Persian Loanwords in Turkish: A Diachronic Study

Literature Survey Report

Mohammadzadehazarabadi, Sadra 22101018

Türkmen, Celal Salih 22102498

March 23, 2025

Introduction

The relationship between Farsi (Persian) and Turkish is one of the most significant cases of language contact in Eurasia. Since the Ottoman era, Persian-derived terms entered Turkish and influenced areas as varied as literature, administration, and spoken usage. However, ultimately, there came to be changes, replacements, or semantic narrowing of many of these loanwords through important language reforms and sociological changes. Advances in Natural Language Processing (NLP) now enable scholars to trace these lexical and semantic developments quantitatively, especially through diachronic word embedding techniques. This survey synthesizes the most relevant studies, covering historical linguistics, corpus-based methods, and computational linguistics, to form a foundation for investigating Persian loanword evolution in Turkish.

Diachronic Computational Linguistics

Diachronic computational linguistics has been a highly significant methodology in semantic shifts. For instance, Kutuzov *et al.* introduce a rich amount of diachronic word embeddings and semantic shifts, which explain how prediction-based word embedding models can monitor lexical semantics

over various time frames [1]. This research will lead us to evaluate the pertinent methodologies to compute the semantic shifts with word embeddings.

Yazar and Kutlu also propose different methods of incorporating alignment trained from Turkish corpora of different decades [2]. Two different methods they use are "Orthogonal Procrustes" and "Spearman correlation". These methods will help us in determining synonyms or replacement words over time. With these methods, we can also determine Persian loanwords changing in Turkish language.

Hamilton's work shows how the meanings of words shift with the use of word embeddings [3]. There are two paradigms outlined: rarer words shift their meaning faster and words that are polysemic are shifting quicker. These rules will help us in understanding the differentiation of words over time. For our work on Turkish loanwords from Persian, we are able to conduct this approach when studying how such words evolved meanings over time. By comparing old Turkish texts with word embeddings, we can see whether Persian loanwords follow these rules or exhibit other patterns.

Contact Between Persian and Turkish

It is worth to study the Persian–Turkish connection since they always influenced each other throughout history. If one examines the sentence structures of Turkish and Persian, the similarities of Turkish and Persian, especially structural, are remarkable. This is particularly surprising since Turkish is an additive language and belongs to the Ural–Altai language family, whereas Persian is a part of the Indo-European language family [4]. Most of the loans were made in the Ottoman period when Persian was a prestigious literary language. One such older work from 1967 mentions Anatolian Turkish and classifies borrowings as literary and colloquial varieties [5]. It argues that "Ottoman Turkish literature, from its beginning to the early twentieth century, was largely dependent on classical Persian literature."

Güzel's research on Khalaj (a Turkic language spoken in Iran) shows how lexical and grammatical elements from Persian incorporate various cultures and their dialects [6]. Even though Khalaj is a different language from Anatolian Turkish, the mechanisms of borrowing remain identical: younger people

tend to borrow more new Persian words, while older communities tend to carry on with preceding forms. This borrowing pattern shows that the use of Persian loanwords in Turkish is influenced by several factors which contribute to linguistic interaction.

Besides the literary and conversational borrowings, Turkish also borrowed many Farsi loanwords for bureaucracy and government. In the Ottoman era, titles of departments in the government, official titles, and state decrees frequently had Persian vocabulary. They preserved their original sense or acquired new senses afterwards. With the Turkish language reforms in the 1920s, numerous Persian loans were deliberately substituted with Turkish versions [2].

Diachronic Resources for Turkish

Robust diachronic corpora are crucial for systematically studying lexical change. Yazar *et al.* developed *Turkronicles*, a large-scale diachronic corpus based on the Official Gazette of Türkiye. For this purpose, they analyzed over 45,000 documents that have been from the 1920s onward [2]. By breaking texts into decades, they illustrate how the vocabularies of two different periods diverge more as time between them increases, and recently emerged Turkish words replace their old equivalents [2]. This corpus records significant events like the transition from Perso-Arabic to Latin script and the simplification campaign against Arabic and Persian terms. Such developments are visible by examining frequencies, contexts, and neighbors of some Persian-origin lexemes over decades.

Furthermore, Tohma and Kutlu point out technical issues for Turkish NLP, describing how its agglutinative nature and vowel harmony pose tokenization, stemming, and lemmatization issues [7]. Researchers who work with diachronic corpora need to overcome the morphological complexities in mapping historical Farsi loanword forms to their modern counterparts (e.g., changes in spelling from *kitab* to *kitap*). We will use this paper to analyze the potential issues we may encounter while working on Turkish corpora.

Cross-Lingual Embedding Alignment

Beyond simple monolingual solutions, cross-lingual word embedding alignment can help us to understand new ways to contrast loanwords in Turkish with their target equivalents in Persian. Artetxe *et al.*'s work gives us *VecMap*, a technique that learns bilingual relations without massive parallel corpora, effectively combining the two language embedding spaces [8]. Doing this for Turkish–Persian can potentially help us to understand whether the sense of any particular Persian-origin word in Turkish is no longer equivalent to the sense of its equivalent in contemporary Persian.

Amtrup *et al.* highlight the need for robust NLP resources, such as dictionaries and morphological analyzers, for Persian and Turkish in their English translation systems [9]. Their prototypes show that morphological complexity and script handling are significant challenges in developing these systems.

Khashabi *et al.*introduce *ParsiNLU*, a benchmark designed to improve natural language understanding (NLU) research in Persian [10]. ParsiNLU has a diverse set of tasks, such as reading comprehension and textual entailment, with over 14,500 human-annotated examples. The authors also compare monolingual and multilingual pre-trained language models to human baselines on this benchmark. Their findings are important observations on the challenges as well as potential opportunities in Persian language understanding.

