In-Depth Analysis of Arabic-Origin Words in the Turkish Morpholex

Mounes Zaval[1,2], **Abdullah Ihsanoğlu**[2], **Asım Ersoy**[1], **Olcay Taner Yıldız**[2]
[1]Sestek [2]Özyeğin University
{mounes.zaval, asim.ersoy}@sestek.com, abdullah.ihsanoglu@ozu.edu.tr
olcay.yildiz@ozyegin.edu.tr

Abstract

MorphoLex is an investigation that focuses on analyzing the roots, prefixes, and suffixes of words. Turkish Morpholex, for example, analyzes 48,472 Turkish words. Unfortunately, it lacks in-depth analysis of the Arabic-origin words, and does not include their accurate and correct roots. This study analyzes Arabic-origin words in the Turkish Morpholex, annotating their roots, morphological patterns, and semantic categories. The methodology developed for this work is adaptable to other languages influenced by Arabic, such as Urdu and Persian, offering broader implications for studying loanword integration across linguistic contexts.

1 Introduction

Morphological lexicons (Arıcan et al., 2022; Sánchez Gutiérrez et al., 2017; Mailhot et al., 2019) play a vital role in understanding the structure of languages, particularly in agglutinative languages like Turkish, where complex words are formed through the combination of multiple morphemes. By analyzing these structures, we can gain insights into how words are constructed. Arıcan et al. (2022) built the first Turkish morphological lexicon that includes an analysis of 48,472 words categorized by their roots, prefixes, and suffixes. As Turkish contains some loanwords from languages such as Arabic and Persian, the analysis of those words needs to follow the grammar of that language. Turkish Morpholex, however, does not process the loanwords accurately.

In this work, we address this problem and analyze the Arabic loanwords to Turkish according to the Arabic grammar. In addition to finding the accurate roots for those words, we analyzed the words across other dimensions as well, such as morphological pattern and semantic categories. We open-source all the annotations done in this work[1].

The methodology used in this study not only deepens our understanding of Turkish Morpholex but also provides a framework that can be applied to other languages with significant Arabic influence, including Urdu and Persian. This highlights the potential for broader applications of this research in multilingual and cross-linguistic studies.

2 Literature Review

The investigation of Arabic roots in the Turkish language, particularly through extensions of the Turkish WordNet, builds on a foundation of research in morphological lexicons and linguistic borrowings. The MorphoLex Turkish project (Arıcan et al., 2022) provides a significant contribution by developing a lexicon for Turkish morphology, inspired by earlier work on morpholexical resources for languages like English (Sánchez Gutiérrez et al., 2017) and French (Mailhot et al., 2019). Studies on Turkish morphological analysis highlight its unique agglutinative structure, which relies heavily on suffixation. However, Turkish has also been profoundly influenced by Arabic due to historical contact, leading to the adoption of numerous loanwords, especially in religious, legal, and administrative contexts.

Existing research in loanwords, such as Serigos (2017)'s work on Anglicisms in Spanish, introduces the concept of semantic specificity. Serigos' study reveals that loanwords often carry more nuanced or specific meanings compared to their native counterparts, a hypothesis that can be extended to Arabic loanwords in Turkish. For example, the Arabic-origin word in Turkish *Adalet* (عدالة in Arabic and

[1] https://github.com/mouneszawal/turkish-lexicon-arabic-roots

Processings of the 1st Workshop on NLP for Languages Using Arabic Script (AbjadNLP), pages 29–36
January 19, 2025. ©2025 Association for Computational Linguistics

Justice in English) has a specific meaning compared to the native Turkish word *Doğruluk*, which is a broader term that can mean correctness, honesty, or truthfulness in general, without necessarily referring to legal justice.

Alshammari and Alshammari (2020) conducted an in-depth analysis of 250 Turkish loanwords of Arabic origin, shedding light on the phonological and morphological adaptations these words undergo during their integration into Turkish. This study highlights the impact of native speaker knowledge on the borrowing process and offers a detailed exploration of phonological modifications, morphological markings, and compound forms in Arabic-origin loanwords.

