واژه‌های دستوری به‌مثابه نشانگرهای گویش فردی: رویکردی پیکره‌ای به شناسایی هویت نویسنده در زبان فارسی

رامین گلشائی*

استادیار زبان‌شناسی، دانشگاه الزهرا، تهران، ایران

چکیده

شناسایی هویت نویسنده یکی از حوزه‌های تحقیقاتی مهم در حیطه‌ی زبان‌شناسی حقوقی است که موضوع پژوهش‌های زبان‌شناختی و رایانشیِ 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 نویسنده‌ی 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 نشانگرهای موثری برای کدگذاری گویش فردی و ممیز سبک نویسندگان باشند. در این پژوهش، ابتدا پیکره‌هایی متنی از نوشته‌های ۵ محقق و نویسنده‌ی معاصر جمع‌آوری و سپس با استفاده از نرم‌افزار خطایاب وفا، استانداردسازی شدند. با استفاده از بسته‌ی 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 مولفه‌های 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 کمینه‌ی متن برای شناسایی موفقیت‌آمیز نویسنده در متون فارسی حدود ۴۰۰۰ واژه بر اساس ۲۰ واژه‌ی دستوری پربسامد است.

کلیدواژه‌ها: گویش فردی، شناسایی نویسنده، تحلیل پیکره‌ای، زبان‌شناسی حقوقی، روش دلتا

---

* نویسندهٔ مسئول مقاله: E-mail: golshaie@alzahra.ac.ir

۱. مقدمه

شناسایی اصالت نوشته‌ها و کشف هویت نویسنده‌ی متون از دیرباز موردتوجه انسان بوده است. یکی از قدیمی‌ترین نمونه‌های آن، اثبات جعلی‌بودن سند «بخشش کنستانتین» توسط لورنزو والا[۱] در سال ۱٤٤۰ 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 (فرونتینی[۲] و همکاران، ۲۰۰۸). پژوهش‌هایی که امروزه در این حوزه انجام می‌شوند، عمدتاً با عنوان سبک‌سنجی[۳] شناخته می‌شوند. در عصر کنونی با تحولی که رایانه در پردازش متون حجیم ایجاد کرده و کاربردهایی که پژوهش‌های سبک‌سنجی در زبان‌شناسی حقوقی [٤] (توتی و هاردکَسل[٥]، ۱۹۸۷) و کشف سرقت علمی و ادبی[٦] (اشتاین و مِیر چو آیسن[۷]، ۲۰۰۹) پیدا کرده است، بررسی نظام‌مند و دقیق متون نوشتاری در زبان‌های مختلف می‌تواند منجربه افزایش درک کنونی محققان نسبت به جوانب مختلف سبک‌های زبانی گردد.

پژوهش‌های نوین سبک‌سنجی که مبتنی بر تحلیل‌های رایانشی هستند، بر مولفه‌های مختلفی از متون تمرکز می‌کنند. این مولفه‌ها می‌توانند در سطح واژگانی (غنای واژه‌ها، بسامد واژه‌ها، غیره)، نحوی (ساخت جمله، بسامد قواعد نحوی، خطاها، اجزای کلام)، و معنایی (هم‌معناها، وابستگی‌های معنایی) باشند (استاماتاتوس[۸]، ۲۰۰۹). تحقیقات انجام‌گرفته درباب مولفه‌های واژگانی نشان می‌دهد که نوع بخصوصی از واژه‌ها، یعنی واژه‌های دستوری، ویژگی‌های منحصربه‌فردی برای شناسایی سبک نویسنده دارند. برخی از این ویژگی‌ها عبارتند از: بسامد بالای آنها در متن که دست پژوهشگر را در تحلیل متون کوتاهتر باز می‌گذارد، عدم تاثیرپذیری واژه‌های نقشی از موضوع یا ژانر نوشته، و عدم کنترل خودآگاه نویسنده بر کاربرد واژه‌های دستوری (بینونگو[۹]، ۲۰۰۳). با اینکه در زبان انگلیسی و زبان‌های دیگر تحقیقات متعددی درباره نقش واژه‌های دستوری در شناسایی سبک نویسنده انجام شده (آرگامون و لِویتان[۱۰]، ۲۰۰۵؛ باروز[۱۱]، ۱۹۸۷؛ ژائو و زوبل[۱۲]، ۲۰۰۵؛ سگارا[۱۳] و همکاران، ۲۰۱۵؛ موستلر و والاس[۱٤]، ۱۹٦٤؛ هولمز[۱٥] و همکاران، ۲۰۰۱)، اما در زبان فارسی پژوهش‌های چندانی درباب این موضوع صورت نگرفته است. از سوی دیگر، از آنجا که واژه‌های دستوری در زبان‌های مختلف به شکل‌های مختلفی نمود می‌یابند، پژوهش‌هایی که نقش موثر واژه‌های دستوری در تفکیک سبک‌های نوشتاری در یک زبان خاص را تایید می‌کنند، لزوماً به‌معنای تایید آن در زبان‌های دیگر ازجمله زبان فارسی نمی‌تواند باشد.

در این پژوهش، کارایی واژه‌های دستوریِ زبان فارسی در تفکیک سبک زبانی نویسنده‌های مختلف بررسی شده است. مسئلهٔ این پژوهش در قالب سه پرسش قابل تعریف است که عبارتند از: (۱) آیا واژه‌های دستوری در نثر فارسی قابلیت تفکیک سبک نویسنده‌ها را دارند؟ (۲) کدام‌یک از توالی‌های یک تا سه واژه‌ای (تک‌نگاشتی[۱۶]، دونگاشتی[۱۷] و سه‌نگاشتی[۱۸]) 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 ۵ محقق و نویسندهٔ معاصر جمع‌آوری و سپس با استفاده از نرم‌افزار خطایاب وفا، استانداردسازی شدند. با استفاده از بستهٔ تحلیل آماری stylo در نرم‌افزارR، 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 مولفه‌های اصلی و همچنین تحلیل خوشه‌ای بر اساس مقیاس فاصله‌ای دلتا بررسی شد. در بخش بعد پیشینهٔ برخی پژوهش‌های انجام‌شده در حوزهٔ سبک‌سنجی و موضوعات مرتبط معرفی خواهد شد. در بخش ۳، مبانی نظری و روش پژوهش شرح داده خواهد شد. بخش ٤ به ارائهٔ نتایج تحلیل داده‌ها می‌پردازد و در نهایت بخش ۵ به بحث و جمع‌بندی نتایج پژوهش اختصاص دارد.

