Language Related Research
E-ISSN: 2383-0816
https://lrr.modares.ac.ir
https://doi.org/10.48311/LRR.16.6.2

Tarbiat Modares University

Vol. 16, No. 6, Tome 90
pp. 37-72
January & February
2026

Application of Distributional Semantics in the Qur'an, A Case Study of the Root-word "*Farah*"

Asieh Zouelm[1] & *Nosrat Nilsaz**[2]

Received: 22 September 2022
Received in revised form: 17 April 2023
Accepted: 10 May 2023

Abstract

Distributional semantics is a neoconstructionist method that focuses on the contextual use of words in real texts to determine the approximate meaning of a word relative to other words. Since one of the goals of applying semantics to the Qur'an is to ascertain the meaning of words according to their practical context, the use of this method in Qur'anic studies becomes important.In this research, in order to introduce the application of distributional semantics in the Qur'an, the descriptive-analytical method is employed to explain the implementation steps, the challenges, and the solutions to overcome them, by examining a case study.The steps are: determining comparable words for the main target word; determining distributional features (surah, equivalent to document, and phrase); linguistic pre-processing of the Qur'anic text and removal of stopwords; forming the context vector and co-occurrence matrix and weighting its elements; and finding and analyzing the semantic similarity of words.The most important challenges of using this method in the Qur'an are the small volume of the Qur'anic text, the lack of suitable software for calculations within the Qur'anic context, and the great difference in the length of surahs in the word-document pattern.Solutions to overcome some of these challenges include: paying closer attention to the basis of the distributional method (distributional hypothesis); avoiding the use of this method for very low-frequency words; and comparing the results obtained from the word-document and word-phrase patterns.For the root-word *Farah*, it is found that the meaning derived from the Qur'anic context (closer to the meaning of pride) is different from the meaning mentioned in standard dictionaries (joy).

Key words: Distributional Semantics, Context Vector, Semantical Similarity, Quran, Farah

[1]. Ph.D candidate, Department of Qur'an and Hadith sciences, Faculty of Humanities, Tarbiat Modares University, Tehran, Iran; ORCID ID: http://orcid.org/0000-0001-8061-5774

[2]. Corresponding author: Associate Professor, Department of Qur'an and Hadith sciences, Faculty of Humanities, Tarbiat Modares University, Tehran, Iran; *Email: nilsaz@modares.ac.ir,* ORCID ID: http://orcid.org/0000-0002-8580-5574

1. Introduction

Distributional semantics, as one of the neo-constructural methods in semantics, identifies the meaning of words by examining their use in natural texts.This method is based on the distributional hypothesis, which states: '*Word with similar distributional properties have similar meanings.*' Applying this method to the Quranic corpus is helpful in understanding the meaning of Quranic words by examining their usage context, especially in the case of words that are disputed by philologists. Thus, by comparing different words from the perspective of their distribution in the Quran (including co-occurrence with other words or use in different *Surahs*), and finding their distributional similarity, it identifies the scope of the meaning of the disputed word by obtaining the semantic distance between the words. Applying the distributional method in the Quran, due to the characteristics of the Arabic language and the Quranic corpus, requires localizing the method and overcoming its possible challenges. The method of implementing the method in the Quran is demonstrated practically using the case example of the root word "*Farah*", the meaning of which is disputed.

Research Question(s)

This research includes two main questions about the method of implementing distributional semantics in the Quran :

1- What are the practical steps for implementing distributional semantics in the Quran, considering its linguistic and textual properties ?

2- What are the challenges of applying distributional method in the Quran and how to overcome them ?

After that, in the practical implementation of the method in the Quran, another question is also raised :

3- What does the use of distributional semantics reveal about the meaning of the root word 'Farah' in the Quran?

2. Literature Review

Many works have been written in the field of distributional semantics, but we have not found a work in Persian that fully introduces this method and specifically implements it in the Quran. However, the translator of the two books *Machine learning for text* (Aggarwal, 2018) and *Networks* (Newman, 2018) into Persian (Ayub Turkian), added the appendices "Practical Text Mining Training" (Turkian, 2019a) and "Text Network" (Turkian, 2019b), respectively, which also examined the semantical similarity of verses of the Holy Quran and the clustering of nouns in the Quran. These two cases can be considered the closest to the subject of this research.

3. Methodology

The present study can be divided into two main parts. In the first part, which includes the explanation of how to implement distributional semantics in the Quran and the examination of its challenges, the research method is descriptive-analytical and the data collection method is documentary. In the second part, which is the implementation of distributional semantics on a case study in the Quran, first a number of words are selected for comparison with the main word "farah", which, based on the views presented in dictionaries, are in the two semantic domains of 'joy' and 'arrogance'. Then, two models of the distributional method are implemented for comparing words. In the word-phrase model, considering a 15-word context window (n=15), a number of phrases are selected as features and words are compared with each other, based on how they co-occur. In the word-document model, surahs are considered as documents (distributional features), and the occurrence of words in surahs is compared. Then, the results of implementing these two patterns of the distribution method on words are compared to obtain the final result.

4. Results

Distributional semantics is applied to the Quranic corpus with the aim of identifying the approximate meaning, i.e., determining the semantic domain of the disputed words. The implementation of this method in the present study includes the following steps:

1. Finding words to compare with the original word based on dictionaries;

2. Determining the distributional features, including documents (Surahs) and phrases that co-occur with the word (by grammatical or non-grammatical relations);

3. Linguistic pre-processing of words, including finding roots and effective forms of each root, disambiguation and merging words;

4. Removing stopwords;

5. Forming the co-occurrence matrix and its weighting (in the word-surah model with the TFIDF rule and in the word-phrase model with the PPMI rule);

6. Calculating the similarity of words by the cosine of the angle between the weighted context vectors;

7. Interpretation of the similarity between words using dictionaries and semantic relations between sentences.

The challenges of applying distributional semantics, especially in the Quran, and some ways to overcome them include:

1. General challenges of the distributional semantics, including the problem of automatic language preprocessing, interpretation of word similarity, and speaker intention recognition.

2. The challenges of applying distributional semantics to the Quran including the small size of the Quranic corpus for statistical calculations, the

lack of appropriate software for effective rooting, disambiguation, and recognizing stopwords, and the large variation in the length of surahs in the word-surah model. Some solutions to overcome the challenges include: paying more attention to the distributional hypothesis as the basis of the method, which, based on the wisdom of the speaker, results in greater credibility in the Quran; and correcting the error caused by the difference in the length of the documents in the word-surah model by comparing the results with the word-phrase model.

Applying the distributional method to the root-word Farah in the Quran shows the difference in the meaning obtained from the Quranic corpus (in the semantic domain of arrogance) compared to the dominant meaning mentioned in dictionaries, especially modern dictionaries (joy); that represents the semantic evolution of this word.

دوماهنامهٔ بین‌المللی
د۱۵، ش٦ (پیاپی ۹۰)، بهمن و اسفند ١٤٠٤، صص ۳۷-۷۲
مقاله پژوهشی
https://lrr.modares.ac.ir/article_7139.html

کاربست معنی‌شناسی توزیعی در قرآن، مطالعهٔ نمونهٔ موردی ریشه – واژهٔ «فرح»

آسیه ذوعلم[۱]، نصرت نیل ساز[۲*]

۱. دانشجوی دکتری رشته علوم قرآن و حدیث، دانشگاه تربیت مدرس، تهران، ایران
۲. دانشیار گروه علوم قرآن و حدیث، دانشگاه تربیت مدرس، تهران، ایران

تاریخ دریافت: ١٤٠١/٦/٣١ تاریخ پذیرش: ١٤٠٢/٢/٢٠

چکیده

معنی‌شناسی توزیعی، روشی نوساختگراست که برای شناخت معنی، به کاربرد واژگان در متون واقعی توجه دارد و به‌کارگیری آن، حدود معنایی یک واژه را در مقایسه با دیگر واژگان مشخص می‌کند. ازآنجاکه یکی از اهداف معنی‌شناسی قرآن، شناخت واژگان با توجه به بافت کاربردی آن‌هاست، بهره‌گیری از این روش در قرآن اهمیت می‌یابد. در این پژوهش به‌منظور معرفی چگونگی کاربست معنی‌شناسی توزیعی در قرآن، با روش توصیفی ـ تحلیلی، گام‌های پیاده‌سازی این روش در قرآن، چالش‌ها و راهکارهایی برای برون‌رفت از آن، با بررسی یک نمونهٔ موردی تشریح شده است. گام‌های اجرای روش در قرآن عبارت‌اند از: تعیین واژگانی برای مقایسه با واژهٔ اصلی، تعیین ویژگی‌های توزیعی (سورهٔ معادل سند، و عبارت)، پیش‌پردازش زبانی متن قرآنی و حذف ایست‌واژه‌ها، تشکیل بردار بافتی و ماتریس هم‌وقوعی و وزن‌دهی عناصر آن، اندازه‌گیری مشابهت معنایی واژگان و تحلیل آن. مهم‌ترین چالش‌های استفاده از روش توزیعی در قرآن، حجم کم پیکرهٔ قرآنی، عدم وجود نرم‌افزار مناسب جهت انجام محاسبات و تفاوت زیاد طول سوره‌ها در الگوی واژه ـ سند است. توجه بیشتر به مبنای روش توزیعی در متن قرآن، عدم استفاده از این روش برای واژگان بسیار کم‌بسامد و مقایسهٔ نتایج به‌دست‌آمده از الگوی واژه ـ سند و واژه ـ عبارت، راهکارهایی برای برون‌رفت از برخی چالش‌هاست. اجرای روش توزیعی برای ریشه ـ واژهٔ فرح در قرآن، تفاوت معنای به‌دست‌آمده از بافت قرآنی (نزدیک به حوزهٔ معنایی تکبر)، نسبت به معنای غالبی ذکرشده در لغت‌نامه‌ها (شادی) را نشان می‌دهد.

واژه‌های کلیدی: معنی‌شناسی توزیعی، بردار بافتی، مشابهت معنایی، قرآن، فرح.

---

* نویسندهٔ مسئول مقاله: Email: nilsaz@modares.ac.ir

۱.مقدمه

معنی‌شناسی توزیعی[۱] روشی نوساختگرا در حوزهٔ زبان‌شناسی پیکره‌ای[۲] است؛ که با الهام از برخی مبانی ساختگرایی مانند توجه به روابط هم‌نشینی و جانشینی و بسط آن شکل گرفته، اما در عمل از برخی مبانی آن، مانند تمایز کاربردشناسی از معنی‌شناسی، فاصله گرفته است. این روش به‌عنوان رویکردی کاربردمدار، با استفاده از داده‌های متون واقعی و به‌عبارتی «زبان طبیعی»،[۳] به شناخت معنی واژگان با توجه به بافت و نسبت به سایر واژه‌ها می‌پردازد.

یکی از اهداف معنی‌شناسی قرآن، شناخت معنی واژگان قرآنی با استفاده از بافت آن است. احتمال تطور معنایی واژگان از عصر نزول تا ظهور اولین کتاب‌های لغت عربی (حدود دو قرن بعد) و وجود اختلاف نظر میان لغت‌شناسان، ایجاب می‌کند که برای شناخت واژگان قرآن، از روش‌های غیرمستقیم شناخت معنی، مانند رویکردهای تاریخی یا مراجعه به سایر متون استفاده شود. معنی‌شناسی توزیعی یکی از این روش‌هاست؛ که با نگاهی آماری به چگونگی کاربرد واژگان در مقایسه با یکدیگر در متن، حدودی از معنای واژگان به‌دست می‌دهد.

