همبستگی‌های بلندبرد آماری در زبان بشر: بررسی موردی زبان پارسی

علی مهری[*]

گروه فیزیک، دانشکده علوم، دانشگاه صنعتی نوشیروانی بابل، بابل، ایران

تاریخ دریافت: 1396/01/20 ویرایش نهائی: 1396/04/04 پذیرش: 1396/10/09

چکیده

ساختار پیچیدهٔ زبان انسان، توانایی تبادل اطلاعات پیچیده را به‌ما می‌دهد. این سامانه ارتباطی از برخی قواعد آماری غیرخطی پیروی می‌کند. ما چهار ویژگی آماری زبان پارسی را بررسی می‌کنیم. یافته‌های ما با محاسبات روی شش اثر ارزشمند از اندیشمندان پارسی‌گوی به‌دست آمده‌اند. دو قانون توانی زیف و هیپس در زبان پارسی برقرار هستند و با هم یک رابطهٔ معکوس دارند. محتوای اطلاعاتی نوشتار، ناشی از چیدمان واژه‌ها توسط نویسنده، به‌کمک آنتروپی اندازه‌گیری می‌شود. از این معیار می‌توان در مرتب سازی واژه‌ها برحسب ارتباط‌شان با موضوع نوشتار بهره برد. همچنین ما بعد فرکتالی هر واژه در نوشتار را با روش جعبه‌شماری محاسبه می‌کنیم. بعد فرکتالی هر واژه، که یک مقدار مثبت کوچک‌تر یا مساوی یک است، توزیع مکانی واژه در نوشتار را نمایش می‌دهد. به‌طور کلی می‌توان ادعا کرد که زبان پارسی مانند دیگر زبان‌های بررسی شده در پژوهش‌های پیشین از قوانین آماری ذکر شده پیروی می‌کند.

کلیدواژگان: متن کاوی، همبستگی بلندبرد، قانون زیف، قانون هیپس، آنتروپی، بعد فرکتالی

مقدمه

معمولاً بسیاری از دنباله‌های نمادین طبیعی مانند زبان، موسیقی، کدهای ژنتیکی و سیگنال‌های عصبی، برای انتقال اطلاعات به‌کار می‌روند. زبان انسان یکی از نمونه‌های مهم زبان‌های طبیعی است و ظهور آن به‌عنوان یک تحول مهم در تکامل انسان به‌شمار می‌آید [1]. زبان انسان اهمیت ویژه‌ای در ارتباطات انسانی، فرهنگ و حتی هوش بشر دارد. بشر زبان را به‌عنوان ابزاری برای ارتباط و بیان ایده‌هایش به‌کار می‌گیرد. مغز گنجایش محدودی برای ذخیرهٔ واژه‌ها دارد [2]. از سویی دیگر بشر نیازمند مفاهیمی بسیار زیاد برای دستیابی به ارتباط موفق است. نیاز روزافزون برای مفاهیم جدید توسط پیچیدگی‌های نحوی و روابط معنایی میان مجموعه‌ای محدود از واژه‌ها و نمادها برطرف شده است.

زبان انسان به‌عنوان یک حامل اطلاعات بسیار پیچیده در عین انتقال اطلاعات زیاد، خطاها را نیز رفع و تعدیل می‌کند. پیچیدگی دستوری و معنایی زبان انسان باعث تمایز بین انسان و سایر گونه‌ها می‌شود. شناخت ویژگی‌های آماری و دینامیک زبان انسان در توصیف جامعیت و ممتازی آن، و تحول فرهنگ‌ها و تمدن‌ها به‌کار می‌آید [3]. بخش عظیمی از دانش بشر در بخش نوشتاری زبان گنجانده شده است. نوشتارها را می‌توان به‌صورت توالی نمادها در نظر گرفت. بخش زیادی از دانش انسان در قسمت نوشتاری زبان گنجانده شده است. چندین ویژگی کلی، بر پیچیدگی دست‌نوشته‌های انسان دلالت دارند. برخی از آن‌ها مانند قانون زیف و قانون هیپس به روابط توانی اشاره دارند

[*] نویسنده مسئول: alimehri@nit.ac.ir

که به شمار رخدادهای (بسامد) کلیدواژه‌ها در نوشتار مرتبط هستند [4و5].

بارزترین ویژگی آماری در نوشتارهای معنادار، توزیع تصادفی واژه‌های کم‌اهمیت و توزیع خوشه‌ای واژه‌های مهم است. هر نویسنده هنگام نگارش یک مطلب از واژه‌های دستوری و کم‌اهمیت به دلایل نگارشی در جای‌جای نوشته استفاده می‌کند، ولی از واژه‌های مهم و مرتبط با هدف نگارش تنها در بخش‌های خاصی از متن بهره می‌گیرد [6].

در فیزیک آماری، آنتروپی به‌عنوان یک مفهوم کلیدی برای استخراج ویژگی‌های درشت مقیاس سامانه‌های طبیعی و مصنوعی از جزئیات ریزمقیاس آنها به‌کار می‌رود. همچنین می‌توان از آن به‌عنوان معیار سنجش نظم (بی‌نظمی) سامانه‌ها در نظریه اطلاعات استفاده کرد [7]. در زبان‌شناسی محاسباتی، آنتروپی مقدار اطلاعات منتقل شده توسط پیام را اندازه‌گیری می‌کند. به‌علاوه تحلیل فرکتالی معمولاً برای اندازه‌گیری پیچیدگی یک سامانه به‌کار می‌آید. بعد فرکتالی سامانه یا طیف چند فرکتالی آن می‌تواند رفتار خود متشابه آن را معین کند. توزیع مکانی نامتعارف واژگان در نوشتار، انتقال یک مفهوم خاص را ممکن می‌سازد. رخدادهای یک کلیدواژه در داخل نوشتار، یک الگوی فرکتالی با بعد مشخص تشکیل می‌دهد [9]. به‌نظر می‌رسد که تحلیل فرکتالی، راه کار خوبی برای بررسی الگوهای فضایی در نوشتارها و یافتن ساختار خود متشابه آنها است. بعد فرکتالی واژه‌ها می‌تواند به سادگی توسط روش جعبه‌شماری محاسبه شود [10].

