



دوماهنامه علمی- پژوهشی

د ۱۰، ش ۳ (پیاپی ۵۱)، مرداد و شهریور ۱۳۹۸، صص ۲۹۳-۳۱۷

واژه‌های دستوری به مثابه نشانه‌های گویش فردی: رویکردی پیکره‌ای به شناسایی هویت نویسنده در زبان فارسی

رامین گلشائی *

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 گسترده 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 به‌کار گرفته می‌شوند، مستقل از موضوع متن به‌کار می‌روند و بسامد بالایی در متون کوتاه دارند، می‌توانند نشانه‌های مؤثری برای کدگذاری گویش فردی و ممیز سبک نویسندگان باشند. در این پژوهش، ابتدا پیکره‌هایی متنی از نوشته‌های 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 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. نتایج نشان داد که واژه‌های دستوری در زبان فارسی قابلیت تفکیک متون متعلق به یک نویسنده را دارند و عملکرد واژه‌های تکنگاشتی بهتر از دونگاشتی و سه‌نگاشتی‌ها در متون کم‌حجم است. همچنین نتایج پژوهش نشان داد که حجم کمیته متن برای شناسایی موفقیت‌آمیز 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.

E-mail: golshaie@alzahra.ac.ir

* نویسنده مسئول مقاله:

۱. مقدمه

شناسایی اصالت نوشته‌ها و کشف هویت نویسنده متون از دیرباز مورد توجه انسان بوده است. یکی از قدیم‌ترین نمونه‌های آن، اثبات جعلی بودن سند «بخشش کنستانتین» توسط لورنزو والا^۱ در سال ۱۴۴۰ میلادی است که بر اساس این سند ادعا شده بود کنستانتین (امپراتور 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^۲ شناخته می‌شوند. در عصر کنونی با تحولی که رایانه در پردازش متون حجیم ایجاد کرده و کاربردهایی که پژوهش‌های سبک‌سنجی در زبان‌شناسی حقوقی^۳ (Totty & Hardcastle, 1987) و کشف سرقت علمی و ادبی^۴ (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 (غنا و واژه‌ها، بسامد واژه‌ها، غیره)، نحوی (ساخت جمله، بسامد قواعد نحوی، خطاها، اجزای کلام)، و معنایی (هم‌معناها، وابستگی‌های معنایی) باشند (Stamatatos, 2009). تحقیقات انجام‌گرفته در باب مؤلفه‌های واژگانی نشان می‌دهد که نوع به‌خصوصی از واژه‌ها، یعنی واژه‌های دستوری، ویژگی‌های منحصر به فردی برای شناسایی سبک نویسنده دارند. 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 پژوهش‌های 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^۵، دونگاشتی^۶ و سه‌نگاشتی^۷) 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، پیکره‌هایی 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 stylo در نرم‌افزار R، 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^۸ متعددی با عنوان نشانگرهای واژگانی، نحوی و معنایی به‌عنوان 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 به‌طور ناشناس منتشر شدند. در مطالعه‌ای که با استفاده از روش‌های آماری و به‌کارگیری بسامد واژه‌های دستوری به‌مثابه 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 پربسامد 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^۱ است. این شاخص که تنوع واژه‌های مورد استفاده 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 حوزه روان‌شناسی زبان نشان می‌دهد افراد مختلف غنای واژگانی متفاوتی دارند (Carroll, 2008). شاخص غنای واژگانی معمولاً در کنار 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 اندازه متن، در به کارگیری آن باید جانب احتیاط نگه داشته شود و حجم متون مورد مقایسه حداقل امکان باید یکسان باشد (ibid).

