بررسی پیکره‌بنیاد واژه‌های هم‌معنی

شهرام مدرس خیابانی[1]

چکیده

چرچ و دیگران (۱۹۹۱: ۱۲-۱۳) با معرفی برخی ابزار آماری همچون «آزمون اطلاعات دوسویه»[2] و «آزمون تی»[3]، اهمیت چنین ابزاری را در تحلیل‌های زبان‌شناختی نشان می‌دهند. از سوی دیگر لاینز[4] (۱۹۹۵: ۶۲) تفاوت در باهم‌آیندهای دو واژهٔ «big» و «large» را از دلایل نبود هم‌معنایی مطلق میان این دو واژه برمی‌شمارد. در این مقاله سعی بر آن است تا با استفاده از دو ابزار ذکرشده، ضمن اشاره به اهمیت پیکره‌های زبانی و ابزار آماری در پژوهش‌های زبان‌شناختی، تفاوت واژه‌های هم‌معنی از منظر باهم‌آیی بررسی شود.

واژه‌های کلیدی: آزمون تی، آزمون اطلاعات دوسویه، هم‌معنایی، باهم‌آیی، پیکرهٔ زبانی.

۱- مقدمه

چرچ[5] و دیگران(۱۹۹۱: ۱) در مقاله‌ای با عنوان «کاربرد آمار برای تحلیل واژگانی» به نقش مهم ابزار آماری در بررسی‌های زبان‌شناختی و به ویژه واژه‌شناسی می‌پردازند. به عقیدهٔ آنها پیکره[6]های زبانی و به ویژه خطوط واژه‌نما[7] به تنهایی نمی‌تواند در مطالعات زبان‌شناختی و به طور اخص فرهنگ‌نگاری، راهگشا باشد. با مراجعه به خطوط واژه‌نمای یک واژه، گرچه می‌توان برحسب شمّ زبانی به برخی اطلاعات دست یافت، ذهن بشر قادر به کشف همهٔ الگوها، جمع‌آوری نمونه‌ها و طبقه‌بندی الگوها برحسب اهمیت آنها نخواهد بود. ضمن اینکه با توجه محض به شمّ زبانی، احتمال بروز خطا نیز بیشتر خواهد شد(همان). به همین دلیل چرچ و دیگران به شماری از ابزار آماری مانند

---

[1]- استادیار، دانشگاه آزاد اسلامی واحد کرج shahram.modarres@kiau.ac.ir

[2]- Mutual Information (MI) Test
[3]- T- Score
[4]- Lyons, J.
[5]- Church, K.
[6]- corpus
[7]- concordance lines

«آزمون اطلاعات دوسویه»، «آزمون تی»، «میانگین»[۱] و «واریانس»[۲] و همچنین نقش هر یک از این ابزار در بررسی بعضی حوزه‌های زبانی همچون واژگان اشاره می‌کنند (همان: ۴).

از طرفی لاینز (۱۹۹۵: ۶۲) با اشاره به تفاوت باهم‌آیندهای دو واژهٔ «big» و «large» به یکی از دلایل نبود هم‌معنایی مطلق میان واژه‌ها اشاره می‌کند.

در مقالهٔ حاضر سعی بر آن است تا ضمن معرفی دو «آزمون اطلاعات دوسویه» و «تی»، به نقش چنین ابزاری در یافتن باهم‌آیندهای یک واژه و همچنین تفاوت برخی واژه‌های هم‌معنی به‌واسطهٔ باهم‌آیندهای آنها اشاره شود. با توجه به اینکه پژوهش حاضر مبتنی بر رویکردی پیکره‌بنیاد[۳] است، لازم است به اهمیت رویکرد احتمالاتی به زبان و پیکره‌های زبانی و همچنین ساختار پیکرهٔ مورد استفاده در پژوهش حاضر اشاره شود.

۲- رویکرد احتمالاتی به زبان

رویکرد احتمالاتی با احیای «تجربه‌گرایی»[۴] شکل و قوت بیشتر گرفت (منینگ[۵] و شوتز[۶]، ۱۹۹۹: ۴). ۴). در سال‌های ۱۹۶۰ تا ۱۹۸۵ میلادی مطالعات زبان‌شناسی، روان‌شناسی، هوش مصنوعی[۷] و پردازش زبان طبیعی کاملاً تحت تأثیر رویکرد «خردگرایی»[۸] بود. این رویکرد بر این پایه استوار است که بخش اعظم دانش انسان براساس آنچه وراثت ژنتیکی نامیده می‌شود و ماهیتی پیش ساخته دارد شکل گرفته است و به کمک حواس به وجود نمی‌آید(همان: ۴-۵). در همین دوران رویکرد خردگرایانهٔ چامسکی، مطالعات زبان‌شناختی را به شدت تحت تأثیر خود قرار داد.

چامسکی سعی بر تأیید ذاتی بودن ماهیت زبان دارد و معتقد است که کودکان علیرغم برخورد با داده‌های محدود زبانی یا به عبارت دیگر «فقر محرک»[۹] می‌توانند پدیده‌ای پیچیده همچون زبان را بیاموزند. وی سعی دارد با طرح این موضوع که بخش‌های اصلی زبان ماهیتی ذاتی دارد و همچون شبکه‌ای پیچیده از بدو تولد به صورت ژنتیکی با کودک همراه است، مسئله فقر محرک را حل کند (همان: ۵). اما در رویکرد تجربه‌گرایی این اعتقاد وجود دارد که گرچه برخی توانایی‌های شناختی از بدو تولد در ذهن انسان وجود دارد، ذهن انسان دربرگیرنده مجموعه‌ای از اصول و مواردی نیست که به بخش‌های مختلف زبان و سایر حوزه‌های شناختی مربوط می‌شود (مانند نظریه‌های صرفی یا

---

[1]- mean
[2]- variance
[3]- corpus-based
[4]- empiricism
[5]- Manning, C. D.
[6]- Schütze, H.
[7]- artificial intelligence
[8]- rationalism
[9]- poverty of stimulus

حالت‌دهی[۱] و غیره)؛ بلکه ذهن انسان با بهره‌گیری از عملکردی کلی، تشخیص الگوها[۲] و تعمیم[۳] و و استفاده از حواس ورودی قوی، قادر به فراگیری ساختارهای ظریف زبان طبیعی خواهد بود (همان). از ۱۹۲۰ تا ۱۹۶۰ میلادی، رویکرد تجربه‌گرا، تحت تأثیر رویکرد خردگرا دچار افول شد، اما بـه نظـر می‌رسد که از دهۀ ۱۹۹۰ میلادی به بعد این رویکرد احیا شده است (همان).

در پردازش زبان طبیعی بر پایه رویکرد تجربه‌گرا، می‌توان با مشخص کردن الگوی زبانی عام و با استفاده از روش‌های آماری به منظور «اسـتنتاج ارزش پارامترهـا»[۴] و همچنـین تشـخیص الگـو و روش‌های یادگیری ماشینی و اعمال این موارد بر حجم زیاد «کاربرد زبانی»[۵]، ساختارهای پیچیـده و گستردۀ زبان را فرا گرفت(همان). بنابر این در چنین رویکردی به کاربرد زبـانی یـا بـه عبـارت بهتـر پیکره‌های زبانی توجهی خاص می‌شود.

نگرش پیکره بنیاد بیشتر با آراء فرث[۶] (۱۹۵۷) و شعار معروف او که «واژه را به کمک همنشینان همنشینان آن می‌شناسیم» شکل گرفت، هر چند چنین نگرشی در آثار ساختگرایان آمریکایی یـا بـه عبارت دیگر زبان‌شناسان پسابلومفیلدی و به ویژه آثار هریس[۷] نیز مشاهده می‌شود. هریس بـا طـرح روال‌های کشف سعی بر کشف خودکار ساختار زبانی دارد(منینگ و شوتز، ۱۹۹۹: ۶۵).