پژوهش حاضر به‌منظور استفاده از معنی‌شناسی توزیعی در قرآن با هدف شناخت معنای واژگان با استفاده از بافت، با روش توصیفی ـ تحلیلی، در درجهٔ اول به دو سؤال اصلی پاسخ می‌دهد: گام‌های عملیاتی معنی‌شناسی توزیعی در قرآن، با توجه به ویژگی‌های زبانی و متنی آن چیست؟ در به‌کارگیری این روش در قرآن چه چالش‌هایی وجود دارد و راه‌های برون‌رفت از آن‌ها چیست؟ با توجه به هدف پژوهش، پیاده‌سازی معنی‌شناسی توزیعی در قرآن، با ذکر نمونهٔ موردی ریشه ـ واژهٔ «فرح» به‌طور دستی و غیررایانه‌ای صورت می‌گیرد. بنابراین سؤال دیگری که این پژوهش به صورت ضمنی درصدد پاسخ به آن است، چیستی معنای ریشه ـ واژهٔ فرح در قرآن است که درمورد آن میان لغت‌شناسان اختلاف نظر وجود دارد. فرضیهٔ مرتبط با سؤال اخیر این است که با استفاده از معنی‌شناسی توزیعی حوزهٔ معنایی فرح در قرآن، به آنچه مفسران و پژوهشگران قرآنی بیان کرده‌اند (معنایی نزدیک به تکبر) نزدیک‌تر است؛ تا آنچه لغت‌شناسان بدون لحاظ کاربردهای قرآن ذکر کرده‌اند (شادی). جمع‌آوری اطلاعات پژوهش کتابخانه‌ای است.

۲.پیشینهٔ پژوهش

در زبان فارسی آثار بسیاری با موضوع معنی‌شناسی با رویکردها و روش‌های مختلف عرضه شده است. اگرچه پژوهش‌های حوزهٔ معنی‌شناسی توزیعی در خارج از ایران، با نگرش زبان‌شناسانه، دامنه‌ای گسترده دارد، اما به‌دلیل نیاز این روش به ابزارهای ریاضیاتی و رایانه‌ای، در ایران به‌صورت محدودی با این نگاه به آن پرداخته شده است. برای نمونه می‌توان به بخشی از کتاب *نظریه‌های معنی‌شناسی واژگانی*[٤] از دیرک گیررتس (2010)، ترجمهٔ کوروش صفوی (۱۳۹۳) اشاره کرد که ضمن پرداختن به سیر تاریخی معنی‌شناسی، در میان روش‌های نوساختگرا به «تحلیل توزیع پیکره‌ای» می‌پردازد. از جملهٔ برخی آثار این حوزهٔ پژوهشی که از سوی متخصصان رایانه ارائه‌شده نیز می‌توان به کتاب‌های *متن‌کاوی یادگیری ماشین*[٥] (آگاروال، ۲۰۱۸، ترجمهٔ ترکیان، ۱۳۹۸) و *شبکه‌ها*[٦] (نیومن، ۲۰۱۸، ترجمهٔ ترکیان، ۱۳۹۸) اشاره کرد. مترجم این دو اثر ضمن ترجمه، به‌ترتیب به آن‌ها پیوست‌های «آموزش عملی متن‌کاوی» (ترکیان، ۱۳۹۸الف) و «شبکهٔ متن» (ترکیان، ۱۳۹۸ب) را افزوده، و ضمن آن به بررسی مشابهت معنایی آیات قرآن کریم و نیز خوشه‌بندی اسم‌ها در قرآن پرداخته است. اما اثری به زبان فارسی که روش توزیعی را به صورت کامل و با غلبهٔ نگرش زبان‌شناسانه معرفی کند یافت نشد.

در حوزهٔ پژوهش‌های قرآنی نیز، آثار فراوانی به روش‌های متعدد معنی‌شناسی در قرآن پرداخته‌اند. پژوهش‌های قرآنیِ سال‌های اخیر با هدف شناخت معنی واژگان قرآنی با استفاده از بافت، عمدتاً با نام معنی‌شناسی ساختگرا، یا ساختاری رواج یافته است. از نخستین آثار در معرفی این روش، پایان‌نامهٔ *معناشناسی تدبر در قرآن کریم* است؛ که ضمن تقسیم هم‌نشینان واژه به انواع مکملی، اشتدادی، تقابلی و توزیعی، جانشینان واژه‌ها را براساس هم‌نشینان مکملی شناسایی و به توصیف معنای واژه براساس آن‌ها می‌پردازد (سلمان‌نژاد، ۱۳۹۱). این روش در آثار دیگری نیز ازجمله پایان‌نامهٔ *معناشناسی کلمه در قرآن کریم* (شفیع‌زاده، ۱۳۹۱) و رسالهٔ *معناشناسی کرم در قرآن* (قاسم‌احمد، ۱۳۹٦) به کار رفته است و در پژوهش‌های مرتبط با شناخت معنا مانند بررسی ترجمه نیز، به نوعی استفاده از آن مشاهده می‌شود (گلی ملک‌آبادی و همکاران، ۱۳۹٦). در برخی پژوهش‌ها تصریح شده که این روش، بومی‌سازی روش‌های معنی‌شناسی در علم زبان‌شناسی است. به‌نظر می‌رسد روش رایجِ موسوم به معنی‌شناسی ساختگرا، از جهاتی برگرفته از معنی‌شناسی توزیعی است. اما نقدهایی به آن وارد است که بررسی آن مجال گسترده‌تری را می‌طلبد. اما به‌طور خلاصه می‌توان

به چند نکته اشاره کرد. اولاً چگونگی استناد کلیت این روش و نیز جزئیات گام‌های اجرایی آن به منابع اصلی مشخص نشده است. به‌عبارتی، بومی‌سازی یک روش، مستلزم شناخت اصل آن و سپس ایجاد تغییراتی متناسب با فضای مورد استفاده است؛ که در معرفی این روش، مشاهده نمی‌شود. همین امر سبب شده است که اجرای روش موسوم به ساختگرا، با نقدهایی منطقی روبه‌رو باشد. به‌علاوه مبنای روش نام‌برده، یعنی توجه به زبان کاربردی برای دریافت معنی، بر خلاف نام آن، اساساً با ساختگرایی در مفهوم اصیل تفاوت دارد.

نوآوری این پژوهش، پس از تبیین گام‌های معنی‌شناسی توزیعی با غلبهٔ نگرش زبان‌شناسانه، معرفی چگونگی به‌کارگیری آن ـ به عنوان روشی برای شناخت معنی از بافت ـ در قرآن، معرفی چالش‌های این کاربست و راه‌های برون‌رفت از آن، با ذکر نمونهٔ موردی ریشه ـ واژهٔ فرح است. گفتنی است ریشه ـ واژهٔ فرح پیش از این با روش موسوم به معنی‌شناسی ساختگرا در پایان‌نامهٔ *معناشناسی فرح در قرآن و روایات* (ذوعلم، ۱۳۹۷) و مقالهٔ «معناشناسی فرح در قرآن کریم؛ پاسخی به شبههٔ نکوهش شادی در قرآن» (ذوعلم و دیگران، ۱۳۹۹) بررسی شده است، اما فارغ از اعتبار نتایج پژوهش نام‌برده، در پژوهش حاضر، استفاده از روشی مستند به منابع اصلی و با رعایت گام‌های اصولی معنی‌شناسی توزیعی، در شناخت معنای این واژه مورد نظر است.

۳. مفاهیم نظری: معنی‌شناسی توزیعی و گام‌های اجرای آن

معنی‌شناسی توزیعی رویکردی کاربردمدار به زبان است که با توجه به بافت کاربردی، به شناخت معنی واژه در مقایسه با سایر واژه‌ها می‌پردازد (گیررتس، ۲۰۱۰، ترجمهٔ صفوی، ۱۳۹۸، صص. ۳٤٥ و ۳٤٦). نمونه‌هایی از کاربردهای این روش شاملِ شناخت معنای واژگان مورد اختلاف، ابهام‌زدایی از واژگان چندمعنا، و مقایسهٔ معنای یک واژه در دوره‌های زمانی مختلف می‌شود (Boleda, 2020, p.217-229).

ایدهٔ بنیادین در پس معنی‌شناسی توزیعی به زبان، فرضیهٔ توزیعی[۷] است که با توجه به روش توزیعیِ مطرح شده از سوی هریس[۸] در دههٔ ۱۹۵۰ به‌وجود آمده است. این فرضیه بیان می‌کند: «واژگان با ویژگی‌های توزیعی مشابه، معانی مشابه دارند» (Sahlgren, 2006, p. 21). بر این اساس اگر دو واژه، ویژگی[۹] توزیعی مشابه در متن داشته باشند، می‌توان گفت به یک دستهٔ واژگانی تعلق دارند (Harris, 1954, p. 156). بعدها دو نوع ویژگی توزیعی برای واژگان، شامل متن (سند[۱۰])هایی که واژه

در آن وقوع یافته و هم‌نشینان یک واژه معرفی شد. بر پایهٔ ایدهٔ هریس در هر نوع، تعدادی ویژگی توزیعی، مشخص شده، مشابهت واژگان براساس هم‌وقوعی[۱۱] هر واژه با این ویژگی‌ها سنجیده می‌شود. بر این اساس، یک مدل واژه ـ فضا، یا مدلی محاسباتی از معنی شکل می‌گیرد؛ که ویژگی‌های توزیعی، با ابعاد فضا بازنمایانده می‌شود. در این مدل، شباهت‌های میان واژگان، با مشاهدهٔ شواهد تجربی کاربرد واقعی زبان، استخراج می‌شود (Sahlgren, 2006, p. 21). هر چه تعداد ویژگی‌های توزیعی (ابعاد فضای واژه‌ای) بیشتر باشد، پیچیدگی‌های بیشتری از معنای واژه‌ها بازنمایانده می‌شود (Schutze, 1993, p. 896).

در این قسمت گام‌های اجرایی معنی‌شناسی توزیعی معرفی می‌شود.

۳ـ۱.تشکیل پیکره

اولین گام روش توزیعی تشکیل پیکره به‌معنای مجموعه متونی با ویژگی‌های موقعیتی مشخص است. ویژگی‌های موقعیتی شامل مشارکت‌کنندگان (گوینده، نویسنده، مخاطبان) و روابط میان آن‌ها، شیوهٔ ارتباط (نوشته، سخنرانی و ...)، شرایط تولید (وجود یا عدم برنامه‌ریزی، ویراستاری و ...)، وضعیت (فضا و زمان)، اهداف ارتباط و موضوع می‌شود (کرافورد و سزومی، ۲۰۱٦، ترجمهٔ صفوی، ۱۳۹۹، صص. ۲٤ و ۲۵). بسته به هدف پژوهش و براساس میزان پرسش‌ها، پیکره‌ها عمومی یا تخصصی هستند (کرافورد و سزومی، ۲۰۱٦، ترجمهٔ صفوی، ۱۳۹۹، صص. ۹۳ و ۹٤). تعداد واژگان برخی پیکره‌های عمومی، از ۱.۹ میلیارد تا ۵۰ میلیون واژه گزارش شده است (کرافورد و سزومی، ۲۰۱٦، ترجمهٔ صفوی، ۱۳۹۹، ص. ۷۲). اما پیکره‌های تخصصی می‌تواند تعداد واژگانی از درجهٔ میلیون واژه داشته باشد.[۱۲]

۳ـ۲.تعیین ویژگی‌های توزیعی واژه‌ها

دو دسته ویژگی توزیعی واژه، سندها و واژگان هم‌وقوع با یک واژه هستند؛ که دو الگوی مستقل «واژه ـ سند» و «واژه ـ واژه» به‌دست می‌دهند. دستهٔ اول ویژگی‌ها، علاوه بر سندهایی مانند مقاله می‌تواند به بافت‌های کوتاه‌تری مانند بخشی از یک مقاله یا یک بند در متن تعمیم یابد (Turney, 2010, p. 148). دستهٔ دوم ویژگی‌ها نیز می‌تواند از واژه، به عبارت[۱۳] (مانند ریشه یا عبارت‌های چندواژه‌ای) گسترش یافته، الگوی واژه ـ عبارت را تشکیل دهد (Sahlgren, 2006, p. 31) و با لحاظ رابطهٔ دستوری با واژه

و یا بدون لحاظ این رابطه تعیین شود (Lenci, 2018, p. 158). اگر رابطهٔ دستوری در این الگو درنظر گرفته نشود، قرار گرفتن در محدودهٔ جمله، یا در زنجیره‌ای با اندازهٔ مشخص از واژه‌ها و به‌عبارتی «پنجرهٔ بافتی»[۱۴] اهمیت دارد (Schutze, 1992, pp. 787, 793)؛[۱۵] که البته هیچ اصل نظری برای مشخص کردن بهترین اندازهٔ پنجره وجود ندارد (Lenci, 2018, p. 153).

