Journal of Linguistic and Rhetorical Studies
Volume 17, Consecutive Number 43, Spring 2026
Pages 1-7(research article)
Received: August 16, 2025 **Revised** October 10, 2025 **Accepted** October 15, 2025
Journal Homepage: https://rhetorical.semnan.ac.ir/?lang=en

Bidirectional Interaction of Psycholinguistics and Corpus Linguistics; Some Challenges Persian Frequency Dictionaries Meet

Masoumeh Mehrabi [1*] & Behrooz Mahmoodi-Bakhtiari [2]
1: Associate Professor of Linguistics at Ayatollah Boroujerdi University, Broujerd, Iran (Corresponding Author): m.mehrabi@abru.ac.ir
2: Associate Professor of Linguistics at Fine Art College of Tehran University, Thran. Iran.

1. Introduction

The interaction between psycholinguistics and corpus linguistics has gained increasing importance in recent decades, particularly in the design of linguistic resources such as frequency dictionaries. Psycholinguistics, concerned with the cognitive processes underlying language comprehension and production, heavily relies on well-designed linguistic materials for experiments—especially lexical decision tasks. Corpus linguistics, on the other hand, provides empirical data on actual language use, most notably frequency of occurrence, which is among the most influential predictors of word recognition speed. Despite the clear conceptual connection between the two fields, the integration of psycholinguistic principles into corpus design—especially frequency dictionaries—remains underdeveloped.

This study is motivated by two key questions:

(1) Why has the methodological interaction between psycholinguistics and corpus linguistics, despite their conceptual overlap, remained limited?

(2) Why are psycholinguistic principles insufficiently incorporated into existing Persian frequency dictionaries?

By comparing two major Persian frequency dictionaries—the Routledge *Frequency Dictionary of Persian* (Miller & Aghajanian-Stewart, 2018) and the *Frequency Dictionary According to a Written Corpus of Today's Persian* (Bijankhan & Mohseni, 2018)—the study identifies methodological shortcomings in both their corpus design and their usability for psycholinguistic

experimentation. The investigation is corpus-driven in nature: patterns observed in real linguistic data motivate hypotheses regarding cognitive processing, which are then examined through lexical decision tasks.

2. Literature Review

A large body of psycholinguistic research demonstrates that lexical decision times and other recognition measures are strongly influenced by several variables:

- Word frequency, one of the most powerful predictors of reaction time;
- Word prevalence, reflecting how widely a word is known within a speech community;
- Word length in letters or syllables;
- Age of Acquisition (AoA), which is highly correlated with frequency but captures developmental learning patterns;
- Similarity to other words, including neighborhood density;
- Concreteness and imageability, which reflect semantic richness.

These variables play an essential role in the design and interpretation of psycholinguistic experiments. For corpora or dictionaries to be useful in experimental settings, they must include or at least be compatible with these measures. However, most Persian frequency dictionaries fail to integrate such psycholinguistic parameters, reducing their usefulness for experimental design.

For a long time, researchers had few options for word frequency measures, with only one or two lists per language due to the labor-intensive counting of printed texts. The advent of digital texts simplified corpus compilation and word counting, typically drawing frequencies from nonfiction sources like books, newspapers, and periodicals (Zolfaghari & Hasini Kochaki, 2023). For undergraduates—commonly studied in psychology—corpora from television subtitles (Brysbaert & New, 2009), social media (Gimenes & New, 2016; Herdağdelen & Marelli, 2017), and blogs (Gimenes & New, 2016) provide the most effective measures, with combined sources outperforming individual ones. Traditional book-based frequencies suit older adults (Brysbaert & Ellis, 2016). Customized frequency lists reflecting participants' media and reading habits may offer further benefits (Johns, Jones, & Mewhort,

2016). Reliable lists require large corpora (≥20 million words), include syntactic information, and must be validated through mega-study data (Brysbaert, Mandera, & Keuleers, 2018).

