متون یک نویسنده، نشانگر غالب بودن آن (توالی) واژه دستوری در متون آن نویسنده است. مثلاً در نمودار ۴، واژه‌های دستوری «و»، «یا» و «با» در متون «پارسانیا» دارای بسامد بیشتری نسبت به متون نویسنده‌های دیگر بوده‌اند. نمودار ۵ و نمودار ۶ نیز به همین ترتیب، توزیع توالی‌های دوتایی و سه‌تایی واژه‌های دستوری را در متون نویسنده‌ها نشان می‌دهد. همچنین شاخص‌های کمی مؤلفه‌های اصلی اول و دوم در نمودار ۴ نشان می‌دهد که در مجموع ۶۴ درصد از گونه‌گونی داده‌ها در قالب این دو مؤلفه اصلی قابل توضیح است؛ مؤلفه اول ۲۸/۸ درصد از گونه‌گونی را در محور افقی و مؤلفه دوم ۲۰/۲ درصد از گونه‌گونی را در محور عمودی توضیح می‌دهد. در همین راستا، واژه‌های دستوری «هم»، «خود» و «نیز» عامل بیشترین تفکیک متون در محور افقی و واژه‌های «آن»، «و»، «در»، «تا» و «برای» عامل بیشترین تمایز متون در محور عمودی بوده‌اند.



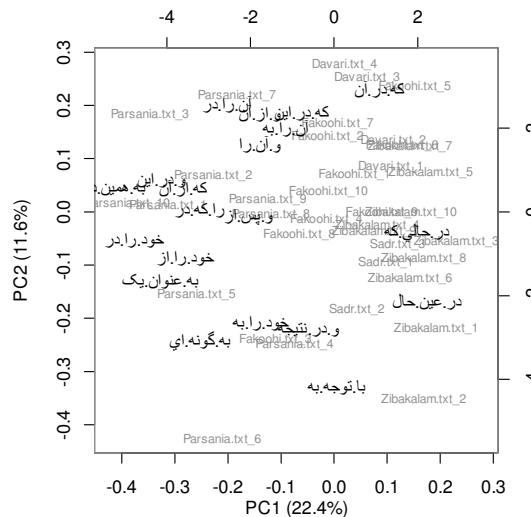
نمودار ۴: تحلیل مؤلفه‌های اصلی و بار عاملی برای واژه‌های دستوری تکنگاشتی در متون ۱۰۰۰۰ واژه‌ای

Figure 4. PCA and loadings of monogram function words in 10000 word texts.



**نمودار ۵:** تحلیل مؤلفه‌های اصلی و بار عاملی برای واژه‌های دستوری دونگاشتی در متن ۱۰۰۰۰ واژه‌ای

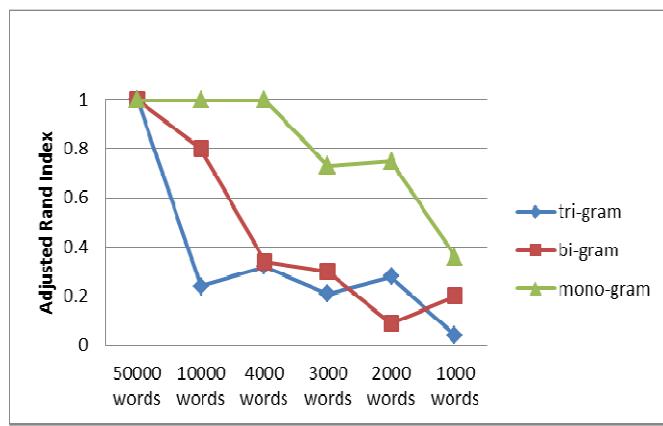
**Figure 5.** PCA and loadings of bigram function words in 10000 word texts.



**نمودار ۶: تحلیل مؤلفه‌های اصلی و بار عاملی برای واژه‌های دستوری سه‌نگاشتی در متون ۱۰۰۰۰ واژه‌ای**

**Figure 6.** PCA and loadings of trigram function words in 10000 word texts.

با کنار هم قراردادن آزمایش‌هایی که هر کدام بر روی مقدار متفاوتی از داده‌های متغیر انجام شد تصویری روشن‌تر از کارایی مقیاس دلتا در طبقه‌بندی متون نویسندهان در این پژوهش به دست می‌آید. نمودار ۷ تغییرات شاخص تعديل شده رکد را برای واژه‌های دنگاشتی در حجم‌های متغیر آزمایش شده در این پژوهش نشان می‌دهد.



