Mühendislik Bilimleri ve Tasarım Dergisi
14(1), 33-44, 2026
e-ISSN: 1308-6693

Journal of Engineering Sciences and Design
DOI: 10.21923/jesd.1746365

Araştırma Makalesi Research Article

OSMANLICA DERLEM VE KELİMELERİN UZUNLUK DAĞILIMININ ANALİZİ

İshak DÖLEK[1*], Atakan KURT[2]

[1] İstanbul Medeniyet Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği, İstanbul, Türkiye

[2] İstanbul Üniversitesi-Cerrahpaşa, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye

Anahtar Kelimeler

Osmanlıca Derlem, Kelime Sıklığı, Harf Dağılımı, Harf Grubu Dağılımı, İstatistiksel Dil Analizi

Öz

Bu çalışma, Osmanlıca metinlerin istatistiksel özelliklerini ve kelime dağılımlarını inceleyerek doğal dil işleme (DDİ) araştırmalarına katkı sağlamayı amaçlamaktadır. Osmanlıca, Arap alfabesine dayalı yapısı ve yoğun Arapça ile Farsça etkileri nedeniyle dil işleme süreçlerinde zorluklar sunmaktadır. Çalışmada, Osmanlıca metinlerden oluşturulan bir derlem kullanılarak kelime, harf ve harf grubu dağılımları detaylı olarak analiz edilmiştir. Derlemde, yaklaşık 2,34 milyon kelime, 5,7 milyon harf grubu ve 14 milyon harf içeren kapsamlı bir söz varlığı ortaya konmuştur. Derlem üzerinde yapılan istatistiksel analizler sonucunda, Osmanlıcada ortalama kelime uzunluğunun 3 ile 7 harf arasında yoğunlaştığı tespit edilmiştir. Çalışma, Osmanlıcanın nicel olarak belgelenmesine katkı sağlamanın yanı sıra, OCR, HTR ve alfabe çevirisi gibi uygulamalara yönelik dil modeli geliştirme süreçlerinde kullanılabilecek nitelikli bir kaynak sunmaktadır. Elde edilen bulgular, tarihi metinlerin dijitalleştirilmesi, otomatik işlenmesi ve kültürel mirasın korunması açısından da önemli bir temel teşkil etmektedir.

ANALYSIS OF OTTOMAN CORPUS AND LENGTH DISTRIBUTION OF WORDS

Keywords

Ottoman Corpus, Word Frequency, Letter Distribution, Ligature Distribution, Statistical Language Analysis.

Abstract

This study aims to contribute to natural language processing (NLP) research by analyzing the statistical properties and word distributions of Ottoman texts. Ottoman presents difficulties in language processing due to its structure based on the Arabic alphabet and its heavy Arabic and Persian influences. In this study, a corpus of Ottoman texts was used to analyses word, letter, and letter group (ligature) distributions in detail. The corpus has approximately 2.34 million words, 5.7 million letter groups, and 14 million letters. As a result of the statistical analysis of the Ottoman corpus, it was determined that the average word length in Ottoman is concentrated between 3 and 7 letters. In addition to contributing to the quantitative documentation of Ottoman, the study provides a qualified resource that can be used in language model development processes for applications such as OCR, HTR and Ottoman-Turkish transliteration. The findings also constitute an important basis for the digitization and automatic processing of historical texts and the preservation of cultural heritage.

Alıntı / Cite

Dölek, İ., Kurt, A., (2026). Osmanlıca Derlem ve Kelimelerin Uzunluk Dağılımının Analizi, Mühendislik Bilimleri ve Tasarım Dergisi, 14(1), 33-44.

Yazar Kimliği / Author ID (ORCID Number)
İ. Dölek, 0000-0002-5823-0103
A. Kurt, 0000-0002-9549-8475

Makale Süreci / Article Process	
Başvuru Tarihi / Submission Date	19.07.2025
Revizyon Tarihi / Revision Date	01.01.2026
Kabul Tarihi / Accepted Date	05.02.2026
Yayım Tarihi / Published Date	20.03.2026

[*] İlgili yazar / Corresponding author: ishak.dolek@medeniyet.edu.tr, +90- 534-483-4533

