حشو در زبان با رویکرد نظریهٔ اطلاعات

اردوان میرزایی*

چکیده: این مقاله ابتدا به نمونهٔ مکالمهٔ دو نفر پرداخته و برخی از مشکلاتی را که از این گفتگو، بین آن دو ظهور کرده بازگو می‌کند. این مشکلات به دو بخش شکلی و ماهوی تقسیم می‌شود. بخش مشکلات شکلی (ظاهری)به دو قسمت فیزیکی بدن و قسمت زبان و ساختار آن می‌پردازد. و در اینجاست که مشکل انتقال اطلاعات در زبان مطرح می‌شود. به لحاظ ارتباط تنگاتنگی که بین نظریهٔ اطلاعات (در ریاضی) و نظریه بهینگی (در زبان) دیده می‌شود زبان به مجموعه‌ای از محدودیتها تعریف شده و نقش محدودیتها در زبان به اختصار آمده است. سپس به ارائهٔ مختصر تئوری اطلاعات[1] اشاره شده و کانال انتقال اطلاعات از مبدأ به مقصد بیان شده است. در این مقاله دو مفهوم آنتروپی[2] و حشو[3] مطرح می‌شود. محاسبهٔ آنتروپی و حشو در زبان انگلیسی و مثالی از یک کلمه انگلیسی آورده شده است. پس از آن به محاسبهٔ آنتروپی و حشو در زبان فارسی پرداخته می‌شود و این نتیجه به دست می‌آید که آنتروپی زبان فارسی روی حرف اول کلمه ۵ بیت/ نماد و زوائد آن حدود ۸۰ در صد است.

کلیدواژه: تئوری اطلاعات، نظریهٔ بهینگی، آنتروپی، حشو، ساختار زبان، دستور زبان، بیت(Bit).

مقدمه

این مقاله به بررسی مختصر یکی از مشکلات زبان با استفاده از ریاضی پرداخته و کاربرد ریاضی (نظریهٔ اطلاعات) در زبان را نشان می‌دهد. اهمیت ریاضی و کاربرد آن در علوم مختلف بر کسی پوشیده نیست. دانشمندان به این نتیجه رسیده‌اند که هرگاه علمی از ریاضی به صورت بهتر و کاراتر استفاده کند پیشرفتش اصولیتر، مستحکمتر و سریعتر خواهد بود. اصولاً، چنانچه بتوان مقادیر کیفی را به کمی تبدیل کرد، در این صورت، علم ریاضی، که عموماً بر مقادیر کمی حکمفرمایی می‌کند، به کار می‌آید و با نظریه‌های مستحکم خود نتایج درخور توجه، گاه غیرعادی، به دست می‌دهد. کاربرد ریاضی نه تنها فی‌نفسه برای خود ریاضی مفید و قابل توجه است، بلکه‌برای علوم دیگر نیز راهگشا خواهد بود. از این‌رو،

*عضو هیئت‌علمی دانشگاه پیام‌نور، سازمان مرکزی.

1. Information Theory 2. Antropy
3. Redundancy

لازم است مختصری از نظریه اطلاعات گفته شود.

