A Study on the Improved Techniques of Corpus-based Frequency Approaches in Automatic Term Extraction (ATE)
(The Case Study: Basic Medicine Vocabulary)

Zohreh Zolfaghar
PhD Candiatate in General Linguistics; Payame Noor University; Tehran, Iran Email: zohreh.zolfaghar@gmail.com

Tayebeh Mosavi Miangah*
PhD in Computational Linguistics; Associate Professor; Payame Noor University; Tehran, Iran Email: mosavit@pnu.ac.ir

Belghis Rovshan
PhD in General Linguistics; Associate Professor; Payame Noor University; Tehran, Iran Email: Bl_rovshan@pnu.ac.ir

Amir Reza Vakilifard
PhD in Second Languages Didactics; Associate Professor; Imam Khomeini University of Qazvin; Qazvin, Iran; Email: vakilifard@hum.ikiu.ac.ir

Iranian Journal of Information Processing and Management

Received: 06, Sep. 2018 **Accepted: 24, Feb. 2020**

Abstract: Nowadays we are witnessing the dramatic growth of utilizing corpus-based studies in linguistics known as corpus linguistics. The current research aims to study the improvement of frequency techniques in Farsi Language and has been conducted in order to achieve a scientific approach in automatic term extraction focused on extracting basic medicine terms. Using statistical approaches along with corpus linguistic tools (hybrid extraction methods) for automatic term extraction purposes have become quite common in a number of languages such as English, French, Japanese and Korean. So far, these approaches have not been utilized in Farsi language widely and most of the efforts for term extraction have been conducted in traditional ways. On the other hand, these approaches are language specific and it is not possible to use them for a different language. They should be modified based on the properties of the target language in order to achieve an extraction method which is appropriate for that language. To do so, a group of

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 35 | No. 4 | pp. 1039-1064
Summer 2020
https://doi.org/10.35050/JIPM010.2020.028

* Corresponding Author

frequency models with approaches to count frequency in a main corpus and a special corpus and their improved methods have been utilized. The frequency method used in this study has counted the terms in a general and a main corpus which is created by the researcher. These corpuses are formed from the texts in science textbooks of Iran highschools (grades 9-12), science textbooks of Iran middle schools (grade 7-8), the science texts taught in Qazvin Imam Khomeini Farsi Language Center and some journals and articles on general science. Achieved results show that there is a potential possibility to extract terms automatically in Farsi. Among the major challenges of utilizing the simple methods we can refer to the process of separating high frequency words such as coordinators or prepositions. Therefore, to increase the power of this model, we improved the basic models by applying some techniques on them. It is observed that the improved frequency method has shown a better performance in the special corpus as opposed to other methods and has been able to predict up to 60% of the special vocabulary in the first 50 high frequency extracted vocabulary. On the other hand, other results of the study show that the presence of low frequency vocabulary in the general corpus with a frequency similar to the frequency of special vocabulary, has led to achieving weaker results than the simple method.

Keywords: Automatic Term Extraction, Medicine Vocabulary, Corpus, Hybrid Extraction Methods, Farsi Language Teaching, Information Retrievals

بررسی تکنیک‌های بهبود عملکرد روش‌های بسامدشماری پیکره‌بنیاد در استخراج خودکار واژگان (مورد مطالعه: واژگان پایهٔ علوم پزشکی)

زهره ذوالفقار
دانشجوی دکتری زبان‌شناسی همگانی؛
گروه زبان‌شناسی؛ دانشگاه پیام نور؛ تهران، ایران؛
zohreh.zolfaghar@gmail.com

طیبه موسوی میانگاه
دکتری زبان‌شناسی رایانشی؛ دانشیار؛
گروه زبان‌شناسی؛ دانشگاه پیام نور؛ تهران، ایران؛
پدیدآور رابط mosavit@pnu.ac.ir

بلقیس روشن
دکتری زبان‌شناسی همگانی؛ دانشیار؛
گروه زبان‌شناسی؛ دانشگاه پیام نور؛ تهران، ایران؛
Bl_rovshan@pnu.ac.ir

امیررضا وکیلی‌فرد
دکتری آموزشکاوی زبان‌های دوم؛ دانشیار؛
گروه آموزش زبان فارسی به غیرفارسی‌زبانان؛
دانشگاه بین‌المللی امام خمینی قزوین؛ قزوین، ایران؛
vakilifard@hum.ikiu.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۷/۰۶/۱۵ **پذیرش:** ۱۳۹۸/۱۲/۰۵ مقاله برای اصلاح به مدت ۴ ماه و نیم نزد پدیدآوران بوده است.

چکیده: امروزه، شاهد گسترش استفاده از روش‌های پیکره‌بنیاد در زبان‌شناسی هستیم. پژوهش حاضر به بررسی تکنیک‌های بهبود عملکرد روش‌های بسامدشماری در زبان فارسی و با منظور دستیابی به رویهٔ علمی جهت استخراج خودکار واژگان پایهٔ علوم پزشکی انجام پذیرفته است. استفاده از روش‌های آماری در کنار ابزار زبان‌شناسی پیکره‌ای (روش‌های استخراج خودکار ترکیبی) جهت استخراج خودکار واژگان در تعدادی از زبان‌های دنیا همچون انگلیسی، فرانسه، ژاپنی، و کره‌ای طی چند دههٔ اخیر بسیار رایج بوده است، حال آن‌که این روش‌ها در زبان فارسی تاکنون به‌صورت جدی مورد استفاده قرار نگرفته و اغلب استخراج‌ها در زبان فارسی به روش سنتی انجام گرفته است؛ ضمن آن‌که به‌کارگیری این روش‌ها در هر زبانی متفاوت است و برون‌داد روش‌های آماری در هر زمان با توجه به ویژگی‌های

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۵ | شمارهٔ ٤ | ص‌ص ۱۰۳۹-۱۰٦٤
تابستان ۱۳۹۹
https://doi.org/10.35050/JIPM010.2020.028

زبان‌شناختی آن زبان متفاوت است. از این رو، باید مطابق با ویژگی‌های هر زبان در این روش‌ها تغییراتی اعمال کرد تا در نهایت، بتوان به روشی جهت استخراج خودکار واژگان دست یافت. جهت نیل به این هدف در زبان فارسی، از خانوادهٔ مدل‌های بسامدشماری با رویکردهای بسامدشماری پیکرهٔ عمومی، بسامدشماری پیکرهٔ اختصاصی و روش‌های بهبودیافتهٔ آن‌ها استفاده شده است. بسامدشماری به‌کاررفته در پژوهش، برپایهٔ پردازش اطلاعات واژگان در دو پیکرهٔ اصلی و اختصاصی، که محقق آن را ایجاد کرده است و از این پس آن را پیکرهٔ محقق‌ساخته می‌نامیم، صورت گرفته است. پیکرهٔ محقق‌ساخته شامل متون درس زیست‌شناسی دورهٔ اول تا چهارم دبیرستان، متون درس علوم دوم و سوم راهنمایی، متون تدریس‌شده در «مرکز آموزش زبان فارسی امام خمینی قزوین»، مجلات و مقالات حوزهٔ پزشکی عمومی و پیکرهٔ عمومی مورد استفاده، پیکرهٔ روزنامهٔ همشهری (نسخهٔ دوم) است. نتایج به‌دست‌آمده نشان می‌دهد که قابلیت استفاده از روش‌های بسامدشماری پیکره‌بنیاد در زبان فارسی برای دست یافتن به شیوه‌ای واحد در استخراج خودکار واژگان وجود دارد. شیوهٔ به‌کارگیری روش‌های آماری کلاسیک و مدرن و روش‌های بهبودیافتهٔ آن‌ها به‌یقین می‌تواند گامی مؤثر در تهیه و تدوین متون آموزشی زبان فارسی و گسترش آموزش این زبان به شمار آید. از عمده‌ترین مشکلات استفاده از روش‌های ساده، می‌توان جداسازی واژگان پرتکرار، همچون حروف ربط را نام برد. بنابراین، جهت بالابردن توان این مدل با اعمال روش‌هایی می‌توان روش‌های اولیه را بهبود بخشید. مشاهده می‌شود که روش بسامدشماری بهبودیافته در پیکرهٔ اختصاصی از سایر روش‌ها عملکرد بهتری داشته و تا ۶۰ درصد واژگان تخصصی را در ۵۰ واژهٔ پربسامد شناسایی می‌کند. از سوی دیگر، مشاهده می‌شود که با افزایش دامنهٔ واژگان مورد بررسی در پژوهش از ۵۰ به ۱۰۰، ۱۵۰ و ۲۰۰، دقت مدل‌ها افزایش یافته و درصد واژگان تخصصی انتخاب‌شده به ثبات می‌رسد.

کلیدواژه‌ها: استخراج خودکار، واژگان علوم پزشکی، پیکره، روش‌های ترکیبی استخراج، آموزش زبان فارسی، بازیابی اطلاعات

۱. مقدمه

طی چند دههٔ اخیر استخراج خودکار واژگان[1] از پیکره‌ها[2] مورد توجه بسیاری از پژوهشگران بوده است. در اوایل دههٔ ۹۰ میلادی، پیکره‌های متنی بزرگ رایانه‌ای ساخته شدند که منجر به ایجاد نخستین برنامه‌های استخراج واژگان[3] گردید. زبان‌شناسان رایانه‌ای و کاربردی، مترجمان، مفسران، مهندسان علوم رایانه و نرم‌افزار و دست‌اندرکاران آموزش

1. automatic term extraction

۲. پیکره به مجموعه‌ای از متون گفته می‌شود؛ مجموعه‌ای خام از داده‌های زبانیِ نوشتاری یا گفتاری که می‌توان در توصیف و تحلیل زبان از آن بهره گرفت.

3. term extractors

زبان به استخراج واژگان از پیکره‌ها و ساخت پیکره‌ها علاقهٔ بسیار نشان داده و استقبال از رویکرد زبان‌شناسی پیکره‌ای روزبه‌روز در حال افزایش است، به‌طوری که بسیاری از کشورهای دنیا اقدام به تهیه و تدوین پیکرهٔ ملی زبان‌های بومی خود کرده‌اند که از میان شناخته‌شده‌ترین آن‌ها می‌توان به پیکرهٔ ملی زبان بریتانیا، پیکرهٔ ملی زبان روسی، پیکرهٔ ملی لهستانی، و پیکرهٔ ملی انگلیسی استرالیا اشاره نمود. علاوه بر پیکره‌های عظیم ملی، با توجه به کاربرد گستردهٔ تحلیل‌های پیکره‌بنیاد، پیکره‌های خردتر با اهداف پژوهشی خاص مانند پیکرهٔ تاریخی زبان انگلیسی، پیکره‌های متون ادبی، و پیکره‌های متون ترجمه‌شده و نظایر آن نیز ایجاد شده‌اند. نوع دیگری از پیکره‌ها که از اهمیت بالایی در آموزش زبان خارجی برخوردار هستند، پیکره‌های متنی زبان‌آموزان با زبان‌های مادری و سطوح توانش زبانی مختلف است. اهداف مختلف محققان زبان‌شناسی پیکره‌بنیاد ساختن واژه‌نامه‌ها، سیستم‌های متنی و تحلیل متن، ترجمهٔ ماشینی، ایجاد پایگاه داده‌ها، نمایه‌سازی و جز آن بوده است. در این میان یکی از کاربردهای نوین زبان‌شناسی پیکره‌ای و رایانه‌ای فراهم‌آوردن بستری مناسب جهت انجام پژوهش‌های آموزشی است. امروزه، بهترین ابزار برای تهیهٔ متون آموزشی، استفاده از داده‌های زبانی واقعی و موثق است که می‌توان این داده‌ها را به کمک پیکره‌ها در اختیار گرفت. پژوهشگران نشان داده‌اند که این شاخهٔ نوین در زبان‌شناسی می‌تواند به‌خوبی و به شکلی کارآمد جهت امور آموزشی مورد استفاده قرار گیرد. می‌دانیم که یکی از حوزه‌های مهم در آموزش زبان، واژگان است. ضرورت و اهمیت پژوهش حاضر نیز دستیابی به روشی خودکار و کارآمد برای استخراج واژگان سطح‌بندی‌شده جهت تهیه و تدوین متون آموزشی است. از آن رو که در زبان فارسی با وجود آثار منتشرشده برای آموزش زبان فارسی، هنوز هم عدم حضور مواد آموزشی استاندارد احساس می‌گردد، بنابراین، با مد نظر قرار دادن اهداف تعیین‌شده توسط «شورای گسترش زبان فارسی»، از جمله «تهیه و تدوین متون آموزشی و کمک‌آموزشی برای خارجیان و نیز ایجاد بانک اطلاعاتی و منابع آموزشی»[1]، لازم است ابعاد مربوط به تدوین مطالب، به‌ویژه در متون آموزشی حوزه‌های تخصصی که تاکنون کمتر مورد توجه بوده‌اند، مورد توجه و دقت بیشتری قرار گیرند.