ANALYSIS OF OTTOMAN CORPUS AND LENGTH DISTRIBUTION OF WORDS

İshak DÖLEK[1†], Atakan KURT[2]
[1] İstanbul Medeniyet University, Faculty of Engineering and Natural Sciences, Department of Computer Engineering, İstanbul, Türkiye
[2] İstanbul University-Cerrahpaşa, Faculty of Engineering, Department of Computer Engineering, İstanbul, Türkiye

Highlights

- A large-scale Ottoman corpus containing approximately 2.34 million words, 5.7 million letter groups, and 14 million characters was constructed for the first time.
- Detailed statistical analyses were conducted on the corpus, including frequency distributions of letters, ligatures, and word lengths, with average word length found to be between 3 and 7 characters
- The corpus and its analyses offer a valuable linguistic resource for OCR, HTR, transliteration, and language model development aimed at the digital processing of historical Ottoman texts
- The study confirms that Ottoman word frequencies follow Zipf's law, indicating a natural language structure similar to modern languages

Purpose and Scope

The primary purpose of this paper is to support natural language processing (NLP) research by developing a comprehensive Ottoman corpus and analyzing its statistical and linguistic features. The study aims to fill a gap in the literature by providing the first large-scale, structured dataset for Ottoman, a language that presents unique challenges due to its Arabic-based script and heavy use of Arabic and Persian vocabulary. The research focuses on analyzing word, letter, and ligature distributions within the corpus and aims to provide foundational data that can be used in OCR, HTR, transliteration, and historical text digitization projects. Ultimately, the study contributes to the preservation and computational accessibility of Ottoman cultural and linguistic heritage.

Design/methodology/approach

To achieve the research objectives, a large-scale Ottoman corpus was constructed using digitized texts sourced from various periods and document types, including handwritten and printed materials. The study adopts a computational linguistics and corpus-based approach, focusing on statistical analysis of linguistic units such as letters, ligatures (connected letter groups), and words. Quantitative methods were applied to calculate frequency distributions, word length statistics, and character-level patterns. The research also employs rule-based morphological assumptions to infer the origin of words (Turkish, Arabic, Persian), contributing to a deeper linguistic analysis. This methodological framework supports applications in natural language processing tasks such as OCR, HTR, and transliteration, while also offering insights into the structural properties of Ottoman.

Findings

The study revealed that Ottoman exhibits a distinctive statistical structure, with the majority of words falling within the 3–7 character length range and an average word length of approximately 4.9 characters. Character frequency analysis showed that certain letters such as ا (elif) and ى (ye) are used with notably high frequency, reflecting both phonetic and morphological characteristics of the language. The distribution of words in the corpus closely follows Zipf's law, indicating that Ottoman, like modern languages, has a small number of highly frequent words and a long tail of infrequent ones.

Originality

This study presents the large-scale, structured Ottoman corpus of its kind, encompassing approximately 2.34 million words and 14 million characters, thereby addressing a significant data gap in historical natural language processing (NLP). By offering the first statistical analysis of Ottoman ligatures, letter frequencies, and word length distributions, this research provides a foundational linguistic resource that was previously unavailable. This study establishes a robust computational framework for future OCR/HTR, and machine translation technologies dedicated to Ottoman cultural heritage.

† Corresponding author: ishak.dolek@medeniyet.edu.tr, +90- 534-483-4533

1. Giriş (Introduction)

Osmanlıca, Osmanlı İmparatorluğu döneminde kullanılan Arap alfabesine dayalı bir yazı dilidir. Yoğun bir şekilde Arapça ve Farsça kelime ve gramer yapıları içermekle birlikte, bu kelimelerin yapılarını muhafaza etmiştir (Özkan, 2007). Osmanlı İmparatorluğu'nun resmi ve edebi dili olarak uzun bir dönem boyunca geniş bir coğrafyada kullanılmıştır. Bu zenginlik, Osmanlıca'nın kültürel, edebi, sosyolojik, tarihsel, askeri ve ekonomik açıdan önemli bir miras olarak kabul edilmesini sağlar. Osmanlıca metinler, dönemin dini, sanatsal ve gündelik yaşamına ışık tutan önemli kaynaklardır. Dini eserler, hutbeler ve fermanlar, dönemin inanç ve değer sistemlerini anlamak için kritik öneme sahiptir. Bu dil, farklı kültürler arasında köprü görevi görmüş ve Osmanlı medeniyetinin kendine özgü kimliğini oluşturmasında büyük rol oynamıştır. Ayrıca, Osmanlıca belgeler, imparatorluk ile diğer devletler arasındaki diplomatik ilişkilerin incelenmesinde de kritik bir rol oynar. Bugün Osmanlı arşivlerinde milyonlarca Osmanlıca belge, tarih araştırmacıları için önemli kaynaklardır.