نمودار ۷: تغییرات شاخص تعديل شده رکد بر اساس تعداد کلمات متون به تفکیک واژه‌های دستوری تکنگاشتی، دونگاشتی و سه‌نگاشتی

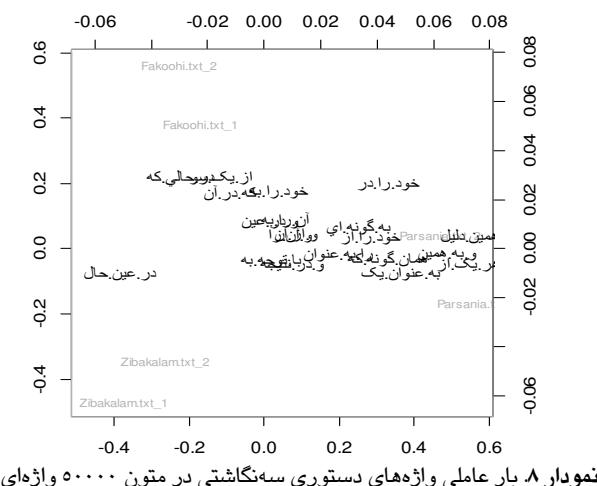
**Figure 7.** The trend for Adjusted Rand Index of monogram, bigram, and trigram function words per text size

با توجه به نمودار ۷، مشاهده می‌شود که به‌طور کلی با کاهش تعداد کلمات متون، دقت دسته‌بندی متون متعلق به یک نویسنده در گروه مربوطه که ما آن را دقت شناسایی نویسنده می‌نامیم کاهش می‌یابد. اما به‌نظر می‌رسد واژه‌های تکنگاشتی، دونگاشتی و سه‌نگاشتی در این میان اندکی متفاوت عمل می‌کنند. نخست اینکه عملکرد واژه‌های تکنگاشتی (یا همان واژه‌های متعارف متغیر) بهترین عملکرد را نسبت به واژه‌های دونگاشتی و سه‌نگاشتی داشته است. در سطح ۵۰۰۰ واژه‌ای، ۱۰۰۰۰ واژه‌ای و ۴۰۰۰ واژه‌ای، دقت واژه‌های تکنگاشتی در شناسایی متون ۱۰۰ ادرصد است و تغییر نمی‌کند. نتایج نشان می‌دهد که متون ۴۰۰۰ واژه‌ای احتمالاً حد کمینه متومنی است که تماماً به درستی شناسایی می‌شوند. البته چنانچه آستانه دقت را تا حد ۷٪ پایین بیاوریم، آنگاه می‌توان گفت که تا حد

۲۰۰۰ واژه نیز عملکرد الگوریتم با استفاده از واژه‌های تکنگاشتی پذیرفتی است.

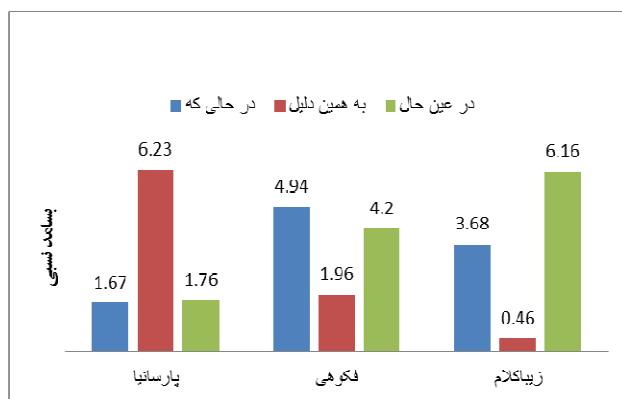
نکته دوم اینکه، در میان واژه‌های دونگاشتی و سه‌نگاشتی، به نظر می‌رسد به‌طور کلی عملکرد واژه‌های دونگاشتی در همه متون (به استثنای متون ۲۰۰۰ و واژه‌ای) اندکی بهتر از واژه‌های سه‌نگاشتی بوده است. نکته دیگری که در عملکرد واژه‌های دستوری سه‌نگاشتی بیشتر مشهود است این است که با کاهش تعداد متون از ۵۰۰۰۰ به ۱۰۰۰ واژه، شاهد فوسان در دقت شناسایی نویسنده هستیم، مثلاً در متون ۴۰۰۰ و واژه‌ای، با آنکه تعداد کلمات نسبت به متون ۱۰۰۰۰ واژه‌ای بیش از نصف کاهش یافته است؛ اما دقت شناسایی اندکی بیشتر بوده است که البته به نظر می‌رسد معنی‌دار نیست.

