بررسی تکوین مباحث مقالات فارسی‌زبان و زبان‌شناسی با کمک رایانه

مسعود قیومی[1] (استادیار زبان‌شناسی، پژوهشگاه علوم انسانی و مطالعات فرهنگی)

تاریخ دریافت مقاله: ۱۴۰۱/۳/۱۶، تاریخ پذیرش: ۱۴۰۱/۰۵/۲۱، تاریخ انتشار: تابستان ۱۴۰۱

چکیده: از زمان پیدایش اینترنت تاکنون، با حجم زیادی از داده‌هایی مواجه هستیم که در این بستر قرار گرفته است. این امر سبب شده است تا شیوهٔ مطالعهٔ داده‌ها و بررسی سیر تحول آنها از روش دستی به خودکار تغییر کند. هدف از انجام این پژوهش، استخراج موضوعات مطرح‌شده در مقالات رشتهٔ زبان‌شناسی و روندشناسی تحول موضوعات در طول زمان با کمک الگوریتم مدل‌سازی موضوعی است. برای این هدف از الگوریتم تخصیص پنهان دیریشله استفاده می‌شود. پیکرهٔ این پژوهش ازطریق خزش به‌دست آمده و پس‌از پالایش و پیش‌پردازش داده‌ها، تعداد ۵، ۱۰ و ۱۵ موضوع از مقالات استخراج شده و براساس واژه‌های هر موضوع به‌صورت دستی برچسب‌گذاری شده است. توزیع مقالات از سال ۱۳۰۶ تا ۱۳۹۹ سبب شد تا این مدت به ۵ مقطع زمانی تقسیم و موضوعات مربوط به هر برش زمانی پس‌از فرایند برچسب‌گذاری مشخص شود. رشد و افول موضوعات استخراج‌شده از مقالات در پردازش‌های با تعداد ۵، ۱۰ و ۱۵ موضوع در بازه‌های زمانی قابل مشاهده است. دستاورد کاربردی این پژوهش سیاست‌گذاری در حوزهٔ علم است که علاوه‌بر مطرح‌کردن یک روش‌شناسی فناورانهٔ کاربردی در پژوهش، می‌توان موضوعات داغ میان پژوهشگران یک رشتهٔ علمی را مشخص کرد و خلأهای موضوعات پژوهشی را یافت و بر متنوع‌سازی و متوازن‌سازی موضوعات پژوهشی اهتمام ورزید.

کلیدواژه‌ها: تکوین، پردازش زبان طبیعی، مدل‌سازی موضوع، زبان‌شناسی پیکره‌ای، مقاله علمی، تحلیل محتوایی.

1. m.ghayoomi@ihcs.ac.ir

۱ مقدمه

هدف اصلی علوم انسانی دیجیتال، ایجاد داده، مدیریت داده و کاربرد داده است (هوگس[1]، ۲۰۱۵). با پیدایش اینترنت و قرارگرفتن اطلاعات در این بستر، با گذشت زمان با حجم زیادی از داده‌ها مواجه هستیم. ازاین‌رو، با گذشت زمان و داده‌های جمع‌آوری‌شده می‌توان به روند تحول موضوعات مختلف ازنظر آماری پرداخت. در این شیوهٔ بررسی سیر تحول به‌صورت دستی امکان‌پذیر نیست و این امر می‌تواند با به‌کارگیری الگوریتم‌های پردازش داده و هوش مصنوعی تحقق یابد. تحلیل داده‌های پردازش‌شده می‌تواند به ارزش افزوده‌ای منجر شود که این ارزش افزوده باتوجه به نوع تحلیل می‌تواند به‌صورت عملی در تصمیم‌گیری‌های کلان و مسئله‌مندی در پژوهش مورد استفاده قرار گیرد.

در حوزهٔ رایانه، داده‌ها به چهار دسته تقسیم می‌شود: داده‌های متنی، تصویری، صوتی و عددی. این تنوع داده را می‌توان در داده‌های موجود در وب مشاهده کرد که پردازش هرکدام از دسته‌ها به الگوریتم‌های مختص به خود نیاز دارد. ازمیان این داده‌ها، حجم زیادی از داده‌ها به‌صورت دادهٔ متنی موجود است. داده‌های متنی نیز در دامنه‌های مختلفی موجود هستند، مانند متن خبری، ادبی، علمی و مانند آن. امروزه حجم زیادی از مقالات علمی توسط پژوهشگران منتشر می‌شود و معمولاً در محیط وب دردسترس است. با گردآوری این مقالات می‌توان به تهیهٔ پیکرهٔ تخصصی از مقالات علمی اقدام کرد. شایان ذکر است که با وجود انواع رشته‌های علمی، پیکرهٔ تخصصی علمی می‌تواند محدود به یک رشته علمی ازمیان علوم تجربی، پزشکی یا انسانی یا محدود به یک حوزه در یک رشته، مانند علوم اجتماعی، علوم سیاسی، علوم کتابداری، زبان و زبان‌شناسی و ادبیات و غیره در علوم انسانی باشد.

هدف از انجام این پژوهش، روندشناسی و بررسی سیر تحول موضوعات مطالعه‌شده در حوزهٔ زبان‌شناسی به‌صورت الگوریتمی است. در این پژوهش تلاش می‌شود تا ضمن تهیهٔ یک پیکرهٔ تخصصی از چکیده‌های متون علمی، مقالات حوزهٔ زبان‌شناسی به‌صورت الگوریتمی پردازش و سیر تحول موضوعات بررسی شود. شایان ذکر است که این پژوهش به جریان‌شناسی و بررسی دلایل سیر تحولات نمی‌پردازد و خارج از موضوع روندشناسی است.

ساختار مقالهٔ حاضر به این شرح است: پس از مقدمه، در بخش ۲ به پیشینهٔ مطالعات

1. L. Hughes

انجام‌شدهٔ مرتبط با روندشناسی پرداخته می‌شود. در بخش ۳ الگوریتم استخراج موضوع پنهان از متن توضیح داده می‌شود. در بخش ۴ پیکرهٔ تهیه‌شده از چکیدهٔ مقالات فارسی در حوزهٔ زبان‌شناسی توصیف می‌شود. بخش ۵ به تحلیل الگوریتمی موضوعی می‌پردازد؛ و در انتها مقاله با نتیجه‌گیری در بخش ۶ به پایان می‌رسد.

۲ پیشینهٔ مطالعاتی

روندشناسی رویدادها، چه رویدادهای اجتماعی و چه علمی، در طی گذشت زمان یکی از مسائل مورد توجه پژوهشگران بوده‌است. از دیدگاه روندشناسی مقالات علمی، مطالعات متنوعی انجام گرفته‌است که ابتدا به مطالعات انجام‌شده روی مقالات به زبان انگلیسی و سپس مقالات فارسی می‌پردازیم.

بلای[1] و لافرتی[2] (۲۰۰۶) مدلی پردازشی پیشنهاد داده‌اند که کار خوشه‌بندی موضوعی مقالات را باتوجه‌به مؤلفهٔ زمان انجام می‌دهد. در این پژوهش سی‌هزار مقاله از مجلهٔ Science برای مدت ۱۲۰ سال در بازهٔ زمانی ۱۸۸۱ تا ۱۹۹۹ از بایگانی JSTOR گردآوری و به‌طور رایانشی بررسی شده‌اند. ازآنجاکه این داده‌ها به‌صورت متنی موجود نبود، تلاش شده است تا با استفاده از تشخیص نوری حروف[3] به پیکرهٔ متنی تبدیل گردد. در فرایند پیش‌پردازش، ستاک واژه‌ها به‌دست آمده و واژه‌های با بسامد کمتر از ۲۵ از متن حذف شده است. سپس، مقالات براساس موضوع خوشه‌بندی شده و پربسامدترین واژه‌های مربوط به هر موضوع در دهه‌های مختلف استخراج شده است.

ونگ[4] و مک‌کالوم[5] (۲۰۰۶) به بررسی موضوعی مقالات در طول زمان پرداخته‌اند. در این پژوهش از الگوریتم بهبودیافته با ارائهٔ یک الگوریتم «تخصیص دریشله پنهان[6]» استفاده شده است. داده‌های این پژوهش بایگانی ۹ ماههٔ ایمیل شخصی، ۱۷ سال مقالات نیپس و بیش از ۲۰۰ سال خطابه‌های ایالتی رئیس‌جمهور بوده است.

ونگ[7] و همکاران (۲۰۰۸) خوشه‌بندی موضوعی مقالات باتوجه‌به مؤلفه زمان را به دو دسته

1. D. M. Blei
2. J. D. Lafferty
3. Optical Character Recognition (OCR)
4. X. Wang
5. A. McCallum
6. Latent Dirichlet Allocation
7. C. Wang

تقسیم کرده‌اند. یکی زمان ناپیوسته است و دیگری زمان پیوسته و پژوهش خود را در زمان پیوسته انجام داده‌اند. در این پژوهش از دو دادهٔ خبری استفاده شده‌است که یکی حاوی ۱۳۴۲ خبر در بازهٔ ۱۹۸۸/۵/۱ تا ۱۹۸۸/۶/۳۰ است و دیگری اخبار مناظرهٔ ریاست جمهوری سال ۲۰۰۸ است که در بازهٔ زمانی ۲۰۰۷/۲/۲۷ تا ۲۰۰۸/۲/۲۲ منتشر شده‌است.

ژو[1] و همکاران (۲۰۱۶) به بررسی تکوین موضوعات در مجلات مرتبط با علم اطلاعات در طول زمان پرداخته‌اند. در این پژوهش، ابتدا مقالات چندین مجلهٔ چینی مرتبط با حوزهٔ فناوری اطلاعات در بازهٔ زمانی سال‌های ۲۰۰۰ تا ۲۰۱۵ خزش شده و یک پیکرهٔ زبانی متشکل از ۲۹۵۵۲ مقاله به‌دست آمده‌است. پس‌از پیش‌پردازش و هنجارسازی داده‌ها، تا ۱۰۰ موضوع از داده‌ها استخراج شد و در هر مرحله سرگشتگی[2] محاسبه شد. براساس متوازن‌شدن سرگشتگی و تعداد موضوعات به‌عنوان معیار، تعداد ۳۵ موضوع مشخص شده و توزیع آماری مستندات این موضوع در مقاطع زمانی مختلف به‌دست آمده است. براساس نتایج، موضوعات فناوری بازیابی ادبیات و همچنین شبکه و ساختار اطلاعات کتابخانه رو به افول بوده و موضوعاتی مانند حمایت از مالکیت معنوی مورد توجه قرار گرفته است. تحلیل‌های معنایی و احساسی ازجمله دیگر موضوعات پژوهشی دیگری بوده است که مورد توجه قرار گرفته است.

زوسا[3] و گرانروث‌ویلدینگ[4] (۲۰۱۹) در پژوهش خود مدل‌سازی موضوع پویای چندزبانه‌ای را پیشنهاد داده‌اند که از ویژگی‌های بین‌زبانی یک پیکرهٔ خبریِ دوزبانه آلمانی-انگلیسی و همچنین پیکرهٔ خبری تطبیقی فنلاندی و سوئدی استفاده می‌کنند.

اگرچه پژوهش‌های متعددی، مانند افشارنیا و اللهیاری فرد (۱۳۸۵) و سواری و بهمنی (۱۳۸۹)، در حوزهٔ تبیین جایگاه علم و فناوری و ترسیم مسیر رشد و بهبود نظام علم و فناوری در ایران و آسیب‌شناسی در حوزهٔ علم انجام پذیرفته است، وضعیت علم زبان‌شناسی در کشور نیز مستثنی نبوده و مورد بررسی قرار گرفته است. یارمحمدی و همکاران (۱۳۷۷؛ ۱۳۷۸) در گزارشی به بررسی وضعیت علم زبان‌شناسی در ایران پرداخته‌اند. در این پژوهش مجموعه‌ای از کتاب‌های تألیف یا ترجمه، مقالات، پایان‌نامه‌های کارشناسی ارشد و دکتری را به‌همراه اطلاعات

1. M. Zhu
2. perplexity
3. E. Zosa
4. M. Granroth-Wilding

کتابخانه‌ای‌شان به‌صورت دستی گردآوری و از نظر کمّی و کیفی تحلیل شده‌اند. در این تحلیل، کار مقوله‌بندی موضوعی منابع انجام پذیرفته و آمار مربوطه استخراج شده است. این موضوعات در دو سطح زبان‌شناسی نظری و کاربردی عبارتند از: صوت‌شناسی، ساخت واژه، نحو، معناشناسی، کاربردشناسی و گفتمان، گویش‌شناسی، کاربردهای ادبی، فرهنگ‌نگاری، عصب‌شناسی زبان، زبان‌شناسی مقابله‌ای، زبان‌شناسی عمومی، زبان‌شناسی رایانشی، زبان‌شناسی تاریخی-تطبیقی، زبان و منطق، روان‌شناسی، جامعه‌شناسی زبان، آموزش زبان و اصول و روش ترجمه. براساس بررسی منابع، این نتایج به‌دست آمده است که موضوعات زبان‌شناسی عمومی، گویش‌شناسی، اصول و روش ترجمه و آموزش زبان مطالب بیشتری را مورد توجه قرار داده‌اند و موضوعاتی مانند معناشناسی، زبان‌شناسی تاریخی-تطبیقی، زبان‌شناسی مقابله‌ای و زبان‌شناسی رایانشی حاوی آثار محدودی هستند.

