

## نظریه اطلاعات و تقریبهایی از زبان فارسی

مهدی ملک‌آرائی و شهریار بروجردیان

### ۱ مقدمه

در سال ۱۹۴۸ کلود شانون (Shannon) مقاله‌ای با عنوان «نظریه ریاضی مخابرات» منتشر کرد و موفق به طرح و ارائه روشی بنیادی و جدید در شاخه ریاضی نظریه مخابرات شد. نظریه مخابرات به ریاضیات (خصوصاً شاخه احتمالات) شدیداً وابسته است و با مسائل به صورتی عام و انتزاعی برخورد می‌کند: معیاری جهانی برای اندازه‌گیری اطلاعات بر مبنای عدم قطعیت، یعنی تردید کاربر در قبال ماهیت پیام، ارائه می‌دهد (بیت)، میزان تولید اطلاعات منبع پیام را بررسی می‌کند، کدگذاری بهینه را برای انتقال حداکثر اطلاعات قابل اطمینان معرفی می‌کند و جز آن.

قالب ریاضی نظریه مخابرات (اطلاعات) کاربرد آن را وسعت زیادی بخشیده است و هم‌اکنون از این نظریه در زبان‌شناسی، فیزیک، روان‌شناسی و جز آن استفاده می‌شود، اما خاستگاه نظریه مخابرات را باید در فن تلگراف و مخابرات الکتریکی جست‌وجو کرد. الفبای مورس، تلگراف ادیسون، فعالیت‌های هارتلی (Hartley) و نایکوئیست (Nyquist) و تحقیقات وینر (Wiener) هر یک به نحوی در شکل‌گیری این نظریه مؤثر بوده‌اند، هر چند نباید فراموش کنیم که نظریه مخابرات قبل از هر چیز «نظریه ریاضی مخابرات» است و مفاهیم آن در قالب فرمولهای ریاضی ارائه شده است.

زبان مجموعه‌ای از نشانه‌های آوایی است که برای ایجاد ارتباط میان انسانها به کار می‌رود و به همین دلیل ساختار آن در نظریهٔ مخابرات مورد توجه و بررسی قرار گرفته است. البته این برداشت از ساختار زبان با برداشت زبان‌شناسها تفاوتی دارد. در این مقاله براساس بررسیهای آماری انجام شده در زبان فارسی، تقریبهای مختلفی از آن معرفی و ارائه خواهد شد.

## ۲ مدل ریاضی

علم و ریاضیات دوشادوش هم رشد کرده‌اند و هر یک پیشرفت دیگری را سرعت بخشیده است. نظریه‌های ریاضی برای توضیح و پیش‌بینی رویدادهای اطراف ما مدل‌های ساده‌شده‌ای ارائه می‌دهند که دقت و محدودهٔ کاربرد متفاوتی دارند. این مدلها عموماً به دو دسته کلی ارادی و احتمالی تقسیم می‌شوند و در بررسی فعالیتها و رفتارهای انسان (چه فردی و چه گروهی) - مانند زبان که یکی از وسایل ارتباط میان بشر است - عموماً از مدل‌های احتمالی استفاده می‌شود.

برای مثال شمارش حروف زبان انگلیسی و فارسی و محاسبهٔ فراوانی آنها مشخص می‌کند که E به‌عنوان پرتکرارترین حروف حدود ۱۳٪ از کل حروف زبان انگلیسی و «الف» حدود ۱۱٪ از حروف زبان فارسی را تشکیل می‌دهند و برخی صامت‌ها مانند R و N و T در انگلیسی و «ر»، «م»، «ن» در زبان فارسی، از فراوانی نسبی زیادی برخوردارند. البته اگر پیام خیلی کوتاه باشد، یا بسته به موضوع متن و سبک نگارش آن، ممکن است فراوانی نسبی حروف در متون مختلف تفاوتی داشته باشد،<sup>۱</sup> ولی بررسی متونی که به اندازهٔ کافی طولانی باشد ما را به نتایج تقریباً مشابهی می‌رساند.