ابتدا دو نفر را در نظر بگیرید که می‌خواهند با هم گفتگو کنند. برای انتقال مطالب از یک ذهن به ذهن دیگر ناچار به استفاده از زبان (گفتاری ـ نوشتاری) هستیم. اطلاعاتی در ذهن وجود دارد، ابتدا این اطلاعات به صورتی خاص (ذاتی و اکتسابی) در ذهن (گوینده) مرتب می‌شود، سپس با ساختار زبان به آن شکل داده می‌شود و با دستور مغز به سیستم تکلم ادا می‌شود (با فرض اینکه عیناً همین صدا به شنونده منتقل می‌شود و مشکلی در تکلم گوینده و شنوایی شنونده وجود نداشته باشد بحث را ادامه می‌دهیم). صدا به شنونده می‌رسد و او با گوش و اعصاب که به مغز منتقل می‌شود، می‌شنود. اکنون شنونده مطالبی را در ذهن خود دارد. آیا این مطالب دقیقاً منظور گوینده است؟ آیا گوینده توانسته مطلب خود را طوری تنظیم و ادا نماید که شنونده دقیقاً همان منظور را دریابد؟ آیا گوش شنونده همان چیزی را که گوینده گفته به ذهن انتقال داده؟ آیا ذهن شنونده دقیقاً همان منظور ذهن گوینده را درک کرده است؟ آیا دقیقاً همان اطلاعاتی که در ذهن گوینده بوده، همان اطلاعات را ذهن شنونده در خود دارد؟ اکنون فرض کنید شنونده مطالب را درک کرده و در صدد پاسخگویی است این بحث در واقع به مباحث زیادی کشیده می‌شود. در این انتقال اطلاعات ذهن گوینده در تنظیم مطالب چقدر نقش دارد؟ نقش گفتار گوینده چقدر است؟ نقش ارتباط (فاصله ـ نوع ارتباط، یعنی تلفن، حضوری و...) چقدر است؟ گوش شنونده چقدر نقش دارد؟ گیرایی مطلب توسط ذهن شنونده چه میزان نقش دارد؟ شبکهٔ اعصاب انتقال اطلاعات از ذهن گوینده به زبان او و شبکهٔ اعصاب انتقال اطلاعات از گوش شنونده به ذهن او چقدر است؟ این انتقال اطلاعات از مبدأ گوینده به مقصد شنونده را به دو بخش می‌کنیم. بخش اول قسمت ماهوی و بخش دوم قسمت شکلی (یا ظاهری) است. در قسمت ماهوی چگونگی تنظیم مطالب در ذهن گوینده (شنونده) و موضوعاتی از این دست مطرح است. در این مقاله به بخش اول (قسمت ماهوی) وارد نمی‌شویم و به قسمتی از بخش دوم (قسمت شکلی یا ظاهری) می‌پردازیم. بخش ظاهری، یا بهتر بگوییم شکلی، خود به دو قسمت می‌شود. قسمت اول به موارد فیزیکی، مانند مکان مکالمه، زمان مکالمه، عمل نمودن صحیح یا غلط دستگاه عصبی طرفین و اعضایی مانند دهان و گوش، که این انتقال را انجام می‌دهند، مربوط می‌شود. در این مقاله به قسمت دوم، که موارد غیر فیزیکی یعنی زبان و ساختار زبان است، توجه می‌کنیم.

زبان به منزلهٔ مجموعه‌ای از قواعد [۴]

در انقلاب‌دوم در تاریخ زبانشناسی، که همان پیدایش زبانشناسی نوین، یعنی زبانشناسی سوسوری و ساخت‌گرایی اروپایی است، زبان (language) به عنوان نظامی از نشانه‌ها (signs) تعریف شد. به بیانی دیگر، در این رویکرد بین نشانه‌ها شبکه‌ای از روابط متصور بود و ارزش هر نشانه در این شبکه تعیین می‌شد. اگرچه واج (phoneme)به منزلهٔ آوای ممیز را ابتدا بودوئن دوکورتنی،[۵] زبانشناس لهستانی تبار، مقیم روسیه تزاری، مطرح کرد؛ تروبتسکوی[۶] با بهره‌گیری از مفاهیم ارزش مطرح‌شده از سوی سوسور اقدام به معرفی انواع تقابلها بین واجها کرد و بدین‌ترتیب واج‌شناسی (phonology) به مثابهٔ شاخه‌ای از زبانشناسی در مکتب پراگ پا به عرصهٔ وجود گذاشت. در واج‌شناسی مکتب پراگ تعیین مشخصه‌های آوایی (phonetic features) تشکیل‌دهندهٔ واجها توسط تروبتسکوی و یاکوبسن[۷] پیگیری‌شد و در همین مقطع زمانی تعریف

۴. نک: دبیرمقدم، ۱۳۸۳.

5. Baudouin de Courtenay
6. Terobestscoy
7. Yacobscon

واج و روالهـای کشـف واج در کـانون توجـه زبانشناسـان ساخت‌گرای آمریکا قرار داشت (Twaddell, 1966: 55-80) واج‌شناسی زایشی معیار (Chomsky& Hall, 1968) بازگشتی بـــود بـه دستور پانینی (قرن ٤ ق . م) و مشخصاً، کشف قـاعده‌هـای واجی زبانها و نحوهٔ صورتبندی و فرموله نمـودن آنها مورد توجه ویژهٔ واج‌شناسان زایشی قرار گرفت.