در برخی دیگر از پژوهش‌ها از نشانگرهای نحوی نظیر اجزای کلام^۲ 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. پیش فرض استفاده از نشانگرهای نحوی در تشخیص هویت نویسنده این است که افراد در تولید زبان از الگوهای نحوی کمابیش مشابهی استفاده می‌کنند. نکته‌ای که استفاده از رویکردی خوبنیاد را با چالش مواجه می‌کند، استفاده از ابزارهای خودکار تحلیل نحوی است. درواقع برای اینکه بتوان در تحلیل متنی از ساختار نحوی بهره جست، نیاز است که با استفاده از نرم‌افزارهای رایانشی خاص (نظیر 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 نظیر شمار دستوری اسامی، زمان و نمود دستوری فعل و مشخصه‌های

زیرمقوله‌ای فعل را شامل شوند (Gamon, 2004). برخی پژوهش‌ها همچنین از رویکردهای نظام‌مندتر و دستوربنیان‌تری به این مسئله استفاده کرده‌اند. مثلاً در برخی پژوهش‌ها (Whitelaw & Argamon, 2004; Whitelaw & Patrick, 2004)، مجموعه‌ای از مشخصه‌های معنایی بر اساس دستور نظام‌مند نقشگرا^{۱۱} تعریف شده‌اند که کلمات یا عبارات زبانی به‌خصوصی را با اطلاعات معنایی مرتبط می‌کنند. برخی مطالعات اخیر از معناشناسی قاب‌بنیان نیز در شناسایی نویسنده، به‌ویژه در آثار ترجمه‌شده، استفاده کرده‌اند. معناشناسی قاب‌بنیان نظریه‌ای جدید در معناشناسی است که معنای یک واژه را در دایره دانش دایره‌المعارفی آن واژه و روابط معنایی با واژه‌های دیگر تعریف می‌کند. برای نمونه، طبق این نظریه، معنای واژه‌ای نظیر «فروختن» بدون در نظرگرفتن دانش دایره‌المعارفی قاب^{۱۲} دادوستد 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 تعریف‌شده بر اساس شبکه قاب‌ها دقت شناسایی نویسنده اصلی متن را افزایش دادند. فرض بنیادین آن‌ها در استفاده از معناشناسی قاب‌بنیان این بوده که ترجمه یک متن بسته به سبک مترجم می‌تواند نشانگرهای نحوی و واژگانی نویسنده اصلی را دگرگون کند؛ اما قاب‌های معنایی به‌کاررفته در متن اصلی در متن ترجمه‌شده دگرگون 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 گویش فردی^{۱۳} تبیین کنند که به شیوه منحصربه‌فرد افراد در به‌کارگیری عناصر زبانی نظیر آواها، واژه‌ها، ساختارهای 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 گویش فردی تشکیک کرده‌اند (Barthes, 1977 cited by Barlow, 2010; Jakobson, 1971: 82). مطالعه اخیری که بر روی گفتار پنج سخنگوی خبری کاخ سفید انجام شده (Barlow, 2010) شواهد محکمی از رد پای گویش فردی در

ساخت‌های دستوری به‌دست می‌دهد. فرضیه مورد آزمون این پژوهش، که بعداً تأیید می‌شود، این است که تنوع در گویش فردی گویشوران در قالب مؤلفه‌های اصلی واژگی‌دستوری 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 به‌نگاشتی، سعی می‌کنند مصادیقی از گویش فردی را به‌دست دهند. در این پژوهش پیکره‌ای ۲/۵ میلیون کلمه‌ای متشکل از ۶۳۰۰ ایمیل که توسط ۱۷۶ کارمند شرکت امریکایی انرون نوشته شده است، آماده و به‌لحاظ 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 به‌کار می‌گیرند. در ادامه تحلیل‌ها پژوهشگران همچنین با استفاده از یک آزمایش آماری نشان می‌دهند که با استفاده از به‌نگاشت‌ها می‌توان نویسندگان رایانامه‌های مجهول‌الهویت 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 نویسندگان مختلف در زبان فارسی به دست داده شود و قابلیت تفکیک سبک نویسنده‌ها بر اساس واژه‌های دستوری بررسی شود. بر اساس شواهد موجود، واژه‌های دستوری به‌دلیل اینکه مجموعه‌ای بسته از عناصر زبانی را تشکیل می‌دهند، مستقل از محتوا و موضوع متن به‌کار گرفته می‌شوند، و سرنخ‌هایی از ساخت‌های کلان نحوی به‌دست 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 متونی را که توسط نویسنده‌های مختلف نوشته شده، از یکدیگر 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 جست‌وجوی فرد از میان نویسندگان ناشناخته و بی‌شمار دلالت ندارد. برای مقایسه متنی که هویت نویسنده آن نامعلوم است با متونی که نویسندگان