ما در این پژوهش قصد داریم که ویژگی‌های آماری زبان پارسی را بررسی نماییم. زبان پارسی متعلق به خانواده زبان‌های هند و اروپایی است. این زبان عمدتاً در ایران، افغانستان، تاجیکستان و برخی مناطق دیگر که از لحاظ تاریخی تحت نفوذ پارسیان بوده، استفاده می‌شود [11]. تقریباً 110 میلیون نفر در سرتاسر جهان پارسی صحبت می‌کنند. به‌علاوه پارسی، زبان رسمی در ایران، افغانستان و تاجیکستان نیز است. زبان پارسی تأثیر قابل توجهی بر زبان کشورهای همسایهٔ ایران داشته است. از نظر تاریخی سه دورهٔ شاخص برای تحولات ساختاری زبان پارسی شناخته می‌شود: دوران باستانی، میانی و معاصر.

جدول1: شش اثر برگزیده از شش نویسندهٔ بنام پارسی‌گوی که در این پژوهش ویژگی‌های آن‌ها را استخراج خواهیم کرد [8].

	کتاب	نویسنده	دوره زندگی (خورشیدی)	شمار واژگان	شمار کلیدواژه‌ها
1	شاهنامه	فردوسی	397-319	606556	14233
2	رباعیات	خیام	510-427	4722	1519
3	لیلی و مجنون	نظامی	588-520	39574	6658
4	مثنوی	مولوی	652-586	417702	29794
5	بوستان	سعدی	671-589	48866	7268
6	غزلیات	حافظ	769-706	64304	7809

نوشتارهای پارسی مدرن با استفاده از یک نوع اصلاح شدهٔ الفبای تازی نوشته می‌شوند، که دارای تلفظ متفاوت و چند حروف اضافه شده نسبت به زبان تازی

است. برخی از آثار مشهور ادبیات پارسی در جدول1 فهرست شده‌اند. ما این نوشتارهای ادبی را تحلیل خواهیم کرد و چهار ویژگی آماری آن‌ها را ارزیابی می‌کنیم. در این بررسی نوشتارهای پارسی به‌شکل دنباله‌ای از 33 نماد در نظر گرفته می‌شوند، که شامل 32 حرف الفبای پارسی استاندارد و فاصلهٔ بین واژه‌ها است. بقیه علائم نوشتاری نادیده گرفته می‌شوند.

ساختار مقاله به شرح زیر است. در بخش2، ما قانون زیف را برای آثار ادبی پارسی ذکر شده بررسی می‌کنیم. سپس در بخش3، نمای هیپس را برای آن‌ها به‌دست می‌آوریم. در بخش4، آنتروپی کلیدواژه‌های نوشتار با استفاده از توزیع مکانی آن‌ها محاسبه شده است. پس از آن، ما محتوای اطلاعاتی نوشتار را با استفاده از میانگین آنتروپی همهٔ واژگان به‌دست می‌آوریم. بعد فرکتالی واژه‌ها در بخش5 محاسبه شده است. نهایتاً در بخش6، ما خلاصه‌ای از بحث و نتیجه‌گیری آن را ارائه می‌دهیم.

قانون زیف

جرج زیف چندین قانون توانی در زمینه‌های گوناگون فعالیت‌های انسانی کشف کرد [12و13]. برجسته‌ترین آن‌ها که به شمار رخدادهای کلیدواژه‌ها در زبان انسان اشاره داشته، بیان می‌کند که اگر به کلیدواژه با بالاترین شمار رخداد (بسامد) در نوشتار رتبهٔ1 و به دومین کلیدواژه با بالاترین بسامد رتبهٔ2 اختصاص داده شود و این کار به‌ترتیب برای همهٔ کلیدواژه‌های نوشتار تکرار شود. آن‌گاه بسامد هر واژه با رتبه‌اش در نوشتار رابطهٔ توانی معکوس دارد [14]:

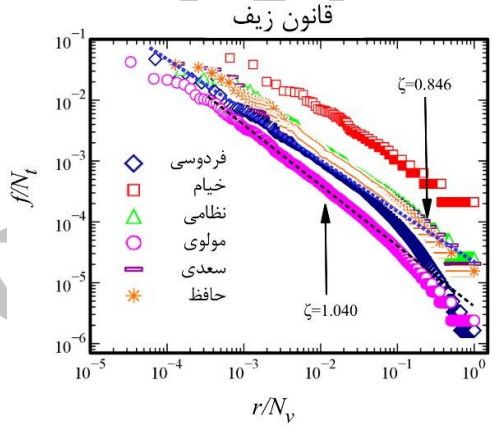
$$f \propto r^{-\zeta} \qquad 1$$

در این معادله، r رتبهٔ کلیدواژه، f بسامد آن در یک مجموعهٔ طبیعی و ζ نمای زیف است. واژگان یک نوشتار را برحسب میزان تکرارشان می‌توان به سه دسته تقسیم کرد: شمار کمی از واژه‌های پرتکرار، شمار متوسطی از واژه‌ها با میزان تکرار متوسط و شمار زیادی واژه‌های کم‌تکرار. بنا بر اصل کمترین تلاش، گوینده و شنونده هر دو تلاش می‌کنند کم‌ترین زحمت را متحمل شوند. از این رو گوینده (نویسنده) سعی دارد با بهره‌گیری از کم‌ترین شمار کلیدواژه‌ها که زیاد هم از آنها استفاده می‌کند، منظور خود را بیان کند. از سوی دیگر شنونده (خواننده) سعی دارد برای درک روشن سخنان گوینده شمار زیادی کلیدواژهٔ کم‌تکرار و نادر در ذهن داشته باشند. رقابت میان این دو فرآیند باعث می‌شود واژگان زبان دسته‌بندی بالا را داشته باشند و رابطهٔ توانی میان فراوانی و رتبهٔ واژگان نوشتار برقرار باشد. جنبه‌های گوناگون پیچیدگی یک سامانهٔ ارتباطی ممکن است به مقدار نمای زیف بستگی داشته باشد [15].

قانون زیف یکی از جذاب‌ترین و همین‌طور شناخته شده‌ترین قوانین تجربی زبان‌شناسی است، و به‌خاطر دلالت بر رفتار توانی و اثرات بلندبرد در دانش سامانه‌های پیچیده بسیار کاربرد دارد. شمار زیادی از سامانه‌هایی که در علوم اجتماعی، اقتصاد، علوم شناختی، زیست‌شناسی و فن‌آوری ارائه شده‌اند، از قانون زیف پیروی می‌کنند. همهٔ آنها از برخی واحدهای بنیادین ساخته شده‌اند، که می‌توانند در نهادهای بزرگ‌تر، گروه‌بندی شوند [16]. در مورد زبان، کلیدواژه‌ها واحدهای بنیادین هستند که با یک الگوی خاص به‌منظور بیان یک ایده توزیع می‌شوند. نوشتارها را می‌توان به حروف الفبا، واژک‌ها، واژه‌ها، عبارت‌ها و اصطلاحات، جملات، پاراگراف‌ها و غیره بخش‌بندی کرد. ولی بیشتر پژوهش‌ها در زبان‌شناسی، واحد پایه را کلیدواژه در نظر می‌گیرند [17]. در بسیاری از زبان‌ها فاصله یا علائم نگارشی مرز میان کلیدواژه‌ها را مشخص می‌کند.

