Journal of Teaching Persian to Speakers of Other Languages
Vol. 8, No. 1 (Tome 17), Spring & Summer 2019
Print Issn:5394-2322
Electronic Issn:2676-3354

Extraction of Core Medical Terms Using Frequency Approach

Zohreh Zolfaghar Kondori[1]
PhD Candidate in General Linguistics, Payam-e-Noor Higher Education Center, Tehran, Iran.
Tayebeh Mosavi Miangah[2]
Corresponding author, Associate Professor, Department of General Linguistics, Payam-e-Noor University, Yazd, Iran.
Belgheys Rowshan[3]
Associate Professor, Department of General Linguistics, Payam-e-Noor University, Tehran, Iran.

Abstract:

Over the past few decades, with the advancement of technology, the use of corpora in linguistic studies has dramatically increased. Linguistic corpuses provide linguistic experts with the possibility to apply different methods for linguistic analysis by providing large collections. Most of the studies that have been done so far have been in English, French, and Japanese, and limited research has been conducted in Farsi language, and this lack, especially in specialized fields such as medical sciences, mathematics, science, tourism and so on is so tangible. So far most of the term or vocabulary extractions in Farsi have been done by using non-automatic methods and through reading and collecting data by the researchers; however, due to the technical properties of Farsi language, using non-Farsi term extractors which have been quite successful in other languages such as English, French and Japanese, have been impossible to use in Farsi so far. This is because of the particularities and specific features of languages. Each of these extractors is defined based on the features and properties of language they have been used for. In order to improve teaching materials in Farsi, paying attention to this problem was of paramount importance and we decided to apply some of these extraction methods and devise an extraction method for Farsi language which works properly. Since Iran's universities admit a lot of non-native Farsi international students annually whose goal is to study at fields such as medicine, engineering and humanities, preparing standard modern teaching materials in Farsi, which are based on the most modern technologies, is significantly important .The purpose of this study was to improve the resources used in teaching Farsi language at university levels, especially for non -native Farsi speakers and to explore the feasibility of using frequency-based methods in the automatic extraction of core medical terms and comparing the capabilities of each method. Findings of the research reveal the strengths and weaknesses of these methods in Farsi language and explore the possibility of using

Received on: 29/08/2018 Accepted on: 30/03/2019
[1]. Email: zohreh.zolfaghar@gmail.com
[2]. Email: mosavit@pnu.ac.ir
[3]. Email: bl_rowshan@pnu.ac.ir
DOI: 10.30479/jtpsol.2019.9257.1393
pp.227-244

each of these methods in Farsi and provide technical solutions for the improvement of the results.

Research Methodology:

The frequency counting approaches utilized in this study included the general and a specialized corpus which was created by the researcher. The general corpus used in this study was the Hamshahri Corpus and the specialized researcher made corpus included: texts from the science books of grades 1-4 of senior high schools and grades 1-3 of junior high schools in Iran, science courses in Imam Khomeini Farsi language center, general medicine texts from journals and internet. After the formation of the corpus, preparation and tokenization, the research introduced two methods of frequency i.e. classical and modern categories. Then, in the next step, the capabilities of each method were compared. The methods used in the classical frequency approach were the frequency of the main general corpus, the frequency of the specialized corpus and their improved approaches. Also, modern methods used in the research were: PMI and Chi-square. Pearson correlation analysis and trend analysis were also used to compare the methods used in the research.

Research findings

The results showed that classical methods in their general form, have little accuracy in identifying specialized vocabulary, however, by applying some techniques, it was possible to improve the process of selecting specialized vocabulary, among which the best performance related to the improved numerical method in the specialized corpus which resulted in extracting 60% of the specialized vocabulary in the first 50 high-frequency words. This result improved by increasing the scope of the study to 100, 150 and 200 first extracted words and it was observed that the percentage of specialized vocabulary identified increased by about 75%. Moreover, the results obtained for modern methods indicated that these methods can be used in Farsi. It can be seen that chi-square method with 32% and PMI method with 52% extraction of specialized vocabulary in the first 50 high frequency words showed a good function in automatic term extraction in Farsi. They automatically detected specialized vocabulary and by increasing the scope of the study to 200 first words, these percentages improved.

Conclusion:

The results of the research showed that frequency-based methods are applicable in Farsi. If we use classic frequency methods, we will need to utilize improved classic frequency methods in order to increase the accuracy of extracted words. Also, in order to achieve reliable results in modern frequency approaches, it is necessary to choose large enough vocabulary scope for the extracted vocabulary.

Key words: Automatic extraction of medical terms, corpus, Mixed extraction approaches, Teaching Persian language

پژوهش‌نامه‌ی آموزش زبان فارسی به غیر فارسی‌زبانان (علمی- پژوهشی)
سال هشتم، شماره‌ی اول (پیاپی۱۷)، بهار و تابستان ۱۳۹۸
شاپا چاپی: ۵۳۹۴-۲۳۲۲
شاپا الکترونیکی: ۲۳۵۴-۲۶۷۶

مقایسه روش‌های کلاسیک و روش‌های مبتنی بر اندازه‌های آماری پیکره‌بنیاد در استخراج خودکار واژه‌های پایه علوم پزشکی به روش بسامدی

زهره ذوالفقارکندری[۱]
دانشجوی زبان‌شناسی همگانی، مرکز تحصیلات تکمیلی پیام نور تهران
طیبه موسوی میانگاه [۲]
نویسنده‌ی مسئول، دانشیارگروه زبان‌شناسی همگانی، دانشگاه پیام نور واحد یزد
بلقیس روشن[۳]
دانشیارگروه زبان‌شناسی همگانی، دانشگاه پیام نور واحد تهران

چکیده

طی دو دهه‌ی اخیر با پیشرفت علم و فناوری، استفاده از روش‌های پیکره‌بنیاد در آموزش زبان و تدوین منابع درسی گسترش چشمگیری داشته است. پژوهش حاضر با هدف دستیابی به روشی خودکار در استخراج واژه از پیکره‌ها در زبان فارسی صورت گرفته است. برای دستیابی به هدف پژوهش روش‌های بسامدشماری در دو گروه کلاسیک و روش‌های مبتنی بر اندازه‌های آماری موردبررسی قرارگرفته و توانمندی هریک از که عبارتند از بسامدشماری پیکره‌ی عمومی، بسامدشماری پیکره‌ی تخصصی و روش‌های بهبودیافته‌ی آن‌ها موردمقایسه قرارمی گیرند. نتایج نشان می‌دهد که در روش‌های کلاسیک با اعمال تکنیک‌هایی می‌توان فرایند انتخاب واژه‌های تخصصی را بهبود بخشید و در این میان بهترین عملکرد مربوط به روش بسامدشماری بهبودیافته در پیکره‌ی تخصصی بوده است. روش‌های به‌کاررفته در پژوهش عبارتند از اطلاعات متقابل نقطه‌ای و مجذور کا[۴]. نتایج به‌دست آمده برای این دو روش نیز قابلیت استفاده از روش‌های بسامدشماری پیکره‌بنیاد در زبان فارسی را مورد تأیید قرار می‌دهد. روش مجذور کا با استخراج ۳۲٪ واژه‌ی تخصصی و روش اطلاعات متقابل نقطه‌ای با استخراج ٪۵۲ واژه‌ی تخصصی، عملکرد مناسبی در تشخیص خودکار واژه‌های تخصصی از خود نشان می‌دهند. نتایج حاصل از اعمال این روش‌ها روی پیکره‌ها و مقایسه آن‌ها نشان می‌دهند که می‌توان از روش‌های مبتنی بر اندازه‌گیری‌های آماری برای استخراج خودکار واژه در زبان بهره جست و به این ترتیب تحولی نوین در تهیه و تدوین متون آموزشی حاصل خواهد شد و آموزش‌دهندگان می‌توانند به فهرست واژگانی دسترسی داشته باشند که دانستن آن برای زبان آموزانشان مفید و گاه ضروری است.

