The Use of Synonyms in *Persian Subject Headings* and *Asfa Thesaurus* Based on *Farsnet* Lexical Tool

Farzaneh Shadanpour[1]

Abstract

Purpose: Synonymy is one of the important features of natural languages. Since a single concept may be expressed by two or more lexical forms, and it is not predictable which lexical form of a single concept will be searched for, the retrieval system must be able to refer from all synonyms of the same idea to the document in which the concept is discussed. This research aimed to investigate the use of synonyms in non-preferred headings/ terms in *Persian subject headings* and *Asfa Thesaurus*, using *Farsent*as a comprehensive lexical source of the Persian language.

Method: This was an applied research in terms of its goals, and used content analysis as a general methodology, specifically Natural Language Processing techniques and tools to measure the extent to which synonyms are used to build non-preferred headings/ terms in both controlled vocabulary, by measuring the similarity of the two groups of data. 3270 main subject headings and 2020 main thesaurus terms were selected, in a purposive sampling procedure, from *Persian Subject Headings*, and *Asfa Thesaurus*, as two controlled vocabulary used in the process of compiling the *Iran National Bibliography*. Non-preferred headings/ terms related to each main heading/ term, as well as synonyms of each, were also extracted from *Farsent*. Reliability was obtained by repeating the extraction of a part of the headings/ terms by a second researcher with a score of 0.618 and 0.706 between zero and 1 respectively. The similarity between the two data sets of non-preferred headings/terms with the synonyms of main headings/ terms related to them in *Farsnet* was measured using Cosine Similarity.

Findings: In the sample taken from *Persian subject headings*, 2561 main subject headings (78.3%) have non-preferred headings that refer to them. 2316 main subject headings (70.8%) also have synonyms in *Farsent*. The similarity score between non-preferred headings and synonyms of the corresponding main headings was 0.125, thus very low. Also, in the sample taken from *Asfa*, 545 main terms in *Asfa* (about 27%) have non-preferred terms. 1376 terms (68%) of these main terms also have synonyms in *Farsnet*. Thus, 1475 main terms (73%) do not have non-preferred terms (which refer to the main term). The similarity score between non-preferred terms in *the Asfa Thesaurus* and the synonyms of the corresponding main terms was 0.131, very low as well.

Conclusion: More commitment to the construction and use of subject references in the form of non-preferred headings is observable in *Persian Subject Headings*, but a small number of referential headings and terms (non-preferred) have been selected from among the synonyms of main subjects/terms in the *Persian language*. This research recommends the introduction of synonyms of terms for all users, including catalogers and those involved in the creation of controlled vocabularies, both during the search for concepts and in the creation of terms, because it can be a step towards improving subject authority databases and, ultimately, a more exhaustive user subject search and retrieval experience.

Keywords

Semantic Relations, Synonym, *Persian Subject Headings*, *Persian Cultural Thesaurus (Asfa)*, *Farsnet*, Similarity Measuring, Cosine Similarity

Citation: Shadanpour, F. (2025). The Use of Synonyms in *Persian Subject Headings and Asfa Thesaurus* Based on Farsnet Lexical Tool. *Librarianship and Information Organization Studies*, 35(4): 7-38.

Doi: 10.30484/NASTINFO.2024.3629.2288

Article Type: Research Article
Article history:
Received: 27 July 2024
Accepted: 20 Oct. 2024

1. Assistant professor, Knowledge and Information Science, Research Department of Data Science, Information and Artificial Intelligence, National Library and Archives Islamic Republic of Iran, Tehran, Iran
f-shadanpour@nlai.ir

Publisher: National Library and Archives of I.R. of Iran

کاربرد مترادف‌ها در *سرعنوان‌های موضوعی فارسی* و *اصطلاح‌نامهٔ اصفا* بر مبنای ابزار واژگانی فارس‌نت

فرزانهٔ شادانپور[1]

چکیده

هدف: ترادف یا هم‌معنایی از ویژگی‌های مهم زبان‌های طبیعی است. ازآنجاکه یک مفهوم واحد ممکن است با دو یا چند شکل واژگانی بیان شود و معلوم نیست کدام شکل واژگانی بازگوکنندهٔ یک مفهوم واحد در سامانهٔ بازیابی مورد جستجو قرار خواهد گرفت، سامانه باید بتواند از همه مترادف‌های یک مفهوم به مدرکی که مفهوم در آن موردبحث قرارگرفته ارجاع دهد. این پژوهش با هدف بررسی وضعیت به‌کارگیری مترادف های سرعنوان‌های گزیده/ اصطلاحات مرجح را در ساخت سرعنوان‌های ناگزیده/ اصطلاحات نامرجح در *سرعنوان‌های موضوعی فارسی* و *اصطلاح‌نامهٔ فرهنگی فارسی «اصفا»*، در تطبیق با *فارس‌نت*، به‌عنوان منبع واژگانی جامع زبان فارسی، انجام شد.

روش: پژوهش از حیث هدف کاربردی و از جنبهٔ روش‌شناسی، تحلیل محتوا بود و از فنون متن‌کاوی و پردازش زبان طبیعی برای سنجش میزان کاربرد مترادف‌ها در هر دو واژگان کنترل‌شده با سنجش شباهت دو گروه داده استفاده کرده است. ۳۲۷۰ سرعنوان موضوعی و ۲۰۲۰ اصطلاح اصلی به‌صورت هدفمند از دو منبع *سرعنوان‌های موضوعی فارسی* و *اصطلاح‌نامهٔ اصفا*، به‌عنوان دو مجموعه واژگان کنترل‌شده مورداستفاده در تدوین *کتابشناسی ملی ایران*، انتخاب شد. سرعنوان‌های ناگزیده، اصطلاحات نامرجح مربوط به هر سرعنوان/ اصطلاح اصلی و مترادف‌های هر یک از *فارس‌نت* نیز استخراج شد. پایایی با تکرار استخراج بخشی از سرعنوان‌ها/اصطلاحات توسط پژوهشگر دوم با شباهت ۰/۶۱۸ و ۰/۷۰۶ از بازهٔ میان صفر تا ۱ به ترتیب برای سرعنوان‌ها و اصطلاحات به دست آمد. با استفاده از زبان برنامه‌نویسی پایتون شباهت میان هریک از دو دستهٔ داده سرعنوان‌های ناگزیده و اصطلاحات نامرجح با مترادف‌های سرعنوان‌ها/ اصطلاحات اصلی مربوط به آن‌ها در *فارس‌نت* با سنجهٔ کسینوس شباهت اندازه‌گیری شد.

یافته‌ها: در نمونهٔ گرفته‌شده از *سرعنوان‌های موضوعی فارسی*، ۲۵۶۱ سرعنوان اصلی (۷۸/۳ درصد) دارای سرعنوان ناگزیده بوده‌اند که به سرعنوان گزیده ارجاع می‌دهد. ۲۳۱۶ سرعنوان اصلی (۷۰/۸ درصد) نیز دارای مترادف در *فارس‌نت* بوده‌اند. نمرهٔ شباهت میان سرعنوان‌های ناگزیده و مترادف‌های سرعنوان اصلی مربوط به هر یک ۰/۱۲۵ به دست آمد که نشان از شباهت پایین آن‌هاست. همچنین در نمونهٔ گرفته‌شده از *اصطلاح‌نامهٔ اصفا*، ۵۴۵ اصطلاح (حدود ۲۷ درصد) دارای اصطلاح ارجاعی نامرجح بوده‌اند. ۱۳۷۶ اصطلاح (۶۸ درصد) از این اصطلاحات نیز دارای مترادف در *فارس‌نت* هستند؛ یعنی تعداد ۱۴۷۵ اصطلاح (۷۳ درصد) فاقد اصطلاح نامرجح (که ارجاع به اصطلاح اصلی می‌دهند) بوده‌اند. نمرهٔ شباهت میان اصطلاحات نامرجح در *اصطلاح‌نامهٔ اصفا* و مترادف‌های اصطلاح اصلی مربوط به هر یک ۰/۱۳۱ به دست آمد که نمرهٔ پایینی است.

نتیجه‌گیری: در *سرعنوان‌های موضوعی فارسی* تقید بیشتری در ساخت و به‌کارگیری ارجاعات موضوعی دیده می‌شود، ولی در هر دو واژگان کنترل‌شده تعداد کمی از سرعنوان‌ها و اصطلاحات ارجاعی (ناگزیده و نامرجح) از میان مترادف‌های مفاهیم در زبان فارسی انتخاب‌شده‌اند. این پژوهش معرفی مترادف‌های عبارت‌ها را برای همه کاربران، ازجمله فهرست‌نویسان و متصدیان ساخت مستندات موضوعی، چه هنگام جستجوی مفاهیم و چه در ساخت اصطلاحات توصیه می‌کند، چراکه می‌تواند به بهبود وضعیت بانک‌های مستند موضوعی و درنهایت تجربه متکامل‌تر کاربر در جستجوی موضوعی و بازیابی منابع کمک کند.

کلیدواژه‌ها

روابط معنایی، مترادف، سرعنوان‌های موضوعی فارسی، *اصطلاح‌نامهٔ فرهنگی فارسی (اصفا)*، فارسنت، شباهت‌سنجی، کسینوس شباهت

۱. استادیار، علم اطلاعات و دانش‌شناسی، گروه پژوهشی علوم داده، اطلاعات و هوش مصنوعی، سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران، تهران، ایران

f-shadanpour@nlai.ir

* این مقاله برگرفته از طرح پژوهشی موظف خاتمه‌یافته در سازمان اسناد و کتابخانه ملی ایران با عنوان «مترادف‌ها در سرعنوان‌های موضوعی فارسی، اصطلاح‌نامهٔ اصفا، و فارس‌نت: مطالعه تطبیقی» است که در تاریخ ۱۴۰۲/۰۳/۲۱ مصوب شده است.

فصلنامهٔ مطالعات کتابداری و سازماندهی اطلاعات، ۳۵ (٤)، زمستان ۱٤۰۳

نوع مقاله: پژوهشی

تاریخ دریافت: ۱٤۰۳/۰۵/۰٦

تاریخ پذیرش: ۱٤۰۳/۰۷/۲۹

ناشر: سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

استناد: شادانپور، فرزانه (۱٤۰۳). کاربرد مترادف‌ها در *سرعنوان‌های موضوعی فارسی و اصطلاح‌نامهٔ اصفا* بر مبنای ابزار واژگانی فارس‌نت. *مطالعات کتابداری و سازماندهی اطلاعات*، ۳۵(٤): ۷-۳۸.