یکی از دلایلی که می‌توان بر رفتار متفاوت واژه‌های آنگاشتی و افت عملکرد شناسایی نویسنده با حرکت از واژه‌های تکنگاشتی تا سه‌نگاشتی بیان کرد، کاهش تعداد کلمات متون و بدنبال آن کاهش رخداد توالی‌های واژگانی در پیکره است. درواقع با کاهش تعداد کلمات در پیکره متون، احتمال اینکه توالی‌های چندگانه در متن تکرار شوند، کاهش بیشتری می‌یابد. اگرچه توالی‌های چندگانه واژگانی، نشانگرهای دقیق‌تری از روال‌های خودکارشده واژگانی می‌توانند باشند؛ اما کاهش تعداد کلمات متون شناسایی آن‌ها را با مشکل مواجه می‌سازد. از این‌رو، عملکرد مطلوب توالی‌های واژگانی در متون حجمی می‌تواند بروز کند.



**Figure 8.** Loadings of trigram function words in 50000 word texts.

به منظور کسب اطلاعات بیشتر از نحوه بهکارگیری واژه‌ها در جملات و بافت دستوری، بررسی توالی‌های دونگاشتی و سه‌نگاشتی پربرسامد در مرحله بعد می‌تواند اطلاعات بیشتری در این راستا به دست دهد. برای مثال، نمودار ۸ که نمودار بار عاملی واژه‌های سه‌نگاشتی در متون ۵۰۰۰ واژه‌ای است، اطلاعات جالب توجهی را در اختیارمان قرار می‌دهد. بر اساس این نمودار می‌توان دریافت که در سطح مولفه اصلی اول (محور افقی) سه‌نگاشتی‌هایی نظری «در عین حال»، «به همین دلیل»، «و به همین»، «هر یک از»، «همان گونه که»، «در حالی که»، «خود را در» و «به عنوان یک» بیشترین تأثیر را در پراکندگی متون در طول این محور داشته‌اند. از میان این سه‌نگاشتی‌ها، برخی نظری «در عین حال»، «در حالی که» و «به همین دلیل» را می‌توان جزو حروف ربطی مرکب دانست که با وجود صورت چندگاهی‌اشان، ماهیت باهم‌آیی<sup>۲۲</sup> دارند و معنای واحدی را می‌رسانند. نمودار ۹ بسامدهای نسبی این سه حرف ربطی مرکب را در متون سه نویسنده نشان می‌دهد که از پیکرهای مربوطه‌شان استخراج شده‌اند. حرف ربط «در عین حال» بیشترین بسامد را در متون زیبایکلام و سپس در متون فکوهی دارد. حرف ربط «به همین دلیل» ممیزه سبکی متون پارسانیاست. همچنین حرف ربط «در حالی که» بیشتر در متون فکوهی، سپس در متون زیبایکلام و کمتر در متون پارسانیا به چشم می‌خورد.



نمودار ۹: بسامد نسبی برخی سه‌نگاشتی‌های پربرسامد ممیز سبکی در پیکره متنی سه نویسنده

**Figure 9.** Relative frequency of some frequent trigrams used as style markers in the corpus of three authors

بر اساس اطلاعات نمودار ۹ و همچنین مشاهدات پیکره‌ای، ساختار گفتمانی علت و معلولی که به‌وسیله حرف ربط «به همین دلیل» بیان شده، در متون پارسایی غلبه داشته است. در حالی که سبک گفتمانی در متون زیباقلام و فکوهی بیشتر در راستای تقابل آراء و تضارب ایده‌های ناسازگار پیش رفته و در قالب حرف ربط «در حالی که» بیان شده است. این معنای گفتمانی تا حدودی در حرف ربط «در عین حال» نیز دیده می‌شود که مشخصه بارز متون زیباقلام است. در متون زیباقلام این حرف ربط گاهی در معنای «همزمان» یا «در همین حال» ظاهر شده و گاهی ایده‌های ناسازگار با مطالب پیش‌گفته را نشانه‌گذاری کرده است.