۲. پیشینهٔ پژوهش

در بسیاری از پژوهش‌های انجام‌شده در حوزهٔ تشخیص هویت نویسنده، محققان از نشانگرهای سبکی[۱۹] متعددی تحت عنوان نشانگرهای واژگانی، نحوی و معنایی به‌عنوان شاخص‌های ممیزِ سبک نویسنده استفاده کرده‌اند. نمونهٔ معروف مطالعاتی که از نشانگرهای واژگانی در شناسایی نویسنده استفاده کرده، مربوط به شناسایی نویسندهٔ ۱۲ مقالهٔ 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 رای مثبت دهند. این مقالات توسط سه نفر یعنی «الکساندر همیلتون»، «جان جِی» و «جیمز مَدیسون» نوشته شده بودند اما به‌طور ناشناس منتشر شدند. در مطالعه‌ای که با استفاده از روش‌های آماری و به‌کارگیری بسامد

واژه‌های دستوری به‌مثابهٔ نشانگرهای واژگانی صورت گرفت (موستلر و والاس، ۱۹٦٤)، مشخص شد که ۱۲ مقاله توسط «مدیسون»، یکی از سه نویسندهٔ مقالات، نوشته شده است. در حوزهٔ 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 (مدبر دباغ[۲۰]، ۲۰۰۷). نتایج این پژوهش که بر روی نظم و نثر فارسی انجام شده نشان می‌دهد واژه‌های دستوری قادر به تفکیک سبک نویسنده‌ها هستند. نشانگر واژگانی دیگری که مبتنی بر بسامد واژگانی عمل می‌کند، نشانگر غنای واژگانی[۲۱] است. این شاخص که تنوع واژه‌های مورداستفاده توسط فرد را نشان می‌دهد معمولاً با تقسیم تعداد واژه‌های منحصربه‌فرد (گونه) متن بر تعداد کل کلمات (نمونه) به‌کار رفته به‌دست می‌آید (یوهانسن[۲۲]، ۲۰۰۸). در مورد غنای واژگانی نیز، تحقیقات صورت‌گرفته در حوزهٔ روان‌شناسی زبان نشان می‌دهد افراد مختلف غنای واژگانی متفاوتی دارند (کارول[۲۳]، ۲۰۰۸). شاخص غنای واژگانی معمولاً در کنار نشانگرهای دیگر متنی برای تفکیک سبک نگارشی نویسنده‌ها مورد استفاده قرار گرفته است (ن.ک. فرهمندپور و نیک‌مهر[۲٤]، ۲۰۱۵). اگرچه شاخص غنای واژگانی می‌تواند تصویری از تنوع واژگانی افراد به‌دست دهد اما به‌دلیل تحت‌تاثیر قرارگرفتن این شاخص از اندازهٔ متن، در به‌کارگیری آن باید جانب احتیاط نگه داشته شود و حجم متون موردمقایسه حدالامکان باید یکسان باشد (همان).

در برخی دیگر از پژوهش‌ها از نشانگرهای نحوی نظیر اجزای کلام[۲۵] به‌عنوان ممیز سبک نویسنده استفاده کرده‌اند (گامون[۲٦]، ۲۰۰٤؛ کوپل و شلر[۲۷]، ۲۰۰۳). 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 مشابهی استفاده می‌کنند. نکته‌ای که استفاده از رویکردی نحوبنیاد 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 (نظیر تقطیع‌گر نحوی یا برچسب‌زن نحوی) بازنمایی نحوی را جمله به‌دست دهیم. این پیش‌پردازش در بسیاری از زبان‌ها به‌دلیل محدودیت در دسترسی به نرم‌افزارهای لازم امکانپذیر نیست.

مشخصه‌های معنایی متن، یکی دیگر از مولفه‌هایی هستند که این پتانسیل را دارند، برخی مشخصه‌های سبکی نویسندهٔ 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 شود (گامون، ۲۰۰٤). برخی پژوهش‌ها همچنین از رویکردهای نظام‌مندتر و دستوربنیان‌تری به این مسئله استفاده کرده‌اند. مثلاً در برخی پژوهش‌ها (وایتلا و آرگامون[۲۸]، ۲۰۰٤؛ وایتلا و پاتریک[۲۹]، ۲۰۰٤)، مجموعه‌ای از مشخصه‌های معنایی بر اساس دستور نظام‌مند نقشگرا[۳۰] تعریف شده‌اند که کلمات یا عبارات زبانی بخصوصی 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 دایرهٔ دانش دایرةالمعارفی آن واژه و روابط معنایی با واژه‌های دیگر تعریف می‌کند. برای نمونه، طبق این نظریه، معنای واژه‌ای نظیر «فروختن» بدون در نظرگرفتن دانش دایرةالمعارفیِ قاب[۳۱] دادوستد و ارتباط آن با دیگر واژه‌ها نظیر «خریدن»، «خریدار»، «کالای مورد معامعله» و غیره قابل درک نخواهد بود. در پژوهشی اخیر (هدگارد و سیمونسن[۳۲]، ۲۰۱۱) که برای شناسایی نویسندهٔ یک متن ترجمه‌شده انجام شده، پژوهشگران با استفاده از روابط معناییِ تعریف‌شده بر اساس شبکهٔ قاب‌ها دقت شناسایی نویسندهٔ اصلی متن را افزایش دادند. فرض بنیادین آنها در استفاده از معناشناسی قاب‌بنیان این بوده که ترجمهٔ یک متن بسته به سبک مترجم می‌تواند نشانگرهای نحوی و واژگانی نویسندهٔ 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 زبان‌شناختیِ گویش فردی[۳۳] تبیین کنند که به شیوهٔ منحصربه‌فرد افراد در به‌کارگیری عناصر زبانی نظیر آواها، واژه‌ها، ساخت‌های دستوری و غیره گفته می‌شود (وارداف و فولر[۳٤]، ۲۰۱۵: ۹). اهمیت این پژوهش‌ها آنجا