ناصح (۱۳۸۰؛ ۱۳۸۳؛ ۱۳۸۶الف؛ ۱۳۸۶ب) صرفاً اقدام به گردآوری منابع مربوط به زبان‌شناسی و موضوعات مربوط به آنها مانند گویش‌شناسی، دستور زبان، زبان‌شناسی عمومی و ادبیات پرداخته است و تحلیلی از تکوین موضوعات ارائه‌شده در منابع گردآوری‌شد ننموده است.
احدی (۱۴۰۰) به‌صورت نظام‌مند مطالعات انجام‌شده در حوزهٔ زبان‌شناسی بالینی و مشکلات کودکان دارای اختلالات رشدی، اعم از کودکان دچار اوتیسم، کم‌توان ذهنی و نارساخوان، را ازنظر موضوعی به‌صورت دستی در بایگانی‌های مختلف مقالات، مانند پرتال جامع علوم انسانی، نورمگز، مگیران، ایرانداک و مرکز علمی جهاد دانشگاهی، بررسی کرده و سپس به فراتحلیل این مطالعات پرداخته است.

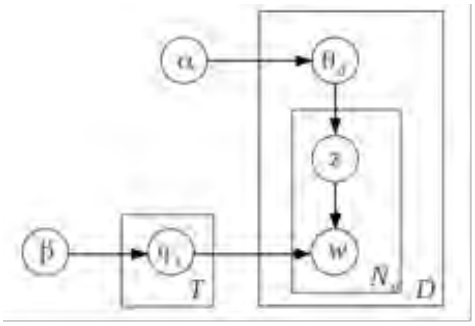
۳ مدل‌سازی موضوع

متن عنوان پر واضح است که زبان ابزار ارتباطی است که در یک کانال ارتباطی تفکر، ایده یا نظر مشخصی از فرد گوینده به دریافت‌کننده منتقل می‌شود. اگرچه این تفکر، ایده یا نظر یک کلیت واحد در قالب یک متن دارد، ولی موضوعات مختلفی می‌توانند در شکل‌گیری آن متن سهیم باشند. «مدل‌سازی موضوع»[1] یکی‌از الگوریتم‌هایی است که برای تحلیل محتوایی متن به‌کار می‌رود. این الگوریتم که توسط پاپادیمیتریو[2] و همکاران (۲۰۰۰) معرفی شده‌است به‌دنبال کشف

1. topic modeling
2. C. P. Papadimitriou

ساختار معنایی پنهان در متن است. شاکلهٔ کلی این الگوریتم این است که هر متن از چند موضـوع انتزاعی تشکیل شده و هر موضوع انتزاعی نیز از تعدادی واژه‌هـای خاصـی متشـکل شده‌اسـت. بـا یافتن این واژه‌ها می‌توان متن را براساس موضوع‌های انتزاعی خوشه‌بندی[1] نمود. برای مثال موضـوع «نحو» در زبان‌شناسی با واژه‌هایی مانند «هسته»، «سـازه»، «گـروه»، «سـاخت» و ماننـد آن بیـان می‌شود و موضوعی مانند آواشناسی با واژه‌هایی مانند «آوا»، «آوانگاری»، «تلفظ» و مانند آن.

پاپادیمیتریو و همکارانش «مدل‌سازی موضوع» را براسـاس جبـر خطـی پیشـنهاد دادنـد، ولـی هوفمن[2] (۱۹۹۹) برای این هدف یک مدل احتمالاتی ارائه کـرد. بـلای و همکـاران (۲۰۰۳) ایـن مدل احتمالاتی را تعمیم دادند و مدلی از الگوریتم «مدل‌سازی موضوع» که به «تخصیص دریشله پنهان» معروف است را معرفی کردند. الگوریتم معرفی‌شده یک مدل آمـاری زایشـی اسـت کـه در اصل «توزیع اولیه دریشله»[3] بوده و برای توزیع احتمـالی مـتن ـ موضـوع و موضـوع ـ واژه بـه‌کار می‌رود. در شکل (۱) مدل «تخصیص دریشله پنهان» نشان داده شده‌است که از دو مـاتریس φ و θ تشکیل یافته‌است. در این مدل φ ماتریسی است که توزیع موضوع T برروی واژه‌های W را براساس توزیع اولیه دریشله با پارامتر β بیان می‌کند. θ ماتریسی است که توزیع متن d برروی موضـوع‌های T را براساس توزیع اولیه دریشله با پارامتر α بیان می‌کند. برای زایش هر واحد واژگـانی w در مـتن d، یک موضوع z از توزیع موضوعی مربوط به متن θ_d به‌دست می‌آید؛ درحالی‌که خود آن واژه w از توزیع واژه‌های موضوع انتخاب‌شدهٔ φ_z به‌دست می‌آید.



شکل ۱ - نمایش تصویری مدل «تخصیص دریشله پنهان»

1. cluster
2. T. Hofmann
3. Dirichlet prior distribution

برای استخراج موضوعات با مدل تخصیص دریشله پنهان نیاز است دو توزیع φ و θ تخمین زده شوند تا اطلاعات دربارهٔ توزیع متن نسبت‌به موضوعات و موضوعات نسبت‌به متن به‌دست آید. برای تخمین این دو، الگوریتم‌های مختلفی، مانند انتشار انتظار (مینکا[1] و لافرتی، ۲۰۰۲)، استنباط تغییرات (بلای و همکاران، ۲۰۰۳) یا نمونه‌گیری گیبس[2] (گریفیث[3] و استیورس[4]، ۲۰۰۴)، پیشنهاد شده‌است که از میان این الگوریتم‌ها، نمونه‌گیری گیبس به‌عنوان یک رویکرد ساده و مؤثر برای مدل‌سازی موضوع استفاده می‌شود. در نمونه‌گیری گیبس، احتمال انتخاب یک موضوع برای یک واژه در یک متن به واژهٔ قبل و دو واژهٔ قبل در بافت و موضوعاتی که به آن موضوعات تخصیص داده می‌شود مشروط شده است که با استفاده از تساوی (۱) محاسبه می‌شود:

(۱)

$$P(z_i = t|w_i = w, z_{-i}, w_{-i}) = \frac{N^{WT}_{wt,-i}+\beta}{\sum_{w'} N^{WT}_{w't,-i}+W\beta} \times \frac{N^{TD}_{td,-i}+\alpha}{\sum_{t'} N^{TD}_{t'd,-i}+T\alpha}$$

در این تساوی، $w_i = w$ نشان می‌دهد iامین واژه در متن واژه w است و $z_i = t$ نشان می‌دهد که واژه w به موضوع t تخصیص داده شده‌است. w_{-i} و z_{-i} بیانگر تمام واژه‌ها و تمام موضوعات تخصیص‌داده‌شده به‌جز واژه iامین است. $N^{WT}_{wt,-i}$ تعداد دفعاتی است که واژهٔ w، به‌جز درنظرگرفتن آمار واژهٔ کنونی، به موضوع t تخصیص داده شده است. $N^{TD}_{td,-i}$ تعداد دفعاتی است که موضوع t، به‌جز موضوع کنونی، به متن d تخصیص داده شده است. با استفاده از نمونه‌گیری گیبس، برای هریک از نمونه‌های این مدل، φ و θ در تساوی (۲) و (۳) محاسبه می‌شوند:

(۲)

$$\emptyset_{wt} = \frac{N^{WT}_{wt}+\beta}{\sum_{w'} N^{WT}_{w't}+W\beta}$$

(۳)

$$\theta_{td} = \frac{N^{TD}_{td}+\alpha}{\sum_{t'} N^{TD}_{t'd}+T\alpha}$$

در این تساوی‌ها، φ_{wt} احتمال کاربرد واژه w در موضوع t است و θ_{td} احتمال کاربرد موضوع t در

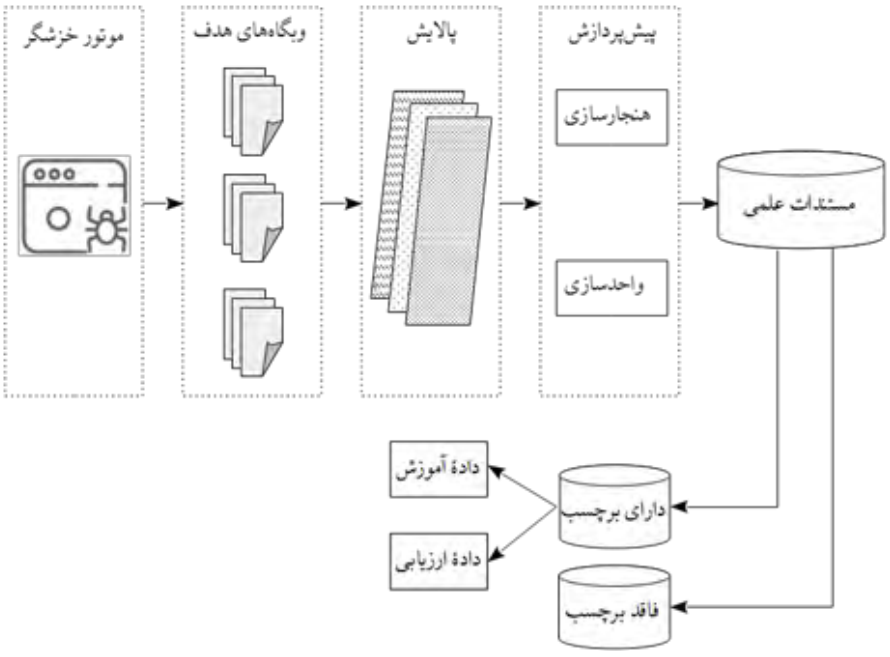
1. T. Minka
2. Gibbs sampling
3. T. L. Griffiths
4. M. Steyvers

متن d است.

در پژوهش حاضر تلاش می‌شود ساختار معنایی نهفته در چکیده‌های مقالات در حوزهٔ زبان‌شناسی با استفاده از الگوریتم «مدل‌سازی موضوع» به‌دست آید و پس‌از این تحلیل اولیه خودکار، به فراتحلیل آن توسط انسان خبره پرداخته شود.

۴ داده‌های پژوهش

برای انجام این پژوهش به پیکرهٔ تخصصی حاصل از مقالات و مستندات علمی نیاز است تا بتوان موضوعات مطرح‌شده در آنها را از لابه‌لای مطالب یافت و براساس موضوع قالب مستندات خوشه‌بندی کرد. اگرچه در مطالعات گذشته، مانند کامیابی‌گل و همکاران (۱۳۹۷) و علایی و همکاران (۱۴۰۰)، کار گردآوری داده انجام شده و براساس نیازهای خود برچسب‌گذاری شده‌است، پیکرهٔ تمام متن که مورد نیاز الگوریتم مدل‌سازی موضوع است دردسترس عموم قرار ندارد؛ بنابراین نیاز است پیکرهٔ مورد نیاز پژوهش تهیه شود. در شکل (۲)، فرایند تهیهٔ داده‌های پژوهش حاضر ترسیم شده‌است. ابتدا نیاز به گردآوری و سپس آماده‌سازی داده داریم که در ادامه توضیح داده خواهد شد.