In many psycholinguistic researches there is a need for (auditory or visual) stimuli which are matched across frequency of occurrence in order to minimize and control the frequency effect which can interfere with the other specified factors and variables. That is why frequency dictionaries are of significance and use in psycholinguistic researches, especially in researches which are mainly based on lexical decision tasks. In Persian there are two main frequency dictionaries (called here, RPFD and PFD) which have been used in some psycholinguistic researches by the authors of this article (Mehrabi, 2014; 2015; Mehrabi and Zaker, 2013; Mehrabi and Mahmoodi-Bakhtiari, 2020, 2021a, 2021b, 2022a,2022b; Mehrabi et al. 2021) which are mainly based on lexical decision tasks. Using this kind of frequency dictionaries they have confronted some difficulties listed in the article body.

3. Methodology

The study adopts a mixed-method approach, combining corpus analysis with experimental psycholinguistic testing. Two Persian frequency dictionaries—based on different corpora and compilation principles—serve as the main sources of frequency data. Their methodological characteristics, including corpus size, genre distribution, tokenization standards, and part-of-speech tagging practices, are critically evaluated.

To assess the adequacy of these dictionaries for psycholinguistic research, several lexical decision experiments were conducted. These experiments tested target words with multiple grammatical functions (e.g., “bozorg” ‘big’; “xub” ‘good/well’), each associated with different frequency counts depending on their part of speech. The experimental items were constructed using frequency values extracted from the dictionaries, and participants' reaction times were recorded.

Statistical analyses included Friedman’s test for repeated measures, followed by Wilcoxon signed-rank tests for pairwise comparisons. These analyses aimed to determine whether frequency differences across grammatical categories corresponded to measurable differences in processing time.

4. Results and Discussion

The results reveal significant discrepancies between the frequency values reported in the Persian dictionaries and the actual psycholinguistic behavior of speakers. For several high-frequency Persian words used across multiple grammatical categories (adjective, noun, adverb), the dictionaries often failed to differentiate between the frequencies of their distinct functions. This lack of part-of-speech–specific frequency data hindered accurate predictions of processing difficulty.

For example, the adjective form of *xub* (‘good’) was processed significantly faster than its adverbial or nominal forms, consistent with the higher frequency of adjectival usage in contemporary Persian. However, the dictionaries either did not separate frequencies by grammatical category or reported incomplete values. As a result, they could not accurately predict processing times.

Similarly, the word *bozorg* (‘big’) demonstrated substantial variation across adjectival, adverbial, and comparative uses. Experimental results showed that participants processed the adjective form more quickly, but the dictionaries did not reflect these distinctions with adequate granularity.

Several broader issues emerged:

1. Corpus imbalance: Overrepresentation of formal written registers reduced the ecological validity of frequency counts for everyday spoken language.
2. Lack of psycholinguistic variables: Neither dictionary incorporated AoA, prevalence, concreteness, or neighborhood density.
3. Tagging inconsistencies: Multi-function Persian words were often tagged under a single category, obscuring meaningful grammatical variation.
4. Insufficient metadata: Lack of information about corpus composition limits replicability and interpretation.

5. Conclusion

The study demonstrates a clear need for stronger integration between corpus linguistics and psycholinguistics in the creation of Persian linguistic resources—especially frequency dictionaries.

While corpus linguistics offers invaluable empirical data, the current Persian dictionaries fail to incorporate crucial psycholinguistic variables, reducing their applicability in experimental contexts. The lexical decision experiments reveal that processing differences across grammatical categories align with actual usage patterns but are not adequately captured in existing resources.

To address these shortcomings, the study suggests several improvements:

- developing multi-register corpora incorporating spoken and informal data;
- including psycholinguistically relevant variables such as AoA, concreteness, prevalence, and neighborhood density;
- providing part-of-speech–specific frequency counts;
- increasing transparency regarding corpus composition and annotation.

Such enhancements would enable the creation of cognitive-oriented, experimentally valid frequency dictionaries that better serve both linguistic researchers and Persian language learners. Ultimately, closer cooperation between the two fields promises more accurate tools, more reliable experimental materials, and a deeper understanding of how Persian is processed in the mind.