زبان به منزلهٔ مجموعه‌ای از محدودیتها[۸]

جـان رابـرت راس (1986) در رسالهٔ دکترای خود، که در ۱۹۸٦ بـه چـاپ رسـید، بحثی در باب محدودیتها (constraints) را بـر عملکـرد گشتارها مطرح می‌کند. ایـن محدودیتها که گاهی شرط (condition) و گاهی اصـل(principle) نـیز خوانـده شـده اسـت، همـان هماهنگـی زبانی‌اند و از این‌رو، متعلق به دستور همگانی (universal grammar) می‌باشند. کودکان این محدودیتها و شـرطها و اصـلها را نمـی‌آموزنـد؛ بلکه آنها بخشی از دانش زیسـتی و فطـری آنـان را تشـکیل مـی‌دهند. تعدادی از این محدودیتها در اشر(Ross, 1986) و چامسکی (1973) دیده می‌شوند(دبیرمقدم، ۱۳۸۳).

ایـن محدودیتها عملکرد گشتارها و مشخصاً گشتارهای حرکتـی، مانند مجهول‌سازی، مبتداسازی، حرکت پرسشواژه و حرکت بند، موصولی‌سازی و ارتقاء را محدود می‌کنند. در نظـریهٔ حاکمیـت و مـرجع‌گزینـی (Chomsky,1981) گـام مهمـی به سوی کمینه‌گرایی (minimalism) برداشته شد و آن ایـن بـود کـه تمـام گشـتارهای حرکتـی موارد خاصـی از یـک فرایـند کلی به نام گشتار حرکت آلفا تلقـی شـدند. تحـول دیگر این بود که در بخش نحو دیگر سخنی از گشتارهای اختیاری و اجباری در میان نبود، بلکه همهٔ گشتارهای این بخش اجباری بودند.

در ایـن نظـریه زبـان توسـط محدودیـتها کنترل مـی‌شود و تفاوت زبانها در محدودیتهایی است که بر ساختار زبان تأثیر می‌گذارند. از این جهت، چگونگی گویش در زبانهای مختلف مربوط به نوع و چگونگی محدودیـتهای حـاکم بـرآن اسـت. در نظـریهٔ بهینگی (optimality theory) محدودیتها دسته‌بندی واولویت‌بندی شـده‌انـد و هـر کـدام اولویت داشته باشد در زبان نقش اساسـیتری بـازی می‌کند. اولویتها در زبانهای مختلف مـتفاوت است؛ بنابراین ساختارهای زبانها تا حدودی مـتفاوت خواهـد بـود، ولـی اصول کلی حاکم بر آنها همچـنان باقـی اسـت. به این ترتیب، نقش گوینده به محدودیتهایـی که بر ساختار زبان حاکم است محدود می‌گردد. در ادامه، بحث را بر اساس نظریهٔ اطلاعات، دنـبال مـی‌کنـیم و بـه بررسـی اختیار گوینده در زبان می‌پردازیم.

رویکرد نظریهٔ اطلاعات

نظـریهٔ اطلاعـات بـه کمیّـت، کدگذاری(نصب علائم) و انـتقال اطلاعات از مبدأ به مقصد مربوط می‌شود. قسمتی از ایـدهٔ اصـلی این نظریه به توسط نی‌کوئیست[۹] (1924) (Beerbower & Jordan, 1960: 43/1184-1198) و هارتـلی[۱۰] (۱۹۲۸) (Elliott, 1970: 3/319-335) فرمول‌بندی شد و بعد از جـنگ جهانی دوم این نظریه پیشرفت قابل توجهی کرد. کار اصلی را شانن[۱۱] انجام داد (1949). برخی از کارهـای شـانن، بـه سرعت، علمی شد و کاربرد پیدا کـرد. بـا توسـعهٔ تکـنولوژی در سـالهای اخیر، شاهد تغییراتـی هسـتیم. ایـن تغیـیرات سـبب ارتقای سطح تـئوری اطلاعـات شد و این نظریه نیز کمک بسزایی در بخشـی از توسـعهٔ تکـنولوژی ایفـا کـرده اسـت؛ تأثـیرات ایـن رشـد را بر جامعهٔ جهانی کنونی شاهد هسـتیم و در آینده شاهد پیشرفتهای بیشتری خواهیم بود.