قانون زیف در بسیاری از زبان‌های بشری از جمله انگلیسی، روسی، اسپانیایی و غیره با نماهای متفاوت،

وابسته به زبان، برقرار است [18]. در این پژوهش ما این قانون را برای زبان پارسی بررسی می‌کنیم. قانون زیف معمولاً به‌روش کمی بررسی می‌شود، توسط رسم لگاریتم بسامد واژه‌ها نسبت به لگاریتم رتبهٔ آن‌ها و یافتن برخی نواحی با رفتار خطی با شیب کم و بیش نزدیک به 1-. شکل1، قانون زیف برای شش کتاب ارزشمند از ادبای مشهور پارسی را نمایش می‌دهد. محور عمودی بسامد بهنجارشدهٔ کلیدواژه‌ها و محور افقی رتبهٔ بهنجارشدهٔ آن‌ها را نشان می‌دهد.



شکل1: فراوانی بهنجار شدهٔ (f/N_t) واژگان کتاب‌های برگزیده از شش ادیب برجستهٔ پارسی برحسب رتبهٔ بهنجار شدهٔ آن‌ها (r/N_v). نقطه‌چین و خط‌چین تیره به‌ترتیب نمایانگر رابطهٔ توانی با نمای ζ=0.846 و ζ=1.040 هستند.

به‌منظور بهنجار کردن بسامد، ما بسامد واژه را به‌طول نوشتار تقسیم می‌کنیم. رتبهٔ بهنجارشدهٔ هر واژه می‌تواند با تقسیم رتبه به شمار کلیدواژه‌های نوشتار به‌دست آید. برای استخراج نمای زیف، بزرگ‌ترین بخش توانی نمودار زیف را با تابع توانی برازش می‌دهیم. محدوده‌های رتبهٔ بهنجار شده (r/N_v) برای برازش جهت یافتن نمای زیف آثار بررسی‌شده در جدول2 گزارش شده‌اند. همهٔ برازش‌ها با دقت بالا ($R^2 \geq 0.99$) انجام شده‌اند. همان‌طور که در شکل1 دیده می‌شود، کوچک‌ترین نمای زیف مربوط به شاهنامهٔ فردوسی (ζ=0.846) و بزرگ‌ترین نمای زیف مربوط به مثنوی مولوی (ζ=1.040) است.

نمای زیف می‌تواند به‌عنوان معیاری از سبک نگارش، برای شناسایی نویسنده به‌کار گرفته شود [19]. رفتار پله‌ای بسامد برای رتبه‌های پایین به‌واژه‌های کم‌تکرار اشاره دارد. این رفتار پله‌ای نتیجهٔ طول محدود سامانه است. معمولاً واژه‌ها به دو گروه واژه‌های محتوایی و دستوری تقسیم می‌شوند.

واژه‌های دستوری برای ساخت ساختار دستوری به‌کار می‌روند و طبعاً بسامد آن‌ها به ساختار جمله وابسته است. از آنجایی که درصد واژه‌های دستوری نسبتاً زیاد است، قسمت بالای نمودار زیف عمدتاً شامل واژه‌های دستوری است. اما اکثر واژه‌ها در ناحیهٔ مرکزی نمودار دارای محتوای اطلاعاتی بالا هستند. بخشی از واژه‌های پرمحتوا به‌علت ارتباط تنگاتنگ با موضوع نوشتار، می‌توانند به‌عنوان واژه‌های کلیدی آن به‌شمار آیند [20]. قانون زیف در مورد واژه‌های کم‌تکرار و پرتکرار با داده‌های تجربی سازگاری ندارد. برای رفع این مشکل مندلبروت، تعمیمی از قانون زیف بیان کرد [21]:

$$f \propto (r + r_0)^{-\zeta} \qquad 2$$

در این رابطه r_0 یک پارامتر قابل تنظیم است. افزون بر این داده‌های مذکور را می‌توان با توابع پیچیده‌تری چون وایبول و یا تابع نمایی دیگرگون شده برازش داد [22].

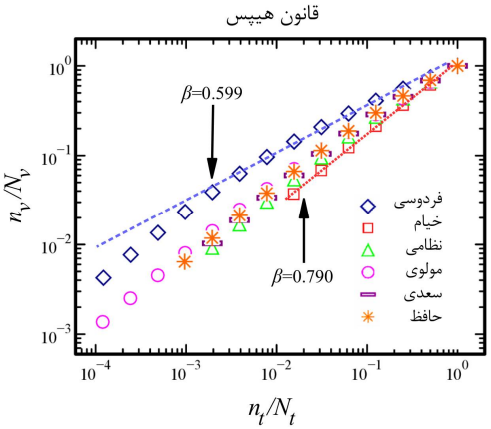
قانون هیپس

قانون هیپس یک قانون تجربی مهم در زبان‌شناسی است. این قانون بیان می‌کند که شمار کلیدواژه‌ها با افزایش طول نوشتار افزایش پیدا می‌کند [5 و 23]. به بیان دیگر این قانون تعیین می‌کند که چگونه شمار واژگان متمایز به‌کار رفته با طول نوشتار ارتباط دارد. در حالت کلی، قانون هیپس به یک رابطهٔ توانی میان عناصر متمایز در یک مجموعه و طول آن مجموعه اشاره دارد. این قانون بیان می‌کند که شمار واژگان متمایز (شمار

کلیدواژه‌ها)، N_v، در یک نوشتار یا مجموعه‌ای از نوشتارها با اندازهٔ، N_t، با معادلهٔ زیر تعیین می‌شود:

$$N_v \propto N_t^{\beta} \qquad 3$$

به $\beta \leq 1$ نمای هیپس گفته می‌شود. در اینجا ما قانون هیپس را برای آثار ادبی پارسی ذکر شده در جدول1 بررسی می‌کنیم. بدین منظور نوشتار به P بخش برابر با اندازهٔ $n_t = N_t/P$ تقسیم می‌شود. میانگین شمار کلیدواژه‌های بخش‌های گوناگون نوشتار توسط $n_v = \sum_{i=1}^{P} n_{vi}/P$ محاسبه می‌شود، که n_{vi} شمار نسبی کلیدواژه‌ها در بخش iام را مشخص می‌کند. ما این فرآیند را برای بخش‌بندی‌هایی با اندازه‌های مختلف از کتاب تکرار می‌کنیم تا نمودار هیپس را به‌دست آوریم.