اما از دیگر تفاوت‌های میان دو رویکرد خردگرایی و تجربه‌گرایی نوع مقوله‌های مورد توجه در هر یک از این دو رویکرد است. زبان‌شناسان طرفدار چامسکی یا به عبارتی «زبان‌شناسان زایشی»[۸] به دنبال توصیف حوزۀ زبانی ذهنی انسان یا زبان درونی‌اند. این زبان درونی به کمک داده‌های زبانی همچون متون زبان یا به عبارت دیگر زبان بیرونی که شاهدی غیرمستقیم از زبان درونی است، به همراه شمّ زبانی گویشوران بومی، مشخص خواهد شد. در مقابل تجربه‌گرایان به دنبال توصیف زبان بیرونی‌اند یعنی آنچه واقعاً اتفاق افتاده است(همان: ۷). تفاوت دیگر این دو رویکرد این است که نگرش چامسکی نسبت به زبان، نگرش «مقوله‌ای»[۹] است که بر اساس آن عناصر زبانی به روش مقوله‌ای تقسیم می‌شود. برای نمونه جمله، یا دستوری است یا غیردستوری. در مقابل در رویکرد تجربه‌گرا و رویکرد احتمالاتی که ریشه در رویکرد تجربه‌گرا دارد، این مسئله اهمیت دارد که آیا جمله طبیعی است یا غیرطبیعی(همان: ۷). براساس این نگرش تشخیص دستوری یا غیردستوری بودن جمله همواره امکان‌پذیر نیست (همان: ۹). برخلاف چامسکی که سعی دارد مفهوم دستوری

1- case marking
2- pattern recognition
3- generalization
4- inducing the values of parameters
5- language use
6- Firth, J. R.
7- Harris, Z.
8- generative linguists
9- categorical

بودن را حتی با جملات نادر و غیرطبیعی چون «colorless green ideas sleep furiously» [ایده‌های سبز بی‌رنگ با عصبانیت می‌خوابند] توجیه کند، در نگرش تجربه‌گرا و در رأس آن در نگرش احتمالاتی، بر آنچه در کل زبان اتفاق می‌افتد تأکید می‌شود.

به عقیدهٔ منینگ و شوتز در نظر گرفتن احتمالات به عنوان بخشی از بررسی علمی زبان به این مفهوم است که شناخت انسان اساساً احتمالاتی است و از آنجا که زبان بخش اصلی شناخت انسان را تشکیل می‌دهد، پس زبان نیز باید احتمالاتی باشد(همان: ۱۵). اما در مقابل این نگرش، چامسکی (۱۹۵۷) معتقد است که نظریهٔ احتمالات برای صوری ساختن مفهوم دستوری بودن مناسب نیست. به عقیدهٔ وی با محاسبهٔ احتمال وقوع جملات از پیکره‌های زبانی، جملات تولید نشدهٔ دستوری و جملات غیردستوری از ارزش احتمالاتی کم و یکسان برخوردارند و در نتیجه زایایی زبان نادیده گرفته می‌شود (منینگ و شوتز، ۱۹۹۹: ۱۵-۱۶).

منینگ و شوتز معتقدند که چنین نگرشی در مورد فردی صادق است که در مقابل نمود احتمالاتی، موضع‌گیری افراطی می‌کند(همان: ۱۶). البته باید توجه داشت که الگوهای احتمالاتی ارائه شده در دو دههٔ ۱۹۴۰ و ۱۹۵۰ میلادی بسیار ساده بود و شاید از آنجا که با چنین الگوهایی نمی‌توان پیچیدگی‌های زبان را توصیف کرد، نامناسب تلقی می‌شدند(همان). اما امروزه بخش اصلی پردازش آماری زبان طبیعی، یافتن روش‌های تخمینی مناسب برای موارد مشاهده نیست و این فرض که تمام موارد مشاهده نشده در چارچوب احتمالاتی یکسانی قرار می‌گیرد، وجود ندارد. به عقیدهٔ منینگ و شوتز گرچه الگوهای احتمالاتی پیشرفته مانند الگوهای غیراحتمالاتی پیشرفته از ارزش تبیینی برخوردارند، می‌توانند پدیده‌های نامشخص و ناقص را که به طور عام در شناخت و به طور خاص در زبان مشاهده می‌شوند، نیز تبیین کنند (همان).

علاوه بر بحث رویکرد احتمالاتی، ذکر این نکته ضروری است که موضوع باهم‌آیی اساساً در سنت زبان‌شناسی چامسکی و حتی سوسور، مورد توجه قرار نگرفته است. در حالی که در زبان‌شناسی ساختگرا توجه بیشتر به «انتزاع کلی»[1] دربارهٔ ویژگی‌های گروه‌های نحوی و جملات زبان است، در سنت زبان‌شناسی انگلیسی که با نام افرادی چون فرث، هلیدی[2] و سینکلر[3] مطرح می‌شود، به موضوعی مانند باهم‌آیی توجهی خاص می‌شود (منینگ و شوتز، ۱۹۹۹: ۱۵۲). در نگرش فرث نسبت به معنی، که نگرشی بافت‌مقید محسوب می‌شود، بر اهمیت بافت تأکید می‌شود که این بافت در

[1]- general abstraction
[2]- Halliday, M. A. K.
[3]- Sinclair, J.

تقابل با «سخنگوی آرمانی»[1] مورد نظر چامسکی، موقعیت اجتماعی[2] را دربر می‌گیرد و در تقابل با «جملهٔ مجزای»[3] مورد نظر وی، بافت کلامی و متنی را شامل می‌شود (همان).

از طرف دیگر فرث با بیان عبارت مشهور خود که «واژه را به کمک همنشینانش می‌شناسیم» به مفهوم باهم‌آیی توجهی خاص دارد. به عقیدهٔ منینگ و شوتز این ویژگی‌های بافتی در نگرش انتزاعی زبان‌شناسیِ زایشی، نادیده گرفته شده است (همان).

۳- پیکرهٔ زبانی

به مجموعه‌ای از متون نوشتاری و گفتار آوانویسی شده، پیکره گفته می‌شود. پیکره‌ها در تحلیل و توصیف زبان‌شناختی مورد استفاده قرار می‌گیرند. طی سه دههٔ پایانی قرن بیستم میلادی تدوین و تحلیل پیکره‌های رایانه‌ای شاخهٔ علمی موسوم به «زبان‌شناسی پیکره‌ای»[4] را به وجود آورده است (کندی[5]، ۱۹۹۸: ۱). پیکره، مبنایی تجربی[6] دارد که به کمک آن می‌توان توزیع ویژگی‌های دستوری، آوایی یا کلامی خاص و محل کاربرد آنها را نشان داد (همان: ۴). تحلیل متون بزرگ بدون استفاده از رایانه امری طاقت‌فرسا است که احتمال بروز اشتباه را نیز درپی دارد. به کمک رایانه امکان دریافت اطلاعات از پیکره بدون هیچ نوع خطا، امکان‌پذیر است. امتیاز زبان‌شناسی پیکره‌ای این است که در بررسی دقیق‌تر مواردی که صرفاً بر اساس درون‌نگری یا به عبارت دیگر شمّ زبانی توجیه شده‌اند، ما را یاری می‌کند (همان: ۵).

اما پیکرهٔ مورد استفاده در پژوهش حاضر بخشی از پیکره‌ای است که توسط «پژوهشکدهٔ پردازش هوشمند علائم» به نام «دادگان زبان فارسی» ارائه شده است که شامل ۱۰,۱۳۶,۴۱۵ «نمونه»[7] از ۱۵۴,۵۹۵ «نوع»[8] است. منظور از نمونه، نمود یا نمودهای یک نوع در متن است. برای برای مثال واژهٔ «جمهوری» نوعی است که ۸,۲۶۱ بار در پیکره دیده شده است که در چنین شرایط می‌توان گفت که این نوع دارای ۸,۲۶۱ نمونه است.

هر واژه یا نوعِ مورد بررسی، بر اساس فاصلهٔ درج شده در دو سوی آن، از دیگر صورت‌ها متمایز شده است و بنابر این مواردی چون نشانهٔ «جمع» یا «یای نکره» که به واژهٔ مورد نظر چسبیده‌اند به‌عنوان صورتی مجزا در نظر گرفته نمی‌شوند. برای نمونه صورت «کتاب‌هایِ» علیرغم اینکه دارای

---

[1]- idealized speaker
[2]- social setting
[3]- isolated sentence
[4]- corpus linguistics
[5]- Kennedy, G.
[6]- empirical
[7]- token
[8]- type

نشانهٔ جمع و کسرهٔ اضافه است، یک نوع محسوب می‌شود. البته از بعد زبان‌شناختی «کسرهٔ اضافه» یا «نشانهٔ جمع» صورت‌های دستوری متمایز محسوب می‌شوند؛ اما از آنجا که تحلیل حاضر تنها به صورت‌های واژگانی اختصاص دارد، از تفکیک این موارد صرف نظر شده است.