کلیدواژه‌ها: استخراج خودکار واژه‌های پزشکی، پیکره، روش‌های ترکیبی استخراج، آموزش زبان فارسی

تاریخ دریافت مقاله: ۱۳۹۷/۰۶/۰۷ تاریخ پذیرش نهایی مقاله: ۱۳۹۸/۰۱/۱۰

[۱]. رایانامه: zohreh.zolfaghar@gmail.com
[۲]. رایانامه: mosavit@pnu.ac.ir
[۳]. رایانامه: bl_rowshan@pnu.ac.ir

شناسه دیجیتال (DOI): 10.30479/jtpsol.2019.9257.1393
صص: ۲۴۴-۲۲۷

[4]. PMI & Chi-square

۱. مقدمه

نزدیک به سه دهه است که استخراج خودکار واژه مورد توجه پژوهشگران در علوم گوناگون بوده است. در اوایل دهه‌ی ۹۰ میلادی به‌کارگیری پیکره‌های متنی بزرگ به ایجاد نخستین برنامه‌های استخراج اصطلاح و یا واژه کمک کردند (واژه و اصطلاح در روش‌های استخراج به یک شیوه استخراج می‌شوند و در مطالعات دیده می‌شود که مواردی که تحت عنوان اصطلاح استخراج می‌شوند در واقع همان واژه هستند و تفاوتی ندارند). برای مثال مترجمان جهت انتخاب بهترین معنی برای یک واژه‌ی چندمعنایی[1] با بسامدشماری، بر اساس بیشترین بسامد مناسب‌ترین معنی را رتبه‌بندی می‌کنند. مفسرین با شمارش واژه‌های خاص به‌کاررفته در متن یک سخنرانی به اهمیت موضوعات و محتوای آن پی می‌برند. مهندسین علوم رایانه و نرم‌افزار در موتورهای جستجوی اینترنتی هنگام جستجو برای واژه‌های پیشنهادی، بر اساس واژه‌هایی با بیشترین بسامد در جستجوهای پیشین، واژه احتمالی را پیشنهاد می‌دهند.

همچنین پژوهشگران نشان داده‌اند که این شاخه‌ی نوین در زبانشناسی می‌تواند به خوبی و به شکلی کار آمد جهت امور آموزشی مورد استفاده قرار گیرد. دست اندرکاران آموزش زبان نیز از این ابزار بسیار توانمند در زبان‌شناسی بهره بسیار برده‌اند. برای نمونه امروزه در حوزه‌ی آموزش زبان‌های خارجی و تخصصی، می‌توان از بسته‌های آموزشی موفق بسیاری که به کمک روش‌های پیکره‌ای تهیه شده‌اند نام برد. از شناخته شده‌ترین بسته‌های آموزشی که به موفقیت تجاری بالایی نیز دست یافته‌اند، می‌توان به مجموعه کتاب‌های «واژه‌های ضروری زبان انگلیسی برای آزمون تافل و یا آیلتس» و بسته‌های مشابه آن اشاره نمود. پیکره‌ها امکان استفاده از داده‌های زبانی را در مقیاسی بزرگ بر اساس متون واقعی و موثق فراهم می‌آورند. شناسایی و استخراج واژه‌های تخصصی پرکاربرد به شیوه‌ای علمی و بدون اعمال نظر شخصی در آموزش زبان‌های خارجی و تخصصی، از اهمیت بالایی برخوردار است. از مزایای استخراج خودکار واژه‌ها از پیکره‌ها به روش بسامدشماری می‌توان به استفاده‌ی دامنه‌ی گسترده‌ای از واژه‌ها، به‌روزرسانی متن‌ها با گذر زمان، بررسی اثر پویایی زبان، بالا رفتن سرعت تهیه‌ی متون آموزشی، تهیه‌ی متون آموزشی کاربردی و نظایر آن اشاره کرد که تهیه‌ی آن از توانایی نیروی انسانی به تنهایی خارج است. هرچند پژوهشگران زبان فارسی طی چند سال گذشته اقدام به بررسی امکان استفاده از روش‌های بسامدی در زبان فارسی کرده‌اند، اما تلاش‌های صورت گرفته در این زمینه تاکنون بسیار محدود بوده است. مسئله‌ی اصلی در این زمینه داشتن متن‌های آموزشی مناسب و استاندارد است، بدین معنا که متن‌هایی با محتوای مناسب بر اساس داده‌های واقعی و پیکره‌ایِ زبان فارسی به‌صورت منظم و ساخت‌یافته تهیه شوند تا بتوان آن‌ها را به‌عنوان مرجع آموزش زبان تخصصی قرار داد.

---

[1]. polysemy

یکی از کمبودهای مسلم در حوزه‌ی آموزش زبان فارسی عدم دسترسی زبان‌آموزان و نیز اساتید آموزش زبان فارسی به واژه‌های تخصصی حوزه‌ی مورد مطالعه‌شان و یا واژه‌ها و متن‌های سطح‌بندی شده و یا طبقه‌بندی شده در زمینه‌های گوناگون است. اهمیت نقش واژه‌های انتخاب شده به‌ویژه در دروس تخصصی جهت تدریس بر کمتر کسی پوشیده است و پژوهش‌هایی از این دست در زبان‌هایی چون انگلیسی و فرانسه و ژاپنی بسیار فراوان هستند، اما تاکنون پژوهش جدی و عمیقی در زمینه‌ی تهیه این نوع واژه‌های طبقه‌بندی شده در زبان فارسی انجام نشده است. برای مثال، مراکز آموزش زبان فارسی جهت تدریس به داوطلبان حضور در رشته‌های پزشکی در ایران، از متونی استفاده می‌کنند که معیار مشخصی در تعیین سطح آن‌ها وجود ندارد و ممکن است با سطح زبان‌آموز همسو نباشند و درنتیجه منجر به کاهش بازدهی آموزش شوند و یا این‌که زبان‌آموز در پایان دوره احساس کند که نیازها و اهداف آموزشی‌اش تأمین نشده است. از این رو، مدرسان زبان فارسی نیازمند آگاهی از واژه‌های پرکاربرد و کلیدی در موضوع مورد تدریس هستند تا امکان تهیه‌ی مواد درسی معتبر و کارآمد را داشته باشند. همچنین آگاهی از این واژه‌ها می‌تواند مبنای سطح‌بندی توانایی زبان‌آموزان در رده‌های مبتدی، متوسط و پیشرفته قرار گیرد و امکان آشنایی جامع‌تر زبان‌آموزان را با واژه‌های پرکاربرد در سطوح بالاتر فراهم آورد.

بنابراین در این پژوهش در پی آن هستیم که به روشی در استخراج خودکار واژه دست یابیم که با ویژگی های ساختاری زبان فارسی مطابقت داشته باشد. برای این کار از بین روش‌های مورد استفاده روز دنیا، روش های مورد استفاده از سوی محققان ژاپنی را که بسیار موفق بوده است برگزیدیم. شوجو، یوتیاما و ناکامارو توانستند در پژوهش‌های خود به استخراج واژه در سطوح و حوزه‌های مختلف دست یابند از جمله استخراج واژه‌های علوم و تکنولوژی درسطوح خاص از پیکره انگلیسی CPE[1] .

۲. چارچوب نظری

با ظهور زبان‌شناسی رایانشی[2] و پیکره‌ای، شاهد گسترش روش‌های تجزیه و تحلیل پیکره‌ای در زبان‌های گوناگون هستیم. محققان بزرگی در رشد و توسعه‌ی زبان‌شناسی پیکره‌ای تاثیرگذار بوده‌اند، زبان‌شناسانی چون لیچ، بایبر، جوهانسون، فرانسیس، کنراد، و مک‌کارتی. بسیاری از زبان‌شناسان جان سینکلر را یکی از تأثیرگذارترین افراد در زبان‌شناسی نوین می‌دانند. (Sinclair,1998) بر این باور است که یک واژه به خودی خود و به تنهایی معنا ندارد، بلکه این معنا را از طریق توالی چندین واژه به‌دست می‌آورد. همین تفکر اوست که پایه و اساس زبان‌شناسی پیکره‌ای را تشکیل می‌دهد. به اعتقاد او زبان‌شناسی پیکره‌ای، دیدگاهی عینی‌تر

[1]. Extracting Level-Specific Science and Technology Vocabulary from the Corpus of Professional English (CPE)
[2]. computational linguistics

از زبان ارائه می‌دهد و این موضوع به دلیل آن است که گویشوران به الگوهای نیمه‌خودآگاهِ تولید شده از طریق زبان دسترسی ندارند، اما یک تحلیل پیکره‌بنیاد می‌تواند تقریباً هر نوع الگوی زبانی از جمله واژگانی، ساختاری، واژی ـ دستوری، کلامی، واجی و صرفی را بررسی کند و این کار را با دستورالعمل‌های بسیار خاصی همچون تفاوت میان کاربردهای زنان و مردان انجام می‌دهد.