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

این پژوهش، همسو با تحقیقات مرور شده، مبتنی بر این فرض زبان‌شناختی است که هر فردی دارای گویش فردی منحصر به خود است که می‌تواند در نشانگرهای مختلف متنی از جمله 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 به‌کارگیری صداها، کلمات، دستور 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 اوست، 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 نگاشته‌شده توسط یک نویسنده را استخراج کرد، می‌توان با استخراج مشخصه‌های سبکی یک متن فاقد هویت و مقایسه آن با مؤلفه‌های سبکی استخراج‌شده از متون دارای هویت، مشخص کرد که سبک متن فاقد هویت به سبک کدام نویسنده نزدیک‌تر است.

۳-۱. 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 برمی‌گیرند که محتوای معنایی پررنگ‌تری دارند، تعدادشان نامحدود است و حجم عمده 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 بالایی 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 به‌کار 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 به‌طور ناخودآگاه به‌کار 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 متمایز می‌شدند، ۲۰ واژه دستوری پربسامد تعیین شد.

۳-۲. پیکره پژوهش

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

- سعی بر آن بود موضوعاتی که نویسندگان 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 به‌کارگرفته 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 اطمینان حاصل شود متن به قلم خود نویسنده نوشته شده است. از آنجا که این پژوهش، به‌صورت آزمایشگاهی انجام می‌شود و هدف از آن آزمودن کارایی واژه‌های دستوری در متمایز کردن سبک نوشته‌های فارسی است، اطمینان از اصالت نوشته‌ها بسیار حائز اهمیت است. این دغدغه موجب شد به جای استفاده از مقالات یا نوشته‌جات علمی معمول که حاوی ارجاع به منابع دیگر هستند، از مقالات و

خودنگاشته‌های 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 گردآوری‌شده در آزمایش‌های سبک‌سنجی 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 نویسندگان 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 پیکره پژوهش کنار گذاشته شوند.

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

پس از مرحله پیش‌پردازش، پیکره آماده استفاده شد. تعداد کلمات پیکره اصلی پژوهش به تفکیک نویسندگان آن به‌صورت جدول ۱ است.

جدول ۱: تعداد کلمات موجود در پیکره به تفکیک نویسندگان
Table 1. Number of words per author in the corpus

نویسنده	تعداد کلمات پیکره
صادق زیباکلام	۱۰۸۶۹۰
حمید پارسانیا	۱۰۷۴۹۹
ناصر فکوهی	۱۰۷۱۰۷
رضا داوری اردکانی	۴۴۸۵۴
نیوشا صدر	۳۴۲۴۶

۳-۳. روش تحلیل

برای تحلیل سبک نویسندگان از بسته تخصصی استاتیلو (Eder et al., 2013) که در زبان برنامه‌نویسی آماری R (R Core Team, 2015) ایجاد شده است، استفاده شد. این بسته قابلیت دریافت فایل‌های پیکره، نمونه‌گیری تصادفی از آن‌ها، تهیه لیست واژه‌های پربسامد n -نگاشتی، و انجام تحلیل‌های مورد نظر بر اساس برخی از الگوریتم‌های ازپیش‌موجود را دارد. در این پژوهش از دو روش بی‌نظارت تحلیل مؤلفه‌های اصلی^{۱۸} و تحلیل خوشه‌ای^{۱۹} استفاده شده است. تحلیل مؤلفه‌های اصلی برای بررسی نقش هر یک از واژه‌های دستوری در تمایز سبک نویسنده‌های مختلف و تحلیل خوشه‌ای برای خوشه‌بندی آثار نویسندگان مورد استفاده قرار گرفت. مقیاس فاصله‌ای به‌کار رفته برای محاسبه فاصله آثار نویسندگان با یکدیگر بر اساس مقیاس دلتا (Burrows, 2002) معین شد.