## ۵. نتیجه

نتایج این پژوهش نشان می‌دهد که نویسنده‌های مختلف واژه‌های دستوری را به‌طور مشابه به‌کار نمی‌گیرند. درواقع با وجود اینکه برخی واژه‌های دستوری پربسامد توسط همه نویسنده‌ها به‌کار گرفته می‌شوند، اما اولویت به‌کارگیری آن‌ها توسط نویسنده‌ها متفاوت است. همانطور که در بخش تحلیل‌های پیکره‌ای نیز به آن پرداخته شد، برخی واژه‌های دستوری تقریباً به‌طور کمابیش یکسانی توسط همه نویسنده‌گان استفاده می‌شود؛ اما برخی دیگر با بسامد بیشتری در متون یک نویسنده نسبت به متون دیگر نویسنده‌گان به‌کار گرفته می‌شوند.

در توضیح علت اولویت‌دهی نویسنده‌های مختلف به واژه‌های دستوری متفاوت، می‌توان دو دلیل را مد نظر قرار داد: دلیل نخست در گویش فردی گویشوران ریشه دارد. گویش فردی هر گویشوری در بردارنده عادات زبانی وی است که در سطوح مختلف آوایی، واژگانی و نحوی زبان می‌تواند نمود پیدا کند. بنابراین، به‌کارگیری و اولویت‌دهی به برخی واژه‌های دستوری خاص می‌تواند مبتنی بر الگوهایی خودکارشده و عادت‌وار باشد که شخص در طول سالیان در مواجهه با درونداد زبانی محیط خود آن‌ها را درونی کرده است. نتایج بدست آمده در پژوهش کنونی در راستای این تبیین است.

دلیل محتمل دوم در اولویت‌دهی نویسنده‌های مختلف به واژه‌های دستوری متفاوت می‌تواند به‌دلیل نگارش متن در یک ژانر خاص باشد. درواقع می‌توان این احتمال را مدنظر داشت که ساختار گفتمانی یک ژانر به‌خصوص بر گزینش واژه‌های دستوری خاص تأثیر بگذارد. برای مثال، بسیاری از پژوهش‌های سبک‌سنجی نشان داده است که به‌کارگیری ضمایر شخصی هم-

بستگی معنی‌داری با برخی عوامل متغیر نظری را، چشم‌انداز روایی نویسنده، جنسیت نویسنده و حتی موضوع متن دارد (Kestemont, 2014). با توجه به اینکه در پژوهش کنونی ضمایر شخصی و هر نوع واژه‌ی مرتبط با موضوع متن از فهرست واژه‌های دستوری پرسامد حذف شدند، تفاوت موضوعی و ژانری نمی‌تواند تبیین مناسبی برای نتایج این پژوهش ارائه دهد. قضاوت درباره همبستگی احتمالی برخی واژه‌های دستوری (غیر از ضمایر شخصی) با ژانر و عوامل متغیر نویسنده بیشتر و دقیق‌تر را می‌طلبد. در مجموع برای پاسخ به پرسش نخست پژوهش، می‌توان این فرضیه را تأیید کرد که واژه‌های دستوری در نثر فارسی قابلیت تفکیک سبک نویسنده‌های مختلف را دارد.