شکل ۲- فرایند تهیهٔ داده‌های پژوهش

۴-۱ گردآوری داده

برای گردآوری داده‌های این پیکره، از روش خزش وبگاه‌هایی که حاوی بایگانی مقالات علمی است، مانند بایگانی پرتال جامع علوم انسانی[1]، نسخهٔ پیشین پایگاه اطلاعات علمی ایران (گنج)[2] و نورمگز[3]، استفاده شده است. ازآنجاکه چکیدهٔ مقالات به‌صورت متنی در این بایگانی‌ها وجود دارد، فرایند خزش را به چکیدهٔ مقالات و نه متن کامل مقالات محدود می‌کنیم. در جدول (۱) داده‌های گردآوری‌شده از وبگاه‌های مختلف گزارش شده‌است. در این جدول وضعیت بایگانی مستندات علمی تقریباً یک قرن، از ۱۲۸۵ تا ۱۳۹۹، مشخص است. پرتال جامع علوم انسانی حاوی چکیدهٔ ۳۲۶،۳۴۵ مقاله در بازهٔ زمانی ۱۲۸۵ تا ۱۳۹۹، پایگاه اطلاعات علمی ایران حاوی ۷۴۵،۲۵۵ چکیدهٔ مقاله، پایان‌نامه کارشناسی ارشد و رسالهٔ دکتری در بازهٔ زمانی ۱۳۰۵ تا ۱۳۹۹ و نورمگز نیز حاوی ۷۵۸،۹۶۵ چکیدهٔ مقاله در بازهٔ زمانی ۱۳۰۰ تا ۱۳۹۹ است. در این فرایند خزش حدود دو میلیون سند علمی گردآوری شده‌است. آنچه از آمار تعداد مستندات علمی وبگاه‌ها مشخص است، افزایش چشمگیر تعداد اسناد علمی در دو دههٔ متأخر ۱۳۸۰ و ۱۳۹۰ است که بیانگر توجه جامعهٔ علمی کشور به بیان دستاورهای علمی و پژوهشی خود ازطریق انتشار مقاله، پایان‌نامه و رساله است.

جدول ۱ - بازه‌های زمانی اسناد موجود در مستندات علمی خزش‌شده

بازه زمانی	پرتال جامع علوم انسانی	پایگاه اطلاعات علمی ایران	نورمگز
۱۲۸۵-۱۲۸۱	۱۴	۰	۰
۱۲۹۰-۱۲۸۶	۱۵۲	۰	۰
۱۲۹۵-۱۲۹۱	۰	۰	۰
۱۳۰۰-۱۲۹۶	۱۵۸	۰	۵۶
۱۳۰۵-۱۳۰۱	۸۷۲	۲	۱۵۶۹
۱۳۱۰-۱۳۰۶	۸۱۸	۱	۲۸۷۸
۱۳۱۵-۱۳۱۱	۱۲۴۰	۲۳	۵۷۶۳
۱۳۲۰-۱۳۱۶	۱۷۳۸	۴۶۸	۶۵۹۱
۱۳۲۵-۱۳۲۱	۷۸۲	۵۲۳	۷۵۲۶
۱۳۳۰-۱۳۲۶	۱۶۲۸	۷۳۸	۱۱۶۴۵

1. http://www.ensani.ir
2. https://ganj-old.irandoc.ac.ir
3. https://www.noormags.ir/

۱۳۳۵-۱۳۳۱	۱۵۳۶	۱۶۷۳	۶۵۴۵
۱۳۴۰-۱۳۳۶	۳۲۸۵	۲۴۸۳	۹۳۷۹
۱۳۴۵-۱۳۴۱	۴۵۲۱	۳۷۰۳	۲۱۴۱۹
۱۳۵۰-۱۳۴۶	۵۲۶۲	۵۹۰۹	۲۸۸۶۵
۱۳۵۵-۱۳۵۱	۶۴۶۰	۷۴۶۸	۳۰۷۳۷
۱۳۶۰-۱۳۵۶	۳۷۳۳	۷۲۶۱	۲۶۹۰۴
۱۳۶۵-۱۳۶۱	۴۰۴۶	۱۷۱۵۷	۲۵۱۵۵
۱۳۷۰-۱۳۶۶	۸۲۱۶	۲۱۸۹۵	۳۷۵۸۹
۱۳۷۵-۱۳۷۱	۱۶۲۱۲	۵۳۲۵۳	۵۱۲۷۷
۱۳۸۰-۱۳۷۶	۳۵۲۰۵	۶۵۷۵۸	۸۶۹۵۱
۱۳۸۵-۱۳۸۱	۶۰۶۳۹	۸۷۳۸۹	۱۳۵۲۵۶
۱۳۹۰-۱۳۸۶	۷۷۹۷۴	۱۵۳۹۲۴	۱۵۵۲۹۹
۱۳۹۵-۱۳۹۱	۵۸۰۷۲	۳۰۷۰۲۰	۱۰۱۰۸۸
۱۳۹۹-۱۳۹۶	۳۳۷۸۲	۸۶۰۷	۶۴۷۳

پس‌از بررسی اولیهٔ داده‌های حاصل از فرایند خزش، با چند نکتهٔ قابل توجه مواجه شدیم:

۱. بعضی از مستندات سال انتشار مشخص نداشت. این موارد از تعداد کل اسناد خزش‌شده کنار گذاشته شده‌است.

۲. بعضی از اسناد چکیده نداشت. دو دلیل عمده برای نداشتن چکیده یافت شد. یکی این که ساختار مقالات قدیمی با ساختار مقالات امروزی متفاوت بوده و در این دسته از مقالات چیزی به‌عنوان چکیده وجود ندارد. دلیل دیگر این که اگرچه برای بعضی از مقالات چکیده در فایل PDF موجود بود، به‌دلیل نبود چکیده در صفحهٔ مربوط به سند علمی به‌صورت متن حروفچینی‌شده، امکان خزش داده میسر نشد.

۳. در داده‌های خزش‌شده از منابع مختلف، اسناد علمی به زبان‌هایی جز فارسی چون عربی، انگلیسی، آلمانی، روسی و غیره موجود است که مستندات علمی مربوط به این موارد در پژوهش حاضر قابل استفاده نیست.

٤. مقالات خزش‌شده از پرتال جامع علوم انسانی حاوی برچسب‌هایی است که نوع حوزهٔ آن سند علمی براساس یک مجموعهٔ برچسب ۱۶گانه، مانند اقتصاد، ادبیات، زبان‌شناسی،

علوم سیاسی و غیره، مشخص شده‌است. درحالی‌که هر دو منبع پایگاه اطلاعات علمی ایران و نورمگز فاقد چنین برچسب‌های محتوایی است.

۵. مستندات علمی پرتال جامع علوم انسانی همگی در حوزهٔ علوم انسانی متمرکز هستند؛ درحالی‌که پایگاه اطلاعات علمی ایران و نورمگز، علاوه‌بر مستندات در حوزهٔ علوم انسانی، حاوی مستندات علمی مربوط به حوزه‌های علوم فنی-مهندسی و پزشکی نیز هستند.

٦. باتوجه‌به این که وبگاه‌های هدف هریک جداگانه تلاش کرده‌است در حداکثر امکان مستندات علمی را در خود ذخیره نماید، امکان تکرار در مستندات وجود دارد.

داده‌هایی که در این پژوهش نیاز داریم مستندات علمی به زبان فارسی است که حاوی اطلاعات پایه چون سال انتشار، عنوان و چکیدهٔ مستند علمی و همچنین مقولهٔ سند علمی (درصورت وجود) است. بنابراین نیاز است در طی مرحلهٔ پیش‌پردازش، ضمن یکپارچه‌سازی داده‌ها، داده‌هایی که ممکن است فاقد اطلاعات پایهٔ مورد نیاز باشند یا به زبانی به‌جز فارسی نوشته شده باشند را به‌عنوان داده‌های پرت تلقی کنیم. طی این فرایند دو دسته داده خواهیم داشت. یک دستهٔ داده از مستندات علمی پرتال جامع علوم انسانی به‌دست می‌آید و حاوی برچسب مقولهٔ هر مقاله است که می‌تواند به‌عنوان دادهٔ آموزش[1] برای ساخت مدل پردازشی مورد استفاده قرار گیرد. دستهٔ دیگر، مجموعهٔ داده‌های حاصل از پایگاه اطلاعات علمی ایران و نورمگز است که نوع مقوله یا حوزهٔ علمی مستندات علمی مشخص نشده‌است و به‌عنوان دادهٔ فاقد نشانه‌گذاری در فرایند پردازش داده مورد استفاده قرار می‌گیرند. برای کاربردی‌شدن این مجموعهٔ داده، فرایند پالایش و پیش‌پردازش جهت هنجارسازی و واحدسازی و همچنین برچسب‌دهی مقولهٔ مستندات را انجام داده‌ایم که در ادامه توضیح داده می‌شود.

۴-۲ پالایش و آماده‌سازی داده‌ها

مجموعهٔ دادهٔ گردآوری‌شده را در چند مرحله پالایش کردیم تا در قالب یک پیکرهٔ زبانی که آن را «پیکرهٔ مستندات علمی» می‌نامیم به‌دست آوریم و در این پژوهش مورد استفاده قرار می‌دهیم. پالایش‌های اعمال‌شده عبارت است از:

۱. حذف مستندات علمی تکراری از داده‌های خزش‌شده از سه منبع بایگانی مستندات.

1. training data

۲. جداسازی مستندات علمی به زبان فارسی با استفاده از الگوریتم تشخیص زبان. در راستای این هدف به‌صورت تجربی دریافتیم که کتابخانهٔ پلی‌گلات[1] با صحت ۹۶ درصد کارایی مناسبی برای تشخیص زبان دارد؛ بنابراین از این کتابخانه برای پالایش زبانی مستندات استفاده کردیم[2]. این کتابخانه که با زبان برنامه‌نویسی پایتون[3] نوشته شده است قابلیت شناسایی ۱۹۶ زبان، ازجمله فارسی، را دارد. نتیجهٔ اجرای این الگوریتم تفکیک مقالات به فارسی و غیر فارسی است که در جدول (۲) فراوانی مقالات گزارش شده است.

جدول ۲- پالایش زبانی مستندات علمی

	پرتال جامع علوم انسانی	گنج و نورمگز
فارسی	۸۱۹،۲۹۴	۸۵۹،۵۰۷،۱
غیرفارسی	۳۹۲،۳۳	۹۸۴،۲۰۹

۳. انتخاب مستندات علمی دارای تاریخ انتشار، عنوان و چکیده ازمیان داده‌های خزش‌شدهٔ وبگاه‌های هدف. در جدول (۳) فراوانی مقالات دارای تاریخ انتشار، عنوان و چکیده گزارش شده‌است.

جدول ۳- پالایش مستندات علمی از نظر وجود اطلاعات پایه

پرتال جامع علوم انسانی	گنج و نورمگز
۱۷۰،۱۱۴	۲۴۳،۸۵۱

۴-۳ پیش‌پردازش

داده‌های گردآوری‌شده پس‌از پالایش همچنان غیرقابل استفاده هستند. دو مرحلهٔ پیش‌پردازش بایستی برروی این پیکره انجام پذیرد تا قابل استفاده شود: یکی هنجارسازی داده و دیگری واحدسازی است. در هنجارسازی تلاش می‌شود نوعی یکدستی در داده ازنظر رفع مشکل تداخل حروف فارسی و عربی و اعداد، حذف علایم زائد و غیره به‌دست آید. در واحدسازی تلاش می‌شود با فاصله‌گذاری مناسب، امکان تشخیص واحدهای واژگانی به‌دست آید؛ چراکه عدم درج فاصلهٔ کامل پس‌از حروف «ا–آ»، «د»، «ذ»، «ر»، «ز»، «ژ»، «و» و «ۀ» سبب چسبندگی واژه‌ها به یکدیگر، یا درج فاصلهٔ

1. Polyglot
2. https://polyglot.readthedocs.io/en/latest/Detection.html
3. Python

کامل به‌جای نیم‌فاصله سبب تفکیک یک واحد واژگانی به بیش‌از یک واحد می‌شود. این دو مشکل موجب می‌شوند رایانه به‌هنگام خوانش پیکره نتواند واژه را به‌درستی تشخیص دهد. برای این هدف، از الگوریتم معرفی‌شده توسط قیومی (۱۳۹۷) که کارایی ۹۷/۸۰ درصدی در واحدسازی داده‌های فارسی دارد استفاده می‌کنیم.