Derlem, yazılı veya sözlü olarak birçok metnin bir araya getirilerek belli standartlara göre oluşturulan büyük metin kümeleridir (Agan ve Diri, 2016; Özkan, 2020). Derlemler yapay zeka, makine çevirisi ve derin öğrenme gibi araştırma alanlarında kullanılırken, hesaplamalı dilbilim ve doğal dil işleme çalışmalarında yürütülen araştırmalar da derlemlerden yararlanmaktadır (Çöltekin, 2020; Schweter, 2020; Uludoğan vd., 2024). Derlem özellikle doğal dil işlemede dil modellerinin eğitilmesi için gerekli olan büyük miktarda dil verisini sağlar. Derlemin büyüklüğü, yüksek kaliteli veri içermesi, dengeli ve çeşitli olması dil modelleri için önem arz etmektedir. Ayrıca bir dildeki kelime dağılımı, dilbilgisi kuralları, deyimler ve diğer dilsel özellikler, derlemler üzerinden analiz edilebilir. Derlem, kelimelerin kullanım sıklığı gibi istatistiksel bilgileri ortaya çıkararak dil modelinin tahmin yeteneğini artırırken, bir cümledeki kelimelerin sırasını, anlamını ve ilişkilerini öğrenmesinde önemli katkı sağlar.

Bu çalışmada, Osmanlıca için bir derlem hazırlanmış ve bu derlem üzerinde harf, harf grubu, kelime ve kelime uzunluk dağılımları hesaplanmış ve Osmanlıca özelinde analizler yapılmıştır. Osmanlıca metinlerde harf sıklıkları, harf grubu (ligature, connected component), kelime ve kelime grupları sıklıkları incelenmiştir. Osmanlıca kelime uzunlukları ele alınmış ve derlemde ortalama kelime uzunluğu tespit edilmiştir. Derlemler, doğal dil işlemede önemli bir yere sahip olmakla birlikte; makine çevirisi, alfabe çevirisi, yazım düzeltme, metin özetleme, metin üretme, optik karakter tanıma (OCR) ve el yazısı tanıma (HTR) gibi birçok doğal dil işleme uygulamasının temel yapı taşını oluşturmaktadır (Dölek ve Kurt, 2021; Dölek ve Kurt, 2022; Dölek ve Kurt, 2023).

Bu makale şu şekilde organize edilmiştir: Bölüm 2'de benzer çalışmalar ele alınmıştır. Bölüm 3'te derlem hakkında bilgi verilmiştir. Bölüm 4'te derlemdeki harf, harf grupları, kelime ve kelime grupları incelenmiştir. Bölüm 5'te derlem üzerinde istatiksel analizler yapılmıştır. Bölüm 6'da sonuçlar paylaşılmıştır.

2. Benzer Çalışmalar (Literature Review)

Derlem ve dil modelleri, DDİ alanında kritik bir öneme sahiptir. Bu alandaki araştırmalar, çeşitli yenilikçi çalışmaların temelini oluşturarak ilerlemelere öncülük etmektedir. Özellikle bilgi çıkarımı, OCR, HTR ve makine çevirisi gibi uygulamalarda, DDİ yöntemleri vazgeçilmez bir rol oynamaktadır. Osmanlıca alanında derlem ve dil modeli çalışmalarının, mevcut literatürde bilindiği kadarıyla henüz yapılmadığı görülmektedir. Osmanlıca, Arap harflerine dayalı bir alfabe kullandığından, Arapça, Farsça ve Urduca gibi benzer alfabe yapısına sahip diller üzerine yapılan çalışmalar bu alanda yol gösterici nitelikte olabilir. Bununla birlikte, Osmanlıca, Osmanlı İmparatorluğu döneminde kullanılan bir yazı dili olup Türkçenin bir formu olduğundan, Türkçe üzerine gerçekleştirilen çalışmalar da analiz edilmiştir.