۳-۴. 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 به‌عنوان پیکره از کل پیکره پژوهش نمونه‌گیری شده‌اند. جدول ۲ میزان حجم پیکره برای هر آزمایش را به همراه نویسندگان 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 متون
۱	۵۰۰۰	فکوهی، پارسانیا، زیباکلام
۲	۱۰۰۰	فکوهی، پارسانیا، زیباکلام، صدر، اردکانی
۳	۱۰۰۰	فکوهی، پارسانیا، زیباکلام، صدر، اردکانی
۴	۲۰۰۰	فکوهی، پارسانیا، زیباکلام، صدر، اردکانی
۵	۳۰۰۰	فکوهی، پارسانیا، زیباکلام، صدر، اردکانی
۶	۴۰۰۰	فکوهی، پارسانیا، زیباکلام، صدر، اردکانی

روال انجام آزمایش‌ها به این صورت بوده که ابتدا در آزمایش اول متون ۵۰۰۰۰ واژه‌ای از نویسندگان تحلیل شدند. برای این منظور، متون سه نویسنده دارای تعداد واژه کافی یعنی ۱۰۰۰۰۰ واژه بودند که برای هر نویسنده دو پیکره ۵۰۰۰۰ واژه‌ای به‌طور تصادفی نمونه‌گیری شدند. در آزمایش دوم، حجم نمونه‌متن‌ها به ۱۰۰۰۰ کاهش داده شد و متون همه پنج نویسنده در تحلیل قرار گرفت. در آزمایش سوم، حجم نمونه‌متن‌ها 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 دچار 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 مشخص‌شدن پیکره‌های تحلیل، اولین گام در تحلیل سبک‌های نویسنده‌ها، استخراج واژه‌های دستوری پربسامد بود که نقش نشانگر واژگانی را در متمایز کردن سبک نویسندگان ایفا می‌کردند. واژه‌های دستوری پربسامد با استفاده از فرامین موجود در بسته استایلو استخراج و سپس به‌صورت دستی بررسی شدند تا واژه‌های غیردستوری از آن پالایش شوند. واژه‌های غیردستوری که از فهرست واژه‌های پربسامد حذف شدند عمدتاً شامل مواردی می‌شدند که در جدول ۳ آمده است.

جدول ۳: فهرست واژه‌های محذوف از جدول واژه‌های پربسامد
Table 3. List of words removed from the high-frequency words list

واژه‌های محذوف	مثال
اعداد	یک، دو، سه، ...
علامت جمع بدون نیم‌فاصله	ها، های
وندهای فاقد نیم‌فاصله	می، ای
افعال (ربطی و غیرربطی)	است، بود، شد، ...
ضمایر شخصی	او، آن‌ها، ...
واژه‌های محتوایی	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

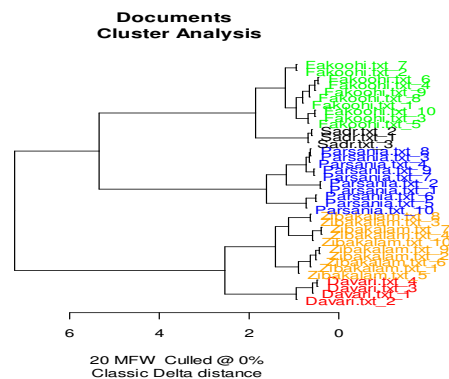
در همه متون به‌کار نرفته بودند، کنار گذاشته می‌شدند. حد کمینه این گزینش ۲۰ درصد و حد بیشینه آن ۸۰ درصد در نظر گرفته شد. گزینش واژه‌های پربسامد با حد ۲۰ درصد بدین معناست که واژه‌هایی که در حداقل ۲۰ درصد از متن‌ها به‌کار رفته باشند، در تحلیل لحاظ خواهند شد. پس از محاسبه تعداد واژه‌های پربسامد و حذف واژه‌های غیردستوری، جدول حاصله آماده تحلیل سبک‌سنجی است. در مرحله بعد، نحوه تمایز نویسندگان و سهم هر کدام واژه‌ها در تمایز سبک‌ها به‌واسطه تحلیل مؤلفه‌های اصلی بررسی شدند. سپس با استفاده از تحلیل خوشه‌ای و بر اساس مقیاس فاصله‌ای دلتا، خوشه‌بندی متون مشابه به‌لحاظ سبکی به دست داده شد. برای ارزیابی دقت خوشه‌بندی، از شاخصی به‌نام شاخص تعدیل‌شده رند^۲ (Hubert & Arabie, 1985) استفاده شد. این شاخص، نسخه بهبودیافته شاخص رند (Rand, 1971) است که خوشه‌بندی تولیدشده توسط الگوریتم را با خوشه‌بندی ایده‌آل (که آثار هر نویسنده باید در خوشه‌های مجزا قرار گیرند) مقایسه می‌کند و میزان مطابقت آن را به‌دست 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.