Stachowskı (2020) investigated phonetic renderings Arabic- and Persian-origin words in Turkish, analyzing 1,748 loanwords to identify both typical and unusual phonetic changes during the borrowing process. The research provides insights into how foreign words adapt to the Turkish phonological system, offering a deeper understanding of linguistic integration mechanisms.

Furthermore, Procházka (2009) investigated Turkish loanwords in Arabic, offering a comparative perspective on the bidirectional nature of linguistic borrowing between Turkish and Arabic. The study sheds light on how Turkish words are adapted into Arabic, enriching the understanding of cross-linguistic influence.

Moreover, Fattakhova and Mingazova (2015) explored how Arabic loanwords have been integrated into Tatar and Swahili. Both languages share similarities in loanword assimilation due to their agglutinative nature but exhibit differences, such as Swahili's postposition of adjectives and Tatar's compound verbs. The study highlights the diverse semantic fields Arabic loanwords cover, such as religion, science, and culture, revealing the historical impact of Arabic in shaping both languages' lexicons.

There are many studies that examine Arabic loanwords in Turkish and other languages, focusing on their linguistic integration, phonological and morphological adaptation (Al-Hashmi, 2016; Perry, 1984; Corriente, 2008; Sayahi, 2005). These studies highlight how Arabic-origin words have been absorbed into recipient languages, often filling semantic gaps and contributing to the linguistic richness of languages like Turkish, Spanish, Tatar, and many others.

Building on these works, this study aims to further explore how Arabic-origin words integrate within the Turkish language by enriching the root-based analysis in the Turkish Morpholex. This work contributes to understanding the semantic and morphological interactions between Arabic and Turkish, as well as the mechanisms by which Arabic loanwords have been absorbed and adapted into the modern Turkish lexicon.

3 Turkish Morpholex

Since Turkish is an agglutinative language, where words are formed by adding suffixes to a base root, Arıcan et al. (2022) emphasizes the importance of analyzing Turkish separately from other languages like English and French, which have different morphological structures. In their work, they develop a Turkish Morpholex, which is morphological lexicon for Turkish that contains 48,472 words, taken from the Turkish KeNet wordnet (Ehsani et al., 2018; Bakay et al., 2021), analyzed based on their roots, prefixes, and suffixes. The creation of this lexicon involved manual annotation, where each word is carefully analyzed for its semantic and morphological structure, unlike the case for the English and French ones where all the analysis was not done manually.

Turkish language originally does not have prefixes. However, prefixes exist and are used currently in Turkish due to the influence of other languages on Turkish such as Arabic, Persian, French, and English. The existence of such loanwords makes the task harder when building morphological lexicons since those would require the analysis of the loaned word according to that language's grammar. Arıcan et al. (2022), for instance, did not analyze the Arabic loanwords in depth and treated them as any other Turkish words. For example, for Arabic-origin word adaletli (fair), they only remove the Turkish suffix (li), which makes the word adalet (justice) an adjective, and consider the word adalet to be the root. Therefore, we analyze in this work those Arabic-origin words in depth to increase the accurateness and depth of the Turkish Morpholex.

4 Arabic Morphology

Arabic is a semitic language, and its morphology is quite different from that of Turkish. While Turkish is an agglutinative language, Arabic uses a root-and-pattern system where words are constructed by formalizing roots into specific patterns.

Arabic words typically derive from triliteral or quadriliteral roots that convey the core meaning. Roots are combined with specific patterns, involving fixed vowels and sometimes additional consonants, to form words in different grammatical categories, such as verbs, nouns, and adjectives. For instance, the root ”ك-ت-ب” (k-t-b, ”to write”) can form words like ”كَتَبَ” (kataba, ”he wrote”) and ”كتّاب” (kitāb, ”book”) based on different patterns. This root-and-pattern system allows for a vast number of word forms derived from a single root.

In addition to roots and patterns, Arabic morphology involves the use of prefixes, suffixes, and infixes to modify words grammatically. Prefixes and suffixes indicate tense, voice, plurality, and other grammatical features, while internal vowel changes (infixes) often reflect tense or voice changes in verbs. For example, the verb ”كَتَبَ” (kataba, ”he wrote”) changes to the passive form ”كُتِبَ” (kutiba, ”it was written”). Understanding these modifications is essential for determining a word’s root and meaning.