مسئلهٔ اصلی در حوزهٔ آموزش زبان فارسی، داشتن متون آموزشی مناسب و استاندارد

1. http://www.tasnimnoor.com/Default.aspx?tabid=298&EntryID=779

است. بدین‌معنا که متون به‌کاررفته در آموزش زبان فارسی باید با محتوای مناسب و بر اساس داده‌های واقعی و پیکره‌ایِ زبان فارسی به‌صورت منظم و ساخت‌یافته تهیه گردند.

می‌دانیم که یکی از معتبرترین مراکز فعال در زمینهٔ آموزش زبان فارسی به غیرفارسی‌زبانان در ایران، «مرکز زبان فارسی دانشگاه بین‌المللی امام خمینی قزوین» است. هر ساله زبان‌آموزان بسیاری از کشورهای مختلف آسیایی، آفریقایی و اروپایی برای یادگیری زبان فارسی جهت اخذ پذیرش در دانشگاه‌های ایران و ادامهٔ تحصیل در ایران وارد این مرکز می‌شوند. این زبان‌آموزان اغلب داوطلب رشته‌های فنی و پزشکی و یا علوم انسانی هستند و یکی از کمبودهای مسلم این زبان‌آموزان عدم دسترسی آن‌ها و نیز اساتید این مرکز به واژگان تخصصی حوزهٔ مورد مطالعهٔ آن‌ها در سطوح مختلف است. اهمیت نقش واژگان انتخاب‌شده جهت تدریس بر کمتر کسی پوشیده است، اما تاکنون پژوهش جدی و عمیقی در زمینهٔ تهیهٔ این واژگان در سطوح مختلف در زبان فارسی و یا درگرایش‌های مختلف رشتهٔ پزشکی و یا حوزه‌ای خاص انجام نگرفته است. این در حالی است که یکی از جنبه‌های مهم آموزش زبان، داشتن متون متناسب با سطح زبان‌آموز است. مدرسان زبان فارسی به زبان‌آموزان متقاضی حضور در رشته‌های پزشکی در ایران، نیازمند آگاهی از واژه‌های پرکاربرد و کلیدی هستند تا بر مبنای آن از توانایی سطح‌بندی زبان‌آموزان در رده‌های مبتدی، متوسط و پیشرفته برخوردار گردند و امکان آشنایی جامع‌تر زبان‌آموزان را با واژه‌های پرکاربرد در سطوح بالاتر فراهم آورند. از این رو، انتخاب و به‌کار گرفتن کارآمدترین روش‌ها جهت تهیه و تدوین این منابع امری ضروری است.

پژوهش‌های گوناگونی در زبان‌های مختلف دنیا برای دستیابی به فهرست واژگان تخصصی و یا اصطلاحات صورت گرفته است که در غالب آن‌ها از روش‌های آماری، زبان‌شناختی و یا ترکیبی از هر دو بهره برده‌اند. در این میان، یکی از پرکاربردترین روش‌های مورد استفاده در استخراج خودکار واژگان، خانوادهٔ روش‌های بسامدشماری[1] است. منظور از روش‌های بسامدشماری، شمارش وقوع واژه‌ها در متن است. این دسته از روش‌ها تاکنون بیشتر در زبان‌های انگلیسی، فرانسه و ژاپنی مورد استفاده قرار گرفته و مسیر تکامل خود را پیموده‌اند. در زبان فارسی نیز شاهد به‌کارگیری این خانواده از روش‌ها طی چند سال اخیر بوده‌ایم. با به‌کارگیری روش‌های بسامدشماری، رفته‌رفته

1. frequency based approaches

ضعف‌های هر یک از این روش‌ها در مسائل واقعی و کاربردی آشکار می‌گردد و نیاز است تا اصلاحاتی در روش‌های اولیه اعمال گردد.

۲. پیشینهٔ پژوهش

زبان‌شناسی پیکره‌ای را می‌توان شاخه‌ای نوین در زبان‌شناسی کاربردی به حساب آورد که در آن ابعاد متنوع تولیدات زبانی مورد بررسی قرار می‌گیرد. زبان‌شناسی پیکره‌ای در واقع، استفاده از متن زبانی دیجیتالی‌شده برای تحلیل‌های زبانشناختی است. از نظر «مک‌انری و ویلسون» پیکره را می‌توان مجموعه‌های نسبتاً بزرگی از متون الکترونیکی دانست که حاشیه‌نویسی، برچسب‌گذاری و نیز دسته‌بندی سنجیده‌ای دارند و از همین رو، امکان بررسی‌های زبانشناختی را برای کاربر فراهم می‌آورند، اما جهت بالابردن دقت پردازش پیکره‌ها و نتایج حاصل از آنها لازم است تا پیکرهٔ مورد استفاده، ابتدا اصلاح و ویرایش شود. برای مثال، جهت خوانش و شمارش پارامترهای مورد نیاز مانند جمله‌ها، عبارات و واژه‌ها توسط کامپیوتر لازم است تا پیکره‌ها فاقد هر نوع عیب و نقصی باشند (McEnery and Wilson 2001). یکی از استفاده‌هایی که از پیکره‌ها به عمل آمده، به‌کارگیری آن‌ها در استخراج واژه یا اصطلاح است. پژوهشگران مختلفی اقدام به استخراج خودکار واژگان، به‌ویژه پس از ظهور زبان‌شناسی رایانه‌ای و پیکره‌ای کرده‌اند و در زبان‌های گوناگونی همچون فرانسه، انگلیسی و به‌ویژه ژاپنی مقالات و پایان‌نامه‌های متعددی در این حوزه تحریر شده است. در دسته‌ای از این آثار، پژوهشگران در صدد ایجاد سیستمی برای استخراج خودکار واژگان برآمده‌اند و این کار را به کمک الگوریتم‌هایی معین و با به‌کارگیری نظام‌های ترکیبی انجام داده‌اند. برای مثال، Enguehard & Pantera (1995) با استفاده از شاخص اطلاعات متقابل در زبان انگلیسی و Nakagawa & Mori (2002) با کمک سیستم رتبه‌دهی موفق به استخراج واژگان از پیکره‌های زبان ژاپنی شده‌اند. «دایلی» از ترکیب دانسته‌های زبانی و روش‌های آماری به استخراج الگوهای نحوی واژگان در زبان انگلیسی پرداخته است (Daille 1994). مبنای اصلی محاسبات در مدل‌های مذکور برای استخراج واژگان اغلب کاندید واژه‌های گروه اسمی است و از یک روش آماری برای تعیین بسامد آن استفاده می‌گردد.

«گرینجر گیلکوویین» بیان می‌دارند که چنین پیکره‌هایی این امکان را به پژوهشگران می‌دهد که به بررسی دقیق نوع و میزان استفاده از واژگان و ساختارهای دستوری توسط

زبان‌آموزان بپردازند و با بهره گرفتن از چنین پیکره‌هایی بتوانند به مقایسهٔ جنبه‌های مختلف زبانی از قبیل انواع اشتباهات، بسامد استفاده از انواع واژگان، عبارات، جمله‌واره‌ها و ساختارهای دستوری در بین گروه‌های مختلف زبان‌آموزان بپردازند (Gilquin & Granger 2015).

«لوسیو ونچورا» و همکاران ضمن برشمردن دشواری‌هایی همچون اختلال، سکوت، بسامد پایین، پیکره‌های بزرگ، و پیچیدگی‌های فرایند استخراج اصطلاحات چندواژه‌ای که روش‌های زبانشناختی و آماری در استخراج واژه دارند، روش دیگری معرفی می‌کنند. روش آن‌ها اندازه‌گیری‌های متعددی را بر مبنای جنبه‌های زبان‌شناختی، آماری، نموداری[1] و تارنما[2] معرفی می‌کند. آن‌ها به‌کمک این روش‌های نوین به استخراج اصطلاحات پزشکی پرداخته و داده‌های استخراج‌شدهٔ خود را با نتایج روش‌های پیشین مقایسه کرده‌اند. آن‌ها روش خود را بسیار کار آمد می‌دانند (Lossio-Ventura et al. 2016).

چند سالی است که پژوهشگران در زبان فارسی نیز با استفاده از روش‌ها و ابزارهای پیکره‌شناسی به تجزیه و تحلیل زبان فارسی پرداخته‌اند. «سپهری» فراهم‌کردن مجموعه‌ای از متون مستخرج از زبان گفتاری و نوشتاری برای استفادهٔ تهیه‌کنندگان فرهنگ‌های لغت را از دستاوردهای بسامدنگاری می‌داند. وی معتقد است که این مجموعهٔ متون را می‌توان در زمرهٔ مطمئن‌ترین منابعی دانست که فرهنگ‌نویسان از آن در جمع‌آوری لغات برای تهیهٔ فرهنگ لغت استفاده می‌کنند (۱۳۹۵). «گزنی» در پژوهش خود در زمینهٔ پردازش خودکار متن و بازیابی اطلاعات به بررسی دشواری‌های نگارش زبان فارسی و تأثیر آن‌ها در محیط‌های دیجیتال و در بازیابی اطلاعات پرداخته است. به باور وی لازم است در طراحی الگوریتم‌های سامانه‌های جست‌وجو و بازیابی فارسی، دشواری‌های نگارش فارسی مورد توجه قرار گیرد (۱۳۸۵). موضوع مطالعه‌شده در پژوهش «گزنی» مطلبی است که در پژوهش حاضر نیز مشاهده شد. چون نرم‌افزارهای آماری و روش‌های آماری روی متون فارسی قابل اجرا نبود و پس از برنامه‌نویسی و اعمال برنامه‌های گوناگون، در نهایت، امکان اجرای روش‌های بسامدشماری روی پیکره فراهم شد. در نتیجه، ضروری است در طراحی الگوریتم‌های سامانه‌های جست‌وجو و بازیابی فارسی، به‌هنجارسازی تنوعات نگارشی و دستوری مد نظر قرار گیرد. تدوین استاندارد نگارش فارسی، استفاده از سیاهه‌های از پیش

1. graph
2. web

تعیین‌شده، تجهیز پایگاه اطلاعاتی به اصطلاحنامه و فرهنگ‌های املایی، و تدوین دستنامه یا راهنمای جست‌وجو، از جمله راهکارهای ارائه شده است. این راهکارها با وجود جامع نبودن، کم‌وبیش اثربخش به نظر می‌رسند.