شکل2. شمار بهنجار شدهٔ واژگان متفاوت (n_v/N_v) در مقابل طول بهنجارشدهٔ نوشتار (n_t/N_t) در کتاب‌های شش ادیب نامی پارسی. نقطه‌چین و خط‌چین نمایانگر برازش داده‌ها با رابطهٔ توانی هستند. افزایش توانی کلیدواژه‌ها با افزایش طول نوشتار ($\beta \leq 0.790$ ≤ 0.599) قانون هیپس را در نوشتارهای بررسی شده تأیید می‌نماید.

در شکل2، نمودار هیپس را برای شش اثر از شش نویسندهٔ مشهور پارسی نشان داده‌ایم. برای از بین بردن وابستگی طولی می‌توان n_t و n_v را به‌ترتیب بر طول کتاب و شمار کلیدواژه‌های آن تقسیم کرد. بازهٔ برازش داده‌ها (حدود پایین و بالای طول بهنجارشدهٔ نوشتار، n_t/N_t) برای یافتن نمای هیپس هم در جدول2 گزارش شده است. مقدار بالای ضرایب تعیین ($R^2 \geq 0.99$) اعتبار برازش‌ها را تأیید می‌کند. مقادیر مرزی توان هیپس در نوشتارهای بررسی شده عبارتند از β=0.599 برای شاهنامهٔ فردوسی و β=0.790 برای رباعیات خیام. نمای هیپس بزرگتر برای خیام نشان دهندهٔ گنجینهٔ واژگان غنی‌تر وی است. به نظر می‌آید، قانون هیپس از میزان حافظه و طبیعت مولد زبان انسان سرچشمه می‌گیرد. از نمای هیپس می‌توان برای شناسایی سبک نگارش نویسندگان بهره برد [24].

آنتروپی

نویسندگان واژه‌ها را در نوشتار بر اساس یک الگوی ویژه پخش می‌کنند تا پیام خود را انتقال دهند. واژه‌ها در N_t جایگاه (طول نوشتار) طبق یک روال خاص توزیع شده‌اند تا یک مفهوم خاص را به‌وجود آورند. قوانین دستوری تعیین می‌کنند که واژه‌ها در کجای جمله قرار بگیرند، و مکان فعل‌ها، اسم‌ها، قیدها و ... را مشخص می‌کنند. این قوانین همبستگی کوتاه‌برد میان واژه‌ها در جمله ایجاد می‌کنند. از سوی دیگر، یک نوشتار معنی ویژه‌ای را با استفاده از چینش معنایی واژه‌ها در داخل نوشتار می‌رساند. بنابراین همبستگی بلندبرد نیز می‌تواند در رخدادهای کلیدواژه‌ها دیده شود [25].

ما برای بررسی توزیع واژه‌ها در نوشتار از فاصلهٔ نسبی میان رخدادهای متوالی کلیدواژه‌ها استفاده می‌کنیم. فاصلهٔ نسبی میان رخدادهای متوالی یک کلیدواژه از تقسیم کردن فاصلهٔ میان دو رخداد متوالی آن در متن بر شمار کل واژه‌های متن (طول نوشتار) به‌دست می‌آید. اگر انتهای متن را به ابتدای آن بچسبانیم، هنگام محاسبهٔ فاصلهٔ آخرین رخداد یک کلیدواژه از رخداد بعد، دوباره از ابتدای نوشتار وارد آن شده به نخستین رخداد آن کلیدواژه می‌رسیم. بدین ترتیب برای هر کلیدواژه به تعداد رخدادهایش، M، در نوشتار احتمال

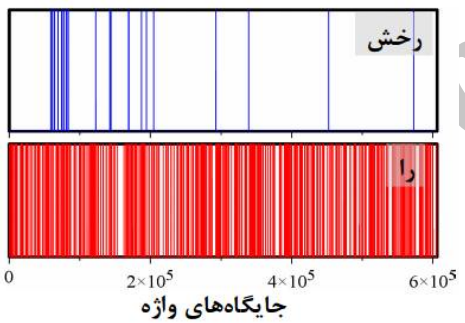
توزیع مکانی حاصل می‌شود. فرض کنیم که تمام مکان‌های رخداد یک کلیدواژهٔ به‌کار رفته در نوشتار در مجموعهٔ $T=\{t_1, t_2,..., t_M\}$ قرار گیرد. فاصلهٔ مکانی میان رخدادهای متوالی یک کلیدواژه می‌تواند با $D=\{d_1, d_2,..., d_M\}$ مشخص شود ($d_i=t_{i+1}-t_i$). به‌کمک فاصلهٔ تعریف شده، می‌توان مجموعهٔ احتمال توزیع مکانی برای هر کلیدواژه را چنین نوشت: $P=\{p_1, p_2,..., p_M\}$، که $p_i=d_i/N_t$ به‌عنوان احتمال وجود یک چرخه با طول d_i حول یک کلیدواژه در شبکهٔ رخداد واژه‌ها تفسیر می‌شود. برای نمونه در این بیت زیبا از حافظ "*صلاح کار کجا و من خراب کجا ببین تفاوت ره کز کجاست تا به کجا*"، با طول $N_t=15$ واژه و شامل $N_v=13$ کلیدواژهٔ متمایز، با چسباندن ابتدا و انتهای جمله به‌هم احتمال‌های مرتبط با توزیع مکانی کلیدواژهٔ "کجا" با توجه به فاصلهٔ نسبی سه رخداد آن از همچنین به‌دست می‌آید: $P=\{p_1=4/15, p_2=8/15, p_3=3/15\}$. اکنون آنتروپی شانون برای توزیع مکانی کلیدواژهٔ، w، در نوشتار داده شده به‌شکل زیر تعریف می‌شود:

$$H(w) = -\sum_{i=1}^{p} p_i(w)\, log(p_i(w)) \quad (4)$$

آنتروپی، میزان (نا)آگاهی دربارهٔ یک متغیر تصادفی مربوط به یک فرآیند طبیعی را به‌دست می‌دهد. مقدار صفر در آنتروپی نشان دهندهٔ نتیجهٔ قطعی برای یک فرآیند تصادفی است. از سوی دیگر، اگر تمام نتایج حاصل از یک فرآیند دارای احتمال یکسان باشند، آنتروپی مربوط به متغیر تصادفی مربوطه بیشینهٔ مقدارش را خواهد داشت. در این پژوهش، آنتروپی نشان دهندهٔ درجهٔ نظم (بی‌نظمی) در الگوی به‌کار رفتن یک کلیدواژه در نوشتار است و بیشینهٔ آن ($H_{max}(w)=log\ (M)$) زمانی رخ می‌دهد که توزیع کلیدواژه در نوشتار همگن باشد.

