University of Tehran Prees

Journal of Language Research

Online ISSN: 2676-3362

https://jolr.ut.ac.ir

The Relationship between Lexical Frequency and the Distribution of Word-Initial Phonological Classes in High-Frequency Persian Words

Azam Estaji [1]

1. Corresponding author, Associate Professor of Linguistics, Department of Linguistics, Ferdowsi University of Mashhad, Iran. Email: estagi@um.ac.ir

Article Info

Article Type:
Research Article

Article History:
Received:
8, July, 2026

In Revised Form:
16, August, 2026

Accepted:
11, September, 2026

Published Online:
21, September, 2026

Keywords: lexical frequency; word-initial phonemes; phonological classes; phonotactics; Persian; corpus linguistics.

Abstract

Lexical frequency is considered an important factor in usage-based phonology, yet its influence on the distribution of word-initial phonemes in Persian has received limited empirical attention. This study investigates the distributional pattern of word-initial phonological classes in Persian and evaluates the extent to which lexical frequency affects this pattern. The dataset consisted of the first 1,000 entries in *Routledge Frequency Dictionary of Persian*, divided into two frequency groups comprising the 500 most frequent words and the following 500 words. The initial segment of each word was classified into seven phonological categories: stops, fricatives, affricates, nasals, liquids, glides, and vowels. The distributions of these categories were compared using the Chi-square test of independence, and effect size was assessed with Cramer's V. The findings reveal a remarkably stable phonological organization across both frequency groups. Stops constitute the dominant word-initial class, followed by fricatives, while the overall proportion of obstruents and sonorants remains highly consistent. The distribution of individual phonemes is also strongly skewed, with only a small number of phonemes—particularly /m/, /e/, /s/, /t/, /d/, /b/, and /h/—accounting for a substantial proportion of all word onsets, whereas many others occur only rarely in initial position. These results highlight the central role of Persian phonotactic constraints and syllable structure in shaping the organization of word-initial phonemes. Although minor differences in the frequencies of some phonological classes were observed between the two lexical-frequency groups, these differences were not statistically significant, and the associated effect size was negligible. The findings therefore suggest that lexical frequency plays only a limited role in determining the distribution of word-initial phonological classes within the high-frequency Persian lexicon, whereas structural phonological constraints provide a more robust explanation. The study contributes empirical evidence to Persian phonology and offers implications for corpus linguistics, computational language processing, speech technology, and Persian language education.

The Author(s): Estaji, A., (2026): The Relationship between Lexical Frequency and the Distribution of Word-Initial Phonological Classes in High-Frequency Persian Words. Journal of Language Researches, No. 1, Vol.17 , Serial No.32, Spring& Summer-(23-36). https://doi.org/10.22059/jolr.2026.410222.666954

Publisher: University of Tehran Press.

© Author(s) retain the copyright. Azam Estaji

DOI: https://doi.org/10.22059/jolr.2026.410222.666954

1. Introduction

The present study investigates the relationship between lexical frequency and the phonological organization of word-initial position in Persian. Within usage-based approaches, lexical frequency is recognized as an important factor shaping linguistic structure. In phonology, however, it remains unclear whether differences in lexical frequency are reflected in the distribution of phonemes in specific structural positions, or whether such distributions remain stable because of phonotactic and syllable-structural constraints. Word-initial position is especially relevant for spoken-word recognition, while syllable onsets are subject to phonotactic constraints related to sonority. Previous research on English, notably Fenk-Oczlon (2001), reported that highly frequent words may exhibit a different phonological profile from less frequent words. Persian provides an interesting testing ground because its phonotactic system and syllable structure differ from those of English, and the question has not been directly investigated. The present study asks whether the first 500 most frequent Persian words differ significantly from the following 500 in their distribution of word-initial phonological classes, and whether any differences indicate a systematic frequency effect or relative stability.The study addresses three questions: (1) does the distribution of word-initial phonological classes differ significantly between the first 500 and the second 500 most frequent Persian words? (2) if differences occur, which phonological classes contribute most strongly? (3) do the observed patterns indicate a systematic frequency-related change, or relative stability across the high-frequency lexicon? The study hypothesized that the two groups would differ significantly, that some classes would be more sensitive to frequency differences, and that the overall organization of high-frequency Persian vocabulary would remain relatively stable. The study adopts a corpus-based descriptive-analytical design. The dataset consists of the first 1,000 entries in A Frequency Dictionary of Persian: Core Vocabulary for Learners (Miller & Aghajanian-Stewart, 2018), divided into two equal frequency groups: Ranks 1–500 and Ranks 501–1000, a division that also permits comparison with Fenk-Oczlon's (2001) analysis. The unit of analysis was the initial phonological segment of each word, classified into one of seven classes—stops, fricatives, affricates, nasals, liquids, glides, and vowels—with initial vowels treated as a separate category rather than as initial glottal stops. The statistical comparison used a Pearson chi-square test of independence; because liquids and glides contained relatively few observations, they were combined for the inferential analysis. Cramér's V was calculated as a measure of effect size, with significance set at $\alpha = .05$.

The descriptive results show a remarkably similar distribution in the two groups. In the first 500 words, stops constitute 31.2% of all word-initial segments, followed by fricatives (26.0%), vowels (18.0%), nasals (15.2%), affricates (5.2%), liquids (3.2%), and glides (1.2%). In the second 500 words, stops account for 30.2%, fricatives for 28.4%, vowels for 15.6%, nasals for 17.6%, affricates for 3.6%, liquids for 4.2%, and glides for 0.4%. The overall ranking and distributional profile remain highly similar, and the inferential analysis confirms that these differences are not statistically significant: the Pearson chi-square test yielded $\chi^2(5, N = 1000) = 3.823$, $p = .575$, with a negligible effect size (Cramér's $V = .062$). The first hypothesis, predicting a significant difference, is therefore not supported, providing evidence for substantial stability within the first 1,000 high-frequency Persian words. At a broader structural level, obstruents constitute approximately two-thirds of word-initial segments in both groups (62.4% and 62.2%), sonorants account for 19.6% and 22.2%, and initial vowels for 18.0% and 15.6%. The distribution of individual phonemes is highly asymmetric: the seven most frequent initial phonemes (/b/, /d/, /t/, /s/, /e/, /m/, /h/) account for approximately 50% of all 1,000 word onsets, while several other phonemes occur only rarely in initial position. /m/ is the single most frequent initial phoneme (11.6% of the entire sample), while /e/ is the most frequent initial vowel (7.1%).

The discussion suggests that lexical frequency does not substantially reorganize the distribution of word-initial phonological classes within the high-frequency Persian lexicon. Although

frequency-related differences occur at the level of individual categories, their magnitude is small and not statistically significant. This contrasts with English, where Fenk-Oczlon (2001) reported systematic differences across frequency levels, including more highly sonorous or weak initial segments among frequent words. The absence of a comparable effect in the Persian data suggests that the relationship between lexical frequency and phonological organization may be language-specific rather than universal. One possible explanation is the stronger role of phonotactic and syllable-structural constraints: the predominance of obstruent onsets is consistent with the preference for low-sonority segments at the beginning of syllables and a steep sonority rise toward the nucleus. The skewed distribution of individual phonemes indicates that phonotactic constraints operate alongside frequency.