Doi: 10.30484/NASTINFO.2024.3629.2288

مقدمه

اطلاعات که بخشی مهمی از آن در قالب کلام انسان پدید می‌آید، ماده مشترکی است که حوزه‌های مختلفی در علم به آن می‌پردازند. محتواهای اطلاعاتی عمدتاً در قالب زبان‌های طبیعی انسان‌ها پدید می‌آیند و یا دست‌کم به‌نوعی با آن ربط پیدا می‌کنند. زبان در اصل مأموریت انتقال معنا را بر عهده دارد، اما ویژگی‌های ذاتی عناصر تشکیل‌دهندهٔ آن گاهی فرایند انتقال معنا را تحت تأثیر قرار می‌دهد: ابهام، ترادف (هم‌معنایی) و چندمعنایی برخی از این ویژگی‌ها هستند. این ویژگی‌های زبانی نه‌تنها بر محتوای معنایی منابع اطلاعاتی، بلکه بر جنبه‌های مختلف سازماندهی محتوا تأثیرگذارند. در حوزهٔ سازماندهی اطلاعات، مهم‌ترین ابزارهای سازماندهی موضوعی، سرعنوان‌های موضوعی و اصطلاح‌نامه‌ها هستند. مهم‌ترین هدف استفاده از زبان‌های کنترل‌شده در بازنمایی موضوعات منابعی که پردازش می‌شوند، حفظ یکدستی در کاربرد اصطلاحات و پرهیز از تشتت و پراکندگی است (سلطانی، ۱۳۸۵). ویژگی‌های مهم زبان طبیعی انسان‌ها که در هر دو نوع ابزار گفته شده در بالا نقش مهمی دارد، ویژگی ترادف یا هم‌معنایی است. مترادف‌ها علاوه بر اهمیتی که در حوزه‌هایی مانند زبان‌شناسی رایانشی، بازیابی اطلاعات – ازجمله سازماندهی اطلاعات (فهرست‌نویسی و نمایه‌سازی)، نظام‌های پرسش و پاسخ و ترجمه ماشینی دارند، بخش مهمی از انواع منابع واژگانی مانند وردنت‌ها و اصطلاح‌نامه‌ها محسوب می‌شوند (Miller et al., 1990). همچنین در تعیین سرعنوان‌های موضوعی مترادف‌ها در نظر گرفته می‌شوند. یک مفهوم واحد ممکن است با دو شکل واژگانی بیان شود و معلوم نیست کاربر احتمال دارد کدام شکل واژگانی را در جستجو به کار ببرد؛ بنابراین، برای پرهیز از ریزش منابع در بازیابی به علت مترادف‌ها لازم

است «تمامی ارجاع‌های لازم که عمدتاً مترادف‌ها و متشابه‌ها را در برمی‌گیرد، ذیل هر موضوع ذکر شود و برای ایجاد حداکثر قابلیت انعطاف، موضوع‌ها و ارجاع‌ها روی برگه‌های مخصوص، ثبت و در محل الفبایی خود برگه‌آرایی شود تا بتوان تغییرات و اضافات لازم را به‌آسانی در آن ادغام کرد» (سلطانی، ۱۳۸۵). عبارتی که توسط نظام انتخاب شده «گزیده» و سایر مترادف‌ها را «ناگزیده» می‌نامند که از آن‌ها به سرعنوان موضوعی گزیده ارجاع داده می‌شود و با عبارت «به‌جای» قابل‌شناسایی هستند (سلطانی و همکاران، ۱۳۹۷). همین مسئله در اصطلاح‌نامه‌ها مطرح است و از مهم‌ترین مراحل در انتخاب اصطلاحات، مهار مترادف‌هاست که با استفاده از ارجاع «به کار ببرید» اصطلاحات مرجح و نامرجح در نظام مشخص می‌شوند (*معرفی و تاریخچه اصفا*، بی‌تا). استفاده از مترادف‌ها در نظام بازیابی اطلاعات منجر به بهبود کارایی بازیابی می‌شود (Kellbessa, 2021). در شرایط ایدئال بازیابی اطلاعات – چه در آن از زبان کنترل‌شده برای ساختن نمایه استفاده شود، چه از زبان طبیعی – باید بتواند از همه مترادف‌های یک مفهوم به مدرکی که مفهوم در آن موردبحث قرارگرفته ارجاع دهد. در زبان‌های کنترل‌شده، مانند سرعنوان‌های موضوعی فارسی و اصطلاح‌نامهٔ اصفا که به‌صورت دستی تهیه‌شده‌اند، نیروی انسانی متصدی تهیه ارجاعات از مترادف‌هاست. قدرمسلم، تکمیل مترادف‌ها بر اساس یک منبع واژگانی مرجع که از حیث شمول معانی و مفاهیم زبان قابل‌اتکا باشد می‌تواند به کارآمدی بالاتر در سازمان‌دهی اطلاعات و بازیابی کمک کند. مقدمه این کار، بررسی وضعیت استفاده از مترادف‌ها در این دو زبانِ کنترل‌شده است که این پژوهش درصدد آن است. به تعبیر دیگر، این پژوهش در حوزهٔ موضوعی تطبیق و مقایسهٔ ابزارهای واژگانی و معنایی مورداستفاده در پردازش، ذخیره و بازیابی اطلاعات قرار می‌گیرد که موردعلاقه و وجه مشترک پژوهش‌های علوم اطلاع‌رسانی و پردازش زبان طبیعی است.

پیشینهٔ پژوهش

موضوع پژوهش با حوزه‌های اصطلاح‌نامه‌ها، سرعنوان‌های موضوعی، ابزارهای واژگانی مانند وردنت‌ها مرتبط است که منابع بسیاری در این حوزه‌ها وجود دارد، ولی درباره مسئله خاص پژوهش، یعنی وضعیت مترادف‌ها در دو منبع مشخص: *سرعنوان‌های موضوعی فارسی و*

اصطلاح‌نامهٔ فرهنگی فارسی «اصفا»[1] بر مبنای *فارس‌نت* تاکنون پژوهش مستقلی صورت نگرفته است. لازم به ذکر است که ترادف جایگاه خاصی در پژوهش‌های زبان‌شناسی و معناشناسی دارد که هرچند مهم، در پیشینه‌های این پژوهش ذکر نمی‌شوند. همچنین پژوهش‌های بسیاری در داخل و خارج از کشور به مسائل مربوط به هم‌ترازی و مقایسهٔ وردنت‌ها در زبان‌های مختلف پرداخته‌اند[2] که به علت عدم ربط با موضوع اصلی این پژوهش از پرداختن به آن‌ها به‌عنوان پیشینه نیز خودداری شد. برخی از پژوهش‌های مرتبط و تا حدودی مرتبط در داخل و خارج از کشور را می‌توان ذکر کرد. به‌عنوان نمونه، شمس فرد[3] و همکاران (۲۰۰۹) در پژوهشی به تشریح فرایند توسعه *فارس‌نت* پرداخته‌اند که هستی‌نگاری واژگانی برای زبان فارسی است. در مرحلهٔ اول رویکرد نیمه‌خودکار برای ایجاد *فارس‌نت* مورداستفاده قرارگرفته است. *فارس‌نت* همچنین دارای روابط بین‌زبانی است که مجموعه مترادف‌های فارسی را به انگلیسی در *وردنت پرینستون*[4] نسخه ۳/۰ متصل می‌کند. *فارس‌نت* بر اساس اصول *وردنت پرینستون*، *یوروردنت*[5] و *بالکان نت*[6] ساخته‌شده و با آن‌ها سازگاری دارد و می‌تواند به وردنت‌های دیگر دنیا متصل شود تا عملکردهای پردازش زبان و توسعه لغت‌نامه‌ها و اصطلاح‌نامه‌های چندزبانه را تسهیل کند. برنجیان و رئیسی (۱۳۹۳) باهدف ارائهٔ واژه‌های مصوب فرهنگستان زبان و ادب فارسی در زمان بازیابی معادل رایج آن‌ها در سامانه بازیابی اطلاعات مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری و همچنین ارائهٔ معادل‌های رایج واژه‌ها در زمان جستجوی واژه‌های مصوب، سامانه‌ای را طراحی کردند که در آن فهرست واژه‌های مصوب فرهنگستان زبان و ادب فارسی در پایگاه اطلاعاتی مقاله‌های الکترونیکی

---

[1]. این دو منبع، ابزار سازمان‌دهی موضوعی در سازمان اسناد و کتابخانه ملی ج. ا. ا. هستند و به علت اهمیت جایگاهی که در سازمان‌دهی موضوعی و پردازش کتابشناسی ملی و منابع موجود در این سازمان دارند، موضوع این پژوهش قرار گرفته‌اند.

[2]. مانند حسابی (۱۳۹٤): مشکلات انطباق دسته‌های هم‌معنای فارس‌نت با دسته‌های هم‌معنای وردنت پرینستون؛ حسابی (۱۳۹٤): واژه‌های ویژهٔ زبان فارسی: خلأهای موجود در انطباق واژگانی *فارس‌نت* با شبکهٔ واژگانی پرینستون (*وردنت*).

[3]. Shamsfard et al.

[4]. Princeton WordNet

[5]. EuroWordNet

[6]. BalkaNet

سامانه رایسست[1]مرکز منطقه‌ای علوم و فناوری اطلاعات شیراز نصب و مورد آزمون‌های لازم قرار گرفت. در این برنامه، جستجو از هر دو طریق ممکن است. هم با واژه‌های مصوب فرهنگستان می‌توان واژه مصطلح را بازیابی کرد و هم با جستجوی واژه مصطلح واژه مصوب فرهنگستان بازیابی می‌شود. حسابی (۱۳۹۵) روابط معنایی درون‌زبانی اسم‌ها را در سه شبکه واژگانی *فارس‌نت*، *یوروورودنت* و *وردنت پرینستون* مقایسه کرده است. نتایج نشان داد یوروورودنت بیشترین تنوع رابطه و *وردنت پرینستون* کمترین تنوع رابطه را دارد و *فارس‌نت* میان این دو قرار دارد. فرناندز لانزا و همکاران[2](۲۰۰۳) در نظام بازیابی اطلاعات برای افزایش ربط مدارک بازیابی شده از اصطلاح‌نامه استفاده کردند. همچنین یک فرهنگ لغت الکترونیکی اسپانیایی از مترادف‌ها که درجه‌های ترادف را محاسبه می‌کند، به نظام افزودند استفاده از این فرهنگ لغت الکترونیکی در جستجوگر وب اسپانیایی باعث افزایش جامعیت بدون کاهش مانعیت و بدون صرف بیش از حد زمان شد. بهاراتی و ونکاتسان[3](۲۰۱۲) روش نمایه‌سازی مبتنی بر مترادف معنایی جدیدی ارائه کردند که در آن اسناد بر اساس نزدیک‌ترین همسایگان از مجموعه اسناد خوشه‌بندی و سپس با ارتباط معنایی عبارت پرس‌وجو با اسناد بازیابی شده با استفاده از یک اصطلاح‌نامه یا مدل هستی‌شناسی برای بهبود عملکرد نظام بازیابی اطلاعات و با افزایش تعداد اسناد مربوطه بازیابی شده پالایش می‌شوند. نتایج نشان می‌دهد که روش پیشنهادی نسبت به روش‌های موجود پیشرفت قابل‌توجهی دارد و می‌تواند سند مرتبط‌تری را در رتبه برتر بازیابی کند. لی و همکاران[4](۲۰۱۳) با نظر به این واقعیت که در میان تمام پرس‌وجوها در وب، زیرمجموعه مهمی از پرس‌وجوها وجود دارد که حاوی ویژگی‌هایی از موجودیت‌های مورد جستجوست که کاربران به آن علاقه‌مند هستند، مانند «سنِ اوباما» – با مترادف‌هایی مانند «اوباما چند سال دارد؟» یا «اوباما، سن و سال»– و با این فرض که شناسایی مترادف‌های این ویژگی‌ها می‌تواند عملکرد تمام رویکردهای حاشیه‌نویسی پرس‌وجو را بهبود بخشد و برای برنامه‌هایی مانند پاسخ‌های فوری و پیشنهاد پرس‌وجو مبتنی بر هدف نیز مفید باشد، یک چهارچوب خوشه‌بندی با چندین توابع هسته برای شناسایی

[1]. Regional Information Center for Science and Technology (RICeST)
[2]. Fernandez Lanza et al.
[3]. Bharathi & Venkatesan
[4]. Li et al.