Arapça dil kaynakları üzerine yapılan çalışmalar arasında (Zaghouani, 2017) ve (Saad and Ashour, 2010) öne çıkmaktadır. Zaghouani (2017) çalışmasında, Arapça için 66 farklı dil kaynağının sistematik biçimde derlendiği bir katalog sunulmaktadır. Saad ve Ashour (2010) ise, kaliteli ve yapısal açıdan güçlü bir açık kaynak Arapça derlemi olan Open Source Arabic Corpora (OSAC)'yi tanıtmaktadır. Hashemi (2010) çalışmasında İngilizce BBC News ve Farsça Hamshahri haberlerinden Farsça-İngilizce paralel bir derlem oluşturulmuştur. Becker ve Riaz (2002) çalışmasında, doğal dil işleme topluluğuna katkı sağlamak amacıyla, Unicode karakter kümesinde saklanan ve yerel Arap alfabesiyle yazılmış Urduca derlem geliştirmeyi hedeflemektedir.

Latin-tabanlı Türkçede derlem oluşturma çalışmaları 1990'lardan başlayıp günümüze kadar uzanmaktadır. ODTÜ Türkçe Derlemi (Say vd., 2002), ODTÜ-Sabancı Ağaç Yapılı Türkçe Derlemi, BOUN Corpus (Sak vd., 2011), Türkçe Ulusal Derlemi (Aksan vd., 2012) ve TS Corpus (Sezer ve Sezer, 2013) mevcut Türkçe derlemlerin en bilinenleridir. Türkçe derlemler üzerine yapılan çalışmalar günümüze kadar devam etmiştir (Dalkılıç ve Çebi, 2003; Demirhan ve Aksan, 2023; Sak vd., 2008). Dalkılıç ve Çebi (2003) çalışmasında Türkçe içeriğe sahip 10 ayrı web sitesinden yararlanılarak, 30 MB büyüklüğünde bir derlem oluşturulmuş ve bu derlemde kelime sayıları hesaplanmış, kelime uzunluk dağılımları incelenmiştir. Aksan vd. (2012), Sezer ve Sezer (2013) çalışmalarında TS Corpus ve Türkçe

Ulusal Derlemi (TNC) oluşturulurken karşılaşılan teorik ve pratik sorunları ele almaktadır. TNC, 50 milyon kelimelik büyük ölçekli, dengeli ve genel amaçlı bir derlem olarak tasarlanmıştır. Sak vd. (2008) çalışmasında, morfolojik ayrıştırma, belirsizlik giderme ve büyük ölçekli web derlemleri gibi dil işleme araçlarına yönelik kapsamlı kaynaklar sunulmuştur.

Arap-tabanlı Osmanlıca derlem çalışmaları, Osmanlıcanın daha iyi anlaşılması ve bu dil üzerinde daha kapsamlı çalışmalar yapılması için kritik öneme sahiptir. Derlenen veriler, Osmanlıca metinlerin dijital dünyaya taşınması ve bu zengin kültürel mirasın korunması için büyük bir potansiyel sunmaktadır. Bu çalışmada Osmanlıca derlem ve kelime sıklığı analizleri ele alınacaktır.

3. Osmanlıca Derlemin Hazırlanması (Preparation of Ottoman Corpus)

Günümüzde Arap-tabanlı Osmanlıca bir derlem hazırlamak oldukça meşakkatli ve zorlu bir görevdir. Osmanlıca dokümanlar kütüphanelerde, arşivlerde, müzelerde ve özel koleksiyonlarda tutulmaktadır. Milyonlarca belge, kitap, ferman, dergi gibi dokümanlara ulaşmak zordur ve bu bilgiler bilgisayar ortamında işlenebilir değildir. Bu sebeplerden dolayı büyük, geniş ve kapsamlı bir derlem oluşturmak ve Osmanlıca özelinde detaylı analiz yapmak ve başta doğal dil işleme olmak üzere teknolojik çözümler üretmek zor görünmektedir. Bu çalışmamızda Osmanlıca ile farkı dönemlerde yazılmış metinler bir araya getirilmiş ve bir derlem oluşturulmuştur. Bu derlem bu alanda ilk defa yapılması bağlamında değerli bir çalışma olacağı öngörülmektedir. Bu derlemde yapılan harf, harf grubu ve kelime analizleri Osmanlıca alanında yapılacak olan OCR, HTR ve dil çevirisi, alfabe çevirisi uygulamalarında önemli bir rol oynayacaktır (Dölek ve Kurt, 2025).