2.Conclusion

This study examined the relationship between lexical frequency and word-initial phonological organization in 1,000 high-frequency Persian words. The comparison of the first 500 and the following 500 frequency-ranked words revealed no statistically significant difference in the distribution of phonological classes, $\chi^2(5, N = 1000) = 3.823$, $p = .575$, with a negligible effect size (Cramér's $V = .062$), indicating substantial stability across the two levels of high lexical frequency. The findings support the view that, although lexical frequency may influence particular aspects of linguistic structure, the broad distribution of word-initial phonological classes in Persian is more strongly constrained by phonotactics and syllable structure, with the predominance of obstruent onsets pointing to a robust organization relatively insensitive to small differences in lexical frequency. The study provides empirical evidence that frequency effects in other languages cannot necessarily be generalized to Persian.

دانشگاه تهران

پژوهش‌های زبانی

شاپای الکترونیکی: ۳۳۶۲-۲۶۷۶

https://jolr.ut.ac.ir/



بررسی رابطه‌ی بسامد واژگانی و توزیع طبقات واج آغازین در واژه‌های پربسامد فارسی

اعظم استاجی[1]✉

۱.نویسنده مسئول، دانشیار زبانشناسی، گروه زبانشناسی، دانشگاه فردوسی مشهد، مشهد، ایران.رایانامه: estagi@um.ac.ir

اطلاعات مقاله

نوع مقاله:
علمی – پژوهشی

تاریخ دریافت:
۱۴۰۵/۰۴/۱۷

تاریخ بازنگری:
۱۴۰۵/۰۵/۲۵

تاریخ پذیرش:
۱۴۰۵/۰۶/۲۰

تاریخ انتشار:
۱۴۰۵/۰۶/۳۰

واژه‌های کلیدی:
بسامد واژگانی، واج آغازین، واج‌آرایی، زبان فارسی، رسایی، زبان‌شناسی پیکره‌ای

چکیده

بسامد واژگانی یکی از عوامل اثرگذار در رویکردهای مبتنی بر کاربرد برای تبیین سازمان واجی زبان به شمار می‌رود، اما نقش آن در توزیع واج‌های آغازین واژه‌های فارسی تاکنون کمتر به‌صورت تجربی بررسی شده است. پژوهش حاضر با هدف تحلیل الگوی توزیع واج‌های آغازین و ارزیابی تأثیر بسامد واژگانی بر این الگو انجام شد. داده‌های پژوهش از هزار واژه نخست «فرهنگ بسامدی زبان فارسی» استخراج و بر اساس میزان بسامد به دو گروه شامل ۵۰۰ واژه نخست و ۵۰۰ واژه بعدی تقسیم شدند. واج آغازین هر واژه بر پایه ویژگی‌های واج‌شناختی در هفت طبقه شامل انسدادی، سایشی، انسایشی، خیشومی، روان، غلت و واکه طبقه‌بندی شد. سپس توزیع این طبقات در دو گروه با استفاده از آزمون خی‌دو و اندازه اثر کرامر بررسی گردید. نتایج نشان داد که در هر دو گروه، همخوان های انسدادی و پس از آن همخوان‌های سایشی بیشترین سهم را در آغاز واژه‌ها دارند و نسبت کلی همخوان‌های گرفته و رسا از ثبات بالایی برخوردار است. همچنین توزیع واج‌های آغازین به‌شدت نامتقارن است؛ به‌گونه‌ای که تنها چند واج، به ویژه، /m/، /e/، /s/، /t/، /d/، /b/و/h/ ، بخش عمده‌ای از آغاز واژه‌های پربسامد فارسی را تشکیل می‌دهند، در حالی که بسیاری از واج‌ها سهم اندکی دارند. این الگو بیانگر نقش پررنگ محدودیت‌های واج‌آرایی و ساخت هجایی در سازمان‌دهی آغاز واژه‌های فارسی است. مقایسه دو گروه بسامدی نیز نشان داد که اگرچه تغییرات جزئی در فراوانی برخی طبقات واجی مشاهده می‌شود، این تفاوت‌ها از نظر آماری معنادار نیست و اندازه‌ی اثر نیز بسیار کوچک است. بنابراین، بسامد واژگانی در میان واژه‌های پربسامد فارسی تأثیر چشمگیری بر توزیع طبقات واجی آغازین ندارد و این توزیع بیش از آنکه تابع بسامد باشد، از محدودیت‌های ساختاری نظام واجی زبان فارسی پیروی می‌کند. یافته‌های این پژوهش می‌تواند در مطالعات واج‌شناسی، زبان‌شناسی پیکره‌ای، آموزش زبان فارسی و مدل‌های پردازش رایانشی گفتار و زبان فارسی مورد استفاده قرار گیرد.

استناد: استاجی، اعظم ؛(۱۴۰۵): بررسی رابطه‌ی بسامد واژگانی و توزیع طبقات واج آغازین در واژه‌های پربسامد فارسی. پژوهش‌های زبانی، سال ۱۷، شمارهٔ ۱، بهار و تابستان- پیاپی ۳۲، (۳۶-۲۳).
DOI: 10.22059/jolr.2026.410222.666954

ناشر: مؤسسهٔ انتشارات دانشگاه تهران. © نویسندگان: اعظم استاجی

۱. مقدمه

بسامد واژگانی یکی از مهم‌ترین متغیرهای زبانی است که نقش آن در شکل‌گیری، سازمان‌دهی و تحول نظام‌های زبانی در چند دهه‌ی اخیر در زبان‌شناسی شناختی، واج‌شناسی و زبان‌شناسی پیکره‌ای مورد توجه زیادی قرار گرفته است. در زبانشناسی کاربرد–بنیاد[1] زبان نظامی پویا تلقی می‌شود که ساختار آن از تجربه‌ی زبانی گویشوران و تکرار الگوهای زبانی شکل می‌گیرد. بر این اساس، بسامد عاملی مؤثر در شکل‌گیری بازنمایی‌های ذهنی، تثبیت الگوهای واجی و صرفی و حتی تغییرات تدریجی ساختار زبان به شمار می‌رود (بایبی[2]، ۲۰۲۳).

شواهد فراوانی نشان داده‌اند که بسامد واژگانی بر جنبه‌های مختلف پردازش زبان تاثیرگذار است. واژه‌های پربسامد معمولاً سریع‌تر شناسایی می‌شوند، با سهولت بیشتری از واژگان ذهنی بازیابی می‌شوند، در برابر فراموشی مقاوم‌ترند و در فرایند تغییرات تاریخی نیز رفتار متفاوتی نسبت به واژه‌های کم‌بسامد از خود نشان می‌دهند (ژورافسکی و همکاران[3]، ۲۰۰۱). افزون بر این، پژوهش‌های متعددی نشان داده‌اند که بسامد با پدیده‌هایی مانند کاهش آوایی، حذف یا ادغام واج‌ها، طول واژه و پیچیدگی ساخت هجایی ارتباط دارد (بایبی، ۲۰۰۱: ۱۱). با وجود این، هنوز مشخص نیست که آیا بسامد می‌تواند بر سازماندهی واجی واژگان نیز اثر بگذارد و الگوی توزیع واج‌ها، به‌ویژه در جایگاه آغاز واژه را دستخوش تغییر کند. آغاز واژه از دیدگاه واج‌شناسی و روان‌شناسی زبان جایگاه ویژه‌ای دارد. نخستین واج یا خوشه‌ی آغازین معمولاً نخستین اطلاعی است که در جریان ادراک گفتار دریافت می‌شود و نقش مهمی در فعال‌سازی مدخل‌های واژگانی، دسترسی به واژگان ذهنی و شناسایی واژه ایفا می‌کند (مارسلن – ویلسون[4]، ۱۹۸۷).