۴-۴ برچسب‌زنی داده‌ها

بایگانی مستندات علمی حاوی منابع علمی از رشته‌های گوناگون است که ممکن است رشتهٔ علمی هریک از سندهای علمی به‌عنوان فراداده تعیین نشده باشد. برای این که در این پژوهش بتوانیم از داده‌های خزش‌شده بهره ببریم، نیاز است تا داده‌های بدون برچسب خزش‌شده از گنج و نورمگز را برچسب‌زنی کنیم. برای رسیدن به این هدف، از مدل پرسپترون مبتنی‌بر بازنمایی معنایی پارس‌برت (فراهانی و همکاران، ۲۰۲۱) که توسط قیومی و موسویان (۱۴۰۱) معرفی شده‌است و کارایی ۷۴/۷۱ درصدی دارد استفاده می‌کنیم.

۵ تحلیل الگوریتمی داده‌ها

در این بخش به تحلیل موضوعی مقالات به‌صورت الگوریتمی و همچنین فراتحلیل آنها پرداخته خواهد شد. در این بررسی، ابتدا مقالات بدون توجه به مؤلفهٔ زمان و سپس با درنظرگرفتن مؤلفهٔ زمان بررسی می‌شوند.

۵-۱ تحلیل موضوعی مقالات بدون توجه به مؤلفهٔ زمان

در بخش ۳، مدل‌سازی موضوعی و استخراج موضوعات پنهان در متن‌ها که از وابستگی‌های مخفی مفاهیم واژه‌ها حاصل می‌شود معرفی شد. در این نوع تحلیل، مجموعهٔ روابط و وابستگی‌های بین متن و موضوع و همچنین موضوع و واژه در یک متن تخمین زده شده و هریک از موضوع‌ها در قالب یک مجموعه از واژه‌ها که ازنظر مفهومی بیانگر یک موضوع هستند بازنمایی می‌شوند. موضوع‌های حاصل از اجرای الگوریتم انتزاعی هستند و ممکن است برای انسان نامفهوم باشند. برای رفع این مشکل می‌توان با استفاده از واژه‌های استخراج‌شدهٔ هر موضوع، با کمک ناظر انسانی و به‌صورت دستی یک برچسب به هر موضوع تخصیص داد. براساس این دستورالعمل می‌توان به فراتحلیل تحلیل‌های الگوریتمی ارائه‌شده پرداخت.

گفته شد که الگوریتم مدل‌سازی موضوع نوعی الگوریتم بی‌نظارتِ پارامتری است که بایستی

پیش‌از اجرای الگوریتم، تعداد موضوعات انتزاعی مشخص شود تا خوشه‌بندی مستندات در این تعداد خوشه (موضوع) انجام پذیرد. به‌صورت تجربی و با فراتحلیل خروجی الگوریتم می‌توان تعداد موضوع‌ها (خوشه‌ها) را مشخص کرد. برای این هدف می‌توان براساس تعداد مشخص موضوع، الگوریتم را اجرا کرد و به بررسی کیفی نتایج به‌دست‌آمده پرداخت. برای این منظور، تعداد ۵، ۱۰ و ۱۵ موضوع در مقالات زبان‌شناسی را هدف قرار می‌دهیم و به فراتحلیل تحلیل‌های الگوریتمی می‌پردازیم.

ازآنجاکه زبان‌شناسی به دو دستهٔ کلی زبان‌شناسی نظری و کاربردی تقسیم می‌شود، تلاش می‌شود در برچسب‌گذاری‌ها این تفکیک اعمال شود. برای شروع، ۵ موضوع از مستندات علمی موجود را به‌صورت الگوریتمی استخراج می‌کنیم. نتایج به‌دست‌آمده از موضوع انتزاعی ماشین، واژه‌های کلیدی هر موضوع، تعداد اسناد هر موضوع و برچسب تخصیص‌یافته دستی به آنها در جدول ۴ گزارش شده است. براساس تحلیل ماشینی انجام‌شده، از مجموع ۱۱۷۷۹ مقالهٔ موجود در حوزهٔ زبان‌شناسی در پیکرهٔ مستندات علمی، ۶۰/۹۶ درصد از مقالات متعلق به حوزهٔ زبان‌شناسی کاربردی بوده و مابقی در حدود ۳۹/۰۴ به حوزهٔ زبان‌شناسی نظری تعلق دارند. در پژوهش‌های کاربردی زبان‌شناختی اکثراً موضوع‌های «روان‌شناسی»، «گویش‌شناسی» و «آموزش» بررسی شده و ازمیان آنها، بیشترین توجه بر روی «روان‌شناسی» متمرکز بوده است. ازمیان موضوعات مربوط به حوزهٔ زبان‌شناسی نظری، موضوعات «تحلیل گفتمان» و «نحو» مورد توجه بوده و موضوع «تحلیل گفتمان» بیشترین توجه را درمیان پژوهشگران به خود اختصاص داده است. سه بخش دیگر زبان‌شناسی، اعم‌از «آواشناسی و واج‌شناسی»، «صرف» و «معناشناسی» نمود بارز درمیان موضوع‌های استخراج‌شده نداشتند. این نتیجه بیانگر این نکته است که مستندات مربوط به این موضوع‌ها در دادهٔ آموزش نتوانسته است بر الگوریتم یادگیری تأثیرگذار باشد.

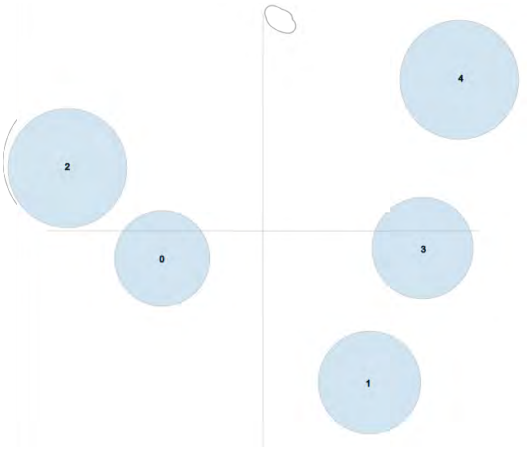
چگونگی تفکیک موضوع‌ها در شکل ۳ به‌صورت بصری نمایش داده شده‌است. این تفکیک سبب می‌شود بتوان به‌راحتی براساس واژه‌های کلیدی هر موضوع، برچسبی را به موضوع تخصیص داد. برای این نمایش بصری در فضای دو بُعدی از کتابخانهٔ pyLDAvis که توسط سیورت[1] و شرلی[2] (۲۰۱۴) تهیه و به زبان پایتون نوشته شده، استفاده شده‌است.[3] محل قرارگرفتن این دایره‌ها در

1. C. Sievert
2. K. Shirley
3. https://pyldavis.readthedocs.io/en/latest/readme.html

فضای دوبعدی براساس مرکزی است که از فاصلهٔ بین موضوع‌ها به‌دست آمده است. دایره‌ها در سه اندازه است که شامل ۱۲ درصد، ۹ درصد و ۶ درصد از حجم پیکره است.

جدول ۴- استخراج ۵ موضوع و کلیدواژه‌های مربوط از مستندات علمی زبان‌شناسی

موضوع انتزاعی	واژه‌های کلیدی	تعداد اسناد	برچسب تخصیص‌یافته
Topic0	خارجی، یادگیری، معلمان، زبان‌آموز، جنسیت	۱۴۲۴	کاربردی: آموزش
Topic1	فرهنگ، نقد، گفتمان، اجتماعی، انتقادی	۲۸۰۸	نظری: تحلیل گفتمان
Topic2	درک، مهارت، خواندن، آگاهی، فراگیری	۳۱۷۳	کاربردی: روان‌شناسی
Topic3	نحوی، ساخت، دستور، جایگاه، ساختار	۱۷۹۰	نظری: نحو
Topic4	گویش، فعل، معنایی، دستوری، آوایی، واجی	۲۵۸۴	کاربردی: گویش‌شناسی



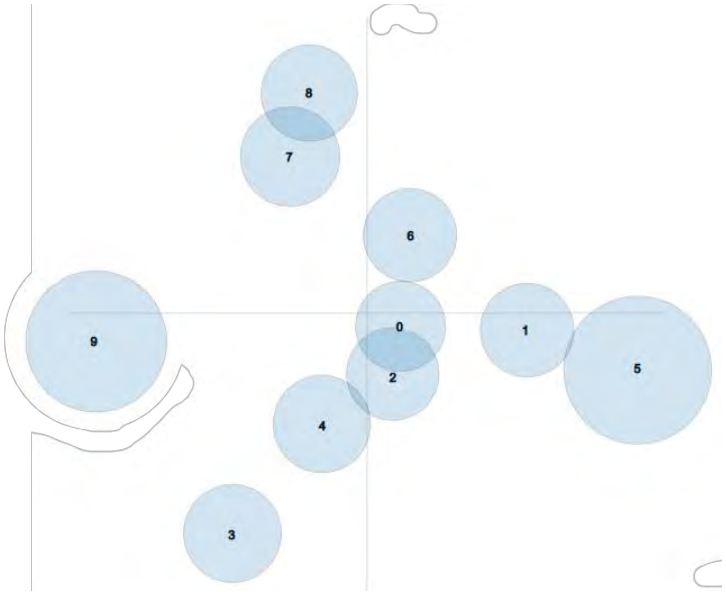
شکل ۳- نمایش بصری ۵ موضوع مستندات علمی زبان‌شناسی

در آزمایش بعدی، تعداد ۱۰ موضوع را از مقالات حوزهٔ زبان‌شناسی استخراج کردیم که نتایج به‌دست‌آمده در جدول ۵ قابل مشاهده است. نکتهٔ قابل توجه خروجی الگوریتم این است که به‌دلیل همپوشانی بعضی موضوع‌های مقالات، نمی‌توان موضوع واحدی را به این‌گونه خوشه‌ها تخصیص داد. ازاین‌رو، در این موارد تلاش شده است با هر دو رویکرد کاربردی و نظری در زبان‌شناسی برچسب موضوع‌ها مشخص شود. برای مثال، دربارهٔ واژه‌های «جنسیت»، «اجتماعی»، «گفتار»، «روایی» و «گفتمان»، اگرچه موضوع از جنبهٔ زبان‌شناسی نظری می‌تواند «تحلیل گفتمان» باشد، از جنبهٔ زبان‌شناسی کاربردی می‌تواند «جامعه‌شناسی زبان» باشد. حتی بعضی از موضوع‌ها، مانند

«تحلیل گفتمان»، با یک برچسب ولی با واژه‌های مختلف تکرار شده است که بیانگر تنوع در در یک موضوع است.

جدول ۵: استخراج ۱۰ موضوع و کلیدواژه‌های مربوط از مستندات علمی زبان‌شناسی

موضوع انتزاعی	واژه‌های کلیدی	تعداد اسناد	برچسب تخصیص‌یافته
Topic0	جنسیت، اجتماعی، گفتار، روایی، گفتمان	۸۰۰	نظری: تحلیل گفتمان؛ کاربردی: جامعه‌شناسی
Topic1	یادگیری، مادری، خارجی، زبان‌آموزان، معلمان	۹۴۴	کاربردی: آموزش
Topic2	ادبیات، تطبیقی، تاریخی، پهلوی، ترجمه‌ها	۱۲۷۲	کاربردی: تاریخی-تطبیقی
Topic3	گویش، معنایی، آوایی، واجی، ساخت	۱۱۲۰	کاربردی: گویش‌شناسی
Topic4	ترکی، مقابله‌ای، معلمان، آوایی، واجی	۱۲۵۹	نظری: آواشناسی؛ کاربردی: آموزش
Topic5	تفکر، فراگیران، هوش، آگاهی، خواندن	۳۱۲۵	کاربردی: روان‌شناسی
Topic6	فرهنگ، نشانه‌های، استعاره‌های، اجتماعی، رفتار	۱۳۳۶	نظری: معنی‌شناسی
Topic7	گفتمان، انتقادی، مفهوم، اندیشه، گفتمانی	۱۲۴۶	نظری: تحلیل گفتمان
Topic8	نقد، نشانه، گفتمان، معنا، مفهوم	۱۱۸۸	نظری: تحلیل گفتمان
Topic9	فعل، نحوی، مرکب، دستوری، مجهول،حالت	۲۲۳۵	نظری: نحو