مشخص می‌شود که برخی از زبان‌شناسان در گذشته همواره نسبت به واقعیتِ گویش فردی تشکیک کرده‌اند (یاکوبسن[۳۵]، ۱۹۷۱: ۸۲؛ بارت[۳۶]، ۱۹۷۷: ۲۱ به نقل از بارلو[۳۷]، ۲۰۱۰). مطالعه‌ای اخیری که بر روی گفتار پنج سخنگوی خبری کاخ سفید انجام شده (بارلو،۲۰۱۰) شواهد محکمی از رد پای گویش فردی در ساخت‌های دستوری به‌دست می‌دهد. فرضیهٔ مورد آزمون این پژوهش، که بعداً موردتایید قرار می‌گیرد، این است که تنوع در گویش فردی گویشوران در قالب مولفه‌های اصلی واژی‌دستوری نمود می‌یابد و محدود به استعمال برخی واژه‌ها و عبارات خاص و منحصربه‌فرد نمی‌شود.

در پژوهشی دیگر (جانسون و رایت[۳۸]، ۲۰۱٤) که بر روی شناسایی نویسندهٔ رایانامه‌های شرکت انرژی اِنرون انجام گرفته، محققان با استفاده از توالی‌های واژگانی n-نگاشتی، سعی می‌کنند مصادیقی از گویش فردی را به‌دست دهند. در این پژوهش پیکره‌ای ۲/۵ میلیون کلمه‌ای متشکل از ٦۳۰۰۰ ایمیل که توسط ۱۷٦ کارمند شرکت آمریکایی اِنرون نوشته شده، آماده شده و به‌لحاظ سبکی مورد تحلیل واقع می‌شوند. نتایج تحلیل نشان می‌دهد که هر یک از کارمندان شرکت الگوهای سبکی متمایز از الگوی دیگران را در نگارش خود به‌کار می‌گیرند. در ادامهٔ تحلیل‌ها پژوهشگران همچنین با استفاده از یک آزمایش آماری نشان می‌دهند که با استفاده از n-نگاشت‌ها 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 موثری 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 فرآیندی تعریف کرد که طی آن با اندازه‌گیری برخی مولفه‌های متنی می‌توان متونی را که توسط نویسنده‌های مختلف نوشته شده، از یکدیگر متمایز کرد (استاماتاتوس، ۲۰۰۹). پیش‌فرض مهمی که در تحقیقات مربوط به شناسایی نویسنده وجود دارد این است که نویسندهٔ یک متن فاقد نویسنده، از بین چند نویسندهٔ مظنون انتخاب می‌شود و بنابراین بدیهی است که فرآیند «شناسایی» به جستجوی فرد از میان نویسندگان ناشناخته و بی‌شمار دلالت ندارد. برای مقایسهٔ متنی که هویت نویسندهٔ 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 مرورشده، 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 شیوهٔ منحصربفرد [هر گویشور] در صحبت‌کردن، شامل به‌کارگیری صداها، کلمات، دستور زبان و سبک زبانی است» (وارداف و فولر، ۲۰۱۵: ۹). اگر بخواهیم از استعارهٔ 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 حوزهٔ شناسایی نویسندهٔ 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 مقایسهٔ آن با مولفه‌های سبکیِ 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 دستهٔ عمدهٔ واژه‌های محتوایی و واژه‌های دستوری تقسیم می‌شوند. واژه‌های محتوایی تمام واژه‌هایی را در می‌گیرند که

محتوای معنایی پررنگ‌تری دارند، تعدادشان نامحدود است و حجم عمدهٔ 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، به‌دلایل زیر نشانگرهای واژگانی از مقولهٔ واژه‌های دستوری انتخاب شده‌اند (کستمونت، ۲۰۱٤):

- واژه‌های دستوری بسامد بالایی نسبت به واژه‌های محتوایی در متون دارند. این ویژگی امتیاز بزرگی در حوزهٔ تشخیص سبک نویسنده محسوب می‌شود، چرا که در دنیای واقعی و در موقعیت‌های کاربردی، حجم متون ایده‌آل نیست و گاهی محقق ناچار است با داده‌های اندکی که در اختیار دارد، سبک نویسندهٔ 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. به‌همین 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 برخلاف 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 واژهٔ دستوری پربسامد تعیین شد.

۳-۲. پیکرهٔ پژوهش

پیکرهٔ 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»[۳۹] ، وبگاه شخصی صادق زیباکلام[٤۰]، وبگاه شخصی رضا داوری اردکانی[٤۱]، و وبگاه شخصی حمید پارسانیا[٤۲] 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 بالابردن روایی پژوهش مدنظر بوده است.

– در جمع‌آوری متون همواره این موضوع مورد تاکید 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 متمایزکردن سبک نوشته‌های فارسی است، اطمینان از اصالت نوشته‌ها بسیار حائز اهمیت است. این دغدغه موجب شد به‌جای استفاده از مقالات یا نوشتجات 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 دستیِ اولیه، کل پیکره به‌طور خودکار توسط نرم‌افزار خطایاب وفا (فیلی[۴۳] و همکاران، ۲۰۱۶) استانداردسازی شد. این خطایاب به‌صورت افزونه‌ای در نرم‌افزار مایکروسافت وورد 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 املایی در متون بوده است.