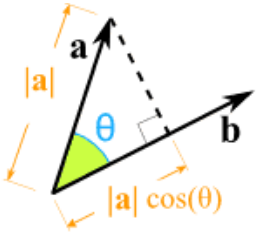
در الگوی واژه ـ سند، مشابهت واژگان براساس هم‌نشینی آن‌ها در سندهای مختلف بررسی می‌شود. اما در الگوی واژه ـ عبارت، که ویژگی‌های توزیعی، هم‌نشینان واژه هستند، مشابهت جانشینان واژه با آن سنجیده می‌شود.

۳ـ۳. تشکیل بردار بافتی[۱۶] واژه و ماتریس هم‌وقوعی[۱۷]

«بُردار»[۱۸] در تعریف پاره‌خطی جهت‌دار است که در فضای n بُعدی، به‌صورت مجموعه‌ای از مؤلفه‌ها، نقطه‌ای از فضا را تعریف می‌کند. بردار بافتی واژه، به‌صورت مجموعه‌ای از بسامدهای هم‌وقوعی واژه با ویژگی‌های توزیعی تعریف می‌شود. هر مؤلفهٔ این بردار در الگوی واژه ـ سند، تعداد وقوع واژه در یک سند؛ و در الگوی واژه ـ عبارت، بسامد هم‌وقوعی واژه با یک عبارت است (Sahlgren, 2006, pp. 29, 30). مجموعهٔ مؤلفه‌های بردارهای بافتی واژه‌ها در متن، به‌صورت جدولی با محور عمودی واژه‌ها و محور افقی ویژگی‌ها، «ماتریس هم‌وقوعی» را تشکیل می‌دهد.

۳ـ۴. محاسبهٔ مشابهت[۱۹] واژگان

مشابهت معنایی واژه‌ها با مقایسهٔ بردارهای بافتی آن‌ها تعیین می‌شود؛ که کاربردی‌ترین معیار آن، تعیین کوسینوس زاویهٔ میان بردارهاست (Schutze, 1992, p. 787). هرچه این زاویه کم‌تر باشد، کوسینوس آن نزدیک‌تر به یک، و مشابهت دو بردار بیشتر است و اگر دو بردار عمود (بدون مؤلفهٔ مشترک) باشند، این مقدار صفر است.



شکل ۱: ضرب داخلی دو بردار

Figure 1: Inner product of two vectors

کوسینوس زاویهٔ دو بردار، از تقسیم ضرب داخلی آن‌ها (جمع حاصل‌ضرب مؤلفه‌های متناظر)، بر ضرب اندازه‌های آن‌ها[۲۰] تعریف می‌شود.

$$a = (a_1. a_2. \dots)$$
$$b = (b_1. b_2. \dots)$$
$$a \cdot b = a_1 b_1 + a_2 b_2 + \cdots$$
$$\cos\theta = \frac{a \cdot b}{|a| \cdot |b|} = \frac{a_1 b_1 + a_2 b_2 + \cdots}{\sqrt{(a_1^2 + a_2^2 + \cdots) \cdot (b_1^2 + b_2^2 + \cdots)}}$$

و مشابهت بردارهای a و b از این رابطه به‌دست می‌آید:

$$sim(a. b) = \cos\theta$$

۳ـ۵. گام‌های اصلاحی اجرای روش توزیعی

برای بهبود عملکرد روش توزیعی از جهت سهولت اجرا و نتیجه‌گیری مؤثرتر، اقداماتی اصلاحی صورت می‌گیرد که عبارت‌اند از:

۳ـ۵ـ۱.پیش‌پردازش[۲۱] زبانی

این عمل به‌منظور به‌کارگیری مؤثرتر عناصر ماتریس هم‌وقوعی، بر پایهٔ اطلاعات زبانی بر روی واژگان صورت می‌گیرد (Sahlgren, 2006, p. 71, 72). نمونه‌هایی از این پیش‌پردازش، ریشه‌یابی[۲۲] واژگان (Ibid, p. 12)، برچسب زدن پاره گفتار[۲۳] (به‌معنی تعیین مقوله‌های واژگان) و ابهام‌زدایی[۲۴] از واژگان چندمعناست.

۳ـ۵ـ۲. کاهش بُعد ابتدایی ماتریس هم‌وقوعی

ماتریس هم‌وقوعی در شکل اولیه، با ابعاد بسیار زیاد، تمام واژگان را بر حسب تمام ویژگی‌ها بیان می‌کند. اما بسیاری از ویژگی‌ها، در تعیین مشابهت واژگان تأثیر اندکی دارند. کاهش بُعد، بازنمایی اطلاعات مفید در فضایی با ابعاد کوچک‌تر است؛ که در ابتدا با اطلاعات زبانی صورت می‌گیرد. حذف واژه‌های متعلق به مقوله‌های دستوری بسته،[۲۵] و یا دارای خواص آماری نامطلوب (مانند بسیار پربسامد یا کم‌بسامد)[۲۶] از آن جمله است. هم‌وقوعی واژگان بسیار پربسامد ـ که معمولاً همان واژه‌های متعلق به

مقوله‌های دستوری بسته هستند ـ با واژهٔ مورد بررسی، اطلاعات معنایی مفیدی دربر ندارد و از این رو ایست‌واژه[۲۷] نامیده می‌شوند (Sahlgren, 2006, pp. 38, 39).

۳ـ۵ـ۳. وزن‌دهی[۲۸] عناصر ماتریس هم‌وقوعی

میزان ارزش عناصر مختلف ماتریس هم‌وقوعی، برای شناخت معنی، مساوی نیست. از این رو با انجام عملیاتی روی بسامدهای هم‌وقوعی اولیه (بسامد خام[۲۹])، هر یک از آن‌ها ارزش‌گذاری می‌شود (Turney, 2010, p.156).

در الگوی واژه ـ سند، هرچه طول سندِ شاملِ واژه کوتاه‌تر باشد، و نیز هرچه واژه در سندهای کم‌تری وقوع یافته باشد، هم‌وقوعی سند و واژه اهمیت بیشتری برای یافتن مشابهت با دیگر واژگان دارد. بر همین اساس نوعی از وزن‌دهی (خانوادهٔ TFIDF) بر روی ماتریس پیاده می‌شود. در نمونه‌ای از این وزن‌دهی، بسامد خام هم‌وقوعی واژهٔ i با سند j (TF_{ij}[۳۰]) در دو ضریب IDF_i (تابعی از معکوس تعداد سندهایی دارای عبارت i)؛[۳۱] و S_j[۳۲] (معکوس تعداد واژگان سند j) ضرب می‌شود و هر عنصر از ماتریس $f_{ij}=TF_{ij}.IDF_i.S_j$ به دست می‌آید (Sahlgren, 2006, pp. 31, 32).

در الگوی واژه ـ عبارت، اگر بسامد عبارت هم‌وقوع با واژهٔ مورد نظر، در کل متن زیاد نباشد، هم‌وقوعی آن با واژه مهم‌تر است. برای وزن‌دهی، احتمال هم‌وقوع شدن مورد انتظارِ واژه و عبارت (p_{exp})[۳۳] (براساس بسامد هر یک از آن‌ها در کل متن)، با بسامد هم‌وقوعی مشاهده شدهٔ این دو در متن (p_{obs})[۳۴]، مقایسه می‌شود. اگر واژه و عبارتی از لحاظ آماری مستقل باشند، هم‌وقوع شدن آن‌ها کاملاً تصادفی است و انتظار می‌رود احتمال هم‌وقوعی آن‌ها، حاصل‌ضرب احتمال وقوع هریک در متن باشد (بسامد مورد انتظار). وزن‌دهی باید به گونه‌ای باشد که اگر بسامد هم‌وقوعی مشاهده شده، برابر با بسامد مورد انتظار باشد، این عنصر تأثیری در تعیین معنی واژه نداشته، عنصر متناظر در ماتریس صفر شود (Turney, 2010, p.158). مدل‌های مختلف وزن‌دهی ماتریس هم‌وقوعی بر همین اساس معرفی شده؛ که در نمونه‌ای از آن، موسوم به قاعدهٔ PPMI، هر عنصر از ماتریس هم‌وقوعی این‌گونه تعریف می‌شود:

$$\begin{cases} \log_2\left(\frac{p_{obs}}{p_{exp}}\right) & \text{اگر } p_{obs} > p_{exp} \\ 0 & \text{اگر } p_{obs} \leq p_{exp} \end{cases}$$

۳ـ۵ـ۴. کاهش بُعد پیشرفته در ماتریس هم‌وقوعی

ماتریس هم‌وقوعی به‌طور طبیعی عناصر بسیاری با مقدار صفر دارد؛ بعضی ویژگی‌ها اطلاعات تکراری دارند؛ و ویژگی‌های توزیعی پنهانی وجود دارند که با لحاظ ویژگی‌های صرفاً ظاهری، از دست می‌روند (Lenci, 2018, p.156). بنابراین از روش‌های پیشرفته‌ای برای کاهش بُعد، مانند تجزیهٔ مقدار تکین[۳۵] (SVD) استفاده می‌شود؛ که با عملیاتی ریاضی، داده‌های کم‌اهمیت ماتریس را حذف و ویژگی‌های مؤثر پنهان را بازمی‌نمایاند (Lenci, 2018, p.156).

۳ـ۶. ترسیم نقشهٔ معنایی[۳۶]

نقشهٔ معنایی، شامل تصویر واژگان مورد مقایسه در یک صفحه است، که با نمایش دوری و نزدیکی واژگان نسبت به هم می‌تواند تا حد خوبی، حوزه‌های معنایی واژگان را از یکدیگر متمایز سازد. هر چه دو واژه در این نقشه به هم نزدیک‌تر (همسایه[۳۷]) باشند، مشابهت معنایی بیشتری دارند.

۳ـ۷. تفسیر مشابهت بردارهای بافتی واژگان

نتیجهٔ اصلی روش توزیعی، مدل کردن شباهت‌های معنایی است. اما «مشابهت معنایی»، خود یک مفهوم مبهم است؛ که شامل انواع روابط واژگانی مانند شمول معنایی، تقابل، جزءواژگی، روابط مکانی و دیگر روابط دسته‌بندی نشده[۳۸] می‌شود (Lenci, 2018, p.161). شناخت این رابطه بخش مهمی از معنی‌شناسی توزیعی است. در این میان رابطهٔ «تقابل» یکی از چالش‌های مهم است؛ که تمایل به وقوع در ویژگی‌هایی مشابهِ «هم‌معنایی» دارد. یکی از راه‌های تمیز آن از هم‌معنایی، استفاده از مدل‌های ترکیبی است که از اطلاعات منابع واژگانی نیز بهره می‌گیرد (Lenci, 2018, p.162).

۴.به‌کارگیری معنی‌شناسی توزیعی در قرآن، بررسی نمونهٔ موردی ریشه ـ واژهٔ «فرح»

آغاز به‌کارگیری روش توزیعی، تشکیل پیکره با جمع‌آوری متون طبیعیِ دارای ویژگی‌های بافتی و موقعیتی مشترک است. ویژگی‌های موقعیتی مشترک در قرآن شامل این موارد است: مشارکت‌کنندگان و رابطهٔ میان آن‌ها (گوینده خداوند، و مخاطب پیامبر و مردم)، شیوهٔ ارتباط (وحی)، شرایط تولید

(انتقال کامل پیام و تنظیم متن با ترتیب توقیفی)، هدف ارتباط (هدایت انسان‌ها) و وضعیت ارتباط (بازهٔ زمانی و مکان مشخص). بر این اساس متن قرآن می‌تواند به عنوان پیکره‌ای طبیعی با رویکرد توزیعی مورد بررسی قرار گیرد.