Arabic words can be categorized into verbs, nouns, adjectives, and particles, with each category following specific morphological rules. Verbs, for example, change according to tense, voice, and mood, while nouns reflect gender, number, and definiteness. Derivation, or Ishtiqaq, is a key feature of Arabic, where multiple related words are derived from a single root. For instance, from the root ”ع-ل-م” (‘-l-m, ”to know”), we get words like ”عَلَّمَ” (‘allama, ”to teach”) and ”علوم” (‘ulum, ”sciences”).

The process of identifying the root of an Arabic word involves stripping away affixes and recognizing weak letters that may change form or disappear in different word structures. This process is crucial in understanding the word’s meaning and forming new words from the same root.

5 Annotation

We initially identified Arabic-origin words found in the Turkish Morpholex by utilizing the official digital dictionary of the Turkish Language Association (TDK)[2], which provides information about the etymological roots of words. We ended up with 4,687 unique words of Arabic-origin according to TDK’s classification.

Subsequently, we started the manual annotation and analysis of each word, drawing primarily from the Riyadh Dictionary[3], a contemporary digital resource for the Arabic language. For some instances, we also consulted the Doha Dictionary[4], another Arabic digital lexicon.

The annotation process, however, presented several challenges. A significant portion of these Arabic-origin words entered the Turkish lexicon during periods of Ottoman rule over Arabic-speaking territories. As a result, many of these terms are now considered outdated in modern Arabic. In some cases, words had experienced a complete shift in meaning, while in others, the terms had been entirely abandoned. Due to these changes, it was often difficult to locate the exact words in contemporary Arabic dictionaries. To overcome this, we had to identify Arabic words with similar morphological and semantic characteristics to complete the annotation.

To address semantic shifts, we relied on historical and contemporary Arabic lexicons, such as the Riyadh and Doha dictionaries, to trace the original meanings of words. For example, the Turkish word ”adalet” (justice) retains its semantic alignment with the Arabic root ”ع-د-ل”, while the word ”şebabet” (youth) has no direct Arabic equivalent but derives from the Arabic root ”ش-ب-ب”. Orthographic changes were handled by identifying consistent patterns of adaptation, such as the omission of weak letters or changes in vowel placement, en-

[2] https://sozluk.gov.tr/

[3] https://dictionary.ksaa.gov.sa/

[4] https://www.dohadictionary.org/

suring accurate root identification.

Three primary challenges emerged during the annotation process:

- Obsolete Words: many Arabic-origin words in Turkish are no longer in active use in modern Arabic. For these, we identified semantically similar roots using historical texts.

- Turkish-Neologisms: some Turkish words, like ”şebabet,” were created using Arabic morphological patterns but have no Arabic counterpart. These were annotated to reflect their hybrid nature.

- Compound Words: words like ”alelacele” (hastily), which combine multiple Arabic roots, were annotated with detailed notes on their composition.

During the annotation process, some words classified as Arabic-origin by the TDK were found not to be of Arabic origin upon further investigation. For example, terms such as *Patlıcan* (eggplant) and *Sabun* (soap) were incorrectly categorized as Arabic-origin. These words were excluded from the annotation process, and their misclassification was documented.

The annotations were carried out by the first three authors, all of whom are native Arabic speakers and fluent in Turkish. Their linguistic expertise ensured a deep understanding of both Arabic roots and Turkish adaptations. To maintain consistency, each annotator independently reviewed a subset of the words, and any disagreements were resolved collaboratively during weekly discussions. This collaborative approach ensured that the final annotations were accurate and reflective of both languages' morphological and semantic systems. The annotation task was evenly distributed among the three annotators, resulting in the successful annotation of 3,855 Turkish words from the total of 4,687 identified Arabic-origin words. Due to time constraints, 338 words were left for future analysis. Each annotated word which include its Arabic root (جذر), morphological pattern (وزن - wazn), and semantic category (قسم الكلمة).

To evaluate the accuracy of our annotations, we conducted a pilot study with 100 randomly selected words, achieving 93% agreement between the annotated roots and the consensus reached among the annotators. This process ensured a high degree of reliability in our dataset.