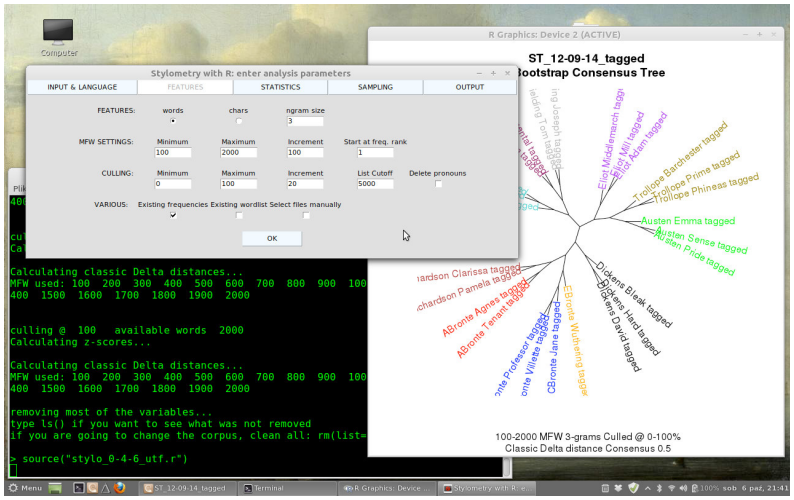
پس از مرحلهٔ پیش‌پردازش، پیکره آمادهٔ استفاده شد. تعداد کلمات پیکرهٔ اصلی پژوهش به‌تفکیک نویسندگان آن به‌صورت جدول ۱ است.

جدول ۱. تعداد کلمات موجود در پیکره به‌تفکیک نویسندگان

نویسنده	تعداد کلمات پیکره
صادق زیباکلام	۱۰۸٦۹۰
حمید پارسانیا	۱۰۷٤۹۹
ناصر فکوهی	۱۰۷۱۰۷
رضا داوری اردکانی	٤٤۸٥٤
نیوشا صدر	۳٤۲٤٦

۳-۳. روش تحلیل

برای تحلیل سبک نویسندگان از بستهٔ تخصصی استایلو (اِدِر[٤٤] و همکاران، ۲۰۱۳) 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، تهیهٔ 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 مولفه‌های اصلی[٤٥] و تحلیل خوشه‌ای[٤٦] استفاده شده تحلیل مولفه‌های 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 محاسبهٔ فاصلهٔ آثار نویسنده‌ها با یکدیگر بر اساس مقیاس دلتا (باروز، ۲۰۰۲) معین شد. شکل ۱ برخی امکانات قابل تنظیم در رابط کاربری بستهٔ استایلو را نشان می‌دهد.



شکل ۱. برخی امکانات رابط کاربردی در بستهٔ استایلو (اِدِر و همکاران، ۲۰۱۳)

۳-۴. روال انجام پژوهش

این پژوهش در قالب ٦ 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 پیکرهٔ پژوهش نمونه‌گیری شده‌اند. جدول ۲ میزان حجم پیکره برای هر آزمایش به همراه نویسندگان نمونه متن‌ها نشان می‌دهد.

جدول ۲. حجم تقریبی پیکره مورد استفاده در هر آزمایش

آزمایش	حجم پیکره برای هر نمونه متن	نویسنده متون
۱	۵۰۰۰۰	فکوهی، پارسانیا، زیباکلام
۲	۱۰۰۰۰	فکوهی، پارسانیا، زیباکلام، صدر، اردکانی
۳	۱۰۰۰	فکوهی، پارسانیا، زیباکلام، صدر، اردکانی
٤	۲۰۰۰	فکوهی، پارسانیا، زیباکلام، صدر، اردکانی
٥	۳۰۰۰	فکوهی، پارسانیا، زیباکلام، صدر، اردکانی
٦	٤۰۰۰	فکوهی، پارسانیا، زیباکلام، صدر، اردکانی

روال انجام آزمایش‌ها به این صورت بوده که ابتدا در آزمایش اول متون ۵۰۰۰۰ کلمه‌ای از نویسندگان تحلیل شدند. برای این منظور، متون سه نویسنده دارای تعداد واژهٔ کافی یعنی ۱۰۰۰۰۰ واژه بودند که برای هر نویسنده دو پیکرهٔ ۵۰۰۰۰ کلمه‌ای به‌طور تصادفی نمونه‌گیری شدند. در آزمایش دوم، حجم نمونه‌متن‌ها به ۱۰۰۰۰ کاهش داده شد و متون همهٔ پنج نویسنده در تحلیل قرار گرفت. در آزمایش سوم، حجم نمونه‌متن‌ها به ۱۰۰۰ 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 ۱۰۰۰ واژه به حجم متون اضافه شد تا دقت شناسایی نویسنده‌ها حداقل در سطح واژه‌های دستوری تک‌نگاشتی به ۱۰۰ درصد برسد.

پس از مشخص‌شدن پیکره‌های تحلیل، اولین گام در تحلیل سبک‌های نویسنده‌ها، استخراج واژه‌های دستوری پربسامد بود که نقش نشانگر واژگانی را در متمایزکردن سبک نویسندگان ایفا می‌کردند. واژه‌های دستوری پربسامد با استفاده از فرامین موجود در بستهٔ 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

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، ...

در طی فرآیند 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 همهٔ متون به‌کار نرفته بودند، کنار گذاشته می‌شدند. حد کمینهٔ این گزینش ۲۰٪ و حد

بیشینهٔ آن ۸۰٪ در نظر گرفته شد. گزینش واژه‌های پربسامد با حد ۲۰٪ بدین معناست که واژه‌هایی که در حداقل ۲۰٪ از متن‌ها به‌کار رفته باشند، در تحلیل لحاظ خواهند شد.