با این حال، ساخت واجی واژگان تنها تابع بسامد نیست. هر زبان مجموعه‌ای از محدودیت‌های واج‌آرایی را بر توزیع واج‌ها و ساخت هجا اعمال می‌کند و این محدودیت‌ها تعیین می‌کنند که کدام واج‌ها یا ترکیب‌های واجی در جایگاه‌های مختلف واژه مجاز یا غیرمجاز باشند. به عنوان مثال کلمنتس[5] (۱۹۹۰) با طرح اصل توالی رسایی[6] نشان داد که ساخت آغاز هجا در زبان‌های مختلف تابع محدودیت‌های نظام‌مند رسایی است. از دیدگاه نظری، پرسش اساسی این است که آیا افزایش یا کاهش بسامد کاربرد می‌تواند این الگوهای واجی را دگرگون کند یا این‌که ساخت واجی واژگان، به‌رغم تفاوت‌های بسامدی، از پایداری نسبی برخوردار است.

یکی از نخستین پژوهش‌هایی که این پرسش را به‌صورت تجربی بررسی کرد، مطالعه‌ی فنک– اتسلن[7] (۲۰۰۱) بود. وی با تحلیل داده‌های فرهنگ بسامدی ثورندایک– لورج[8] (۱۹۴۴) نشان داد که توزیع برخی طبقات واجی در آغاز واژه‌های انگلیسی در گروه‌های مختلف بسامدی یکسان نیست و با کاهش بسامد، ترکیب واج‌های آغازین تغییر می‌کند. این یافته نشان داد که میان بسامد واژگانی و سازماندهی واجی آغاز واژه می‌تواند رابطه‌ای وجود داشته باشد.

زبان فارسی از نظر تعداد همخوان‌ها، محدودیت‌های واج‌آرایی، ساخت هجایی و الگوهای آغاز واژه تفاوت‌های قابل توجهی با زبان انگلیسی دارد. در فارسی، آغاز هجا در واژگان بومی تقریباً همواره با یک همخوان شکل می‌گیرد و نظام واج‌آرایی این زبان در مقایسه با انگلیسی محدودیت‌های بیشتری بر ساخت آغاز هجا اعمال می‌کند. افزون بر این، توزیع طبقات واجی در آغاز واژه‌های فارسی با انگلیسی یکسان نیست. چنین تفاوت‌هایی این احتمال را مطرح می‌کنند که رابطه‌ی میان بسامد و سازماندهی واجی نیز در فارسی از الگوی متفاوتی پیروی کند.

---

1. Usage-based linguistics
2.Bybee
3.Jurafsky, Bell, Gregory, & Raymond
4.Marslen-Wilson
5.Clemnts
6.Sonority Sequencing Principle
7.Fenk-Oczlon
8.Thorndike–Lorge

با وجود گسترش مطالعات پیکره‌ای در زبان فارسی، تاکنون پژوهشی که توزیع طبقات واج آغازین را در سطوح مختلف بسامدی به‌صورت نظام‌مند و با استفاده از آزمون‌های آماری بررسی کرده باشد، گزارش نشده است. بیشتر مطالعات موجود بر بسامد واژه، ساخت هجایی، واج‌آرایی، پردازش واژگان یا تغییرات آوایی تمرکز داشته‌اند و ارتباط میان بسامد و سازماندهی واجی آغاز واژه همچنان ناشناخته باقی مانده است. این خلأ، ضرورت انجام پژوهش در این حوزه را آشکار می‌سازد.

پژوهش حاضر با هدف بررسی این مسئله انجام شده است که آیا کاهش نسبی بسامد در میان واژگان پربسامد زبان فارسی با تغییر در توزیع طبقات واج آغازین همراه است یا آنکه سازماندهی واجی این واژگان از پایداری نسبی برخوردار است. برای این منظور، هزار واژه‌ی نخست فرهنگ بسامدی زبان فارسی روتلج[1] (۲۰۱۸) انتخاب و به دو گروه شامل ۵۰۰ واژه‌ی پربسامد نخست و ۵۰۰ واژه‌ی پربسامد بعدی تقسیم شد. سپس واج آغازین هر واژه بر اساس طبقات واجی (انسدادی، سایشی، انسایشی، دماغی، روان، نیم‌واکه و واکه) دسته‌بندی و توزیع این طبقات در دو گروه با استفاده از آزمون‌های آماری مقایسه شد.

سؤالات پژوهش

پژوهش حاضر در پی پاسخ به پرسش‌های زیر است:

۱.آیا توزیع طبقات واج آغازین در ۵۰۰ واژه نخست و ۵۰۰ واژه بعدی فرهنگ بسامدی فارسی تفاوت معناداری دارد؟

۲.در صورت وجود تفاوت، کدام طبقات واجی بیشترین سهم را در ایجاد این تفاوت دارند؟

۳.آیا تفاوت‌های مشاهده‌شده نشان‌دهنده تغییر نظام‌مند ساخت واجی در واژگان پربسامد فارسی است یا بیانگر پایداری نسبی الگوی توزیع واج‌های آغازین در این واژگان است؟

فرضیه‌های پژوهش

با توجه به پیشینه پژوهش ، فرضیه‌های زیر مطرح می‌شوند:

فرضیه اول: (H_1) توزیع طبقات واج آغازین در دو گروه بسامدی با یکدیگر تفاوت معنادار دارد.

فرضیه دوم: (H_2) سهم طبقات مختلف واجی در ایجاد این تفاوت یکسان نیست و برخی طبقات واجی نسبت به سایر طبقات حساسیت بیشتری به تغییرات بسامدی نشان می‌دهند.

فرضیه سوم: (H_3) با وجود تفاوت‌های احتمالی در فراوانی برخی طبقات واجی، الگوی کلی سازماندهی واجی واژگان پربسامد فارسی از پایداری نسبی برخوردار است.

۲. مروری بر پیشینه پژوهش

پژوهش درباره رابطه میان بسامد کاربرد و ساخت واجی از دهه ۱۹۸۰ به بعد به یکی از محورهای مهم زبان‌شناسی شناختی و واج‌شناسی کاربرد- بنیاد تبدیل شده است. هرچند بخش عمده این پژوهش‌ها بر نقش بسامد در پردازش واژگان، تغییرات واجی و کاهش آوایی تمرکز داشته‌اند، برخی مطالعات نیز به بررسی ارتباط میان بسامد و سازماندهی واجی واژگان پرداخته‌اند.