6 Statistics

Arabic Roots	
# Distinct Arabic Roots	1430
# Source Turkish Roots	3855

Table 1: Turkish roots linked to distinct Arabic roots

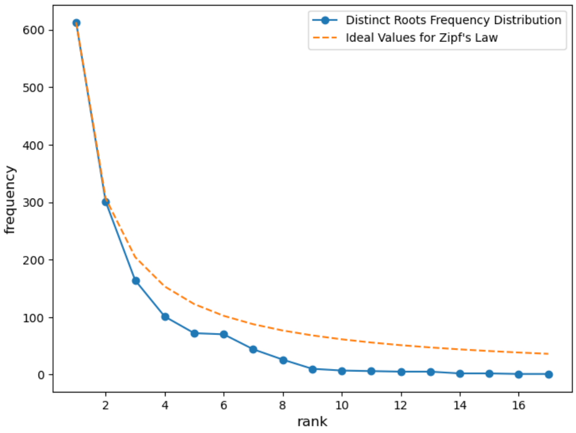

Figure 1: Distribution of the distinct Arabic roots compared to the ideal zipf's law values.

Table 1 provides an overview of the number of distinct Arabic roots and their corresponding Turkish source roots. The table reveals that there are 1,430 distinct Arabic roots, which their frequency distribution quite follows the ideal Zipf's law (Human, 1949) values as shown in Figure 1, associated with 3,855 Turkish roots in total. This suggests a significant lexical borrowing from Arabic, indicating the deep historical and cultural connections between the Arabic and Turkish languages. The fact that 3,855 Turkish words are connected to these 1,430 Arabic roots highlights the Arabic influence on the Turkish vocabulary.

The most common Arabic roots, shown in Table 2, are some specific Arabic roots that have the highest number of Turkish derivatives. For example, the Arabic root قوم is connected to 18 Turkish words, including Takvim (calendar), Kıvam (consistency), and Kayyum

Arabic Root	# Of Words	Meaning	Example Words
قوم	18	Refers to standing, rising, or establishing. It covers meanings such as to stand up, rise, set up, lead, establish, or correct.	takvim, kıvam, kayyum
حكم	16	Tied to judgment, wisdom, or authority. It includes ruling, governing, giving verdicts, and acting with wisdom.	mahkeme, hikmet, hakem
ملك	16	Associated with ownership, control, or kingship, signifying possession, dominion, power, authority, and being a king.	emlak, mülk, melek
حول	15	Focuses on transformation, movement, or change, covering concepts like shifting, transferring, or circling.	tahavvül, mütehavvil, istihale
عرض	15	Deals with presenting, displaying, or exposing. It can also refer to width or breadth and encompasses concepts like honor or reputation.	arz, maruz, taarruz
ولي	14	Focuses on closeness, support, and guardianship, including meanings such as protecting, being close, allying, or acting as a guardian.	vali, vilayet, mütevelli
حقق	13	Relates to achieving or realizing, implying the act of making something true or bringing it into existence.	elhak, hakikat, hakiki
قدر	13	Relates to measuring, determining, or decreeing. It also signifies power, capability, fate, or predestination.	kadar, kadir, kudret
عرف	13	Involves knowledge or recognition, implying knowing, recognizing, or understanding.	muarefe, örf, tarif
جمع	13	Relates to gathering or collecting, implying the act of bringing together or assembling.	cami, camia, cemaat
حلل	13	Encompasses resolving, analyzing, or making something permissible. It can mean to untie, explain, or make lawful.	mahal, mahalle, inhilal

Table 2: Most common Arabic roots along with Turkish example words.

(guardian). Other roots such as ملك (related to ownership or kingship), and عرض (meaning ”offer” or ”show”) each is related to several Turkish word.

We also show the most common semantic categories in Table 3, categorizing the Arabic-rooted words in Turkish by grammatical function with examples of Turkish words for each category. The most frequent category is معني اسم (meaning noun), with 1,789 occurrences, including words like Abes (absurd) derived from the Arabic root عبث (meaning ”nonsense” or ”absurdity”). Other categories include اسم ذات (concrete noun), and صفة فاعل (Subjective Adjective), صفة مفعول (Objective Adjective), and صفة مشبهة (Comparable Adjective), each illustrating the variety of ways Arabic roots are integrated into Turkish vocabulary. These categories reflect how Arabic words were adapted not only semantically but also grammatically into Turkish, indicating a sophisticated linguistic integration process. Similarly, we show in