پس از محاسبهٔ تعداد واژه‌های پربسامد و حذف واژه‌های غیردستوری، جدول حاصله آمادهٔ تحلیل سبک‌سنجی است. در مرحلهٔ بعد، نحوهٔ تمایز نویسندگان و سهم هر کدام واژه‌ها در تمایز سبک‌ها به‌واسطهٔ تحلیل مولفه‌های 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 تعدیل‌شدهٔ رَند[۴۷] (هوبرت و ارابی[۴۸]، ۱۹۸۵) استفاده شد. این شاخص، نسخهٔ بهبودیافتهٔ شاخص رَند (رَند[۴۹]، ۱۹۷۱) است 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 اشیاء مشخص در نظر بگیریم، شاخص رَند را می‌توان طبق فرمول زیر برای مقایسهٔ 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؛

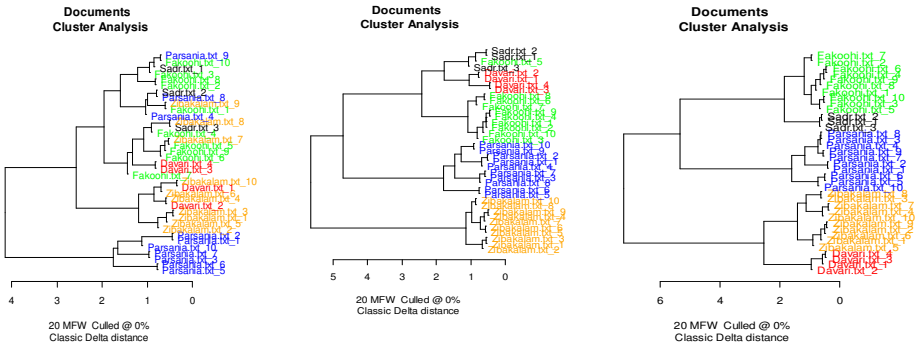
b. جفت‌هایی که در طبقه‌بندی U هم‌گروه ولی در طبقه‌بندی V غیرهم‌گروه هستند؛

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. برای نمونه، شکل ۲ نمودار درختی تحلیل خوشه‌ای را برای واژه‌های دستوری تک‌نگاشتی، دونگاشتی و سه‌نگاشتی در سطح ۱۰۰۰۰ واژه نشان می‌دهد.



الف‌ب‌ج

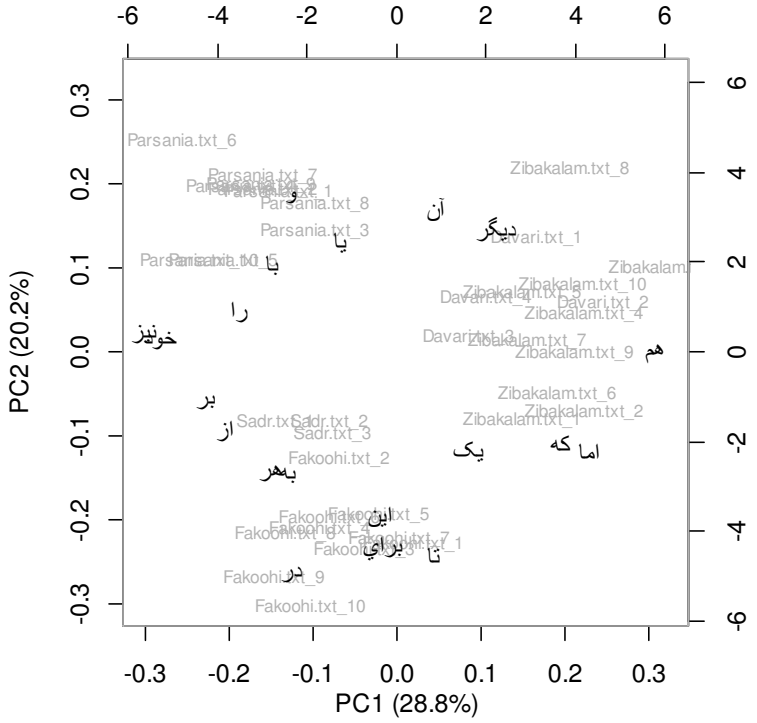
شکل ۲. کارایی واژه‌های دستوری تک‌نگاشتی (الف)، دونگاشتی (ب) و سه‌نگاشتی (ج) در تفکیک نویسنده‌ها در سطح متون ۱۰۰۰۰ واژه

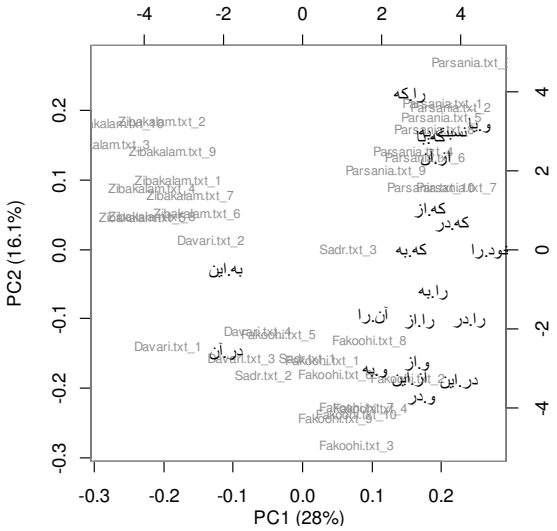
با توجه به نمودار درختی شکل ۲ می‌توان دریافت که واژه‌های دستوری یک، دو و سه‌نگاشتی عملکرد متفاوتی در تفکیک نوشته‌های نویسندگان داشته‌اند. در واقع واژه‌های دستوری تک‌نگاشتی (شکل ۲-الف) بهترین عملکرد را در مقایسه با واژه‌های دستوری دونگاشتی و سه‌نگاشتی داشته‌اند. با توجه به شکل ۲، واژه‌های دستوری دونگاشتی در تفکیک متون مربوط به «داوری»، «صدر» و «فکوهی» اندکی دچار خطا شده است. میزان خطا در واژه‌های دستوری سه‌نگاشتی 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 روبرو شود.