عمده پژوهش‌های صورت گرفته در این حوزه در زبان‌هایی همچون فرانسه، انگلیسی و به ویژه ژاپنی بوده است و استخراج خودکار واژه یکی از اهداف اصلی این پژوهش‌ها بوده است. برای این کار روش‌های گوناگونی وجود دارد که از شناخته شده‌ترین آن‌ها، می‌توان به بسامدشماری اشاره نمود. روش‌های بسامدشماری اولیه که با عنوان «روش‌های بسامدشماری کلاسیک» شناخته می‌شوند با تمرکز بر روی یک پیکره اقدام به استخراج واژه می‌نمایند. توجه به این نکته ضروری است که برای استفاده از روش‌های پیکرهبنیاد، نیاز است تا مجموعه-ای از اصلاحات و ویرایش‌های اولیه بر روی پیکره‌ها اعمال گردد تا پردازش آن‌ها توسط رایانه میسر و نتایج حاصل از آن‌ها قابل استناد گردد. بنابراین جهت خوانش و شمارش پارامترهای مورد نیاز مانند جمله‌ها، عبارات و واژه‌ها، پیکره‌ها باید فاقد هر نوع اشتباه نگارشی، املایی و نظایر آن باشند. همچنین برای کاهش خطای محاسباتی و تسریع فرایند استخراج، شکل‌ها و جدول‌ها نیز باید پیش از آغاز فرایند پردازش از پیکره حذف گردند.

۳. پیشینه‌ی پژوهش

از شاخص‌ترین پژوهش‌های صورت گرفته در زمینه‌ی بسامد شماری کلاسیک می‌توان به تلاش‌های (Nation, 2001) اشاره کرد. وی در پژوهش خود در زمینه‌ی استخراج واژه‌های دانشگاهی، واژه‌ها را به چهار دسته پربسامد، دانشگاهی، فنی و کم‌بسامد تقسیم‌بندی کرد و با استفاده از بررسی بسامد و دسته‌بندی مفروض اقدام به استخراج واژه‌های دانشگاهی کرد. گروه دیگری از روش‌ها به نتایج حاصل از یک پیکره اکتفا نکرده و با استفاده از دست‌کم دو پیکره اقدام به استخراج واژه می‌کنند. این روش‌ها با عنوان «روش‌های بسامدشماری مبتنی بر اندازه‌های آماری» شناخته می‌شوند. (Enguehard et al.,1994) با استفاده از شاخص اطلاعات متقابل که روشی است بر پایه‌ی مقایسه دو پیکره، اقدام به استخراج خودکار واژه در زبان انگلیسی نمودند. (Daille,1994) از ترکیب دانسته‌های زبانی و روش‌های آماری به استخراج الگوهای نحوی واژه در زبان انگلیسی پرداخت. (Nakagawa et al., 2002) با کمک سیستم رتبه‌دهی، موفق به استخراج واژه از پیکره‌های زبان ژاپنی شدند. (Frantzi et al.,2000) با استفاده از روش‌های بسامدشماری و تلفیق روش‌های زبان‌شناسی و آماری به شناسایی عبارت‌های چند واژه‌ای پرداختند. (Swales, 2004) نشان داد که تحلیل‌های پیکره‌ای بر مبنای روش‌های آماری فراتر از کارهای نظری بوده و در کاربرد نیز می‌توان از آن

در زبان انگلیسی استفاده کرد. در مثالی از تلاش‌های سازمانی و گسترده در زمینه‌ی مطالعات زبانشناسی انتشارات (Harper & Collins, 2005) از روش اطلاعات متقابل نقطه‌ای در تحلیل‌های پیکره‌ای خود استفاده کردند. (Chujo et al., 2006) به تعیین سطوح مختلف واژه‌های تخصصی با استفاده از روش‌های آماری پرداختند و (Gries ,2010) به معرفی آماره‌های کاربردی در زبان‌شناسی پیکره‌ای پرداخت.

از منظری دیگر به دلیل کاربردهای فراوان تحلیل‌های پیکره‌بنیاد از یک سو و دشواری‌های اجرایی و مالی جهت ایجاد پیکره‌ها از سوی دیگر، امروزه دولت‌ها در سطوح ملی و کلان اقدام به تهیه پیکره‌های زبانی نموده‌اند تا مبنای پژوهش‌های آتی قرار گیرند. از شناخته شده‌ترین این پیکره‌ها می‌توان به پیکره‌ی ملی بریتانیا[1] پیکره‌ی ملی زبان روسی [2]، پیکره‌ی ملی لهستانی[3]، و پیکره‌ی ملی استرالیا[4] اشاره نمود.

(Reppen, 2010) نشان داد که پیکره‌ها می‌توانند در امر آموزش زبان نیز به کمک مدرسان زبان خارجی بیایند. پیکره‌هایی که در بخش آموزش زبان تهیه می‌شوند، معمولاً خردتر بوده و با اهداف خاصی تهیه می‌شوند. مانند پیکره‌ی تاریخی زبان انگلیسی، پیکره‌های متون ادبی و پیکره‌های متون ترجمه شده و نظایر آن. (Granger,2015) بیان می‌دارد که چنین پیکره‌هایی این امکان را به پژوهشگران می‌دهند تا به بررسی دقیق نوع و میزان استفاده از واژه‌ها و ساختارهای دستوری توسط زبان‌آموزان بپردازند و با بهره‌گرفتن از آن‌ها، می‌توان به مقایسه‌ی جنبه‌های مختلف زبانی از قبیل انواع اشتباهات، بسامد استفاده از انواع واژه‌ها، عبارات، جمله‌واره‌ها و ساختارهای دستوری در بین گروه‌های مختلف زبان‌آموزان پرداخت.

هرچند پژوهش‌های پیکره‌بنیاد در زبان فارسی بسیار محدود هستند، اما طی یک دهه‌ی گذشته رشد چشمگیری داشته‌اند. از شاخص‌ترین این پژوهش‌ها می‌توان به تلاش‌های (Nemat Zadeh, 2013) اشاره کرد که به بررسی واژه‌های پایه در کتاب ریاضی اول دبستان پرداخته است و با طبقه‌بندی واژه‌ها و مقایسه‌ی آن‌ها با واژه‌های پایه به معرفی واژه‌های مناسب‌تری برای کتاب درسی ریاضی می‌پردازد. (Sepehri,2006) در پژوهش خود با عنوان «بسامدنگاری و دستاوردهای آن در آموزش» نشان داد که بسامدنگاری به مثابه یک فناوری بدیع آموزشی است که متخصصان زبان از طریق آن به بعضی از اهداف خود در مطالعات زبانی و آموزش زبان نائل خواهند شد. (Rasooli et al., 2008) در پژوهش خود با عنوان «روشی جدید در خطایابی املایی در زبان فارسی» به بررسی روش‌های مختلف خطایابی واژه‌ها در زبان فارسی و روشی برای یافتن خطاهای املایی واژه‌ها پرداختند. آن‌ها با اشاره به روش‌های مختلف برای خطایابی واژه‌ها، چالش‌ها و مشکلات پیش رو برای این روش‌ها را در زبان فارسی تشریح کردند و به مواردی همچون ویژگی‌های خاص

---

1. BNC
2. RNC
3. NKJP
4. AusNC

زبان فارسی، مشکلات موجود در رسم‌الخط رایانه‌ای زبان فارسی و حروف دارای چند نوع رسم‌الخط رایانه‌ای پرداختند. همچنین پیشنهادهای سازنده‌ای به کاربران در مورد بررسی و نحوه‌ی پیاده‌سازی برنامه‌ای برای پیدا کردن واژه‌های نادرست از واژه‌های قبلی و بعدی واژه‌ی مورد نظر و خود واژه‌ی نادرست ارائه کردند.