شکل ۴- نمایش بصری ۱۰ موضوع مستندات علمی زبان‌شناسی

براساس تحلیل ماشینی انجام‌شده در جدول ۵، از مجموع مقالات موجود در حوزهٔ زبان‌شناسی، ۷۲/۳۳ درصد از مقالات متعلق به حوزهٔ زبان‌شناسی کاربردی است و ۶۸/۴۶ درصد به حوزهٔ زبان‌شناسی نظری تعلق دارند. در این محاسبه، توزیع آماریِ موضوعاتی که متعلق به هر دو حوزه هستند در آمار هر دو حوزه لحاظ شده است. در زبان‌شناسی نظری، موضوع‌هایی مانند «تحلیل گفتمان»، «نحو»، «معنی‌شناسی» و «آواشناسی» مورد توجه بوده است؛ درحالی‌که در زبان‌شناسی کاربردی، موضوع‌هایی مانند «روان‌شناسی زبان»، «آموزش زبان»، «زبان‌شناسی تاریخی و تطبیقی»، «گویش‌شناسی» و «جامعه‌شناسی زبان» مورد توجه بوده است. ازمیان موضوع‌های زبان‌شناسی نظری و کاربردی، «تحلیل گفتمان» و «روان‌شناسی زبان» کانون توجه بوده و به‌ترتیب حدود ۲۱ و ۲۷ درصد از موضوع‌های مستندات در حوزهٔ زبان‌شناسی را به خود اختصاص داده‌اند. موضوعی مانند «صرف» به‌عنوان یکی دیگر از بخش‌های زبان‌شناسی، نمود بارز درمیان موضوع‌های استخراج شده نداشته است. این نتیجه بیانگر این نکته است که مستندات مربوط به این موضوع در دادهٔ آموزش نتوانسته است بر الگوریتم یادگیری تأثیرگذار باشد و در قالب یک موضوع مشخص،

هویت مستقل پیدا کند. تفکیک بعضی موضوعات و همپوشانی آنها در شکل ۴ قابل مشاهده است. همان‌طور که ملاحظه می‌شود، موضوع Topic2 با برچسب «کاربردی: تاریخی-تطبیقی» با دو موضوع Topic0 با برچسب «نظری: تحلیل گفتمان؛ کاربردی: جامعه‌شناسی» و Topic4 با برچسب «نظری: آواشناسی؛ کاربردی: آموزش» ارتباط محتوایی دارد. ارتباط بین دو موضوع Topic0 و Topic2 که به نظر می‌رسد معناشناختی باشد بیشتر از ارتباط دو موضوع Topic2 و Topic4 است که به نظر می‌رسد آواشناختی باشد. دو موضوع Topic7 و Topic8 که با یکدیگر ارتباط محتوایی دارند دارای برچسب «نظری: تحلیل گفتمان» هستند. تحلیل گفتمان انجام‌شده در موضوع Topic7 با رویکرد نشانه‌شناختی و معنایی است درحالی‌که در موضوع Topic8، تحلیل گفتمان با رویکرد نحوی و ادبی انجام پذیرفته است.

در آزمایش دیگری، تعداد ۱۵ موضوع را از مقالات در حوزهٔ زبان‌شناسی استخراج کردیم که نتایج به‌دست‌آمده در جدول ۶ قابل مشاهده‌اند. نکتهٔ قابل توجه خروجی الگوریتم این است که به‌دلیل همپوشانی موضوع‌های مقالات، به اکثر خوشه‌ها بیش‌از یک برچسب تخصیص داده شده و حتی بعضی از موضوع‌ها با یک برچسب ولی با واژه‌های مختلف تکرار شده‌اند. برای مثال می‌توان به موضوع «گویش‌شناسی» اشاره کرد که ممکن است در آن یک گویش با رویکرد زبان‌شناسی ازجنبهٔ نحو، آوا، واج، معنا و تحلیل گفتمان و کاربردشناسی مورد مطالعه قرار گیرد. بنابراین، برای این موضوع نمی‌توان یک برچسب مشخص تخصیص داد و براساس واژه‌های آن موضوع می‌توان علاوه بر برچسب در زبان‌شناسی کاربردی، برچسبی در زبان‌شناسی نظری به آن موضوع تخصیص داد. نکتهٔ قابل توجه در این نتایج این است که موضوعات «صرف» و «آواشناسی» در Topic4 و Topic14 هویت مستقل پیدا کرده و در قالب دو موضوع مشخص دو خوشه تشکیل داده‌اند.

جدول ۶- استخراج ۱۵ موضوع و کلیدواژه‌های مربوط از مستندات علمی زبان‌شناسی

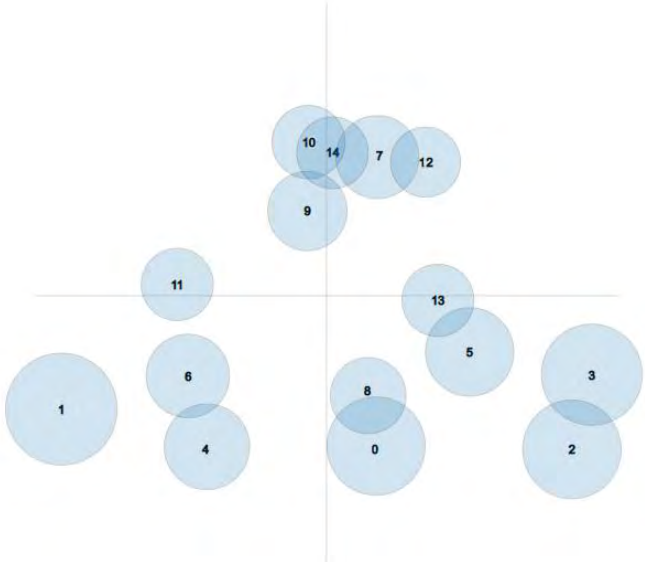
موضوع انتزاعی	واژه‌های کلیدی	تعداد اسناد	برچسب تخصیص‌یافته
Topic0	تدریس، مدرس، دبیرستان، پایه، مادری	۱۹۱۵	کاربردی: آموزش
Topic1	فعل، مرکب، نحوی، مجهول، اسم	۱۵۲۵	نظری: نحو
Topic2	تست، بهبود، آگاهی، نوشتاری، تفکر	۱۴۴۴	کاربردی: روان‌شناسی
Topic3	یادگیری، خارجی، فراگیران، خواندن، توانش	۱۶۲۸	کاربردی: آموزش
Topic4	واژه‌های، واژه‌ها، وندهای، اشتقاقی، وند	۱۲۴۵	نظری: صرف

Topic5	خارجی، خواندن، درک، یادگیری، فراگیری، گفتار	۹۸۲	کاربردی: آموزش
Topic6	دستور، ساخت‌های، وجه، سببی، نحوی	۷۲۰	نظری: نحو
Topic7	گفتمان، انتقادی، ادبی، انسجام، گفتمانی	۹۱۱	نظری: تحلیل گفتمان
Topic8	معنایی، یادگیری، معنای، مفاهیم، درک	۵۶۹	نظری: معنی‌شناسی؛ کاربردی: آموزش
Topic9	گویش، واجی، آوایی، واج، واکه	۷۴۲	نظری: آواشناسی؛ کاربردی: گویش‌شناسی
Topic10	گفتمان، زنان، مردان، روایت، پیام	۶۴۰	نظری: تحلیل گفتمان
Topic11	ساخت، ضمایر، بند، حرکت، گویش	۶۴۸	نظری: نحو؛ کاربردی: گویش‌شناسی
Topic12	جنسیت، نقد، درک، کلامی، زن	۶۰۸	نظری: تحلیل گفتمان
Topic13	تطبیقی، شفاهی، گفتمان، انتقادی، سغدی	۵۷۶	نظری: تحلیل گفتمان؛ کاربردی: تاریخی-تطبیقی
Topic14	گفتاری، ترکی، بازشناسی، الگوی، لهجه	۵۴۲	نظری: آواشناسی؛ کاربردی: رایانشی

براساس داده‌های تحلیل‌شده، ازمیان کاربردهای زبان‌شناسی می‌توان به موضوع‌های روان‌شناسی زبان، آموزش زبان، گویش‌شناسی و زبان‌شناسی تاریخی-تطبیقی اشاره کرد که ازمیان آنها موضوع آموزش حدود ۴۳ درصد از موضوع مستندات علمی را تشکیل داده است. ازمیان بخش‌های مربوط به زبان‌شناسی نظری، موضوع نحو حدود ۲۵ درصد از موضوع مستندات علمی را تشکیل داده است. موضوع دیگری که در فرایند برچسب‌گذاری مشاهده شد، این بود که وجود واژه‌های تخصصی مانند «بازشناسی» که معمولاً در بازشناسی گفتار و پردازش صوت و زبان‌شناسی رایانشی کاربرد دارد می‌تواند رد پایی از میان‌رشته‌ای‌های کاربردی را در تحلیل موضوعی بر جای بگذارد. ازاین‌رو، شناخت واژه‌های تخصصی هر رشته و میان‌رشته از اهمیت شایانی برخوردار است.

یکی‌از چالش‌هایی که با تعداد ۱۵ موضوع در خوشه‌بندی داده با آن مواجه شدیم وجود ابهام در مواردی بود که واژه‌های کلیدی موضوع ازنظر نظری به دو بخش مرتبط در زبان‌شناسی، مانند «نحو» و «معناشناسی»، مربوط می‌شدند. برای حفظ یکدستی در فرایند برچسب‌گذاری، خود را به این‌گونه ملزم و محدود کردیم که حداکثر یک برچسب در حوزهٔ زبان‌شناسی نظری و یک برچسب در حوزهٔ زبان‌شناسی کاربردی به یک موضوع تخصیص یابد. چنانچه در موضوعی که ممکن است به دو بخش زبان‌شناسی نظری، مانند نحو و معناشناسی یا معناشناسی و تحلیل گفتمان، مربوط شود باید

فقط یک برچسب تخصیص یابد. برای رفع ابهام از کلیدواژه‌های متعددی استفاده می‌شود که بالاترین احتمال را در آن موضوع داشته باشد و ابهام‌زدایی صورت می‌پذیرد. نمونه چالش دیگر، بررسی‌های رده‌شناختی و نحوی در زبان است که بسیار ممزوج یکدیگر شده و نمی‌توان به‌روشنی میان آنها مرزی را ترسیم کرد و یک یا هردو برچسب را به آن موضوع تخصیص داد. همپوشانی ۱۵ موضوع استخراج‌شده به‌صورت بصری در شکل ۵ به‌عنوان یکی دیگر از خروجی‌های سامانه نمایش داده شده است. در این تصویر، موضوع Topic14 که «نظری: آواشناسی؛ کاربردی: رایانشی» برچسب خورده است با موضوعات Topic8 با برچسب «نظری: معنی‌شناسی؛ کاربردی: گویش‌شناسی»، Topic9 با برچسب «نظری: آواشناسی؛ کاربردی: گویش‌شناسی» و Topic11با برچسب «نظری: نحو؛ کاربردی: گویش‌شناسی» همپوشانی دارد. همان‌طور که در برچسب‌ها مشخص است، ویژگی مطالعات آواشناختی حلقه مشترک این موضوعات است که از جنبه‌های مختلف مورد بررسی قرار گرفته‌است.



شکل ۵- نمایش بصری ۱۵ موضوع مستندات علمی زبان‌شناسی

۵-۲ تحلیل و فراتحلیل موضوعی در گذر زمان

در بخش قبل، به فراتحلیل تحلیل‌های ماشینی که به‌واسطهٔ استخراج موضوعات مخفی کلیهٔ مستندات علمی در حوزهٔ زبان‌شناسی انجام شده بود پرداختیم. یک متغیری که می‌توان در این نوع

بررسی مورد توجه قرار داد، مؤلفهٔ زمان است؛ به این مفهوم که موضوعات مستندات علمی در بازه‌های زمانی مشخص بررسی شود تا امکان مقایسهٔ آنها با یکدیگر فراهم شود. برای رسیدن به این هدف، زمان انتشار مقالات در استخراج موضوع مستندات لحاظ می‌شود. پیش‌از آن که به تحلیل موضوعات مستخرج بپردازیم، اطلاعات آماری مربوط به تعداد مستندات در حوزهٔ زبان‌شناسی که در بازهٔ زمانی ۱۳۰۶ تا ۱۳۹۹ در پیکره موجود است را استخراج کردیم. ازآنجاکه حجم زیادی از مقالات حوزهٔ زبان‌شناسی به دهه‌های ۱۳۸۰ و ۱۳۹۰ تعلق دارد و قبل‌از این مدت، مقالات چندان زیادی نیافتیم، کل داده را به ۵ دورهٔ زمانی تقسیم کردیم که تقسیم زمانی و تعداد مقالات زبان‌شناسی در بازه‌های زمانی مورد نظر در جدول ۷ گزارش شده‌است.