همچنین علاوه بر تحلیل خوشه‌ای، برای شناسایی واژه‌ها یا توالی واژه‌هایی که بیشترین تاثیر را در تفکیک متون نویسنده‌ها داشته‌اند، داده‌ها با استفاده از تحلیل مولفه‌های اصلی بررسی شدند. شکل ۳ تا ۵ نتایج تحلیل مولفه‌های اصلی را در قالب نمودار بار عاملی[۵۰] برای واژه‌های دستوری تک‌نگاشتی (۳)، دونگاشتی (٤) و سه‌نگاشتی (٥) در سطح ۱۰۰۰۰ واژه نشان می‌دهد. در این نمودارها، متون نویسنده‌ها با رنگ خاکستری و واژه‌های دستوری (بار عاملی) با رنگ مشکی متمایز شده‌اند. نزدیکی برخی (توالی) واژه‌های دستوری با متون یک نویسنده، نشانگر غالب‌بودن آن (توالی) واژه دستوری در متون آن نویسنده است. مثلاً در شکل ۳،

واژه‌های دستوری «و»، «یا» و «با» در متون «پارسانیا» دارای بسامد بیشتری نسبت به متون نویسنده‌های دیگر بوده‌اند. شکل ٤ و شکل ٥ نیز به‌همین ترتیب، توزیع توالی‌های دوتایی و سه‌تایی واژه‌های دستوری را در متون نویسنده‌ها نشان می‌دهد. همچنین شاخص‌های کمّی مولفه‌های اصلی اول و دوم در شکل ۳ نشان می‌دهد که در مجموع ٤۹٪ از گونه‌گونی داده‌ها در قالب این دو مولفهٔ اصلی قابل توضیح است؛ مولفهٔ اول ۲۸/۸٪ از گونه‌گونی را در محور افقی و مولفهٔ دوم ۲۰/۲٪ 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.

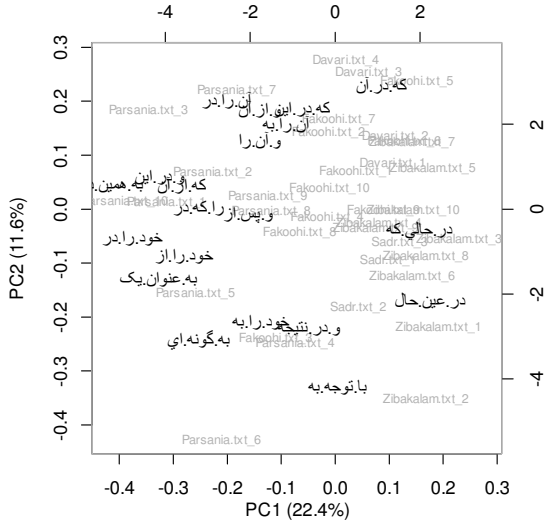
شکل ۳. تحلیل مولفه‌های اصلی و بار عاملی برای واژه‌های دستوری تک‌نگاشتی در سطح متون ۱۰۰۰۰ واژه





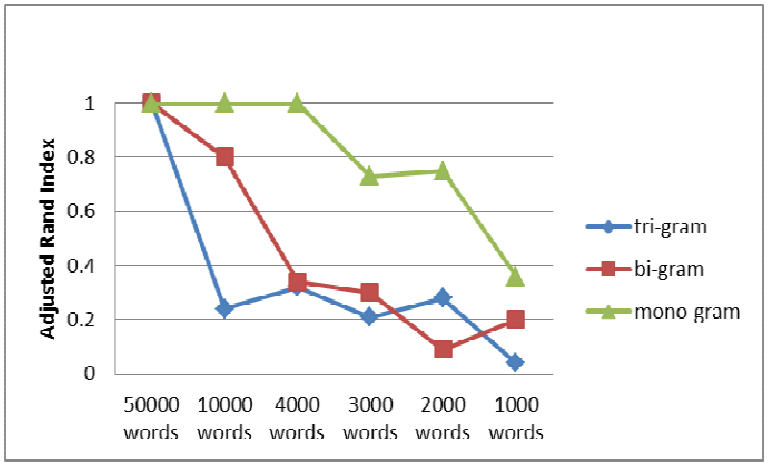
شکل ٤. تحلیل مولفه‌های اصلی و بار عاملی برای واژه‌های دستوری دونگاشتی در سطح متون

۱۰۰۰۰ واژه



شکل ٥. تحلیل مولفه‌های اصلی و بار عاملی برای واژه‌های دستوری سه‌نگاشتی در سطح متون ۱۰۰۰۰ واژه

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. شکل ٦ نمودار تغییرات شاخص تعدیل‌شدهٔ رَند را برای واژه‌های nـنگاشتی در حجم‌های متنی آزمایش‌شده در این پژوهش نشان می‌دهد.

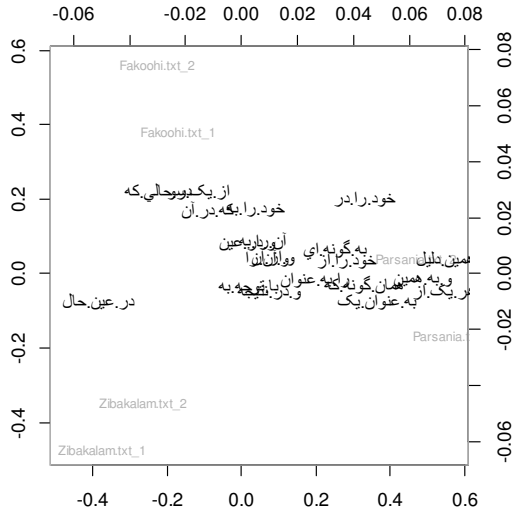


شکل ٦. نمودار nـنگاشتی تغییرات شاخص تعدیل‌شدهٔ رَند بر اساس تعداد کلمات متون

با توجه به شکل ٦، 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 واژه‌ای ٥٠٠٠٠ کلمه، ١٠٠٠٠ کلمه و ٤٠٠٠ کلمه، دقت واژه‌های تک‌نگاشتی در شناسایی متون ١٠٠٪ است و تغییر نمی‌کند. نتایج نشان می‌دهد که متون ٤٠٠٠ کلمه‌ای احتمالاً حد کمینهٔ متونی است که تماماً به‌درستی شناسایی می‌شوند. البته چنانچه آستانهٔ دقت را تا حد ٠/٧ پایین بیاوریم، آنگاه می‌توان گفت که تا حد ٢٠٠٠ کلمه نیز عملکرد الگوریتم با استفاده از واژه‌های تک‌نگاشتی پذیرفتنی است.