واژه‌های پرمحتوا که مرتبط با موضوع نوشتار هستند، در بخش‌های خاصی از نوشتار به‌منظور رساندن ایدهٔ مورد نظر، ظاهر می‌شوند ($H_{پرمحتوا} \ll log\ M$). در مقابل، واژگان کم‌اهمیت و کم‌محتوا (مانند حروف تعریف، حروف اضافه، حروف ربط و ...) با توجه به ضرورتهای دستوری، تقریباً به‌صورت همگن و یکنواخت در سرتاسر نوشتار توزیع می‌شوند ($H_{کم‌محتوا} \approx logM$).

از دیدگاه فیزیکی، در نوشته‌های معنادار طبیعی یا مصنوعی، واژه‌های پرمحتوا (مرتبط با موضوع نوشتار) یکدیگر را جذب می‌کنند و تمایل دارند که خوشه‌هایی در نواحی محدودی از سامانه (نوشته) ایجاد کنند. در حالی که واژه‌های کم‌محتوا (نامرتبط با موضوع نوشتار) همدیگر را نمی‌بینند و ارتباطی با هم ندارند، بنابراین به‌صورت تصادفی در نوشتار ظاهر می‌شوند [26]. در نتیجه توزیع مکانی تقریباً همگن در نوشتار ایجاد می‌کنند (شکل3).

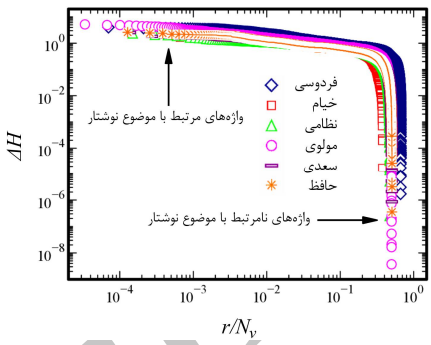


شکل3. نحوهٔ توزیع یک واژهٔ کلیدی "رخش" و یک واژهٔ کم‌محتوا "را" در شاهنامهٔ فردوسی. "رخش" در جایگاه‌های خاصی از متن و به شکل غیریکنواخت توزیع شده، ولی "را" در جای‌جای متن حضور نسبتا یکنواخت دارد.

با توجه به نکات اشاره شده، اکنون می‌توان به‌کمک آنتروپی معیاری برای جداسازی واژه‌های پرمحتوا از واژه‌های کم‌محتوا و دستوری ساخت:

$\Delta H=|H(w)-log\ (M)|$ در عمل بدین منظور تمام کلیدواژه‌های نوشتار را بر اساس مقدار ΔH شان به‌ترتیب نزولی مرتب می‌کنیم. واژه‌های پرمحتواتر در ابتدای این لیست قرار می‌گیرند [25]. اگر مقدار بهنجار

شدهٔ ΔH واژه‌ها را برحسب رتبهٔ بهنجار شدهٔ آنها رسم کنیم، از نظمی زیف‌گونه پیروی می‌کند (شکل 4).



شکل 4: ΔH بهنجار شدهٔ واژه‌های کتاب‌های بررسی شده، از شش شاعر بنام پارسی‌گوی، در مقابل رتبهٔ بهنجار شدهٔ آنها (r/N_v). تمام نمودارها رفتار زیف‌گونه با دو نمای متفاوت در دو بخش ابتدایی و انتهایی دارند. بخش نخست نمودارها با شیب ملایم مربوط به واژه‌های پرمحتوا و بخش انتهایی با شیب تند مربوط به واژه‌های کم‌محتوا است.

واضح است که بخش آغازین نمودارهای معیار بهنجار آنتروپی-رتبه با نمای زیف کوچک، مربوط به واژه‌های مهم، با محتوای اطلاعاتی بالاست که با موضوع نوشتار همبستگی زیاد دارند. دنبالهٔ نمودارها با نمای بزرگ زیف مربوط به واژه‌های کم‌اهمیت با محتوای اطلاعاتی ضعیف است که به موضوع نوشتار بستگی ندارند.

میانگین آنتروپی نوشتار (حاصل از میانگین‌گیری روی تمام کلیدواژه‌های آن) محتوای اطلاعاتی مربوط به توزیع مکانی واژه‌های آن نوشتار را نتیجه می‌دهد. میانگین آنتروپی وابسته به آثار بررسی شده در جدول 3 گزارش شده است. دیوان حافظ دارای بیش‌ترین آنتروپی ($H=0.105$) میان سایر آثار بررسی شده است. این مقدار بالای آنتروپی نشان دهندهٔ این است که حافظ کلیدواژه‌های به‌کار رفته را با الگوی همگن‌تری در دیوان اشعارش توزیع کرده است (در قیاس با دیگر آثار بررسی شده).

بعد فرکتالی

مندل‌بروت ریاضیدان لهستانی، پایه‌گذار هندسهٔ جدیدی شد که به آن هندسهٔ بدون اندازه یا هندسه فرکتالی می‌گویند [27]. در این هندسه اشکالی مورد بررسی قرار می‌گیرند، که بسیار نامنظم به‌نظر می‌رسند. اما اگر با دقت به آنها نگاه کنیم متوجه می‌شویم که تکه‌های کوچکشان کم و بیش شبیه به کل شکل هستند، به‌عبارتی جزء در این اشکال، نماینده‌ای از کل است. چنین اشکالی را خود متشابه نیز می‌نامند.

در هندسهٔ اقلیدسی خطوط و منحنی‌ها ساختارهای یک بعدی بوده، سطوح و حجم‌ها به‌ترتیب دو و سه‌بعدی به‌شمار می‌آیند. ساختارهای فرکتالی برخلاف اشکال اقلیدسی ابعاد غیرصحیح دارند (شکل 5).