Semantic Shifts in Loanwords

The history of individual Farsi borrowings in Turkish is diverse. Some words continued to be used in everyday language (e.g., *divan*), retaining or somewhat limiting their sense, while others gained extended or metaphorical senses. Nourzaei's diachronic study of *-ak* suffixes in Persian shows how evaluative structures can develop into grammatical markers over centuries [11]. Therefore, it suggests that morphological or semantic analysis might take place in Turkish as well. This analysis can help us for our research on analyzing Farsi loanwords in Turkish.

Ziaei et al. state that Persian itself has undergone structural transformations with time, (i.e., syntactic and semantic) which can also affect cross-lingual comparisons [12]. For example, a Persian word in the 15th century can vary from its modern-day use.

Baltazani's "Greek spoken prosody" research also suggests how phonetic development can be simultaneous with loanword transmission, and demonstrates methodologies we can use for the Persian language. [13]. Therefore, comparing historical dictionaries and modern-day corpora for Turkish and Persian provides an insight for whether Turkish use has diverged from current state of Persian evolution.

Conclusion

A comprehensive history of research points to the long history of Farsi's enormous impact on Turkish language, with computational and analytical power of NLP. This will allow us to analyze the development of these loanwords more precisely than it has ever been possible. Diachronic word embeddings allow quantitative monitoring of lexical evolution, and cross-lingual alignment methodologies allow direct comparisons of a loanword's meaning in Turkish to the original in Farsi. Sources like Turkronicles record such critical turning points, showing how language reforms accelerated lexical replacement.

By integrating these corpus-based approaches with structural investigation of Turkish and Persian, we are better able to determine how particular Farsi loanwords have survived or changed over the centuries. This interdisciplinary strategy—crossing computational semantics, sociolinguistics, and diachronic linguistics—holds the potential to provide deeper understanding into the history of Persian-derived vocabulary in Turkish. Ultimately, such study will increase our understanding of language contact and the reasons behind it.

References

[1] A. Kutuzov, L. Øvrelid, T. Szymanski, and E. Velldal, "Diachronic word embeddings and semantic shifts: A survey," *Proceedings of the 27th International Conference on Computational Linguistics (Tutorials)*, pp. 1–11, 2018. [Online]. Available: https://arxiv.org/abs/1806.03537

[2] T. Yazar, M. Kutlu, and I. K. Bayırlı, "Turkronicles: Diachronic resources for the fast evolving Turkish language," *arXiv preprint*, arXiv:2405.10133, 2024. [Online]. Available: https://arxiv.org/abs/2405.10133

[3] W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Diachronic word embeddings reveal statistical laws of semantic change," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1489–1501, 2016. [Online]. Available: https://cs.stanford.edu/people/jure/pubs/diachronic-acl16.pdf

[4] O. Gedik, "The interaction between Turkish and Persian in the context of language-culture relationship," *Journal of Turkish Studies*, vol. 11, no. 4, pp. 351–369, 2020. [Online]. Available: https://www.academia.edu/114840262

[5] A. Tietze, "Persian loanwords in Anatolian Turkish," *Middle Eastern Philology*, vol. 2, no. 1, pp. 11–29, 1967. [Online]. Available: https://www.azargoshnasp.net/languages/Persianpersianloanwordsistanbulturkish.pdf

[6] E. Güzel, "Some Observations on Persian Copies in Khalaj: Case of Talkhab Dialect," *International Review of Turkic Linguistics*, vol. 7, no. 3, pp. 45–57, 2022. [Online]. Available: https://dergipark.org.tr/tr/download/article-file/3350737

[7] K. Tohma and Y. Kutlu, "Challenges encountered in Turkish natural language processing studies," *Natural and Engineering Sciences*, vol. 5, no. 3, pp. 204–211, 2020. [Online]. Available: https://dergipark.org.tr/en/download/article-file/1421687

[8] M. Artetxe, G. Labaka, and E. Agirre, "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings," *Proceedings of the 56th Annual Meeting of the Association*

for Computational Linguistics, pp. 789–798, 2018. [Online]. Available: https://aclanthology.org/P18-1073.pdf

[9] J. W. Amtrup, K. Megerdoomian, and R. Zajac, "Rapid development of translation tools: Application to Persian and Turkish," *Proceedings of the ACL Workshop on Machine Translation*, pp. 1–8, 2000. [Online]. Available: https://aclanthology.org/www.mt-archive.info/Coling-2000-Amtrup.pdf

[10] D. Khashabi *et al.*, "PARSINLU: A suite of language understanding challenges for Persian," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1147–1162, 2021. [Online]. Available: https://doi.org/10.1162/tacl_a_00419

[11] M. Nourzaei, "Diachronic development of the K-suffixes: Evidence from classical new Persian, contemporary written Persian, and contemporary spoken Persian," *Iranian Studies*, vol. 55, no. 1, pp. 115–160, 2022. [Online]. Available: https://doi.org/10.1017/irn.2021.27

[12] S. Ziaei, B. Hadian, V. Rezai, and M. Jafari, "Diachronic study of information structure in Persian," *Journal of Researches in Linguistics*, vol. 15, no. 2, 2024. [Online]. Available: https://jrl.ui.ac.ir/article_27849.html?lang=en

[13] M. Baltazani, "Intonation of Greek–Turkish contact: A real-time diachronic study," *Speech Prosody*, 2020. [Online]. Available: https://www.isca-archive.org/speechprosody_2020/baltazani20_speechprosody.pdf