اصـول بنـیادی ایـن تـئوری، تـا حـد زیـادی، کـار

---

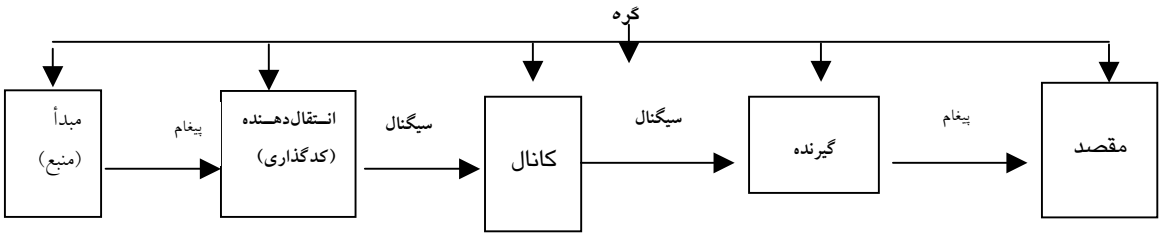
۸. نک: دبیرمقدم، ۱۳۸۳.

9. Nyquist
10. Hartley
11. Shannon

شانن و همکارانش در آزمایشگاههای تلفن بل در امریکاست. شانن در ۱۹۴۸ نظریهٔ ریاضی مخابرات را منتشر کرد و به دنبال آن، کاربردهای زیادی در زمینه‌های کامپیوتری، به خصوص ICT [۱۲] و مخابرات، زمین‌شناسی، تلویزیون و رادیو، الکترونیک پیدا کرد (Shannon, 1984: 27/37-423; Guiasu, 1977; Pelto, 1954: 62/501-511; Tasch, 1980; Beerbower…, 1960: 43/1184-1198; Ferguson, 1980: 137/107-108; Idem, 1982; Idem, 1983).

کـار اصلی شانن طراحی سیستمهای مخابراتی بود کـه بـه کمـک آنها بتوان اطلاعاتی معین را از نقطه‌ای بـه نقطـه‌ای دیگـر انتقال داد و درآنجا، به طور دقیق، بازسـازی کرد. به کمک ریاضی می‌توان اطلاعات را، چـه در مـبدأ و چه در مقصد، تعریف کمّی کرد و از طریق کانال، اطلاعات را منتقل کرد.

نمونه‌ای کلی از یک سیستم ارتباطی در شکل ۱ نشان داده شده است.



شکل۱. نمای کلی یک سیستم ارتباطی

مغز گوینده که مبدأ یا منبع اصلی تولید اطلاعات است، اطلاعـات (پـیغام) را از طـریق نـرون عصـبی به مجموعهٔ انـتقال‌دهـنده، یعنـی دهان، زبان و ...، به صورت کدگذاری شـده، یعنی زبان فارسی (یا انگلیسی یا ...) با رعایت قواعد حاکم بر آن ارسال کرده و دستور انتقال را می‌دهد. زمانی که شـروع بـه صـحبت مـی‌کند، این پیغام به شکل سیگنالهای صـوتی خـارج شـده و بـه مجموعـهٔ گیرنده، یعنی گوش شنونده می‌رسد. گوش شنونده، که با زبان فارسی (انگلیسی یـا ...) و قواعد آن از طریق مغز آشناست، پیغام را به توسط سلسـلهٔ اعصـاب به مغز شنونده انتقال می‌دهد. مغز شنونده اطلاعـات را دریافـت مـی‌کـند و به پردازش و بررسی آن می‌پردازد.

معـیاری کـه شـانن (1984:27/380) بـرای اطلاعـات پیشنهاد کرد، بیت یا رقم در مبنای ۲ است. بنابراین، اگر وسـیله‌ای مـثل کلید داشته باشیم که دو حالت پایه‌ای روشن و خاموش دارد، داریم :

$$-\log_2 \frac{1}{2} = 1 \text{ بیت}$$

n تـا از ایـن وسـایل مـی‌توانـند n بیت اطلاعات ذخیره کنند، عده کل حالات ممکن 2^n است. پس :

$$-\log_2 \frac{1}{2^n} = n \quad \text{بیت}$$

برای مثال تعداد حروف الفبای انگلیسی ۲۶ است بنابراین :

$$-\log_2 \frac{1}{26} = 4/7 \quad \text{بیت}$$

بـرای درک نظریه باید خانه‌هایی را تصور کنیم که هر کـدام مـی‌توانـند دو حالـت بـه خود بگیرند. در باب حروف الفبای فارسی شکل ۲ را ببینید:

شکل۲. در ۳۲ خانه ۵ بیت اطلاعات ذخیره می‌شود

اکنون دو نکته در خور ذکر است: یکی اینکه، می‌تـوان بـه جـای مبنای ۲، که شانن آن را در نظـر گرفت،