جدول ۷- تعداد مقالات زبان‌شناسی در ۵ دورهٔ زمانی

سال انتشار	تعداد سند
۱۳۸۰-۱۳۰۶	۲۲۰۷
۱۳۸۵-۱۳۸۱	۱۳۸۵
۱۳۹۰-۱۳۸۶	۲۷۷۶
۱۳۹۵-۱۳۹۱	۴۷۴۹
۱۳۹۹-۱۳۹۶	۶۶۲

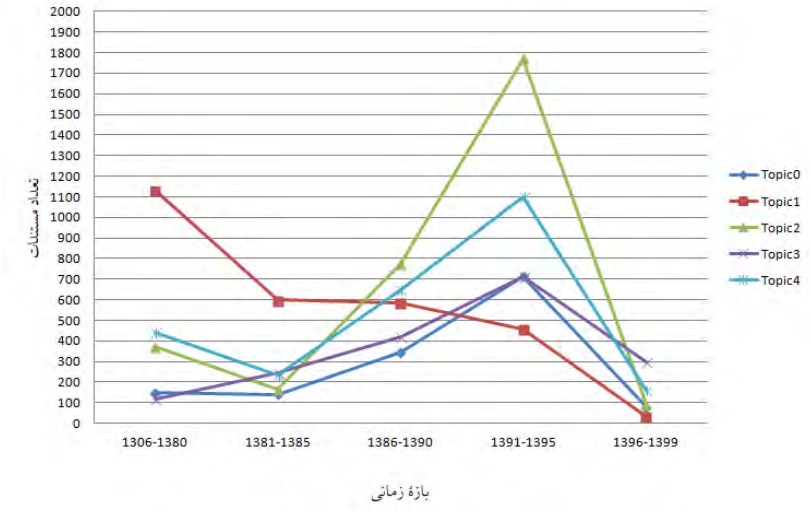
پس‌از اجرای الگوریتم مدل‌سازی موضوع پویا از نظر زمانی، به تعداد موضوع‌های تعیین‌شده برای مقاطع زمانی مشخص خروجی به‌دست می‌آید. بنابراین، در این بخش از تحلیل، در هر آزمایش برای استخراج موضوع پنج نمودار خواهیم داشت که هر نمودار به یک بازهٔ زمانی تعلق دارد. در ادامه، تعداد ۵، ۱۰ و ۱۵ موضوع باتوجه‌به مؤلفهٔ زمان به‌صورت الگوریتمی استخراج شده و تحلیل می‌شود.

در جدول ۷ پنج بازهٔ زمانی را برای استخراج موضوع‌ها در نظر گرفتیم و سپس اطلاعات آماری مربوط به فراوانی موضوعات در بازه‌های زمانی مشخص را استخراج و در جدول ۸ گزارش کردیم. برای آنکه بتوانیم موضوعات زبان‌شناسی را در مقاطع زبانی مختلف بسنجیم و مقایسه کنیم، آن‌ها را به‌صورت نمودار در شکل ۶ نمایش داده‌ایم. براساس این نمودار دو بازهٔ زمانی ۱۳۰۶ تا ۱۳۸۰ و ۱۳۹۶ تا ۱۳۹۹ در این بررسی چندان قابل اعتماد نیستند. دلیل اصلی این است که در بازهٔ زمانی اول، فاصله دو بازهٔ زمانی بسیار طولانی است و به‌طور دقیق نمی‌توان سخن گفت. در بازهٔ زمانی

دوم، تعداد مستندات برای تمامی موضوعات کاهش یافته‌است. دو دلیل برای این کاهش وجود دارد. زمان خزش داده‌های این پژوهش در سال ۱۳۹۸ انجام پذیرفته است و مقالات منتشرشده سال ۱۳۹۸ و ۱۳۹۹ در زمان خزش چندان زیاد نبوده است. دلیل دیگر که با دلیل قبلی مرتبط است، به به‌روزرسانی وبگاه‌های مختلف بایگانی مستندات علمی در طول زمان مربوط است. به این مفهوم که انتهای سال ۱۳۹۷ به مفهوم بایگانی کامل مستندات متعلق به سال ۱۳۹۷ در آن مقطع زمانی نیست. ازاین‌رو، این بازهٔ زمانی نیز نمی‌تواند به‌طور دقیق گویای وضعیت موضوعات مطرح‌شده در مستندات علمی باشد.

جدول ۸- توزیع ۵ موضوع مقالات زبان‌شناسی در ۵ دورهٔ زمانی

موضوع انتزاعی	برچسب	۱۳۰۶ تا ۱۳۸۰	۱۳۸۱ تا ۱۳۸۵	۱۳۸۶ تا ۱۳۹۰	۱۳۹۱ تا ۱۳۹۵	۱۳۹۶ تا ۱۳۹۹
Topic0	کاربردی: آموزش	۱۴۷	۱۴۱	۳۴۵	۷۱۲	۷۹
Topic1	نظری: تحلیل گفتمان	۱۱۳۰	۵۹۸	۵۸۶	۴۵۹	۳۵
Topic2	کاربردی: روان‌شناسی	۳۷۳	۱۶۶	۷۷۵	۱۷۶۹	۹۰
Topic3	نظری: نحو	۱۱۶	۲۴۷	۴۲۰	۷۱۰	۲۹۷
Topic4	کاربردی: گویش‌شناسی	۴۴۱	۲۳۳	۶۵۰	۱۰۹۹	۱۶۱



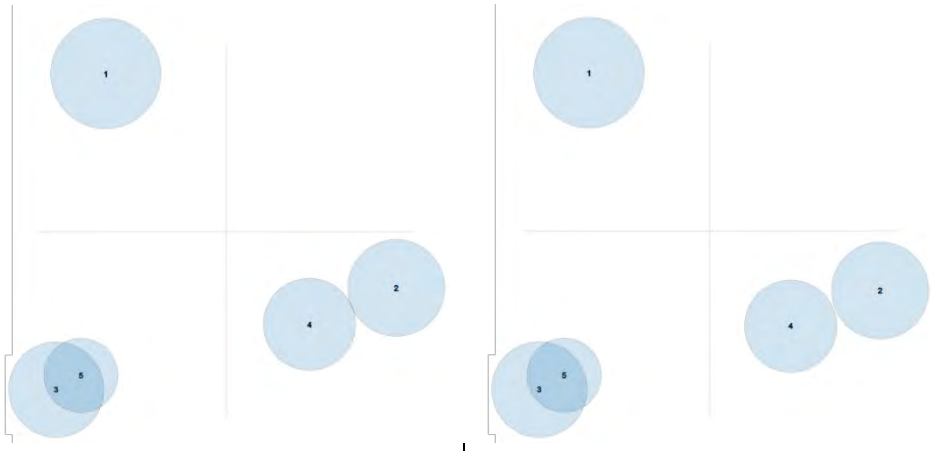
شکل ۶- تنوع ۵ موضوع مستندات علمی زبان‌شناسی در گذر زمان

همان‌طور که در نمودار شکل ۶ مشاهده می‌شود، تعداد مقالات منتشرشده در اکثر موضوعات، به‌خصوص موضوعات کاربردی مانند آموزش زبان، روان‌شناسی زبان و گویش‌شناسی، و همچنین موضوعات نظری، مانند نحو، رو به رشد بوده‌است؛ ولی با گذشت زمان، موضوع نظری تحلیل گفتمان کمتر مورد توجه قرار گرفته‌است. حتی اگر مقطع زمانی طولانی‌مدت ۱۳۰۶ تا ۱۳۸۰ و کوتاه‌مدت ۱۳۹۶ تا ۱۳۹۹ را نادیده بگیریم، همچنان این کاهش به چشم می‌خورد. درحالی‌که در این شکل، موضوع روان‌شناسی زبان از ۱۳۸۰ تا ۱۳۹۵ با شیب زیادی رو به رشد بوده و مورد توجه پژوهشگران قرار گرفته است.

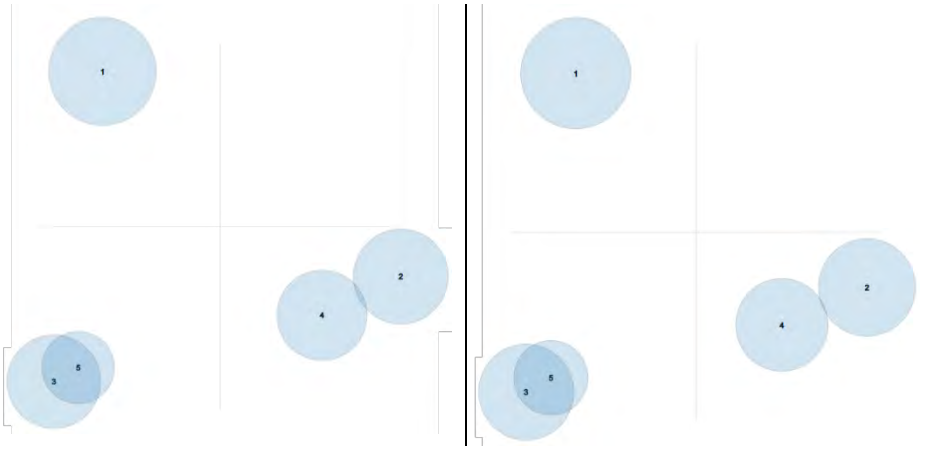
گویش‌شناسی نیز موضوع دیگری است که پس‌از روان‌شناسی زبان مورد توجه پژوهشگران قرار گرفته است. توجه به موضوع‌های نحو و آموزش زبان در رتبه‌های بعدی قرار می‌گیرد. نکته قابل توجه درمورد موضوع نحو این است که این موضوع از ۱۳۰۶ مورد توجه قرار گرفته و در بازهٔ سال‌های ۱۳۹۶ تا ۱۳۹۹ بیشترین موضوعات زبان‌شناسی را به خود اختصاص داده است.

از جنبهٔ دیگری به بررسی موضوعات بازه‌های زمانی مختلف پرداختیم که در شکل‌های ۷-الف تا ۷-د قابل مشاهده است. نکتهٔ قابل توجه در این شکل‌ها این است که دو موضوع آموزش زبان و روان‌شناسی زبان در بازهٔ ۱۳۰۶ تا ۱۳۸۰ بسیار به یکدیگر نزدیک بوده‌اند ولی اشتراکی نداشته‌اند. از سال ۱۳۸۰ تا ۱۳۸۵ میان این دو موضوع از نظر واژه‌های به‌کاررفته در این دو موضوع، اشتراکی ایجاد شده و این اشتراک تا سال ۱۳۹۹ ادامه یافته و بیشتر شده است. استفاده از واژه‌های مشترک در این دو موضوع بیانگر این نکته است که با گذشت زمان، نوعی همگرایی بین دو موضوع آموزش زبان و روان‌شناسی زبان ایجاد شده و سبب همسوشدگی آنها شده است. اگر قبلاً به‌صورت دو موضوع مجزا از یکدیگر بررسی می‌شد، امروزه ماهیت بین‌رشته‌ای بین این دو موضوع شکل گرفته و سبب شده است سهم واژه‌هایی مشترک در آموزش زبان و روان‌شناسی زبان افزایش یابد.

برای بررسی مطالعات مربوط به موضوعات آموزش زبان (Topic0) و روان‌شناسی زبان (Topic2) که براساس شکل‌های ۷-الف و ۷-د نوعی اشتراک موضوعی در مقالات ایجاد شده است، ۱۰ واژهٔ پرکاربرد این دو موضوع متعلق به بازهٔ زمانی ۱۳۰۶ تا ۱۳۸۰ را با ۱۰ واژهٔ پرکاربرد این دو موضوع متعلق به بازهٔ زمانی ۱۳۹۶ تا ۱۳۹۹ مقایسه کردیم. واژه‌های این دو موضوع در دو برش زمانی مورد نظر در جدول ۹ فهرست شده‌اند.