اما آیا دانستن این فراوانیها محدودیتی در نوشتن ایجاد می‌کند؟ به عبارت دیگر آیا ملزم به استفاده از ۱۳ حرف E در هر ۱۰۰ حرف هستیم؟ در حقیقت، کم نبوده‌اند کسانی که از این الگوها و قالبها سر باز زده‌اند. برای مثال، ارنست وینسنت رایت در سال ۱۹۳۹ رمانی ۲۶۷ صفحه‌ای با نام Gadsby منتشر کرد که در آن (حدود ۵۰۰۰۰ کلمه) از حروف E استفاده نشده است! این چند سطر از این رمان است:

Upon this basis I am going to show you how a bunch of bright young folks did

۱. برای مثال، تکرار صامت‌ها و مصوت‌های مشابه می‌تواند الفاکتندهٔ معنی یا ایجادکنندهٔ نوعی موسیقی باشد.

find a champion; a man with boys and girls of his own; a man of so dominating<sup>1</sup>

برای مدل‌سازی زبان آن را منبعی در نظر می‌گیریم که نشانه‌های معینی را صادر می‌کند. هر چه ویژگی‌های بیشتری از زبان در منبع منظور شود، مدل بهتری خواهیم داشت. این نشانه‌ها برای زبان فارسی جای خالی، ۳۲ حرف الفبا، «آ»، «ا»، «ؤ»، «ئ»، همزه، تشدید و تنوین و برای انگلیسی جای خالی و ۲۶ حرف الفباست.

الف) به‌عنوان ساده‌ترین مدل زبان را منبعی در نظر می‌گیریم که نشانه‌هایی را مستقل از هم و با احتمال مساوی صادر می‌کند (شنون آن را تقریب مرتبه صفرم حروف نامیده است)؛ رشته‌های زیر نمونه رشته‌های تولیدی چنین منبعی است.

و پلظ اگاآلاشخصخزذفیطش طظمجملدز

xfoml rxkhrjffjuj zlpwefwkeyj

همان‌طور که مشخص است در این رشته‌ها تعداد «ط»، «ظ» و Z و W در مقایسه با متون واقعی بسیار زیاد است، در حالی که تعداد «الف»، E و جای خالی به اندازه کافی نیست.

ب) در مرحله بعد نشانه‌ها را مستقل اما بر مبنای فراوانی‌شان در زبان تولید می‌کنیم (شنون آن را تقریب مرتبه اول حروف نامیده است). بدین منظور ابتدا باید احتمال صدور نشانه‌های مختلف زبان محاسبه شود. در پیوست (۱) احتمال صدور نشانه‌های فارسی آمده است. از این پیوست اطلاعات سودمند بسیاری به دست می‌آید، مثلاً اینکه پنج نشانه پرمسامد فارسی به ترتیب عبارت‌اند از: جای خالی، «الف»، «ی»، «ر»، «د» یا اینکه با توجه به احتمال صدور جای خالی (حدود ۲۰٪) متوسط طول کلمه‌های فارسی ۴ حرف است.<sup>۲</sup> رشته‌های زیر نمونه رشته‌های تولیدی است:

و ضهییگ چم عگتوا اشقرخرا ادا لا شهبوا یج

ocro hli rgwr nmieewis eu ll nbnese bya

اما در زبان می‌توان پیش‌بینی کرد که چه حروفی به دنبال چه حروفی می‌آیند و بسامد آنها چه قدر است. مثلاً در خط فارسی آمدن جای خالی پس از همزه (به‌عنوان یک نشانه مستقل «ء»، نه روی کرسی «ئ» و ...) حتمی است یا پس از «م» احتمال آمدن «ی» از دیگر نشانه‌ها بیشتر است و در انگلیسی آمدن U به دنبال Q صددرصد قابل پیش‌بینی است یا پس از H، T، بیش از دیگر نشانه‌ها آمده است.

۱. برای مثال، گذشته افعال با قاعده و اعداد ۲۹-۷ و بسیاری اعداد دیگر در این رمان به‌کار نرفته است.

۲. اگر جای خالی بعد از هر کلمه را هم جزئی از آن محسوب کنیم، طول متوسط پنج نشانه خواهد بود.

ج) در مرحله بعد نشانه‌ها را بر مبنای فراوانی دو حرفی‌ها (digraph) تولید می‌کنیم (شنون آن‌را تقریب مرتبه دوم حروف نامیده است) به این ترتیب که با در نظر گرفتن حروف مجاور در نوشته‌ها و شمارش دو حرفی‌ها نتیجه می‌گیریم که مثلاً در یک متن ۱۰۰۱ حرفی (با احتساب جای خالی) - که شامل ۱۰۰۰ دو حرفی است - ده دو حرفی «م + ی» است و لذا احتمال آن که یک دو حرفی «م+ی» باشد ۰.۱٪ است. با توجه به اینکه احتمال آن که یک دو حرفی با «م» شروع شود همان احتمال صدور «م» است (۰.۴۴٪ از پیوست (۱))، احتمال آن که دو حرفی‌ای که با «م» شروع شده، «م+ی» باشد تقریباً ۰.۲٪ (۰.۴۴٪ ÷ ۰.۱٪) است. در پیوست (۲) هر عدد نشان‌دهنده احتمال صدور نشانه مربوط به سطر، پس از نشانه مربوط به ستون خود است به درصد. از این پیوست هم، اطلاعات سودمندی قابل استخراج است، مثلاً اینکه کلمه‌های فارسی بیشتر با چه حروفی آغاز می‌شود، چرا که ستون اول جدول احتمال صدور نشانه‌ها پس از جای خالی را نشان می‌دهد؛ یا اینکه بیشتر به چه حروفی ختم می‌شود،<sup>۲</sup> که سطر اول نشان‌دهنده آن است. رشته‌های زیر نمونه رشته‌های تولیدی است:

آز هیدهء رایت یار طهوجنر پذیرق مست درا

on ie antsoutinys are t intcore st be s deamy

د) به طریقی مشابه، با بررسی سه حرفی‌ها (trigraph) و محاسبه فراوانی آنها به تقریب مرتبه سوم می‌رسیم و مرتبه تقریب را می‌توان به همین ترتیب بالاتر برد. رشته زیر نمونه رشته مرتبه سوم است:

in no ist lat whey cractict froure birs

### ۳ قانون زیپف (Zipf)

بر اساس این قانون اگر در یک متن کلمه‌ها را به ترتیب فراوانی‌شان رتبه‌بندی کنیم - به پر بسامدترین کلمه رتبه (۱)، به بعدی (۲) و به همین ترتیب رتبه بدهیم - در آن صورت حاصل ضرب رتبه در فراوانی تقریباً عددی ثابت خواهد بود و احتمال صدور  $r$  امین کلمه، به طور تقریبی برابر است با:

$$P(r) = \frac{1}{r^2}$$

۱. ممکن است برخی از اعداد جدول از دقت کافی برخوردار نباشد که یکی از علتهای آن فاصله‌گذارهای متون

بررسی شده است مثلاً اگر «او را» بدون فاصله نوشته شده باشد «و + ر» به عنوان یک جزء شمارش شده است.

۲. البته با در نظر گرفتن تکرار کلمه‌ها در متن

۳. این رابطه تقریبی است. صورت دقیقتر آن Modified Zipf's law نام دارد.

به این دو مثال<sup>۱</sup> توجه کنید:

Rank	Word	Freq
1	the	68315
2	of	35716
3	and	27856
262	public	345
263	within	344
2731	steadily	37
2732	necessity	37

رتبه	کلمه	فراوانی
۳	از	۲۲۴۷
۵۷	خود	۲۳۰
۵۸	اگر	۲۱۳
۵۲۱	نشان	۲۵
۵۲۲	نگاه	۲۵
۱۰۲۷	نعمت	۱۲
۱۰۲۸	وجود	۱۲

#### ۴ نتیجه‌گیری

با بررسی تقریبها متوجه می‌شویم که با بالا رفتن مرتبه تقریب، رشته‌ها به متن واقعی شبیه‌تر می‌شود. مثلاً در تقریب مرتبه اول کلمه‌های ocro و «یج» شبیه کلمه‌های<sup>۲</sup> انگلیسی و فارسی است و در تقریب مرتبه دوم بسیاری از کلمه‌ها از جمله on و «یار» در زبان وجود دارد. البته می‌توان رشته‌هایی را براساس فراوانی کلمه‌ها هم تولید کرد. این تقریبها در بررسی حشو زبان، و اطلاعات پیوسته‌ها در کدبندی زبان، رمزگشایی، شناسایی زبان (Language identification) و جز آن به‌کار می‌آید. اطلاعات جداول از شمارش ۳۸۲۵۹۴۶ حرف (در ۹۶۳۹۷۹ کلمه) و براساس رسم‌الخط مرکز نشر دانشگاهی استخراج شده است.

#### سپاسگزاری

از دکتر سعید نادر اصفهانی، دکتر مصطفی عاصی، مهندس علی کافی و از مسئولین T<sub>E</sub>X-پارک مرکز نشر دانشگاهی (مهندس سجادی، صدیقه مسعودی و آزاده ابری) تشکر می‌شود.

#### مراجع

J. R. Pierce. *Symbols, signals and noise*. 1st ed. 1962.

- کارلسون، ا. ب. سیستمهای مخابراتی. مرکز نشر دانشگاهی، ۱۳۷۳.  
 باطنی، محمدرضا. مسائل زبان‌شناسی نوین. انتشارات آگاه، ۱۳۷۴.  
 نجفی، ابوالحسن. مبانی زبان‌شناسی. انتشارات نیلوفر، ۱۳۷۶.