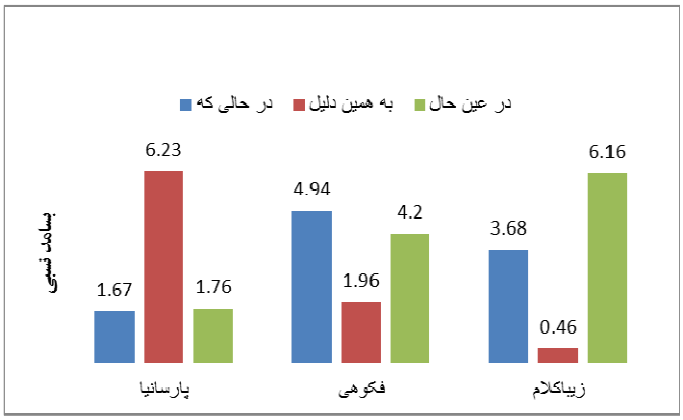
نکتهٔ دوم اینکه، در میان واژه‌های دونگاشتی و سه‌نگاشتی، به نظر می‌رسد به‌طور کلی عملکرد واژه‌های دونگاشتی در همهٔ متون (به استثناء متون ۲۰۰۰ کلمه‌ای) اندکی بهتر از واژه‌های سه‌نگاشتی بوده است. نکتهٔ دیگری که در عملکرد واژه‌های دستوری سه‌نگاشتی بیشتر مشهود است این است که با کاهش تعداد متون از ۵۰۰۰۰ به ۱۰۰۰ واژه، شاهد نوسان در دقت شناسایی نویسنده هستیم. مثلاً در سطح متون ٤۰۰۰ کلمه‌ای، با آنکه تعداد متون نسبت به سطح ۱۰۰۰۰ کلمه‌ای 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 nـنگاشتی و افت عملکرد شناسایی نویسنده با حرکت از واژه‌های تک‌نگاشتی تا سه‌نگاشتی عنوان کرد، کاهش تعداد کلمات متون و به‌دنبال آن کاهش رخداد توالی‌های واژگانی در پیکره است. در واقع با کاهش تعداد کلمات در پیکرهٔ متون، احتمال اینکه توالی‌های چندگانه در متن تکرار شوند، کاهش بیشتری می‌یابد. اگرچه توالی‌های چندگانهٔ واژگانی، نشانگرهای دقیق‌تری از روال‌های خودکارشدهٔ واژگانی می‌توانند باشند، اما کاهش تعداد کلمات متن شناسایی آنها را با مشکل مواجه می‌سازد. از این‌رو، عملکرد مطلوب توالی‌های واژگانی در متون حجیم می‌تواند بروز کند.



شکل ۷. نمودار بار عاملی واژه‌های سه‌نگاشتی در متون ۵۰۰۰۰ کلمه‌ای

به‌منظور کسب اطلاعات بیشتر از نحوهٔ به‌کارگیری واژه‌ها در جملات و بافت دستوری، بررسی توالی‌های دونگاشتی و سه‌نگاشتی پربسامد در مرحلهٔ بعد می‌تواند اطلاعات بیشتری در این راستا به‌دست دهد. برای مثال، شکل ۷ که نمودار بار عاملی واژه‌های سه‌نگاشتی در متون ۵۰۰۰۰ کلمه‌ای است، اطلاعات جالب توجهی را در اختیارمان قرار می‌دهد. بر اساس این نمودار می‌توان دریافت که در سطح مولفهٔ 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 تاثیر را 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 علارغم صورت چندکلمه‌ای‌شان، ماهیت باهم‌آیی[۵۱] دارند و معنای واحدی را می‌رسانند. شکل ۸ 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 پارسانیا است. همچنین حرف ربط «در حالی که» بیشتر در متون فکوهی، سپس در متون زیباکلام و کمتر در متون پارسانیا به چشم می‌خورد.



شکل ۸. بسامد نسبی برخی سه‌نگاشتی‌های پربسامد ممیز سبکی در پیکرهٔ متنی سه نویسنده

بر اساس اطلاعات شکل ۸ 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. بحث و نتیجه‌گیری

نتایج این پژوهش نشان می‌دهد که نویسنده‌های مختلف واژه‌های دستوری را به‌طور مشابه به‌کار نمی‌گیرند. در واقع با وجود اینکه برخی واژه‌های دستوری پربسامد توسط همهٔ 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 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 ریشه در گویش فردی گویشوران 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 زبانیِ 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 بخصوص باشد. در واقع می‌توان این احتمال را مدنظر داشت که ساختار گفتمانی یک ژانر بخصوص بر گزینش واژه‌های دستوری خاص تاثیر بگذارد. برای مثال، بسیاری از پژوهش‌های سبک‌سنجی نشان داده است که به‌کارگیری ضمایر

شخصی همبستگی معنی‌داری با برخی عوامل متنی و غیرمتنی نظیر ژانر، چشم‌انداز روایی نویسنده، جنسیت نویسنده و حتی موضوع متن دارد (کستمونت، ۲۰۱٤). با توجه به اینکه در پژوهش فعلی ضمایر شخصی و هر واژه‌ای که احتمال نوعی رابطه بین موضوع متن و واژه‌های مربوطه وجود داشت از فهرست واژه‌های دستوری پربسامد حذف شدند، تفاوت موضوعی و ژانری نمی‌تواند تبیین مناسبی برای نتایج این پژوهش ارائه دهد. قضاوت دربارهٔ همبستگی احتمالی برخی واژه‌های دستوری (غیر از ضمایر 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 تایید 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 دستوریِ تک‌واژه‌ای بهترین کارایی را در سطح محافظه‌کارانهٔ ٤۰۰۰ واژه و با دقت ۱۰۰٪ از خود نشان دادند (رجوع کنید به شکل ٦). دقت واژه‌های دونگاشتی و سه‌نگاشتی با کاهش حجم متون، بلافاصله با افت کارایی در تفکیک صحیح متون مواجه شد. با آنکه متونی با کمینه‌حجم ٤۰۰۰ واژه به‌عنوان آستانهٔ بهترین عملکرد الگوریتم دلتا در متون فارسی معین شد، اما به‌نظر می‌رسد این آستانه ماهیت حدی ندارد و اتفاق نظری در مورد کمینهٔ متون برای تحلیل مطمئن سبک‌شناختی وجود ندارد. برای مثال، برخی پژوهشگران ٥۰۰۰ واژه را به‌عنوان حد کمینهٔ قابل اعتماد برای متون داستانی در انگلیسی و دیگر زبان‌ها تعیین کرده‌اند (اِدِر، ۲۰۱۳)، در حالی که برخی دیگر قائل به کمینهٔ ۱۰۰۰ واژه (هولمز و همکاران، ۲۰۰۱) 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 کمینهٔ ٤۰۰۰ واژه را به‌عنوان حد مطلوب برای الگوریتم دلتا و با مدنظر قراردادن تعداد ۲۰ واژهٔ 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 موثر باشند و موجب همگرایی در تخمین کمینه واژه‌های موردنیاز برای تفکیک سبک نویسندگان شوند.