الف) بازۀ زمانی ۱۳۰۶ تا ۱۳۸۰ ب) بازۀ زمانی ۱۳۸۱ تا ۱۳۸۵



ج) بازۀ زمانی ۱۳۸۶ تا ۱۳۹۰ د) بازۀ زمانی ۱۳۹۶ تا ۱۳۹۹

شکل ۷- روند رشد و توسعۀ موضوعات مقالات علمی زبان‌شناسی در گذر زمان

جدول ۹- ۱۰ واژهٔ مشترک موضوعات روان‌شناسی و آموزش زبان در دو برش زمانی

رتبه	بازهٔ زمانی ۱۳۰۶ تا ۱۳۸۰		بازهٔ زمانی ۱۳۹۶ تا ۱۳۹۹	
	روان‌شناسی زبان	آموزش زبان	روان‌شناسی زبان	آموزش زبان
۱	یادگیری	زبان‌آموز	یادگیری	زبان‌آموز
۲	اجتماع	آزمون	معلمان	مهارت
۳	معلمان	مهارت	اجتماع	آزمون
۴	جنسیت	فراگیر	خارج	یادگیری
۵	خارج	یادگیری	زبان‌آموز	درک
۶	زبان‌آموز	درک	جنسیت	فراگیر
۷	مدرس	خواندن	مدرس	خواندن
۸	هوش	خارج	کلاس	خارج
۹	الگو	نوشتار	تدریس	نوشتار
۱۰	کلاس	دانش	هوش	دانش

از بررسی واژه‌های دو موضوع روان‌شناسی زبان و آموزش زبان در دو برش زمانی ۱۳۰۶ تا ۱۳۸۰ و ۱۳۹۶ تا ۱۳۹۹ نتایج زیر به‌دست آمد:

- در برش زمانی اول بین این دو فهرست سه واژهٔ «یادگیری»، «خارج» و «زبان‌آموز» مشترک هستند که می‌توانند گویای نزدیکی دو موضوع باشند.
  - واژهٔ «یادگیری» دارای رتبهٔ نخست در روان‌شناسی زبان و رتبهٔ پنجم در آموزش زبان است.
  - واژهٔ «خارج» دارای رتبهٔ پنجم در روان‌شناسی زبان و رتبهٔ هشتم در آموزش زبان است.
  - واژهٔ «زبان‌آموز» دارای رتبهٔ ششم در روان‌شناسی زبان و رتبهٔ نخست در آموزش زبان است.
- در برش زمانی دوم بین این دو فهرست، سه واژهٔ «یادگیری»، «خارج» و «زبان‌آموز» مشترک هستند که تغییر رتبهٔ واژه‌های مشترک می‌تواند گویای همپوشانی دو موضوع باشد.
  - واژهٔ «یادگیری» دارای رتبهٔ نخست در روان‌شناسی زبان و رتبهٔ چهارم در

آموزش زبان است که رتبهٔ این واژه در موضوع آموزش زبان نسبت‌به برش زمانی اول یک پله افزایش داشته و به رتبهٔ این واژه در موضوع روان‌شناسی زبان نزدیکتر شده است.

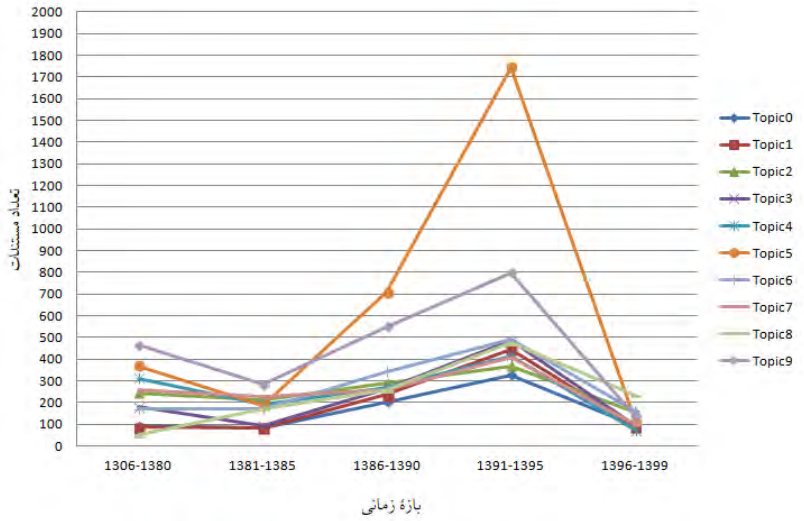
- واژهٔ «خارج» دارای رتبهٔ چهارم در روان‌شناسی زبان و رتبهٔ هشتم در آموزش زبان است. این واژه در برش زمانی اول در روان‌شناسی زبان یک پله افزایش رتبه داشته ولی در برش زمانی دوم در آموزش زبان هیچ تغییری نداشته است.
- واژهٔ «زبان‌آموز» دارای رتبهٔ پنجم در روان‌شناسی زبان و رتبهٔ نخست در آموزش زبان است. این واژه در برش زمانی اول در روان‌شناسی زبان یک پله افزایش رتبه داشته ولی در برش زمانی دوم در آموزش زبان هیچ تغییری نداشته است.

براساس شواهد چنین می‌توان نتیجه گرفت که کاهش رتبهٔ «زبان‌آموز» در روان‌شناسی زبان و همچنین کاهش رتبهٔ «یادگیری» در آموزش زبان گویای ایجاد اشتراک و همپوشانی این دو موضوع است.

نکات قابل توجه دیگری که در جدول ۹ قابل مشاهده است عبارتند از این که رتبه‌بندی واژه‌های پراهمیت موضوع روان‌شناسی زبان در دو برش زمانی تغییر داشته است؛ ولی در آموزش زبان واژه‌های رتبه هفتم تا دهم ثابت باقی مانده است. واژهٔ «جنسیت» در روان‌شناسی زبان در بازهٔ ۱۳۰۶ تا ۱۳۸۰ که دارای رتبهٔ چهارم است در برش زمانی دوم با دو پله سقوط به رتبهٔ ششم رسیده است. به‌نظر می‌رسد در مقالات متأخر نسبت‌به مقالات قدیمی‌تر توجه کمتری به متغیر جنسیت معطوف شده است. عدم توجه به «هوش» نیز به همین صورت است؛ چراکه از رتبهٔ هشتم در برش زمانی اول به رتبهٔ دهم سقوط کرده است. اگرچه در برش زمانی اول، بررسی «الگو» در روان‌شناسی زبان رتبهٔ نهم را به خود اختصاص داده است، در دورهٔ متأخر بررسی آن کمتر مورد توجه قرار گرفته و با دو پله سقوط در رتبهٔ یازدهم قرار گرفته است. توجه به «معلمان»، «تدریس» و «کلاس» در دورهٔ متأخر روان‌شناسی زبان بیشتر از برش زمانی اول است. در برش زمانی اول آموزش زبان توجه بر «آزمون» بیشتر از «مهارت» بوده است ولی در دورهٔ متأخر تغییر رویکرد صورت گرفته و توجه به «مهارت» بیشتر از «آزمون» است. ازاین‌رو، «درک» که یکی‌از مهارت‌های آموزش زبان است در دورهٔ متأخر بیشتر مورد توجه قرار گرفته و سبب کاهش رتبهٔ این واژه شده است.

جدول ۱۰- توزیع ۱۰ موضوع مقالات زبان‌شناسی در ۵ دورهٔ زمانی

موضوع انتزاعی	برچسب	۱۳۰۶ تا ۱۳۸۰	۱۳۸۱ تا ۱۳۸۵	۱۳۸۶ تا ۱۳۹۰	۱۳۹۱ تا ۱۳۹۵	۱۳۹۶ تا ۱۳۹۹
Topic0	نظری: تحلیل گفتمان؛ کاربردی: جامعه‌شناسی	۹۳	۸۳	۲۰۴	۳۲۷	۹۳
Topic1	کاربردی: آموزش	۸۸	۸۵	۲۳۶	۴۴۵	۹۰
Topic2	کاربردی: تاریخی-تطبیقی	۲۴۶	۲۱۰	۲۹۰	۳۶۹	۱۵۷
Topic3	کاربردی: گویش‌شناسی	۱۸۲	۹۶	۲۶۳	۴۸۶	۹۳
Topic4	نظری: آواشناسی؛ کاربردی: آموزش	۳۱۳	۱۹۲	۲۶۸	۴۱۵	۷۱
Topic5	کاربردی: روان‌شناسی	۳۷۰	۱۸۸	۷۰۸	۱۷۴۵	۱۱۴
Topic6	نظری: معنی‌شناسی؛ کاربردی: جامعه‌شناسی	۱۷۰	۱۷۰	۳۴۴	۴۹۰	۱۶۲
Topic7	نظری: تحلیل گفتمان	۲۵۷	۲۲۶	۲۵۷	۴۰۷	۹۹
Topic8	نظری: تحلیل گفتمان	۵۰	۱۷۳	۲۵۷	۴۷۴	۲۳۴
Topic9	نظری: نحو	۴۶۶	۲۸۴	۵۵۱	۷۹۷	۱۳۷



شکل ۸- تنوع ۱۰ موضوع مستندات علمی زبان‌شناسی در گذر زمان

برای آنکه بهتر بتوانیم موضوعات را در بازه‌های زمان پنج‌گانه مورد نظر بررسی کنیم، ۱۰ موضوع

استخراج‌شده در جدول ۱۰ را مورد بررسی و فراتحلیل قرار می‌دهیم. اطلاعات آماری مربوط به فراوانی موضوعات در بازه‌های زمانی منتخب را استخراج و در جدول ۱۰ گزارش کردیم. سنجش و مقایسه موضوعات زبان‌شناسی در مقاطع زبانی مختلف، به‌صورت نمودار در شکل ۸ نمایش داده شده است. براساس دلایلی که پیشتر توضیح داده شد، دو بازهٔ زمانی ۱۳۰۶ تا ۱۳۸۰ و ۱۳۹۶ تا ۱۳۹۹ در این بررسی چندان قابل اعتماد نیستند؛ بنابراین در فراتحلیل خود آنها را لحاظ نمی‌کنیم. همان‌طور که در نمودار شکل ۱۳ مشاهده می‌شود، تعداد مقالات منتشرشده در همهٔ موضوعات در بازهٔ ۱۳۸۰ تا ۱۳۹۵ افزایشی است. در میان موضوعات مستندات علمی زبان‌شناسی متعلق به این بازه، روان‌شناسی زبان مهمترین موضوعی بوده است که در مستندات علمی این بازه مورد توجه قرار گرفته است. موضوع نحو نیز در رتبه دوم قرار دارد و کانون توجه پژوهشگران زبان‌شناسی است. موضوع Topic0 که حاوی برچسب تحلیل گفتمان و جامعه‌شناسی زبان است موضوعی است که کمترین توجه زبان‌شناسان را به خود جلب کرده است.

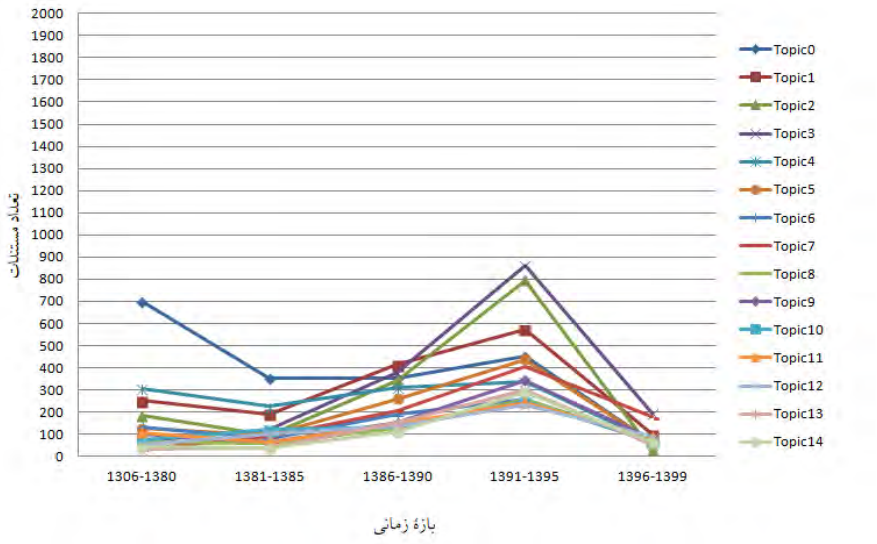
در ادامه، ۱۵ موضوع استخراج‌شده در جدول ۶ را مورد بررسی و فراتحلیل قرار می‌دهیم. اطلاعات آماری مربوط به فراوانی موضوعات در بازه‌های زمانی منتخب را استخراج و در جدول ۱۱ گزارش می‌کنیم.