پیکرهٔ مورد بررسی نمونهٔ انتخابی[۱] مناسب از زبان فارسی است، که به منظور متعادل‌سازی آن سبک‌ها و گونه‌های[۲] مختلف زبانی مورد استفاده قرار گرفته است. ضمن اینکه ۲۰ درصد از کل پیکره را زبان گفتاری تشکیل می‌دهد.

در دادگان مورد بررسی، نمونه‌ها به صورت «تک‌نگاشتی»[۳](یک واژه) یا «چندنگاشتی»[۴] از قبیل «دونگاشتی»[۵] (دو واژهٔ هم‌وقوع)، «سه‌نگاشتی»[۶] (سه واژهٔ هم‌وقوع) و «چهارنگاشتی»[۷] (چهار واژهٔ واژهٔ هم‌وقوع) به همراه فراوانی وقوع هر یک قابل استخراج است که در پژوهش حاضر از دونگاشتی‌ها استفاده شده است.

۴- آزمون اطلاعات دوسویه و آزمون تی

چرچ و دیگران (۱۹۹۱) به این نکته اشاره می‌کنند که توجه محض به پیکره‌های زبانی به تنهایی نمی‌تواند زبان‌شناس را در یافتن روابط ظریف میان واحدهای زبانی یاری کند. در رابطه با این مسئله آنان ابزار آماری متفاوتی ارائه کردند که در این بخش دو مورد، یعنی «آزمون‌های اطلاعات دوسویه» و «تی» معرفی می‌شود.

۴- ۱- اطلاعات دوسویه

منظور از اطلاعات دوسویه که به صورت $I(x,y)$ نشان داده می‌شود، مقایسهٔ میزان هم‌وقوعی مشاهده شدهٔ دو واژهٔ x و y با میزان احتمال وقوع دو واژه براساس فراوانی هر یک به تنهایی در متن است که به صورت فرمول ۱ نشان داده می‌شود:

$$I(x,y) = log_{۲} \frac{P(x,y)}{P(x)P(y)} = \frac{f(x,y)/N}{(f(x)/N)(f(y)/N)} = \frac{f(x,y)N}{f(x)f(y)}$$

فرمول ۱- نحوهٔ محاسبهٔ اطلاعات دوسویه (چرچ و دیگران، ۱۹۹۱: ۳)

---

[1]- representative Sample
[2]- genre
[3]- unigram
[4]- *n*-gram
[5]- bigram
[6]- trigram
[7]- quadrogram

اگر میان x و y رابطه قوی وجود داشته باشد، در این صورت میزان احتمال وقـوع دو واژه x و y بـا هم یعنی$P(x,y)$ از میزان وقوع تصادفی x و y بیشـتر اسـت و در نتیجـه $I(x,y)$ بزرگتـر از صـفر خواهد بود. اگر هیچ نوع رابطهٔ معنی‌داری میان x و y نباشد، در این صورت میزان وقـوع بـا هـم دو واژه تقریباً با میزان وقوع تصادفی آنها برابر و در نتیجه مقدار I تقریباً صفر خواهد شد. اگـر x و y در توزیع تکمیلی[1] باشند یعنی هرگز در کنار یکدیگر به‌کار نروند، در این صورت میزان وقوع دو واژه بـا هم، بسیار کمتر از میزان وقوع تصادفی آنها و در نتیجـه مقـدار I کمتـر از صـفر مـی‌شـود (چـرچ و دیگران، ۱۹۹۱: ۴).

برای محاسبهٔ احتمال وقوع دو واژه x و y کافی است که فراوانی هر یک از دو واژه یعنی $f(x)$ و $f(y)$ در متن بر تعداد کل واژه‌های متن یعنی N تقسیم شود. احتمال وقوع دو واژه با هم نیز می‌توان با محاسبه فراوانی مواردی که دو واژه در کنار یکدیگرند یعنی $f(x,y)$ و تقسیم آن بر تعداد کل واژه‌ها یعنی N به‌دست آورد (همان: ۵). چرچ و دیگران برای نشان دادن باهم‌آیندهای دو واژهٔ «strong» و« powerful » از پیکرهٔ ۴۴/۳ میلیون واژه‌ای (N=۴۴,۳۰۰,۰۰۰) آرشیو خبری خبرگزاری آسوشیتد پرس مربوط به سال ۱۹۸۸ استفاده کردند. برای نمونه اگر فراوانی واژه « strong»، ۷۸۰۹ و فراوانی واژهٔ «northerly»، ۲۸ و فراوانی توالی«strong northerly»، ۷ باشد، در چنین شرایطی میزان اطلاعات دوسویهٔ این توالی یعنی$I(strong, northerly)$ برمبنای فراوانی کل پیکره، ۱۰/۴۷ خواهد بود که این عدد به شکل زیر محاسبه می‌شود (همان):

$$I(strong;northerly) = log_{۲} \frac{۷ \times N}{۷۸۰۹ \times ۲۸} = ۱۰/۴۷$$

چرچ و دیگران با اشاره به این جمله فرث که «ما واژه را با همنشینان آن می‌شناسـیم» بـه ایـن مسئله اشاره می‌کنند که اطلاعات دوسویه، خلاصه‌ای از آنچه با واژه‌هـای مـورد نظـر مـا همنشـین می‌شود، در اختیارمان خواهد گذاشت (همان).

اما منینگ و شوتز (۱۹۹۹: ۱۷۸–۱۷۹)، تاریخچه‌ای از اطلاعات دوسویه را ارائـه مـی‌دهنـد. بـه عقیدهٔ آنها برای نخستین بار فانو[2] (۱۹۶۱)، اطلاعات دوسویه را دربارهٔ دو پدیـدهٔ x و y ایـن چنـین توصیف کرده است «اطلاعات دوسویه میزان اطلاعاتی است که وقوع پدیده‌ای مانند x دربارهٔ وقـوع پدیده‌ای چون y ارائه می‌دهد».

برای نمونه اگر مقدار اطلاعات دوسـویهٔ دو واژهٔ x و y در ترکیـب xy، ۱۸/۳۶ باشـد، بـه ایـن مفهوم است که وقوع واژهٔ x در موقعیت i در پیکره، به مقدار ۱۸/۳۶ بار افزایش می‌یابد اگـر واژهٔ y در موقعیت۱+i قرار گیرد. عکس این حالت نیـز وجـود دارد؛ بـه ایـن صـورت کـه وقـوع واژهٔ y در موقعیت۱+i به مقدار۱۸/۳۶ بار افزایش می‌یابد اگر واژه x در موقعیت i باشد؛ به عبارت دیگـر ایـن

[1]- complementary distribution
[2]- Fano, R. M.

مقدار بالای اطلاعات به ما کمک خواهد کرد که اگر x واژهٔ مورد نظر ما باشد، به احتمال بسیار واژهٔ y پس از آن حضور خواهد داشت (همان).

به عقیدهٔ منینگ و شوتز، یکی از ایرادات وارد بر اطلاعات دوسویه این است که مقدار اطلاعات دوسویه در توالی‌های دوتایی که از واژه‌های کم بسامد تشکیل شده است (و همیشه این واژه‌های کم‌بسامد کنار یکدیگر قرار می‌گیرد)، نسبت به توالی‌هایی که از واژه‌های پربسامد تشکیل شده است (و این واژه‌های پربسامد معمولاً کنار یکدیگر قرار نمی‌گیرد) بیشتر است. به عقیدهٔ منینگ و شوتز، این برخلاف آن چیزی است که یک ابزار آماری باید برای استخراج باهم‌آیی‌ها ارائه دهد، زیرا موارد پربسامد از اهمیت بیشتر برخوردار است. برای حل این مشکل منینگ و شوتز پیشنهاد می‌کنند که بهتر است واژه‌هایی که فراوانی وقوع کمتر از ۳ مورد در پیکره دارند، محاسبه نشود. به عقیدهٔ آنها گرچه مشکل مطرح شده حل نمی‌شود، نتایج بهتری به‌دست خواهد آمد (همان: ۱۸۲). کلیر[1] (۱۹۹۳) نیز با بررسی اطلاعات دوسویه نشان می‌دهد که این آزمون موارد خاص همچون اسامی افراد و موارد منحصر به فرد را بیشتر نمایان می‌سازد.