همان طور که پیشتر اشاره شد، یکی از نخستین پژوهش‌هایی که رابطه‌ی میان بسامد واژگانی و توزیع طبقات واجی را بررسی کرد، مطالعه‌ی فنک- اتسلون (۲۰۰۱) بود . وی با تحلیل رابطه‌ی میان درجه‌ی رسایی همخوان‌های آغازین (بر اساس مقیاس رسایی) و بسامد واژگانی در داده‌های بسامدی ثورندایک-لورج (۱۹۴۴)، نشان داد که توزیع همخوان‌های آغازین در طبقه‌ی واژگان بسیار پربسامد (طبقه‌ی AA با تکرار ۱۰۰ بار یا بیشتر در میلیون) تفاوت معناداری با توزیع کلی آن‌ها داشت. بر اساس یافته‌های وی، واژه‌هایی که با همخوان غلت آغاز می‌شدند، با اختصاص ۱۱,۵ درصد، بیشترین سهم را در طبقه‌ی بسیار پربسامد داشتند؛ در حالی که واژه‌های آغازشده با همخوان‌های انسدادی تنها ۴,۵

1. Routledge

درصد از این طبقه را شامل شدند. این نتایج بیانگر این بود که در بالاترین سطح بسامدی، میزان حضور همخوان‌های گرفته به شکل چشمگیری پایین‌تر و سهم همخوان‌های رسا بسیار بالاتر بود.

علاوه بر این، فنک- اتسلون(۲۰۰۱) نشان داد همبستگی معناداری میان بسامد واژگانی و ویژگی‌های آوایی همخوان‌های آغازین دیده می‌شود. مقایسه توزیع همخوان‌های آغازین ضعیف ([ɣ], [w], [h], [θ], [ð]) در دو طبقه‌ی بسامدی (۱-۵۰۰ و ۵۰۰-۱۰۰۰) نشان داد که بسامدِ وقوع این همخوان‌ها در واژه‌های بسیار پربسامد (طبقه‌ی اول) به‌طور محسوسی بالاتر از طبقه‌ی دوم است. نتایج این پژوهش نشان داد ساخت واجی واژگان می‌تواند تحت تأثیر ویژگی‌های آماری آنها قرار گیرد و همین امر زمینه را برای پژوهش‌های بعدی در حوزه تعامل میان بسامد و سازماندهی واجی فراهم ساخت. پژوهش‌های جدید در واج‌شناسی کاربرد-بنیاد نیز رابطه‌ی بین بسامد و بازنمایی واجی را تایید و تقویت کرده‌اند. بایبی (۲۰۲۳) بسامد را یکی از ارکان اصلی رویکرد مذکور می‌داند و میان بسامد رخداد و بسامد نوع تمایز قائل می‌شود. به نظر وی، تکرار مداوم واحدهای زبانی موجب تقویت بازنمایی ذهنی و خودکارشدن تولید گفتار می‌شود؛ ازاین‌رو، تغییرات آوایی کاهشی معمولاً زودتر در واژه‌ها و عبارت‌های پربسامد رخ می‌دهند. به بیان لو و سونگ[1] (۲۰۲۴) واژه‌های پربسامد در پردازش ادراکی زبان انگلیسی، نسبت به واژه‌های کم‌بسامد رفتار متفاوتی دارند و اثر بسامد حتی بر ادراک ساخت آغاز هجا نیز مشاهده می‌شود.

در حوزه بازشناسی واژه‌های گفتاری نیز پژوهش‌های اخیر حاکی از آن هستند که اطلاعات موجود در آغاز واژه همچنان مهم‌ترین عامل در فعال‌سازی واژگان ذهنی به شمار می‌رود. ازاین‌رو، توزیع واج‌های آغازین با شیوه‌ی سازمان‌دهی واژگان ذهنی و فرایند بازشناسی واژه‌های گفتاری نیز ارتباط دارد، زیرا فعال‌سازی و رقابت میان مدخل‌های واژگانی بر پایه‌ی اطلاعات واج آغازین صورت می‌گیرد (کاپنولا و همکاران،[2] ۲۰۲۴؛ مارسلن- ویلسون، ۱۹۸۷).

مطالعات رده‌شناختی جدید نیز بر این نکته تاکید دارند که توزیع واج‌های آغاز هجا در زبان‌های مختلف کاملاً تصادفی نیست بلکه از الگوهای همگانی تبعیت می‌کند. پایگاه فونوتکتیکون[3] که در سال ۲۰۲۵ معرفی شده، اطلاعات واج‌آرایی بیش از ۵۰۰ زبان را گردآوری کرده و امکان مقایسه ساخت آغاز هجا، هسته و پایانه هجا را فراهم ساخته است. این پایگاه نشان می‌دهد که نوع و فراوانی طبقات همخوانی در آغاز هجا از ویژگی‌های پایدار نظام واجی زبان‌هاست (جو و شو[4]، ۲۰۲۵).

در مطالعات زبان‌شناسیِ فارسی، آواشناسی زبان فارسی اثر ثمره (۱۳۷۱) یکی از مهمترین منابعی است که به ویژگیهای توزیعی واجهای زبان فارسی پرداخته است. ثمره با تحلیل پیکره‌ی واژگانیِ «فرهنگ فارسی معین»، بسامد وقوع همخوان‌ها و واج آرایی آنها را بررسی کرده است. با این حال تاکید پژوهش مذکور بیشتر بر توزیع واجها در هجاهای مختلف زبان فارسی است. ثمره (۱۳۷۱) در بررسی هجای cvcc به رابطه‌ی همخوان آغازین و سایر اجزاء هجا پرداخته است و تعداد وقوع تک تک همخوان‌های زبان فارسی را در جایگاه آغازین این نوع هجا بررسی کرده است. با تجمیع این داده‌ها در طبقات بسامدیِ گوناگون، می‌توان به الگوی دقیق‌تری از میزان وقوع و توزیع همخوان‌ها در ساخت هجایی دست یافت، اما همان طور که اشاره شد این تحلیل تنها برای هجای cvcc انجام شده است و همخوان آغازین سایر هجاها مورد بررسی قرار نگرفته است (ثمره، ۱۳۷۱: ۲۰۳).

اسلامی و بی‌جن خان (۱۳۸۲، به نقل از بیجن خان،۱۳۸۴: ۱۲۶) با تحلیل حدود ۱۷ هزار واژه در بخشی از تحلیل خود توزیع همخوانهای رسا و گرفته را در آغاز انواع هجاها بررسی کرده‌اند و به این نتیجه رسیده‌اند که بخش اعظمی از هجاهای CV ۶۵٪، CVC ۷۰٪ و CVCC ۷۷٪ با همخوان‌های گرفته آغاز می‌شوند. این توزیع آماری نشانگر یک گرایش جهانی در نظام‌های واجی است که بر اساس آن، حضور آواهایی با کمترین میزان رسایی در جایگاه آغازه، به دلیل ایجاد

1 . Lu & Song
2 . Kapnoula et al.
3. Phonotacticon
4.Joo & Hsu

حداکثر تقابل با هسته هجا، از منظر ادراک شنیداری بهینه محسوب می‌شود.اسلامی و همکاران (۱۳۹۴) با رویکردی کمّی و پیکره‌محور، نظام آوایی زبان فارسی را مورد تحلیل آماری قرار داده‌اند. نویسندگان پس از استخراج داده‌ها از یک پیکره، ویژگی‌های آوایی واژگان را با تکیه بر شاخص‌هایی چون الگوهای تکیه، ساختار هجایی، واج‌آرایی و خوشه‌های همخوانی بررسی کرده‌اند. این پژوهشگران در تحلیل داده‌های خود به بررسی فراوانی واج‌های زبان فارسی نیز پرداخته‌اند اما در این بررسی جایگاه واج‌ها (از جمله بسامد وقوع واج‌ها در جایگاه آغازین واژه) از یکدیگر تفکیک نشده است. بر پایه بررسی منابع در دسترس، تاکنون پژوهشی که به‌طور مستقیم توزیع طبقات واج آغازین را در سطوح مختلف بسامدی واژگان فارسی مقایسه کرده باشد، گزارش نشده است. پژوهش حاضر با تکیه بر داده‌های هزار واژه‌ی نخست فرهنگ بسامدی فارسی روتلج، این خلأ را هدف قرار می‌دهد. در این پژوهش، توزیع طبقات واج آغازین در دو گروه از واژگان پربسامد فارسی (۱—۵۰۰ و ۵۰۱—۱۰۰۰) مقایسه شده و با استفاده از آزمون‌های آماری، فرضیه وجود یا نبود تفاوت میان این دو گروه بررسی می‌شود.