الگوهای مربوط به مترادف‌های هدف از پرس‌وجو پیشنهاد کردند. آن‌ها از یک نظام وزن‌دهی بدون نظارت هم استفاده کردند. نتایج پژوهش کارآمدی چهارچوب خوشه‌بندی پیشنهادی را در یافتن الگوهای مترادف برای عبارت پرس‌وجو نشان داد. زنگ و همکاران[1](۲۰۱۲) سه روش را در بسط پرس‌وجو در بازیابی اسناد بالینی به آزمون گذاشتند که عبارت بودند از ۱) بسط مترادف‌ها، ۲) کاربرد مدل‌سازی موضوعی برای شناسایی اصطلاحات مرتبط برای بسط پرس‌وجو، ۳) کاربرد اصطلاحات مرتبط از یک پایگاه داده برگرفته از چکیده‌ها. آزمایش هر سه روش بر مجموعه‌ای از یادداشت‌های بالینی هرچند در مقایسه با روش معمول موفقیت‌آمیز بود، اما در بین سه روش بسط، روش مبتنی بر مدل موضوعی ازنظر فراخوانی و سنجه f بهترین عملکرد را داشت. سوتو و همکاران[2](۲۰۰۸) روشی برای گسترش پرس‌وجوهای کاربر ارائه کردند که در آن برای هر عبارت در پرس‌وجوی اصلی، تمام مترادف‌های آن با معنای خاصی با حداکثر فراوانی مفهومی معرفی می‌شوند. برای اندازه‌گیری میزان حضور مفهوم در یک سند (حتی در یک مجموعه سند)، فرمول بسامد مفهومی معرفی‌شده است. فرمول‌های فازی جدید نیز برای محاسبه درجه مترادف بین اصطلاحات برای مدیریت با مفاهیم (معانی) معرفی‌شده‌اند که کاربرد آن‌ها، حتی برای اصطلاحی که در یک سند ظاهر نمی‌شود، درجاتی از حضور آن را بر اساس میزان ترادف آن با اصطلاحاتی که در سند آمده است، تخمین می‌زند. بازیل[3](۲۰۱۵) باتوجه‌به ساختار باز وردنت که شبکه‌ای معنایی مبتنی بر مجموعه‌های مترادف و روابط میان آن‌هاست، معتقد است که وردنت را می‌توان به‌عنوان جایگزینی برای هستی‌شناسی‌ها در فرایندهای تولید زبان طبیعی مانند تولید چکیده، عبارت های کلیدی و کلیدواژه‌ها و یا در ترکیب با سایر پایگاه‌های دانش به کار گرفت. عبدالحسن و هادی[4](۲۰۱۷) از منطق فازی برای بهبود دقت ابهام‌زدایی معنایی از طریق تنظیم وزن‌های مترادف و از هوش ازدحامی[5] به‌ویژه رویکرد کلونی زنبورهای مصنوعی[6] برای حل مشکل

[1]. Zeng et al.
[2]. Soto et al.
[3]. Basile
[4]. Abdul Hassan1 & Hadi
[5]. Swarm Intelligence
[6]. Artificial Bee Colony (ABC)

میزان دقت ابهام‌زدایی معنایی و سرعت اجرای الگوریتم در مجموعه اسناد بزرگ‌مقیاس استفاده کردند. نتایج تجربی برتری مدل پیشنهادی را ازنظر جامعیت، مانعیت و سرعت اجرای الگوریتم[1]نسبت به مدل سنتی نشان داد. لی و همکاران[2](۲۰۱۹) با توجه به اینکه بسط مترادف، فنی است که کلمات مرتبط را به جستجو اضافه می‌کند و جامعیت را بهبود می‌بخشد، به کاربرد بسط مترادف برای جستجوی رایانامه پرداخته‌اند که برای شرکت‌های خصوصی مانند ایمیل، چندین چالش پژوهشی منحصربه‌فرد را به همراه دارد. کلبسا[3](۲۰۲۱) گزارش کرده که با افزودن مترادف‌ها به اصطلاح‌نامه مورداستفاده در یک نظام بازیابی اطلاعات، عملکرد بازیابی از حیث جامعیت بالاتر بوده ولی از حیث مانعیت کاهش داشته، اما عملکرد کلی سیستم یا سنجه f1 را تا حدودی بهبود بخشیده است.

از مجموع پیشینه‌ها ملاحظه می‌شود که گزارش برنجیان و رئیسی و پژوهش کلبسا را می‌توان مرتبط‌ترین با این پژوهش دانست. درباره وضعیت ابزارهای سازمان‌دهی معنایی از حیث میزان کاربرد مترادف‌ها – به‌ویژه درباره زبان‌های کنترل‌شده مورداستفاده در تدوین *کتابشناسی ملی ایران*– پژوهش مشابهی چه در خارج و چه در داخل کشور یافت نشد که باتوجه‌به نقش این دو ابزار از دیدگاه مرجعیت سازمان، ضروری است این جنبه موردبررسی قرار گیرد. باتوجه‌به اینکه پیشینهٔ خاصی که به‌طور مشخص به وضعیت مترادف‌ها در این دو ابزار معنایی پرداخته باشد، یافت نشد، این پژوهش می‌کوشد تا وضعیت استفاده از مترادف‌ها را در *سرعنوان‌های موضوعی فارسی* و *اصطلاح‌نامهٔ اصفا*، به‌عنوان دو ابزار سازمان‌دهی معنایی که مورداستفاده بسیاری از کتابخانه‌ها و مراکز اطلاع‌رسانی است، در تطبیق با *فارسنت*، به‌عنوان منبع واژگانی جامع زبان فارسی، بررسی و به این سه پرسش پاسخ دهد: ۱) وضعیت کاربرد مترادف‌ها در *سرعنوان‌های موضوعی فارسی* در تطبیق با *فارسنت* چگونه است؟ ۲) وضعیت کاربرد مترادف‌ها در *اصطلاح‌نامهٔ اصفا* در تطبیق با *فارسنت* چگونه است؟ ۳) وضعیت کاربرد مترادف‌ها در *سرعنوان‌های موضوعی فارسی* در مقایسه با *اصطلاح‌نامهٔ اصفا* چگونه است؟

---

[1]. Latency
[2]. Li et al.
[3]. Kellbessa

روش پژوهش

این پژوهش از حیث هدف کاربردی و از جنبه روش‌شناسی، تحلیل محتوا با استفاده از فنون متن‌کاوی و پردازش زبان طبیعی است. پیش از ورود به بحث جامعه پژوهش و ویژگی‌های آن، ذکر این نکات و پیش‌فرض‌ها ضروری است: نخست اینکه در این پژوهش، منظور از مترادف‌ها واژه‌هایی هستند که از حیث ظاهر باهم متفاوت‌اند ولی برای بیان مفهوم یکسان به کار می‌روند و در فارس‌نت در مجموعه‌های «هم‌معنا» قرار دارند. دوم، سرعنوان‌های موضوعی، اصطلاح‌نامه و وردنت‌ها هر یک از ابزارهایی هستند که کارکردهای خاص خود را دارند و از جهات بسیاری باهم متفاوت‌اند، ولی وجه مشترک آن‌ها این است که از مجموعه‌ای از مفاهیم تشکیل‌شده‌اند که در زبان انسان رمزگذاری شده‌اند و روابط معنایی را در یک زبان به نمایش می‌گذارند. سوم، چنین پیش‌فرض گرفته‌شده که سرعنوان‌های ناگزیده و اصطلاحات نامرجح صحیح هستند و به‌درستی برای هر سرعنوان و اصطلاح اصلی انتخاب شده‌اند، بنابراین صحت‌وسقم آن‌ها موردبررسی نبوده است. در این پژوهش سه دسته داده گردآوری و در فایل اکسل در سه ستون درج‌شده‌اند.

نخست، مفاهیمی هستند که در قالب سرعنوان موضوعی و اصطلاح در دو منبع *سرعنوان‌های موضوعی فارسی* (سلطانی و همکاران، ۱۳۹۷) و نسخه برخط *اصطلاح‌نامهٔ اصفا* گردآمده‌اند و با نمونه‌گیری هدفمند (قضاوتی) بر اساس معیارهایی که در ادامه ذکر می‌شود انتخاب‌شده‌اند. ازآنجاکه بانک در مستندات موضوعی *سرعنوان‌های موضوعی فارسی*، پیوسته در حال تغییر و تحول است، برای اجرای این پژوهش نیاز به مستندی بود که طی مدت اجرا دست‌خوش تغییرات نشود. علاوه بر این، در متن ویراست مورداستفاده از *سرعنوان‌های موضوعی فارسی* (۱۳۹۷، ویراست ٤، ص ۱۵ و ١٦) چنین ذکرشده:

«کاربران محترم توجه داشته باشند که در صورت تفاوت معادل و ارجاعات و یادداشت‌های رکوردهای مستند در شکل چاپی، *سرعنوان‌های موضوعی فارسی* در مقایسه با رکوردهای بانک مستند موضوعی در سیستم پیوسته، مرجع مورد اعتماد و استفاده همواره همین متن چاپی خواهد بود. تا به یاری خدا در آینده نزدیک کلیه سرعنوان‌ها و تقسیمات فرعی آن‌ها در بانک مستند موضوعی اصلاح شود و سیستم پیوسته به‌هیچ‌وجه با

متن چاپی تفاوت نداشته باشد»؛ بنابراین شکل چاپی مورداستفاده قرار گرفت.

دوم، سرعنوان‌های ناگزیده و اصطلاحات نامرجح مربوط به هر سرعنوان/ اصطلاح اصلی که هر یک در ستونی مجزا مقابل سرعنوان‌ها و اصطلاحات اصلی درج‌شده‌اند. این‌ها همان ارجاعات «به‌جای» برای هر سرعنوان و عبارت‌هایی هستند که انتظار می‌رود به کاربر در فرایند جستجو و بازیابی کمک کنند، به‌ویژه، هنگامی‌که ممکن است چند بیان برای یک مفهوم وجود داشته باشد، یا قصد این باشد که چندین مفهوم مرتبط با سرعنوان اصلی (که لزوماً با آن مترادف معنایی و لغت‌نامه‌ای نیستند) ذیل یک سرعنوان جمع شوند؛ بنابراین سرعنوان‌های ناگزیده انواع مختلفی دارند، ولی درمجموع از مترادف‌ها و مفاهیم نزدیک یا مرتبط به مفهوم سرعنوان اصلی تشکیل‌شده‌اند.

سوم، مترادف‌های سرعنوان‌ها و اصطلاحات اصلی در فارس‌نت که در ستون سوم وارد شدند. *فارس‌نت* وردنت[1] فارسی است. وردنت‌ها پرکاربردترین منبع معناشناسی و ابزار بازنمایی معنا در یادگیری ماشین و پردازش زبان طبیعی و شبکه‌ای متشکل از واحدهای معنایی با روابط تعریف‌شده و به‌شدت ساختاریافته هستند که نه یک واژه‌نامه سنتی است و نه یک اصطلاح‌نامه؛ بلکه ویژگی‌هایی از هر دو نوع منبع واژگانی را در خود ترکیب کرده است (Fellbaum, 1998, p. 7). وردنت‌ها پایگاه‌های واژگانی الکترونیکی هستند که در آن‌ها اکثر انواع کلمات، شامل اسم، فعل، صفت و قید در مجموعه‌های مترادف[2] نگه‌داری می‌شوند که هرکدام یک مفهوم واژگانی‌شده را نشان می‌دهند و روابط معنایی گوناگونی مانند ترادف، تضاد، اعم و اخص، جزء و کل و... این مجموعه‌های مترادف را به هم پیوند می‌دهد (Miller, 1995). اولین وردنت، *وردنت پرینستون* بود که در حال حاضر در نسخه ۳ آن حدود ۱۱۰۰۰ مجموعه مترادف وجود دارد. وردنت‌ها به «استاندارد طلایی»[3] برای دانش واژگانی تبدیل‌شده‌اند که معانی و مفاهیم کلمات یک یا چند زبان را بازنمایی می‌کنند و منبع دانش مهمی برای عملکردهای پردازش زبان طبیعی

---

[1]. WordNet

[2]. Synset که شکل کوتاه شده برای عبارت Synonym set است.

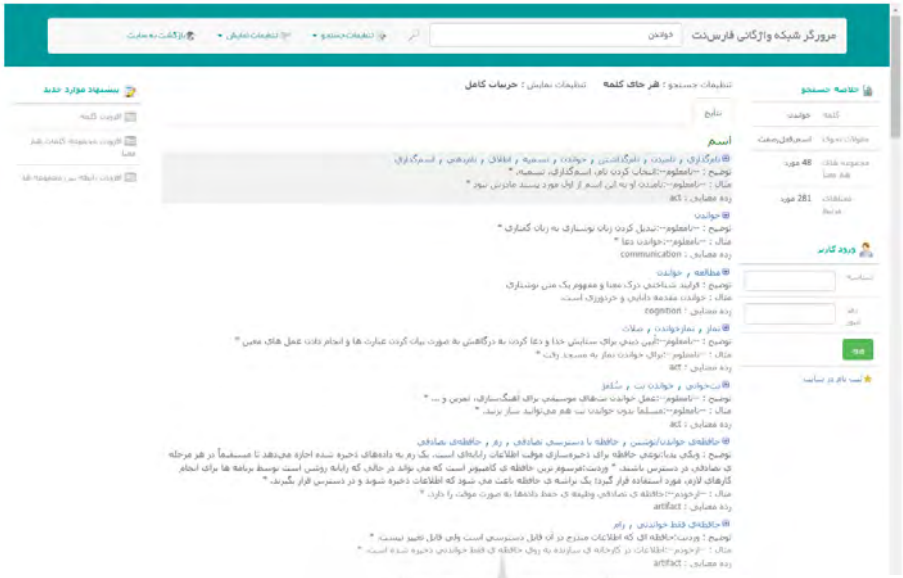
[3]. Golden Standard: استانداردی که موردپذیرش عموم است و از آن استفاده می‌شود و می‌تواند مرجع مقایسه و ارزیابی قرار گیرد.