برای نمونه ممکن است ترکیبی که تنها ۳ بار در متن مشاهده شده است از مقدار اطلاعات دوسویهٔ زیاد برخوردار باشد. کلیر معتقد است مواردی که فراوانی‌شان کمتر از سه مورد است منحصر به فرد و تصادفی است و باید حذف شوند. به نظر استابز[2] (۱۹۹۵) هیچ معیار زبان‌شناختی برای حذف موارد کمتر از ۳ مورد وجود ندارد و احتمالاً این مقدار براساس مشاهدات کلیر درنظر گرفته شده است.

با این حال کلیر به این نکته اشاره می‌کند که اطلاعات دوسویه در برجسته ساختنِ اصطلاح‌ها، ضرب المثل‌ها و اصطلاح‌های فنی نقش مهمی ایفا می‌کند (کلیر، ۱۹۹۳).

۴- ۲- آزمون تی

براساس آزمون تی میزان تفاوت هم‌وقوعی مشاهده شدهٔ دو عنصر و هم‌وقوعی احتمالی دو عنصر مشخص خواهد شد؛ به عبارت دیگر می‌توان تشخیص داد که هم‌وقوعی مشاهده شدهٔ دو عنصر اتفاقی است یا خیر. این مقدار به صورت فرمول ۲ محاسبه می‌شود:

$$t = \frac{\frac{f(w1w2)}{N} - \frac{f(w1)f(w2)}{N^2}}{\frac{\sqrt{f(w1w2)}}{N}}$$

فرمول ۲- نحوهٔ محاسبهٔ تی برای یافتن رابطهٔ میان هم‌وقوعی احتمالی و هم‌وقوعی مشاهده شده (چرچ و دیگران، ۱۹۹۱: ۸)

[1]- Clear, J.
[2]- Stubbs, M.

در این فرمول f(w1) فراوانی واژهٔ نخست، f(w2) فراوانی واژهٔ دوم، f(w1w2) فراوانی دو واژهٔ پیاپی و N تعداد کل نمونه‌ها در پیکره است.

اگر مقدار تیِ به دست آمده از ۲/۵۷۶ بیشتر باشد با ۹۹ درصد اطمینان و اگر از ۱/۹۶۰ بیشتر باشد با ۹۵ درصد اطمینان می‌توان گفت که میان آنچه مشاهده شده و آنچه پیش‌بینی شده است اختلاف معنی‌داری وجود دارد و به این ترتیب می‌توان ادعا کرد که مقدار مشاهده شده بسیار بیش‌تر از مقدار پیش‌بینی شده است و در نتیجه توالی مورد بررسی معنی‌دار است و می‌توان آن را باهم‌آیی تلقی کرد. در غیر این‌صورت یعنی در صورتی که این مقدار کمتر از ۲/۵۷۶ (برای اطمینان ۹۹ درصد) یا کمتر از ۱/۹۶۰ (برای اطمینان ۹۵ درصد) باشد می‌توان گفت که میان آنچه مشاهده شده و آنچه پیش‌بینی شده اختلاف معنی‌داری وجود ندارد و توالی مشاهده شده به صورت تصادفی رخ داده است.

چرچ و دیگران (۱۹۹۱: ۸) برای مشخص کردن احتمال باهم‌آیی «powerful» [قوی] و «support » [حمایت]، از آزمون تی به صورت فرمول ۳ استفاده می‌کنند. در این بررسی میزان N برابر۴۴/۳ میلیون واژه، f(powerful support) برابر ۲ ، f(support) برابر ۱۳,۴۲۸ و f(powerful) برابر۱۹۸۴ است.

$$t = \frac{p(w1w2) - p(w1)p(w2)}{\sqrt{\sigma^2 p(w1w2) + \sigma^2 (p(w1)p(w2))}} = \frac{\frac{f(w1w2)}{N} - \frac{f(w1)f(w2)}{N^2}}{\frac{\sqrt{f(w1w2)}}{N}}$$

فرمول ۳- نحوهٔ محاسبهٔ تی برای یافتن رابطهٔ میان هم‌وقوعی احتمالی و هم‌وقوعی مشاهده شده (چرچ و دیگران، ۱۹۹۱: ۸)

$$t = \frac{p(powerful\ support) - p(powerful)p(support)}{\sqrt{\sigma^2(powerful\ support) + \sigma^2(p\ (powerful)p(support))}}$$

$$= \frac{\frac{f(powerful\ support)}{N} - \frac{f(powerful)f(support)}{N^2}}{\frac{\sqrt{f(powerful\ support)}}{N}}$$

$$\approx \frac{2 - \frac{1984 \times 13428}{N}}{\sqrt{2}} \approx 0/99$$

مقدار تی به دست آمده بسیار کمتر از ۱/۹۶۰ است و بنابر این می‌توان ادعا کرد که چنین ترکیبی به شکل تصادفی رخ داده است.

در ادامه چرچ و دیگران به این مسئله اشاره می‌کنند که گرچه معیار اطلاعات دوسویه برای نشان دادن ارتباط میان واژه‌ها سودمند است، تفاوت میان دو واژهٔ هم‌معنی را نشان نمی‌دهد؛ بنابر این بهتر است از معیاری دیگر همچون آزمون تی استفاده شود(همان: ۶-۸). آنها برای نشان دادن تمایز میان «powerful support » و «strong support» از آزمون تی استفاده می‌کنند. بر اساس فرض صفر میان«powerful support »و «strong support» اختلاف معنی‌داری وجود ندارد. این مقدار به صورت زیر محاسبه خواهد شد (در اینجا f(strong support) برابر با ۱۷۵ است و سایر مقادیر مشابه نمونهٔ قبلی است):

$$t = \frac{p(powerful\ support) - p(strong\ support)}{\sqrt{(p(powerful\ support))^{2} + (p(\ strong\ support))^{2}}}$$

$$= \frac{\frac{f(powerful\ support)}{N} - \frac{f(strong\ support)}{N}}{\sqrt{\frac{f(powerful\ support)}{N^{2}} + \frac{f(strong\ support)}{N^{2}}}} \approx \frac{۲ - ۱۷۵}{\sqrt{۲ + ۱۷۵}} = -۱۳$$

مقدار تیِ برابر ۱۳- نشان می‌دهد که میان دو ترکیب« strong support» و« powerful support» اختلاف فاحش وجود دارد و بنابر این فرض صفر رد می‌شود.

بدین ترتیب مشاهده می‌شود که آزمون تی نه تنها در مشخص کردن باهم‌آیند بودن دو عنصر زبانی سودمند است، بلکه برای مشخص کردن تمایز دو واژهٔ هم‌معنی نیز مورد استفاده قرار می‌گیرد. در بخش پنجم با اعمال دو ابزار آماری ذکر شده بر پیکرهٔ زبانی پژوهش [رک بخش دوم مقالهٔ حاضر]، مشخص خواهد شد که باهم‌آیندهای صورت‌های هم‌معنی به لحاظ آماری متفاوت‌اند و این خود می‌توان دلیل محکمی بر نبود هم‌معنایی مطلق باشد.

۵- بررسی باهم‌آیندهای صورت‌های هم‌معنی

لاینز(۱۹۹۵: ۶۲) تفاوت باهم‌آیندهای دو واژهٔ «big » و «large» را یکی از دلایل نبود هم‌معنایی مطلق میان این دو واژه برمی‌شمارد. در این پژوهش سعی بر آن است تا به کمک دو ابزار آماری یعنی آزمون اطلاعات دوسویه و آزمون تی ضمن نشان دادن باهم‌آیندهای واژه‌ها، تفاوت معنی واژه‌های هم‌معنی بررسی شود. برای دست یافتن به این مهم، ابتدا نمونه‌هایی از صورت‌های واژگانیِ

هم‌وقوعِ دو واژهٔ نسبتاً هم‌معنیِ «مشکی» و «سیاه» را که به ترتیب در جدول‌های ۱ و ۲ ارائه شده است، بررسی خواهد شد[۱].