شناخت معنی واژه با استفاده از اطلاعات توزیع واژه در پیکرهٔ قرآنی، عملاً برای واژه‌های مورد اختلاف معنایی سودمند است. به‌علاوه، ازآنجاکه نتیجهٔ معنی‌شناسی توزیعی، شناخت حدودی معنی واژگان در مقایسه با یکدیگر است و تفاوت‌های ظریف معنایی در این روش کمتر مورد توجه قرار می‌گیرد، برای شناخت واژگانی کاربرد دارد که درمورد معنای آن‌ها اختلاف نظر کلی وجود دارد و به عبارتی حوزهٔ معنایی آن‌ها مورد اختلاف است. ویژگی‌های متن قرآن و نیز زبان عربی سبب می‌شود در کاربست قرآنی معنی‌شناسی توزیعی نکاتی مورد توجه قرار گیرد. در این بخش گام‌های پیاده‌سازی روش توزیعی در قرآن، به ترتیب اجرا، با نمونهٔ موردی ریشه ـ واژهٔ «فرح» بررسی می‌شود؛ که در قرآن ۲۲ بار به دو شکلِ فعل ثلاثی مجرد (۱۶ بار) و صفت مشبهه فَرِح (شش بار) به صورت مفرد یا جمع به‌کار رفته است (عبدالباقی، ۱۳۹۳، صص. ٦٦٢، ٦٦٣).

۴ـ۱. انتخاب واژگانی برای سنجش مشابهت با واژهٔ اصلی

از آنجا که در روش غیر رایانه‌ای، عملاً بررسی مشابهت تمام واژگان قرآنی با یک واژه امکان‌پذیر نیست، تعداد محدودی واژه به این منظور انتخاب می‌شود. برای واژه‌ای که درمورد معنی آن اختلاف وجود دارد، معانی ذکرشده در لغت‌نامه‌ها و تفاسیر، و واژگان مشمول حوزهٔ معنایی آن‌ها، می‌تواند برای مقایسه با واژهٔ اصلی گزینش شوند.

لغت‌شناسان برای واژهٔ فَرَح، دو معنای اصلی بیان کرده‌اند: ۱ ـ سُرور (ابن درید، ۱۹۸۸، ج۱، ص. ۵۱۸؛ ازهری، ١٤٢١ق، ج۵، ص. ۱۵؛ جوهری، ١٣٧٦ق، ج۱، ص. ۳۹۰؛ آذرنوش، ۱۳۸۳، ص. ۵۰۰)؛ ۲ ـ تکبر و واژگان حوزهٔ معنایی آن (ابوعبیده، ۱۳۸۱ق، ج۲، ص. ۱۱۱؛ ابن قتیبه، ١٤١١ق، ص. ٢٨٦؛ جوهری، ١٣٧٦ق، طبرسی، ۱۳۷۲، ج۷، ص. ٤١٦). بنابراین واژگان حوزه‌های معنایی سرور (مانند استبشار، ضحک، رضایت) و تکبر (مانند مرح، فخر، اختیال، علو، بغی) و نیز واژگان و اصطلاحات نزدیک به این حوزه‌ها، برای مقایسه با فرح مورد توجه هستند. شایان ذکر است با دقت در ارتباط میان معنای ذکرشده و نوع لغت‌نامه، به نظر می‌رسد لغت‌شناسان، و به ویژه در دورهٔ معاصر بیشتر به‌معنای

سرور برای فرح تأکید دارند و معنای تکبر غالباً از سوی مفسران یا لغت‌شناسانی بیان شده است، که کاربردهای قرآنی را مبنای شناخت معنا دانسته‌اند.

۴ـ۲.تعیین نوع ویژگی توزیعی

در الگوی واژه ـ سند، چگونگی مطابقت بخش‌های قرآن با سندها اهمیت دارد. در پیکره‌های زبانی مورد بررسی در روش توزیعی، سندها متن‌هایی جداگانه دارای ماهیتی نسبتاً مستقل‌اند؛ به‌عنوان نمونه مقالات یک نشریه (نک: Dumais et.al,1988). در قرآن اگرچه برخی، واحد «آیه» را به‌عنوان سند برگزیده‌اند (ترکیان، ۱۳۹۸الف، ص. ٦۲۳)، اما در بسیاری موارد آیه‌ها، وابسته به هم و حتی گاهی مکمل یکدیگر برای تشکیل جمله‌اند.[۳۹] با توجه به قرارگیری توقیفی آیات در واحدهای مستقل سوره، سوره به‌عنوان سند، گزینهٔ مناسب‌تری به نظر می‌رسد.

علاوه بر مفهوم اصطلاحی سوره که بخش‌های متمایز آیات (عمدتاً همراه با بسمله) است، برخی پژوهشگران «رکوعات قرآنی» را با عنوان «سوره‌وار»، برای گونه‌ای از دسته‌بندی آیات با واحدهایی مستقل مشابه سوره، از زمان پیامبر (ص) معرفی می‌کنند. در این دسته‌بندی مجموعه آیات قرآن به ۵۵۵ رکوع تقسیم می‌شود که در برخی نسخه‌های قدیمی با حرف «ع» از هم جدا شده‌اند (لسانی فشارکی و مرادی زنجانی، ۱۳۹۵، صص. ۸۸ ـ ۹٦). هرچند با دقت در محدودهٔ رکوعات قرآنی (لسانی فشارکی و مرادی زنجانی، ۱۳۹۵، صص. ۲۲۸ ـ ۲۳۲) به نظر می‌رسد هدف از این دسته‌بندی، سهولت در قرائت و حفظ آیات، و نه دسته‌بندی موضوعی است، اما به‌هرحال سبب تشکیل سندهایی بسیار کوچک‌تر از سوره و پررنگ شدن نقش هم‌وقوعی دو واژهٔ نزدیک در متن به‌ویژه برای سوره‌های طولانی می‌شود. این روش را می‌توان معادل تعمیم سندها به بخش‌ها یا بندهایی از یک مقاله در این الگو (Turney, 2010, p. 148) دانست.

در الگوی واژه ـ عبارت، دو نوع هم‌وقوعی واژگان می‌تواند مورد توجه باشد: قرارگیری واژه‌ها در یک پنجرهٔ بافتی، یا رابطهٔ نحوی میان واژگان. در نوع اخیر می‌توان علاوه بر روابط نحوی تعریف شده (مانند رابطهٔ «فعل و فاعل»)، رابطه‌های دیگری را نیز، براساس ارتباط معنایی جملات (مانند شرط و جواب شرط) در نظر گرفت. شناخت این نوع هم‌وقوعی به دلیل لزوم توجه به معنا دشوارتر از نوع اول است.

در این پژوهش برای معنی‌شناسی ریشه ـ واژهٔ فرح در قرآن، الگوی واژه ـ عبارت (بدون لحاظ

روابط دستوری) و الگوی واژه ـ سند (معادل سوره) مورد توجه قرار می‌گیرد.

۴ـ۳. پیش‌پردازش زبانی متن

در ریشه‌یابی واژگان قرآنی برای قرار گرفتن در ماتریس هم‌وقوعی، ریشه‌محور بودن ساخت واژگان در زبان عربی نقش مهمی دارد. وجود باب‌ها و صیغه‌های متعدد از یک ریشه، معانی متفاوتی ایجاد می‌کند که گاهی تنها برخی از آن‌ها مورد نظر است. به‌عنوان مثال برای بررسی مشابهت واژگان حوزهٔ معنایی سرور با فرح، مصدر«استبشار» به‌معنای «سرور به‌سبب بشارت» (عسکری، ۱٤۰۰ق، ص. ۲٦۰) و واژگان مرتبط با این مصدر مورد توجه قرار می‌گیرد و سایر باب‌های ریشهٔ بشر به‌دلیل ارتباط اندک معنایی بررسی نمی‌شود. از ریشهٔ کبر نیز، مصدرهای «تکبر» و «استکبار»، با اندکی اختلاف معنایی به‌معنای «خودبزرگ‌بینی» (نک: ایزوتسو، ۱۳۷۸، ص. ۲۸۹)، مورد نظر است و از سایر واژگان این ریشه صرف‌نظر می‌شود. در ویژگی‌های توزیعی فرح در ماتریس واژه ـ عبارت نیز، از میان واژه های ریشهٔ فضل، اسم مصدر «فَضل» به‌دلیل بار معنایی ویژه‌ای که آن را از صیغه‌های دیگر این ریشه متمایز می کند، انتخاب می‌شود. در توجه به یک صیغهٔ ویژه از یک ریشه، برخی معتقدند ریشه‌یابی اسم‌ها در قرآن تنها تبدیل جمع به مفرد است و شامل یافتن حروف اصلی واژه نمی‌شود (ترکیان، ۳۹۸ الف، ص. ٦۱۹ و ٦۲۷). با این حال در بسیاری موارد، کاربرد صیغه‌های فعلی و اسمی یک ریشه، تفاوت عمده‌ای ندارد.[٤۰] به‌طور کلی بسته به مورد پژوهش، توجه به کل واژگان ریشه و یا لحاظ واژه یا واژه‌هایی از یک ریشه، می‌تواند مقبول باشد.

در ابهام‌زدایی، واژگانی با ظاهر مشابه اما با معانی مختلف[٤۱] از هم متمایز؛ و به تناسب نیاز تنها برخی در ماتریس هم‌وقوعی بررسی می‌شوند. از واژه‌های مورد مقایسه با فرح، ریشهٔ «سرر» در قرآن دارای دو معنای «پنهان کردن» و «شادی» است؛ که تنها معنای دوم، با فرح مقایسه می‌شود. ریشهٔ «بَغی» نیز در شکل ثلاثی مجرد در اصل به‌معنای طلب و در بسیاری کاربردها «طلب تجاوز از میانه‌روی» است؛ که به معانی دیگری مانند ظلم، زنا، حسد و تکبر گسترش یافته (فراهیدی، ۱٤۰۹ق، ج٤، ص. ٤٥۳؛ ازهری، ۱٤۲۱ق، ج۸، ص. ۱۷۹؛ راغب، ۱٤۱۲ق، ج۱، صص. ۱۳٦ و ۱۳۷) و به‌علاوه گاهی به‌معنای «خواستن و جست‌وجو کردن»[٤۲] (متعدی) و گاه به‌معنای «چیزی را به صورتی خواستن»[٤۳] (دومفعولی) به‌کار رفته است. برخی از این معانی در معنی‌شناسی فرح حذف می‌شوند. در میان

ویژگی‌های توزیعی فرح در الگوی واژه ـ عبارت نیز، واژهٔ «سیئه» تنها به‌معنای «اتفاق ناگوار» و نه به‌معنای «عمل بد» مورد توجه است.

علاوه بر ریشه‌یابی و ابهام‌زدایی، ادغام واژه‌های هم‌معنا یا قریب‌المعنا نیز ازجمله موارد پیش‌پردازش زبانی پیکره است. در ماتریس هم‌وقوعی فرح، می‌توان واژه‌های جهنم و عذاب، و نیز واژه‌های اِعراض، تَوَلی، اِدبار و تصعیر خَد (به معنی رویگردانی) را ادغام کرد.

۴ـ۴. کاهش بُعد ابتدایی ماتریس هم‌وقوعی

کاهش بُعد ابتدایی، شامل حذف ایست‌واژه‌هاست که به دلیل هم‌وقوعی با بسیاری واژگان، هم‌نشینی آن‌ها، اطلاعات چندانی برای شناخت معنا در بر ندارد. واژه‌های متعلق به مقوله‌های دستوری بسته، مانند حروف، ضمائر و افعال ناقصه از این جمله اند. در برخی بررسی‌های زبان‌شناسانهٔ قرآن، تعدادی از این ایست‌واژه‌ها (با عنوان «واژه‌های توقف») ذکر شده که برخی ترکیب دو یا چند ایست‌واژه‌اند (ترکیان، ۱۳۹۸ب، ص. ٦۲۳).

جدول ۱: تعدادی از ایست‌واژه ها در قرآن

Table 1: some of the stopwords in the Qur'an

نمونه واژه‌های توقف								
الا	إنهم	علينا	فلا	كل	لنا	منها	وعلى	على
الذي	أم	عليه	فهم	كله	له	منهم	ولا	والذي
الذين	أولئك	عليهم	في	كلهم	له	منهما	ولكن	والذين
إذ	أيها	عند	فيه	كما	لهم	هم	وما	وإذ
إذا	بما	عندنا	ق	كنتم	لهم	هما	ومما	وإذا
إن	به	عنده	قال	لا	ما	هو	وهذا	ولا
إنا	بهم	عندي	قالوا	لك	مما	هي	وهذه	ولهم
إنما	بين	عنه	قل	لك	من	والذي	وهم	إلي
إننا	ثم	عنها	قول	لكم	منا	والذين	يا	إليك
إنه	علي	عنهم	قيل	لم	منه	وإن	ذلك	

۴ـ۵.تشکیل ماتریس هم‌وقوعی و وزن‌دهی عناصر آن

تشکیل ماتریس هم‌وقوعی، در دو الگوی ذکرشده، پس از پیش‌پردازش زبانی و حذف ایست‌واژه‌ها، با شمارش هم‌وقوعی‌ها صورت می‌گیرد. در ماتریس واژه ـ عبارت برای ریشهٔ فرح، پنجرهٔ هم‌وقوعی ۱۵واژه‌ای پیشنهاد می‌شود. پس از آن وزن‌دهی عناصر این دو ماتریس مورد نظر است.