«بیرنگ و بشیری» در پژوهش خود به بررسی بسامد شعرهای فارسی در نفثةالمصدور پرداختند. آن‌ها نشان دادند که از مجموع ۸۵ شاهد شعری که در کتاب آورده شده، بیشترین بسامد به لحاظ قالب شعری به رباعی و پس از آن به قصیده اختصاص دارد. در اوزان شعری هم پس از وزن رباعی، بحر مضارع بیشترین کاربرد را دارد (۱۳۹۶).« جهانگردی» و همکاران با استفاده از روش‌ها و ابزارهای زبان‌شناسی پیکره‌ای به سنجش میزان همپوشانی و انطباق واژه‌های ارائه‌شده در کتاب‌های آموزش زبان فارسی به غیرفارسی‌زبانان، با پُربسامدترین واژه‌های زبان فارسی پرداختند. آن‌ها بر اساس متن‌های موجود در پایگاه داده‌های زبان فارسی، یک پیکرهٔ زبانی متوازن طراحی کرده و پُربسامدترین واژه‌های آن را به‌عنوان مبنای کار قرار دادند. نتایج و یافته‌های پژوهش آن‌ها نشان می‌دهد که به لحاظ سطوح زبان‌آموزی، میزان همپوشانی واژگانی هر یک از گروه‌های مورد بررسی با گروه‌های متناظر آن‌ها در پیکرهٔ مبنا بسیار پایین است (۱۳۹۵).

«راد» و همکاران به ارائهٔ روشی نو برای شاخص‌گذاری خودکار و استخراج کلمات کلیدی پرداختند. آن‌ها با اشاره به دشواری‌های زبان فارسی به‌خصوص ویژگی‌های نگارشی و دستوری بیان می‌کنند که استخراج خودکار در زبان فارسی دشوار است. مبنای کار آن‌ها خوشه‌بندی بود؛ بدین‌معنا که اطلاعات خوشه‌بندی می‌شدند و با جدا کردن نمونه‌ها از یکدیگر و قرار دادن آن‌ها در گروه‌های شبیه‌به‌هم، واژگان کلیدی استخراج می‌شدند. نتایج آزمایش آن‌ها روی چندین متن نشان‌دهندهٔ دقت روش پیشنهادی آن‌هاست (۱۳۹۵). از آنجا که در زبان فارسی پژوهش جامع در حوزهٔ استخراج واژگان تخصصی از پیکره‌ها انجام نشده است، بنابراین، پژوهش حاضر را می‌توان پژوهشی نو در این زمینه دانست.

در اینجا توجه به چند نکته ضروری به نظر می‌رسد. نخست آن‌که اغلب روش‌های یادشده در زبان‌های فرانسه و انگلیسی مورد ارزیابی قرار گرفته‌اند و برای زبان‌های دیگر همچون فارسی تاکنون به‌صورت جدی مورد ارزیابی قرار نگرفته‌اند. دیگر آن‌که روش‌هایی که در بالا به آن‌ها پرداخته شد، درصدهای موفقیت گوناگونی داشته‌اند و در مواردی نیز نتایج، دلالت بر عدم توانایی این روش‌ها در استخراج واژگان داشته است

کـه از عمـدهٔ دشـواری‌های ایـن فرایندهـا می‌تـوان بـه تشـخیص واژه‌هـای مرکـب، اصطلاحـات مرکـب، ماهیـت اصطلاحـی واژه و متناسـب‌بودن آن‌هـا اشـاره نمـود. بنابرایـن، از نظـر کارایـی، ایـن سیسـتم‌ها بسـته بـه روش مـورد اسـتفاده و حجـم پیکـره متفـاوت هسـتند و اغلـب بـرای پیکره‌هـای کوچـک طراحـی شـده‌اند. در پژوهـش حاضـر بـا در نظـر گرفتـن مـوارد مذکـور، امـکان اسـتفاده از روش‌هـای پیکـره‌ای بسـامدی در زبـان فارسـی و روش‌هـای بهبـود عملکـرد آن‌هـا مـورد ارزیابـی قـرار گرفتـه اسـت.

۳. مدل و روش آزمون

تمرکـز اصلـی پژوهـش بـر روی تحلیـل داده‌هـای پیکـرهٔ علـوم پزشـکی اسـت. از آنجـا کـه تـا پیـش از ایـن، چنیـن پیکـره‌ای به‌صـورت تخصصـی در زبـان فارسـی وجـود نداشـته، جهـت ایجـاد آن تـلاش شـده تـا به‌صورتـی جامـع منابـع پرکاربـرد آموزشـی در تهیـهٔ آن به‌ترتیـب زیـر مـورد توجـه قـرار گیـرد:

- ◇ متون درس زیست‌شناسی دورهٔ اول تا چهارم دبیرستان؛
- ◇ متون درس علوم دوم و سوم راهنمایی؛
- ◇ متون تدریس‌شده در مرکز زبان قزوین؛
- ◇ مجلات و مقالات حوزهٔ پزشکی عمومی.

پـس از گـردآوری متـون بـه میـزان کافـی ویرایـش آغـاز شـد. آن دسـته از متـون استخراج‌شـده از وب‌سـایت مربـوط بـه کتـب درسـی (www.roshd.ir) کـه بـا فرمـت «پی‌دی‌اف» بودنـد، در مرحلـهٔ نخسـت بـه فرمـت قابـل ویرایـش یعنـی «وُرد» تبدیـل شـدند. سـپس، بـا اسـتفاده از نرم‌افـزار سفارشـی کـه بـرای همین‌منظـور طراحـی شـده بـود، متـون مـورد نظـر پیش‌ویرایـش شـدند؛ بدیـن مفهـوم کـه حـروف و کلمـات انگلیسـی، کلمـات تک‌حرفـی، اعـداد، اشـکال، جـداول، نمودارهـا، نمادهـا، علامـات نگارشـی و ماننـد آن حـذف شـدند. عمـل تقطیـع[1] نیـز بـا اسـتفاده از همیـن نرم‌افـزار صـورت گرفـت و تمـام کلمـات متـن جداسـازی شـد. سـپس، بـار دیگـر متـون به‌صـورت دسـتی نیـز ویرایـش شـدند تـا احتمـال خطـا یعنـی اسـتخراج غیرواژه‌هـا بـه حداقـل برسـد. پـس از اصلاحـات صورت‌گرفتـه، متـون مـورد بررسـی به‌صـورت یـک متـن کامـل و به‌صـورت یـک پیکـرهٔ زبانـی تنظیـم گردیـد.

1. tokenization

۳-۱. روش‌های بسامدشماری

روش‌های بسامدشماری از اوایل دههٔ ۹۰ میلادی همواره مورد توجه پژوهشگران و زبان‌شناسان پیکره‌ای بوده است. روش‌های بسامدشماری به‌کاررفته در این پژوهش را می‌توان مطابق جدول ۱، به ۴ دستهٔ اصلی تقسیم‌بندی کرد. در تحلیل‌های مبتنی بر بسامدشماری، تأکید اصلی بر فراوانی واژگان به‌کاررفته در یک پیکره است. همان‌گونه که جدول نیز نشان می‌دهد، این روش‌ها شباهت زیادی به یکدیگر داشته و هر یک جهت رفع مشکلات پدیدآمده در کار با داده‌های واقعی سیر تکاملی خود را پیموده‌اند.

۳-۲. بسامدشماری پیکرهٔ اصلی عمومی

شمارش فراوانی واژگان را می‌توان از نخستین روش‌های به‌کاررفته در بسامدشماری دانست. در این روش به شمارش واژگان در یک پیکرهٔ عمومی پرداخته می‌شود و واژگان به‌ترتیب فراوانی از بیشترین به کمترین مرتب‌سازی می‌شوند. از آنجا که پیکرهٔ مورد بررسی یک پیکرهٔ عمومی است، بنابراین، می‌توان انتظار داشت واژگان نقشی[1] که بیشترین تکرار را دارند، مانند حروف ربط، حروف اضافه و نظایر آن از فراوانی بیشتری برخوردار باشند و واژگان محتوایی[2] که واژگان تخصصی نیز بخشی از آن‌ها هستند، با فراوانی کمتر در انتهای جدول قرار گیرند. منظور از واژگان محتوایی همان انواع اسم، صفت، قید و افعال هستند.

۳-۳. بسامدشماری پیکرهٔ اختصاصی

رویکرد اصلی در این روش بسامدشماری در یک پیکرهٔ تخصصی است. علت انتخاب پیکرهٔ تخصصی در این است که نتایج تجربی نشان می‌دهد که در یک پیکرهٔ عمومی ممکن است واژگانی غیرتخصصی و کم‌کاربرد نیز با بسامدی مشابه بسامدِ واژگان تخصصی مشاهده گردد. بنابراین، متخصصان جهت رفع این مشکل بررسی پیکره‌های اختصاصی را پیشنهاد داده‌اند.

۳-٤. بسامدشماری پیکرهٔ اصلی عمومی بهبودیافته

یکی از عمده‌ترین چالش‌ها که در بررسی‌های تجربی خیلی زود خود را نمایان

1. function words 2. content words

ساخت، جداسازی واژگان تخصصی و غیرتخصصی پس از مرتب‌سازی به کمک بسامدشماری بود. بنابراین، استفاده از رویّه‌ای که به کمک آن بتوان این فرایند را تسریع نمود، بسیار ارزشمند خواهد بود. از این رو، روشی ابداع گردید که به کمک آن تفاوت فراوانی کل واژگان در پیکرهٔ عمومی از مقدار فراوانی واژگان در پیکرهٔ عمومی کسر می‌گردد و نتایج به‌دست‌آمده به‌ترتیب صعودی مرتب‌سازی می‌شود. علت این امر آن است که با کسر کردن فراونی کل واژگان از فراوانی هر یک از واژگان پس از مرتب‌سازی به‌ترتیب صعودی، بهبود چشمگیری در بالا آمدن رتبهٔ واژگان تخصصی در پیکرهٔ عمومی حاصل می‌شود و این امر می‌تواند موجب تسریع فرایند انتخاب واژگان گردد. فرمول‌های مربوط به این روش‌ها در جدول ۱، آورده شده است.

٥-۳. بسامدشماری پیکرهٔ اختصاصی بهبودیافته

پس از کسب نتایج موفقیت‌آمیز در به‌کارگیری روش کسر کردن واژگان با بسامد بالا از فراوانی در پیکرهٔ اصلی عمومی، متخصصان اقدام به اجرای روش مذکور در پیکرهٔ اختصاصی نمودند. نتایج تجربی حاصل دلالت بر بهبود انتخاب و بالا آمدن رتبهٔ واژگان تخصصی در این روش دارد.