نوع اول	نوع دوم	فراوانی نوع اول	فراوانی نوع دوم	فراوانی توالی نوع اول و دوم	اطلاعات دوسویه	مقدار تی
لباس	مشکی	۶۵۰	۴۷	۶	۱۰,۹۵	۲,۴۴
پیراهن	مشکی	۱۱۱	۴۷	۵	۱۳,۲۴	۲,۲۳
چادر	مشکی	۹۰	۴۷	۳	۱۲,۸۱	۱,۷۳

جدول ۱- صورت‌های پیش از واژهٔ «مشکی» (ترتیب دونگاشتی‌ها براساس مقدارتی)

نوع اول	نوع دوم	فراوانی نوع اول	فراوانی نوع دوم	فراوانی توالی نوع اول و دوم	اطلاعات دوسویه	مقدار تی
دریای	سیاه	۹۹۵	۶۹۹	۶۳	۹,۸۴	۷,۹۲
جعبه	سیاه	۱۱۶	۶۹۹	۲۱	۱۱,۳۵	۴,۵۸
بازار	سیاه	۲۳۳۵	۶۹۹	۱۹	۶,۸۸	۴,۳۲
رنگ	سیاه	۱۱۷۲	۶۹۹	۱۳	۷,۳۲	۳,۵۸
غلام	سیاه	۱۴۱	۶۹۹	۹	۹,۸۵	۲,۹۹
طنز	سیاه	۲۷۸	۶۹۹	۸	۸,۷۰	۲,۸۲
کنیز	سیاه	۱۹	۶۹۹	۸	۱۲,۵۷	۲,۸۲
دوران	سیاه	۲۷۹۵	۶۹۹	۷	۵,۱۸	۲,۵۷
ابرهای	سیاه	۴۳	۶۹۹	۶	۱۰,۹۸	۲,۴۴
تخته[۲]	سیاه	۸۶	۶۹۹	۶	۹,۹۸	۲,۴۴
چشمه‌ای	سیاه	۱۵۲	۶۹۹	۵	۸,۸۹	۲,۲۳

[۱]- در پژوهش حاضر تنها توالی‌هایی با فراوانی ۳ و بیشتر بررسی شده است. از سوی دیگر به دلیل تعداد زیاد نمونه‌ها، تعداد زیادی از نمونه‌هایی که مقدار تی در آنها کمتر از ۱/۹۶۰ بود، حذف شده است.

[۲] - به دلیل درج فاصله میان عناصر واژهٔ مرکب «تخته سیاه» این نمونه، توالی دونگاشتی محسوب شده است و این مسئله یکی دیگر از مشکلات پژوهش‌های پیکره‌بنیاد را نشان می‌دهد.

نوع اول	نوع دوم	فراوانی نوع اول	فراوانی نوع دوم	فراوانی توالی نوع اول و دوم	اطلاعات دوسویه	مقدار تی
اسب	سیاه	۲۷۷	۶۹۹	۵	۸,۰۳	۲,۲۲
آفریقای	سیاه	۳۶۱	۶۹۹	۵	۷,۶۴	۲,۲۲
قاره	سیاه	۴۰۸	۶۹۹	۵	۷,۴۷	۲,۲۲
لباس	سیاه	۶۵۰	۶۹۹	۵	۶,۸۰	۲,۲۱
فهرست	سیاه	۶۸۲	۶۹۹	۵	۶,۷۳	۲,۲۱
صورت	سیاه	۱۰۹۲۵	۶۹۹	۶	۲,۹۹	۲,۱۴
دود	سیاه	۱۹۲	۶۹۹	۴	۸,۲۳	۱,۹۹
جگر	سیاه	۴۹	۶۹۹	۴	۱۰,۲۰	۱,۹۹
سنگ‌های	سیاه	۲۱۴	۶۹۹	۴	۸,۰۸	۱,۹۹
استبداد	سیاه	۲۳۶	۶۹۹	۴	۷,۹۴	۱,۹۹
دودکش‌های	سیاه	۷	۶۹۹	۴	۱۳,۰۱	۱,۹۹
مخمل	سیاه	۲۲	۶۹۹	۴	۱۱,۳۶	۱,۹۹
چای	سیاه	۴۰۵	۶۹۹	۴	۷,۱۶	۱,۹۸
جمعه	سیاه	۱۶۳۸	۶۹۹	۴	۵,۱۴	۱,۹۴
نفت	سیاه	۳۲۷۳	۶۹۹	۴	۴,۱۴	۱,۸۸
آب	سیاه	۵۲۳۹	۶۹۹	۴	۳,۴۶	۱,۸۱
خال	سیاه	۳۳	۶۹۹	۳	۱۰,۳۶	۱,۷۳
لکه‌های	سیاه	۴۰	۶۹۹	۳	۱۰,۰۸	۱,۷۳

جدول ۲- صورت‌های پیش از واژهٔ «سیاه» (ترتیب دونگاشتی‌ها براساس مقدار تی)

همانگونه که در جدول ۱ مشاهده می‌شود، واژهٔ «مشکی» تنها با سه صورت واژگانی «لباس»، «پیراهن» و «چادر» به‌کار رفته است و در میان این سه مورد، دو توالی «لباس مشکی» و «پیراهن مشکی» از مقدار تی بیش از ۱/۹۶۰ برخوردارند و بنابر این به لحاظ آماری (با سطح اطمینان ۹۵ درصد) معنی‌دار تلقی می‌شوند. این در حالی است که مقدار اطلاعات دوسویه در هر سه مورد نسبتاً زیاد (بیش از ۱۰) است و این میزان زیاد حاکی از پیوستگی عناصر این نمونه‌ها است.

اما جدول ۲ نشان می‌دهد که برخلاف واژهٔ «مشکی»، تعداد صورت‌های هم‌وقوع واژهٔ «سیاه» بسیار زیاد است و این مسئله یعنی گستردگی صورت‌های هم‌وقوع واژهٔ «سیاه» و محدود بودن هم‌وقوع‌های واژهٔ «مشکی» را می‌توان یکی از تفاوت‌های دو واژهٔ «مشکی« و «سیاه» تلقی کرد.

در صورت‌های ارائه شده در جدول ۲ نیز شاهد توالی «لباس سیاه» هستیم. گرچه در نمونه‌های ارائه شده، هر دو صورت «مشکی» و «سیاه» پس از واژهٔ «لباس» مشاهده می‌شود، مقدار متفاوت تی و اطلاعات دوسویه، تفاوت دو توالی «لباس سیاه» و «لباس مشکی» را نشان خواهد داد. در حالی که مقدار تی در نمونه «لباس سیاه» ۲/۲۱ است، مقدار تی در نمونهٔ «لباس مشکی» ۲/۴۴ است. از سوی دیگر مقدار اطلاعات دوسویه در نمونهٔ «لباس سیاه» ۶/۸۰ است و این در حالی است که این مقدار در نمونهٔ «لباس مشکی» ۱۰/۹۵ است. مقدار اطلاعات دوسویهٔ بیشتر در نمونهٔ «لباس مشکی» در مقایسه با «لباس سیاه» حکایت از همنشینی بیشتر واژهٔ «مشکی» با واژهٔ «لباس» در مقایسه با واژهٔ «سیاه» دارد.

مسئلهٔ نبود هم‌معنایی مطلق را می‌توان میان چند واژه نیز بررسی کرد. صفوی (۱۳۷۹: ۱۹۷) هنگام بحث دربارهٔ باهم‌آیی واژگانی به نقش واژه‌های باهم‌آیند در تشخیص تفاوت صورت‌های هم‌معنی همچون «پیر»، «کهنه»، «قدیمی» و «کهنسال» اشاره می‌کند. در جدول‌های ۳ تا ۶ به ترتیب صورت‌های واژگانی هم‌وقوع واژه‌های «کهنسال»، «پیر»، «کهنه» و «قدیمی» ارائه شده است.