به شمار می‌روند (Li et al., 2022). جایگاه و کارکرد وردنت‌ها را در پردازش‌های زبانی و متن‌کاوی، می‌توان هم‌سنگ اصطلاح‌نامه‌هایی دانست که به قالب هستی‌شناسی درآمده‌اند، ولی باوجود شباهت‌ها، تفاوت‌های مهمی نیز با آن‌ها دارند.

در ایران، وردنت فارسی با عنوان *فارس‌نت*[1] در سال ۱۳۸۷ در آزمایشگاه پردازش زبان طبیعی دانشگاه شهید بهشتی و با حمایت مرکز تحقیقات مخابرات ایران توسعه‌یافته است. آخرین نسخه *فارس‌نت* (نسخه ۳/۰) دارای بیش از صد هزار مدخل واژگانی (واژه یا عبارت) است که در حدود چهل هزار مجموعه مترادف جا گرفته‌اند. برای هر مدخل حداقل یک معنی تعریف‌شده و هر معنی در یک و فقط یک مجموعه مترادف شرکت می‌کند. کلیه مجموعه‌های مترادف یا در سلسله‌مراتب شرکت می‌کنند و یا به‌عنوان سرگروه معرفی می‌شوند. در ضمن هر مجموعه مترادف، یا حداقل یکی از اعضا آن، در حداقل یک رابطه غیر سلسله‌مراتبی شرکت نموده است. هر مجموعه مترادف در صورت امکان به مجموعه مترادف نظیر در *وردنت پرینستون* - نسخه ۳/۰- نگاشت شده است (*فارس‌نت*، بی‌تا).

علت انتخاب *فارس‌نت* این بود که مبنای آن ترادف بوده و ساختاری شبیه اصطلاح‌نامه دارد و تقریباً همه معانی موجود در لغت‌نامه‌های موجود را گردآوری کرده است و به‌صورت برخط در اختیار همگان قرار دارد (شکل ۱).

[1]. FarsNet



شکل ۱- نمایی از نتایج جستجو در *فارسنت*

سنجش اصلی این پژوهش مقایسهٔ شباهت میان سرعنوان‌های ناگزیده و اصطلاحات نامرجع با مترادف‌های موجود در *فارسنت* برای هر سرعنوان/ اصطلاح اصلی است. شکل ۲ برشی از فایل اکسل به‌دست‌آمده را نشان می‌دهد.

گزیده در س ع	ناگزیده در س ع	مترادفهای س ع در ف ن
آب	آب هاي طبيعي	پهنه آب آب
آبازورها	حبابهای چراغ نورافشانها	نورتاب
آبای کلیسا	پدران کلیسا پیشوایان آغازین مسیحیت پیشوایان آغازین مسیحیت، فلسفه	نداشت

شکل ۲- برشی از فایل اکسل داده‌ها

به پیروی از نوع پژوهش، به‌منظور اینکه هم جامعه محدود شود و هم مفاهیم بیشتری در پوشش پژوهش قرار گیرد، از سرعنوان‌ها و اصطلاحات به‌صورت هدفمند (قضاوتی) و با لحاظ پاره‌ای ویژگی‌ها نمونه‌هایی انتخاب شد که در ادامه ویژگی‌های آن‌ها ذکر می‌شود:

از سرعنوان‌های موضوعی:

- سرعنوان‌های اصلی که در کتاب *سرعنوان‌های موضوعی فارسی* با قلم بزرگ و تیره همراه با معادل انگلیسی آن مشخص‌شده‌اند، در جامعه وارد شدند.
- سرعنوان‌های مقلوب (مانند اعداد، مفهوم) و دارای تقسیمات فرعی در جامعه وارد نشده‌اند.
- به علت رعایت حداکثری یکدستی در کل پژوهش از حیث موضوعات و مفاهیم، سرعنوان‌های حوزهٔ پزشکی (ازجمله دامپزشکی و داروسازی)، به‌جز اصطلاحات بسیار رایج که به حیطه زبان عمومی تعلق دارند و واژگان خاص پزشکی نیستند (مانند «آفت» یا «آکنه» یا «آلزایمر»)، از جامعه خارج شدند.
- بعضی سرعنوان‌ها یا اصطلاحات لزوماً به همان شکل در لغت‌نامه‌ها وجود ندارند و مفاهیمی هستند که زایش اندیشه در زبان به وجود می‌آورد که پشتوانه انتشارات نیز دارد، مانند سرعنوان «اسرار تجاری». سرعنوان‌های اصلی از حیث شکل ظاهر ممکن است تک‌واژه‌ای یا عبارتی حاوی چند واژه باشند. در این موارد، با دو هدف محدودکردن تعداد جامعه پژوهش و درعین‌حال حفظ کلمات حاوی مفاهیم اصلی در آن، واژه اصلی معنایی در جامعه پژوهش واردشده است. به‌عنوان‌مثال از عبارت‌های ترکیب‌شده از کلمه «آب» و سایر کلمات، مانند «آب در هنر»، «آب دریا»، «آب‌رسانی شهری» و غیره، فقط کلمه «آب» در جامعه واردشده است.
- بعضی از کلمات فقط به‌صورت ترکیب با سایر کلمات در *سرعنوان‌های موضوعی فارسی* وجود دارند؛ ولی در لغت‌نامه‌ها شکل منفرد و غیرترکیبی آن‌ها نیز وجود دارد که فقط یکی از ترکیب‌ها در جامعه وارد شد، مثل «رسانه»، «رژیم»، «رزم»، «زیبایی»، «روابط»، «زیست» و بسیاری دیگر.
- اغلب کلمات چه به‌صورت اسم، چه فعل و چه صفت معنای واحدی را می‌رسانند، مثل «امانت‌داری» و «امانت‌داری کردن»، «انضباط» و «منضبط بودن» و ازاین‌قبیل که در این صورت مترادف‌های ممکن یک کلمه در هر نقشی در جامعه واردشده است.
- بعضی از سرعنوان‌ها ترکیبی از چند کلمه بودند، ولی کلمات تشکیل‌دهنده آن یا خود مترادف بودند یا نزدیک به معنای آن‌که در جامعه وارد شدند، مانند «اندیشه و

تفکر»، «بخت و اقبال»، «بزهکاران و مجرمان، تبعید و تبعیدیان»، «تـزیین و آرایـش»، «خلق‌وخو»، «نظم و ترتیب»، «صمغ و رزین» و ازاین‌قبیل. در صورت وجود تفـاوت اندک در معنا، مترادف‌های هر دو یا چند کلمه در ستون مترادف‌ها واردشده است.

- درصورتی‌که حوزهٔ معنایی سرعنوان مشخص‌شده باشد، مترادفِ همان معنا در جامعه واردشده و درصورتی‌که برای سرعنوان اصلی حوزهٔ معنایی خاصی ذکر نشده باشـد، همه مترادف‌های ممکن در جامعه واردشده‌اند.
- در مواردی که چند تـک‌واژه در یـک ترکیـب دال بـر یـک مفهـوم واحـد بوده‌انـد، حتی‌الامکـان در جامعـه واردشـده‌اند، ماننـد «آب‌زمین‌شناسـی»، «آب‌مروایـد»، «رأی‌گیری»، «رنگین‌کمان».
- کلمات همنام[1](که به آن‌ها هم‌آوا – هم‌نویسه هم گفتـه می‌شـود) دو کلمـه بـا ظـاهر کاملاً یکسان و معانی متفاوت هستند که اغلب به حوزه‌های موضـوعی متفـاوتی نیـز تعلق دارند و در سرعنوان‌های موضوعی فارسی با پرانتز بـه ایـن حوزه‌هـا اشاره‌شـده است. این کلمات حتی‌الامکان در همه معانی و صـورت‌ها در چهـارچوب معیارهـای تعیین‌شده در جامعه واردشده‌اند، مانند «عود» و «عود (ماده معطر)».
- واژه‌های لاتین فارسی نوشته‌شده که ارجاع «بـه‌جای» نداشـتند، حـذف شـدند، ماننـد «آمازون‌ها»، «تردمیل» یا «پاراگلایدرها».
- نام ورزش‌های مختلف حذف شد.
- سرعنوان‌های حاوی نام و اسامی خاص، مانند نام اشخاص، قومیت‌ها، قبایل، طوایف، نژادها و خاندان‌ها، نام گویش‌ها و زبان‌هـا، نـام خـدایان و ایـزد بـانوان اسـاطیری و شخصیت‌های افسانه‌ای و داستانی، تنالگان‌ها، اسامی جغرافیـایی، ماننـد نـام نـواحی، مناطق، کشورها، شهرها و روستاها، مناظر و بناهـای قـدیم و جدیـد، نـام تأسیسـات شهری، مثل بزرگراه‌ها، میدان‌ها، خیابان‌ها، نـام پارک‌هـا، جنگل‌هـا، بیشـه‌ها، منـاطق حفاظت‌شده، از جامعه موردمطالعه حذف‌شده‌اند.

[1]. در انگلیسی به آن‌ها Homonym گفته می‌شود.

- اسامی ماه‌ها و روزهای هفته، جشن‌ها و مناسبت‌ها از جامعه حذف شدند، مانند «شنبه»، «نوروز»، «تیرگان» و غیره.
- درجایی که در واردکردن یا نکردن سرعنوانی تردید وجود داشت، سرعنوان در جامعه وارد شد تا جامعیت کار حفظ شود. به‌عنوان‌مثال، ترکیبات «خود» مانند «خودکاری»[1] یا «خون» مانند «خون‌خواهی» و ازاین‌قبیل.

همه سرعنوان‌های پذیرفته‌نشده که به هر سرعنوان اصلی ارجاع می‌دهند و با اصطلاح «به‌جای» در مقابل سرعنوان اصلی مشخص‌شده‌اند، در ستون دوم مقابل هر سرعنوان ذکر شدند و لزوماً مترادف لغت‌نامه‌ای سرعنوان پذیرفته‌شده نیستند. مقایسهٔ مترادف‌ها با این دسته از سرعنوان‌ها صورت گرفته است.

اصطلاحات برگرفته از *اصطلاح‌نامهٔ اصفا* با همان معیارهای در نظر گرفته‌شده در *سرعنوان‌های موضوعی فارسی* در جامعه وارد و در ستون مربوط درج شدند. بدین ترتیب، سرعنوان موضوعی ۳۲۷۰ و اصطلاح ۲۰۲۰ در جامعه پژوهش واردشده‌اند.