۴ـ۵ـ۱. ماتریس واژه ـ سوره

در وزن‌دهی TFIDF در ماتریس واژه ـ سند (در قرآن واژه ـ سوره)، پس از به‌دست آوردنِ بسامد هم‌وقوعی واژه در سوره‌ها، دو ضریب اعمال می‌شود: ۱ ـ $IDF_{\text{واژه}}$، که نشان می‌دهد که هرچه واژه یا عبارتی در سوره‌های کم‌تری تکرار شده باشد از نظر زبان‌شناسی اهمیت بیشتری دارد؛[۴۴] ۲ ـ $S_{\text{سوره}}$، که معکوس طول سوره[۴۵] است و نشان می‌دهد هم‌وقوعی دو واژه در سورهٔ کوتاه‌تر، بیانگر مشابهت بیشتر آن‌هاست.

درمجموع عنصرِ متناظر با یک سوره و یک واژه در ماتریس هم‌وقوعی، چنین به‌دست می‌آید:

$$\text{TFIDF} = \frac{\text{تعدادوقوع‌واژه‌درسوره}}{\text{طول‌سوره}} \times \log_2 \left(\frac{\text{تعدادکل‌سوره‌ها}}{\text{تعدادسوره‌هایی‌که‌واژه‌درآن‌آمده}} \right)$$

برای نمونه ابتدا ماتریس هم‌وقوعی واژه ـ سوره فرح، برای مقایسه با واژه‌های فخر، اختیال، مرح و سرور، بدون اعمال ضریب نشان داده می‌شود. در زیرِ نام هر سوره، تعداد واژگان آن (فرهنگ نامهٔ علوم قرآنی، ۱۳۹٤، صص. ۲٦۲۱ ـ ۲۷۳۳ و ۲۷۵۱)؛ و در مقابلِ نام واژه‌ها تعداد سوره‌هایی که این واژه‌ها در آن‌ها وقوع یافته (نک: عبدالباقی، ۱۳۹۳ش) درج شده است.

جدول ۲: ماتریس واژه ـ سورهٔ فرح با بسامدهای خام

Table 2: Word-Sura matrix of Farah with raw frequencies

	بقره ٦٢٢١	آل‌عمران ٣٤٨٠	نساء ٣٧٤٥	انعام ٣٨٥٠	توبه ٤٠٩٨	یونس ١٨٣١	هود ١٧١٥	رعد ٨٥٥	اسراء ١٥٣١	مؤمنون ١٨٤٠	نمل ١١٤٩	قصص ١٤٤١	روم ٨١٩	لقمان ٥٤٢	غافر ١١٩٩	شوری ٨٦٦	حدید ٥٤٤	انسان ٢٤٠	انشقاق ١٠٩
فرح (١٣)	٠	٣	٠	١	٢	٢	١	٢	٠	١	١	٢	٣	٠	٢	١	١	٠	٠
فخر (٤)	٠	٠	١	٠	٠	٠	١	٠	٠	٠	٠	٠	٠	١	٠	٠	٢	٠	٠
اختیال (٣)	٠	٠	١	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	١	٠	٠	١	٠	٠
مرح (٣)	٠	٠	٠	٠	٠	٠	٠	٠	١	٠	٠	٠	٠	١	١	٠	٠	٠	٠

	بقره ٦٢٢١	آل‌عمران ٣٤٨٠	نساء ٣٧٤٥	انعام ٣٨٥٠	توبه ٤٠٩٨	یونس ١٨٣١	هود ١٧١٥	رعد ٨٥٥	اسراء ١٥٣١	مؤمنون ١٨٤٠	نمل ١١٤٩	قصص ١٤٤١	روم ٨١٩	لقمان ٥٤٢	غافر ١١٩٩	شوری ٨٦٦	حدید ٥٤٤	انسان ٢٤٠	انشقاق ١٠٩
سرور (٥)	١	١	٠	٠	١	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	١	٢

ماتریس وزن‌دهی شده چنین به‌دست می‌آید (در زیرِ نام هر سوره، طول آن[٤٦] و در مقابلِ نام واژه‌های مورد مقایسه با فرح، عامل دوم در فرمول وزن‌دهی، $\log_2\left(\frac{114}{\text{تعدادسوره‌هایی‌که‌واژه‌درآن‌آمده}}\right)$ ذکر شده است):

جدول ۳: ماتریس واژه ـ سوره فرح وزن‌دهی‌شده

Table 3: Weighted Word- Sura matrix of Farah

	بقره ١١٤	آل‌عمران ٦٤	نساء ٦٩	انعام ٧١	توبه ٧٥	یونس ٣٤	هود ٣٢	رعد ١٦	اسراء ٢٨	مؤمنون ٣٤	نمل ٢١	قصص ٢٦	روم ١٠٩	لقمان ١	غافر ٢٢	شوری ١٦	حدید ١	انسان ٤٠	انشقاق ٢
فرح (٣.١)	٠	١.٤٥	٠	٠.٤٤	٠.٨٣	١.٨٢	٠.٩٧	٣.٨٨	٠	٠.٩١	١.٤٨	٢.٣٨	٦.٠٧	٠	٢.٨٢	١.٩٤	٣.١	٠	٠
فخر (٤.٨)	٠	٠	٠.٧٠	٠	٠	٠	١.٥	٠	٠	٠	٠	٠	٠	٤.٨	٠	٠	٩.٦	٠	٠
اختیال (٥.٢)	٠	٠	٠.٧٥	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٥.٢	٠	٠	٥.٢	٠	٠
مرح (٥.٢)	٠	٠	٠	٠	٠	٠	٠	٠	١.٨٦	٠	٠	٠	٠	٥.٢	٢.٤	٠	٠	٠	٠
سرور (٤.٥)	٠.٣٩	٠.٧٠	٠	٠	٠.٦٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	١١.٢٥	٤٥

۴ـ۵ـ۲.ماتریس واژه ـ عبارت

در این پژوهش الگوی پنجرهٔ بافتی (بدون لحاظ رابطهٔ دستوری) برای هم‌وقوعی واژگان مورد نظر است؛ که پس از حذف ایست‌واژه‌ها محاسبه می‌شود. برای وزن‌دهی ماتریس واژه ـ عبارت، بسامد مشاهده شدهٔ هم‌وقوعی دو واژه (p_{obs}) با بسامد مورد انتظار برای هم‌وقوعی (p_{exp}) آن‌ها مقایسه می‌شود.

در پیکره‌ای دارای N واژه، که واژهٔ الف در آن m بار تکرار شده است، احتمال وقوع واژهٔ الف در پنجرهٔ بافتی n واژه‌ای، با فرض بزرگ بودن مقدار N، برابر است با $\frac{n\times m}{N}$؛ که اگر دو واژه الف و ب، با بسامد وقوع m و m´ در متن، از لحاظ آماری مستقل باشند، انتظار می‌رود احتمال وقوع هر دو آن‌ها در یک پنجرهٔ بافتی n واژه‌ای (p_{exp})، حاصل‌ضرب احتمال وقوع هر یک در آن باشد. احتمال مشاهده شدهٔ هم‌وقوعی دو واژهٔ الف و ب (p_{obs}) نیز، از تقسیم بسامد هم‌وقوعی مشاهده شده (f)، بر تعداد پنجره‌های بافتی به‌دست می‌آید ($p_{obs} = \frac{n\times f}{N}$)؛ که در محاسبهٔ هم‌وقوعی دو واژه، در صورت نادیده گرفتن جهت هم‌وقوعی، یک پنجره از طرف راست و یک پنجره از طرف چپ واژهٔ اصلی لحاظ می‌شود.

براساس قاعدهٔ PPMI، عناصر ماتریس هم‌وقوعی (برای دو واژه با بسامد وقوع m و m´ در متن و بسامد هم وقوعی f در پنجرهٔ بافتی) چنین به‌دست می‌آید:

$$\begin{cases} \log_2\left(\frac{p_{obs}}{p_{exp}}\right) = \log_2\left(\frac{f\times N}{n\times m\times m'}\right). & \text{اگر } p_{obs} > p_{exp} \\ 0 & \text{اگر } p_{obs} \le p_{exp}. \end{cases}$$

برای پیاده‌سازی در قرآن، به منظور حذف ایست‌واژه‌ها فرض می‌شود نسبت آن‌ها به کل واژه‌ها در بخش‌های مختلف پیکرهٔ قرآنی تقریباً مساوی است. بنابراین نسبتِ «تعداد کل واژگان» به «تعداد واژگان باقی‌مانده پس از حذف ایست‌واژه‌ها»، برای تعدادی از سوره‌ها محاسبه، و به کل قرآن تعمیم داده می‌شود. این نسبت به‌طور تقریبی حدود ٦٧ درصد به‌دست می‌آید[٤٧] و بنابراین اندازهٔ پیکرهٔ قرآنی با حذف ایست‌واژه‌ها، شامل ٥٢١٣٠ واژه از تعداد کل ٧٧٨٠٧ واژه (فرهنگ نامهٔ علوم قرآنی، ١٣٩٤، ص. ١٧) است. وزن‌دهی عناصر ماتریس هم‌وقوعی براساس قاعدهٔ PPMI، برای نمونه با احتساب پنجرهٔ

هم‌وقوعی ۱۵واژه‌ای (n=۱۵) در قرآن، چنین به‌دست می‌آید:

$$\begin{cases} \log_2\left(\frac{f \times 3475}{m \times m'}\right). & \text{اگر } f \times 3475 > m \times m' \\ 0 & \text{اگر } f \times 3475 \leq m \times m' \end{cases}$$

به‌کارگیری دقیق معنی‌شناسی توزیعی مستلزم درنظر گرفتن تمام هم‌نشینانِ واژگانِ مورد مقایسه است. هرچه ویژگی‌های توزیعی بیشتری در ماتریس لحاظ شود، طول بردار بافتی واژهٔ مورد مقایسه افزایش، و طبق محاسبات، مشابهت دو واژه کم‌تر به‌دست می‌آید.

واژه‌های مورد مقایسه با فرح، براساس اشتراک با آن در هم‌نشینی با واژه‌ها و عبارت‌هایی (ویژگی‌های توزیعی)، مورد مقایسه قرار می‌گیرند. تعدادی از عبارت‌های چندواژه‌ای که به‌عنوان ویژگی توزیعی فرح استفاده شده است عبارت‌اند از: «فی الارض»، «الله لایحب» و «اذاقهٔ (چشاندن) رحمهً».

نمونهٔ کوچکی از ماتریس هم‌وقوعی واژه ـ عبارت برای مقایسهٔ فرح، فخر، اختیال، مَرَح و سرور در زیر آمده است. به‌دلیل محدودیت اندازه‌گیری بدون نرم‌افزار، تنها تعداد کمی ویژگی توزیعی بیان شده است. در مقابل هر واژه میزان تکرار آن درکل قرآن ذکر شده است.