جدول ۱۱- توزیع ۱۵ موضوع مقالات زبان‌شناسی در ۵ دورهٔ زمانی

موضوع انتزاعی	برچسب	۱۳۰۶ تا ۱۳۸۰	۱۳۸۱ تا ۱۳۸۵	۱۳۸۶ تا ۱۳۹۰	۱۳۹۱ تا ۱۳۹۵	۱۳۹۶ تا ۱۳۹۹
Topic0	کاربردی: آموزش	۶۹۷	۳۵۳	۳۵۳	۴۵۱	۶۱
Topic1	نظری: نحو	۲۵۰	۱۹۱	۴۱۳	۵۷۲	۹۹
Topic2	کاربردی: روان‌شناسی	۱۸۵	۹۳	۳۴۲	۷۹۵	۲۹
Topic3	کاربردی: آموزش	۷۰	۱۲۳	۳۷۹	۸۶۰	۱۹۶
Topic4	نظری: صرف	۳۰۵	۲۲۴	۳۱۰	۳۳۶	۷۰
Topic5	کاربردی: آموزش	۱۲۴	۹۶	۲۶۰	۴۳۷	۶۵
Topic6	نظری: نحو	۱۳۲	۸۰	۱۸۸	۲۵۹	۶۱
Topic7	نظری: تحلیل گفتمان	۳۱	۹۰	۲۰۷	۴۰۶	۱۷۷
Topic8	نظری: معنی‌شناسی؛ کاربردی:آموزش	۶۳	۵۶	۱۲۹	۲۵۸	۶۳

Topic9	نظری: آواشناسی؛ کاربردی:گویش‌شناسی	۱۰۱	۶۴	۱۵۲	۳۴۴	۸۱
Topic10	نظری: تحلیل گفتمان	۷۶	۱۲۰	۱۳۰	۲۴۷	۶۷
Topic11	نظری: نحو؛ کاربردی: گویش‌شناسی	۱۰۷	۶۶	۱۴۶	۲۴۳	۸۶
Topic12	نظری: تحلیل گفتمان	۵۳	۹۸	۱۳۷	۲۳۲	۸۸
Topic13	نظری: تحلیل گفتمان؛ کاربردی: تاریخی-تطبیقی	۳۲	۳۸	۱۵۲	۳۰۰	۵۴
Topic14	نظری: آواشناسی؛ کاربردی: رایانشی	۴۳	۳۵	۱۱۳	۲۸۷	۶۴

برای آنکه بتوان موضوعات زبان‌شناسی را در مقاطع زبانی مختلف بسنجیم و مقایسه کنیم، به‌صورت نمودار در شکل ۹ نمایش دادیم. براساس دلایلی که پیشتر درمورد آنها توضیح داده شد دو بازۀ زمانی ۱۳۰۶ تا ۱۳۸۰ و ۱۳۹۶ تا ۱۳۹۹ در این بررسی چندان قابل اعتماد نیست؛ بنابراین در فراتحلیل خود آنها را لحاظ نمی‌کنیم.



شکل ۹- تنوع ۱۵ موضوع مستندات علمی زبان‌شناسی در گذر زمان

همان‌طور که در نمودار شکل ۹ مشاهده می‌شود، تعداد مقالات منتشرشده در همۀ موضوعات در بازۀ ۱۳۸۰ تا ۱۳۹۵ افزایشی است. در میان موضوعات مستندات علمی زبان‌شناسی متعلق به این بازه، آموزش زبان مهمترین موضوعی است که در مستندات علمی این بازه مورد توجه قرار گرفته

است. موضوع روان‌شناسی زبان در رتبه دوم قرار داشته و موضوع Topic1 که نحو است رتبهٔ سوم کانون توجه پژوهشگران زبان‌شناسی را به خود اختصاص داده است. موضوع Topic12 که حاوی برچسب تحلیل گفتمان است موضوعی است که کمترین توجه زبان‌شناسان را به خود جلب کرده است. در این موضوع، تحلیل گفتمان جنسیتی بررسی و مطالعه شده است. موضوع Topic4 که در مورد صرف است اگرچه در بازهٔ زمانی ۱۳۸۰ تا ۱۳۹۰ رشد داشته است، در بازهٔ ۱۳۹۰ تا ۱۳۹۵ رشد چشمگیری نداشته و تقریباً ثابت مانده است. موضوع Topic6 که درمورد نحو و بررسی ساخت‌های نحوی است شاهد کمترین رشد انتشار مقالات علمی در این بازه‌های زمانی بوده است. درحالی‌که بررسی‌های نحوی در سطح واژه در موضوع Topic1 جایگاه سوم مقالات علمی را به‌دست آورده است.

۶ نتیجه‌گیری

در این مقاله، به تکوین و تحلیل موضوعی مقالات علمی در حوزهٔ زبان و زبان‌شناسی پرداخته شد. در همین راستا، از الگوریتم مدل‌سازی موضوع استفاده کردیم تا موضوعات انتزاعی مقالات باتوجه‌به مؤلفهٔ زمان به‌دست آید. سپس، هریک از خوشه‌ها را برچسب‌گذاری کردیم تا موضوعات انتزاعی معنادار شوند. برای این هدف، تعداد ۵، ۱۰ و ۱۵ موضوع را مدنظر قرار دادیم و سه مدل تحلیل موضوعی ساختیم. در مرحلهٔ بعد، موضوعات را از جنبهٔ زمان به‌عنوان یک متغیر که در روند انتشار مقالات وجود داشته و غیرقابل انکار است، بررسی کردیم. در بررسی زمانی موضوعات مشخص شد که در طول زمان ممکن است همگرایی در موضوعات رقم بخورد، مانند موضوعات آموزش و روان‌شناسی زبان، که این مسئله براساس تحلیل‌های انجام‌شده به‌صورت بصری مورد بررسی قرار گرفت و از نظر کیفی مطالعه شد. همچنین موضوعاتی که در گذر زمان بیشتر از سایر موضوعات مورد توجه پژوهشگران این حوزه بود مشخص و بسامد آنها از پیکرهٔ برچسب‌گذاری‌شدهٔ دستی استخراج شد. دستاورد کاربردی این پژوهش سیاست‌گذاری در حوزهٔ علم است که علاوه‌بر مطرح‌کردن یک روش‌شناسی فناورانهٔ کاربردی در پژوهش، می‌توان موضوعات داغ میان پژوهشگران یک رشتهٔ علمی را مشخص کرد و خلأهای موضوعات پژوهشی را یافت و بر متنوع‌سازی و متوازن‌سازی موضوعات پژوهشی اهتمام ورزید.

منابع

احدی، حوریه (۱۴۰۰). کاربست علم زبان‌شناسی در حل مشکلات کودکان دارای اختلالات رشدی: مرور

نظام‌مند و فراتحلیل معیارهای زبانی و فرازبانی این کودکان و ارائه پیشنهادهایی جهت تدوین کتب آموزشی مناسب آنها. گزارش فنی. تهران: پژوهشگاه علوم انسانی و مطالعات فرهنگی.

افشارنیا، سعید و نجف الهیاری‌فرد (۱۳۸۵). «تبیین جایگاه علم و فناوری (بررسی وضعیت ایران و مقایسه آن با سایر کشورها) پیش‌نیاز اساسی تدوین استراتژی و ترسیم مسیر رشد و بهبود نظام علم و فناوری». مجموعه مقالات کنفرانس بین‌المللی استراتژی‌ها و تکنیک‌های حل مسئله. تهران.

سواری، کریم و شمس‌الله بهمنی (۱۳۸۹). «آسیب‌شناسی تولید علم در موسسات و مراکز آموزشی کشور». همایش ملی مدیریت پژوهش و فناوری. تهران: دانشگاه امام صادق.

علایی ابوذر، الهام و همکاران (۱٤۰۰). «معرفی یک پیکره متنی تخصصی: پیکره پژوهشنامه». مجله پژوهش‌های زبان‌شناسی تطبیقی. س۱۱، ش۲۲، ۲۷۱-۲۸۹.

قیومی، مسعود (۱۳۹۷). «ارائه یک روش مبتنی‌بر مدل زبانی برای واحدسازی پیکره فارسی». زبان و زبان‌شناسی. س۱٤، ش۲۷، ۲۱-۵۰.

قیومی، مسعود و مریم موسویان (۱٤۰۱). «کاربرد یادگیری ماشینی مبتنی‌بر شبکه عصبی برای دسته‌بندی مستندات علمی». پژوهشنامه پردازش و مدیریت اطلاعات. س٤، ش۳۷، ۱۲۱۷-۱۲٤٦.

کامیابی‌گل، عطیه و همکاران (۱۳۹۷). «استخراج اطلاعات از پیکره زبانی: معرفی پیکره مقالاه‌های علمی-پژوهشی دانشگاه فردوسی مشهد». کتابداری و اطلاع‌رسانی. س۲، ش۲۱، ۳-۲۵.

ناصح، محمدامین (۱۳۸۰). فهرست پایان‌نامه‌های کارشناسی ارشد و دکتری در زمینه گویش‌های ایران. تهران: فرهنگستان زبان و ادب فارسی.

ناصح، محمدامین (۱۳۸۳). فهرست پایان‌نامه‌های دانشگاهی در زمینهٔ دستور زبان فارسی. تهران: فرهنگستان زبان و ادب فارسی.

ناصح، محمدامین (۱۳۸٦الف). فهرست پایان‌نامه‌های دانشگاهی در عرصه زبان و ادب فارسی و مسائل زبان‌شناسی. ضمیمه مجله ۳۱ نامه فرهنگستان. تهران: فرهنگستان زبان و ادب فارسی.

ناصح، محمدامین (۱۳۸٦ب). چکیدةپایان‌نامه‌های حوزه زبان و زبان‌شناسی. تهران: انتشارات دانشگاه علامه طباطبائی.

یارمحمدی، لطف‌الله، علی‌محمد حق‌شناس و رضا نیلی‌پور (۱۳۷۷) بررسی وضعیت علم زبان‌شناسی در ایران. گزارش فنی، فرهنگستان علوم جمهوری اسلامی ایران.

یارمحمدی، لطف‌الله، علی‌محمد حق‌شناس و رضا نیلی‌پور (۱۳۷۸) «بررسی وضعیت علم زبان‌شناسی در ایران». نامه فرهنگ بهار. ۳٤، ۱۱۷-۱۲۵.

Blei, D. M. et al. (2003). "Latent Dirichlet allocation". *Journal of Machine Learning Research*. 3: 993–1022.

Blei, D. M. & J. D. Lafferty (2006). "Dynamic topic models". *Proceedings of the 23rd International Conference on Machine Learning*. W. Cohen, & A. Moore (eds.), Pittsburgh, PA, 113-120.

Farahani, M. et al. (2021). "ParsBERT: Transformer-based model for Persian language understanding". *Neural Processing Letters*. 53: 3831–3847.

Griffiths, T. L. & M. Steyvers (2004). "Finding scientific topics". Proceedings of the National Academy of Sciences. 5228–5235.

Hofmann, T. (1999). "Probabilistic latent semantic indexing". *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research & Development in Information Retrieval.* F. Gey, M. Hearst, & R. Tong (eds.), California, Berkeley, USA, 211-218.

Hughes, L. (2015). "Digital humanities, big data, and new research methods". *Presentation at the Workshop on Digital Music Lab* - Analyzing Big Music Data.

Minka, T., & J. Lafferty (2002). "Expectation-propagation for the generative aspect model". *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*. A. Darwiche, & N. Friedman (eds.), San Francisco, CA, USA, Morgan Kaufmann Publishers Inc, 352–359.

Papadimitriou, C. et al. (2000). "Latent semantic indexing: A probabilistic analysis". *Journal of Computer and System Sciences*. 61(2): 217-235.

Sievert, C. & K. Shirley (2014). "LDAvis: A method for visualizing and interpreting topics". *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. J. Chuang, S. Green, M. Hearst, J. Heer, & P. Koehn (eds), Baltimore, Maryland, USA, Association for Computational Linguistics, 63–70.

Wang, C., D. Blei, & D. Heckerman (2008). "Continuous time dynamic topic models". *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*. D. McAllester, & P. Myllymaki (eds.), AUAI Press, Arlington, Virginia, USA, 579–586.

Wang, X. & A. McCallum (2006). "Topics over time: A non-Markov continuous-time model of topical trends". *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. T. Eliassi-Rad, G. Chair, L. Ungar, M. Craven, & D. Gunopulos (eds.), ACM, 424-433.

Zhu, M., X. Zhang, & H. Wang (2016). "A LDA based model for topic evolution: Evidence from information science journals". *Modeling, Simulation and Optimization Technologies and Applications, Advances in Computer Science Research*. 58: 49-54.

Zosa, E. & M. Granroth-Wilding (2019). "Multilingual dynamic topic model". *Proceedings of the International Conference on Recent Advances in Natural Language Processin*. R. Mitkov, & G. Angelova (eds.), Varna, Bulgaria. INCOMA Ltd., 1388–1396.