Bu çalışmada kullanılan derlem, Osmanlica.com çatısı altında geliştirilen ve farklı amaçlara yönelik olarak yapılandırılmış özgün veri kümelerinden türetilmiştir. Söz konusu veri kümeleri, öncelikli olarak Osmanlıca el yazması ve matbu belgelerin OCR ile HTR işlemlerine yönelik olarak oluşturulmuş; aynı zamanda Osmanlıca-Türkçe alfabe çevirisi süreçlerinde kullanılmak üzere kurgulanmıştır. Genel itibarıyla, toplanan dokümanların yaklaşık %30'u matbu, %70'i ise el yazısı içeren dokümanlardan oluşmaktadır. Veri setlerinin inşasında, farklı dönemlere, yazı stillerine (rik'a, nesih, talik vb.) ve içerik türlerine (resmi belgeler, edebi metinler, dini metinler, gazeteler vb.) ait çok çeşitli Osmanlıca dokümanlar taranmış ve bu dokümanlardan transkripsiyon ve eşleştirme verileri elde edilmiştir. Derlemin oluşturulmasında, hem tarihsel dil çeşitliliğini hem de yazı sistemlerindeki yapısal farklılıkları yansıtacak şekilde temsiliyetin sağlanmasına özen gösterilmiştir. Bu amaçla örnek olarak XVI. ve XVII. yüzyıllara tarihlenen erken dönem Osmanlı metinleri ile 1557–1911 yılları arasını kapsayan daha geç dönem belgeleri derleme dahil edilmiştir. Bu kapsamda, Osmanlı hukuk ve idari yapısını yansıtan kadı sicilleri önemli bir örnek teşkil etmektedir; mesela İstanbul (sur içi) ile bilâd-i selâse (Üsküdar, Galata, Eyüp) mahkemelerine ait kadı sicilleri, farklı dönemlere ait el yazması metinler sunarak derlemin tarihsel temsiliyetini güçlendirmektedir. Bu doğrultuda, farklı dönemleri temsil eden Osmanlıca metinlerden oluşan veri kümeleri hazırlanarak derlem oluşturulmuştur. . Tablo 1'de derlemde bulunan kelime, tekil kelime, harf grubu ve harf sayısı bilgileri sunulmuştur. Osmanlıca derlemde yaklaşık olarak 2,34 milyon kelime, 5,7 milyon harf grubu ve 14 milyon harf bulunmaktadır. Bu çalışma kapsamında oluşturulan Osmanlıca derlem, Türkçe için hazırlanan büyük ölçekli derlemlerle, örneğin Turkish National Corpus (TNC) (50 milyon sözcük; Aksan vd., 2013) ve TS Corpus (yaklaşık 1.8 milyar sözcük; Sezer ve Sezer, 2013) ile kıyaslandığında, sözcük sayısı bakımından daha sınırlıdır. Bununla birlikte, söz konusu Osmanlıca derlem; harf, harf grubu ve kelime yapıları hakkında ayrıntılı bilgiler sunmaktadır. Elde edilen bu veriler, metinlerin hem dilbilgisel hem de tipografik özelliklerinin anlaşılması açısından önem taşımakta olup, özellikle dijital metin işleme, dil modelleme ve tarihi metinlerin analizi gibi alanlarda değerli katkılar sağlamaktadır.

Tablo 1. Derlem özeti (Corpus summary)

Kelime	2.340.620
Tekil kelime	205.598
Harf grubu (ligature)	5.687.619
Harf	14.215.805
Harf kümesi	35
Rakam	10
Noktalama	11

4. Osmanlıca Derlemin İncelenmesi (Analysis of Ottoman Corpus)

Osmanlıca harf, harf grubu ve kelimelerin detaylı bir şekilde incelenmesi; OCR , HTR , alfabe çevirisi ve dil çevirisi gibi doğal dil işleme ve belge dijitalleştirme süreçleri için kritik bir öneme sahiptir. OCR ve HTR sistemleri, Osmanlıca gibi bağlamsal ve bağlayıcı harf yapısına sahip yazı sistemlerinde harflerin konuma göre farklı biçimlerini ve harf gruplarını doğru tanımlayabilmek için karakter düzeyinde hassas analizlere ihtiyaç duyar. OCR