شکل 5: مثلث سرپینسکی به‌عنوان یک نمونهٔ ساختار فرکتالی با بعد فرکتالی D=1.58. این ساختار با حذف مثلث میانی از مثلث‌های بزرگ‌تر در هر گام به‌دست می‌آید.

با ملاحظهٔ اشکال موجود در طبیعت، مشخص می‌شود که هندسهٔ اقلیدسی قادر به تبیین و تشریح اشکال پیچیده و ظاهراً بی‌نظم طبیعی نیست. بنابراین مجبوریم از هندسهٔ فرکتالی برای مطالعهٔ بسیاری از ساختارها مانند کوه‌ها، ابرها، کهکشان‌ها و ... بهره ببریم. فرکتال، یک ساختار ریاضی نامتعارف و تکه‌تکه است که می‌تواند به بخش‌هایی تقسیم شود که هر یک از آنها، کپی ساده شده‌ای از کل مجموعه است.

فرکتال‌ها دارای طبیعت خودتشابهی هستند و بعد فرکتالی ویژگی مهم آنهاست. بعد فرکتالی یک مجموعه با بعد هندسی فضایی که مجموعه در آن قرار گرفته تفاوت دارد [28]. این بعد وابستگی جزییات ساختار فرکتالی را به مقیاس آن آشکار می‌سازد. بعد فرکتالی می‌تواند غیرصحیح هم باشد و خودتشابهی سامانه را

نمایان می‌کند. نوشتار، یک آرایش خاص از واژه‌ها در یک آرایهٔ تک‌بعدی دارای یک معنای مشخص است. هر به‌هم‌ریختگی تصادفی واژه‌ها در نوشتار، معنای آن را به‌طور قابل ملاحظه از بین می‌برد. یعنی آرایش واژه‌ها برای رساندن مفهوم مهم است. به دیگر سخن، نظم در الگوی رخدادهای کلیدواژه‌ها در نوشتار نشانه‌ای بامعنا بودن آن است.

اگر نوشتار را به‌عنوان یک سامانه در فضای تک‌بعدی در نظر بگیریم، الگوی رخدادهای هر کلیدواژه، یک الگوی فرکتالی تشکیل می‌دهد. یکی از روش‌های محاسبهٔ بعد فرکتالی جعبه‌شماری است. به‌کمک این روش می‌توان به هر کلیدواژه در نوشتار داده شده بعد فرکتالی اختصاص داد. در روش جعبه‌شماری، نوشتار با طول N_t به N_l جعبه با طول برابر l تقسیم می‌شود: ($N_l = [N_t / l]$)، که $[x]$ بخش صحیح x را می‌دهد. شمار جعبه‌هایی که حاوی یک کلیدواژهٔ خاص هستند (جعبه‌های پر) با $n_l(w)$ مشخص می‌شود. شکل6 روش جعبه‌شماری را برای یک رباعی از خیام نمایش می‌دهد. در شکل7، ما برای کلیدواژهٔ "گفت" شمار بهنجار شدهٔ جعبه‌های پر، ($n_l(w) /N_l$)، را به‌شکل تابعی از طول بهنجار شدهٔ جعبه‌ها، l/N_t ، رسم کرده‌ایم. نتایج گزارش شده مربوط به کتاب‌های جدول1 هستند. همان‌طور که مشاهده می‌شود، تمام نمودارها، برای جعبه‌های کوچک، از رابطهٔ توانی پیروی می‌کنند. به‌منظور محاسبهٔ بعد فرکتالی کلیدواژهٔ انتخاب شده، باید شیب بخش خطی نمودار log-log جعبه‌شماری را پیدا کنیم:

بنگر ز صبا دامن گل چاک شده بلبل ز جمال گل طربناک شده
در سایه گل نشین که بسیار این گل در خاک فرو ریزد و ما خاک شده

شکل6. یک رباعی از خیام به‌طول N_t=29 واژه که بهN_7=[29/7] =4 بخش با طول l = 7 افراز شده است. کلیدواژهٔ "گل" در n_7 =3 بخش ظاهر شده است.

$$\frac{n_l(w)}{N_l} \propto \left(\frac{l}{N_t}\right)^D \qquad 5$$

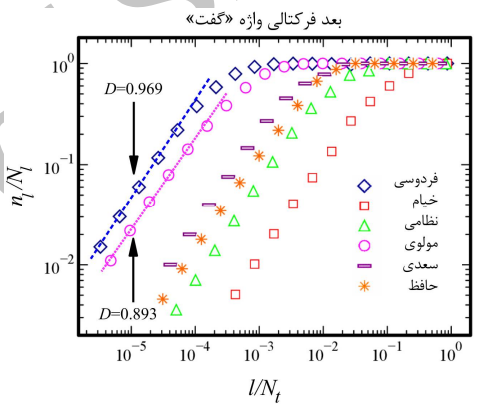
که D بعد فرکتالی کلیدواژهٔ مورد نظر است.

محدوده‌های انتخاب شده از اندازهٔ بهنجارشدهٔ جعبه‌ها (l/N_t) جهت برازش با تابع توانی به‌منظور یافتن بعد فرکتالی واژهٔ "گفت" در آثار شش نویسندهٔ پارسی‌گوی در جدول2 گزارش شده‌اند. ضریب تعیین همهٔ برازش‌ها بالاتر از نود و نه صدم است. با بهره‌گیری از این روش، بعد فرکتالی همهٔ کلیدواژه‌های نوشتار به‌صورت $0 \leq D \leq 1$ به‌دست می‌آید.

جدول2. محدوده‌های برازش داده‌ها با توابع توانی برای استخراج نمای زیف (ζ)، نمای هیپس (β) و بعد فرکتالی (D) برای کتاب‌های شش نویسندهٔ شهیر پارسی‌گوی. اعداد نوشته شده در هر بازه به‌ترتیب حد پایین و بالای کمیت نشان داده شده در محور افقی نمودار مربوط به نمای مورد نظر هستند، که برای برازش استفاده شده‌اند. ضریب تعیین بالا ($R^2 \geq 0.99$) برای همهٔ برازش‌ها، معتبر بودن آنها را تأیید می‌کند.