در این پژوهش، مرحلهٔ استخراج سرعنوان‌ها و اصطلاحات بر اساس چهارچوب تعیین‌شده و استخراج مترادف‌های آن‌ها در فارس‌نت به‌صورت دستی توسط پژوهشگر صورت گرفته است؛ اما مرحلهٔ سنجش شباهت با استفاده از ابزارهای ماشینی پردازش زبان طبیعی صورت گرفته در ادامه گفته خواهد شد. پردازش‌های ماشینی، هرچند بار هم که با یک مجموعه داده تکرار شوند، نتایج یکسان به بار می‌آورند و الزامات پایایی را به‌خوبی برآورده می‌کنند؛ اما در قسمت گردآوری، یا همان انتخاب سرعنوان‌ها و اصطلاحات بر اساس معیارهای داده‌شده، سنجش پایایی به این صورت انجام شد که بخشی از اصطلاحات را پژوهشگر دیگری استخراج کرده و میزان شباهت آن با کار پژوهشگر اصلی به‌عنوان پایایی در نظر گرفته‌شده است. از *سرعنوان‌های موضوعی فارسی* حرف آ و الف و از *اصطلاح‌نامهٔ اصفا*، اصطلاحات حوزهٔ علوم و فناوری مجدداً توسط پژوهشگر دیگری استخراج شد که با توجه به اسمی (از نوع کلمه) بودن داده‌ها و تعداد آن‌ها (بالای ۲۰۰۰)، میزان توافق دو پژوهشگر با کسینوس شباهت سنجیده و به ترتیب توافق ۰/٦۱۸ و ۰/۷۰٦ از بازه میان صفر و ۱ به دست آمد که نمره قابل‌قبولی محسوب می‌شود.

---

[1]. معادل Automation

برای شرح ویژگی‌های جامعه و نمونه موردمطالعه از آمار توصیفی شامل میانگین و انحراف معیار استفاده شد. برای سنجش اینکه مترادف‌ها تا چه میزان در سرعنوان‌های ناگزیده و اصطلاحات نامرجح به کار گرفته‌شده‌اند، این سرعنوان‌ها و اصطلاحات که به سرعنوان یا اصطلاح اصلی ارجاع می‌دهند، با مترادف‌های هر سرعنوان یا اصطلاح در فارس‌نت مقایسه شده‌اند. این مقایسه با استفاده از شباهت‌سنجی صورت‌گرفته است. سنجش شباهت یا تفاوت روش‌های مختلفی دارد که بنا بر ویژگی این پژوهش از روش کسینوس شباهت[1] با استفاده از زبان برنامه‌نویسی پایتون استفاده‌شده است. در کسینوس شباهت فاصلهٔ بین دونقطه شباهت (شباهت میان سرعنوان یا اصطلاح مترداف آن) به‌طور ساده، کسینوس زاویهٔ بین دو بردار است. کسینوس شباهت حاصل تقسیم بین ضرب نقطه‌ای بردارها و حاصل‌ضرب نرم‌های اقلیدسی یا اندازه هر بردار است:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$

در صورت انطباق دو بردار که در این معیار نشانه شباهت کامل است، زاویهٔ بین دو بردار صفر است و مقدار شباهت ۱ خواهد بود و در کمترین میزان شباهت دو بردار، یعنی هنگامی‌که زاویهٔ بین دو بردار ۱۸۰ درجه باشد نتیجه ۱- است. نقاط با زوایای کمتر، شباهت بیشتری دارند و نقاط با زوایای بزرگ‌تر از همدیگر متفاوت‌تر هستند. کسینوس شباهت به خود زاویهٔ مربوط نیست، بلکه کسینوس زاویه در نظر گرفته می‌شود؛ بنابراین، زاویه‌های کمتر (زیر ۹۰ درجه) شباهت بیشتری دارند. در حوزهٔ متن‌کاوی و اطلاعات متنی، کسینوس شباهت مدارک بین صفر و یک درجه‌بندی می‌شود، چراکه بسامد واژگان را نمی‌توان به‌صورت نمره منفی نشان داد و زاویهٔ میان دو بردار بسامد واژگان بزرگ‌تر از ۹۰ درجه نخواهد بود. مثلاً اگر زاویهٔ بین دو بردار ۹۰ درجه باشد (متعامد یا عمود بر یکدیگر باشند)، کسینوس شباهت صفر بوده و هیچ شباهتی میان این دو بردار نیست. هر چه زاویهٔ میان دو بردار کوچک‌تر یا کمتر از ۹۰

[1]. Cosine similarity

درجه شود، شباهت بیشتر است و نمره شباهت به ۱ میل می‌کند. کسینوس شباهت برای استفاده در داده‌هایی با ابعاد و اندازه بالا ولی ویژگی‌های تُنُک مناسب است، بنابراین تفاوت در مقادیر داده در نظر گرفته نمی‌شود. این معیار یکی از پرکاربردترین‌ها در پردازش متن، متن‌کاوی و سنجش شباهت یا عدم شباهت متون است (برنتی، ۱۳۹۹).

در هر عملکرد پردازش زبان طبیعی با استفاده از زبان‌های برنامه‌نویسی و کتابخانه‌های آن‌ها یک مرحلهٔ اصلی پیش از ورود داده در الگوریتم‌ها وجود دارد که به آن پیش‌پردازش می‌گویند. داده‌ها همان اصطلاحات استخراج‌شده و مترادف‌های آن‌ها هستند و در ستون‌های اکسل ردیف شده‌اند و بالاتر برشی از آن نشان داده شد. داده‌ها باید به‌گونه‌ای پردازش شوند که بتوان آن‌ها را در الگوریتم وارد کرد. درباره پیش‌پردازش‌های لازم نظرات گوناگونی ارائه‌شده است. آنچه مسلم است اینکه برای پیش‌پردازش داده‌های متنی و ورود به الگوریتم روش یکسانی وجود ندارد (Shi, 2019) و پیش‌پردازش‌ها به‌شدت به محتوای پیکره و ویژگی‌های آن وابسته‌اند. پیش‌پردازش‌های به‌کاربرده شده در این پژوهش عبارت‌اند از:

- نرمال‌سازی متن که عبارت است از یکسان‌کردن نویسه‌های استفاده‌شده در متون و تبدیل آن به قالب مشترک یونی‌کد؛
- فیلترینگ یا پاک‌سازی: بررسی و حذف مدارک مخدوش و غیرمرتبط. حذف مدارک مخدوش و غیرمرتبط در مرحلهٔ گردآوری صورت گرفته است. پاک‌سازی از عناصر نامطلوب در متن، شامل حذف علائم سجاوندی، اعداد و سایر علائم، حذف حروف انگلیسی و لاتینی و حذف ایست‌واژگان از مدارک نیز جزء این مرحله است. این علائم با استفاده از کدهای عبارت‌های منظم و کدهای مناسب دیگر از پیکره حذف می‌شوند. به لحاظ نوع داده‌ها در آن‌ها حذف ایست‌واژه صورت نگرفت؛ ولی برخی علائم مانند (،)، [،]، –، و فضاهای اضافی در این مرحله حذف شدند.
- توکن‌بندی؛ یا تبدیل متن به کلمات تشکیل‌دهنده آن.

ابزارهای پیش‌بینی‌شده برای گردآوری و اجرای پژوهش عبارت‌اند از:

- نرم‌افزار اکسل نسخه ۲۰۱۳ (۲۵)؛

- زبان برنامه‌نویسی و کتابخانه‌های متنوع پایتون، در محیط اجرای Jupyter notbook و Conda. در مراحل مختلف، بسته به نوع تحلیل موردنیاز از کتابخانه‌های مختلف پایتون آن (آخرین نسخه هر کتابخانه) به شرح زیر استفاده‌شده است:
- کتابخانه‌های پایتون مانند os برای کار با سیستم‌عامل و فایل سیستم‌ها، کتابخانه Nltk برای پردازش و پیش‌پردازش‌های پیکره،
- کتابخانه Scikit-Learn برای محاسبه کسینوس شباهت؛
- Pandas و Numpy برای ایجاد تغییرات ضروری برای تحلیل، ساخت ماتریس‌های چندبعدی، آرایه‌بندی؛
- جعبه‌ابزار پارسیور تحت پایتون برای پیش‌پردازش‌های زبان فارسی.[1]
- نرم‌افزار spss نسخه ۲۵ برای محاسبات آمار توصیفی.

یافته‌ها

سرعنوان‌های موضوعی فارسی (۱۳۹۷) که برای اجرای این پژوهش مورداستفاده قرارگرفته است، حاوی ۲۷۸۰٤ سرعنوان اصلی بوده که از این تعداد مطابق با معیارهای گفته شده در روش‌شناسی ۳۲۷۰ سرعنوان برگزیده شد. در اثر اصلی به‌طور متوسط ۸٦۹ سرعنوان در هر حرف الفبا با انحراف معیار به ۸٥٤ وجود دارد که نشان از میزان بالای تفاوت در توزیع ذیل هر حرف الفباست و در نمونه‌های انتخابی شده به‌صورت هدفمند به‌طور متوسط ۱۰۲ سرعنوان در هر حرف الفبا و انحراف معیار ۱۰۱ وجود دارد و در جدول ۱ نشان‌داده‌شده است.

جدول ۱- میانگین و انحراف معیار داده‌ها در سرعنوان‌های موضوعی

		حرف الفبا	س ع در اثر اصلی	س ع در نمونه
داده معتبر		۳۲	۳۲	۳۲
میانگین			۸٦۹	۱۰۲/۱۸۷٥
انحراف استاندارد			۸٥٤/٥٦۰۷	۱۰۱/۲۳۹۸
جمع			۲۷۸۰٤	۳۲۷۰

[1]. https://github.com/ICTRC/Parsivar

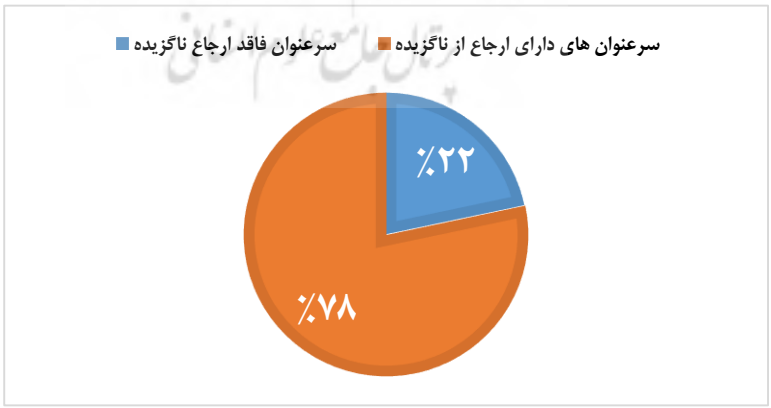
اصطلاح‌نامهٔ اصفا ذیل ۱۷ حوزهٔ موضوعی تنظیم‌شده است که تا اجرای این پژوهش مجموعاً ۱۱۲۱۰ اصطلاح داشته و مطابق با محدودهٔ تعیین‌شده برای نمونه‌ها، تعداد ۲۰۲۰ از آن‌ها برای پژوهش انتخاب‌شده‌اند. در *اصطلاح‌نامهٔ اصفا* به‌طور متوسط ٦٥٩ اصطلاح مرجح با انحراف معیار ۳٤۲ در هر حوزهٔ موضوعی وجود دارد. در نمونه موردبررسی از *اصطلاح‌نامهٔ اصفا* به‌طور میانگین ٦۳ اصطلاح ذیل هر حرف الفبا وجود دارد و در جدول ۲ نشان داده‌شده است.