۳. روش پژوهش

پژوهش حاضر از نوع توصیفی ـ تحلیلی و مبتنی بر پیکره است. جامعه آماری شامل ۱۰۰۰ واژه پربسامد فارسی [۱] است که بر اساس رتبه بسامدی به دو گروه مساوی تقسیم شدند. واحد تحلیل در این پژوهش واج آغازی هر واژه است. برای هر واژه، تنها نخستین همخوان آغازین در نظر گرفته شد و ویژگی‌های سایر بخش‌های واژه در تحلیل وارد نشد. در مواردی که واژه با واکه آغاز می‌شد، واکه به عنوان واحد آغازی ثبت گردید و با توجه به جایگاه مورد بحث همخوان انسدادی چاکنایی در زبان فارسی این دسته از واژه ها در دسته همخوان آغازی انسدادی قرار نگرفتند و جداگانه بررسی شدند. بعلاوه، صورت واجی واژه‌ها که در فرهنگ بسامدی ثبت شده بود، مبنای دسته بندی واج‌ها بود و نه صرفاً صورت نوشتاری. پس از استخراج واج آغازین، همه‌ی داده‌ها به صورت دستی توسط نگارنده بر اساس طبقه‌ی تولیدی همخوان‌ها طبقه‌بندی شد. سپس فراوانی هر طبقه‌ی واجی شمارش و درصد هر طبقه نسبت به کل واژگان هر گروه محاسبه شد. در ادامه نیز واج‌های آغازین در دو گروه عمده‌ی همخوان‌های رسا و گرفته مورد تحلیل قرار گرفت.تحلیل داده‌ها در دو مرحله انجام گرفت. در مرحله نخست، برای هر یک از دو گروه، فراوانی و درصد وقوع طبقات مختلف واج آغازی محاسبه و در قالب جدول‌های توصیفی ارائه شد. در مرحله دوم، برای بررسی تفاوت میان دو گروه از آزمون خی دو مستقل استفاده شد. این آزمون مشخص می‌کند که آیا اختلاف مشاهده‌شده در توزیع طبقات واجی میان دو گروه می‌تواند ناشی از تغییر بسامد باشد یا آنکه تفاوت‌های مشاهده‌شده در محدوده نوسانات تصادفی قرار می‌گیرند. برای اجرای آزمون از نرم افزار اس پی اس اس نسخه ۲۷,۰,۱ استفاده شد. سطح معنی‌داری آزمون $\alpha = 0.05$ در نظر گرفته شد. علاوه بر مقدار خی دو و سطح معنی‌داری، برای برآورد شدت رابطه میان متغیرها از شاخص اندازه‌ی اثر کرامر (Cramér's V) نیز استفاده شد تا اندازه‌ی اثر مستقل از حجم نمونه گزارش شود.

۱. تحلیل داده‌ها

در این قسمت به ترتیب به توصیف و تحلیل داده‌های پژوهش می پردازیم.

۴-۱ . توصیف داده‌ها

در این پژوهش، واج آغازین هزار واژه‌ی پربسامد زبان فارسی بررسی شد. این واژگان بر اساس رتبه‌ی بسامدی به دو گروه مساوی شامل ۵۰۰ واژه نخست و ۵۰۰ واژه دوم تقسیم شدند. سپس واج آغازین هر واژه بر اساس شیوه تولید در هفت طبقه واجی طبقه‌بندی شد. جدول ۱ توزیع فراوانی و درصد هر طبقه‌ی واجی را در دو گروه بسامدی نشان می‌دهد.

همان‌گونه که جدول ۱ نشان می‌دهد، در هر دو گروه، همخوان‌های انسدادی بیشترین فراوانی را دارند. پس از آن، همخوان‌های سایشی و سپس واکه‌ها و خیشومی‌ها قرار می‌گیرند. مقایسه دو گروه نشان می‌دهد اگرچه برخی طبقات

تغییرات نسبی چند درصدی را نشان می‌دهند، الگوی کلی توزیع طبقات واجی در دو گروه تا حد زیادی مشابه باقی مانده است.

جدول ۱. توزیع طبقات واج آغازین در دو گروه بسامدی

طبقه واجی	۱–۵۰۰	درصد	۵۰۱–۱۰۰۰	درصد	تغییر
انسدادی	156	31.2	151	30.2	−1.0
سایشی	130	26.0	142	28.4	+2.4
انسایشی	26	5.2	18	3.6	−1.6
خیشومی	76	15.2	88	17.6	+2.4
روان	16	3.2	21	4.2	+1.0
غلت	6	1.2	2	0.4	−0.8
واکه	90	18.0	78	15.6	−2.4
جمع	500	100	500	100	—

۴-۲ .تحلیل آماری

برای بررسی اینکه آیا تفاوت‌های مشاهده‌شده از نظر آماری معنادار هستند یا خیر آزمون خی‌دو مستقل بر روی شش طبقه واجی انجام شد. به منظور رعایت پیش‌فرض‌های آزمون، طبقات «روان» و «غلت» به دلیل فراوانی پایین در یک طبقه ادغام شدند. نتایج نشان داد که تفاوت توزیع طبقات واجی میان دو گروه بسامدی معنادار نیست، $\chi^2(5, N = 1000) = 3.82, p = .575, Cramér's V = .062.$ بنابراین، با وجود نوسان‌های جزئی در فراوانی برخی طبقات، الگوی کلی توزیع طبقات واجی در دو گروه بسامدی از پایداری بالایی برخوردار است و اثر بسامد واژگانی بر این سطح از سازماندهی واجی ناچیز به نظر می‌رسد.

۴-۳ .تحلیل رسایی واج‌های آغازین

یکی از مفاهیم اصلی در تحلیل ساختار آغاز هجا، مفهوم رسایی است. در واج‌شناسی، رسایی به درجه‌ی برجستگی آکوستیکی و گشودگی مجرای گفتار در تولید واج‌ها اشاره دارد. بر اساس اصل توالی رسایی هجاها معمولاً به‌گونه‌ای سازمان می‌یابند که از عناصر کم‌رساتر در حاشیه‌ها به عناصر رساتر در هسته حرکت کنند. از این منظر، نوع واج آغازین واژه می‌تواند در شکل‌دهی به «شیب رسایی» هجا و در نتیجه، در سهولت ادراک یا تولید آن نقش داشته باشد. از یک سو، آغازه‌های کم‌رساتر ممکن است به دلیل ایجاد تضاد بیشتر با هسته‌ی هجایی، مرز واژه را برای شنونده برجسته‌تر کنند؛ و از سوی دیگر، آغازه‌های رساتر ممکن است از نظر تولیدی کم‌هزینه‌تر باشند.