نمونه‌های جدول‌های ۳ تا ۶ حاکی از تمایز صورت‌های هم‌وقوع چهار واژهٔ ذکر شده است. در حالی که واژهٔ «کهنسال» تنها با دو واژهٔ «درختان» و «درخت» به‌کار رفته است، واژهٔ «قدیمی» از طیف گسترده‌ای از صورت‌های هم‌وقوع برخوردار است. از سوی دیگر در حالی که واژهٔ «پیر» بیشتر با جانداران و به ویژه انسان‌ها به‌کار می‌رود، صورت «قدیمی» غالباً با اشیاء و به ویژه انواع «ساختمان» به‌کار می‌رود. این در حالی است که واژهٔ «کهنه» نسبت به سایر نمونه‌ها بیشتر از بار عاطفی منفی برخوردار است.

نوع اول	نوع دوم	فراوانی نوع اول	فراوانی نوع دوم	فراوانی توالی نوع اول و دوم	اطلاعات دوسویه	مقدار تی
درختان	کهنسال	۲۷۵	۵۱	۷	۱۲,۳۰	۲,۶۴
درخت	کهنسال	۴۳۵	۵۱	۳	۱۰,۴۲	۱,۷۳

جدول ۳- صورت‌های پیش از واژهٔ «کهنسال»(ترتیب دونگاشتی‌ها براساس مقدار تی)

نوع اول	نوع دوم	فراوانی نوع اول	فراوانی نوع دوم	فراوانی توالی نوع اول و دوم	اطلاعات دوسویه	مقدار تی
مرد	پیر	۲۰۳۶	۲۸۳	۸	۷,۱۳	۲,۸۰
موش‌های	پیر	۷۰	۲۸۳	۶	۱۱,۵۸	۲,۴۴
مزار	پیر	۱۳۲	۲۸۳	۶	۱۰,۶۶	۲,۴۴
مادر	پیر	۱۳۴۵	۲۸۳	۶	۷,۳۱	۲,۴۳
زن	پیر	۲۴۴۸	۲۸۳	۶	۶,۴۵	۲,۴۲
افراد	پیر	۵۱۴۲	۲۸۳	۵	۵,۱۲	۲,۱۷
امامزاده	پیر	۲۴۸	۲۸۳	۴	۹,۱۷	۱,۹۹
کفتار	پیر	۷	۲۸۳	۴	۱۴,۳۲	۱,۹۹
پدر	پیر	۱۴۸۸	۲۸۳	۴	۶,۵۸	۱,۹۷
گفتار	پیر	۶۶۵	۲۸۳	۳	۷,۳۳	۱,۷۲
ره	پیر	۱۸۴۶	۲۸۳	۳	۵,۸۶	۱,۷۰
جامعه	پیر	۷۵۶۳	۲۸۳	۳	۳,۸۲	۱,۶۱

جدول ۴- صورت‌های پیش از واژهٔ «پیر»(ترتیب دونگاشتی‌ها براساس مقدار تی)

نوع اول	نوع دوم	فراوانی نوع اول	فراوانی نوع دوم	فراوانی توالی نوع اول و دوم	اطلاعات دوسویه	مقدار تی
اطلاع	کهنه	۱۴۸۵	۲۱۸	۶	۷,۵۵	۲,۴۳
کفش	کهنه	۱۵۵	۲۱۸	۵	۱۰,۵۰	۲,۲۳
زخم	کهنه	۱۹۹	۲۱۸	۵	۱۰,۱۹	۲,۲۳
پل	کهنه	۶۹۹	۲۱۸	۵	۸,۳۷	۲,۲۲
روپوش‌های	کهنه	۶	۲۱۸	۴	۱۴,۹۱	۱,۹۹
شیوه‌های	کهنه	۵۳۳	۲۱۸	۴	۸,۴۴	۱,۹۹
سال	کهنه	۲۶۴۸۸	۲۱۸	۵	۳,۱۳	۱,۹۸
وقت	کهنه	۲۵۰۱	۲۱۸	۴	۶,۲۱	۱,۹۷

نوع اول	نوع دوم	فراوانی نوع اول	فراوانی نوع دوم	فراوانی توالی نوع اول و دوم	اطلاعات دوسویه	مقدار تی
کوزه	کهنه	۳۵	۲۱۸	۳	۱۱,۹۶	۱,۷۳
سرای	کهنه	۵۴	۲۱۸	۳	۱۱,۳۳	۱,۷۳
قصر	کهنه	۱۳۶	۲۱۸	۳	۱۰,۰۰	۱,۷۳

جدول ۵- صورت‌های پیش از واژهٔ «کهنه»(ترتیب دونگاشتی‌ها براساس مقدار تی)

نوع اول	نوع دوم	فراوانی نوع اول	فراوانی نوع دوم	فراوانی توالی نوع اول و دوم	اطلاعات دوسویه	مقدار تی
شهر	قدیمی	۹۲۸۹	۸۳۷	۲۲	۴,۸۴	۴,۵۲
بافت	قدیمی	۴۵۷	۸۳۷	۱۸	۸,۸۹	۴,۲۳
بسیار	قدیمی	۷۳۴۰	۸۳۷	۱۷	۴,۸۰	۳,۹۷
پل	قدیمی	۶۹۹	۸۳۷	۱۲	۷,۶۹	۳,۴۴
دوستان	قدیمی	۸۱۷	۸۳۷	۱۲	۷,۴۷	۳,۴۴
دوست	قدیمی	۱۴۴۳	۸۳۷	۱۲	۶,۶۵	۳,۴۲
بازار	قدیمی	۲۳۳۵	۸۳۷	۹	۵,۵۴	۲,۹۳
خودروهای	قدیمی	۳۴۷	۸۳۷	۸	۸,۱۲	۲,۸۱
مدارس	قدیمی	۱۳۷۰	۸۳۷	۸	۶,۱۴	۲,۷۸
ساختمان	قدیمی	۱۶۸۹	۸۳۷	۸	۵,۸۴	۲,۷۷
محلات	قدیمی	۱۰۳	۸۳۷	۷	۹,۶۸	۲,۶۴
قیمت‌های	قدیمی	۲۱۱	۸۳۷	۷	۸,۶۵	۲,۶۳
سرباز	قدیمی	۳۴۱	۸۳۷	۷	۷,۹۵	۲,۶۳
کتاب‌های	قدیمی	۹۴۶	۸۳۷	۷	۶,۴۸	۲,۶۱
بناهای	قدیمی	۲۲۲	۸۳۷	۶	۸,۳۵	۲,۴۴
اشیاء	قدیمی	۳۲۸	۸۳۷	۶	۷,۷۹	۲,۴۳
خانه	قدیمی	۳۲۰۸	۸۳۷	۶	۴,۵۰	۲,۳۴

خانه‌های	قدیمی	۲۸۹	۸۳۷	۶	۷,۹۷	۲,۴۳
قلعه	قدیمی	۲۹۹	۸۳۷	۶	۷,۹۲	۲,۴۳
آثار	قدیمی	۳۵۵۹	۸۳۷	۶	۴,۳۵	۲,۳۲
محله‌های	قدیمی	۵۲	۸۳۷	۵	۱۰,۱۸	۲,۲۳
همکار	قدیمی	۱۱۷	۸۳۷	۵	۹,۰۱	۲,۲۳
سربازان	قدیمی	۴۴۰	۸۳۷	۵	۷,۱۰	۲,۲۱
سنت	قدیمی	۷۵۲	۸۳۷	۵	۶,۳۳	۲,۲۰
نام	قدیمی	۵۹۵۲	۸۳۷	۵	۳,۳۴	۲,۰۱
دروازه‌های	قدیمی	۶۶	۸۳۷	۴	۹,۵۱	۱,۹۹
خانه‌ای	قدیمی	۱۳۹	۸۳۷	۴	۸,۴۴	۱,۹۹
نقشه‌های	قدیمی	۱۴۱	۸۳۷	۴	۸,۴۲	۱,۹۹
بازیگران	قدیمی	۳۸۰	۸۳۷	۴	۶,۹۹	۱,۹۸
باشگاه‌های	قدیمی	۶۰۸	۸۳۷	۴	۶,۳۱	۱,۹۷
منزل	قدیمی	۷۵۰	۸۳۷	۴	۶,۰۱	۱,۹۶