جدول ۲- میانگین و انحراف معیار داده‌ها در اصطلاحات

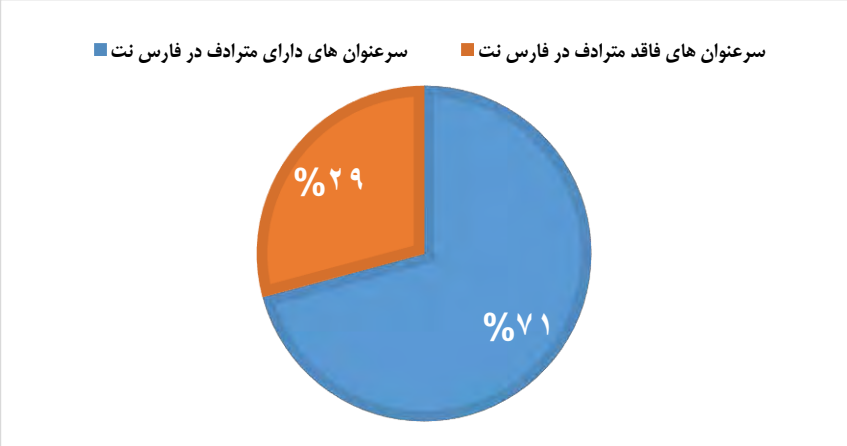
	تعداد حوزهٔ موضوعی	فراوانی	میانگین	انحراف معیار
اصطلاحات در جامعه اصلی	۱۷	۱۱۲۱۰/۰۰	٦٥٩۱/٤۱۱۸	۳٤۲/٦۳
اصطلاحات در نمونه	۱۷	۲۰۲۰	٦۳/۱۲٥۰	٦۳/۳۹۳۷

وضعیت کاربرد مترادف‌ها در *سرعنوان‌های موضوعی فارسی* در تطبیق با *فارس‌نت*

منظور از وضعیت کاربرد مترادف‌ها در *سرعنوان‌های موضوعی فارسی* در تطبیق با *فارس‌نت* این است که بررسی کنیم چه میزان ارجاعات در قالب سرعنوان‌های ناگزیده برای بازیابی بهتر در سامانه به کار گرفته شده و این سرعنوان‌های ناگزیده تا چه میزان از مترادف‌های زبانی گرفته‌شده‌اند. در جامعه پژوهش، ۲٥٦۱ سرعنوان اصلی (۷۸/۳ درصد) دارای سرعنوان ناگزیده بوده‌اند که به سرعنوان گزیده ارجاع می‌دهد. ۲۳۱٦ سرعنوان اصلی (۷۰/۸ درصد) نیز دارای مترادف در *فارس‌نت* بوده‌اند که در نمودارهای ۱ و ۲ نشان داده‌شده است.



نمودار ۱- درصد فراوانی سرعنوان‌های دارای و فاقد سرعنوان ارجاعی



نمودار ۲- درصد فراوانی سرعنوان‌های دارای مترادف در *فارس‌نت*

برای سنجش اینکه چه مقدار از مترادف‌های موجود در زبان فارسی در سرعنوان‌های ناگزیده به کار گرفته‌شده است، ستون سرعنوان‌های ناگزیده در داده‌های گردآوری‌شده با ستون مترادف‌های استخراج‌شده برای هر سرعنوان اصلی از *فارس‌نت* به روش کسینوس شباهت، شباهت‌سنجی شدند. نمره شباهت ۰/۱۲۵ حاصل محاسبات بود که نشان از شباهت اندک این سرعنوان‌ها به مترادف‌های موجود در زبان فارسی، موجود در *فارس‌نت* است (شکل ۳):

```
from sklearn.metrics.pairwise import cosine_similarity
print(cosine_similarity(df, df))
```

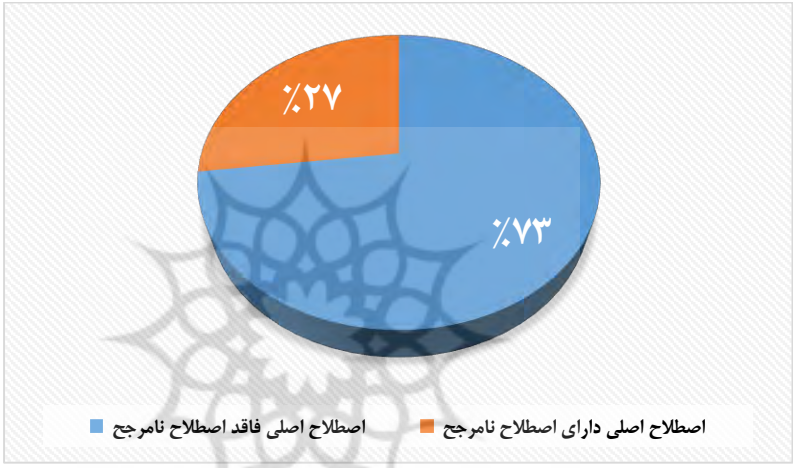
```
[[1.         0.12535118]
 [0.12535118 1.        ]]
```

***شکل ۳*- نمره شباهت میان مترادف‌های موجود در زبان فارسی برای سرعنوان‌های اصلی و سرعنوان‌های ناگزیده: خروجی گرفته‌شده در کتابخانه سایکیت لرن**

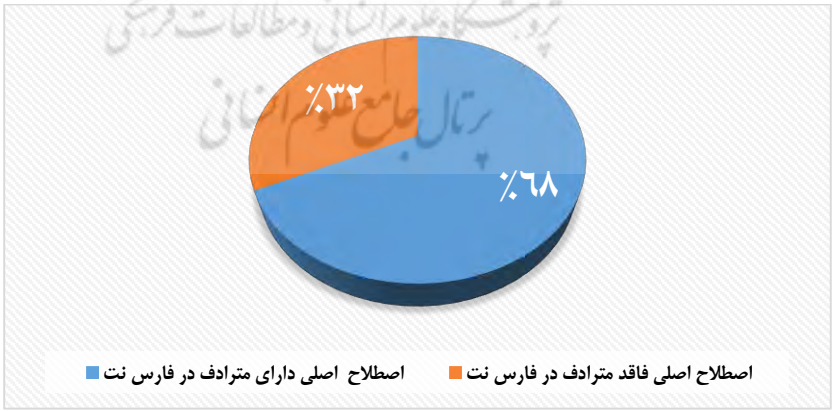
در پیوست ۱ چند نمونه از سرعنوان‌های اصلی، سرعنوان ارجاعی ناگزیده که به آن‌ها ارجاع می‌دهد و مترادف‌های سرعنوان اصلی که در *فارس‌نت* موجود بوده ولی مورداستفاده قرار نگرفته، قابل‌ملاحظه است.

وضعیت کاربرد مترادف‌ها در *اصطلاح‌نامهٔ اصفا* در تطبیق با *فارس‌نت*

در نمونهٔ مستخرج از *اصفا* ٥٤٥ اصطلاح (حدود ٢٧ درصد) دارای اصطلاح ارجاعی نامرجح بوده‌اند. ١٣٧٦ اصطلاح (٦٨ درصد) از این اصطلاحات نیز دارای مترادف در *فارس‌نت* هستند؛ یعنی تعداد ١٤٧٥ اصطلاح (٧٣ درصد) فاقد اصطلاح نامرجح (که ارجاع به اصطلاح اصلی می‌دهند) بوده‌اند. تعداد ٦٤٤ اصطلاح اصلی (٣١ درصد) نیز در *فارس‌نت* فاقد مترادف بوده‌اند که در نمودار ٣ و ٤ نشان داده‌شده است.



نمودار ٣- درصد فراوانی اصطلاحات *اصفا* دارا و فاقد اصطلاح ارجاعی نامرجح



نمودار ٤- درصد فراوانی اصطلاحات دارا و فاقد مترادف در *فارس‌نت*

برای سنجش شباهت میان اصطلاحات نامرجع در اصطلاح‌نامهٔ اصفا و میزان به‌کارگیری مترادف‌های زبان فارسی موجود در فارس‌نت نیز ستون اصطلاحات نامرجع در داده‌های گردآوری‌شده با ستون مترادف‌های استخراج‌شده برای هر اصطلاح اصلی از *فارس‌نت* به روش کسینوس شباهت، شباهت‌سنجی شدند. نمره شباهت ۰/۱۳۱ حاصل محاسبات بود که نشان از شباهت اندک این اصطلاحات به مترادف‌های موجود در زبان فارسی، موجود در *فارس‌نت* است:

```
from sklearn.metrics.pairwise import cosine_similarity
print(cosine_similarity(df, df))
```

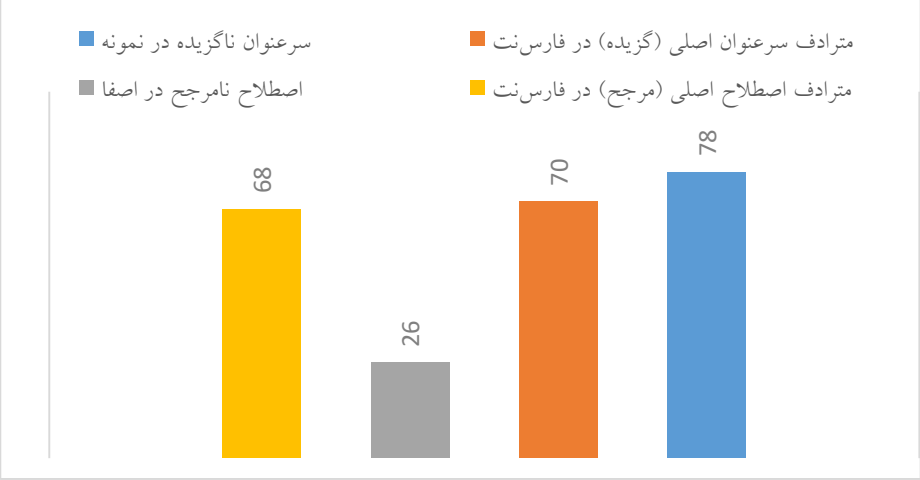
```
[[1.         0.13164624]
 [0.13164624 1.        ]]
```

شکل ٤- نمرهٔ شباهت میان اصطلاحات نامرجع در *اصطلاح‌نامهٔ اصفا* و مترادف‌های اصطلاحات مرجع در *فارس‌نت*: خروجی گرفته‌شده در کتابخانه سایکیت لرن

در پیوست ۲ نمونه‌هایی از اصطلاحات اصلی، اصطلاح ارجاعی نامرجع که به آن‌ها ارجاع می‌دهند و مترادف‌های اصطلاح اصلی که در *فارس‌نت* موجود بوده ولی مورداستفاده قرار نگرفته، ذکرشده است.

مقایسهٔ کاربرد مترادف‌ها در *سرعنوان‌های موضوعی فارسی* در مقایسه با *اصطلاح‌نامهٔ اصفا*

در نمونهٔ موردبررسی مستخرج از سرعنوان‌ها، ۲۵٦۱ سرعنوان اصلی دارای سرعنوان ارجاعی ناگزیده بوده‌اند که حدود ۷۸ درصد از آن‌ها را تشکیل می‌دهد. همچنین ۲۳۱٦ مورد از سرعنوان‌های اصلی نمونه، معادل ۷۰ درصد، دارای مترادف در *فارس‌نت* هستند؛ اما در نمونه مستخرج از *اصفا* ٥٤٥ اصطلاح (حدود ۲٦ درصد) دارای اصطلاح ارجاعی نامرجع بوده‌اند. ۱۳۷٦ (٦۸ درصد) مورد از این اصطلاحات نیز دارای مترادف در *فارس‌نت* هستند. این یافته‌ها در نمودار ٥ نشان داده‌شده‌اند.



نمودار ۵- به‌کارگیری سرعنوان‌های ارجاعی (ناگزیده) و اصطلاحات ارجاعی (نامرجح) و موجودبودن مترادف‌های سرعنوان‌ها و اصطلاحات اصلی در *فارس‌نت*

نتیجه‌گیری

مطابق یافته‌ها، ۲۵۶۱ سرعنوان اصلی (۷۸/۳ درصد) دارای سرعنوان ناگزیده بوده‌اند که به سرعنوان گزیده ارجاع می‌دهد. ۲۳۱۶ سرعنوان اصلی (۷۰/۸ درصد) نیز دارای مترادف در *فارس‌نت* بوده‌اند. نمرهٔ شباهت ۰/۱۲۵ حاصل محاسبات دال بر شباهت اندک این سرعنوان‌ها به مترادف‌های موجود در زبان فارسی، موجود در *فارس‌نت* است. هرچند برخی از سرعنوان‌های اصلی اصولاً فاقد مترادف در زبان فارسی هستند، اما با توجه به مقدار قابل‌توجه مترادف‌های موجود در *فارس‌نت* برای سرعنوان‌های اصلی، این مسئله می‌تواند به علت‌هایی مانند عدم احساس لزوم ساخت ارجاعات از همه مترادف‌های واژه، یا عدم رجوع فهرست‌نویس به ابزارهای واژگانی مانند واژه‌نامه‌ها و فرهنگ‌های اصطلاحات، (که فارس‌نت نیز از همین قبیل است) برای استخراج مترادف‌ها و تکیه‌بر گنجینه زبانی حاضر در ذهن باشد. لازمهٔ قضاوت قطعی‌تر در این خصوص پژوهشی جداگانه است. البته باید توجه داشت که بسیاری از مفاهیم لزوماً به همان شکل در لغت‌نامه‌ها وجود ندارند و مفاهیمی هستند که زایش اندیشه در زبان به وجود می‌آورد که با تبدیل به منبع اطلاعاتی مانند کتاب، از پشتوانهٔ انتشاراتی نیز برخوردار می‌شود. بسیاری از این مفاهیم را می‌توان در دایره‌المعارف‌ها جای داد که توضیح آن‌ها نه به‌صورت مترادف‌هایی با یک کلمه یا عبارتی کوتاه، بلکه با جملاتی

طولانی‌تر قابل‌ارائه است و در لغت‌نامه‌ها به‌ندرت می‌توان مترادف‌های لغت‌نامه‌ای برای آن‌ها یافت. با این وجود استفاده از مترادف‌های موجود به‌عنوان شناسهٔ افزوده موضوعی بر غنای سیستم بازیابی می‌افزاید.