ve HTR sistemlerinde başarı oranı büyük ölçüde karakter tanıma doğruluğuna bağlıdır. Daha açık bir ifadeyle OCR/HTR sistemlerinde temel alınan metriklerden birisi de karakter hata oranıdır (Character Error Rate - CER). Osmanlıca, Arap alfabesi temelli olup harfler bitişik yazıldığı için, aynı harfin farklı konumlarda (başta, ortada, sonda) farklı şekillerde yazılması tanımayı zorlaştırmaktadır. Derlem üzerinde yapılan harf sıklığı analizleri, modellerin en sık karşılaştığı karakterleri daha iyi tanımasına ve karakter temelli hataların azaltılmasına olanak sağlar. Osmanlıca metinlerde harfler sıklıkla birleşik (ligatürlü) yazılır. Bu birleşimlerin sıklığı ve biçimi, özellikle HTR sistemlerinde el yazısı varyasyonlarını tanımak için kritik öneme sahiptir. Harf grubu analizleri, OCR/HTR sistemlerinde segmentasyon hatalarını azaltmak ve bağlı karakterlerin doğru ayrıştırılmasını sağlamak amacıyla kullanılabilir. Alfabe çevirisi sürecinde ise, tek bir Osmanlıca harfin birden fazla Latin harfine karşılık gelebilmesi ve yazımda sesli harflerin çoğunlukla yer almaması, yalnızca harf değil, harf grubu ve kelime düzeyinde çözümleme yapılmasını zorunlu kılar. Örneğin "و" harfi bağlama göre "v", "ö", "o", "u" veya "ü" şeklinde farklı transliterasyonlara sahip olabilir. Ayrıca, Osmanlıca kelimelerdeki yoğun Arapça ve Farsça kökenli yapıların doğru çözümlenebilmesi için morfolojik analiz ve bileşik yapılarının ayrıştırılması da gerekmektedir. Dil çevirisi aşamasında ise bir kelimenin bağlama göre farklı anlamlara gelebilmesi, sadece yüzeysel transliterasyonla değil, bağlamdan anlam çıkarımıyla da desteklenen çok katmanlı bir çözümleme sürecini gerekli kılar. Dolayısıyla Osmanlıca metinlerin sağlıklı biçimde dijitalleştirilmesi, yorumlanması ve Türkçeye aktarılması için harf, harf grubu ve kelime düzeyinde dilsel çözümleme vazgeçilmezdir. Tablo 2, Tablo 3 ve Tablo 4'te derlemdeki karakter, rakam ve noktalama işaretlerinin dağılımı paylaşılmıştır. Bu veriler Osmanlıcanın istatistiksel özelliklerini analiz etmek ve karakter bazlı dil modelleri geliştirmek için temel oluşturabilir. Tablolar Osmanlıcada hangi harflerin daha sık kullanıldığını ve dilin yazı sistemi üzerindeki karakteristik dağılımını ortaya koymaktadır. Bu dağılım, hem harflerin fonetik özelliklerini hem de dilin morfolojik yapısını anlamak için önemli ipuçları sunmaktadır.

1. Elif (ا) ve ye (ى) harfleri, sırasıyla %11.9 ve %11.5 oranıyla metinlerde en yüksek sıklığa sahiptir. Bu durum, elif ve ye harflerin hem ünlü hem de ünsüz kelimelerde sıklıkla kullanılmasından kaynaklanmaktadır.
2. Ra (ر) ve vav (و) harfleri de sırasıyla %7.8 ve %7.4 sıklıklarıyla dikkat çekmektedir. Ra harfi, Türkçedeki birçok köklü kelimenin parçasıdır, vav ise hem ünlü hem de ünsüz olarak işlev görebilmektedir.
3. Lam (ل) ve he (ه) 6.59% ve %6.46 frekanslarıyla orta sıklık grubunda yer alır. Bu harflerin yaygınlığı, özellikle Arapça kökenli kelimelerden kaynaklanmaktadır.
4. Nun (ن) ve dal (د) harfleri de sıklıkla kullanılan harfler arasında olup sırasıyla %6.36 ve %5.60 oranında görülmüştür.
5. Osmanlıcada Arap alfabesinde bulunmayan ve Türkçeye özgü olan gaf (ڭ) harfi %0.98 oranında yer almaktadır. Benzer şekilde, ge (گ) harfi de %0.74 oranında kullanılmıştır. Bu oranlar, Osmanlıca metinlerinde Türkçe kökenli kelimelerin varlığını vurgulamaktadır.
6. Çe (چ) ve pe (پ) gibi Farsça kökenli harflerin de nispeten düşük oranlarda kullanıldığı (%0.58 ve %0.31) görülmektedir. Bu durum, bu harflerin daha çok Farsça kökenli kelimelerle sınırlı kaldığını göstermektedir.
7. Derlem nadir kullanılan harfleri de içermektedir. Özellikle ta-marbuta (ة) ve je (ژ) harfleri toplam karakterlerin yalnızca %0.01 ve %0.01'ini oluşturmaktadır. Bu harfler, çoğunlukla yabancı kökenli (Fransızca, İngilizce gibi) özel terimlerde karşımıza çıkarmaktadır.