Keywords: Corpus Linguistics, Frequency Dictionary, Psycholinguistics, Persian, Frequency of occurrence.

References

Bijankhan, M. and Mohseni, M. (2018). *Frequency Dictionary According to a Written Corpus of Today Persian Language*. Tehran: University of Tehran.

Brysbaert, M., & Ellis, A. W. (2016). Aphasia and age of acquisition: Are early-learned words more resilient? *Aphasiology, 30*, 1240–1263.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word-frequency effect in word processing: An updated review. *Current Directions in Psychological Science, 27*(1), 45–50.

Gimenes, M., & New, B. (2016). WorldLex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods, 48*, 963–972.

Herdağdelen, A., & Marelli, M. (2017). Social media and language processing: How Facebook and Twitter provide the best frequency estimates for studying word recognition. *Cognitive Science, 41*, 976–995. https://doi.org/10.1111/cogs.12392

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2016). Experience as a free parameter in the cognitive modeling of language. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2291–2296). Cognitive Science Society.

Mehrabi, M. (2014). Lexical information of Persian transitive verbs during sentence comprehension. *Language Related Research, 5*(1), 271–295.

Mehrabi, M. (2015). The effect of Farsi verb representational complexity on processing time during listening comprehension. *Journal of Researches in Linguistics, 7*(1), 77–92.

Mehrabi, M., & Mahmoodi-Bakhtiari, B. (2020). A comparative study of comprehension of evidentiality in Persian, English, and Turkish: A psycholinguistic approach. *Comparative Linguistic Researches, 20*, 243–257.

Mehrabi, M., & Mahmoodi-Bakhtiari, B. (2021a). The psychological reality of the evidentiality hierarchy in Persian during sentence listening comprehension. *Language Related Research, 12*(2), 539–566. https://doi.org/10.29252/LRR.12.2.17

Mehrabi, M., & Mahmoodi-Bakhtiari, B. (2021b). Mental representations of Persian and English absolute and relative tenses: A contrastive psycholinguistic approach. *Journal of Researches in Linguistics, 13*(1), 89–112.

Mehrabi, M., & Mahmoodi-Bakhtiari, B. (2022a). Representational complexity of Persian absolute tenses during listening comprehension. *Scientific Journal of Language Research, 13*(41), 55–80.

Mehrabi, M., & Mahmoodi-Bakhtiari, B. (2022b). Representational complexity of Persian relative tenses during listening

comprehension. *Language Related Research, 13*(2), 1–32. https://doi.org/10.52547/LRR.13.2.1

Mehrabi, M., Mahmoodi-Bakhtiari, B., & Vaezi, H. (2021). Auditory perception of Persian interrogatives from a psycholinguistic approach and its application in Persian teaching. *Journal of Teaching Persian to Speakers of Other Languages, 10*(2), 137–156.

Mehrabi, M., & Zaker, A. (2013). Lexical information of Persian transitive verbs during sentence comprehension. *Iranian Studies, 46*(6), 959–971.

Miller, C., and Aghajanian-Stewart, k. (2018). *A Frequency Dictionary of Persian; Core Vocabulary for Learners.* London: Routledge.

Zolfaghari, H., & Hasani Kochaki, E. (2023). Stylistic analysis of the prose of Qajar-period newspapers. *Journal of Linguistic and Rhetorical Studies, 14*(31), 65–98.

About the Authors:

Masoumeh Mehrabi is a graduate in Linguistics and a faculty member at Ayatollah Ozma Borujerdi University. Her main areas of teaching and research include psycholinguistics, (critical) discourse analysis, the linguistic study of Persian literature, and the Persian language.

Behrooz Mahmoudi-Bakhtiari is a graduate in Linguistics and a faculty member at University of Tehran. His main areas of teaching and research include dramatic discourse analysis, theater and film semiotics, Persian language education, and Iranian dialects.