پژوهش‌های واج‌شناسی نشان داده‌اند که در بسیاری از زبان‌ها، آغازه‌هایی که شیب رسایی تندتری به سمت هسته دارند، مطلوب‌ترند. برای مثال هجای پا /pa/ با شیب آغازی زیاد مطلوبتر از هجای یا /ja/ با شیب آغازی ملایم تلقی میگردد. آغاز هجا در پردازش گفتار اهمیت ویژه‌ای دارد و یک گرایش عام در زبان‌ها دیده می‌شود که واج‌های با رسایی کمتر آغازه هجا را پر کنند (اسمیت، ۲۰۰۳؛ به نقل از بی‌جن‌خان ۱۳۸۴: ۱۲۳) . اگر سلسله‌مراتب رسایی را به صورت زیر در نظر بگیریم: انسدادی >انسایشی >سایشی >خیشومی >روان >غلت > واکه.

آن‌گاه فاصله‌ی رسایی تا هسته برای انسدادی‌ها بیشترین و برای غلت‌ها کمترین است. جدول زیر نشان می‌دهد در داده-های مورد بررسی هرچه رسایی واج آغازین کمتر باشد، فراوانی آن در آغاز واژه‌های پربسامد بیشتر است:

جدول۲- ترتیب نزولی رسایی واج‌های آغازین

طبقه واجی	درصد
انسدادی	31.2٪
سایشی	26.0٪
انسایشی	5.2٪
خیشومی	15.2٪
روان	3.2٪
غلت	1.2٪

در مرحله بعد، برای بررسی نقش ویژگی‌های طبقات واجی عمده، داده‌ها در سه گروه اصلی طبقه‌بندی شدند : همخوان‌های گرفته (انسدادی، سایشی و انسایشی)، رساها (خیشومی، روان و غلت) و واکه‌ها . نتایج این دسته‌بندی در جدول ۳ آمده است.

جدول ۳- توزیع سه گروه اصلی واج‌های آغازین در دو گروه بسامدی ۱-۱۰۰۰:

گروه واجی	۱-۵۰۰	درصد	۵۰۱-۱۰۰۰	درصد	کل ۱-۱۰۰۰	درصد
گرفته‌ها	312	62.4٪	311	62.2٪	623	62.3٪
رساها	98	19.6٪	111	22.2٪	209	20.9٪
واکه‌ها	90	18.0٪	78	15.6٪	168	16.8٪
جمع	500	100٪	500	100٪	1000	100٪

جدول ۳ نشان می‌دهد که نزدیک به دو سوم واژه‌های پربسامد فارسی (۶۲/۴ درصد) با همخوان‌های گرفته آغاز می‌شوند. در مقابل، سهم رساها ۱۹/۶ درصد و سهم آغازه‌های واکه‌ای ۱۸ درصد است. بنابراین، در واژه‌های پربسامد فارسی، آغاز هجا با همخوان‌های گرفته به‌مراتب رایج‌تر از آغاز هجا با رساها یا واکه‌هاست. در مجموعِ دو گروه بسامدی نیز این نسبت کمابیش دیده می‌شود.

۴-۴. مقایسه با سایر پیکره ها

جهت انجام این مقایسه، ابتدا داده های مستخرج از ثمره (۱۳۷۱: ۲۰۳) [۲] در مورد همخوان آغازین هجای cvcc در دو گروه همخوان‌های رسا و گرفته دسته بندی شد تا مشخص شود آیا شباهت یا تفاوتی بین این دو گروه از واژه‌ها دیده می‌شود یا خیر. جدول زیر توزیع این دو گروه همخوان را در دو پیکره مورد اشاره نشان می‌دهد؛

جدول ۴- مقایسه نسبت همخوان‌های گرفته، رسا و واکه‌ها در دو دسته داده:

گروه	داده های ثمره	فرهنگ بسامدی
گرفته‌ها	70.0٪	62.3٪
رساها	22.4٪	20.9٪
واکه‌ها	7.6٪	16.8٪

همان طور که ملاحظه می شود رساها تقریباً ثابت مانده‌اند واختلاف آنها فقط ۱,۵ درصد است. بزرگ‌ترین اختلاف مربوط به واکه‌هاست یعنی بیش از دو برابر، این احتمالاً ناشی از تفاوت در واحد تحلیل است. در داده‌های فرهنگ بسامدی روتلج همه انواع هجاها حضور دارند و هجای آغازین واژه تحلیل شده است، در حالی که داده‌های مستخرج از ثمره (۱۳۷۱) تنها مبتنی برهجای CVCC است و قیدی در مورد این‌که هجا در آغاز واژه باشد وجود ندارد. از سوی دیگر، گرفته‌ها نیز کاهش یافته‌اند، اما این کاهش تقریباً معادل همان افزایش واکه‌هاست. اگر واکه‌ها را کنار بگذاریم و فقط همخوان‌ها را مقایسه کنیم اختلاف میان دو مجموعه بسیار کمتر می‌شود.

جدول ۵- مقایسه نسبت همخوان‌های گرفته و رسا در دو دسته داده:

اختلاف	فرهنگ بسامدی	داده های ثمره	گروه
−0.9	74.9٪	75.8٪	گرفته‌ها
+0.9	25.1٪	24.2٪	رساها

ملاحظه می‌شود اختلاف فقط حدود ۰٫۹ درصد است. بنابراین می توان گفت نسبت کلی رساها و گرفته‌ها در آغاز هجاها و واژه‌های فارسی تا حد زیادی پایدار است و تغییر عمده میان دو مجموعه بیشتر ناشی از حضور یا عدم حضور واژه‌های واکه آغازین است، نه تغییر در سازماندهی همخوان آغاز واژه. میانگین درصد همخوان‌های گرفته در آغاز انواع هجا در پژوهش اسلامی و بی‌جن خان (۱۳۸۲) نیز حدود ۷۱٪ است که بسیار به ارقام فوق نزدیک است. در مجموع این مقایسه نشان می‌دهد که در سطح کلانِ رساها و گرفته‌ها، الگوی توزیع بسیار پایدارتر از آن چیزی است که در طبقات جزئی‌تر (مانند انسدادی، سایشی یا روان) دیده می‌شود و در زبان فارسی داشتن شیب تند در آغاز هجا در همه‌ی انواع واژه های فارسی فارغ از بسامد کاربرد آنها ترجیح داده می‌شود. در جدول زیر توزیع تک تک واج‌های آغازین (به جز انسدادی چاکنایی) در هزار واژه‌ی پربسامد به ترتیب نزولی دیده می‌شود:

جدول ۶- توزیع تک تک واج‌های آغازین (به جز انسدادی چاکنایی) در هزار واژه‌ی پربسامد به ترتیب نزولی