Bu karakter sıklıkları, Osmanlıcada Arap alfabesinin yoğun kullanımını yansıtırken, dilin Türkçe, Arapça ve Farsça gibi çeşitli dillerden kelime almış ve zenginleşmiş bir yapı sunduğunu göstermektedir. Ayrıca derlemde bulunan rakam ve noktalama işaretlerin dağılımı Tablo 3 ve Tablo 4'te verilmiştir.

Osmanlıcada sıklıkla kullanılan kelimeler Tablo 5'te verilmiştir. Osmanlıca metinlerde و (ve), بر (bir), بو (bu), او (o) gibi kelime sıklıkları Osmanlıca da sık kullanılan yardımcı kelimeleri (stopwords) göstermektedir. Bu tür sık kullanılan kelimeler, metnin temel anlam yapısına katkı sağlasa da çoğu DDİ görevinde anlam çıkarımı için daha az önemli kabul edilir ve genellikle filtrelenir.

Tablo 6'da derlemde yer alan kelimelerin uzunluklarına göre sıklık dağılımları sunulmaktadır. Derlemde en yüksek sıklığın, 3–7 harf uzunluğundaki kelimelerde yoğunlaştığı ve bu aralıktaki kelimelerin toplam kelimelerin %71.24'ünü oluşturduğu görülmektedir. Bu bulgu, Osmanlıca metinlerde en sık kullanılan kelimelerin ortalama olarak 3–7 harf uzunluğuna sahip olduğunu göstermektedir. Buna karşılık, daha kısa (1–2 harfli) ve daha uzun (10 ve üzeri harfli) kelimelerin kullanım sıklığı, orta uzunluktaki kelimelere kıyasla oldukça düşüktür. Ayrıca Şekil 1'de gösterildiği gibi, derlemin %98.37'sinin 1–10 harf uzunluğundaki kelimelerden oluştuğu tespit edilmiştir.

Türkçede en yüksek sözcük sıklığı aralığı 2–10 harf iken (Dalkılıç ve Çebi, 2003), Osmanlıcada bu aralığın 3–7 harf olması, temelde Osmanlıcanın Arapça gibi sesli harfleri genellikle yazıda kullanmamasından kaynaklanmaktadır. Bunun pratik bir sonucu olarak, Osmanlıcada yazı yazmak için daha az harf kullanılması mümkün olmakta ve bu durum yazının daha hızlı yazılmasını sağlamaktadır. Bir diğer gözlem ise Osmanlıca kelimelerin, Türkçeye kıyasla

daha kısa olmasıdır. Bu durum, Osmanlıca metinlerin ve dolayısıyla kitapların daha az harf ve daha az sayfa kullanılarak yazılabilmesine olanak tanımaktadır.

Tablo 2. Karakter dağılımı (Character distribution)

Karakter	Sıklık	%	Karakter	Sıklık	%
ا	1380458	11.95	ز	151836	1.31
ى	1327379	11.49	ج	149287	1.29
ر	896699	7.76	ح	147402	1.28
و	848715	7.35	ف	125264	1.08
ل	761039	6.59	ك	113291	0.98
ه	745594	6.46	ص	90923	0.79
ن	734486	6.36	گ	85735	0.74
د	647126	5.60	ط	81177	0.70
م	614850	5.32	خ	76949	0.67
ت	442387	3.83	غ	68318	0.59
ب	421263	3.65	چ	66636	0.58
س	310352	2.69	ض	40578	0.35
ق	308761	2.67	پ	36047	0.31
ك	304807	2.64	ظ	27362	0.24
ع	181354	1.57	ث	24532	0.21
ش	175160	1.52	ذ	22752	0.20
ة	1221	0.01	ء	5579	0.05
ژ	1108	0.01			