جدول ۴: ماتریس واژه ـ عبارت فرح، بسامد خام

Table 4: Word- Term matrix of Farah with raw frequencies

	چشاندن (اذاقه)	فضل	الله (٥) لایحب (۲۲)	فی الارض	عذاب و جهنم	مرح (۳)	کبر	فخر (٥)	اختیال	بغی (۲٤)	فساد (٥٠)	هزو (۳٤)	رویگردانی	سیئهٔ (۱٤)	استبشار (۸)	شرک (۱٦۱)	لون (۹)
فرح (۲۲)	٤	۳	۲	۳	4	۱	۱	۲	۱	۱	۲	۲	۳	٤	۱	٦	۰

فخر (٥)	١	١	٣	١	٢	١	٠		٣	٠	٠	١	٢	٠	٠	٠	٠		
اختیال (٣)	٠	١	٣	١	١	١	٠	٣		٠	٠	٠	٢	٠	٠	٠	٠		
مرح (٣)	٠	٠	١	٣	١		١	١	١	٠	٠	٠	١	٠	٠	١	٠		
سرور (٦)	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠	١	٠	٠	١		

اِعمال قاعدهٔ PPMI برای وزن‌دهی به ماتریس هم‌وقوعی، ماتریس زیر را نتیجه می‌دهد:

جدول ٥: ماتریس واژه ـ عبارت فرح وزن‌دهی‌شده

Table 5: Weighted Word- Term matrix of Farah

	چشاندن (ذائقه) رحمت	فضل	الله (ه) لایحب	فی الارض	عذاب و جهنم	مرح	کبر	فخر	اختیال	بغی	فساد	هزو	رویگردانی	سیئة	استبشار	شرک	لون
فرح	٦.٣	٢.٥	٣.٨	١.٤	٠.٥	٥.٧	١.٥	٦.٠	٥.٧	٢.٧	٢.٧	٣.٢	١.٨	٥.٥	٤.٣	٢.٦	٠
فخر	٦.٤	٣.٠	٦.٦	٢.٠	١.٦	٧.٩	٠		٩.٤	٠	٠	١	٣.٣	٠	٠	٠	٠
اختیال	٠	٣.٨	٧.٣	٢.٧	١.٤	٨.٦	٠	٩.٤		٠	٠	٠	٤.١	٠	٠	٠	٠

	چشاندن (اذاقه) رحمت	فضل	الله (ه) لایحب	فی الارض	عذاب و جهنم	مرح	کبر	فخر	اختیال	بغی	فساد	هزو	رویگردانی	سیئة	استبشار	شرک	لون
مرح	۰	۰	٥.٧	٤.٣	١.٤		٤.٣	٧.٩	٨.٦	۰	۰	۰	٣.١	۰	۰	٢.٨	۰
سرور	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	٥.٤	۰	۰	٦

۴ـ۶.محاسبهٔ مشابهت واژگان و تفسیر آن

پس از وزن‌دهی به عناصر ماتریس هم‌وقوعی، مشابهت واژگان براساس مشابهت بردارهای وزن‌دهی‌شده بررسی می‌شود. به‌عنوان یک نمونه مشابهت واژه‌های فخر و فرح در الگوی واژه ـ سوره، با محاسبهٔ کوسینوس زاویهٔ بین بردارهای آن‌ها به‌دست می‌آید.

$$\cos\left(\text{فخر و فرح}\right) = \frac{\text{ضرب‌داخلی‌بردارهای‌فخروفرح}}{\text{اندازهٔ‌بردارفرح} \times \text{اندازهٔ‌بردارفخر}}$$

$$\text{ضرب‌داخلی‌بردارهای‌فخروفرح} = \sum_{n}\left(\text{مؤلفهٔ}n\text{ام‌بردار‌بافتی‌فخر} \times \text{مؤلفهٔ}n\text{ام‌بردار‌بافتی‌فرح}\right)$$

$$\text{اندازهٔ‌بردار‌بافتی‌یک‌واژه} = \sqrt{\sum\left(\text{هرمؤلفهٔ‌بردار‌بافتی‌واژه}\right)^2}$$

$$\text{ضرب‌داخلی‌بردارهای‌فخروفرح} = 0/97 \times 1/5 + 3/1 \times 9/6 = 31/22$$

$$\text{اندازهٔ‌بردار‌بافتی‌فرح} = 9/44$$

$$\text{اندازهٔ‌بردار‌بافتی‌فخر} = 10/86$$

$$\rightarrow \cos\left(\text{فخر و فرح}\right) = 0/30$$

این روش برای محاسبهٔ مشابهت واژگان با فرح در الگوی واژه سوره و نیز وزن‌دهی PPMI در الگوی واژه ـ عبارت، با تعداد بیشتری واژه اجرا شده و نتیجه در جدول ٦ آمده است:

جدول ۶: مشابهت فرح با واژگان دیگر

Table 6: Similarity of Farah with other words

واژه‌های مورد مقایسه	مشابهت از طریق الگوی واژه ـ سوره	مشابهت از طریق الگوی واژه ـ عبارت (وزن‌دهی PPMI)
فخر و فرح	۰.۳۰	0.7۳
اختیال و فرح	۰.۲۳	0.61
مرح و فرح	۰.۱٤	0.57
کبر و فرح	۰.۱۷	۰.۲۸
بغی و فرح	۰.۳۰	۰.٤۳
علو و فرح	۰.۱٥	۰.۲۰
کفر و فرح	۰.۲٦	۰.۷۹
ظلم و فرح	۰.۳٤	۰.۳۹
شرک و فرح	۰.٤۳	۰.٥۷
سرور و فرح	۰.۰۰	۰.۲۷
استبشار و فرح	۰.۱۲	۰.۱٤
رضا و فرح	۰.۰٤	۰.۲۹
ضحک و فرح	۰.۰۳	۰.۲٥
غرور و فرح	۰.۱۷	۰.۲۲

علاوه بر مشابهت فرح با دیگر واژه‌ها، مشابهت سایر واژگان با یکدیگر نیز قابل محاسبه است؛ که بر این اساس، نرم‌افزارهایی قادر به ترسیم نقشهٔ معنایی (در هر الگو به‌صورت جداگانه) هستند. فاصلهٔ واژه‌ها در نقشهٔ معنایی، حوزهٔ معنایی آن‌ها را مشخص می‌کند؛ که از گام‌های ابتداییِ شناخت معنی است. به‌عبارتی معنی‌شناسی توزیعی، در جایگاه یک روش آماری برای شناخت معنا، تنها حدود معنایی یک واژه را در مقایسه با دیگر واژگان مشخص می‌کند و به‌دست آوردن معنای دقیق آن، مستلزم استفاده از روش‌های دیگر است. با مقایسهٔ شباهت فرح با واژگان مختلف، اجمالاً و در نگاهی کلی نتیجه می‌شود که این واژه، در مقایسه با حوزهٔ معنایی سرور، با واژگان حوزهٔ معنایی کبر مشابهت بیشتری دارد.[٤٨] تفاوت نتایج استفاده از الگوهای مختلفِ محاسبهٔ مشابهت معنا در روش توزیعی، که مرتبط با چالش‌های استفاده از این روش در قرآن است، در ادامه بررسی می‌شود.

bike	1			
car	0.16	1		
dog	0	0	1	
lion	0	0	0.17	1
	bike	*car*	*dog*	*lion*

۵.چالش‌های استفاده از معنی‌شناسی توزیعی در قرآن و راهکارهای برون رفت از آن

به‌کارگیری معنی‌شناسی توزیعی دارای چالش‌هایی است؛ که برخی به ماهیت روش توزیعی و برخی به ویژگی‌های پیکرهٔ مورد بررسی مانند زبان، حجم پیکره و ... مربوط می‌شود.

۵ـ۱.محدودیت‌های معنی‌شناسی توزیعی

یکی از مشکلات استفاده از روش توزیعی، پیش‌پردازش زبانی پیکره است که باید به‌صورت دستی و بدون استفاده از نرم‌افزار انجام گیرد. ابهام‌زدایی و لحاظ روابط دستوری در الگوی واژه ـ عبارت از این جمله است.

از دیگر چالش‌های کلی معنی‌شناسی توزیعی، تفسیر مشابهت میان واژگان است. به‌کارگیری روش توزیعی، تنها حوزهٔ معنایی یک واژه را ـ آن هم به صورت تقریبی و نه قطعی ـ مشخص می‌کند و برای شناخت دقیق معنا کافی نیست. شناخت معنی، نیازمند ترکیب با روش‌های دیگر است. به‌عنوان مثال تشخیص تقابل معنایی دو واژه، با استفاده از منابع لغتی؛ و تشخیص دیگر روابط معنایی واژگان، ضمن استفاده از روش‌های دیگری مانند توجه به روابط معنایی میان جملات میسر است (رک: فائز و همکاران، ۱۳۹۸، صص. ۱۰۵ ـ ۱۱۱).[٤٩]

محدودیت دیگر روش توزیعی با وجود همهٔ توانایی‌ها برای شناخت معنی، این است که این روش ادعا نمی‌کند که می‌تواند دقیقاً به مراد گویندهٔ کلام پی برد و اثرات ارجاعی کلام که معنای مورد نظر گوینده است، در آن مدل‌سازی نمی‌شود (Boleda, 2020, p.224).[٥٠]

۵ـ۲.بررسی چالش‌های به‌کارگیری روش توزیعی در قرآن

برخی از چالش‌های به‌کارگیری معنی‌شناسی توزیعی در قرآن به هم‌خوانی شرایط اجرای معنی‌شناسی

توزیعی با ویژگی‌های متن قرآن؛ و برخی به عدم وجود ابزارهای لازم برای اجرای روش بازمی‌گردد.

از مشکلات مربوط به هم‌خوانی شرایط اجرای روش با ویژگی‌های متن، حجم کم متن قرآنی (حدود ۷۸۰۰۰ واژه) نسبت به پیکره‌های معمول برای پیاده‌سازی زبان‌شناسی پیکره‌ای، به‌عنوان روشی آماری و نیازمند به تعداد زیادِ داده است. در غیر این صورت داده‌های توزیعی به تصادفی بودن نزدیک می‌شود. با این حال می‌توان در سطحی کوچک و با حذف داده‌های بسیار کم‌بسامد به‌نوعی آن را اجرا کرد. برای پیاده‌سازی این روش در قرآن، حداقل بسامد ۱۰ بار تکرار برای واژهٔ اصلی، و سه بار تکرار برای واژگان مورد مقایسه با آن پیشنهاد می‌شود. در هر صورت، زیاد بودن بسامد واژگان در قرآن (تا حدی که آن‌ها را به ایست‌واژه تبدیل نکند) و نیز بسامد هم‌وقوعی واژگان، به نتایج واقعی‌تری برای شناخت معنی می‌انجامد. در نمونهٔ مورد بررسی (ریشه ـ واژهٔ فرح) نیز، نتایج به‌دست آمده تنها تا اندازه‌ای حدود معنایی فرح را بیان می‌کند و به‌دلیل زیاد نبودن بسامد آن در قرآن، نمی‌توان دقت زیادی انتظار داشت. مشابهت به‌دست آمده برای واژهٔ فرح و واژگان کم‌بسامد، دقت کم‌تری نیز دارد. نکتهٔ شایان توجه این است که آنچه در به‌کارگیری معنی‌شناسی توزیعی در قرآن، در درجهٔ اول اهمیت دارد، استفاده از «مبنای روش توزیعی»، یا همان فرضیهٔ توزیعی است که با بیان «مشابهت معنایی واژگانِ دارای توزیع مشابه»، می‌تواند در قرآن نیز مورد توجه قرار گیرد. از آنجا که مبنای این فرضیهٔ زبانی، سخن‌گویی حکیمانهٔ انسان است و حکمت گوینده، سبب اعتبار داده‌های توزیع واژگان می شود، حکیمانه‌ترینِ سخن‌ها، سخن پروردگار، دارای داده‌های معتبرتری برای سنجش آماری است.

مشکل مهم دیگر در اجرای روش توزیعی در قرآن مربوط به ابزار به‌کارگیری روش است. در زبان عربی و به‌ویژه در قرآن با وجود نرم‌افزارهایی برای یافتن ریشهٔ واژه و بسامد آن در سوره‌ها، نرم‌افزارهای کارآمدی برای برخی محاسبات (مانند حذف ایست‌واژه‌ها) موجود نیست (ترکیان، ۱۳۹۸الف، ص. ٦٤٦). جداسازی واژگان باب‌های مختلف یک ریشه، و تشخیص معانی متفاوت یک واژه نیز در قرآن عملاً با نرم‌افزار امکان‌پذیر نیست و بنابراین لازم است بسیاری از محاسبات به‌صورت دستی و غیر خودکار انجام شود. براساس این کاستی در الگوی واژه ـ عبارت با محاسبات دستی، عملاً امکان به‌شمار آوردن تمام هم‌وقوعی‌ها وجود ندارد و بنابراین نتایج از دقت لازم برخوردار نیست. به‌عنوان نمونه در پژوهش حاضر، تنها برخی هم‌نشینان فرح مورد توجه قرار گرفته و هم‌نشینان اختصاصی سایر واژه‌های مورد مقایسه با فرح به حساب نیامده است. بنابراین طول بردار بافتی همهٔ واژگان مورد مقایسه، کم‌تر از میزان واقعی، و در پی آن مشابهت واژهٔ فرح با واژگان دیگر عمدتاً بیشتر از مقدار واقعی به‌دست می‌آید.