جدول ۱. شرح محاسبهٔ روش‌های کلاسیک بسامدشماری

روش‌های کلاسیک	
روش	**شرح**
بسامدشماری پیکرهٔ عمومی	فراوانی واژگان در پیکرهٔ اصلی عمومی = a
بسامدشماری پیکرهٔ اختصاصی	فراوانی واژگان در پیکرهٔ اختصاصی = b
بسامدشماری پیکرهٔ عمومی بهبودیافته	a فراوانی کل واژگان در پیکرهٔ اصلی عمومی = c
بسامدشماری پیکرهٔ اختصاصی بهبودیافته	b فراوانی کل واژگان در پیکرهٔ اصلی تخصصی = d

٤. نتایج و برآورد مدل

در جدول ۲، نتایج حاصل از مرتب‌سازی ۵۰ واژهٔ نخست با بیشترین فراوانی در هر روش نمایش داده شده است. در روش بسامدشماری پیکرهٔ اصلی عمومی، فراوانی واژگان در پیکرهٔ اصلی مورد شمارش قرار گرفت و بر اساس بیشترین فراوانی واژگان مرتب‌سازی شد. همان‌گونه که جدول نیز نمایش می‌دهد، مشاهده می‌شود که واژگان غیرتخصصی

در رتبه‌های بالای جدول قرار گرفته‌اند. جهت بررسی دقیق‌تر، روش بسامدشماری در پیکرهٔ اختصاصی مورد استفاده قرار گرفت و نتایج مشابهی حاصل گردید. بنابراین، با توجه به نتایج تجربی به‌دست‌آمده می‌توان استدلال نمود که در پیکرهٔ مورد بررسی، واژگان غیرتخصصی در مقایسه با واژگان تخصصی بیشتر تکرار شده‌اند و بنابراین، هنگام مرتب‌سازی بر اساس فراوانی، جایگاهی بالاتر به خود اختصاص می‌دهند که پدیده‌ای نامطلوب است.

به‌منظور کاهش مشکلات پدیدآمده و بالابردن توانایی‌های مدل‌ها و حفظ رویّهٔ پردازش اطلاعات، دو روش بسامدشماری پیکرهٔ اصلی عمومی بهبودیافته و بسامدشماری پیکرهٔ اختصاصی بهبودیافته مورد بررسی قرار گرفت. همان‌طور که جدول ۲، نیز نشان می‌دهد، روش بسامدشماری پیکرهٔ اصلی عمومیِ بهبودیافته به‌دلیل طیف گستردهٔ واژگان پیکرهٔ عمومی از عملکرد مطلوبی برخوردار نبوده و حتی نتایج ضعیف‌تری نسبت به روش اولیه در پیکرهٔ عمومی حاصل می‌گردد. روش بسامدشماری پیکرهٔ اختصاصیِ بهبودیافته جهت بررسی همین رویکرد برای پیکرهٔ تخصصی مورد استفاده قرار گرفت. نتایج به‌دست‌آمده در جدول ۲، نشان می‌دهد که اِعمال این روش تأثیر بسیار عمده‌ای در شناسایی واژگان تخصصی در پیکرهٔ تخصصی دارد تا آنجا که اعمال آن موجب شناسایی صحیح ۶۴ درصد از واژگان تخصصی در ۵۰ واژهٔ مرتب‌سازی‌شده با بیشترین فراوانی شده است.

از منظری دیگر، جهت بررسی تأثیر اِعمال روش‌های بهبودیافته دامنهٔ واژگان مورد بررسی به ۱۰۰، ۱۵۰ و ۲۰۰ واژهٔ نخست افزایش یافت. مشاهده می‌شود که تأثیر افزایش دامنهٔ واژگان مرتب‌سازی‌شده مثبت بوده و موجب می‌گردد درصد واژگان تخصصی انتخاب‌شده بر اساس روش‌های بهبودیافته افزایش یابد. پدیدهٔ مثبت دیگر حاصل از افزایش دامنهٔ به ثبات رسیدن درصد واژگان تخصصی انتخاب‌شده توسط روش‌های بسامدشماری است. بنابراین، نتایج به‌دست‌آمده دلالت بر این موضوع دارد که رویکرد بسامدشماری با اعمال اصلاحاتی در زبان فارسی قابلیت شناسایی واژگان تخصصی را داشته و می‌تواند مورد استفاده قرار گیرد.

جدول ۲. نتایج روش‌های بسامدی در ۵۰ واژهٔ اول

روش	درصد واژگان در ۵۰ واژهٔ پرتکرار دامنهٔ ۵۰-۱	
	واژگان غیرتخصصی (درصد)	واژگان تخصصی (درصد)
بسامدشماری پیکرهٔ اصلی عمومی	۹۰	۱۰
بسامدشماری پیکرهٔ اختصاصی	۱۰۰	۰
بسامدشماری پیکرهٔ اصلی عمومی بهبودیافته	۹۴	۶
بسامدشماری پیکرهٔ اختصاصی بهبودیافته	۳۶	۶۴

جدول ۳. نتایج روش‌های بسامدی در ۵۰ واژهٔ دوم

روش	درصد واژگان در ۵۰ واژهٔ پرتکرار دامنهٔ ۱۰۰-۵۱	
	واژگان غیرتخصصی (درصد)	واژگان تخصصی (درصد)
بسامدشماری پیکرهٔ اصلی عمومی	۷۴	۲۶
بسامدشماری پیکرهٔ اختصاصی	۱۰۰	۰
بسامدشماری پیکرهٔ اصلی عمومی بهبودیافته	۷۲	۲۸
بسامدشماری پیکرهٔ اختصاصی بهبودیافته	۲۰	۸۰

جدول ٤. نتایج روش‌های بسامدی در ۵۰ واژهٔ سوم

روش	درصد واژگان در ۵۰ واژهٔ پرتکرار دامنهٔ ۱۵۱-۱۰۱	
	واژگان غیرتخصصی (درصد)	واژگان تخصصی (درصد)
بسامدشماری پیکرهٔ اصلی عمومی	۷۶	۲۴
بسامدشماری پیکرهٔ اختصاصی	۹۸	۲
بسامدشماری پیکرهٔ اصلی عمومی بهبودیافته	۵۸	۴۲
بسامدشماری پیکرهٔ اختصاصی بهبودیافته	۲۶	۷۴

جدول ۵. نتایج روش‌های بسامدی در ۵۰ واژهٔ چهارم

روش	درصد واژگان در ۵۰ واژهٔ پرتکرار دامنهٔ ۲۰۰-۱۵۱	
	واژگان غیرتخصصی (درصد)	واژگان تخصصی (درصد)
بسامدشماری پیکرهٔ اصلی عمومی	۷۶	۲۴
بسامدشماری پیکرهٔ اختصاصی	۱۰۰	۰
بسامدشماری پیکرهٔظ اصلی عمومی بهبودیافته	۶۰	۴۰
بسامدشماری پیکرهٔ اختصاصی بهبودیافته	۲۸	۷۲

شکل ۱. مقایسهٔ روند دقت روش‌های بسامدشماری در ٤ دامنهٔ مورد بررسی

دویست واژه استخراج‌شدهٔ پربسامد بر اساس هر یک از روش‌ها در پیوست ۱، نمایش داده شده است.

۵. نتیجه‌گیری و بحث

نتایج به‌دست‌آمده از پژوهش نشان می‌دهد که استفاده از روش‌های آماری در پیکره‌های زبانی به‌طور خودکار از توانایی بالایی در استخراج واژگان پایهٔ علوم پزشکی برخوردار است. در پژوهش حاضر به بررسی تکنیک‌های مرسوم بسامدشماری در زبان فارسی و با منظور دستیابی به رویّهٔ علمی جهت استخراج خودکار واژگان پایهٔ علوم پزشکی

پرداخته شد. در این مسیر از خانوادهٔ مدل‌های بسامدشماری با رویکردهای بسامدشماری پیکرهٔ عمومی، بسامدشماری پیکرهٔ اختصاصی و روش‌های بهبودیافتهٔ آن‌ها استفاده گردید.

نتایج به‌دست‌آمده نشان می‌دهد که قابلیت استفاده از روش‌های بسامدشماری پیکره‌بنیاد در زبان فارسی وجود دارد و با استفاده از این روش‌ها می‌توان واژگان تخصصی را برای اهداف آموزشی و تهیهٔ دانشنامه‌های تخصصی و نظایر آن به‌کار برد. همچنین، از روش‌های بهبودیافته می‌توان جهت رفع چالش‌های عمدهٔ پژوهش، همچون جداسازی واژگان پرتکرار غیرتخصصی از فهرست نهایی استفاده نمود.

و نکتهٔ آخر این‌که برای دستیابی به عملکردی بهتر در روش‌های بسامدشماری نیاز است تا دامنهٔ واژگان نهایی به‌اندازهٔ کافی بزرگ باشد تا خروجی‌های به‌دست‌آمده از ثبات و دقت بالاتری برخوردار گردند.

فهرست منابع

بیرنگ محمدرضا و علی‌اصغر بشیری. ۱۳۹۶. تحلیل بسامدی شواهد شعری فارسی در نفثةالمصدور و بازیابی منابع برخی از این اشعار. *مجموعه مقالات چهارمین همایش متن‌پژوهی ادبی فارسی*، آذر ۱۳۹۶. تهران: حکایت قلم نوین.

جهانگردی کیومرث، مصطفی عاصی، آزیتا افراشی، و امیررضا وکیلی فرد. ۱۳۹۵. واژه در کتاب آموزش زبان فارسی به غیرفارسی‌زبانان: پژوهشی پیکره‌بنیاد. *پژوهشنامه آموزش زبان فارسی به غیر فارسی‌زبانان* ۵ (۲): ۳-۲۶.

راد، فرهاد، حمید پروین، آتوسا دهباشی، و بهروز مینایی بیدگلی. ۱۳۹۵. ارائهٔ روشی جدید برای شاخص‌گذاری خودکار و استخراج کلمات کلیدی برای بازیابی اطلاعات و خوشه‌بندی متون. *فصلنامه پردازش علایم و داده‌ها* ۱ (۲۷): ۱۰۰-۸۷.

رسولی، محمدصادق، و بهروز مینایی بیدگلی. ۱۳۸۷. روشی جدید در خطایابی املایی در زبان فارسی. دومین کنفرانس داده‌یابی ایران، دانشگاه امیرکبیر ۱۱-۱۲ آذر ۱۳۸۷. تهران.

سپهری، مهرداد. ۱۳۹۵. بسامدنگاری و دستاوردهای آن در آموزش. *مجله زبان و زبان‌شناسی*. ۲ (۳۳): ۶۰-۴۷.

گزنی، علی. ۱۳۸۵. استخراج خودکار عبارت‌های کلیدی از متون مقاله‌های فارسی. *مجله کتابداری و اطلاع‌رسانی* ۹، (۳): ۹۵-۱۰۶.

معادی، معین، و کاظم فولادی‌قلعه. ۱۳۹۶. ارائهٔ روشی برای استخراج کلمات کلیدی در زبان فارسی. دومین کنفرانس بین‌المللی و سومین همایش ملی کاربرد فناوری‌های نوین در علوم مهندسی. دانشگاه تربت حیدریه. تربیت حیدریه.

References

Anette, Hulth. 2004. Combining Machine Learning and Natural Language Processing for Automatic Keyword Extraction. Doctoral Dissertation. Stockholm University.

Bin, He, and Zhang Yongzheng. 2018. Automatic Term Extraction in Large Text Corpora. Retrieved from:https://www.cs.dal.ca/~yongzhen/course/6509/report.pdf (accessed Sep 10, 2019)

Castellví, M. T. C., R. E. Bagot, and J. V. Palatresi. 2001. Automatic term detection: A review of current systems. *Recent advances in computational terminology* Natural Language Processing 2.PP:53-88. Johns Benjamin Publishing Co

Chujo, K., M. Utiyama, and K. Oghigian. 2006. Selecting level-specific Kyoto tourism vocabulary using statistical measures. In L. Yiu-nam, M. Jenkins, & H. Chung-shun (Eds.), *New aspects of English language teaching and learning*, pp. 126-138. Taipei, Taiwan: Crane Publishing Company Ltd.