	کتاب	ζ	β	D
1	شاهنامه فردوسی	[3,00e-4 , 3,00e-2]	[9,76e-4 , 1]	[3,30e-6 , 5,28e-5]
2	رباعیات خیام	[6,00e-3 , 3,00e-1]	[1/55e-2 , 1]	[4,24e-4 , 2,71e-2]
3	لیلی و مجنون نظامی	[7,51e-4 , 2,00e-2]	[1,95e-3 , 1]	[5,05e-5 , 6,47e-3]
4	مثنوی مولوی	[4,03e-4 , 1,00e-1]	[1,20e-4 , 1]	[4,79e-6 , 1,53e-4]
5	بوستان سعدی	[2,06e-3 , 1,00e-1]	[1,94e-3 , 1]	[4,09e-5 , 1,31e-3]
6	غزلیات حافظ	[3,07e-3 , 2,00e-1]	[9,64e-4 , 1]	[3,11e-5 , 9,95e-4]

اگر رخدادهای یک کلیدواژه به‌طور یکنواخت در نوشتار توزیع شده باشند، همهٔ جعبه‌ها با احتمال یکسان حاوی رخدادهای آن خواهند بود. بنابراین، در این حالت شمار جعبه‌های پر، بیش‌ترین مقدار ممکن را خواهد داشت. در حالت‌های دیگر، ممکن است برخی جعبه‌ها شامل بیش از یک رخداد و برخی دیگر خالی باشند، که در این صورت شمار جعبه‌های پر کم‌تر از مقدار حدی است. از آنجایی که واژه‌های کم‌محتوا توزیع یکنواخت‌تری در نوشتار دارند، بعد فرکتالی آنها بزرگ‌تر از بعد فرکتالی واژه‌های پرمحتوا و کلیدی خواهد بود.



شکل7: نتایج جعبه‌شماری برای کلیدواژهٔ "گفت" در شش کتاب از ادیبان پارسی‌گوی. محور عمودی شمار بهنجار شدهٔ جعبه‌هایی است که شامل این کلیدواژه هستند (n_l/N_l) و محور افقی طول بهنجار شدهٔ جعبه‌هاست (l/N_t). بعد فرکتالی کلیدواژهٔ "گفت" از $D=0.893$ در مثنوی مولوی تا $D=0.969$ در شاهنامهٔ فردوسی تغییر می‌کند.

بعد فرکتالی یک نوشتار، همان میانگین مقدار بعد فرکتالی واژگان آن است. ما بعد فرکتالی کتاب‌های بررسی شده را در جدول3 گزارش کرده‌ایم. بعد فرکتالی غیرصحیح، وجود خودتشابهی آماری در زبان پارسی را تأیید می‌کند. از آنجایی که بعد فرکتالی همهٔ کتاب‌های بررسی شده کوچک‌تر از یک و نزدیک به آن است، می‌توان نتیجه گرفت که زبان بشر دارای ساختار هندسی نزدیک به ساختارهای یک بعدی مانند خط بوده و از ساختارهای صفر بعدی مانند نقطه فاصلهٔ بیشتری دارد. پیش از این پژوهش‌گران با توجه به فراوانی حروف الفبا در یک نوشتار و جایگاه آنها در میان حروف الفبا برای زبان بعد فرکتالی به‌دست آورده‌اند [29]. بعد فرکتالی، به‌عنوان معیاری از پیچیدگی نوشتار، از سری زمانی حاصل از فراوانی یا طول واژگان نوشتار نیز استخراج می‌شود [30]. به‌کمک رابطهٔ توانی میان فراوانی فاصلهٔ میان رخدادهای متوالی نت‌ها و اندازهٔ این فاصله‌ها می‌توان بعد فرکتالی برای نتهای موسیقی را نیز محاسبه نمود [31]. از بعد فرکتالی می‌توان برای استخراج واژگان کلیدی نوشتارها استفاده نمود [9].

جدول3. نمای زیف (ζ)، نمای هیپس (β)، آنتروپی (H) و بعد فرکتالی (D) برای آثار ادبی شش ادیب بنام پارسی. میانگین این کمیت‌ها برای آثار بررسی شده هم در ردیف پایانی گزارش شده است.

	کتاب	ζ	β	H	D
1	شاهنامه فردوسی	0,846	0,599	0,098	0,969
2	رباعیات خیام	0,864	0,790	0,074	0,948
3	لیلی و مجنون نظامی	0,908	0,758	0,097	0,961
4	مثنوی مولوی	1,040	0,731	0,083	0,893
5	بوستان سعدی	0,859	0,739	0,104	0,950
6	غزلیات حافظ	0,994	0,733	0,105	0,949
	میانگین	0,919	0,725	0,094	0,945

نتیجه‌گیری

به‌طور خلاصه، در این پژوهش ویژگی‌های آماری زبان پارسی را مطالعه نمودیم. ما روی چند ویژگی آماری، که نمایان‌گر همبستگی بلندبرد و روابط پیچیده در این زبان است، تمرکز کردیم. در اینجا قانون زیف، قانون هیپس، آنتروپی و ساختار فرکتالی زبان بررسی نمودیم. ما محاسبات‌مان را روی آثار فاخر شش ادیب

برجستهٔ پارسی‌گوی انجام دادیم که نتایج در جدول3 خلاصه شده است.

در زبان‌شناسی آماری، قانون زیف بیان‌گر یک رابطهٔ توانی میان شمار رخدادهای کلیدواژه‌های نوشتار و رتبهٔ آنهاست که برای سنجش ویژگی‌های بسیاری از سامانه‌های طبیعی به‌کار می‌آید. افزون بر این، قانون هیپس نحوهٔ افزایش شمار کلیدواژه‌ها را به‌عنوان تابعی از طول نوشتار تعیین می‌کند. اگر یک سامانه نمای زیف کوچکی داشته باشد، نمای هیپس آن بزرگ خواهد بود و برعکس. این رابطهٔ معکوس از نظر کیفی بیان‌گر این است که افراد با دایرهٔ لغات غنی (مقدار زیاد نمای هیپس) از کلیدواژه‌های بیشتر با بسامد کمتر (نمای زیف کوچک) استفاده می‌کنند. با توجه به این نکته، خیام دایرهٔ لغات غنی‌تری (با γ کوچک و β بزرگ) در مقایسه با سایر نویسندگان بررسی شده دارد.

ما با استفاده از معیار آنتروپی، یک مقدار کمی به اطلاعات موجود در توزیع مکانی واژه‌ها در نوشتار اختصاص دادیم. همچنین، ما فرض کردیم که توزیع واژه‌ها در نوشتار، از یک الگوی فرکتالی پیروی می‌کند، و با استفاده از روش جعبه‌شماری، بعد فرکتالی برای آنها تعریف کردیم. بعد فرکتالی معرفی شده، یک معیار مهم از توزیع مکانی واژه‌ها ارائه می‌دهد. روابط توانی، وجود همبستگی‌های بلندبرد میان رخدادهای کلیدواژه‌ها را، به‌منظور بیان ایدهٔ نویسنده، تأیید می‌کند. این پژوهش نشان می‌دهد که زبان پارسی، مانند دیگر زبان‌های بررسی شده [32]، از قواعد بنیادین آماری پیروی می‌کند.