واج	۱–۵۰۰	۵۰۱–۱۰۰۰	کل	٪	واج	۱–۵۰۰	۵۰۱–۱۰۰۰	کل	٪	واج	۱–۵۰۰	۵۰۱–۱۰۰۰	کل	٪
/m/	53	63	116	11.6	/dʒ/	16	14	30	3.0	/o/	7	7	14	1.4
/e/	35	36	71	7.1	/ɑ/	17	13	30	3.0	/tʃ/	10	4	14	1.4
/s/	23	44	67	6.7	/q/	10	19	29	2.9	/j/	6	2	8	0.8
/t/	33	31	64	6.4	/f/	12	16	28	2.8	/i/	5	2	7	0.7
/d/	40	23	63	6.3	/v/	17	10	27	2.7	/l/	1	1	2	0.2
/b/	29	31	60	6.0	/z/	11	10	21	2.1	/u/	1	0	1	0.1
/h/	33	26	59	5.9	/x/	18	18	36	3.6	/ʒ/	0	1	1	0.1
/n/	23	25	48	4.8	/r/	15	20	35	3.5	جمع	500	500	1000	100
/a/	25	20	45	4.5	/ʃ/	16	17	33	3.3					
/k/	22	20	42	4.2	/p/	11	21	32	3.2					
/g/	11	6	17	1.7										

با بررسی داده های جدول فوق می‌توان به نکات زیر اشاره کرد:

۱ .توزیع آغازه‌های واژه‌های پربسامد نامتوازن است. تنها چند واج، بخش بزرگی از آغاز واژه‌ها را تشکیل می‌دهند. هفت واج نخست، /m/ ، /e/، /s/، /t/، /d/، /b/ و /h/ در مجموع ۵۰۰ مورد از ۱۰۰۰ آغازه (۵۰٪) را به خود اختصاص داده‌اند. در مقابل، بسیاری از واج‌ها مانند /ʒ/، /u/ و /l/ سهمی ناچیز دارند.

۲ . واج /m/ با فراوانی ۱۱/۶٪، پرتکرارترین واج آغازی است و فاصله‌ی قابل‌توجهی با سایر واج‌ها دارد. لازم به ذکر است که طبقه‌ی خیشومی‌ها فقط دو عضو دارد و بسامد بالای این دو عضو در کنار بسامد پایین‌تر طبقه‌ی انسدادی‌ها با تعداد بیشتر و بسامد کمتر در روند کلی گرایش به آغازه‌هایی با بیشترین اختلاف رسایی چندان مشهود نیست. [۳]

۳ .واکه‌ی /e/شاخص‌ترین واکه آغازی است. در میان واکه‌ها، /e/ با ۷/۱٪ بیشترین فراوانی را دارد و حتی از بسیاری از همخوان‌های پرکاربرد نیز پرتکرارتر است. بنابراین، توزیع واکه‌های آغازی یکنواخت نیست.

۴. واج‌هایی مانند /ʒ/، /u/ و /ا/ تقریباً هرگز در آغاز واژه‌های پربسامد ظاهر نمی‌شوند. این الگو نشان می‌دهد که محدودیت‌های توزیعی آغاز واژه بر برخی واج‌ها شدیدتر از سایر جایگاه‌های واژه است.

۵. بحث و نتیجه

هدف این پژوهش بررسی تأثیر بسامد کاربرد بر توزیع طبقات واج آغازین در واژگان پربسامد زبان فارسی بود. برای این منظور، واج آغازین هزار واژه‌ی پربسامد زبان فارسی استخراج و در دو گروه بسامدی با یکدیگر مقایسه شد. نتایج آزمون خی دو نشان داد که میان دو گروه تفاوت آماری معناداری وجود ندارد. بنابراین بسامد تنها یکی از عوامل مؤثر بر توزیع واج‌های آغازین است و نقش آن در مقایسه با محدودیت‌های نظام واجی فارسی محدود به نظر می‌رسد. در هر دو گروه بسامدی، همخوان‌های انسدادی بیشترین سهم را در آغاز واژه‌ها دارند و پس از آن سایشی‌ها قرار می‌گیرند. حفظ این الگوی کلی نشان می‌دهد که ساخت آغاز واژه در فارسی از ثبات قابل توجهی برخوردار است. از منظر واج‌شناسی، این نتیجه با دیدگاه‌هایی همراستا است که ساخت هجایی و محدودیت‌های واج‌آرایی را عامل اصلی شکل‌گیری الگوهای توزیع واج‌ها می‌دانند. بر اساس این دیدگاه‌ها، گرچه بسامد می‌تواند احتمال وقوع برخی الگوها را تغییر دهد، اما چارچوب کلی توزیع واج‌ها توسط محدودیت‌های واجی زبان تعیین می‌شود. یافته‌های پژوهش حاضر نیز نشان می‌دهد که کاهش بسامد موجب تغییر نسبی برخی طبقات می‌شود، اما این تغییرات در مقیاس کل نظام واجی محدود باقی می‌ماند. مقایسه-ی یافته‌های حاضر با پژوهش‌های انجام‌شده درباره‌ی زبان انگلیسی نشان می‌دهد که تاثیر بسامد در این دو زبان یکسان نیست. بسامد واژگانی در زبان انگلیسی با توزیع واج‌های آغازین ارتباط دارد و تغییرات وابسته به بسامد چشمگیر گزارش شده‌اند، در حالی که در فارسی الگوهای واج‌آرایی سهم بیشتری در حفظ سازماندهی واجی واژگان دارند. در مطالعه‌ی فنک- اتسلون (۲۰۰۱)، سهم واج‌های آغازین ضعیف[j]) ، [w]، [h]، [θ]، ([ð] از گروه بسامدی ۱—۵۰۰ به گروه ۵۰۱— ۱۰۰۰ کاهش می‌یافت، اما در فارسی بین دو گروه بسامدی تفاوت معناداری مشاهده نشد. در داده‌های فارسی، کاهش عمدتاً در رده‌های انسدادی، انسایشی و آغازهای واکه‌ای دیده شد، در حالی که سهم همخوان‌های سایشی،روان و خیشومی افزایش یافت.

در کنار این یافته‌ها، مقایسه‌ی توصیفی با داده‌های گزارش‌شده در برخی پژوهش‌های پیشین مانند ثمره (۱۳۷۱) و اسلامی و بیجن‌خان (۱۳۸۲) نکته‌ی جالب توجهی را آشکار می‌کند. در سطح کلان، نسبت رساها و گرفته‌ها در داده‌های حاضر تفاوت چشمگیری با داده‌های حاصل از پیکره‌های غیربسامدی ندارد. این شباهت نشان می‌دهد که برخی ویژگی‌های توزیع واج‌های آغازین در فارسی، فراتر از تفاوت‌های بسامدی، از ثبات قابل توجهی برخوردارند. بنابراین می-توان گفت با توجه به غلبه همخوان‌های گرفته، الگوی توزیع واج‌های آغازین در واژه‌های پربسامد فارسی با اصل ترجیح شیب رسایی زیاد در آغاز هجا سازگار است.