Chujo, K., K. Oghigian, C. Nishigaki, M. Utiyama, & T. Nakamura. 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., C. Nishigaki, & M. Utiyama. 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, p. 2005.

Chujo K, Utiyama Masao, Nakamura Takahiro. 2007. Extracting Level-Specific Science and Technology Vocabulary From, the Corpus of Professional English (CPE), Retrieved form : http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2007/47Paper.pdf (accessed Sep. 10, 2018).

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

Chujo Kiyomi. 2004. Measuring Vocabulary Levels of English Textbooks and Tests Using a BNC Lemmatised High Frequency Word List, Nihon University. *Language and computers*: 231-249. Amsterdam: Rodopi.

Chujo, Kiyomi, and Nishigaki Chikako. 2006. Creating Spoken Academic Vocabulary Lists from the British National Corpus. *Practical English Studies* 12: 3-19.

Coxhead, A., and P. Nation. 2001. The specialised vocabulary of English for academic purposes. *Research perspectives on English for academic purposes*. Cambridge Applied Linguistics, pp. 252-267. Cambridge: Cambridge University Press.

Daille, B.1994 . Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. *The Balancing Act: Combining Symbolic and Statistical Approaches to Language.* Las Cruces: New Mexico State University.

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

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

Frantzi, K., S. Ananiadou, and H. Mima. 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. In Aquilino Sánchez & M. Almela (Eds.), A mosaic of corpus linguistics: selected approaches (pp. 269–291). Frankfurt am Main: Peter Lang

Gilquin, G., and S. Granger. 2015. From design to collection of learner corpora. *The Cambridge handbook of learner corpus research* 3 (1): 9-34.

Lossio-Ventura, J. A., C. Jonquet, M. Roche, and M. Teisseire. 2016. Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal* 19 (1-2): 59-99.

Herriman, I., and M. B. Aronsson. 2009. Writing in English. *Corpora and language teaching* Philadelphia, USA: John Benjamins Publishing Company.

Nakagawa, H., and T. Mori. 2002. A simple but powerful automatic term extraction method. *In COLING-02 on COMPUTERM 2002: second international workshop on computational terminology.* V. 14, pp. 1-7. John benjamins Publishing Co: United states

McEnery, A. M., and A. Wilson. 2001. *Corpus linguistics: an introduction*. 2nd ed. Edinburgh: Edinburgh University Press.

Patrick, Pantel, and Lin Dekang. 2001. A Statistical Corpus-Based Term Extractor. *Advances in Artificial Language* In Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence (AI'01). Springer-Verlag, Berlin, Heidelberg, 36–46.

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

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

Von Joachim Wermter. 2008. Collocation and Term Extraction Using Linguistically Enhanced Statistical Methods. Retrieved from, http://www.db-thueringen.de/servlets/DerivateServlet/Derivate-17176/Wermter/Dissertation.pdf. (accessed June 16, 2018).

Swales, J. M. 2002. Integrated and fragmented worlds: EAP materials and corpus linguistics. In J. Flowedew (ed). *Academic discourse*. London: Longman. pp.150-164.

Gómez Guinovart, X., & A. Simoes. 2009. Parallel corpus-based bilingual terminology extraction. *In CEUR Workshop Proceedings*, URI: http://hdl.handle.net/1822/17500 (accessed June 20, 2018).

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

پیوست ۱: دویست واژه استخراج‌شدهٔ پربسامد بر اساس هر یک از روش‌ها

روش های کلاسیک							
شماره	بسامد شماری پیکره عمومی	آماره	بسامد شماری پیکره اختصاصی	آماره	بسامد شماری پیکره عمومی بهبودیافته	آماره	بسامد شماری پیکره اختصاصی بهبودیافته	آماره
۱	و	۴۱۱۷۳۹۰	و	۱۷۵۵۲	منع	۱۰۳۴۲۱۱۶۹	مژک	۵۷۳۵۵۰
۲	در	۳۰۷۱۲۸۹	در	۱۵۸۰۲	نبود	۱۰۳۴۲۰۸۰۹	هتروتروف	۵۷۳۵۴۹
۳	به	۲۶۴۹۲۳۷	به	۱۵۶۲۴	دستاوردهای	۱۰۳۴۲۰۷۸۶	پزشکی	۵۷۳۵۴۹
۴	از	۲۱۱۱۵۲۸	از	۱۴۴۸۰	نیت	۱۰۳۴۲۰۵۳۵	کربن	۵۷۳۵۴۶
۵	می	۱۷۶۵۵۲۹	که	۱۱۰۲۹	قزوین	۱۰۳۴۲۰۴۹۴	بیماری‌ها	۵۷۳۵۴۳
۶	را	۱۴۳۶۰۸۱	این	۱۰۶۰۵	یزد	۱۰۳۴۲۰۴۲۱	تراکم	۵۷۳۵۴۲
۷	است	۱۳۳۰۴۳۴	را	۸۱۴۸	تلاش‌های	۱۰۳۴۲۰۳۶۶	چرخه	۵۷۳۵۴۱

روش های کلاسیک								
شماره	بسامد شماری پیکره عمومی	آماره	بسامد شماری پیکره اختصاصی	آماره	بسامد شماری پیکره عمومی بهبودیافته	آماره	بسامد شماری پیکره اختصاصی بهبودیافته	آماره
۸	که	۱۲۱۵۵۲۳	است	۸۰۵۱	آورد	۱۰۳۴۱۹۶۰۷	محرک	۵۷۳۵۴۰
۹	با	۱۰۹۸۶۲۳	می	۵۳۵۹	دل	۱۰۳۴۱۹۱۵۱	باکتری	۵۷۳۵۴۰
۱۰	های	۱۰۷۲۸۰۲	با	۴۴۰۱	کوتاهی	۱۰۳۴۱۹۱۳۷	آندوپلاسمی	۵۷۳۵۳۹
۱۱	این	۵۷۱۷۲۱	های	۴۳۳۸	اول	۱۰۳۴۱۹۰۱۸	کپسول	۵۷۳۵۳۸
۱۲	ها	۵۱۲۹۵۴	برای	۳۸۰۴	دستگاههای	۱۰۳۴۱۸۱۹۷	باکتری‌ها	۵۷۳۵۳۵
۱۳	آن	۴۸۰۳۱۱	آن	۳۶۶۲	اتحادیه	۱۰۳۴۱۸۱۶۱	غیرجنسی	۵۷۳۵۳۵
۱۴	شود	۳۹۴۳۵۹	یک	۳۱۱۹	ساری	۱۰۳۴۱۸۱۱۲	لاکتوز	۵۷۳۵۳۴
۱۵	سلول	۳۷۵۱۳۰	خود	۲۹۴۱	قم	۱۰۳۴۱۸۰۰۹	پرده‌ی	۵۷۳۵۳۳
۱۶	شکل	۳۵۹۶۱۱	شده	۲۸۰۹	خدا	۱۰۳۴۱۷۶۹۵	آبکشی	۵۷۳۵۳۱
۱۷	این	۳۱۶۷۰۷	کرد	۲۴۱۳	قوه	۱۰۳۴۱۷۲۸۸	اندازه	۵۷۳۵۳۱
۱۸	برای	۲۹۰۹۶۹	شود	۲۳۲۸	راهکارهای	۱۰۳۴۱۷۲۳۱	غشای	۵۷۳۵۲۷
۱۹	یک	۲۸۸۸۴۹	بر	۲۳۱۲	عمران	۱۰۳۴۱۷۱۹۵	پانکراس	۵۷۳۵۲۷
۲۰	دارد	۲۷۸۷۹۷	تا	۲۲۸۹	اهدا	۱۰۳۴۱۷۱۷۳	اصطکاک	۵۷۳۵۲۶
۲۱	خود	۲۷۰۲۰۳	سال	۲۱۶۰	نهم	۱۰۳۴۱۶۹۰۷	لولهٔ	۵۷۳۵۲۴
۲۲	که	۲۶۸۰۲۱	شد	۲۱۵۰	او	۱۰۳۴۱۶۸۰۵	جمله‌های	۵۷۳۵۲۳
۲۳	شده	۲۵۶۲۷۰	کشور	۲۰۹۸	سیل	۱۰۳۴۱۶۷۸۲	آهک	۵۷۳۵۱۹
۲۴	دو	۲۵۶۰۹۸	ها	۲۰۶۳	هایش	۱۰۳۴۱۶۱۸۶	آهکی	۵۷۳۵۱۵
۲۵	بدن	۲۳۴۵۰۸	نیز	۱۹۷۵	رود	۱۰۳۴۱۵۹۴۳	باکتریایی	۵۷۳۵۱۵
۲۶	دارند	۲۳۳۳۰۹	گفت	۱۹۳۷	باشیم	۱۰۳۴۱۵۸۶۴	بالک	۵۷۳۵۱۴
۲۷	کند	۲۳۲۴۶۷	بود	۱۸۲۰	ری	۱۰۳۴۱۵۷۸۵	بیماری‌هایی	۵۷۳۵۱۴
۲۸	ای	۲۲۷۷۶۴	هم	۱۸۱۵	امام	۱۰۳۴۱۵۵۹۲	بکرزایی	۵۷۳۵۱۴
۲۹	کنند	۲۱۸۲۶۲	ای	۱۷۵۶	مخالفت	۱۰۳۴۱۵۲۳۰	بندپایان	۵۷۳۵۱۳
۳۰	وجود	۲۱۱۳۲۹	کند	۱۷۴۱	برد	۱۰۳۴۱۵۱۲۵	بیکربنات	۵۷۳۵۱۳
۳۱	هستند	۲۱۱۳۲۶	ایران	۱۶۹۲	گشته	۱۰۳۴۱۴۶۸۸	توکسین	۵۷۳۵۱۱
۳۲	هر	۱۹۴۷۱۹	ما	۱۶۶۳	تیتر	۱۰۳۴۱۴۵۵۱	دریچهٔ	۵۷۳۵۰۹