جدول ۶- صورت‌های پیش از واژهٔ «قدیمی» (ترتیب دونگاشتی‌ها براساس مقدار تی)

نکتهٔ مهم این است که همهٔ نمونه‌های ارائه‌شده در این چهار جدول، با سطح اطمینان ۹۹ درصد، معنی‌دار نیست. در جدول ۷ نمونه‌هایی با مقدار تی بیش از ۲/۵۷۶ ارائه شده است. در میان نمونه‌های جدول ۵ هیچ نمونهٔ معنی‌داری، با سطح اطمینان ۹۹ درصد، در کنار واژهٔ «کهنه» به چشم نمی‌خورد. این در حالی است که از دو نمونهٔ جدول ۳ تنها عبارت «درختان کهنسال» با سطح اطمینان ۹۹ درصد معنی‌دار است. جالب‌تر اینکه مقدار اطلاعات دوسویه در توالی «درختان کهنسال» (۱۲/۳۰) از سایر نمونه‌ها بیشتر است و به همین دلیل است که امکان پیش‌بینی موصوف واژهٔ «کهنسال» برای واژهٔ «درخت» نسبت به سه صفت دیگر ذکر شده، بسیار بیشتر است.

نوع اول	نوع دوم	فراوانی نوع اول	فراوانی نوع دوم	فراوانی توالی نوع اول و دوم	اطلاعات دوسویه	مقدار تی
شهر	قدیمی	۹۲۸۹	۸۳۷	۲۲	۴,۸۴	۴,۵۲
بافت	قدیمی	۴۵۷	۸۳۷	۱۸	۸,۸۹	۴,۲۳
بسیار	قدیمی	۷۳۴۰	۸۳۷	۱۷	۴,۸۰	۳,۹۷

پل	قدیمی	۶۹۹	۸۳۷	۱۲	۷٫۶۹	۳٫۴۴
دوستان	قدیمی	۸۱۷	۸۳۷	۱۲	۷٫۴۷	۳٫۴۴
دوست	قدیمی	۱۴۴۳	۸۳۷	۱۲	۶٫۶۵	۳٫۴۲
بازار	قدیمی	۲۳۳۵	۸۳۷	۹	۵٫۵۴	۲٫۹۳
خودروهای	قدیمی	۳۴۷	۸۳۷	۸	۸٫۱۲	۲٫۸۱
مرد	پیر	۲۰۳۶	۲۸۳	۸	۷٫۱۳	۲٫۸۰
مدارس	قدیمی	۱۳۷۰	۸۳۷	۸	۶٫۱۴	۲٫۷۸
ساختمان	قدیمی	۱۶۸۹	۸۳۷	۸	۵٫۸۴	۲٫۷۷
محلات	قدیمی	۱۰۳	۸۳۷	۷	۹٫۶۸	۲٫۶۴
قیمت‌های	قدیمی	۲۱۱	۸۳۷	۷	۸٫۶۵	۲٫۶۳
سرباز	قدیمی	۳۴۱	۸۳۷	۷	۷٫۹۵	۲٫۶۳
کتاب‌های	قدیمی	۹۴۶	۸۳۷	۷	۶٫۴۸	۲٫۶۱
درختان	کهنسال	۲۷۵	۵۱	۷	۱۲٫۳۰	۲٫۶۴

جدول ۷- باهم‌آیندهای واژه‌های «قدیمی»، «پیر» و «کهنسال» با مقدار تی بیش از ۲/۵۷۶

البته در چهار جدول ۳ تا ۶ نمونه‌هایی به‌چشم می‌خورد که مقدار تی آنها از ۱/۹۶۰ بیشتر است و بنابر این می‌توان با سطح اطمینان ۹۵ درصد، آنها را معنی‌دار تلقی کرد. از جملهٔ این نمونه‌ها می‌توان به «زخم کهنه»، «محله‌های قدیمی» و «کفتار پیر» اشاره کرد که اتفاقاً از مقدار اطلاعات دوسویهٔ زیاد (بیش از ۱۰) نیز برخوردارند. این نمونه‌ها حداقل برحسب شمّ زبانی نگارنده معنی‌دار به نظر می‌رسند. چنین نمونه‌هایی نشان می‌دهند که در پیکره‌های کم‌حجم همچون پیکرهٔ مورد استفاده در پژوهش حاضر، پائین آوردن سطح اطمینان تا ۹۵ درصد برای یافتن صورت‌های معنی‌دار ضروری است. البته امکان متمایز ساختن واژه‌های هم‌معنی مبتنی بر صورت‌های باهم‌آیند همواره امکان‌پذیر نیست. سینکلر (۱۹۹۶ و ۲۰۰۴) اشاره می‌کند که گوینده یا نویسنده بر اساس نوعی رویکرد کاربردشناختی به کل عبارت، نوع خاصی از واژه‌ها را انتخاب می‌کند و این رویکرد توسط شنونده یا خواننده قابل درک است. این رویکرد گوینده یا نویسنده نوای معنایی[1] نامیده می‌شود (لودلینگ[2] و کیتو[3]: ۲۰۰۹، ۹۸۰). به‌عبارت ساده‌تر گوینده یا نویسنده بر اساس نگرش مثبت، خنثی یا منفی خود، نوع خاصی از واحدهای واژگانی را در کنار یکدیگر قرار می‌دهد. برای نمونه از آنجا که

[1]- semantic prosody
[2]- Lüdeling, A.
[3]- Kytö, M.

باهم‌آیندهای اصلی واژهٔ «cause» [سبب] موارد ناخوشایندی چون «problems» [مشکلات]، «trouble» [زحمت، سختی]، «damage» [خرابی، ویرانی]، «pain» [درد] و «disease »[بیماری] را دربر می‌گیرد، این واژه از نوای معنایی منفی برخوردار است (استابز، ۱۹۹۵). سینکلر (۱۹۹۱: ۱۱۲) با بررسی صورت‌های هم‌وقوع فعل «happen» [اتفاق افتادن] نشان می‌دهد که این فعل غالباً با موارد ناگوار همچون «تصادف کردن» همراه می‌شود و نتیجه می‌گیرد که کاربرد زیاد واژه‌ها و عبارت‌ها میل به حضور در یک محیط معنایی خاص را نشان می‌دهد. بنابر این می‌توان ادعا کرد که واژه‌ها یا عبارت‌های هم‌معنی ممکن است به دلیل تفاوت در نوای معنایی خاص خود از یکدیگر متمایز شوند. در پیوند با این ادعا، لودلینگ و کیتو (۲۰۰۹: ۹۸۰) به نقل از پارتینگتون[1] (۱۹۹۸) با بررسی پیکره‌بنیادِ واژه‌های هم‌معنی «sheer» [مطلق]،« pure»[خالص]،«absolute » [مطلق]، «complete»[کامل]، نشان می‌دهد که هر واژه در الگوی عبارتی[2] خاص خود به کار می‌رود؛ به عبارت دیگر هر واحد واژگانی الگوی رفتاری خاص و منحصر به خود را دارد. بنابر این حتی اگر باهم‌آیندهای دو واژه یا عبارت هم‌معنی از ویژگی‌های معنایی مشترک برخوردار باشد، نوای معنایی متفاوت این واژه‌ها یا عبارت‌ها باعث تمایز آنها خواهد شد.

در بررسی صورت‌های هم‌معنی مانند «منجر به ... شدن»، «باعث ... شدن»، «موجب ... شدن» و «سبب ... شدن»، نشان داده شد که تفاوت چنین ساختارهایی نه در نوع باهم‌آیندها بلکه در نوای معنایی این ساختارها است(مدرس خیابانی، ۱۳۹۰). در جدول ۸ برخی باهم‌آیندهای پربسامد مشابه این نمونه‌ها ارائه شده است[3]. همانگونه که ملاحظه می‌شود در این نمونه‌ها باهم‌آیندهایی مشابه با مقدار زیاد تی و اطلاعات دوسویه وجود دارد و شاید نتوان صرفاً مبتنی بر نوع باهم‌آیندها تفاوت چنین ساختارهایی را نشان داد.