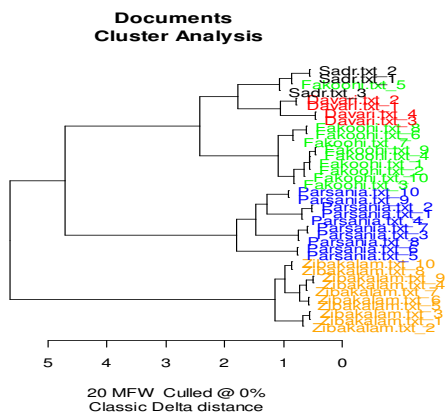
۴. نتایج تحلیل داده‌ها

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



نمودار ۱: تحلیل خوشه‌ای نوشته‌های نویسندگان بر اساس واژه‌های دستوری تک‌نگاشتی در متون ۱۰۰۰۰ واژه‌ای

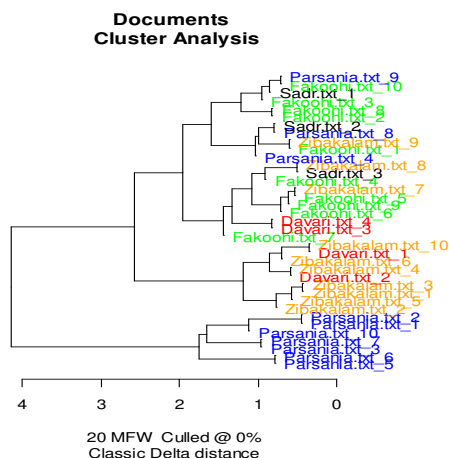
Figure 1. Cluster analysis of authors' writings based on monogram function words in 10000 word texts.



نمودار ۲: تحلیل خوشه‌ای نوشته‌های نویسندگان بر اساس واژه‌های دستوری دو‌نگاشتی در متون ۱۰۰۰۰ واژه‌ای

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

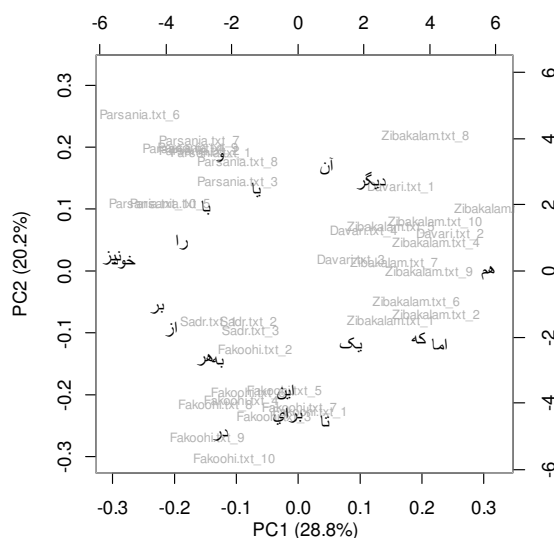
با توجه به تحلیل خوشه‌ای نمودارهای ۱ تا ۳ می‌توان دریافت که واژه‌های دستوری یک، دو و سه‌نگاشتی عملکرد متفاوتی در تفکیک نوشته‌های نویسندگان داشته‌اند. در واقع واژه‌های دستوری تک‌نگاشتی (نمودار ۱) بهترین عملکرد را در مقایسه با واژه‌های دستوری دونگاشتی و سه‌نگاشتی داشته‌اند. با توجه به نمودار ۲، واژه‌های 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. به‌نظر می‌رسد با افزایش توالی واژه‌ها، بسامد توالی‌ها در متون کاهش می‌یابد و این موضوع موجب می‌شود تفکیک متون با افت دقت روبه‌رو شود.