روش های کلاسیک

شماره	بسامد شماری پیکره عمومی	آماره	بسامد شماری پیکره اختصاصی	آماره	بسامد شماری پیکره عمومی بهبودیافته	آماره	بسامد شماری پیکره اختصاصی بهبودیافته	آماره
۳۳	کنید	۱۸۳۵۷۵	اما	۱۴۹۴	زندگی	۱۰۳۴۱۴۳۲۳	رده	۵۷۳۵۰۳
۳۴	شوند	۱۸۲۳۷۱	وی	۱۴۵۳	قبلا	۱۰۳۴۱۴۰۳۱	زمینه	۵۷۳۵۰۰
۳۵	آن ها	۱۷۸۵۴۱	دارد	۱۳۶۶	افراد	۱۰۳۴۱۴۰۰۲	سیرنشده	۵۷۳۴۹۷
۳۶	قرار	۱۷۷۵۵۳	یا	۱۳۳۹	مخرب	۱۰۳۴۱۳۶۳۳	عفونت‌های	۵۷۳۴۹۲
۳۷	اند	۱۷۳۸۱۹	هر	۱۲۹۱	گرگان	۱۰۳۴۱۳۳۸۴	قسمت‌های	۵۷۳۴۹۲
۳۸	آب	۱۷۰۸۲۸	باید	۱۲۷۹	روستاهای	۱۰۳۴۱۳۲۷۵	کننده‌ی	۵۷۳۴۹۱
۳۹	خون	۱۶۹۰۴۱	قرار	۱۲۷۶	باری	۱۰۳۴۱۳۰۷۵	مایه‌کوبی	۵۷۳۴۹۱
۴۰	بر	۱۶۵۹۶۷	دو	۱۱۹۸	گویند	۱۰۳۴۱۲۹۰۸	مردمک	۵۷۳۴۸۹
۴۱	یا	۱۶۵۳۰۴	آنها	۱۱۴۱	نظرهای	۱۰۳۴۱۲۴۵۶	میانبرگ	۵۷۳۴۸۶
۴۲	مواد	۱۶۲۷۰۲	کرده	۱۱۳۰	خارجی	۱۰۳۴۱۲۳۲۷	موکوز	۵۷۳۴۸۵
۴۳	انجام	۱۶۲۵۱۲	او	۱۱۲۵	سفرهای	۱۰۳۴۱۲۱۹۱	میوزین	۵۷۳۴۸۲
۴۴	نام	۱۶۱۶۸۹	مورد	۱۱۱۴	تی	۱۰۳۴۱۲۱۰۰	نرهایی	۵۷۳۴۷۸
۴۵	تا	۱۵۴۱۵۶	خواهد	۱۱۰۹	باورهای	۱۰۳۴۱۲۰۶۰	یابی	۵۷۳۴۷۷
۴۶	هم	۱۵۱۳۴۹	کنند	۱۰۹۳	مولف	۱۰۳۴۱۱۹۴۳	چارگف	۵۷۳۴۷۳
۴۷	درون	۱۴۵۷۵۴	دیگر	۱۰۶۰	محورهای	۱۰۳۴۱۱۸۱۲	کپکهای	۵۷۳۴۷۱
۴۸	استفاده	۱۴۴۹۸۹	کار	۱۰۵۱	اش	۱۰۳۴۱۱۰۰۳	کیموس	۵۷۳۴۷۱
۴۹	گونه	۱۳۹۴۵۴	تهران	۱۰۳۹	دهیم	۱۰۳۴۱۰۵۰۲	گلیکوژن	۵۷۳۴۶۵
۵۰	صورت	۱۳۳۸۵۷	مردم	۹۹۴	نوشتار	۱۰۳۴۱۰۳۲۵	گردهٔ	۵۷۳۴۶۵
۵۱	چه	۱۳۳۰۹۷	باشد	۹۹۴	داخلی	۱۰۳۴۱۰۳۱۱	کربنی	۵۷۳۴۶۴
۵۲	نیز	۱۲۲۷۳۷	عنوان	۹۹۳	بندرعباس	۱۰۳۴۰۹۶۴۸	اتوزومی	۵۷۳۴۶۳
۵۳	مولکول	۱۲۲۲۲۹	وجود	۹۸۳	رفت	۱۰۳۴۰۹۶۱۰	اینترفاز	۵۷۳۴۶۱
۵۴	یا	۱۲۱۵۴۹	سازمان	۹۷۰	شدن	۱۰۳۴۰۹۳۹۲	بساک	۵۷۳۴۶۰
۵۵	مانند	۱۲۰۲۴۰	اسلامی	۹۰۸	پوند	۱۰۳۴۰۹۲۵۹	بزاقی	۵۷۳۴۶۰
۵۶	شدن	۱۱۹۹۸۸	نظر	۸۷۸	سامانه	۱۰۳۴۰۹۰۸۶	جمعیت ها	۵۷۳۴۵۹
۵۷	تولید	۱۱۹۷۲۹	روز	۸۷۴	نخست	۱۰۳۴۰۸۶۷۴	تکاملی	۵۷۳۴۵۷

روش های کلاسیک								
شماره	بسامد شماری پیکره عمومی	آماره	بسامد شماری پیکره اختصاصی	آماره	بسامد شماری پیکره عمومی بهبودیافته	آماره	بسامد شماری پیکره اختصاصی بهبودیافته	آماره
۵۸	حرکت	۱۱۷۴۶۹	دولت	۸۵۷	زاهدان	۱۰۳۴۰۸۶۱۸	جوانه‌های	۵۷۳۴۵۲
۵۹	نشان	۱۱۷۳۴۸	اگر	۸۴۶	یافته	۱۰۳۴۰۸۲۸۰	حدود	۵۷۳۴۵۲
۶۰	میشود	۱۱۴۹۹۸	شرکت	۸۴۱	سپر	۱۰۳۴۰۸۱۳۸	خمیرکره	۵۷۳۴۵۱
۶۱	روی	۱۱۳۵۸۴	پس	۸۳۸	عالی	۱۰۳۴۰۷۶۷۹	درشکل	۵۷۳۴۵۰
۶۲	دیگر	۱۱۳۲۶۰	گروه	۸۲۴	جاهای	۱۰۳۴۰۷۶۰۳	درونپوش	۵۷۳۴۵۰
۶۳	اگر	۱۱۲۴۵۱	است	۸۰۷	باوری	۱۰۳۴۰۷۴۸۶	دیفتری	۵۷۳۴۵۰
۶۴	بخش	۱۱۲۳۱۴	این	۸۰۰	فورا	۱۰۳۴۰۷۴۶۹	دولختی	۵۷۳۴۴۹
۶۵	دستگاه	۱۱۱۵۵۷	دست	۷۹۴	کاستن	۱۰۳۴۰۷۴۵۳	ریشهٔ	۵۷۳۴۴۴
۶۶	ژن	۱۱۱۱۴۴	صورت	۷۹۲	قبلی	۱۰۳۴۰۶۷۷۵	سه‌لختی	۵۷۳۴۴۴
۶۷	رشد	۱۰۷۷۲۰	همه	۷۸۶	نمادهای	۱۰۳۴۰۶۷۰۱	فولیکولی	۵۷۳۴۴۳
۶۸	بعضی	۱۰۷۵۰۲	بین	۷۷۹	هدفمند	۱۰۳۴۰۶۴۷۲	لیتیک	۵۷۳۴۴۳
۶۹	یک	۱۰۶۱۵۳	گزارش	۷۶۵	باشید	۱۰۳۴۰۶۴۱۸	لجیستیک	۵۷۳۴۳۷
۷۰	باشد	۱۰۵۶۲۹	من	۷۶۵	عالی	۱۰۳۴۰۵۹۲۴	ماهیچه‌	۵۷۳۴۳۴
۷۱	اما	۱۰۴۹۷۱	پیش	۷۵۵	ره	۱۰۳۴۰۵۵۴۷	میکند	۵۷۳۴۳۲
۷۲	بین	۱۰۲۴۹۳	تیم	۷۵۲	آلود	۱۰۳۴۰۵۲۰۱	مژکداران	۵۷۳۴۳۱
۷۳	دهد	۱۰۲۲۱۱	انجام	۷۵۰	یاسوج	۱۰۳۴۰۴۸۴۴	میابد	۵۷۳۴۲۸
۷۴	گیاهان	۱۰۰۸۵۶	حال	۷۴۰	آشنایی	۱۰۳۴۰۳۸۴۳	نوکلئوتیدی	۵۷۳۴۲۵
۷۵	می‌شود	۹۸۴۸۴	برنامه	۷۳۳	کوچ	۱۰۳۴۰۳۳۳۰	هسته‌	۵۷۳۴۲۴
۷۶	پس	۹۷۶۴۵	اعلام	۷۲۹	مرگ	۱۰۳۴۰۳۲۳۰	پایدارکننده	۵۷۳۴۱۹
۷۷	انسان	۹۵۶۴۴	یکی	۷۱۹	شعر	۱۰۳۴۰۲۷۹۳	پروتئینی	۵۷۳۴۱۶
۷۸	گوش	۹۵۲۵۱	طرح	۷۱۵	کشنده	۱۰۳۴۰۲۳۱۲	پلاسموسیت	۵۷۳۴۱۴
۷۹	تشکیل	۹۴۶۹۱	بخش	۷۱۲	هفته	۱۰۳۴۰۱۶۸۲	پیام‌ها	۵۷۳۴۱۲
۸۰	ماده	۹۴۴۸۸	چه	۷۱۰	آورند	۱۰۳۴۰۱۶۸۰	پروتوپلاستی	۵۷۳۴۱۲
۸۱	جانوران	۹۳۹۳۵	توجه	۶۹۶	پژوهشگران	۱۰۳۴۰۱۶۰۳	پروتئین‌ها	۵۷۳۴۰۷
۸۲	افراد	۹۲۶۶۵	هستند	۶۸۵	زلزله	۱۰۳۴۰۱۴۸۲	کالبدشناسی	۵۷۳۴۰۱

روش های کلاسیک								
شماره	بسامد شماری پیکره عمومی	آماره	بسامد شماری پیکره اختصاصی	آماره	بسامد شماری پیکره عمومی بهبودیافته	آماره	بسامد شماری پیکره اختصاصی بهبودیافته	آماره
۸۳	دست	۹۱۰۲۱	شهر	۶۶۴	دربرگیرنده	۱۰۳۳۹۸۶۰۳	کریک	۵۷۳۳۹۶
۸۴	چند	۹۰۸۰۹	استفاده	۶۶۲	مفاهیمی	۱۰۳۳۹۸۵۲۹	کلاله	۵۷۳۳۹۵
۸۵	سلولی	۹۰۰۲۸	داشته	۶۵۰	بجز	۱۰۳۳۹۸۲۵۵	شوند	۵۷۳۳۹۴
۸۶	زندگی	۸۹۵۸۹	هزار	۶۳۷	ویژه	۱۰۳۳۹۸۲۳۶	آمینواسیدها	۵۷۳۳۹۰
۸۷	وارد	۸۸۶۳۴	جامعه	۶۳۴	پیشنهادها	۱۰۳۳۹۷۴۰۹	آنتریدی	۵۷۳۳۸۸
۸۸	قلب	۸۸۰۷۷	گذشته	۶۳۴	اعصاب	۱۰۳۳۹۶۷۴۷	اندازه‌ی	۵۷۳۳۸۵
۸۹	سطح	۸۷۹۳۷	درصد	۶۲۴	جداره	۱۰۳۳۹۵۹۰۲	اوریک	۵۷۳۳۷۹
۹۰	دهید	۸۷۲۰۶	نیست	۶۱۶	متون	۱۰۳۳۹۵۲۸۵	بیماریزا	۵۷۳۳۷۸
۹۱	مورد	۸۶۷۷۲	داد	۶۱۴	البته	۱۰۳۳۹۴۶۳۱	تراوایی	۵۷۳۳۷۵
۹۲	کار	۸۱۳۴۷	اقتصادی	۶۰۶	بگیریم	۱۰۳۳۹۴۶۲۳	تجزیه‌ٔ	۵۷۳۳۷۲
۹۳	اثر	۸۱۰۹۴	افزایش	۶۰۴	پژوهش	۱۰۳۳۹۴۶۰۰	تیغک	۵۷۳۳۶۶
۹۴	دهند	۸۰۸۴۹	توسعه	۶۰۳	آلمانی	۱۰۳۳۹۴۳۶۷	دیوارهه	۵۷۳۳۶۵
۹۵	بیشتر	۷۹۵۶۲	داشت	۵۹۵	دیگری	۱۰۳۳۹۴۱۹۹	رادیواکتیو	۵۷۳۳۶۲
۹۶	گیاه	۷۹۴۳۱	منطقه	۵۹۰	مناسبتر	۱۰۳۳۹۳۷۲۴	رابطه‌ٔ	۵۷۳۳۵۴
۹۷	جانداران	۷۸۷۲۰	بوده	۵۷۸	واقعی	۱۰۳۳۹۲۲۴۲	ریشه‌ٔ	۵۷۳۳۵۲
۹۸	سه	۷۸۱۷۱	راه	۵۷۵	نوین	۱۰۳۳۹۲۰۰۶	سارکومر	۵۷۳۳۵۲
۹۹	تعداد	۷۸۱۰۳	ملی	۵۷۳	بیابانی	۱۰۳۳۹۱۷۷۳	شیرهٔ	۵۷۳۳۴۱
۱۰۰	میکنند	۷۶۸۰۸	داده	۵۷۲	مفرط	۱۰۳۳۹۱۲۶۶	لاکتات	۵۷۳۳۳۵
۱۰۱	مختلف	۷۶۴۹۶	همین	۵۶۶	آبگیر	۱۰۳۳۹۱۱۶۳	میسازند	۵۷۳۳۳۲
۱۰۲	گروه	۷۵۳۲۶	همچنین	۵۶۳	آب	۱۰۳۳۸۹۹۱۹	میتوکندری	۵۷۳۳۲۶
۱۰۳	راه	۷۴۹۸۴	دهد	۵۵۷	چون	۱۰۳۳۸۹۵۷۷	ناحیه‌ی	۵۷۳۳۲۵
۱۰۴	ساختار	۷۳۷۴۰	ایجاد	۵۴۶	انسان‌ها	۱۰۳۳۸۸۴۰۷	نیکوتین	۵۷۳۳۲۳
۱۰۵	طور	۷۳۶۳۷	اجتماعی	۵۴۶	بشره	۱۰۳۳۸۸۰۹۵	ویروسی	۵۷۳۳۲۰
۱۰۶	مغز	۷۳۱۳۰	جهان	۵۴۴	بیضه	۱۰۳۳۸۶۸۰۰	پیکان	۵۷۳۳۰۸
۱۰۷	یک	۷۲۸۹۷	قانون	۵۳۶	دارد	۱۰۳۳۸۶۷۳۲	ژن‌های	۵۷۳۳۰۳