اما در الگوی واژه ـ سوره، همهٔ سندها (سوره‌ها) لحاظ می‌شود و از این رو نتایج، مشابهت واژه‌ها را در الگوی واژه ـ سوره کمتر از الگوی واژه ـ عبارت به دست می‌دهد.

علاوه بر این، در برخی گام‌های اجرای روش توزیعی، مانند کاهش بُعد با روش SVD نیز آشنایی با نرم‌افزارهای محاسباتی لازم است و یکی از علل عدم ورود زبان‌شناسان و مشابه آن پژوهشگران قرآنی به حوزهٔ معنی‌شناسی توزیعی، عدم آشنایی با این ابزارهاست.

مشکل دیگر مرتبط با شرایط اجرای روش توزیعی، که در الگوی واژه ـ سند، با لحاظ سوره (در معنای اصطلاحی آن) به عنوان سند نمود می‌یابد، تفاوت بسیار زیاد طول سوره‌هاست. با وجود لحاظ تفاوت طول سوره در وزن‌دهی عناصر ماتریس هم‌وقوعی، عملاً نقش قرارگیری واژه‌ها در سوره‌های طولانی، به‌سبب عدم توجه به دوری و نزدیکی آن‌ها به هم، نادیده گرفته شده، به ویژه میزان مشابهتِ به‌دست‌آمده برای واژه‌هایی که در سوره‌های کوتاه به‌کار رفته‌اند، نسبت به دیگر واژه‌ها دورتر از واقعیت نشان داده می‌شود. در نمونهٔ مورد بررسی، به‌دلیل عدم کاربرد واژهٔ فرح در سوره‌های کوتاه، مشابهت فرح با واژگانی که در سوره‌های کوتاه کاربرد دارند، کم‌تر از میزان واقعی به‌دست می‌آید. به‌عنوان نمونه مشابهت فرح با سرور (که دو مورد از شش وقوع آن، در سورهٔ انشقاق و یک مورد در سورهٔ انسان است) صفر شده است.

یک راهکار برای حل این مشکل، استفاده از رکوعات قرآنی به‌عنوان سند است؛ که طول آن‌ها تفاوت اندکی با هم دارد. اما علاوه بر این‌که لزوماً این دسته‌بندی، واحدهای مستقل دارای موضوع واحد را نشان نمی‌دهد، نرم‌افزارهای موجود در تعیین بسامد وقوع واژه‌ها در رکوعات ناتوانند و این امر محاسبات را دشوار می‌کند. راهکار پیشنهادی برون‌رفت از این چالش، تصحیح خطای مربوط به تفاوت طول سوره‌ها، ضمن مقایسه با الگوی واژه ـ عبارت، پس از انجام محاسبهٔ شباهت میان واژگان است. با توجه به این‌که مشابهت در الگوی واژه ـ عبارت بیش از مقدار واقعی آن به دست می‌آید، می‌توان گفت مقدار واقعی مشابهت، بین دو مقدار به‌دست آمده در الگوی واژه ـ سوره و الگوی واژه ـ عبارت است.

۶. نتیجه

۱- معنی‌شناسی توزیعی در قرآن، با هدف شناخت تقریبی معنا و به‌عبارتی تعیین حوزهٔ معنایی واژگان مورد اختلاف، با رویکردی آماری صورت می گیرد. پیاده‌سازی این روش بر روی متن قرآن به‌منظور

شناخت معنی یک واژه، شامل گام‌های زیر است:

۱. یافتن واژگانی برای مقایسه با واژهٔ اصلی؛

۲. تعیین ویژگی‌های توزیعی واژه‌ها، شامل سندها (سوره‌ها با معنای اصطلاحی یا رکوعات قرآنی) و عبارات هم‌وقوع با واژه (با لحاظ رابطهٔ دستوری یا وقوع در پنجرهٔ بافتی)؛

۳. پیش‌پردازش زبانی متن قرآنی شامل ریشه‌یابی و یافتن صیغه‌های مهم هر ریشه برای حضور در ماتریس هم‌وقوعی، ابهام‌زدایی و ادغام واژگان قریب‌المعنا؛

۴. حذف ایست‌واژه‌ها؛

۵. تشکیل ماتریس هم‌وقوعی و وزن‌دهی آن (در الگوی واژه ـ سوره با قاعدهٔ TFIDF و در الگوی واژه ـ عبارت با قاعدهٔ PPMI)؛

۶. تشکیل بردار بافتی وزن‌دهی‌شده و محاسبهٔ مشابهت واژگان با کوسینوس زاویهٔ آن‌ها؛

۷. تفسیر مشابهت میان واژگان با استفاده از لغت‌نامه‌ها و روابط معنایی میان جملات.

۲- چالش‌های به‌کارگیری معنی‌شناسی توزیعی به‌ویژه در قرآن و برخی راه‌های برون‌رفت از آن عبارت است از:

۱. چالش‌های کلی مربوط به معنی‌شناسی توزیعی؛ شامل عدم سهولت پیش‌پردازش زبانی به‌صورت خودکار، تفسیر مشابهت میان واژگان، محدودیت در تشخیص مراد متکلم.

۲. چالش‌های به‌کارگیری معنی‌شناسی توزیعی در قرآن شامل حجم کم پیکرهٔ قرآنی برای محاسبات آماری، عدم وجود نرم‌افزار مناسب جهت ریشه‌یابی مؤثر، ابهام‌زدایی و محاسبهٔ ایست‌واژه‌ها، و تفاوت بسیار زیاد طول سوره‌ها در الگوی واژه ـ سوره است. توجه بیشتر به مبنای روش توزیعی (فرضیهٔ توزیعی) که بر پایهٔ حکمت گوینده، اعتبار بیشتر در داده‌های قرآنی را نتیجه می‌دهد، و تصحیح خطای ناشی از تفاوت طول سندها در الگوی واژه ـ سوره با مقایسهٔ نتایج با الگوی واژه ـ عبارت؛ راهکارهای پیشنهادی برون‌رفت از برخی چالش‌هاست.

۳- علاوه بر پاسخ به سؤال‌های اصلی پژوهش، اجرای روش توزیعی برای ریشه ـ واژهٔ فرح در قرآن، تفاوت معنای به‌دست‌آمده از بافت قرآنی (نزدیک به حوزهٔ معنایی تکبر)، نسبت به معنای غالبی ذکرشده در لغت‌نامه‌ها، به‌ویژه دیدگاه لغت‌شناسان غیرقرآنی و به طور اخص، لغت‌نامه‌های امروزی (شادی) را نشان می‌دهد. تفاوت این دو معنا، با توجه به مبنای شباهت زبان قرآن به زبان عربی عصر نزول، تطور معنایی این واژه را نتیجه می‌دهد. شناخت سیر تحول درزمانی معنای این واژه، با

به‌کارگیری روش توزیعی در متون دوره‌های مختلف پس از قرآن، مانند روایات، امکان پذیر است و می‌تواند موضوع پژوهش‌های دیگر باشد.[۵۱]

۷. پی‌نوشت‌ها

1. distributional semantics
2. corpus linguistics
3. natural language
4. *Theories of Lexical Semantics*
5. *Machine learning for text*
6. *Networks*
7. Distributional Hypothesis
8. Zellig Harris
9. feature
10. document
11. co - occurance

۱۲. نمونه‌ای از پیکره‌های تخصصی، *پیکرهٔ میشیگان شامل مقالات سطح بالای دانشجویی:* Michigan Corpus of Upper level Student Papers (MICUSP) است (کرافورد و سزومی، ۲۰۱٦، ترجمهٔ صفوی، ۱۳۹۹، ص. ۳۷)؛ که ۸۳۰ مقاله با ۲.٦ میلیون واژه دارد (اقتباس از https://eresources.eli.lsa.umich.edu/micusp - papers/ - academic - written - of - corpus).

13. term
14. context window

۱۵. این انتخاب می‌تواند به‌جهت هم‌نشینی با واژه (راست یا چپ) نیز مقید شود (Lenci, 2018, p.153) و نزدیکی و دوری دو واژه در پنجرهٔ بافتی هم مورد توجه باشد (Sahlgren, 2006,p.41؛ ترکیان، ۱۳۹۸ب، ص. ٤۲٤).

16. context vector
17. co - occurrence matrix
18. vector
19. similarity

۲۰. اندازهٔ یک بردار براساس قضیهٔ فیثاغورث، جذر مجموع مجذور مؤلفه‌های آن است:

$$|a| = \sqrt{(a_1^2 + a_2^2 + \cdots)}$$

21. pre - processing
22. lemmatization
23. part - of - speech (POS) tagging
24. Disambiguation

۲۵.مقوله‌های دستوری بسته، به‌ندرت عضو جدید می‌پذیرند؛ مانند ضمایر و حروف ربط و .... اما در مقوله‌های باز، جایگزینی، تغییر و پذیرش اجزای جدید رخ می‌دهد، مانند اسم، صفت، فعل و ... (Sahlgren, 2006, p. 38).

۲۶.واژگان بسیار کم‌بسامد معمولاً به‌طور تصادفی با واژهٔ اصلی هم‌وقوع شده‌اند.

27. stop words
28. weighting
29. raw frequency
30. Term Frequency

۳۱.Inverse Document Frequency. این تابع به‌صورت لگاریتم نسبت کل اسناد (D)، به سندی که عبارت i در آن به‌کار رفته، تعریف می‌شود: $IDF_i = \log_2 \left(\frac{D}{DF_i}\right)$. در بسیاری موارد که «نسبت» اهمیت دارد، به دلیل گسترهٔ بسیار وسیع مقادیر، از تابع لگاریتم استفاده می‌شود؛ که با درنظر گرفتن مرتبه (توان)، نسبت‌ها را به مقیاس مناسبی درمی‌آورد ($x^z = y$ معادل است با $\log_x y = z$).

32. scaling
33. expected probability
34. observed probability
35. Singular Value Decomposition
36. semantic map
37. neighbor

۳۸.. به عنوان نمونه «ماشین و راننده» دارای مفاهیم مرتبط با هم اند، اما در این دسته‌بندی نمی‌گنجند.

۳۹.به‌عنوان نمونه: إِن شَجَرَةَ الزقومِ، طَعامُ الْأَثيمِ (دخان: ٤٣ و ٤٤).

۴۰.به‌عنوان مثال عبارت‌های «الذین آمنوا» و «مؤمنون» ذیل ریشهٔ «امن» قابل بررسی است.

۴۱.این واژگان دارای رابطهٔ «چندمعنایی» یا «هم‌آوا ـ هم‌نویسی» هستند. ا«چندمعنایی» (polysemy) ارتباط معانی مختلف واژه و «هم‌آوا ـ هم‌نویسی» یا هم‌نامی (homonymy)، چند معنا داشتن واژه بدون ارتباط معنایی را نشان می‌دهد (صفوی، ۱۳۹۹، صص. ۱۱۰ و ۱۱۱؛ افراشی، ۱۳۹۵، صص. ۱۲۹ و ۱۳۰).

۴۲.... أَفَحُكمَ الجاهلية يَبغونَ... (مائده: ۵۰)

۴۳.الذينَ يَصُدونَ عَنْ سَبيلِ اللهِ وَ يَبْغُونَها عِوَجاً ... (اعراف:٤٥)

۴۴.با توجه به فرمول، مقدار این عبارت چنین به دست می‌آید: $IDF_{\text{واژه}} = \log_2 \left(\frac{D}{DF_{\text{واژه}}}\right) =$ $\log_2 \left(\frac{\text{تعدادکل‌سوره‌ها}}{\text{تعدادسوره‌هایی‌که‌واژه‌درآن‌آمده}}\right)$؛ که «تعداد کل سوره‌ها» بسته به لحاظ سوره یا سوره‌وار به‌معنای رکوعات قرآنی، ۱۱٤ یا ۵۵۵ است.