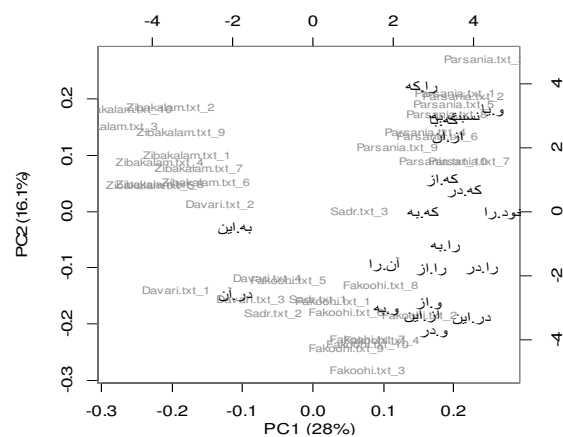


نمودار ۳: تحلیل خوشه‌ای نوشته‌های نویسندگان بر اساس واژه‌های دستوری سه‌نگاشتی در متون ۱۰۰۰۰ واژه‌ای

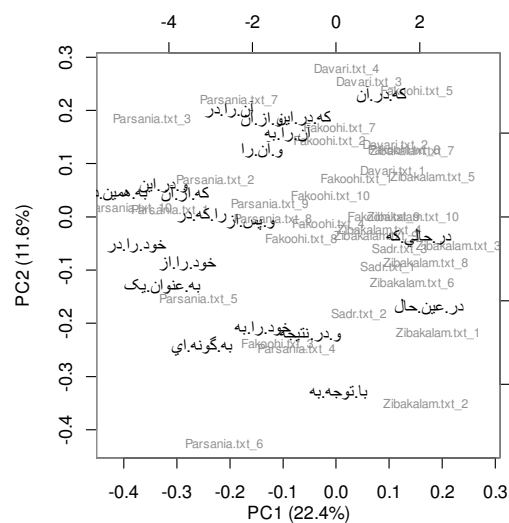
Figure 3. Cluster analysis of authors' writings based on trigram function words in 10000 word texts.

همچنین علاوه بر تحلیل خوشه‌ای، برای شناسایی واژه‌ها یا توالی واژه‌هایی 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^{۲۱} برای واژه‌های دستوری



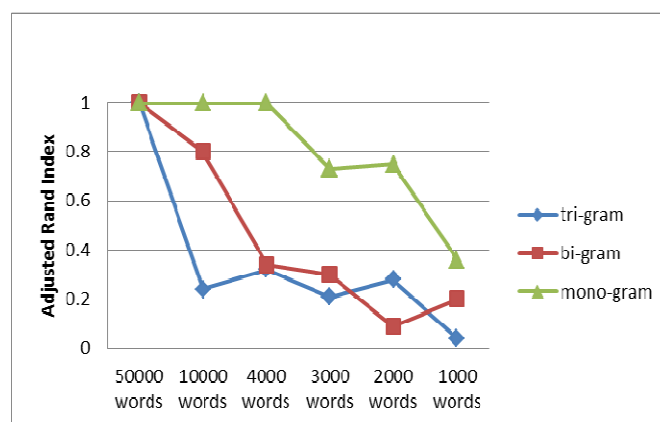


نمودار ۵: تحلیل مؤلفه‌های اصلی و بار عاملی برای واژه‌های دستوری دونگاشتی در متون ۱۰۰۰۰ واژه‌ای
Figure 5. PCA and loadings of bigram function words in 10000 word texts.



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 متنی انجام شد تصویری روشن‌تر از کارایی مقیاس دلتا در طبقه‌بندی متون نویسندگان در این پژوهش به‌دست می‌آید. نمودار ۷ تغییرات شاخص تعدیل‌شده رند را برای واژه‌های تک‌نگاشتی در حجم‌های متنی آزمایش‌شده در این پژوهش نشان می‌دهد.