(Jahangardi et al., 2016) در مقاله‌ای با عنوان «واژه در کتاب‌های آموزش زبان فارسی به غیرفارسی‌زبانان: پژوهش پیکره‌بنیاد» با استفاده از روش‌ها و ابزارهای زبان‌شناسی پیکره‌ای، به سنجش میزان هم‌پوشانی و انطباق واژه‌های ارائه شده در کتاب‌های آموزش زبان فارسی به غیرفارسی‌زبانان، با پربسامدترین واژه‌های زبان فارسی پرداختند. آن‌ها براساس متن‌های موجود در پایگاه داده‌های زبان فارسی، یک پیکره‌ی زبانی متوازن طراحی کرده و پربسامدترین واژه‌های آن را مبنای کار قرار دادند. نتایج و یافته‌های پژوهش آن‌ها نشان می‌دهد که به لحاظ سطوح زبان‌آموزی، میزان هم‌پوشانی واژه‌های هریک ازگروه‌های مورد بررسی با گروه‌های متناظر آن‌ها در پیکره‌ی مبنا بسیار پایین است. از آنجا که در زبان فارسی پژوهش جامعی در حوزه‌ی استخراج واژه‌های تخصصی از پیکره‌ها صورت نگرفته است، بنابراین پژوهش حاضر را می‌توان پژوهشی نوین در این زمینه دانست.

۴. روش پژوهش

تمرکز اصلی این پژوهش بر تحلیل داده‌های پیکره‌ی پزشکی استوار است. از آنجاکه تا پیش از این چنین پیکره‌ای به‌صورت تخصصی در زبان فارسی وجود نداشته است، محقق با جمع‌آوری متونی که به‌صورت رایج با عنوان درس زیست‌شناسی و علوم تجربی در آموزش استفاده می‌شوند، پیکره‌ی مورد بررسی را تهیه کرد. دلیل انتخاب و گردآوری این متون از آن است که دانش‌آموزان فارسی‌زبان با مطالعه همین متن‌ها با واژه‌های ابتدایی و پایه‌ی کلیه رشته‌های علوم پزشکی آشنا می‌شوند و هنگام ورود به دانشگاه در یکی از رشته‌های علوم پزشکی، این واژه‌ها به‌عنوان واژه‌های تخصصی و در عین حال، اولیه و ابتداییِ این رشته‌های تحصیلی در مجموعه واژه‌های ذهن آن‌ها وجود دارد. بنابراین، به نظر محقق زبان‌آموزان غیرفارسی‌زبان نیز به هنگام ورود به دانشگاه‌های ایران و تحصیل در رشته‌های علوم پزشکی و زیرشاخه‌های آن، باید مجموعه واژگانی دانش‌آموزان فارسی‌زبان را در ذهن داشته باشند و یا با آن آشنا باشند تا از آمادگی بهتری برای یادگیری برخوردار باشند. از این رو، آنچه که فارسی‌زبانان می‌آموزند را مبنای کار قرار داده و پیکره را گردآوری کردیم. با این اوصاف، منابع زیر پیکره‌ی ایجاد شده از سوی محقق را تشکیل می‌دهند:

- متون دروس زیست‌شناسی چهار ساله دبیرستان (سال تحصیلی ۹۳ـ۹۴)
- متون درس علوم دوم و سوم راهنمایی (سال تحصیلی ۹۳ـ۹۴)
- متون تدریس شده در مرکز آموزش زبان فارسی امام خمینی قزوین (سال تحصیلی ۹۳ـ۹۴)

مشخصات پیکره‌ی محقق ساخته در جدول ۱ نمایش داده شده است. مشاهده می شود که برای ایجاد پیکره‌ی عمومی [1] ۱۴۰۰ مگابایت و پیکره‌ی تخصصی ۵ مگابایت داده مورد استفاده قرار گرفته است که از نظر حجم داده‌ی به‌کاررفته در ایجاد پیکره از سطحی مطلوب برخوردار بوده و می‌توان از پیکره‌ی ایجاد شده برای تجزیه و تحلیل‌های بسامدشماری در تحلیل‌های کلاسیک و روش‌های مبتنی بر اندازه‌های آماری استفاده کرد.

جدول ۱. آماره‌های توصیفی پیکره‌ی عمومی و تخصصی محقق‌ساخته مورد بررسی در پژوهش

زبان برنامه نویسی	**Scala**
در محیط توزیع شده	**Spark**
حجم پیکره‌ی عمومی	**۱۴۰۰ مگا بایت**
حجم پیکره‌ی تخصصی	**۵ مگا بایت**

مراحل ایجاد پیکره بدین شرح است که در گام نخست پس از مراجعه به وب‌سایت شبکه‌ی ملی مدارس برای دریافت متن کتاب‌های درسی[2] ، متون استخراج شده با فرمت پی‌دی‌اف[3] مرور و اصلاح گردیدند. در گام بعدی این متن‌ها به فرمت قابل ویرایش وُرد[4] تبدیل شده و پس از آن با استفاده از نرم‌افزار سفارشی که برای همین منظور طراحی شده بود، متون مورد نظر پیش ویرایش شدند، بدین مفهوم که حروف و کلمات انگلیسی، کلمات تک‌حرفی، اعداد، شکل‌ها، جدول‌ها، نمودارها، نمادها، علامت‌های نگارشی، حروف اضافه، ربط و مانند آن حذف شدند. عمل نشانه‌گذاری[5] نیز با استفاده از همین نرم‌افزار صورت گرفت و تمام واژه‌های متن جداسازی شدند. به این صورت که فقط واژه‌ها با نشانه‌هایی چون «می» و «ها» و «ی» و مانند آن باقی مانده بودند که این موارد هم به صورت دستی حذف شدند و آنچه که در پیکره باقی ماند بیشتر واژه‌هایی بودند که از هم جدا و قابل پردازش آماری بودند.

در پژوهش‌های پیکره‌بنیاد پایایی[6]، داده‌های استراج‌شده از پیکره با در نظر گرفتن دو عامل اصلی، یعنی دقت و صحت برچسب‌گذاری پیکره و نیز سیستم کدگذاری پیکره، مورد ارزیابی قرار می‌گیرد. پیکره‌های مورد استفاده در پژوهش به‌صورت نیمه خودکار برچسب‌گذاری شده‌اند. بدین مفهوم که ابتدا با استفاده از سیستم فارسی- تگ (Rezai et al., 2017)، داده‌های مورد نیاز به‌صورت خودکار برچسب‌گذاری شدند تا مقولات واژگانی تمام واژه‌ها مشخص شوند. سپس، پیکره‌ی برچسب‌گذاری شده‌ی حاصل، تماماً به‌صورت دستی

1. Hamshahri Collection, version 2
2. www.roshd.ir
3. PDF
4. word
5. tokenization
6. reliability

بازبینی و اصلاح شد. علاوه بر این، مجموعه‌ی برچسبی [1] که در سیستم برچسب‌گذاری خودکار فارسی- تگ استفاده شده است، نسخه‌ی اصلاح شده از مجموعه‌ی برچسب پیکره‌ی زبانی بی‌جن‌خان[2] می‌باشد که در آن برچسب‌های فرعیِ سطح دو به بعد حذف شده‌اند تا هم با نیازهای تحقیقاتی پیکره‌بنیاد بیشتری مطابقت داشته باشد و هم دقت و صحت برچسب‌گذاری به بالاترین حد خود برسد (Vulanovic et al., 2018). در آخرین گام پس از بازبینی نهایی و رفع سایر نواقص، پیکره جهت بسامدشماری آماده شد.

۴ .۱. بسامدشماری

اساس کار بسامدشماری، شمارش بسامد یک متغیر یا ویژگی خاص در پیکره‌هاست و می‌توان بیان داشت که بسامد مبنای قضاوت درباره‌ی وضعیت متغیر مورد بررسی در پیکره است. در پژوهش حاضر متغیر مورد بررسی، واژه‌های حاضر در پیکره هستند که بسامد آن‌ها در پیکره‌ی محقق‌ساخته مورد شمارش قرار می‌گیرند. روش‌های بسامدشماری پس از طی مسیر تکاملی خود امروزه به دو دسته کلاسیک و روش‌های مبتنی بر اندازه‌های آماری طبقه‌بندی می‌شوند که در ادامه به معرفی آن‌ها می‌پردازیم.

۴. ۱،۱. بسامدشماری کلاسیک

همان‌طور که اشاره شد در تحلیل‌های مبتنی بر بسامدشماری، تأکید اصلی بر بسامد واژه‌های به‌کاررفته در یک پیکره است. روش‌های کلاسیک بسامدشماری با قرار دادن تمرکز تحلیل‌ها بر یک پیکره‌ی مشخص و تجزیه و تحلیل بسامد واژه‌ها در آن، اقدام به استخراج واژه می‌نمایند. در روش‌های کلاسیک به‌ویژه مدل‌های اولیه، واژه‌ها در پیکره‌های استاندارد مورد شمارش قرار گرفته و به ترتیب بسامد، مرتب‌سازی می‌شوند. مطالعات تجربی نشان می‌دهند که این رویه‌ی تحلیل بدون نقص نبوده و در فرایند استخراج خودکار مشکلاتی را پدید می‌آورد.