۱. مثال اول از کتاب اسرار التوحید است و اطلاعات آن از پژوهشگاه علوم انسانی گرفته شده و مثال دوم از اینترنت: [Http://linguistlist.org/~ask-ling/msg1076.html](http://linguistlist.org/~ask-ling/msg1076.html)

۲. البته منظور از کلمه اینجا کلمه بامعنی است.

پیوست ۱. جدول احتمال مرتبه اول

حروف چهار شکلی      حروف دو شکلی      حروف یک شکلی

حرف	ابتدا				انتهای		حرف
	میان	انتهای	چسبان	نچسب	کل	کل	
ی	۲٫۲۸	۲٫۴۸	۱٫۹۴	۱٫۳	۸٫۰۰	۱۰٫۸۲	۲۰٫۷۴
ن	۲٫۱	۱٫۴۲	۰٫۸۳	۰٫۹۹	۵٫۳۴	۶٫۳۴	۰٫۵۶
ه	۱٫۲۷	۱٫۰۱	۲٫۱۵	۰٫۷۲	۵٫۱۵	۵٫۳۶	۰٫۱
م	۲٫۸۲	۰٫۸۱	۰٫۵۳	۰٫۲۴	۴٫۴	۴٫۴۹	۰٫۰۹
ت	۱٫۲۸	۱٫۳	۱٫۱۶	۰٫۳	۴٫۰۴	۱٫۵۱	
ب	۲٫۹۵	۰٫۴۷	۰٫۱۵	۰٫۱۶	۳٫۷۳	۰٫۶۰	
ک	۱٫۵۹	۰٫۵۹	۰٫۲۷	۰٫۰۵	۲٫۵۰	۰٫۱۷	
س	۱٫۵۳	۰٫۶۷	۰٫۱	۰٫۰۵	۲٫۳۵	۰٫۰۹	
ل	۰٫۸	۰٫۷	۰٫۴۲	۰٫۳۲	۲٫۲۴	۰٫۰۷	
ش	۱٫۱	۰٫۴۲	۰٫۲۴	۰٫۰۸	۱٫۸۴	۰٫۰۲	
ع	۰٫۴۵	۰٫۵	۰٫۱	۰٫۰۸	۱٫۱۳		
ف	۰٫۵۴	۰٫۳۴	۰٫۱۱	۰٫۰۸	۱٫۰۷		
ق	۰٫۴۴	۰٫۴	۰٫۰۹	۰٫۰۵	۰٫۹۸		
گ	۰٫۶۲	۰٫۲۷	۰٫۰۴	۰٫۰۲	۰٫۹۵		
ج	۰٫۵۳	۰٫۲۳	۰٫۰۲	۰٫۰۵	۰٫۸۳		
خ	۰٫۵۵	۰٫۲۴	۰٫۰۱	۰٫۰۱	۰٫۸۱		
ح	۰٫۴۱	۰٫۲۳	۰٫۰۵	۰٫۰۳	۰٫۷۲		
ط	۰٫۲۴	۰٫۲۷	۰٫۰۷	۰٫۰۴	۰٫۶۲		
ص	۰٫۳۲	۰٫۱۹	۰٫۰۳	۰٫۰۳	۰٫۵۷		
پ	۰٫۴۴	۰٫۰۶	۰٫۰۱	۰٫۰۱	۰٫۵۲		
چ	۰٫۲۸	۰٫۰۵	۰٫۰۲	۰٫۰۱	۰٫۳۶		
ض	۰٫۱۲	۰٫۰۹	۰٫۰۲	۰٫۰۵	۰٫۲۸		
ث	۰٫۰۹	۰٫۰۷	۰٫۰۳	۰٫۰	۰٫۲۰		
غ	۰٫۰۸	۰٫۰۷	۰٫۰	۰٫۰	۰٫۱۶		
ظ	۰٫۰۳	۰٫۱۱	۰٫۰	۰٫۰	۰٫۱۴		
ئ	۰٫۰۵	۰٫۰۳	۰٫۰	۰٫۰	۰٫۰۸		

بیوست ۲. جدول احتمالی شرطی (مرتبہ دوم)