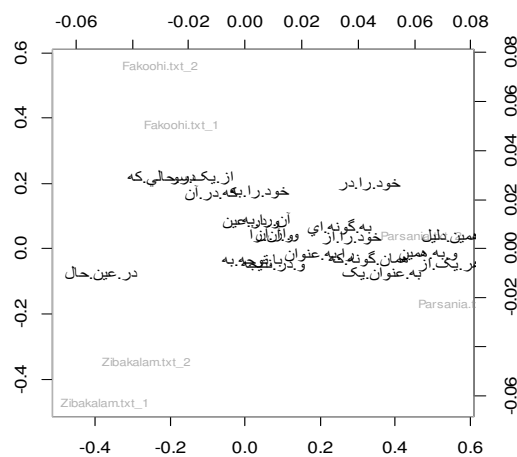
نمودار ۷: تغییرات شاخص تعدیل‌شده رند بر اساس تعداد کلمات متون به تفکیک واژه‌های دستوری تک‌نگاشتی، دو‌نگاشتی و سه‌نگاشتی

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 متنی) بهترین عملکرد را نسبت به واژه‌های دو‌نگاشتی و سه‌نگاشتی داشته است. در سه سطح ۵۰۰۰۰ واژه‌ای، ۱۰۰۰۰ واژه‌ای و ۴۰۰۰ واژه‌ای، دقت واژه‌های تک‌نگاشتی در شناسایی متون ۱۰۰ درصد است و تغییر نمی‌کند. نتایج نشان می‌دهد که متون ۴۰۰۰ واژه‌ای احتمالاً حد کمینه متونی است که تماماً به‌درستی شناسایی می‌شوند. البته چنانچه آستانه دقت را تا حد ۰/۷ پایین بیاوریم، آنگاه می‌توان گفت که تا حد

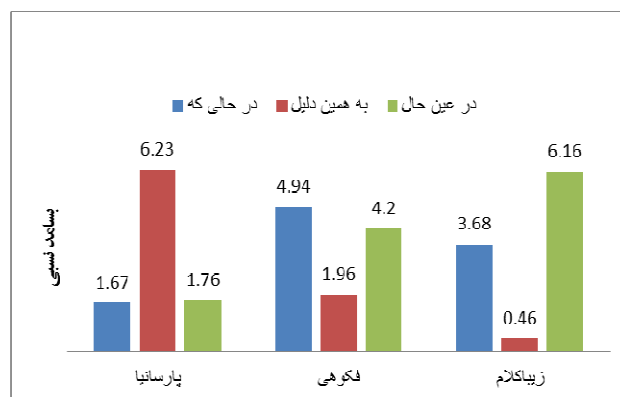
نکته دوم اینکه، در میان واژه‌های دونگاشتی و سه‌نگاشتی، به نظر می‌رسد به‌طور کلی عملکرد واژه‌های دونگاشتی در همه متون (به استثنای متون ۲۰۰۰ واژه‌ای) اندکی بهتر از واژه‌های سه‌نگاشتی بوده است. نکته دیگری که در عملکرد واژه‌های دستوروی سه‌نگاشتی بیشتر مشهود است این است که با کاهش تعداد متون از ۵۰۰۰ به ۱۰۰۰ واژه، شاهد نوسان در دقت شناسایی نویسنده هستیم. مثلاً در متون ۴۰۰ واژه‌ای، با آنکه تعداد کلمات نسبت به متون ۱۰۰۰ واژه‌ای بیش از نصف کاهش یافته است؛ اما دقت شناسایی اندکی بیشتر بوده است که البته به‌نظر می‌رسد معنی‌دار نیست.

یکی از دلایلی که می‌توان بر رفتار متفاوت واژه‌های هـ نگاشتی و افت عملکرد شناسایی نویسنده با حرکت از واژه‌های تکنگاشتی تا سه‌نگاشتی بیان کرد، کاهش تعداد کلمات متون و به‌دنبال 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 متن شناسایی آن‌ها را با مشکل مواجه می‌سازد. از این رو، عملکرد مطلوب توالی‌های واژگانی، در متون حجم می‌تواند بروز کند.