در پاسخ به پرسش دوم پژوهش، با توجه به نتایجی که از خوشبندی واژه‌های دستوری تکنگاشتی، دونگاشتی و سه‌نگاشتی به دست آمد، واژه‌های تکنگاشتی یا همان واژه‌های دستوری تکواژه‌ای بهترین کارایی را در سطح محافظه‌کارانه ۴۰۰۰ واژه و با دقت ۱۰۰ درصد از خود نشان دادند (رجوع کنید به نمودار ۷). دقت واژه‌های دونگاشتی و سه‌نگاشتی با کاهش حجم متون، بلاعده با افت کارایی در تفکیک صحیح متون مواجه شد. با آنکه متون با کمینه حجم ۴۰۰۰ واژه به عنوان آستانه بهترین عملکرد الگوریتم دلتا در متون فارسی معین شد؛ اما به نظر می‌رسد این آستانه ماهیت حدی ندارد و اتفاق نظری در مورد کمینه متون برای تحلیل مطمئن سبک‌شناسی وجود ندارد. برای مثال، برخی پژوهشگران ۵۰۰۰ واژه را به عنوان حد کمینه قابل اعتماد برای متون داستانی در انگلیسی و دیگر زبان‌ها تعیین کرده‌اند (Eder, 2013)، در حالی که برخی دیگر قائل به کمینه ۱۰۰۰ واژه (Holmes et al., 2001) به عنوان حد قابل اعتماد برای پژوهش‌های سبک‌سنجی هستند. نتایج به دست آمده برای زبان فارسی نیز به نوعی همسو با یافته‌ها در دیگر زبان‌هاست. بنابراین در پاسخ به پرسش سوم پژوهش، براساس نتایج این پژوهش فعلًا می‌توان کمینه ۴۰۰۰ واژه را به عنوان حد مطلوب برای الگوریتم دلتا و با مدنظر قراردادن تعداد ۲۰ واژه دستوری برای تفکیک متون فارسی پذیرفت. پژوهش‌های آتی با تغییر تعداد واژه‌های دستوری و نیز حجم پیکره‌های متغیر دستوری، می‌توانند در ارزیابی اعتبار نتایج این پژوهش مؤثر باشند و موجب همگرایی در تخمین کمینه واژه‌های موردنیاز برای تفکیک سبک نویسنده‌گان شوند.

## ۶. سپاسگزاری

این مقاله برگرفته از پژوهشی است که با حمایت مالی پژوهشگاه علوم و فناوری اطلاعات ایران انجام شده است.

## ۷. پی‌نوشت‌ها

1. Lorenzo Valla
2. stylometry
3. forensic linguistics
4. plagiarism
5. monogram
6. bigram
7. trigram
8. style markers
9. lexical richness
10. parts of speech
11. Systemic Functional Grammar
12. frame
13. idiolect
14. www.anthropology.ir
15. www.zibakalam.com
16. www.rezadavari.ir
17. www.parsania.ir
18. Principal Component Analysis
19. Cluster Analysis
20. Adjusted Rand Index
21. loadings
22. collocation

## ۸. منابع

- Argamon, S., & Levitan, S. (2005). "Measuring the usefulness of function words for authorship attribution". In *Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*.
- Binongo, J. (2003). "Who wrote the 15<sup>th</sup> book of Oz? An application of multivariate

analysis to authorship attribution". *Chance*, 16. Pp. 9-17.

- Barlow, M. (2010). "Individual usage: a corpus-based study of idiolects". In *LAUD Symposium*. Landau, Germany.
- Barthes, R. (1977). *Elements of Semiology*. Hill and Wang: New York.
- Burrows, J. F. (1987). "Word patterns and story shapes: The statistical analysis of narrative style". *Literary and Linguistic Computing*. 2. Pp. 61–70.
- Burrows, J. F. (2002). "Delta: A measure of stylistic difference and a guide to likely authorship". *Literary and Linguistic Computing*. 17. Pp. 267–287.
- Carroll, D. (2008). *Psychology of Language* (5<sup>rd</sup> ed.). Wadsworth.
- Eder, M. (2013). "Does size matter? Authorship attribution, small samples, big problem". *Literary and Linguistic Computing*. DOI: <http://dx.doi.org/10.1093/litc/fqt066>.
- Eder, M.; M. Kestemont & J. Rybicki (2013). "Stylometry with R: a suite of tools". In *Digital Humanities 2013: Conference Abstracts*. University of Nebraska-Lincoln, NE. Pp. 487-89.
- Faili, H.; N.; Ehsan; M. Montazery & M. M. Pilehvar (2016). "Vafa spell-checker for detecting spelling, grammatical, and real-word errors of Persian language". *Digital Scholarship in Humanities*. 31 (1). Pp. 95-117.
- Farahmandpour, Z. & H. Nikmehr, (2015), "A study on intelligent authorship methods in Persian language". *Journal of Computing and Security*, 2(1).Pp. 63-76.
- Frontini, F.; G. Lynch & C. Vogel, (2008), "Revisiting the 'Donation of Constantine'". In *Proceedings of AISB 2008*. Pp. 1–9.
- Gamon, M. (2004). "Linguistic correlates of style: Authorship classification with deep linguistic analysis features". In *Proceedings of the 20th International Conference on Computational Linguistics*. Pp. 611-617.
- Hedegaard, S. & J. G. Simonsen (2011). "Lost in translation: Authorship attribution using frame semantics". In *Proceedings of the 49<sup>th</sup> Annual Meeting of the Association for Computational Linguistics*. June 19-24. Portland. Oregon. Pp. 65-70.