منابع

[1] J.M. Smith, E. Száthmáry, The Major Transitions in Evolution, Oxford University Press, Oxford, (1997).

[2] S. Romaine, the Evolution of Linguistic Complexity in Pidgin and Creole Languages, in: The Evolution of Human Languages (ed. J.A. Hawkins, M. Gell-Mann), Addison Wesley, Redwood City, (1992) 213-238.

[3] M.A. Montemurro, D.H. Zanette, Complexity and Universality in the Long-Range Order of Words, *arXiv:* 1503.01129v1 (2015).

[4] G. Zipf, Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology, Addison-Wesley Press, Cambridge, (1949).

[5] H.S. Heaps, Information Retrieval: Computational and Theoretical Aspects, Academic Press, New York, (2001).

[6] M. Ortuño, P. Carpena, P. Bernaola-Galvàn, E. Muñoz, A.M. Somoza, Keyword Detection in Natural Languages and DNA, *Europhysics Letters* **57** (2002) 759-764.

[7] T. Cover, J. Thomas, Elements of Information Theory, John Wiley & Sons, New York, (1991).

[8] http://ganjoor.net/.

[9] E. Najafi, A.H. Darooneh, The Fractal Patterns of Words in a Text: A Method for Automatic Keyword Extraction, *PLoS ONE* **10** (2015) e0130617.

[10] M.F. Barnsley, Fractals Everywhere, Morgan Kaufmann, San Francisco, (1993).

[11] http://www.britannica.com/topic/ Persian-language.

[12] S.T. Piantadosi, Zipf's Word Frequency Law in Natural Language: A Critical Review and Future Directions, *Psychonomic Bulletin & Review* **21** (2014) 1112-1130.

[13] D.H. Zanette, Statistical Patterns inWritten Language, *arXiv:* 1412.3336v1 (2014).

[14] I. Moreno-Sánchez, F. Font-Clos, A. Corral, Large-Scale Analysis of Zipf's Law in English Texts, *arXiv:* 1509.04486v1 (2015).

[15] J. Baixeries, B. Elvevåg, R. Ferrer-i-Cancho, The Evolution of the Exponent of

[24] A. Mehri, A.H. Darooneh, A. Shariati, The Complex Networks Approach for Authorship Attribution of Books, *Physica A* 391 (2012) 2429-2437.

[25] A. Mehri, A.H. Darooneh, The Role of Entropy in Word Ranking, *Physica A* **390** (2011) 3157-3163.

[26] A. Mehri, M. Jamaati, H. Mehri, Word Ranking in a Single Document by Jensen-Shannon Divergence, *Physics Letters A* **379** (2015) 1627-1632.

[27] B.B. Mandelbrot, *The Fractal Geometry of Nature*, W.H. Freeman and Company, New York, (1982).

[28] K. Falconer, *Fractal Geometry*, John Wiley & Sons, Chichester, (2003).

[29] A. Eftekhari, Fractal Geometry of Texts: An Initial Application to the Works of Shakespeare, *Journal of Quantitative linguistics* **13** (2006) 177-193.

[30] M. Ausloos, Measuring Complexity with Multifractals in Texts. Translation Effects, *Chaos, Solitons & Fractals* **45** (2012) 1349-1357.

[31] K.J. Hsu, A.J. Hsu, Fractal geometry of music, *Proceeding of the National Academy of Sciences* **87** (1990) 938-941.

[32] A. Mehri, S.M. Lashkari, Power-Law Regularities in Human Language, *European Physical Journal B* **89** (2016) 241.

Zipf's Law in Language Ontogeny, *PLoS ONE* **8** (2013). e53227.

[16] F. Font-Clos, A. Corral, Log-Log Convexity of Type-Token Growth in Zipf's Systems, *Physical Review Letters* **114** (2015) 238701.

[17] A. Corral, G. Boleda, R. Ferrer-i-Cancho, Zipf's Law for Word Frequencies: Word Forms versus Lemmas in Long Texts, *PLoS ONE* **10** (2014) e0129031.

[18] A. Gelbukh, G. Sidorov, Zipf and Heaps Laws' Coefficients Depend on Language, *Lecture Notes in Computer Science* **2004** (2001) 332-335.

[19] S. Havlin, The Distance Between Zipf Plots, *Physica* A **216** (1995) 148-150.

[20] A.E. Allahverdyan, W. Deng, Q.A. Wang, Explaining Zipf's Law via Mental Lexicon, *Physical Review E* **88** (2013) 062804.

[21] B. Mandelbrot, Information Theory and Psycholinguistics: A Theory of Words Frequencies, *Readings in Mathematical Social Science* (1968) 350-368.

[22] S. Naranan, W.K. Balasubrahmanyan, Models for Power Law Relations in Linguistics and Information Science, *Journal of Quantitative Linguistics* **5** (1998) 35-61.

[23] V.V. Bochkarev, E.Y. Lerner, A.V. Shevlyakova, Deviations in the Zipf and Heaps laws in natural languages, *Journal of Physics: Conference Series* **490** (2014) 012009.

Long Range Statistical Correlations in Human Language: A Case Study of Persian Language

Ali Mehri*

Department of Physics, Noshirvani University of Technology, Babol

Received: 09.04.2017 Final revised: 25.06.2017 Accepted: 30.12.2017

Abstract

The complex structure of human language enables us to exchange very complicated information. This communication system obeys some common nonlinear statistical regularities. We investigate four important statistical features of Persian language. We perform our calculations on six masterpieces from famous Persian scholars. Zipf's law and Heaps' law, which imply well-known power-law behaviors, are established in this language, showing a qualitative inverse relation with each other. Furthermore, the informational content associated with word ordering is measured by using an entropic metric. This metric can be applied in words relevance ranking process. We also calculate fractal dimension of words in the text by using box counting method. The fractal dimension of each word, that is a positive value less than or equal to one, exhibits its spatial distribution in the text. Generally, we can claim that the Persian language follows the mentioned statistical laws, like other languages studied in previous research.

Keywords: Text mining, Long range correlation, Zipf's law, Heaps' law, Entropy, Fractal dimension

* Corresponding author: alimehri@nit.ac.ir