روش های کلاسیک									
شماره	بسامد شماری پیکره عمومی	آماره	بسامد شماری پیکره اختصاصی	آماره	بسامد شماری پیکره عمومی بهبودیافته	آماره	بسامد شماری پیکره اختصاصی بهبودیافته	آماره
۱۰۸	ساخته	۷۲۶۶۱	تنها	۵۳۳	تولید	۱۰۳۳۸۶۱۸۳	کاریوتیپ	۵۷۳۲۹۵
۱۰۹	هنگام	۷۱۱۷۹	حضور	۵۳۲	ازای	۱۰۳۳۸۵۴۷۲	کلسترول	۵۷۳۲۹۴
۱۱۰	باکتری	۷۰۷۰۴	زمینه	۵۲۶	دورانداختن	۱۰۳۳۸۵۳۴۱	گسلنده	۵۷۳۲۹۲
۱۱۱	نظر	۷۰۵۳۶	دارند	۵۲۳	دوره	۱۰۳۳۸۴۰۵۴	یوکاریوت	۵۷۳۲۸۴
۱۱۲	باعث	۷۰۳۰۳	بیشتر	۵۲۰	رگ	۱۰۳۳۸۳۸۰۹	یاخته‌ی	۵۷۳۲۸۲
۱۱۳	پوست	۷۰۲۸۰	بسیار	۵۱۹	شکل	۱۰۳۳۸۳۵۵۶	دیواره	۵۷۳۲۷۴
۱۱۴	عمل	۷۰۲۷۲	سه	۵۱۳	فوق	۱۰۳۳۷۸۱۳۱	آنتی‌بیوتیک	۵۷۳۲۶۴
۱۱۵	آنزیم	۶۹۶۱۸	کردند	۵۱۰	قارچی	۱۰۳۳۷۷۶۹۷	جلبکها	۵۷۳۲۶۴
۱۱۶	دیگر	۶۹۰۰۱	ماه	۵۰۸	مجموعه	۱۰۳۳۷۶۹۶۶	جامعه‌ستیز	۵۷۳۲۶۱
۱۱۷	مثل	۶۸۱۵۶	چند	۵۰۴	میلیون	۱۰۳۳۷۶۸۲۶	دریازی	۵۷۳۲۴۸
۱۱۸	می‌کند	۶۷۴۹۴	گرفته	۵۰۳	میدانید	۱۰۳۳۷۶۲۶۹	رشته‌ها	۵۷۳۲۳۶
۱۱۹	سبب	۶۶۶۶۷	سوی	۵۰۲	نزد	۱۰۳۳۷۵۳۱۴	نحوه	۵۷۳۲۳۴
۱۲۰	ممکن	۶۶۶۴۸	میلیون	۴۹۷	زندگی	۱۰۳۳۷۴۸۷۵	نوترکیب	۵۷۳۲۱۳
۱۲۱	انتخاب	۶۶۳۷۴	ادامه	۴۹۱	طی	۱۰۳۳۷۴۰۹۴	پیک	۵۷۳۲۰۲
۱۲۲	زیر	۶۶۳۰۰	تولید	۴۸۶	هسته	۱۰۳۳۷۳۸۸۲	آسک	۵۷۳۱۸۸
۱۲۳	پاسخ	۶۳۴۲۱	میان	۴۸۶	پربسامد	۱۰۳۳۷۲۲۳۸	آسیب‌شناسی	۵۷۳۱۸۶
۱۲۴	ایجاد	۶۳۳۰۰	روی	۴۸۴	کار	۱۰۳۳۷۰۹۶۸	آمیزشی	۵۷۳۱۸۳
۱۲۵	نیز	۶۳۲۲۳	نسبت	۴۸۲	چاقی	۱۰۳۳۷۰۴۱۵	آنزیمها	۵۷۳۱۷۹
۱۲۶	بزرگ	۶۳۲۲۱	اشاره	۴۷۹	کوچه	۱۰۳۳۷۰۲۱۲	باریک	۵۷۳۱۶۹
۱۲۷	ماهیچه	۶۲۵۹۱	شدن	۴۷۴	چشم	۱۰۳۳۶۹۶۵۲	بیوتیک	۵۷۳۱۶۵
۱۲۸	افزایش	۶۲۱۱۰	کاهش	۴۷۳	گوجه	۱۰۳۳۶۹۲۵۹	حفره	۵۷۳۱۵۸
۱۲۹	شما	۶۱۶۷۳	هیچ	۴۷۰	گوگردی	۱۰۳۳۶۷۲۵۸	روده‌ی	۵۷۳۱۴۸
۱۳۰	کرد	۶۱۵۷۳	کردن	۴۶۷	گیلن	۱۰۳۳۶۶۴۱۹	ساکاریدها	۵۷۳۱۴۶
۱۳۱	بیماری	۶۱۰۶۰	نه	۴۶۶	بجنورد	۱۰۳۳۶۴۰۴۷	شاخک	۵۷۳۱۴۳
۱۳۲	هورمون	۶۰۰۵۹	نشان	۴۶۴	جایگزینی	۱۰۳۳۶۲۶۹۲	صفحه	۵۷۳۱۳۳

روش های کلاسیک

شماره	بسامد شماری پیکره عمومی	آماره	بسامد شماری پیکره اختصاصی	آماره	بسامد شماری پیکره عمومی بهبودیافته	آماره	بسامد شماری پیکره اختصاصی بهبودیافته	آماره
۱۳۳	رنگ	۵۹۷۰۲	زندگی	۴۶۱	روده	۱۰۳۳۶۲۴۱۰	عفونتزا	۵۷۳۱۳۳
۱۳۴	توجه	۵۹۳۵۶	شما	۴۵۵	ویروس	۱۰۳۳۵۹۹۳۲	فسفولیپیدها	۵۷۳۱۱۹
۱۳۵	دانه	۵۸۹۷۹	جهانی	۴۵۵	قدیمی	۱۰۳۳۵۹۲۷۴	قفسه	۵۷۳۱۱۲
۱۳۶	کردن	۵۸۴۳۱	آموزش	۴۵۴	محیط	۱۰۳۳۵۸۷۵۰	مارمولک	۵۷۳۱۰۶
۱۳۷	ما	۵۸۲۰۲	زمان	۴۵۴	شمال	۱۰۳۳۵۷۴۰۱	لیزوژنی	۵۷۳۱۰۴
۱۳۸	فعالیت	۵۸۱۲۸	دلیل	۴۴۹	عمد	۱۰۳۳۵۷۱۸۳	مشاهدهٔ	۵۷۳۰۹۸
۱۳۹	بیماری	۵۷۴۵۰	بعد	۴۴۸	زلزله	۱۰۳۳۵۳۷۵۹	نافلز	۵۷۳۰۹۱
۱۴۰	حاصل	۵۷۴۳۴	اینکه	۴۴۸	مخرب	۱۰۳۳۵۳۳۴۶	نمیشوند	۵۷۳۰۷۴
۱۴۱	داده	۵۷۴۱۷	چنین	۴۴۸	شیرخوار	۱۰۳۳۵۲۵۸۹	هستهٔ	۵۷۳۰۶۰
۱۴۲	یکدیگر	۵۷۳۰۰	حتی	۴۴۷	آخر	۱۰۳۳۵۲۴۵۲	همانندسازی	۵۷۳۰۵۷
۱۴۳	کمک	۵۷۲۲۴	توان	۴۴۶	جگر	۱۰۳۳۵۱۶۴۳	پلاسمودیومی	۵۷۳۰۵۲
۱۴۴	برای	۵۶۷۶۵	ولی	۴۴۶	قبل	۱۰۳۳۵۱۳۱۹	پوسته	۵۷۳۰۴۱
۱۴۵	سال	۵۶۶۲۳	درباره	۴۴۱	آتشفشانی	۱۰۳۳۴۹۹۰۵	کلنی	۵۷۳۰۲۴
۱۴۶	فشار	۵۶۲۸۵	بیش	۴۳۹	گاز	۱۰۳۳۴۷۵۵۵	کندتر	۵۷۳۰۲۰
۱۴۷	دارای	۵۶۲۲۹	برخی	۴۳۸	سیل	۱۰۳۳۴۷۴۳۴	چروکیده	۵۷۲۹۹۰
۱۴۸	رشته	۵۵۸۱۷	توسط	۴۳۸	فلزها	۱۰۳۳۴۵۱۷۴	کات	۵۷۲۹۲۸
۱۴۹	نقش	۵۵۶۴۴	نام	۴۳۷	هدفدار	۱۰۳۳۴۴۹۱۵	آنافاز	۵۷۲۹۱۵
۱۵۰	تر	۵۵۵۱۱	جدید	۴۳۵	یابد	۱۰۳۳۴۴۶۶۳	بساوایی	۵۷۲۹۰۵
۱۵۱	انرژی	۵۵۲۹۳	نفر	۴۳۴	ابزارهای	۱۰۳۳۴۳۳۵۴	بی‌هوازی	۵۷۲۹۰۴
۱۵۲	کروموزوم	۵۵۲۵۵	اول	۴۳۳	صورت	۱۰۳۳۴۲۶۷۴	تغییرگونه	۵۷۲۹۰۴
۱۵۳	حالت	۵۴۵۹۲	آغاز	۴۳۳	همه	۱۰۳۳۴۲۱۶۶	ریشه‌ی	۵۷۲۸۵۹
۱۵۴	پروتئین	۵۴۵۷۸	نظام	۴۲۷	بازنویسی	۱۰۳۳۳۱۸۰۶	فتوسیستم	۵۷۲۸۴۷
۱۵۵	مقدار	۵۴۴۰۱	طور	۴۲۷	اقلیم	۱۰۳۳۳۱۰۴۶	فاصله	۵۷۲۸۳۸
۱۵۶	یکی	۵۳۹۰۰	بازی	۴۲۵	باتلاق ها	۱۰۳۳۲۵۴۴۹	فیزیکی	۵۷۲۸۰۵
۱۵۷	انتقال	۵۳۰۹۱	افراد	۴۲۱	بیماری	۱۰۳۳۱۹۹۱۴	متافاز	۵۷۲۷۸۹