نمودار ۸. بار عاملی واژه‌های دستوری سه‌نگاشتی در متون ۵۰۰۰۰ واژه‌ای

به منظور کسب اطلاعات بیشتر از نحوه به کارگیری واژه‌ها در جملات و بافت دستوری، بررسی توالی‌های دونگاشتی و سه‌نگاشتی پربسامد در مرحله بعد می‌تواند اطلاعات بیشتری در این راستا به دست دهد. برای مثال، نمودار ۸ که نمودار بار عاملی واژه‌های سه‌نگاشتی در متون ۵۰۰۰۰ 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 نظیر «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 نظیر «در عین حال»، «در حالی که» و «به همین دلیل» را می‌توان جزو حروف ربطی مرکب دانست که با وجود صورت چندکلمه‌ای‌شان، ماهیت باهم‌آیی^{۲۲} دارند و معنای واحدی را می‌رسانند. نمودار ۹ بسامدهای نسبی این سه حرف ربطی مرکب را در متون سه نویسنده نشان می‌دهد که از پیکره‌های مربوطه‌شان استخراج شده‌اند. حرف ربط «در عین حال» بیشترین بسامد را در متون زیباکلام 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 زیباکلام و کمتر در متون پارسانیا به چشم می‌خورد.



نمودار ۹: بسامد نسبی برخی سه‌نگاشتی‌های پربسامد 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 به‌وسیلهٔ حرف ربط «به همین دلیل» بیان شده، در متون پارسانیا غلبه داشته است. در حالی که سبک گفتمانی در متون زیباکلام 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 مشخصهٔ بارز متون زیباکلام است. در متون زیباکلام 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 همهٔ نویسندگان به‌کار گرفته می‌شوند، اما اولویت به‌کارگیری آن‌ها توسط نویسندگان متفاوت است. همانطور که در بخش تحلیل‌های پیکره‌ای نیز به آن پرداخته شد، برخی واژه‌های دستوری تقریباً به‌طور کمابیش یکسانی توسط همهٔ نویسندگان استفاده می‌شود؛ اما برخی دیگر با بسامد بیشتری در متون یک نویسنده نسبت به متون دیگر نویسندگان 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 دربردارندهٔ 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 به‌دست 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 ضمائر شخصی هم-

بستگی معنی‌داری با برخی عوامل متنی و غیرمتنی نظیر ژانر، چشم‌انداز روایی نویسنده، جنسیت نویسنده و حتی موضوع متن دارد (Kestemont, 2014). با توجه به اینکه در پژوهش کنونی ضماین شخصی و هر نوع واژه‌ی مرتبط با موضوع متن از فهرست واژه‌های دستوری پریسامد حذف شدند، تفاوت موضوعی و ژانری نمی‌تواند تبیین مناسبی برای نتایج این پژوهش ارائه دهد. قضاوت درباره‌ی همبستگی احتمالی برخی واژه‌های دستوری (غیر از ضماین شخصی) با ژانر و عوامل متنی و غیرمتنی بررسی 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 خوشه‌بندی واژه‌های دستوری تکنگاشتی، دونگاشتی و سه‌نگاشتی به‌دست 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، بلافاصله با افت کارایی در تفکیک صحیح متون مواجه شد. با آنکه متونی با کمینه حجم ۴۰۰۰ واژه به‌عنوان آستانه‌ی بهترین عملکرد الگوریتم دلتا در متون فارسی معین شد؛ اما به‌نظر می‌رسد این آستانه ماهیت حدی ندارد و اتفاق نظری در مورد کمینه متون برای تحلیل مطمئن سبک‌شناختی وجود ندارد. برای مثال، برخی پژوهشگران ۵۰۰۰ واژه را به‌عنوان 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) به‌عنوان حد قابل اعتماد برای پژوهش‌های سبک‌سنجی هستند. نتایج به‌دست‌آمده برای زبان فارسی نیز به‌نوعی همسو با یافته‌ها در دیگر زبان‌هاست. بنابراین در پاسخ به پرسش سوم پژوهش، براساس نتایج این پژوهش فعلاً می‌توان کمینه ۴۰۰۰ واژه را به‌عنوان 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 متنی مورد‌استفاده، می‌توانند در ارزیابی اعتبار نتایج این پژوهش مؤثر باشند و موجب همگرایی در تخمین کمینه واژه‌های موردنیاز برای تفکیک سبک نویسندگان 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 15th 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* (5th 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/lc/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 49th 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 3rd 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* (7th 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.