۴. ۱. ۲. بسامدشماری پیکره‌ی عمومی

شمارش بسامد واژه‌ها از نخستین روش‌های به‌کاررفته در بسامدشماری است. در این روش به شمارش واژه‌ها در یک پیکره‌ی عمومی پرداخته می‌شود و پس از آن واژه‌ها به ترتیب بسامد از بیشترین به کمترین مرتب‌سازی می‌شوند. در پیکره‌های عمومی می‌توان انتظار داشت واژه‌های غیرتخصصی که بیشترین کاربرد را دارند، مانند حروف ربط و نظایر آن از بسامد بیشتری نسبت به واژه‌های تخصصی برخوردار باشند و واژه‌های تخصصی با بسامد کمتر در انتهای جدول قرار گیرند.

---

[1]. tagset
[2]. http://dbrg.ut.ac.ir/Bijankhan/

۴. ۱. ۳. بسامدشماری پیکره‌ی تخصصی

در این روش بسامد واژه‌ها در یک پیکره‌ی تخصصی مورد بررسی قرار می‌گیرد. به دلیل نتایج تجربی به‌دست آمده از مطالعه‌ی پیکره‌ها، مشاهده‌ی واژه‌های عمومی و کم‌کاربرد با بسامدی مشابه بسامد واژه‌های تخصصی، می‌توانند موجب بروز اخلال و بالا رفتن خطا در هنگام استخراج خودکار واژه‌های تخصصی شده و دقت مدل را کاهش دهند. بنابراین، متخصصان جهت رفع این مشکل بررسی یک پیکره‌ی تخصصی را پیشنهاد داده‌اند.

۴. ۱. ۴. بسامدشماری پیکره‌ی عمومی بهبودیافته

همان‌طور که پیشتر اشاره شد از عمده‌ترین چالش‌ها در استخراج خودکار واژه‌ها، جداسازی واژه‌های عمومی از واژه‌های تخصصی پس از مرتب‌سازی به کمک بسامدشماری است. جهت به‌دست آوردن رویه‌ای نظام‌مند که فرایند استخراج را تسریع کند، روش بهبودیافته ابداع گردید. در این روش، بسامد کل واژه‌های موجود در پیکره‌ی عمومی، از مقدار بسامد واژه‌ی مورد بررسی در پیکره‌ی عمومی کسر می‌شود و نتایج به‌دست آمده به ترتیب صعودی مرتب‌سازی می‌شود. با این کار بهبود چشمگیری در بالا آمدن رتبه واژه‌های تخصصی در پیکره‌ی عمومی حاصل می‌گردد و این امر می‌تواند موجب تسریع فرایند انتخاب واژه‌ی تخصصی گردد و متخصصان جهت انتخاب واژه‌ها می‌توانند با سرعت بیشتری اقدام به انتخاب واژه‌های تخصصی نمایند.

۴. ۱. ۵. بسامدشماری پیکره‌ی تخصصی بهبودیافته

نتایج به‌دست آمده نشان می‌دهند که امکان به‌کارگیری روش‌های بهبودیافته در پیکره‌ی عمومی وجود دارد، از این رو، متخصصان اقدام به اجرای روش بهبودیافته در یک پیکره‌ی تخصصی نمودند. آن‌ها در این روش استدلال می‌کنند که به دلیل تخصصی بودن پیکره، واژه‌های تخصصی بیشتری در پیکره حضور دارد و بنابراین بهبود حاصل در مقایسه با اجرای روش در پیکره‌ی عمومی بیشتر است.

جدول ۲ چهار روش بسامد شماری کلاسیک به‌کاررفته در پژوهش را نمایش می‌دهد. همان‌گونه که در جدول نیز مشخص است، این روش‌ها شباهت زیادی به یکدیگر دارند و هریک جهت رفع مشکلات پدید آمده در کار با داده‌های واقعی ایجاد شده‌اند.

جدول ۲. شرح محاسبه روش‌های بسامدشماری کلاسیک

روش‌های کلاسیک	
روش	شرح
بسامد شماری پیکره عمومی	a = بسامد واژه‌ها در پیکره عمومی
بسامد شماری پیکره تخصصی	b = بسامد واژه‌ها در پیکره تخصصی
بسامد شماری پیکره عمومی بهبود یافته	c= بسامد کل واژه‌ها در پیکره عمومی ـ a
بسامد شماری پیکره تخصصی بهبود یافته	d = بسامد کل واژه‌ها در پیکره ی اصلی تخصصی ـb

۴. ۱. ۶. بسامدشماری مبتنی بر اندازه‌های آماری

با توجه به مشکلاتی که پیشتر در روش‌های کلاسیک به آن‌ها اشاره شد ابداع روش‌های کاراتر همواره مورد توجه پژوهشگران بوده است. از این رو، با تلفیق روش‌های کلاسیک و استفاده از تکنیک‌های آماری، خانواده‌ی جدیدی از تحلیل‌ها با عنوان روش‌های مبتنی بر اندازه‌های آماری در بسامدشماری پدید آمدند. اساس کار در بیشتر این روش‌ها مقایسه‌ی بسامد واژه‌ها یا توابع آماری از بسامد واژه‌های استخراج شده در دو پیکره‌ی مجزا است. به بیانی ساده‌تر، روش‌های مبتنی بر اندازه‌های آماری با اعمال توابع آماری، به دنبال تبیین شباهت‌ها میان الگوهای موجود در فهرست‌های استخراج شده از پیکره‌های مورد بررسی هستند. شناسایی الگوهای تکرار واژه در دو پیکره‌ی مورد بررسی (در پژوهش حاضر پیکره‌ی عمومی و تخصصی) در تعیین وضعیت واژه‌های استخراج شده بسیار مهم هستند. در جدول ۳ نحوه‌ی محاسبه‌ی دو روش اطلاعات متقابل نقطه‌ای و مجذور کا نمایش داده شده است. این شاخص‌ها برابری توزیع آماری واژه‌های استخراج شده در دو پیکره را مورد آزمون قرار می‌دهند.

جدول۳. شرح محاسبه روش‌های بسامدشماری مبتنی بر اندازه‌های آماری

روش‌های روش‌های مبتنی بر اندازه‌های آماری	
نام روش	**نحوه محاسبه**
مجذور کا	((a+b+c+d)*((a*d)-(b*c))^2/((a+b)*(c+d)*(a+c)*(b+d))
اطلاعات متقابل نقطه‌ای	log(a*(a+b+c+d))/log((a+c)*(a+b))

مقادیر بزرگ آماره به رد فرض صفر مبنی بر برابری بسامد واژه در دو پیکره‌ی عمومی منتهی می‌گردد که پس از مرتب‌سازی این آماره از مقادیر بزرگ به کوچک می‌توان واژه‌های مورد نظر را استخراج نمود. از آنجاکه آماره به‌دست آمده برای واژه‌های مرتب‌سازی شده جهت استخراج ۵۰ واژه نخست بسیار بزرگ بوده و در سطح آلفا ۵ درصد معنی‌دار هستند (سطح معناداری به‌دست آمده برای همه‌ی ۵۰ واژه‌ی استخراج شده، کمتر از ۰,۰۵ به دست آمده است و از آنجاکه این اعداد ثبت شده همگی برابر ۰,۰۰۰ هستند، بر رد فرض صفر مبنی بر تشابه بسامد واژه‌ها در دو پیکره دلالت دارند (برای پرهیز از پیچیدگی نمایش جدولی این اعداد به اختصار در جدول ۳ نمایش داده شده‌اند)، همچنین جدول‌های ۴ و ۵ که به تبیین واژه‌های استخراج شده می‌پردازند، با الهام‌گیری از جداول پژوهش مشابه محقق ژاپنی (Chujo et al., 2006) در زبان ژاپنی تهیه شده‌اند.