روش های کلاسیک								
شماره	بسامد شماری پیکره عمومی	آماره	بسامد شماری پیکره اختصاصی	آماره	بسامد شماری پیکره عمومی بهبودیافته	آماره	بسامد شماری پیکره اختصاصی بهبودیافته	آماره
۱۵۸	کوچک	۵۲۹۶۰	دوره	۴۲۰	زیادشدن	۱۰۳۳۱۹۱۴۹	ماهیچهٔ	۵۷۲۷۸۴
۱۵۹	موجود	۵۲۸۴۳	بررسی	۴۱۶	ساقه	۱۰۳۳۱۳۵۵۴	مصدر	۵۷۲۷۷۳
۱۶۰	آن‌ها	۵۲۸۰۳	استان	۴۱۳	سپاسگزاری	۱۰۳۳۱۰۷۴۷	می‌شنوید	۵۷۲۷۶۸
۱۶۱	نسبت	۵۲۷۱۲	فرهنگ	۴۱۲	سوختگی	۱۰۳۳۰۳۲۱۴	می‌روند	۵۷۲۷۵۷
۱۶۲	ترشح	۵۲۵۷۶	عمومی	۴۰۹	نجاران	۱۰۳۳۰۲۳۹۱	میوکارد	۵۷۲۷۰۰
۱۶۳	فقط	۵۲۴۴۷	مختلف	۴۰۷	همان‌طور	۱۰۳۳۰۲۲۰۱	نوترکیب	۵۷۲۶۳۵
۱۶۴	جمعیت	۵۱۹۹۵	شوند	۴۰۷	پشه	۱۰۳۲۹۹۵۹۹	هاورس	۵۷۲۶۲۲
۱۶۵	میشوند	۵۱۸۲۸	امور	۴۰۶	نشوند	۱۰۳۲۹۸۹۳۶	وظیفه‌ی	۵۷۲۶۱۹
۱۶۶	بین	۵۱۶۲۸	آنان	۴۰۶	گوسفند	۱۰۳۲۹۵۸۶۲	وراثتی	۵۷۲۶۰۷
۱۶۷	او	۵۱۵۱۹	تواند	۴۰۱	داد	۱۰۳۲۹۴۰۷۵	پیام‌های	۵۷۲۵۷۶
۱۶۸	طول	۵۱۲۷۰	پایان	۳۹۸	مسیرهای	۱۰۳۲۹۱۰۸۴	پلاستیکی	۵۷۲۵۵۹
۱۶۹	بعد	۵۰۹۰۱	تمام	۳۹۵	دندانی	۱۰۳۲۸۷۳۵۰	کشت	۵۷۲۵۳۲
۱۷۰	حدود	۵۰۸۷۲	گفته	۳۸۹	زیستن	۱۰۳۲۸۶۳۶۲	کره،	۵۷۲۴۴۵
۱۷۱	روش	۵۰۵۸۰	آینده	۳۸۷	پرده	۱۰۳۲۸۲۵۳۲	گلوکاگون	۵۷۲۴۰۴
۱۷۲	چگونه	۵۰۳۵۲	بزرگ	۳۸۵	کوشیده	۱۰۳۲۸۱۳۲۸	رونویسی	۵۷۲۲۳۵
۱۷۳	غذا	۵۰۲۱۵	ارائه	۳۸۵	کارا	۱۰۳۲۷۰۱۸۴	آرکگن	۵۷۲۲۰۶
۱۷۴	می‌شوند	۴۹۷۷۸	بسیاری	۳۸۴	بگذاری	۱۰۳۲۵۳۵۷۷	اسمی	۵۷۲۱۵۷
۱۷۵	لازم	۴۹۶۷۳	بار	۳۸۴	ثانوی	۱۰۳۲۵۳۵۷۴	اکتسابی	۵۷۲۱۴۲
۱۷۶	تقسیم	۴۹۳۱۱	جهت	۳۸۴	خارها	۱۰۳۲۴۶۶۴۱	انگل‌های	۵۷۲۰۸۳
۱۷۷	مشاهده	۴۹۱۱۸	نفت	۳۸۳	شیی	۱۰۳۲۳۷۱۳۹	بافت‌ها	۵۷۲۰۷۸
۱۷۸	هسته	۴۹۰۳۹	فیلم	۳۸۳	فولاد	۱۰۳۲۳۲۴۳۶	تلوفاز	۵۷۱۹۶۱
۱۷۹	اندازه	۴۸۹۶۰	کنیم	۳۷۹	استخوانی	۱۰۳۲۳۱۵۹۴	حشره‌ی	۵۷۱۹۲۳
۱۸۰	علت	۴۸۷۱۷	سرمایه	۳۷۵	تغییرها	۱۰۳۲۳۰۳۹۵	دریچه‌های	۵۷۱۸۳۵
۱۸۱	عصبی	۴۸۱۲۱	بدون	۳۷۴	دور	۱۰۳۲۰۸۸۰۵	رنگیزه	۵۷۱۸۰۰
۱۸۲	بسیاری	۴۸۰۹۸	کتاب	۳۷۲	نورسته	۱۰۳۲۰۸۶۳۳	ریبوز	۵۷۱۷۴۸

روش های کلاسیک							
شماره	بسامد شماری پیکره عمومی	آماره	بسامد شماری پیکره اختصاصی	آماره	بسامد شماری پیکره عمومی بهبودیافته	آماره	بسامد شماری پیکره اختصاصی بهبودیافته	آماره
۱۸۳	جسم	۴۷۹۹۶	چون	۳۷۱	آبجو	۱۰۳۱۹۶۸۸۲	سازنده‌ی	۵۷۱۷۳۸
۱۸۴	می‌کنند	۴۷۷۳۰	آب	۳۷۱	اصطلاحی	۱۰۳۱۹۴۷۰۰	فولیکول	۵۷۱۶۰۹
۱۸۵	میکند	۴۷۷۰۸	دوم	۳۶۷	تشخیص	۱۰۳۱۸۶۱۰۶	لاله‌ی	۵۷۱۵۸۶
۱۸۶	برابر	۴۷۶۷۲	حدود	۳۶۷	تخم	۱۰۳۱۷۶۰۵۴	مکانیسم	۵۷۱۵۷۰
۱۸۷	دهید	۴۷۶۱۵	هایی	۳۶۵	تصاویر	۱۰۳۱۷۳۹۳۴	نوکلئیک	۵۷۱۴۸۵
۱۸۸	تولید	۴۷۲۰۸	قیمت	۳۶۴	خرچنگ	۱۰۳۱۴۸۱۹۶	کشش	۵۷۱۰۸۹
۱۸۹	بیشتر	۴۶۸۹۴	روابط	۳۶۳	سیستمهای	۱۰۳۱۰۵۲۹۲	گامتوفیت	۵۷۰۹۵۷
۱۹۰	برگ	۴۶۷۹۱	هفته	۳۶۳	شبنم	۱۰۳۰۸۹۷۷۳	میکروسکوپ	۵۷۰۷۷۹
۱۹۱	نیروی	۴۶۷۴۲	رشد	۳۶۰	مناسبتر	۱۰۳۰۷۰۵۴۴	آمونیاک	۵۷۰۲۳۶
۱۹۲	استخوان	۴۶۷۰۶	روزنامه	۳۵۸	میدانیم	۱۰۲۹۸۴۵۹۲	آنتیکدون	۵۷۰۰۹۴
۱۹۳	زیر	۴۵۸۸۵	قابل	۳۵۸	می یابد	۱۰۲۹۵۱۹۴۹	الکلی	۵۶۹۵۶۰
۱۹۴	جذب	۴۵۷۶۶	شرایط	۳۵۷	هوازی	۱۰۲۸۹۳۱۸۲	اندازهٔ	۵۶۹۴۹۷
۱۹۵	شد	۴۵۷۵۲	دانشگاه	۳۵۶	پا	۱۰۲۳۹۲۱۰۱	اوگلنا	۵۶۸۵۳۹
۱۹۶	بدون	۴۵۲۹۶	همان	۳۵۵	مناسبتر	۱۰۲۳۶۶۲۸۰	بیماریهای	۵۶۵۸۴۷
۱۹۷	محیط	۴۴۵۳۷	کل	۳۵۲	یازدهم	۱۰۲۲۴۹۳۸۰	تتراپلوئید	۵۶۵۷۵۰
۱۹۸	ایجاد	۴۴۴۸۲	فعالیت	۳۴۹	ریاضیات	۱۰۲۱۳۴۴۶۹	کفش	۵۶۳۲۹۳
۱۹۹	زمین	۴۴۴۰۹	حاضر	۳۴۹	تصوّر	۱۰۲۰۲۸۸۲۲	کشاورزان	۵۶۲۸۶۹
۲۰۰	آزاد	۴۴۳۶۸	گرفت	۳۴۸	خورشید،	۱۰۱۶۹۹۳۷۴	گره‌های	۵۵۹۴۱۸

زهره ذوالفقار کندری

متولد ۱۳۵۳، دارای مدرک تحصیلی دکتری زبان‌شناسی همگانی از دانشگاه پیام نور تهران است. ایشان هم‌اکنون مشغول به تدریس زبان فارسی و انگلیسی در کشور کاناداست.

پژوهش‌های پیکره‌بنیاد در زبان‌شناسی از جمله علایق پژوهشی وی است.

طیبه موسوی میانگاه

متولد سال ۱۳۵۱ دارای مدرک دکتری در زبان‌شناسی رایانشی از آکادمی علوم روسیه در مسکو است. ایشان هم‌اکنون دانشیار گروه زبان‌شناسی دانشگاه پیام نور مرکز یزد است و به‌عنوان محقق مدعو در دانشگاه دولتی کنت در ایالت اوهایوی آمریکا مشغول به فعالیت است.

پردازش زبان طبیعی، ترجمهٔ ماشینی، بازیابی اطلاعات دوزبانه، زبان‌شناسی پیکره‌ای و نیز آموزش زبان فارسی به غیرفارسی‌زبانان در محیط‌های دیجیتال از جمله علایق پژوهشی وی است.

بلقیس روشن

متولد ۱۳۳۴ دارای دکتری در رشتهٔ زبان‌شناسی همگانی است. ایشان هم‌اکنون دانشیار گروه زبان‌شناسی دانشگاه پیام نور است.

معناشناسی، عصب‌شناسی و روان‌شناسی زبان از جمله علایق پژوهشی وی است.

امیررضا وکیلی‌فرد

متولد سال ۱۳۴۷ دارای مدرک تحصیلی دکتری در رشتهٔ آموزشکاوی زبان‌های دوم از دانشگاه مونترال کاناداست. ایشان هم‌اکنون دانشیار گروه آموزش زبان فارسی به غیر فارسی‌زبانان دانشگاه بین‌المللی امام خمینی (ره) است.

روان‌شناسی یادگیری زبان خارجی، روش‌های آموزش زبان و ادبیات، آموزش زبان برای اهداف دانشگاهی و ارزشیابی آموزشی از جمله علایق پژوهشی وی است.