نوع اول	نوع دوم	فراوانی نوع اول	فراوانی نوع دوم	فراوانی توالی نوع اول و دوم	اطلاعات دوسویه	مقدار تی
موجب	نگرانی	۳۴۰۷	۷۶۵	۳۳	۷,۰۰	۵,۷۰
باعث	نگرانی	۲۸۲۷	۷۶۵	۲۴	۶,۸۱	۴,۸۶

[1]- Partington, A.

[2]- phraseological pattern

[3] – در دادگان مورد بررسی در ساخت «منجر به ... شدن» پس از واژهٔ «منجر» صورت «به» یک واژهٔ مستقل محاسبه شده و در نتیجه صورت «منجر به» یک دونگاشتی یعنی دو واژهٔ پیاپی محسوب شده و با حضور واژهٔ بعدی یک سه‌نگاشتی یعنی توالی سه واژه شکل می‌گیرد. از آنجا که اطلاعات استخراج شده تنها شامل دونگاشتی‌ها است، میزان تی و اطلاعات دوسویه در توالی‌هایی مانند «منجر به فوت» محاسبه نشده و به همین دلیل است که این ساخت در جدول ۸ ارائه نشده است.

نوع اول	نوع دوم	فراوانی نوع اول	فراوانی نوع دوم	فراوانی توالی نوع اول و دوم	اطلاعات دوسویه	مقدار تی
باعث	مرگ	۲۸۲۷	۱۹۸۸	۴۳	۶,۲۸	۶,۴۷
موجب	مرگ	۳۴۰۷	۱۹۸۸	۳۶	۵,۷۵	۵,۸۹
باعث	کاهش	۲۸۲۷	۴۲۳۹	۸۳	۶,۱۳	۸,۹۸
موجب	کاهش	۳۴۰۷	۴۲۳۹	۸۰	۵,۱	۸,۷۸
سبب	کاهش	۱۵۵۱	۴۲۳۹	۳۱	۵,۵۸	۵,۴۵
موجب	شادی	۳۴۰۷	۳۸۷	۴۶	۸,۴۷	۶,۷۶
باعث	شادی	۲۸۲۷	۳۸۷	۲۱	۷,۶۰	۴,۵۶
باعث	رشد	۲۸۲۷	۳۳۹۱	۳۰	۴,۹۹	۵,۳۰
موجب	رشد	۳۴۰۷	۳۳۹۱	۲۰	۴,۱۳	۴,۲۲
سبب	رشد	۱۵۵۱	۳۳۹۱	۶	۳,۵۳	۲,۲۴
موجب	تقویت	۳۴۰۷	۱۸۳۳	۴۱	۶,۰۶	۶,۳۱
باعث	تقویت	۲۸۲۷	۱۸۳۳	۱۵	۴,۸۷	۳,۷۴
سبب	تقویت	۱۵۵۱	۱۸۳۳	۶	۴,۴۲	۲,۳۳
باعث	افزایش	۲۸۲۷	۷۰۳۵	۱۱۵	۵,۸۷	۱۰,۵۴
موجب	افزایش	۳۴۰۷	۷۰۳۵	۹۹	۵,۳۹	۹,۷۱
سبب	افزایش	۱۵۵۱	۷,۳۵	۲	۴,۸۹	۵,۴۷

جدول ۸- باهم‌آیندهای مشابه در صورت‌های «باعث ... شدن»، «موجب ... شدن» و «سبب ... شدن»

بررسی حدود ۵۰۰ خط واژه‌نما یا نمونه‌جمله از این سه ساخت، نشان می‌دهد که گرچه به لحاظ نوع صورت‌های باهم‌آیند، شباهت زیادی میان این نمونه‌ها وجود دارد، نوای معنایی «منجر به ... شدن» منفی‌تر از سایر موارد است؛ یعنی در جملاتی با بار عاطفی منفی، بیشتر از این نمونه نسبت به سایر موارد استفاده می‌شود. از سوی دیگر نوای معنایی «سبب ... شدن» مثبت‌تر و نوای معنایی «باعث ... شدن» خنثی‌تر از سایر نمونه‌هاست(مدرس خیابانی،۱۳۹۰).

۶- نتیجه‌گیری

حاصل پژوهش حاضر را می‌توان به شرح زیر خلاصه کرد:

۱- با استفاده از ابزار آماری همچون آزمون اطلاعات دوسویه و تی، می‌توان باهم‌آیندهای یک واژه را از پیکره‌های زبانی استخراج کرد.

۲- به‌کمک ابزار آماری امکان متمایز ساختن برخی صورت‌های هم‌معنی مبتنی بر باهم‌آیندهای آنها امکان‌پذیر است.

۳- در شرایطی که بیشتر باهم‌آیندهای دو یا چند صورت هم‌معنی، یکسان باشد، نوای معنایی این صورت‌ها می‌تواند تمایز میان آنها را نمایان سازد.

۴- شناسایی تمایز میان واژه‌های هم معنی در مباحثی مانند آموزش زبان خارجی، فرهنگ‌نگاری و ترجمه از اهمیتی بسیار برخوردار است. بدیهی است با آگاهی از تمایز صورت‌های هم‌معنی نه تنها ساخت دستوری بلکه ساخت طبیعی زبان تولید می‌شود. برای نمونه گرچه ساختی مانند «درخت پیر» ساختی غیر دستوری محسوب نمی‌شود، ساختی مانند «درخت کهنسال» بسیار طبیعی‌تر از آن می‌نماید.

۵- گرچه شمّ زبانی تحلیل‌گر در بررسی‌های زبان‌شناختی از ارزشی ویژه برخوردار است، پژوهش حاضر نشان می‌دهد که نمی‌توان اهمیت پیکره‌های زبانی و استفاده از آمار در مطالعات زبان‌شناختی را نادیده گرفت.

منابع

- صفوی، کورش(۱۳۷۹). *درآمدی بر معنی شناسی*، تهران، پژوهشگاه فرهنگ و هنر اسلامی.
- مدرس خیابانی، شهرام(۱۳۹۰). «نوای معنایی در صورت‌های هم‌معنی؛ رویکردی پیکره‌بنیاد» مقالهٔ چاپ نشده. ارائهٔ سخنرانی در دومین کارگاه آموزشی و پژوهشی معنی‌شناسی، انجمن زبان‌شناسی ایران، تهران، پژوهشگاه علوم انسانی و مطالعات فرهنگی.

- Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton & Co.
- Church, K. and P. Hanks(1989 [1990]). "Word Association Norms, Mutual Information and Lexicography", in *Computational Linguistics*, 16: 22-29.
- Church, K. and W. Gale, P. Hanks and D. Hindle(1991) "Using Statistics in Lexical Analysis", in U. Zernic (ed) *Lexical Acquisition*, Englewood Cliff , NJ: Erlbaum.

- Clear, J. (1993). “From Firth Principles – Computational Tools for Study of Collocation”, in M. Baker, G . Francis, and E. Tognini-Bonelli (eds), *Text and Technology*: In Honor of John Sinclair, pp. 271-292, Amsterdam: Benjamins.
- Fano, R. M. (1961). *Transmission of Information; A Statistical Theory of Communications*. New York: MIT Press.
- Firth, J. R. (1957). “Modes of Meaning”, in J. R. Firth, *Papers in Linguistics*: 1934-51, pp. 190-215, London: Oxford University Press.
- Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London and New York: Longman.
- Lüdeling, A and M. Kytö (2009). *Corpus Linguistics: An International Handbook, Volume 2*. Berlin, New York: Walter de Gruyter.
- Lyons, J. (1995). *Linguistic Semantics :An Introduction* .Cambridge :Cambridge University Press.
- Manning,C. D. and H. Schutze(1999). *Foundations of Statistical Natural Language Processin*. MIT Press, Cambridge, MA.
- Partington, A. (1998). *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.
- Sinclair, J. (1991). *Corpus, Corncordanee, Collocation*. Oxford: Oxford University Press.
- Sinclair, J. (1996) “The Search for Units of Meaning”. In Texas 9(1), 74-106.
- Sinclair, J. (2004) *Trust the Text: Language, Corpus and Discourse*. London: Routledge.
- Stubbs, M. (1995) “Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Studies”, in *Functions of Language* 2,1: 1-25 <http:www.benjamins.com> 18.08.2005