در نمونه مستخرج از *اصفا* ۵۴۵ اصطلاح (حدود ۲۷ درصد) دارای اصطلاح ارجاعی نامرجح بوده‌اند. ۱۳۷۶ اصطلاح (۶۸ درصد) از این اصطلاحات نیز دارای مترادف در *فارسنت* هستند؛ یعنی تعداد ۱۴۷۵ اصطلاح (۷۳ درصد) فاقد اصطلاح نامرجح (که ارجاع به‌اصطلاح اصلی می‌دهند) بوده‌اند. تعداد ۶۴۴ اصطلاح اصلی (۳۱ درصد) نیز در *فارسنت* فاقد مترادف بوده‌اند.

نمره شباهت ۰/۱۳۱ نشان از شباهت اندک این سرعنوان‌ها به مترادف‌های موجود در زبان فارسی، موجود در *فارسنت* است. به‌طورکلی، در حوزه‌های موضوعی هفده‌گانه *اصطلاح‌نامهٔ اصفا* تعداد بسیار کمی ارجاع از نامرجح به مرجح وجود دارد و درنتیجه نمی‌توان انتظار داشت که این مقدار پایین اصطلاح نامرجح نمره شباهت بالایی با مترادف‌ها داشته باشند و نه از مترادف‌ها و نه از تعابیر دیگر برای ساخت ارجاع استفاده چندانی نشده است. ولی برای اصطلاحات اصلی در *فارسنت* مترادف وجود دارد که میزان آن در نمونه‌های موردمطالعه تا حدود زیادی به هم نزدیک است.

از مجموع یافته‌ها دراین‌خصوص می‌توان چنین استنباط کرد که در *سرعنوان‌های موضوعی فارسی* تقید بیشتری در ساخت و به‌کارگیری ارجاعات از مفاهیم واژگانی‌شده (مترادف زبانی موجود در لغت‌نامه‌ها یا ترکیب‌هایی که سازنده سرعنوان موضوعی آن را برای هدایت کاربر به سرعنوان اصلی مناسب تشخیص می‌دهد) دیده می‌شود. دیگر اینکه *اصفا* از حیث کاربرد ارجاعات موضوعی برای افزایش نقاط دسترسی موضوعی کاربر (۲۷ درصد) ضعیف ارزیابی می‌شود. هرچند قضاوت درباره کیفیت این ارجاعات یا مفاهیم نزدیک مطالعه دیگری را می‌طلبد و بسیاری از آن‌ها برگرفته از مترادف‌های لغت‌نامه‌ای نیستند. به‌هرروی، داده‌ها وجود میزان خوبی از مترادف برای مفاهیم را نشان می‌دهد که می‌توان از آن‌ها برای ساخت ارجاعات و کمک به بازیابی بهتر بهره برد.

از حیث شباهت مترادف‌های سرعنوان‌ها/ اصطلاحات اصلی در *فارسنت* با آنچه در هر دو ابزار در ساخت سرعنوان/ اصطلاح ارجاعی (سرعنوان ناگزیده و اصطلاح نامرجح) به‌کاربرده شده، می‌توان گفت که تفاوت چندانی میان این دو ابزار ملاحظه نشد؛ هرچند نمره

شباهت کسینوسی میان مترادف‌ها با اصطلاحات نامرجح در *اصفا* (۰/۱۳۱ درصد) به میزان اندکی از *سرعنوان‌های موضوعی فارسی* (۰/۱۲۵ درصد) وضعیت بهتری دارد، ولی به علت تعداد بسیار کم‌کاربرد اصطلاحات نامرجح در مقایسه با سرعنوان‌های ناگزیده برای ارجاع به موضوعات اصلی، این برتری اندک را نمی‌توان معیار قطعی برای قضاوت دانست؛ بنابراین، هرچند بیش از ۷۰ درصد از سرعنوان‌ها دارای اصطلاح ارجاعی ناگزیده برای بهبود بازیابی دارند، به علت تفاوت آن‌ها با مترادف‌های زبانی می‌توان چنین نتیجه گرفت که جای استفاده بهتر از ابزارهای واژگانی زبانی در تدوین این ابزار وجود دارد.

از منظر مقایسه با پژوهش‌های پیشین، می‌توان گفت که هرچند پژوهش‌هایی مانند فرناندز لانزا و همکاران (۲۰۰۳)، سوتو و همکاران (۲۰۰۸)، عبدالحسن و هادی (۲۰۱۷)، لی و همکاران (۲۰۱۳)، زنگ و همکاران (۲۰۱۲) و کلبسا (۲۰۲۱) نقش مترادف‌ها را در کارآمدی بیشتر بازیابی موردبررسی قرار داده‌اند؛ برنجیان و رئیسی (۱۳۹۳) نیز باهدف ارائهٔ واژه‌های مصوب فرهنگستان زبان و ادب فارسی در زمان بازیابی معادل رایج آن‌ها در سیستم بازیابی اطلاعات مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری و همچنین ارائهٔ معادله‌ای رایج واژه‌ها در زمان جستجوی واژه‌های مصوب، سیستمی را طراحی و ارزیابی کردند؛ و بهاراتی و ونکاتسان(۲۰۱۲) روش نمایه‌سازی مبتنی بر مترادف معنایی جدید ارائه کردند، اما ابزارهای سازمان‌دهی محتوا از حیث کاربرد مترادف‌ها موردبررسی قرار نگرفته‌اند و به همین علت این پژوهش‌ها فقط از حیث لزوم کاربرد ابزارهایی مانند وردنت‌ها در نظام‌های بازیابی که یکی از پیشنهادهای مهم این پژوهش نیز هست، با این پژوهش شباهت دارند و از حیث یافته‌ها و نتایج نمی‌توان مقایسه‌ای صورت داد. همچنین از حیث همین هدف اصلی پژوهش حاضر مبنی بر لزوم استفاده از روابط معنایی زبان و مترادف‌ها به‌طور خاص، این پژوهش با پژوهش سوتو و همکاران (۲۰۰۸) که روشی برای گسترش پرس‌وجوهای کاربر ارائه کردند که در آن برای هر عبارت در پرس‌وجوی اصلی، تمام مترادف‌های آن با معنای خاصی با حداکثر فراوانی مفهومی معرفی می‌شوند و بازیل (۲۰۱۵) که معتقد است که وردنت (ها) را می‌توان به‌عنوان جایگزینی برای هستی‌شناسی‌ها در فرایندهای تولید زبان طبیعی [مانند تولید چکیده، عبارت های کلیدی و کلیدواژه] و یا در ترکیب با سایر پایگاه‌های دانش به کار گرفت، هم‌سویی ویژه‌ای دارد؛ بدین معنا که این پژوهش، معرفی مترادف‌ها را برای همه کاربران، ازجمله فهرست‌نویسان و متصدیان ساخت مستندات موضوعی، چه حین جستجوی مفاهیم و چه در

ساخت اصطلاحات توصیه می‌کند و می‌تواند گامی در جهت بهبود وضعیت بانک‌های مستند موضوعی و درنهایت تجربه متکامل‌تر کاربر در بازیابی منابع موردنظر باشد.

پیشنهادها

در مراحل مختلف اجرای پژوهش، موضوعاتی وجود داشتند که می‌توانند موضوع پژوهش مستقلی باشند. پژوهش‌هایی نیز در تکمیل بحث اصلی این پژوهش قابل‌اجرا هستند که در ادامه ذکر می‌شوند:

- گونه‌شناسی معنایی و شکلی سرعنوان‌های موضوعی و اصطلاحات؛
- روند واژه‌گزینی برای بازنمایی مفاهیم در قالب سرعنوان گزیده و ناگزیده/ اصطلاح مرجح و نامرجح؛
- بررسی ترادف درونی (همپوشانی معنایی) در *سرعنوان‌های موضوعی فارسی*؛
- بررسی ترادف درونی (همپوشانی معنایی) در *اصطلاح‌نامهٔ اصفا*؛
- بررسی نقش مترادف‌ها در عملکرد بازیابی از حیث جامعیت و مانعیت.

قدردانی

بر خود لازم می‌دانم از همکاران محترم سازمان اسناد و کتابخانه ملی ج. ا. ا. که در اجرای این پژوهش یاری‌رسانم بوده‌اند: آقایان دکتر مهدی خادمیان، رئیس محترم گروه مستندسازی (کتابخانه)، محمد ایرانشاهی، رئیس محترم انتشارات، خانم‌ها دکتر نرجس‌خاتون عزیزیان، کارشناس گروه مستندسازی و سمانه خاتمساز، رئیس محترم اداره راهبری و پشتیبانی برنامه‌های کاربردی، دکتر محبوبه قربانی که اطلاعات و منابع موردنیاز اینجانب را در اختیارم نهادند، سپاسگزاری کنم. از جناب آقای مهندس علیرضا میقانی که به‌عنوان پژوهشگر دوم زحمت بررسی مجدد، انتخاب اصطلاحات و تهیه فایل از بخشی از داده‌ها را برای انجام عملکرد پایایی‌سنجی بر عهده گرفتند، صمیمانه متشکرم. نکته‌سنجی‌های دلسوزانه سرکار خانم دکتر عاطفه شریف، ناظر طرح را ارج می‌نهم.

تضاد منافع

هیچ‌گونه تعارض منافع توسط نویسنده بیان‌نشده است.