۶. سپاسگزاری

این پژوهش با حمایت مالی پژوهشگاه علوم و فناوری اطلاعات ایران انجام شده است.

۷. پی‌نوشت‌ها

1. Lorenzo Valla
2. Frontini
3. stylometry
4. forensic linguistics
5. Totty and Hardcastle
6. plagiarism
7. Stein and Meyer zu Eissen
8. Stamatatos
9. Binongo
10. Argamon & Levitan
11. Burrows
12. Zhao & Zobel
13. Segarra
14. Mosteller & Wallace
15. Holmes
16. monogram
17. bigram
18. trigram
19. style markers
20. Modaber Dabagh
21. lexical richness
22. Johansson
23. Carroll
24. Farahmandpour and Nikmehr
25. parts of speech
26. Gamon
27. Koppel and Schler
28. Whitelaw & Argamon
29. Whitelaw & Patrick
30. Systemic Functional Grammar
31. frame
32. Hedegaard & Simonsen
33. idiolect
34. Wardhaugh and Fuller
35. Jakobson

36. Barthes
37. Barlow
38. Johnson and Wright
39. www.anthropology.ir
40. www.zibakalam.com
41. www.rezadavari.ir
42. www.parsania.ir
43. Faili
44. Eder
45. Principal Component Analysis
46. Cluster Analysis
47. Adjusted Rand Index
48. Hubert and Arabie
49. Rand
50. loadings
51. 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,* 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*, 61–70.
- Burrows, J. F. (2002). "Delta: A measure of stylistic difference and a guide to likely authorship". *Literary and Linguistic Computing, 17*, 267–287.
- Carroll, D. (2008). *Psychology of Language* (5rd 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/llc/fqt066.*
- Eder, M., Kestemont, M., and Rybicki, J. (2013). "Stylometry with R: a suite of tools". In *Digital Humanities 2013: Conference Abstracts.* University of Nebraska-Lincoln, NE, pp. 487-89.
- Faili, H., Ehsan, N., Montazery, M., and Pilehvar, M. M. (2016). "Vafa spell-checker for detecting spelling, grammatical, and real-word errors of Persian

language". *Digital Scholarship in Humanities, 31* (1), *95-117.*

- Farahmandpour, Z. and Nikmehr, H. (2015). "A study on intelligent authorship methods in Persian language". *Journal of Computing and Security, 2*(1), 63-76.
- Frontini, F., Lynch, G., and Vogel, C. (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. & Simonsen, J. G. (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., Gordon, L.J., and Wilson, C. (2001). "A widow and her soldier: stylometry and the American civil war". *Literary and Linguistic Computing, 16*(4), 403-420.
- Hubert, L. and Arabie, P. (1985). "Comparing partitions". *Journal of Classification, 2*(1), 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, 61–79.
- Johnson, A. and Wright, D. (2014). "Identifying idiolect in forensic authorship attribution". *Language and Law/Linguagem e Direito*, Vol. 1(1), 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., & Schler, J. (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), 149-163.
- Mosteller, F., & Wallace, D.L. (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,* 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., Eisen, M. and Ribeiro, A. (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), 538-556.
- Stein, B., & Meyer zu Eissen, S. (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. & Hardcastle, J. P. (1987). "Forensic linguistics: the determination of authorship from habits of style". *Journal of the Forensic Science Society*, 27, 13-28.
- Wardhaugh, R. & Fuller, J. M. (2015). *An Introduction to Sociolinguistics* (7th ed.). Wiley-Blackwell.
- Whitelaw, C. & Argamon, S. (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. & Patrick, J. (2004). "Selecting systemic features for text classification". In *Proceedings of Australian Language Technology Workshop,* Sydney, Australia, pp. 93-100.
- Zhao, Y., and Zobel, J. (2005). "Effective and scalable authorship attribution using function words". In *Information Retrieval Technology* (pp. 174–189). Springer.

Function words as idiolect markers:
A corpus-based approach to authorship attribution in Farsi

Abstract

Authorship attribution is one of the key research areas within the field of forensic linguistics that has been the subject of extensive linguistic and computational studies in variety of languages. However, there is limited research on authorship attribution in Farsi. In this paper, the possibility of differentiating texts of different authors has been studied using Farsi function words. Function words can be considered powerful style markers for accommodating idiolect, since they have been shown to be processed unconsciously, have high frequency in texts, and remain independent of text topic. First, a corpus of five Iranian scholars' writings was compiled, normalized and divided into different text samples. Then 20 most frequent words were extracted from the authors' text samples and n-gram sequences (up to tri-grams) were analyzed using principal component analysis and cluster analysis functions of the Stylo R package. Findings show that function words in Farsi are capable of differentiating authors' writings with mono-gram words performing better than bi-gram and tri-grams in small size samples. It is also concluded that based on the experimental conditions of this work the minimum number of words for a text to be successfully attributed to an author is about 4000 words.

Keywords: idiolect, authorship attribution, corpus analysis, forensic linguistics, delta method