- Holmes, D. I.; L.J. Gordon & C. Wilson (2001). "A widow and her soldier: stylometry and the American civil war". *Literary and Linguistic Computing*. 16(4). Pp. 403-420.
- Hubert, L. and Arabie, P. (1985). "Comparing partitions". *Journal of Classification*. 2(1).Pp. 193-218.
- Jakobson, R. (1971). *Studies on Child Language and Aphasia*. The Hague: Mouton.
- Johansson, V. (2008). "Lexical diversity and lexical density in speech and writing: A developmental perspective". *Lund University, Department of Linguistics and Phonetics: Working Papers*. 53.Pp. 61–79.
- Johnson, A. & D. Wright, (2014), "Identifying idiolect in forensic authorship attribution". *Language and Law/Linguagem e Direito*. Vol. 1(1). Pp. 37-69.
- Kestemont, M. (2014). "Function words in authorship attribution: from black magic to theory?" In *Proceedings of the 3<sup>rd</sup> Workshop on Computational Linguistics for Literature*, April 27, Gothenburg, Sweden. Pp. 59-66.
- Koppel, M. & J. Schler (2003). "Exploiting stylistic idiosyncrasies for authorship attribution". In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*. Pp. 69-72.
- Modaber Dabagh, R. (2007). "Authorship attribution and statistical text analysis". *Metodološki zvezki*. 4(2). Pp. 149-163.
- Mosteller, F. & D.L. Wallace (1964). *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley.
- Rand, W. M. (1971). "Objective criteria for the evaluation of clustering methods". *Journal of the American Statistical Association*. 66. Pp. 846-850.
- R Core Team (2015). "R: A language and environment for statistical computing". R Foundation for Statistical Computing, Vienna, Austria. URL <https://R-project.org/>.
- Segarra, S.; M. Eisen & A.Ribeiro, (2015), "Authorship attribution through function word adjacency networks". In *IEEE Transactions on Signal Processing*. Vol. 63, No. 20. Oct. 15.Pp. 5464-5478.

- Stamatatos, E. (2009). "A survey of modern authorship attribution methods". *Journal of the American Society for Information Science and Technology*. 60(3). Pp. 538-556.
- Stein, B. & S. Meyer zu Eissen (2007). "Intrinsic plagiarism analysis with meta-learning". In B. Stein, M. Koppel, & E. Stamatatos (Eds.), *SIGIR workshop on plagiarism analysis, authorship identification, and near-duplicate detection (PAN 07)* (Pp. 45–50). CEUR-WS.org.
- Totty, R. N. & J. P Hardcastle (1987). "Forensic linguistics: the determination of authorship from habits of style". *Journal of the Forensic Science Society*. 27. Pp. 13-28.
- Wardhaugh, R. & J. M. Fuller (2015). *An Introduction to Sociolinguistics* (7<sup>th</sup> ed.). Wiley-Blackwell.
- Whitelaw, C. & S. Argamon (2004). "Systemic functional features in stylistic text classification". In *Proceedings of AAAI Fall Symposium on Style and Meaning in Language, Art, and Music*.
- Whitelaw, C. & J. Patrick (2004). "Selecting systemic features for text classification". In *Proceedings of Australian Language Technology Workshop*, Sydney, Australia. Pp. 93-100.
- Zhao, Y. & J. Zobel (2005). "Effective and scalable authorship attribution using function words". In *Information Retrieval Technology* (Pp. 174–189). Springer.