بررسی توزیع تک تک واج‌های آغازین نیز نشان داد: الف) هفت واج نخست (/m, e, s, t, d, b, h/) از میان ۲۸ واج، دقیقاً نیمی (۵۰٪) از کل واج‌های آغازین را به خود اختصاص می‌دهند. این یعنی توزیع واج آغازین در واژه‌های پربسامد نامتقارن است. ب) واج /m/ پرتکرارترین واج آغازین است. این یافته با دیدگاه کلاسیک یاکوبسن درباره‌ی برجستگی واج /m/ به‌عنوان یکی از بنیادی‌ترین همخوان‌های زبان و از نخستین واج‌های اکتسابی کودک (یاکوبسن، ۱۹۶۸: ۴۷–۴۸) همراستا است. جمع بسامد کاربرد /m/ و /n/ (۱۶۴ مورد، ۱۶٫۴٪) نیز نشان می‌دهد، خیشومی‌ها به‌تنهایی بیش از یک‌ششم کل واژه‌های نمونه را آغاز می‌کنند. در انتها، برای توضیح علت عدم وجود تفاوت معنادار بین دو دسته بسامدی باید به این نکته توجه داشت که واژه‌های رتبه‌ی ۱ تا ۱۰۰۰ در هر زبانی همگی «واژه‌های بسیار پربسامد» به‌شمار می‌روند (در مقایسه با ده‌ها هزار واژه‌ی کم‌بسامدتر در زبان). ممکن است تفاوت واقعی و معنادار فقط در مقایسه‌ی واژه‌های فوق‌العاده پربسامد (رتبه‌ی ۱–۵۰) با واژه‌های کاملاً کم‌بسامد (رتبه‌ی ۵۰۰۰+) پدیدار شود، نه در مقایسه‌ی دو نیمه‌ی یک بازه‌ی

نسبتاً همگن و پربسامد. این نکته می‌تواند مسیر پژوهش‌های بیشتر در این زمینه را پیشنهاد کند. از جنبه کاربردی، نتایج این پژوهش می‌تواند در مدل‌های بازشناسی گفتار، آموزش زبان فارسی، واج‌شناسی رایانشی و تدوین فرهنگ‌های بسامدی مورد استفاده قرار گیرد. همچنین این پژوهش گامی اولیه در بررسی رابطه میان بسامد واژگانی و ساخت آغاز واژه در فارسی محسوب می‌شود و می‌تواند زمینه را برای پژوهش‌های گسترده‌تر درباره تعامل بسامد، ساخت هجایی و سازماندهی واجی در زبان فارسی فراهم سازد. پژوهش حاضر گامی در جهت تبیین این تعامل است، اما دامنه‌ی آن به هزار واژه‌ی پربسامد محدود می‌شود. بررسی واژگان کم‌بسامد، مقایسه‌ی گونه‌های گفتاری و نوشتاری، تحلیل جایگاه‌های دیگر واج در واژه و انجام مطالعات بینازبانی می‌تواند تصویر کامل‌تری از نقش بسامد در سازماندهی واجی زبان ارائه دهد.

اعلامیه تعارض منافع و حمایت مالی: نویسندگان در خلال انجام این پژوهش حمایت مالی دریافت نکرده و هیچ گونه تعارض منافعی برای اعلام نداشته‌اند.

اعلامیه هوش مصنوعی مولد و فناوری‌های مبتنی بر هوش مصنوعی در فرایند نگارش

شمارش، دسته‌بندی و تحلیل تمامی داده‌ها توسط نگارنده انجام شده است. آزمون خی دو و محاسبه اندازه اثر کرامر با استفاده از نرم افزار اس پی اس اس نسخه ۲۷,۰,۱ اجرا شده است. تنها برای ویرایش و یک‌دست‌سازی متن از هوش مصنوعی کلاد (sonnet 5) استفاده شده است. مسئولیت علمی تمامی مطالب مقاله بر عهده نویسنده است.

تعارض منافع

بنا بر اظهار نویسندگان این مقاله تعارض منافع ندارد.

بیانیه دسترسی به داده‌ها

داده‌های پژوهش حاضر از طریق درخواست از نویسندگان قابل دسترسی است.

سپاسگزاری

از داوران محترم به خاطر ارائه نظرهای ساختاری و علمی سپاسگزاری می‌شود.

پی‌نوشت‌ها

۱.دلیل انتخاب این دو رده بسامدی، امکان مقایسه آن با پژوهش فنک –اتسلن (۲۰۰۱) است که در همین دو رده بر روی داده‌های زبان انگلیسی انجام شده است.

۲.لازم به ذکر است ثمره برای بررسی توزیع واج‌ها در انواع هجاهای زبان فارسی مدخل‌های فرهنگ فارسی معین را تحلیل کرده است.

۳. این نکته را داور محترم مقاله یاداوری فرمودند. از ایشان سپاسگزارم.

منابع

اسلامی، م.، رحیمی، ا.، و اسلامی، س. (۱۳۹۲). نظام آوایی زبان فارسی در آینه آمار .*زبان و زبان‌شناسی*، ۱۸. صص ۶۵–۹۰.

اسلامی،م. بی‌جن خان، م. (۱۳۸۲). واج‌آرایی و واژه‌گزینی. *مجموعه مقالات همایش واژه‌گزینی، فرهنگستان ادب و زبان فارسی.*

بی‌جن خان، م. (۱۳۸۴). *واج شناسی نظریه بهینگی*. انتشارات سمت.

ثمره، ی (۱۳۷۱). *نظام آوایی زبان فارسی* .تهران: مرکز نشر دانشگاهی.

Bijankhan, M. (2005). *Phonology Optimality theory*. Tehran: SAMT Publications. [In Persian].

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, J. (2023). What is usage-based linguistics? In M. Díaz-Campos & S. Balasch (Eds.), *The handbook of usage-based linguistics*. https://doi.org/10.1002/9781119839859.ch1

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In J. Kingston & M. Beckman (Eds.), *Papers in laboratory phonology* (Vol. 1, pp. 283–333). Cambridge: Cambridge University Press.

Eslami, M., & Bijankhan, M. (2003). *Phonological patterning and word formation*. In *Proceedings of the Conference on Word Formation*. Academy of Persian Language and Literature. (Reprinted in Bijankhan, 2005). [In Persian].

Eslami, M., Rahimi, A., & Eslami, S. (2013). Phonostatistics of Persian Phonological System. *Zabān va Zabānshenāsi*, 18, 65–90. [In Persian].

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 431–448). Amsterdam: John Benjamins.

Jakobson, R. (1968). *Child language, aphasia and phonological universals*. Mouton. (Original work published 1941).

Joo, I., & Hsu, Y. Y. (2025). Phonotacticon: A cross-linguistic phonotactic database. *Linguistic Typology, 29*(2), 405–431. https://doi.org/10.1515/lingty-2023-0094

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 229–254). Amsterdam: John Benjamins.

Kapnoula, E. C., Jevtović, M., & Magnuson, J. S. (2024). Spoken word recognition: A focus on plasticity. *Annual Review of Linguistics, 10*, 233–256.

Lu, Y., & Song, J. (2024). Effects of word frequency and familiarity on perceptual epenthesis. *Language Research, 41*, 75–107.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word recognition. *Cognition, 25*(1–2), 71–102.

Miller, C., & Aghajanian-Stewart, K. (2018). *A frequency dictionary of Persian: Core vocabulary for learners*. Routledge.

Samareh, Y. (1992). *The phonological system of Persian*. Tehran: Markaz-e Nashr-e Daneshgahi. [In Persian].

Smith, J. (2003). *Onset sonority constraints and subsyllabic structure*. Rutgers Optimality Archive. http://roa.rutgers.edu/

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Teachers College, Columbia University.