12. Information Communication Technology

از مبنای دیگری (مثلاً ۳ یا e یا ۱۰) استفاده کرد؛ دیگر اینکه، در هر خانه بیش از یک حرف نمی‌توان قرار داد و چون n حالت وجود دارد پس احتمال قرار گرفتن یکی از این حروف در خانه اول $\frac{1}{n}$ است. برای حروف الفبای فارسی برای خانهٔ اول $\frac{1}{32}$ احتمال این وجود دارد که یکی از حروف در آن خانه قرار گیرد. رابطه‌ای نزدیک بین اطلاعات و عدم یقین وجود دارد. هر چه پاسخ نامعینتر باشد، با دادن پاسخ، اطلاعات بیشتری به دست می‌آید. اگر چند حالت ممکن وجود داشته باشد، مقدار اطلاعات زمانی ماکزیمم می‌شود که احتمال وقوع همهٔ حالتها یکسان باشد. مجموعه‌ای از n اتفاق را با احتمال هر کدام $p_1, p_2, \ldots\ldots, p_n$ در نظر بگیرید. احتمال اینکه اتفاق اول روی دهد p_1 است که می‌تواند در خانه اول جای گیرد چون n خانه وجود دارد و ممکن است اتفاق اولی در قسمتهای خانه ٤ تایی جای گیرد. پس $n \times p_1$ زمان خانه‌ها توسط اتفاق اول اشغال می‌شوند و به این ترتیب $n \times -\log_2 p_1$ بیت اطلاعات ذخیره می‌شود.

آنتروپی سیستم را با H نشان داده و به صورت زیر تعریف می‌شود:

$$H = -K\sum_{i=1}^{n} P_i \log_2 p_i \qquad \text{بیت}$$

در رابطه فوق K به انتخاب واحد سنجش اطلاعات بستگی دارد (Shannon…,1949: 393; Guiasu, 1977:1ff). در اینجا K=1 انتخاب می‌شود. پس داریم:

$$H = -\sum_{i=1}^{n} P_i \log_2 p_i \qquad \text{بیت}$$

این معادله اساس نظریهٔ اطلاعات است و اطلاعات را به انتخاب و عدم یقین مربوط می‌کند. کلمهٔ آنتروپی در نظریهٔ اطلاعات تقریباً همان مفهوم آنتروپی در ترمودینامیک است. در ترمودینامیک، آنتروپی به مقدار اضمحلال سیستم گفته می‌شود. یکی از مفاهیم اساسی و مهم نظریه اطلاعات زوائد است، که به صورت زیر تعریف می‌شود (Usher, 1984):

$$\text{Re} = \% \frac{\text{آنتروپی واقعی ـ حداکثر آنتروپی}}{\text{حداکثر آنتروپی}}$$

اکنون سیستمی را در نظر می‌گیریم که دو حالت ممکن دارد و احتمال وقوع هر دو نیز برابر است؛ یعنی:

$$p_1 = p_2 = 0/5$$

پس داریم:

$H = (-0/5\log_2{}^{0/5}) + (-0/5\log_2{}^{0/5}) = 1$ اگر $p_1 = 1, p_2 = 0$ (یا بر عکس) آنگاه $H = 0$ و اگر $P_1 = 0/25, P_2 = 0/75$ (یا بر عکس) آنگاه H=0/8

بنابراین، آنتروپی سیستم وقتی ماکزیمم است که احتمال وقوع تمامی رخدادها با هم برابر باشد؛ و وقتی مینیمم است که احتمال وقوع یک رخداد برابر ۱ و احتمال وقوع سایر رخدادها صفر باشد.

محاسبهٔ آنتروپی زبان انگلیسی

زبان انگلیسی دارای ۲٦ حرف است و کل این حروف کلمات را تشکیل می‌دهند. ساختار زبان و دستور زبان انگلیسی کلمات را کنار هم قرار داده و جملات، که منظور نظر است، ساخته می‌شود. از آنجایی‌که ۲٦ حرف دارای احتمالات مساوی است، پس هر کدام به احتمال $\frac{1}{26}$ ممکن است ظاهر شود. مقدار آنتروپی(H) برای حروف انگلیسی به صورت زیر محاسبه می‌شود:

$$H = -\sum_{i=1}^{26} p_i \log_2{}^{p_i}$$

$$H = -(p(A)\log_2{}^{p(A)} + p(B)\log_2{}^{p(B)} + \ldots\ldots\ldots\ldots + P(Z)\log_2{}^{p(z)})$$

$$= \underbrace{-(\frac{1}{26}\log_2{}^{\frac{1}{26}} + \frac{1}{26}\log_2{}^{\frac{1}{26}} + \ldots\ldots\ldots + \frac{1}{26}\log_2{}^{\frac{1}{26}})}_{\text{۲۶ مرتبه}} = -(26(\frac{1}{26}\log_2{}^{\frac{1}{26}})) \approx 4/1$$

واحد آنتروپی(H) اغلب با بیت/ حرف یا بیت/ سمبل نشان داده می‌شوند. به این ترتیب، آنتروپی زبان انگلیسی حدود ٤/١بیت / نماد است.