Semantic Category	Frequency	Turkish Word	Arabic Root
اسم معنى (Meaning Noun)	1789	Abes, acayip	عبث, عجب
اسم ذات (Concrete Noun)	782	Şafak, acemi	شفق, عجم
صفة فاعل (Subjective Adjective)	460	Muavin, acil	عون, عجل
صفة مفعول (Objective Adjective)	284	Muaf, ceriha	عفو, جرح
صفة نسبية (Comparable Adjective)	145	Zayıf, acuze	ضعف, عجز
صفة مستوية (Attributive Adjective)	144	Acem, adedi	عجم, عدد
صفة مبالغة (Exaggerated Form)	45	Abus, acul	عبس, عجل
اسم مكان (Place Noun)	40	Mahal, mahalle	حل, حلل
اسم مرة (Instance Noun)	20	Gamze, gazve	غمز, غزو
فعل (Verb)	17	Acaba, ahraz	عجب, خرس
اسم آلة (Instrument Noun)	14	Makas, mastara	قص, سطر
اسم مبهم (Ambiguous Noun)	12	Badehu, fevk	بعد, فوق

Table 3: Most common semantic categories with example Turkish words.

Morphological Pattern (wazn)	Frequency	Turkish Word	Arabic Root
تفْعِيل (Taf'īl)	217	tabir	عبر
فعْل (Fa'l)	192	af	عفو
فَاعِل (Fā'il)	133	acil	عجل
مَفْعُول (Maf'ūl)	133	mağdur	غدر
إِفْعَال (If'āl)	124	ibraz	برز
تَفَعُّل (Tafa'ul)	115	taaffün	عفن
فَعِيل (Fa'īl)	111	afif	عفف
اِفْتِعَال (Ift'āl)	106	içtihat	جهد

Table 4: Most common morphological patterns with example Turkish words.

Table 4 the most common morphological patterns with example Turkish words.

In summary, these tables demonstrate the profound influence of Arabic on Turkish, showing how many Turkish words have been derived from Arabic roots and illustrating the rich linguistic interchange between the two languages.

7 Discussion

The methodology developed in this study can be adapted for languages like Urdu and Persian, which share similar influences from Arabic. For example, Urdu's reliance on Arabic morphological patterns could benefit from a similar annotation process to enrich its morpholexical resources. By demonstrating the scalability of our approach, this study provides a foundation for analyzing Arabic-origin words across diverse linguistic contexts.

The integration of Arabic-origin words into Turkish reflects a unique interplay between two morphological systems. Words like ”adaletli” illustrate how Turkish suffixation adapts Arabic roots while maintaining their core semantic properties. This insight could guide further research on the morphological interactions between agglutinative and Semitic languages.

Additionally, the findings contribute to understanding how Arabic-origin words are morphologically integrated into Turkish grammar. While Arabic employs a root-and-pattern system, Turkish transforms these roots by applying its suffixation processes, adapting them to its agglutinative structure. This study also demonstrates how Turkish retains Arabic morphological patterns (e.g., *Taf'īl*, *Fa'l*) or modifies them to align with its linguistic framework. Semantic adaptations reveal how borrowed words are aligned with Turkish cultural and linguistic contexts, sometimes resulting in hybrid structures like *şebabet*, which have no direct Arabic equivalent.

By documenting these processes, the study highlights the role of Arabic-origin words in enriching Turkish vocabulary across domains like law, administration, and science. Furthermore, the annotated dataset serves as a valuable resource for enhancing computational models of Turkish grammar, enabling more accurate processing of loanwords in natural language processing (NLP) applications. These findings provide a broader understanding of cross-linguistic borrowing and its impact on language evolution.

8 Conclusion

In conclusion, this study highlights the critical role of Arabic-origin words in enriching the Turkish language, addressing a significant gap in the existing Turkish Morpholex. The insights gained extend beyond Turkish, offering a methodology adaptable to languages like Urdu and Persian. By enhancing our understanding of linguistic adaptation, this work contributes to broader cross-linguistic studies of loanword integration and provides a foundation for further research into the historical and cultural interplay between languages. By meticulously analyzing 4,687 Arabic loanwords, we have identified 1,430 distinct Arabic roots linked to 3,855 Turkish words, demonstrating the deep historical and cultural interconnections between these two languages. Our research not only annotates the roots and morphological patterns of these Arabic words but also categorizes them semantically, revealing a complex landscape of linguistic integration.