جدول۳. شرح محاسبه روش‌های بسامدشماری روش‌های مبتنی بر اندازه‌های آماری

ردیف	روش	سطح معنی‌داری ۵۰ واژه نخست پس از مرتب‌سازی در هر روش
۱	اطلاعات متقابل نقطه‌ای	۰,۰۰
۲	مجذور کا	۰,۰۰
۳	بسامدشماری پیکره‌ی تخصصی بهبود یافته	۰,۰۰
۴	بسامدشماری پیکره‌ی عمومی بهبود یافته	۰,۰۰
۵	بسامدشماری پیکره‌ی تخصصی	۰,۰۰
۶	بسامدشماری پیکره‌ی عمومی	۰,۰۰

۵. ارائه و واکاوی داده‌ها

۵,۱. تحلیل یافته‌های بسامدشماری کلاسیک

در روش بسامد شماری پیکره‌ی عمومی، بسامد واژه‌ها در پیکره‌ی اصلی مورد شمارش قرار گرفت و بر اساس بیشترین بسامد واژه‌ها مرتب‌سازی شدند. همان‌گونه که جدول نیز نمایش می‌دهد، واژه‌های غیرتخصصی در رتبه‌های بالای جدول قرار گرفته‌اند. نتایج به‌دست آمده از اجرای روش بسامدشماری در پیکره‌ی تخصصی، مشابه پیکره‌ی عمومی بوده و هر دو روش خروجی‌های نسبتاً مشابهی پس از بسامدشماری به دست می‌دهند. بنابراین می‌توان استدلال نمود که در پیکره‌های مورد بررسی به روش‌های ساده، واژه‌های غیرتخصصی در مقایسه با واژه‌های تخصصی بیشتر تکرار شده‌اند و بنابراین هنگام مرتب‌سازی بر اساس بسامد، جایگاهی بالاتر به خود اختصاص می‌دهند.

به منظور کاهش مشکلات پدید آمده و بالابردن توانایی‌های مدل‌ها، دو روش بسامدشماری پیکره‌ی عمومی بهبودیافته و بسامدشماری پیکره‌ی تخصصی بهبود یافته، مورد بررسی قرار گرفتند. همانطور که جدول ۴ نیز نمایش می‌دهد روش بسامدشماری پیکره‌ی عمومی بهبودیافته به دلیل طیف گسترده‌ی واژه‌های پیکره‌ی عمومی از عملکرد مطلوبی در شناسایی واژه‌های تخصصی (ده درصد) برخوردار نبوده و حتی نتایج ضعیف‌تری نسبت به روش اولیه در پیکره‌ی عمومی حاصل شده است. روش بسامدشماری پیکره‌ی تخصصی بهبودیافته، جهت بررسی همین رویکرد برای پیکره‌ی تخصصی مورد استفاده قرار گرفت. نتایج به‌دست آمده در جدول ۴ نشان می‌دهد که اعمال این روش تأثیر بسیار عمده‌ای در شناسایی واژه‌های تخصصی در پیکره‌ی تخصصی دارد تا آنجاکه اعمال آن موجب شناسایی صحیح ۶۴٪ واژه‌های تخصصی در ۵۰ واژه‌ی مرتب‌سازی شده با بیشترین بسامد شده است. این واژه‌ها را سه کارشناس از جمله یک پرستار، یک دانشجوی سال اول پزشکی و خود نگارنده (به عنوان فردی که مدرس درس زیست‌شناسی زبان‌آموزان غیرفارسی‌زبان بوده و با نیازهای آنان آشنایی دارد)، به‌عنوان واژه‌های تخصصی تأیید کردند. بنابراین نتایج به‌دست آمده دلالت بر این موضوع دارند که رویکرد بسامدشماری کلاسیک با اعمال اصلاحاتی در شناسایی واژه‌های تخصصی می‌تواند مورد

استفاده قرار گیرد و در میان روش‌های کلاسیک بهترین روش برای رسیدن به این هدف استفاده از روش بسامدشماری پیکره‌ی تخصصی بهبودیافته است.

۵. ۲. تحلیل یافته‌های بسامدشماری روش‌های مبتنی بر اندازه‌های آماری

یافته‌های به‌دست آمده برای روش‌های مبتنی بر اندازه‌های آماری دلالت بر موفقیت نسبی این روش‌ها دارد. با توجه به نتایج به‌دست آمده مشاهده می‌شود که پس از مرتب‌سازی ۵۰ واژه‌ی نخست، روش مجذورکا دارای دقت ٪۳۲ و روش اطلاعات متقابل نقطه‌ای دارای دقت ٪۵۲ در تشخیص صحیح واژه‌های تخصصی است. هرچند دقت هر دو روش در مقایسه با روش پیکره‌ی تخصصی با دقت ٪۶۴ کمتر است، اما پیشنهاد می‌شود در پژوهش‌هایی که تنها از روش‌های بسامدشماری مبتنی بر اندازه‌های آماری استفاده می‌شود از روش اطلاعات متقابل نقطه‌ای استفاده شود.

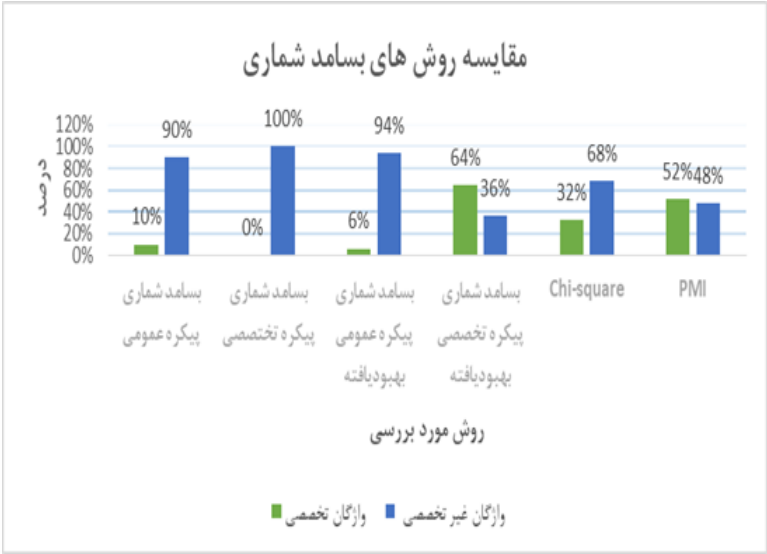
۵. ۳. مقایسه میان روش‌های بسامدشماری کلاسیک و روش‌های مبتنی بر اندازه‌های آماری

یافته‌های به‌دست آمده از پژوهش نشان می‌دهد که امکان استفاده از برخی روش‌های بسامدشماری کلاسیک و روش‌های مبتنی بر اندازه‌های آماری در زبان فارسی وجود دارد. در میان روش‌های مورد بررسی، روش بسامدشماری پیکره‌ی تخصصی بهبودیافته بهترین عملکرد را در هر دو گروه کلاسیک و روش‌های مبتنی بر اندازه‌های آماری با ٪۶۴ تشخیص درست، به خود اختصاص داده است. همچنین با توجه به جدول ۴، به_طور عمومی می‌توان نتیجه گرفت که روش‌های مبتنی بر اندازه‌های آماری از دقت بیشتری در مقایسه با سایر روش‌های کلاسیک از خود نشان می‌دهند و روش‌های اطلاعات متقابل نقطه‌ای با ٪۵۲ و مجذور کا با ٪۳۲ در جایگاه بعدی قرار دارند.

جدول ۴. واژه‌های تخصصی استخراج شده توسط هریک از روش‌های بسامدی

ردیف	روش	درصد واژگان تخصصی	درصد واژگان غیر تخصصی	کل
۱	اطلاعات متقابل نقطه‌ای	٪۵۲	٪۴۸	٪۱۰۰
۲	مجذور کا	٪۳۲	٪۶۸	٪۱۰۰
۳	بسامدشماری پیکره‌ی تخصصی بهبودیافته	٪۶۴	٪۳۶	٪۱۰۰
۴	بسامدشماری پیکره‌ی عمومی بهبودیافته	٪۶	٪۹۴	٪۱۰۰
۵	بسامدشماری پیکره‌ی تخصصی	۰	٪۱۰۰	٪۱۰۰
۶	بسامدشماری پیکره‌ی عمومی	٪۱۰	٪۹۰	٪۱۰۰

همچنین برای نمایش بهتر توانایی روش‌ها، نمودار ستونی مقایسه میان روش‌های بسامدشماری کلاسیک و روش‌های مبتنی بر اندازه‌های آماری ترسیم گردید. همان‌طور که شکل ۱ نیز نشان می‌دهد، روش بسامدشماری پیکره‌ی تخصصی بهبودیافته، روش اطلاعات متقابل نقطه‌ای و مجذور کا به ترتیب دارای بهترین عملکرد هستند.