زوائد در زبان انگلیسی

در نظریهٔ اطلاعات، مفهوم حشو به تکرار یک پیغام مربوط است. اگر N پیغام مختلف وجود داشته باشد و تنها یکی را، مثلاً ۳ بار، ارسال کنیم، با فرض اینکه کانال انتقال بدون گره[۱۳] باشد، تنها اولین پیغام بامعنی است و بقیهٔ پیغامها تکرار اولی بوده و زائد هستند. برای مثال، در یک سیستم دو دویی (مثلاً A=0 و B=1)، وقتی پیغام A را ۳ بار ارسال کنیم، به صورت کد گذارشده OOO، یا پیغام B را سه بار ارسال کنیم، که به صورت کدگذاری شده به شکل 111 می‌باشد، و اولین کد رسیده یعنی O (برای A)یا 1 (برای B) با معنی است و بقیه کدها تکرار پیغام قبلی و زائد است. در این صورت، زوائد به شکل زیر است:

$$\text{Re} = \% \frac{\text{ماکزیمم آنتروپی} - \text{آنتروپی واقعی}}{\text{ماکزیمم آنتروپی}} = \frac{3-1}{3} = \%\frac{2}{3} = \%66/6$$

از آنجایی‌که سه بار پیغام ارسال شده، پس سه بیت اشغال شده که ماکزیمم آنتروپی است و آنتروپی واقعی برابر ۱ است زیرا یک کد (۰ یا ۱) تنها یک بیت اشغال می‌کند.

کلمات بخش وسیعی از احتمالات را تشکیل می‌دهند. به این ترتیب که یک کلمه ممکن است دامنهٔ وسیعی از لغات را تشکیل دهد. یک مکالمهٔ تلفنی را در نظر بگیرید، برای گفتن یک کلمه در زبان انگلیسی ٢٦ انتخاب می‌توان برای حرف اول کلمه در نظر گرفت، که حدود ٤/١بیت / نماد اشغال می‌کند. فرض کنید اولین حرف D باشد. انتخاب حرف بعدی محدودتر است، زیرا در انگلیسی کلمه‌ای که مثلاً با DD شروع شود، وجود ندارد. این محدودیت مربوط به ساختار زبان انگلیسی است.

حدود ۷ حرف وجود دارد که می‌تواند بعد از D واقع شود که حدود ۲/۸ بیت خواهد بود. یعنی برای یک کلمه ۲ حرفی ۲/۸ بیت / نماد خواهیم داشت. فرض کنیم حرف دوم I است. انتخاب زیاد دیگری برای حرف سوم وجود دارد . اگر حروف بعدی به ترتیب R،T،S باشند بنابراین به کلمه DISTR رسیده‌ایم، ولی باز هم برای انتخاب حرف ششم حالتهایی وجود دارد. اگر حروف بعدی U،B،I باشند، بنابراین می‌توانیم کلمه را DISTRIBUTION یا DISTRIBUTIVE حدس بزنیم. ملاحظه می‌کنیم که آخرین اطلاعات کسب‌شده بخش کوچکی از اطلاعات را تشکیل می‌دهند؛ زیرا دنباله‌ای که از

13. Node

حروف DISTRIBU تشکیل می‌شود رویکرد مناسبی برای حدس زدن کلمه پیش روی ما قرار می‌دهد. این دنباله تقریباً همان چیزی است که در دستور زبان به آن بن کلمه گفته می‌شود، البته این دنباله گاهی از نظر تعداد حروف کمتر از تعداد حروف بن کلمه می‌باشد.