۴۵.برای محاسبهٔ طول سوره، تعداد واژگان سوره‌ها، یا نسبت تعداد واژگان سوره‌ها به تعداد واژگان یک

سوره می‌تواند محاسبه شود.

۴۶.برای محاسبهٔ طول سوره، نسبت تعداد واژگان آن با تعداد واژگان سورهٔ حدید، ٥٤٤ واژه (فرهنگ نامهٔ علوم قرآنی، ۱۳۹٤، ص. ۲٦۷٦) مقایسه شده است ($\frac{\text{تعدادواژگان‌سوره}}{544}$ = طول‌سوره).

۴۷.تعداد واژه‌های سوره‌های تغابن، مجادله، قمر و تحریم به‌ترتیب ۲٤۱، ٤۷۳، ۳٤۲ و ۲٤۰ ذکر شده است (فرهنگ‌نامهٔ علوم قرآنی، ۱۳۹٤، صص. ۲٦۷۳ ـ ۲٦۸٥). پس از حذف ایست‌واژه‌ها (با محاسبهٔ دستی)، تعداد واژگان این سوره‌ها به‌ترتیب ۱٦۱، ۲۸٦، ۲٤۰ و ۱٦٦ به‌دست آمد. متوسط نسبت تعداد واژگان این سوره‌ها پس از حذف ایست‌واژه‌ها به تعداد کل واژگان سوره، حدود ٦۷ درصد است.

۴۸.ممکن است تصور شود برخی مقادیر به‌دست‌آمده، نشانگر رابطهٔ معنایی قابل توجه و معناداری میان واژگان نیست. اما توجه به نتایج برخی پژوهش‌های این حوزه، معنادار بودن این مقدار مشابهت معنایی را نشان می‌دهد. در پژوهشی که واژه‌های bike، car، dog و lion، از لحاظ توزیعی بررسی شده، مشابهت میان واژگان چنین به‌دست آمده است (Lenci, 2018, 156):

bike	1			
car	0.16	1		
dog	0	0	1	
lion	0	0	0.17	1
	bike	*car*	*dog*	*lion*

که مشابهت میان واژه‌های bike و car، و نیز واژه‌های dog و lion، را کم‌تر از ۰.۲ نشان می‌دهد؛ درحالی که آن‌ها به یک دستهٔ مفهومی از واژگان تعلق دارند.

۴۹.برای مثال در معنی‌شناسی فرح، ارتباط میان جملات دربردارندهٔ فرح و فخر در آیه‌های ۲۲ و ۲۳ سورهٔ حدید، شمول معنایی یا هم‌معنایی این واژه‌ها را نتیجه می‌دهد (ذوعلم و همکاران، ۱۳۹۹، ص. ۱۳).

۵۰.این محدودیت معنی‌شناسی توزیعی، در کاربست قرآنی، معادل با عدم شناخت مراد متکلم با این روش است. بر این اساس با توجه به برخی تعاریفِ ظاهر و باطن قرآن، معنی‌شناسی توزیعی صرفاً به شناخت معنی ظاهری واژگان قرآنی ـ البته با روشی اصولی و نظام‌مند ـ می‌پردازد و ادعایی درمورد کشف لایه‌های باطنی آن ندارد.

۵۱.در پایان‌نامهٔ *معنی‌شناسی فرح در قرآن و روایات* تا اندازه‌ای به این تحول پرداخته شده است (نک: ذوعلم، ۱۳۹۷).

۸. منابع

- آذرنوش، آ. (۱۳۸۳). *فرهنگ معاصر عربی ـ فارسی*. تهران: نی.
- ابن درید، م. (۱۹۸۸). *جمهرة اللغة*. بیروت: دار العلم للملایین.

- ابن قتیبه، ع. (١٤١١ق). *تفسیر غریب القرآن*. بیروت: دار و مکتبة الهلال.
- ابوعبیده، م. (١٣٨١ق). *مجاز القرآن*. قاهره: مکتبة الخانجی.
- ازهری، م. (١٤٢١ق). *تهذیب اللغة*. بیروت: دار احیاء التراث العربی.
- افراشی، آ. (١٣٩٥). *مبانی معناشناسی شناختی*. تهران: پژوهشگاه علوم اسلامی و مطالعات فرهنگی.
- ایزوتسو، ت. (١٩٦٦). *مفاهیم اخلاقی ـ دینی در قرآن مجید*. ترجمهٔ ف. بدره‌ای (١٣٧٨). تهران: فرزان.
- ترکیان، ا. (١٣٩٨الف). آموزش عملی متن کاوی، پیوست الف *متن کاوی یادگیری ماشین*. چ. آگاروال، ترجمهٔ ا. ترکیان (صص ٦١٧ ـ ٦٥٩). چاپ دوم، تهران: نیاز دانش.
- ترکیان، ا. (١٣٩٨ب). شبکهٔ متن، پیوستِ *شبکه‌ها*. م. نیومن، ترجمهٔ ا. ترکیان (صص ٤٣٣ ـ ٤٥٠). تهران: نیاز دانش.
- جوهری، ا. (١٣٧٦ق). *الصحاح (تاج اللغة و صحاح العربیة)*. بیروت: دارالعلم للملایین.
- ذوعلم، آ. (١٣٩٧). *معناشناسی فرح در قرآن و روایات*. پایان‌نامهٔ کارشناسی ارشد. دانشکدهٔ الهیات دانشگاه تهران.
- ذوعلم، آ.، فائز، ق.، و خوش‌منش، ا. (١٣٩٩). معناشناسی فرح در قرآن کریم؛ پاسخی به شبههٔ نکوهش شادی در قرآن. *تحقیقات علوم قرآن و حدیث*، *٤٧* (٣)، ١ ـ ٣٤.
- راغب اصفهانی، ح. (١٤١٢ق). *مفردات ألفاظ القرآن*. بیروت: دارالقلم.
- صفوی، ک. (١٣٩٩). *درآمدی بر معنی‌شناسی*. چاپ ششم، تهران: پژوهشکدهٔ فرهنگ و هنر اسلامی.
- طبرسی، ف. (١٣٧٢). *مجمع البیان فی تفسیر القرآن*. تهران: ناصر خسرو.
- عبد الباقی، م. (١٣٩٣). *المعجم المفهرس لالفاظ القرآن الکریم*. قم: نوید اسلام.
- عسکری، ح. (١٤٠٠ق). *الفروق فی اللغة*. بیروت: دار الافاق الجدیدة.
- فائز، ق.، خوش‌منش، ا.، و ذوعلم، آ. (١٣٩٨). بهره‌گیری از سیاق متنی در معناشناسی ساختگرای قرآن کریم. *پژوهش‌های قرآن و حدیث*، ٥٢(١)، ٩٥ ـ ١١٥.
- فراهیدی، خ. (١٤٠٩ق). *کتاب العین*. قم: هجرت.
- فرهنگ‌نامهٔ علوم قرآنی، دفتر تبلیغات اسلامی، قم: پژوهشگاه علوم و فرهنگ اسلامی، ١٣٩٤.
- کرافورد، و.، و سزومی، ا. (٢٠١٦). *زبان‌شناسی پیکره‌ای در عمل*. ترجمهٔ م. نوبخت (١٣٩٩). تهران: سمت.
- گلی ملک‌آبادی، ف.، خاقانی اصفهانی، م.، و شکرانی، ر. (١٣٩٦). نقدی معناشناختی بر ترجمه‌های فارسی

واژهٔ «دون» در قرآن کریم. *جستارهای زبانی*، ۳٦ (۱)، ۲۰۷ ـ ۲۳۰.

- گیررتس، د. (۲۰۱۰). *نظریه‌های معنی‌شناسی واژگانی*. ترجمهٔ ک. صفوی (۱۳۹۸). تهران: نشر علمی.
- لسانی فشارکی، م.، و مرادی زنجانی، ح. (۱۳۹۵). *سوره‌شناسی*. چاپ دوم. قم: نصایح.

References

- Afrashi. A. (2016). *Introdusing Cognitive Semantics*. Institute for Humanities and Cultural Studies [In Persian].
- Askari. H. (1979). *Alfuroogh Fillugha*. Dar Al'afagh Aljadida [In Arabic].
- Azarnoosh, A. (2004). *A Dictionary of Modern Written Arabic*. Nei [In Persian].
- Azhari. (2000). *Tahzib Allugha*. Dar Al'ihya Altorath Al'arabiy [In Arabic].
- Abu - ubaida. M. (1963). *Majaz alQuran*. Maktaba Alkhanaji [In Arabic].
- Boleda, G. (2020). *Distributional Semantics and Linguistic Theory*. *Annual Review of Linguistics*, 6, 213– 234.
- Crawford. W., Csomay. E. (2016). *Doing Corpus Linguistics*, Samt [In Persian].
- *Dictionary of Quranic Sciences*. (2015). Islamic Sciences and Culture Academy [In Persian].
- Dumais, S., Furnas, G., Landauer, T. (1988). Using latent semantic analysis to improve access to textual information, In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* New York (pp. 281- 285).
- Faez. Gh., Khoshmanesh. A., Zouelm. A. (2019). The Use of Textual Context in Structural Semantics of the Holy Qur'an, *Quranic Researches and Tradition*, 52(1), 95 - 115 [In Persian].
- Farahidi. Kh. (1988). *Ketab Al'Ain*. Hejrat [In Arabic].
- Geeraerts. D. (2010). *Theories of Lexical Semantics*, Nashre Elmi [In Persian].
- Goli. F, Khaqani. M, Shokrani R. (2017). A Semantic Criticism of the Persian Translation of the Arabic Term "Doon" in the Holy Quran, *Language Relate Research*, 8 (1), 207 - 230 [In Persian].
- Harris, Z. (1954). *Distributional Structure*. *WORD*, 10(2 - 3), 146 - 162.
- Ibn Duraid. M. (1988). *Jumhora allugha*. Dar Al'ilm Almullain [In Arabic].

- Ibn Qutaiba. ‘A. (1990). *Tafsir Qarib alQuran*. Dar va Maktaba Alhilal [In Arabic].
- Jawhari. I. (1956). *Alsehah*. Dar Al’ilm Almullain [In Arabic].
- Izutsu. T. (1966). *The structure of meaning in ethico - religious concepts in Quran*. Farzan [In Persian].
- Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4, 151–171.
- Lesani Fesharaki. M., Moradi. Zanjani. H. (2016). *Surah Shenasi*. Nasayeh [In Persian].
- Raghib Isfahani. H. (1991). *Mufradat Alfaz Alquran*. Dar Alghalam [In Arabic].
- Safavi. K. (2020). *An Introduction to Semantics*. Sourey - e - Mehr [In Persian].
- Sahlgren, M. (2006) *The Word - Space Model*, Doctoral Dissertation, Stockholm University, Sweden.
- Schutze, H. (1992). dimensions of meaning, In *ACM/IEEE conference on Supercomputing*, (pp. 787 - 796).
- Schutze, H. (1993). Word space. In *Proceedings of the 1993 Conference on Advances in Neural Information Processing Systems*, NIPS’93 (pp. 895–902).
- Tabrasi. F. (1993). *Majma’ Albayan fi Tafsir Alquran*. Naserkhosro [In Arabic].
- Turkian, A. (2019a). Learning Practical text mining, *Machine learning for text*. Aggarwal. C. (pp. 617– 659), Niaze Danesh [In Persian].
- Turkian, A. (2019b). Text Network. *Networks*. Newman.M. (pp. 433– 450), Niaze Danesh [In Persian].
- Turney, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics, *Journal of Artificial Intelligence Research*, 37, 141 - 188.
- Zouelm. A. (2018). *Semantical Study of Farah in Quran and Tradition*. Master thesis, Faculty of Theology and Islamic Studies, Tehran University [In Persian].
- Zouelm. A., Faez. Gh., Khoshmanesh. A. (2020). Semantical Study of Farah in the holy Qur'an, A Response to the Doubt of Disapproval of Joy in the Qur'an. *Journal of Researches of Quran and Hadith Sciences*. 17(3), 1 - 34 [In Persian].