By enhancing the Turkish Morpholex with accurate analyses of Arabic-origin words, we hope to facilitate a deeper understanding of the intricate dynamics of language contact and evolution. The implications of this research extend beyond Turkish, as it provides insights into the broader processes of language adaptation and the significance of historical interactions in shaping modern lexicons. Future studies could build upon these findings to enhance language models for the Turkish language, leveraging the enriched dataset for more accurate morphological and semantic analysis. Expanding the annotation process to other languages influenced by Arabic, such as Urdu and Persian, will validate the scalability of our methodology and contribute to comparative linguistic studies. Furthermore, integrating this dataset into universal morpholexical resources, such as multilingual WordNets, will broaden its applicability and utility for NLP tasks in multilingual and cross-linguistic contexts.

References

Shadiya Al-Hashmi. 2016. *The Phonetics and Phonology of Arabic Loanwords in Turkish: residual effects of gutturals*. Ph.D. thesis, University of York.

Wafi Alshammari and Ahmad Alshammari. 2020. Adaptation of turkish loanwords originating from arabic. *International Journal of English Linguistics*, 10:388.

Bilge Nas Arıcan, Aslı Kuzgun, Büsra Marsan, Deniz Baran Aslan, Ezgi Sanıyar, Neslihan Cesur, Neslihan Kara, Oguzhan Kuyrukçu, Merve Ozçelik, Arife Betül Yenice, et al. 2022. Morpholex turkish: A morphological lexicon for turkish. In *LREC 2022 Workshop Language Resources and Evaluation Conference 20-25 June 2022*, page 68.

Özge Bakay, Özlem Ergelen, Elif Sarmış, Selin Yıldırım, Bilge Nas Arıcan, Atilla Kocabalcıoğlu, Merve Özçelik, Ezgi Sanıyar, Oğuzhan Kuyrukçu, Begüm Avar, et al. 2021. Turkish wordnet kenet. In *Proceedings of the 11th global wordnet conference*, pages 166–174.

Federico Corriente. 2008. *Dictionary of Arabic and allied loanwords: Spanish, Portuguese, Catalan, Galician and kindred dialects*, volume 1. Brill.

Razieh Ehsani, Ercan Solak, and Olcay Taner Yildiz. 2018. Constructing a wordnet for turkish using manual and automatic annotation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(3):1–15.

Aida R Fattakhova and Nailya G Mingazova. 2015. Arabic loanwords in tatar and swahili: Morphological assimilation. *Journal of Sustainable Development*, 8(4):302.

ZG Human. 1949. Human behaviour and the principle of least effort.

Hugo Mailhot, Maximiliano Wilson, Joël Macoir, Hélène Deacon, and Claudia Sánchez Gutiérrez. 2019. Morpholex-fr: A derivational morphological database for 38,840 french words. *Behavior Research Methods*, 52.

John R Perry. 1984. -at and-a: Arabic loanwords with the feminine ending in turkish. *Turkish Studies Association Bulletin*, 8(2):16–25.

Stephan Procházka. 2009. Turkish loanwords. *Kees. Versteegh et al.(eds.). Encyclopedia of Arabic Language and Linguistics*, 4:489–594.

Lotfi Sayahi. 2005. Phonological adaptation of spanish loanwords in northern moroccan arabic.

Jacqueline Serigos. 2017. Using distributional semantics in loanword research: A concept-based approach to quantifying semantic specificity of anglicisms in spanish. *International Journal of Bilingualism*, 21(5):521–540.

Kamil Stachowskı. 2020. Phonetic renderings in turkish arabisms and farsisms. *Türkbilig*, 20(40):23–47.

Claudia Sánchez Gutiérrez, Hugo Mailhot, Hélène Deacon, and Maximiliano Wilson. 2017. Morpholex: A derivational morphological database for 70,000 english words. *Behavior Research Methods*, http://link.springer.com/article/10.3758/s13428-017-0981-8:1–13.