شانن و ویور[14] (1949) مطالعات و بررسیهای زیادی در مورد زوائد زبان انگلیسی انجام دادند. ولی محاسبهٔ زوائد یک متن که دارای تعدادی جمله است کار بسیار مشکلی است. زیرا، همانطور که می‌دانید، هر متن از چندین جمله و هر جمله از چندین و هر کلمه از چندین حرف تشکیل شده است و با توجه به اینکه همهٔ متنها یکی نیست و در واقع هر متنی با متن دیگر تفاوت دارد، پس محاسبهٔ زوائد یک متن با متن دیگر متفاوت خواهد بود.

یکی از بهترین کاربردهای حشو در زبان آن است که حدود اختیار گوینده و ساختار زبان را مشخص می‌کند؛ مثلاً، زوائد زبان انگلیسی حدود ۷۸ درصد است (Useher, 1984: 14). پس ۷۸ درصد علائم انگلیسی را ساختار زبان تعیین می‌کند و فقط ۲۲درصد اختیاری است. از دیگر ویژگیهای زوائد می‌توان به ساختن جدولهای بزرگ اشاره کرد. با زبانی که زوائد آن ۳۳درصد باشد، از لحاظ تئوری، می‌توان جدول کلمات متقاطع سه بعدی ساخت (Guiasu, 1977:130; Shannon, 1984:399). قابل ذکر است که دیدگاه نظریهٔ بهینگی، نیز ما را به نقطه‌ای مشابه در نظریهٔ اطلاعات می‌رساند.

محاسبهٔ آنتروپی و حشو در زبان فارسی

اکنون قسمتی از تئوری اطلاعات را برای زبان فارسی پیاده می‌کنیم. از آنجا که زبان فارسی ۳۲ حرف دارد، پس برای اینکه کلمه‌ای را ادا کنیم ابتدا یک حرف از ۳۲ حرف را باید انتخاب نماییم که احتمال چنین انتخابی (با توجه به متساوی بودن همهٔ احتمالها برای تمام حروف $\frac{1}{32}$) است.

مقدار آنتروپی سیستم زبان فارسی در زیر آمده است:

$$H = -\sum_{i=1}^{n} P_i \log p_i$$

$$H = -(p(ا) \log p(ا) + p(ب) \log p(ب) + \ldots\ldots\ldots + p(ی) \log p(ی)$$

$$H = -(\underbrace{\frac{1}{32}\log_2 \frac{1}{32} + \frac{1}{32}\log_2 \frac{1}{32} + \ldots\ldots\ldots + \frac{1}{32}\log_2 \frac{1}{32}}_{\text{۳۲مرتبه}}$$

H= ۵ بیت/سمبل

پس آنتروپی زبان فارسی ۵ بیت / نماد است که در مقایسه با آنتروپی زبان انگلیسی که برابر ٤/١بیت/نماد است به تعداد حدود ۰/۹بیشتر است. این اختلاف ناشی از تعداد بیشتر حروف در زبان فارسی است. برای محاسبهٔ زوائد زبان فارسی فقط روی حرف اول داریم:

$$\mathrm{Re} = \%\frac{5-1}{5} = \%80$$

بنابراین، زوائد زبان فارسی حدود ۸۰ درصد است. زوائد زبان انگلیسی حدود ۷۸ درصد است که، با مقایسه، درمی‌یابیم که حشو زبان فارسی در روی حرف اول حدود ۲ درصد بیشتر از وضعیت مشابه در زبان انگلیسی است. از آنجا که زوائد زبان فارسی ۸۰ درصد است، پس ۸۰ درصد علائم زبان فارسی را ساختار زبان تعیین می‌کند و فقط ۲۰درصد اختیاری است.

تا اینجا تنها منابع مجزای اطلاعاتی بررسی شد و فقط حرف اول کلمه مورد بحث قرار گرفت. چنانچه بحث بر سر حرف دوم کلمه را آغاز کنیم، صحبت از وابسته یا مستقل بودن حروف اول و دوم به میان می‌آید. در این حالت، به احتمالات مشترک

14. W.Weaver

اشاره خواهد شد. ابتدا بحث وقوع یا عدم وقوع اتفاقات متوالی و سپس، احتمالات مربوط را خواهیم داشت. این بحث سنگین و طویل است و در این مقال جای نمی‌گیرد. همین مقدار بدانیم که این بحث را نباید خاتمه یافته تلقی کرد.