شکل ۱. مقایسه‌ی دقت روش‌های بسامدشماری کلاسیک و روش‌های مبتنی بر اندازه‌های آماری

همچنین ۵۰ واژه‌ی استخراج شده‌ی پربسامد، بر اساس هریک از روش‌ها در جدول ۵ نمایش داده شده است تا امکان مقایسه میان واژگان استخراج شده در هر یک از روش‌ها فراهم گردد.

جدول ۵. واژه‌های تخصصی استخراج شده توسط هریک از روش‌های بسامدی

ردیف	روش‌های کلاسیک				روش‌های مبتنی بر اندازه‌های آماری	
	بسامدشماری پیکره‌ی تخصصی بهبودیافته	بسامدشماری پیکره‌ی عمومی بهبودیافته	بسامدشماری پیکره‌ی تخصصی	بسامدشماری پیکره‌ی عمومی	Chi-square	اطلاعات متقابل نقطه‌ای
۱	مژک	منع	و	و	پوش	هیدروژن
۲	هتروتروف	نبود	در	در	است	آنها
۳	پزشکی	دستاوردهای	به	به	هیدروژن	آزمایشگاه
۴	کربن	نیت	از	از	یعنی	مخاطب
۵	بیماری‌ها	قزوین	که	می	آنها	بدن
۶	تراکم	یزد	این	را	آزمایشگاه	افزوده
۷	چرخه	تلاشهای	را	است	مخاطب	نسبت

۸	محرک	آورد	است	که	بدن	رسوبی
۹	باکتری	دل	می	با	افزوده	مناطقی
۱۰	آندوپلاسمی	کوتاهی	با	های	نسبت	میدهند
۱۱	کپسول	اول	های	این	غذای	تیره
۱۲	باکتری‌ها	دستگاههای	برای	ها	رسوبی	ساختمان
۱۳	غیرجنسی	اتحادیه	آن	آن	مناطقی	پیوندهای
۱۴	لاکتوز	ساری	یک	شود	میدهند	صاف
۱۵	پرده‌ی	قم	خود	سلول	تیره	فراوان
۱۶	آبکشی	خدا	شده	شکل	آنفلوآنزا	تفسیر
۱۷	اندازه	قوه	کرد	این	ساختمان	جایگزین
۱۸	غشای	راهکارهای	شود	برای	آنرا	روزهای
۱۹	پانکراس	عمران	بر	یک	پیوندهای	پدیده
۲۰	اصطکاک	اهدا	تا	دارد	اندوخته	اند
۲۱	لولهٔ	نهم	سال	خود	منع	خرید
۲۲	جمله‌های	او	شد	که	شدگی	کربن
۲۳	آهک	سیل	کشور	شده	صاف	انتظار
۲۴	آهکی	هایش	ها	دو	مقداری	اعصاب
۲۵	باکتریایی	رود	نیز	بدن	جوانان	موجود
۲۶	بالک	باشیم	گفت	دارند	فراوان	فقر
۲۷	بیماری‌هایی	ری	بود	کند	تفسیر	پایان
۲۸	بکرزایی	امام	هم	ای	دیافراگم	هرگونه
۲۹	بندپایان	مخالفت	ای	کنند	جایگزین	درستی
۳۰	بیکربنات	برد	کند	وجود	روزهای	یابند
۳۱	توکسین	گشته	ایران	هستند	پدیده	طول
۳۲	دریچهٔ	تیتر	ما	هر	اند	بری
۳۳	رد۵	زندگی	اما	کنید	آماس	تجربه
۳۴	زمینه	قبلا	وی	شوند	خرید	جبران
۳۵	سیرنشده	افراد	دارد	آنها	کربن	ماست
۳۶	عفونت‌های	مخرب	یا	قرار	چسبی	هرگز
۳۷	قسمت‌های	گرگان	هر	اند	انتظار	اتوبوس
۳۸	کننده‌ی	روستاهای	باید	آب	اعصاب	درحال
۳۹	مایه‌کوبی	باری	قرار	خون	مایه	مسئولیت
۴۰	مردمک	گویند	دو	بر	موجود	گمان
۴۱	میانبرگ	نظرهای	آنها	یا	جانوری	طعمه
۴۲	موکوز	خارجی	کرده	مواد	فقر	پاسخ
۴۳	میوزین	سفرهای	او	انجام	پایان	با
۴۴	نرهایی	تی	مورد	نام	هرگونه	می‌یابد

۴۵	یابی	باورهای	خواهد	تا	درستی	تنظیم
۴۶	چارگف	مولف	کنند	هم	یابند	قطعی
۴۷	کپکهای	محورهای	دیگر	درون	طول	بستگی
۴۸	کیموس	اش	کار	استفاده	بری	ملاحظه
۴۹	گلیکوژن	دهیم	تهران	گونه	تجربه	فشار
۵۰	گردۀ	نوشتار	مردم	صورت	برسانند	ورود

۶. نتیجه‌گیری و پیشنهادهای آموزشی پژوهشی

نتایج به‌دست آمده از پژوهش نشان می‌دهد روش‌های بسامدشماری در پیکره‌های زبانی، از توانایی بالایی در استخراج واژه‌های پایه‌ی علوم پزشکی به یوه‌ی خودکار برخوردار هستند. همچنین در زبان فارسی امکان استفاده از هر دو گروه روش‌های کلاسیک و روش‌های مبتنی بر اندازه‌های آماری وجود دارد و با اعمال برخی تغییرات در روش‌های کلاسیک می‌توان شاهد بهبود قابل توجه آن‌ها بود و به نتایج قابل قبولی دست پیدا کرد. مزایای استفاده از روش‌های کلاسیک عبارتند از:

- سهولت محاسبه و تفسیر نتایج به‌دست آمده.
- صرفه‌جویی در زمان و هزینه به دلیل استفاده از یک پیکره در تحلیل‌ها.

و از معایب روش‌های کلاسیک می‌توان به موارد زیر اشاره نمود:

- دقت عمومی کمتر در مقایسه با روش‌های مبتنی بر اندازه‌های آماری.
- توجه تنها بر بسامد واژه.

روش‌های مبتنی بر اندازه‌های آماری به‌طور عمومی در مقایسه با تمامی روش‌های کلاسیک به جز روش بسامدشماری پیکره‌ی تخصصی بهبودیافته در استخراج خودکار واژه‌های پایه‌ی علوم پزشکی از عملکرد بهتری برخوردار هستند. از مزایای این روش‌ها می‌توان به موارد زیر اشاره نمود:

- دقت عمومی بیشتر در مقایسه با روش‌های کلاسیک.
- استفاده از اطلاعات بیشتر و در نظر گرفتن اطلاعات دست‌کم دو پیکره در تحلیل‌ها.

و از معایب روش‌های مبتنی بر اندازه‌های آماری می‌توان به موارد زیر اشاره نمود:

- پیچیدگی محاسبات و دشواری تفسیر نتایج.
- صرف زمان و هزینه بیشتر در مقایسه با روش‌های کلاسیک به دلیل استفاده از دست‌کم دو پیکره در تحلیل‌ها.

با توجه به نتایج به‌دست آمده از پژوهش و نیاز روز افزون به بهینه‌سازی مواد آموزشی زبان فارسی و تدوین مواد درسی معتبر و استاندارد، موارد زیر برای گسترش بحث پیشنهاد می‌گردد:

- روش‌های مورد بررسی و دیگر روش‌های بسامد شماری در سایر زمینه‌های تخصصی همچون علوم مهندسی و انسانی در زبان فارسی.
- بررسی امکان استفاده از روش‌های مورد بررسی در پژوهش برای تهیه‌ی کتاب‌های کمک آموزشی در سنین کودک و نوجوان
- بررسی امکان استفاده از روش‌های مورد بررسی در پژوهشروی پیکره‌ی اشعار کلاسیک و نو فارسی با هدف انتخاب واژگان جایگزین برای واژه‌های بیگانه.
- تهیه‌ی پیکره‌های تخصصی جامع‌تر با هدف گسترش زبان‌شناسی پیکره‌ای با حمایت نهادهای متولی همچون فرهنگستان زبان و ادب فارسی.