د	خ	ح	ج	ج	ت	ت	ب	ب	ا	ا	ا	ا	Space	Space
۲۹۵۳	۲۵۱	۱۰۱۸	۴۲۴	۷۸۱	۱۶۵	۳۵۰۳	۴۱	۸۱۳	۱۳۱۳	۴۲۲۱	۶۰۶	—	۲۷۷	Space
۰۰۶	۰۰۰	۰۰۰	۰۰۰	۰۰۱	۰۰۰	۰۰۳	۰۰۰	۰۰۱	۰۰۱	۰۰۰	۰۰۰	۰۰۰	۰۰۰	آ
۰۰۳	۰۰۰	۰۰۲	۰۰۰	۰۰۱	۰۰۰	۰۵۹	۰۰۰	۰۰۰	۰۰۰	۰۰۰	۰۰۰	۰۰۰	۰۰۰	آ
۱۲۷۲	۱۴۴۹	۱۸۶۸	۲۹۶	۱۷۳۴	۳۴۲۲	۸۴۱	۱۸۲۶	۱۹۸۱	۰۰۲	۰۹	۰۰۰	۱۳۵۱	۱۳۵۱	ا
۰۱۸	۰۲۳	۰۶۴	۰۱۶	۲۱۲	۹۹۸	۱۴۴	۰۰۰	۰۴	۳۶۸	۰۰۰	۶۰۷	۱۲۱۶	۱۲۱۶	ب
۰۰۳	۰۰۰	۰۰۰	۰۲۲	۰۰۲	۰۰۰	۰۰۹	۰۰۱	۰۱۱	۰۱۷	۰۱۷	۰۲۳	۱۹۲	۱۹۲	ب
۰۰۳	۱۶۰۹	۷۶۲	۰۰۹	۰۶	۰۰۰	۰۱۱	۱۲۵	۳۰۸	۲۵۲	۲۲۷	۰۶۱	۴۲۹	۴۲۹	ت
۰۰۳	۰۰۰	۱۰۸	۰۰۰	۰۰۰	۰۰۰	۰۰۷	۰۰۰	۰۰۰	۰۲۷	۱۴۶۸	۰۵۸	۰۲۱	۰۲۱	ت
۰۰۵	۰۰۴	۰۹۷	۰۰۲	۰۰۰	۰۰۰	۰۶۹	۰۰۰	۰۰۸	۰۴۳	۰۰۶	۰۱	۱۵۴	۱۵۴	ج
۰۰۳	۰۰۴	۰۰۰	۰۴۱	۰۰۱	۰۰۰	۰۰۲	۰۰۴	۰۰۳	۰۰۳	۰۰۰	۰۰۴	۱۲۷	۱۲۷	ج
۰۰۳	۰۰۰	۰۰۱	۰۰۰	۰۰۵	۰۰۰	۱۱۲	۰۰۰	۰۴۸	۰۸۵	۰۰۰	۰۰۰	۵	۵	ح
۰۰۹	۰۰۰	۰۰۱	۰۰۱	۰۰۱	۰۰۰	۰۹۸	۰۳۹	۲۰۸	۱۰۲	۲۲۳	۰۸	۱۸۳	۱۸۳	خ
۰۶۶	۲۸۸	۷۴۸	۰۰۷	۴۷۶	۰۰۰	۰۸۷	۲۵۲	۲۵۹	۳۸۹	۰۰۶	۰۲۶	۷۸۱	۷۸۱	د
۰۰۰	۰۰۵	۰۶۲	۰۰۰	۱۳۹	۰۰۰	۰۰۷	۳۹۳	۲۲۲	۰۰۳	۰۰۳	۱۰۹	۰۲۲	۰۲۲	ذ
۱۶۱۴	۳۹۴	۵۹۵	۹۹۹	۶۱۲	۱۸۳۵	۸۶۱	۱۶۵۴	۱۵۲۸	۸۵۸	۰۰۰	۲۷۱	۳۷۸	۳۷۸	ر
۰۱۴	۰۶۲	۰۰۲	۰۰۳	۵۵۷	۰۰۰	۰۲۲	۰۴۶	۰۶۹	۶۸۹	۰۰۰	۳۴۱	۱۲۴	۱۲۴	ز
۰۰۰	۰۰۰	۰۰۰	۰۰۰	۰۰۰	۰۰۰	۰۰۰	۰۵۳	۰۰۰	۰۱۸	۰۰۰	۰۰۱	۰۰۲	۰۰۲	ژ
۱۸۳	۱۳۳	۶۶۵	۰۷۶	۳۹۹	۰۰۰	۰۳	۸۱۸	۱۷۸	۷۳۲	۵۴۲	۲۰۶	۲۱۷	۲۱۷	س
۰۳۳	۷۵۳	۰۳۷	۳۹۳	۰۱۷	۰۰۰	۱۴۲	۰۷۲	۰۶۳	۲۴۴	۰۰۶	۱۴۳	۲۷۳	۲۷۳	ش
۰۰۰	۴۹	۱۱۹	۰۰۰	۰۰۰	۰۰۰	۱۷۳	۰۰۰	۰۱۸	۱۱۱	۰۰۰	۰۰۰	۰۷۷	۰۷۷	ص