نتیجه

این مقاله کاربرد مختصری از ریاضی را در بخش کوچکی از زبان نشان داد. تعریف و به کاربردن مفهوم جدید از نظریهٔ اطلاعات، که مربوط به بخش انتقال اطلاعات از فرستنده به گیرنده است و با زبان ریاضی بیان می‌شود را در قالب زبان فارسی روی حرف اول یک کلمه قرار دادیم و نتیجه آن شد که آنتروپی سیستم زبان فارسی روی حرف اول کلمه ۵ بیت /نماد است و زوائد آن حدود ۸۰ درصد می‌باشد؛ بنابراین، اختیار ما برای سخن گفتن ۲۰ درصد و احاطهٔ زبان فارسی بر گفتار ما حدود ۸۰ درصد است. در زبان انگلیسی آنتروپی سیستم حدود ٤/١ بیت/ نماد است و زوائد آن روی حرف اول کلمه ۷۸ درصد است. این مقاله می‌تواند راهگشای شاخهٔ جدیدی از تحقیقات در زبان فارسی باشد که پایهٔ آن ریاضی است. بسترسازی مناسب ریاضی برای زبان می‌تواند به درک، تسهیل در یادگیری و شناخت بیشتر زبان منجر شود. از این‌رو، توجه به این مسئله در خور اهمیت است.

منابع

جـان نـژاد، محسن(۱۳۸۰)، زبان وجنسیت، *تفاوتهای زبانی میان گویشـوران مـرد و زن ایرانی در تعامل مکالمه‌ای*، رسالهٔ دکترای زبانشناسی دانشگاه تهران؛

دبـیر مقـدم، محمد (۱۳۸۳)، «نظریه بهینگی»، *مجلهٔ زبان و ادب*، دانشگاه علامه‌طباطبایی، شمارهٔ ۲۰؛

ـــــــ (۱۳۸۳)، *زبانشناسـی نظـری، پـیدایش و تکویـن دستور زایشی*، سمت، تهران؛

فارسـیان، محمدرضـا (۱۳۷۸)، *جنسـیت و واژگـان*، پایـان نامـهٔ کارشناسی ارشد، دانشگاه تهران؛

Archangeli, D. (1997), *Optimality Theory :An introduction to linguistics in the 1990s*, Blasil Blackwell;

Atkinsoon, K. (1985), *Elementary numerical analysis*, New York: Wiley;

Beerbower, J. R. & Jordan, D. (1960), "Application of Information Theory to Palaeontological Problems", *Journal of palaeontology*;

Bromberger, S. & Halle, M. (1991), *Why phonology is Different*, Blasil Blackwell;

Chomsky, N. (1973), *Condition on transformation*, New York: Holt, Rinehart & Winston;

ـــــ (1981), *Lectures on Government and Binding*, Mass:MIT press, Cambridge;

Chomsky, N. & Hall, M. (1968), *The Sound Pattern of English*, Harper and Row, New York;

Elliott, R. E. (1970), "Simulation of a Productive Coal Measures Sequence", *the A mercian Geologist*;

Ferguson, J. (1980), "Application of Information Theory: Dataprocessing", *Quarterly Journal of the Geological society of London;*

ـــــ (1982), "The Application of Information Theory to Trend Surface Analysis", *Mathematical Journal*;

Idem, (1983);

Glass, J. C. (1980), *An introduction to mathematical methods in economics*, New York: Mc Graw-Hill;

Gray, J. R. (1967), *Probability*, Oliver and Boyd Edinburgh;

Guiasu, S. (1977), *Information Theory with Application,* New York : Mc Graw - Hill;

Harbaugh, J. W. & Merriam, D. F. (1968), *Computer Applications in Stratigraphic Analysis*, New York: Wiley;

Kager, R. (1999), *Optimality Theory*, Cambridge press;

Pelto, C. R. (1954), "Mapping of Multicomponent System", *Journal of Geology*;

Ross, J. (1967), Constraints on variables in syntax.oh.D.Dissertation MIT;

ـــ (1986), Infinite syntax. Norwood,N.J,Ablex publishing corporation;

Shannon, C. E. & Weaver, W. (1949), *A Mathematical*

Twaddell, W. F. (1935), "On defining the phoneme", *language Monograph*, N.16, reprinted in M. Joos(ed), *Readings in linguistics I*, Fourth edition (1966), the university of chicago press;

Usher, M. J. (1984), *Information Theory of Information Technologists*, London: Mc Millan.■

Theory of Communication, univ, of Illinois press, Champaign , Illinois;

Shannon, C. E. (1984), "A mathematical Theory of Communication", Technical Journal of the Bell system;

Tasch, P. (1980), Palaeobiology of the invertebrates, 2ndedn, New York: Wiley;