فهرست منابع:

جهانگردی، کیومرث، عاصی، مصطفی، افراشی، آزیتا و وکیلی فرد، امیررضا. (۱۳۹۵). واژه در کتاب آموزش زبان فارسی به غیرفارسی زبانان: پژوهشی پیکره‌بنیاد. *پژوهش‌نامه‌ی آموزش زبان فارسی به غیرفارسی‌زبانان*.سال پنجم، شماره۲، صص: ۲۶-۳.

رسولی، محمدصادق و مینایی بیدگلی، بهروز. (۱۳۸۷). روشی جدید در خطایابی املایی در زبان فارسی. *دومین کنفرانس داده‌یابی ایران*، (صص:۴-۲)، دانشگاه امیر کبیر.

سپهری، مهرداد. (۱۳۹۵). بسامدنگاری ودستاوردهای آن در آموزش. *مجله زبان و زبانشناسی*. سال دوم، شماره ۳، صص: ۶۰-۴۷.

نعمت زاده، شهین. (۱۳۹۵). واژگان پایه در خدمت تألیف کتاب ریاضی پایه‌ی اول. *فصلنامه مطالعات برنامه‌ی درسی ایران*. سال هفتم، شماره ۲۷، صص: ۸۴-۶۷.

Refrences:

Biber, D. (1993) .Representativeness in corpus design. *Literary and Linguistic Computing*. 8(4), 243-57.

Biber,D. (1990). *Some methodological Issues in Corpus Based Analyses of Linguistic Variation*, ms. University of Southern California , Los Angeles.

Bin, H & Zhang, **Y.** (2013). *Automatic Term Extraction in Large Text Corpora*, retrieved from:https://www.cs.dal.ca/~yongzhen/course/6509/report.pdf

Chujo, K., Utiyama, M., & Oghigian, K. (2006). Selecting level-specific Kyoto tourism vocabulary using statistical measures .*New aspects of English language teaching and learning* ,Taipei: Crane Publishing Company Ltd.,pp.126-138.

Chujo, K., Oghigian, K., Nishigaki, C., Utiyama, M., & Nakamura, T. (2007). Creating e-learning material with statistically-extracted spoken and written business vocabulary from the British National Corpus .*Journal of the College of Industrial Technology Nihon University*, 40, 1-12.

Chujo, K., Nishigaki, C., & Utiyama, M. (2005). Selecting 500 essential daily-life words for Japanese EFL elementary students from English picture dictionaries and a children's spoken corpus .In *Proceedings of inaugural international conference on the teaching and learning of English in Asia, Penang, Malaysia* Vol. 11, No. 15,1-12.

Chujo, K., Utiyama, M., & Nakamura, T. (2007). *Extracting Level-Specific Science and Technology Vocabulary From, the Corpus of Professional English CPE* ,Retrieved from : http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2007/47Paper.pdf

Chujo, K., Utiyama, M., & Nakamuru, T. O. (2010). Evaluating Statistically Extracted Domain –Specific Word Lists. *Corpus, ICT, and Language Education*, (eds) G. Weir and S. Ishikawa.University of Strathclyde Publishing, Glasgow, UK.

Chujo, K. (2004). Measuring Vocabulary Levels of English Textbooks and Tests Using a BNC Lemmatised High Frequency Word List. English Corpora under Japanese Eyes, Rodopi, pp. 231-249

Chujo, K., & Nishigaki, Ch. (2006). Creating Spoken Academic Vocabulary Lists from the British National Corpus. *Practical English Studies*, Vol.12, 19-34.

Coxhead, A., & Nation, P. (2001). The specialised vocabulary of English for academic purposes. The Specialized Vocabulary of English for Academic Purposes. In J. Flowerdew, & M. Peacock (Eds.), Research Perspectives on English for Academic Purposes, Chapter: The Specialised Vocabulary of English for Academic Purposes). Cambridge: Cambridge University Press. , 252-267.

Daille, B. (1994). Study and implementation of combined techniques for automatic extraction of terminology in *The balancing act: Combining symbolic and statistical approaches to language*, 29-36.

Enguehard, C. & Pantera, L. (1995). Automatic natural acquisition of a terminology. *Journal of quantitative linguistics*, *2*(1), 27-32.

Foo, J. (2012) .*Computational Terminology: Exploring Bilingual and Monolingual Term Extraction* ,Department of Computer and Information Science Linköping University, Linköping University Electronic Press.

Francis, W. N. (1992). Language corpora BC. In *Directions in Linguistics: Proceedings of Nobel Symposium* ,Vol. 82, 17-32.

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal on Digital Libraries*, *3*(2), 115-130.

Gries, S. T. (2010). Useful statistics for corpus linguistics. *A mosaic of corpus linguistics: Selected approaches*, 66, 269-291.

Granger, S. (2015). The contribution of learner corpora to reference and instructional materials design.UK:Cambridge Uiversity Press.

Hulth, **A.** (2004). *Combining Machine Learning and Natural Language Processing for Automatic Keyword Extraction* .Doctoral Dissertation, Stockholm University

Jahangardi ,K., & Asi ,M., & Afrashi, A., & Vakilifard ,A. R. (2016).Vocabulary in the Textbooks of Teaching Persian to Non-Persian speakers: A Corpus-Based Study. *Journal of Teaching Persian to speakers of other Languages*.5(12),3-26 [In Persian]

Leech, G. (1991). The state of the art in corpus linguistics. In Aijmer, K. and Altenberg, B. (Eds.) *English corpus linguistics: Studies in Honour of Jan Svartvik*. London: Longman

Leech, G. (1992). Corpora and theories of linguistic performance. *Directions in corpus linguistics,* 1992: 105-122.

McEnery, T. (2001). *Corpus linguistics: An introduction*. Edinburgh University Press.

Nematzadeh, Sh. (2013). Core Vocabulary Serving the First Grade Primary School Mathematics Textbook. *Journal of Curriculum Studies* (J.C.S.) Vol.7 (27), 84-67 [In Persian]

Nakagawa, H. & Mori, T. (2002). A simple but powerful automatic term extraction method. In *COLING-02 on COMPUTERM 2002: second international workshop on computational terminology-Volume 14* , 1-7.

Pantel, Pand. & Lin, D.(2001). A Statistical Corpus-Based Term Extractor. *Advances in Artificial Language,* Volume 2056, 36-46.

Patry, A. & Langlais, P. (2005). Corpus-based terminology extraction .In *Terminology and Content Development–Proceedings of 7th International Conference on Terminology and Knowledge Engineering*, Litera, Copenhagen.

Rasooli ,M. S. & Minaei-Bidgoli, B. (2008). A new approach for Persian spellchecking. *In IDMC2008* 11-12 Nov, Amir Kabir University,Tehran,Iran [In Persian]

Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge University Press.

Rezai, M. J. & Mosavi Miangah, T. (2017). FarsiTag: a part-of-speech tagging system for Persian. *Digital Scholarship in the Humanities*, 32(3), 632–642.

Sepehri ,M. (2006). Concordancing and its Pedagogical Implications. *Language & Linguistics, Journal of the Linguistic Society of Iran*,Vol 2, 47-60 [In Persian]

Sinclair, J. (1991). *Corpus, concordance, collocation* (Vol. 1). Oxford: Oxford University Press.

Sinclair, J. (1998). Corpus evidence in language description, In Gerry Knowles, Tony Mcenery, Stephen Fligelstone, Anne Wichman, (Eds.) *Teaching and language corpora*. Longman pp. 27-39

Swales, J. M. (2002). Integrated and fragmented worlds: EAP materials and corpus linguistics. *Academic discourse*, 150-164.

Vulanovic, R. & Mosavi Miangah, T. (2018). A Comparison of the Accuracy of Parts-of-Speech Tagging Systems Based on a Mathematical Model, *Journal of Quantitative Linguistics*, DOI: 10.1080/09296174.2018.1474517.

Wermter, V. J. (2008). Collocation and Term Extraction Using Linguistically Enhanced Statistical Methods.https://db_thueringen.de/servlets/MCRfileNodeServlet/dbt_derivate_00017176/warmter/dissertation.pdf

Zhang, C. & Wu, D. (2012).Bilingual terminology extraction using multi-level termhood .*The Electronic Library*, 30(2), 295-309.