پیوست‌ها

پیوست ۱. نمونه‌هایی از سرعنوان‌های اصلی، سرعنوان‌های ناگزیده (ارجاعی) و مترادف‌های آن‌ها در فارس‌نت

	سرعنوان گزیده	سرعنوان ناگزیده	مترادف‌های سرعنوان گزیده در *فارس‌نت*
۱	آبدزدک	نداشت	زمین‌سنبه، خاک‌سنبه
۲	آبشارها	آبشرها	آب‌شیب، آبشار
۳	آب‌مروارید	کاتاراکت	آب‌سفید، کاتاراکت
٤	آرنج	نداشت	مرفق، وارن، زند زیرین، اولنا، زند پایینی، زند اسفل، آرنجگاه
٥	آسمان‌خراش‌ها	ساختمان‌سازی آسمان‌خراش‌ها، ساختمان‌های اداری، ساختمان‌های بلند، ساختمان‌های فلزی، معماری	برج
٦	آغازیان	موجودات تک‌سلولی	تک‌یاخته
۷	ابریشم	ابریشمی، پارچه‌های؛ پارچه‌های حریر، حریر، پارچه‌های ابریشمی	بریشم، ابریشم، سیلک، پرنیان، پرند
۸	اسانس‌ها	گیاهان معطر، جوهر	عطرمایه
۹	استنساخ	نسخه‌برداری	رونویسی، کپی‌برداری، رونوشت برداشتن، نسخه گرفتن، نسخه برداشتن
۱۰	هدایت	راهنمایی	سوق، راهنمایی، ارشاد، سرمشق‌دهی، رهنمود، راه‌نمود، دلالت، هدایت
۱۱	نویسندگی	ادبیات به‌منزله حرفه نویسندگی	انشا کردن، مرقوم‌کردن، نوشتن، مرقوم‌داشتن، نگاشتن، مرقوم فرمودن، تحریر کردن، به رشتهٔ تحریر درآوردن، نویسندگی کردن، برنوشتن، مکتوب کردن، ترقیم کردن، به تحریر درآوردن، به رشتهٔ نگارش درآوردن، کتابت کردن، به رشتهٔ تحریر آوردن، نبشتن، به رشتهٔ تحریر کشیدن، عبارت سازی، قلم‌زنی
۱۲	مهمانی‌ها	نداشت	خوان، سورچرانی، شیلان، مهمانی، بزم، ضیافت، ولیمه،

	سرعنوان گزیده	سرعنوان ناگزیده	مترادف‌های سرعنوان گزیده در *فارس‌نت*
			سور، میهمانی
۱۳	چاشنی‌ها	نداشت	طعم‌دهنده
۱٤	خسوف	ماه‌گرفتگی	گرفت ماه، ماه‌گرفتگی، مه‌گرفت
۱۵	کسوف	خورشیدگرفتگی	گرفت خورشید، کسوف، خورشیدگرفتگی، آفتاب‌گرفتگی، خورگیر، خورگرفت
۱٦	کتاب‌شیدایی	بیماری جمع‌آوری کتاب، جنون کتاب	نداشت
۱۷	کاسنی	شیکوره	تلخک
۱۸	کارآفرینی	آنت رپر نور، کاروری، کسب‌وکار بزرگ	کارسازی، اشتغال‌زایی، شغل‌سازی، اشتغال‌زایی‌کردن
۱۹	قورباغه‌ها	وزغ‌ها	چغز، وزغ، غوک، ضفدع
۲۰	قضاوقدر	تقدیر، سرنوشت، فاتالیسم، قدر و قضا	مقدرات، نصیب، مقدر، قضا، قسمت، قدر، طالع، اقبال، سرنوشت، تقدیر، قضاوقدر، بخت، پیشانی‌نوشت

پیوست ۲. نمونه‌هایی از اصطلاحات اصلی، اصطلاحات نامرجح (ارجاعی) و مترادف‌های آن‌ها در فارس‌نت

	اصطلاح مرجح یا اصلی	اصطلاح نامرجح	مترادف‌های اصطلاح مرجح در فارس‌نت
۱	ابریشم	نداشت	پرنیان، پرند، حریر، بریشم، ابریشم، سیلک
۲	اتلاف	نداشت	تضییع، تبذیر، اسراف، اتلاف، خاصه‌خرجی، ریخت‌وپاش، اتلاف‌کردن، تباهیدن، تضییع‌کردن، برباددادن، تباه ساختن، به هدردادن، تلف نمودن، تباه‌کردن، هدر کردن، تلف کردن، بر بادکردن، حرام کردن، تلف ساختن، هدردادن
۳	اجاره	نداشت	کرایه
٤	احضار	نداشت	فراخوانی
۵	ادغام	نداشت	یکی‌کردن، ادغام کردن، یک‌جا کردن، یک‌کاسه کردن
٦	اسکناس	نداشت	پول کاغذی، پول‌برگ
۷	اسلیمی	عربانه	نداشت
۸	اغراق	نداشت	مبالغه، غلو، بزرگ‌نمایی، اغراق، اغراق‌گویی، درشت‌نمایی کردن، بزرگ گردانیدن، شاخ‌وبرگ دادن، گنده کردن

	اصطلاح مرجع یا اصلی	اصطلاح نامرجع	مترادف‌های اصطلاح مرجع در فارس‌نت
۹	پارچه	قماش	منسوج، قماش
۱۰	تعزیه‌خوانی	تعزیه‌گردانی	تعزیه‌گردانی، شبیه‌خوانی
۱۱	سور	کمیت نما	نداشت
۱۲	شفقت	نداشت	دلسوزی
۱۳	صبر	بردباری، شکیبایی	حلم، بردباری، شکیبایی
۱٤	ضایعات	نداشت	پس‌مانده، پس‌ماند
۱۵	عصاره	اسانس	شیره، عصیر، چکیده، کنه، فشرده، هسته، چیستی، مغز، زبده، ماهیت، عصاره، جوهره، جوهر، جان
۱٦	غسالخانه‌ها	نداشت	مرده‌شوی‌خانه، مرده‌شور خانه، غسل‌خانه، شورخانه، مرده‌شوخانه
۱۷	فراموشی	نداشت	از یاد بردن، به فراموشی سپردن، فراموش کردن، از حافظه زدودن، نسیان کردن، از سر نهادن، از سر افتادن، به دست فراموشی سپردن، از خاطر بردن، به بوتهٔ فراموشی سپردن، نسیان
۱۸	کسوف	نداشت	گرفت خورشید، خورشیدگرفتگی، آفتاب‌گرفتگی، خورگیر، خورگرفت
۱۹	کل‌گرایی	کل‌نگری، کل‌گرایی	نداشت
۲۰	غزوه‌ها	جنگ‌های پیامبر، غزوات	نداشت

منابع

برنتی، سید محمدرضا (۱۳۹۹، ۲۸ فروردین). *داده‌کاوی – ۵ – شباهت کسینوسی: معیارهای شباهت*. بلاگ شخصی سیدمحمدرضا برنتی.
داده-کاوی-۵-شباهت-کسینوسی.https://www.berneti.ir

برنجیان، شاپوررضا و رئیسی، سارا (۱۳۹۳). *بازیابی کلمات معادل در سیستم های اطلاعاتی*. دومین همایش ملی پژوهش‌های کاربردی در علوم کامپیوتر و فناوری اطلاعات، تهران.
https://civilica.com/doc//٤٥٥٤٠١

حسابی، اکبر (۱۳۹۵). مقایسه‌ی روابط معنایی درون زبانی اسامی در فارس‌نت، یورونت و وردنت پرینستون. *جستارهای زبانی*، *۷* (٤): ۱٤۹-۱۷۳.

سلطانی، پوری (۱۳۸۵). سرعنوان‌های موضوعی فارسی. *دائره‌المعارف کتابداری و اطلاع‌رسانی* (ج ۲). تهران: سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران.

سلطانی، پوری، فانی، کامران و زهادی، فیروزان. (ویراستاران) (۱۳۹۷). *سرعنوان‌های موضوعی فارسی* (ویراست ٤). تهران: سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران. بازیابی از http://portal.nlai.ir/

فارس‌نت. (بی تا). بازیابی از:

http://farsnet.nlp.sbu.ac.ir/Site3/Modules/Public/Farsnet.jsp/

معرفی و تاریخچه اصفا. (بی‌تا). بازیابی از https://portals.nlai.ir/thesaurus/

References

Abdul Hassan, A. K., & Hadi, M. J. (2017). Sense-Based Information Retrieval Using Fuzzy Logic and Swarm Intelligence. *International Journal of Multimedia and Ubiquitous Engineering*, *12*(1): 363-376. http://dx.doi.org/10.14257/ijmue.2017.12.1.31

Asfa: Introduction and history (n.d.). https://portals.nlai.ir/thesaurus/ [In Persian]

Basile, Y. (2015). *WordNet as an Ontology for Generation* [Paper presentation]. WebNLG 2015 1st International Workshop on Natural Language Generation from the Semantic Web, June 2015, Nancy, France. hal-01195793

Berenjian, Sh. R., & Reissi, S. (2014). *Retrieving equivalent words in information systems*. The Second National Conference on Applied Research in Computer Science and Information Technology, Tehran. https://civilica.com/doc/455401. [In Persian]

Berneti, M. R. (2020, April, 17). *Data mining-5- Cosine Similarity: Similarity Criteria*. Seyed Mohammad Reza Berneti Personal Blog. https://www.berneti.ir/داده-کاوی-۵-شباهت-کسینوسی/ #. [In Persian]

Bharathi, G., & Venkatesan, D. (2012). Improving information retrieval using document clusters and semantic synonym extraction. *Journal of Theoretical and Applied Information Technology*, 36(2): 167- 172.

Farsnet (n.d.). http://farsnet.nlp.sbu.ac.ir/Site3/Modules/Public/Farsnet.jsp/ [In Persian]

Fellbaum, C. (Ed.). (1998). *WordNet. An Electronic Lexical Database*. MIT Press, MA. https://doi.org/10.7551/mitpress/7287.001.0001

Fernandez Lanza, S., Grana, J., & Sobrino, A. (2003). Introducing FDSA (Fuzzy Dictionary of Synonyms and Antonyms): applications on information retrieval and stand-alone use. *Mathematics & Soft Computing*, *10*(2): 57-70. Available at: https://raco.cat/index.php/Mathware/article/view/84890

Hesabi, A. (2016). A Comparison between Intra lingual Semantic Relations of Nouns in Fars Net, Euro Net and Princeton Word Net. *Language Related Research*, *7*(4): 149 -173. [In Persian]

https://portals.nlai.ir/thesaurus [In Persian]

Kelbessa, I.W. (2021). The effects of having lists of synonyms on the performance of Afaan Oromo Text Retrieval system. *ArXiv*, abs2103.02900

Li, Ch., Zhang, M., Bendersky, M., Deng, H., Metzler, D., & Najork, M. (2019). Multi-view Embedding-based Synonyms for Email Search Multi-view Embedding-based Synonyms for Email Search. In *Proceedings of SIGIR '19*, July 21–25, 2019, Paris, France (pp 575-584). https://doi.org/10.1145/3331184.3331250

Li, S. Li, B., Yao, H., Zhou, S., Zhu, J., & Zeng, Z. (2022). Completing WordNets with Sememe Knowledge. *Electronics*, *11*(79). DOI: 10.3390/electronics11010079

Li, Y., Hsu, B. J., & Zhai, Ch. X. (2013). Unsupervised identification of synonymous query intent templates for attribute intents. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (CIKM'13), San Francisco Ca, USA, 27 October 2013- 1 November 2013*(pp. 2029–2038). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2505515.2505694.

Miller, G. A. (1995). Wordnet: A lexical database for English. *Communication ACM*, *38*(11): 39-41. DOI: 10.1145/219717.219748

Miller, J., Beckwith, R., Fellbaum, C., Gross D., & Miller, K. (1990). *Introduction to Wordnet: An on-line Lexical Database. International Journal of Lexicography*, 3(4): 235-244. Nancy, France. Hal-01195793

Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Noor, P., Famian, A.R., Bagherbeigi, S., Fekri, E., & Monshizadeh, M. (2009). *Semi Automatic Development of FarsNet; the Persian WordNet*. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA). Available at: http://www.lrec-conf.org/proceedings/lrec2010/pdf/784_Paper.pdf

Shi, H. (2019). *A principaled approach to the evaluation of topic modeling algorithms*. [Doctoral dissertation Northwestern University, Illinois].

Retrieved from ProQuest Dissertations & Theses Global database. (UMI No. 13883392)

Soltani, P. (2006). Persian subject headings. *The encyclopedia of library and information science* (vol. II). Tehran: Iran National Library and Archives. [In Persian]Soltani, P., Fani, K., Zohadi, F. (Eds), & Azizian, N., (Assitant editor). (2018). *List of Persian Subject Headings* (4th Ed.). Tehran: National Library and Archives of Iran. [In Persian]

Soto, A., Olivas, J.A., & Prieto, M.E. (2008). Fuzzy Approach of Synonymy and Polysemy for Information Retrieval. In: R. Bello, R. Falcón, W. Pedrycz, and J. Kacprzyk, (eds) *Granular Computing: At the Junction of Rough Sets and Fuzzy Sets* (pp. 179-198) Springer, Berlin, Heidelberg. https://link.springer.com/chapter/10.1007/978-3-540-76973-6_12

Zeng, Q. T., Redd, D., Rindflesch, T., & Nebeker, J. (2012). Synonym, topic model and predicate-based query expansion for retrieving clinical documents. *AMIA Annual Symposium Proceedings, 2012, 3-7 November 2012, Chicago, Illinois, USA* (pp.1050-1059). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3540443/