Tablo 3. Noktalama işaretleri dağılımı (Distribution of punctuation marks)

Noktalama	Sıklık	%
-	36,902	0.32
،	16721	0.14
/	4856	0.04
"	2427	0.02
:	1876	0.02
)	1455	0.01
[	1407	0.01
]	1406	0.01
(	1381	0.01
؟	676	0.01
؛	638	0.01

Tablo 4. Rakamların dağılımı (Distribution of numbers)

Rakam	Sıklık	%
١	11994	0.10
٣	10138	0.09
٢	7744	0.07
٠	5244	0.04
٤	4520	0.04
٥	4333	0.04
٦	3332	0.03
٩	2971	0.03
٨	2939	0.03
٧	2747	0.02

3–7 harf aralığındaki kelimelerin sayısının yüksek olması, Osmanlıcanın kendine özgü bir karakterini yansıtmakta olup, bu tür dağılımların dile göre farklılık gösterebilen özellikler arasında olduğu düşünülmektedir. Buradan çıkarılabilecek bir diğer sonuç, Osmanlıcada 1 ve 2 harfli sözcüklerin sayısının oldukça sınırlı olduğudur. Bu durum, Osmanlıca metin analizi yapılırken 1 ve 2 harfli kelime listelerinin küçük bir diziye sığdırılarak hızlı ve etkin bir şekilde işlenmesini mümkün kılmaktadır. Ayrıca, karakter tabanlı ve altbirim (subword) tabanlı modeller açısından değerlendirildiğinde, bu sözcük uzunluğu aralığı Osmanlıcanın morfolojik yapısının görece dengeli bir sözcük uzunluğu ürettiğini göstermektedir. Özellikle tarihî metinlerde karşılaşılan yazım varyasyonları ve eklemeli yapı dikkate alındığında, orta uzunluktaki kelimelerin baskın olması, modellerin genelleme yeteneğini olumlu yönde etkilemektedir.

Bununla birlikte, OCR ve HTR sistemleri açısından 3–7 harf uzunluğundaki kelimelerin baskınlığı, kelime-seviyesinde hata düzeltme ve bağlamsal doğrulama aşamalarında avantaj sağlamaktadır. Çok kısa kelimeler bağlamsal belirsizlik yaratırken, çok uzun kelimeler tanıma hatalarına daha açıktır; orta uzunluktaki kelimelerin yüksek oranı, bu tür sistemlerde hata oranlarının düşürülmesine katkı sunmaktadır.

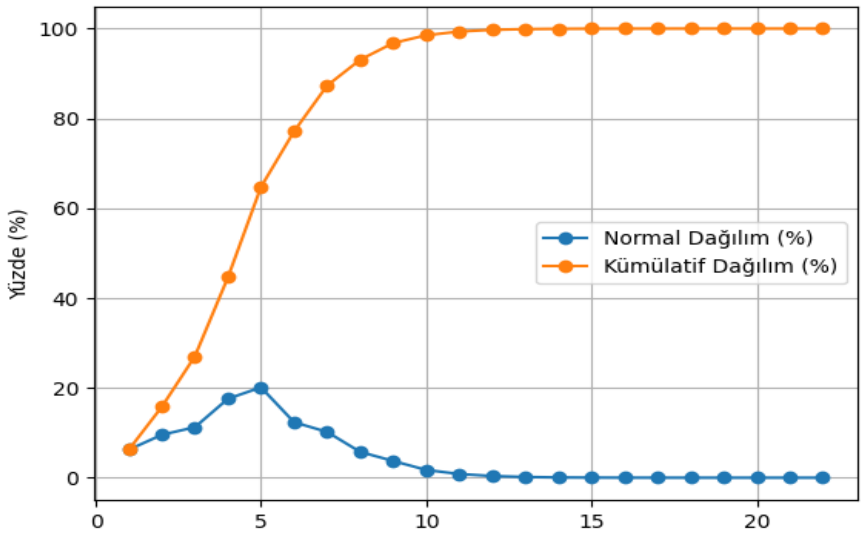
Son olarak, derlemin %98.37'sinin 1–10 harf aralığında yer alması, Osmanlıca için geliştirilecek DDİ sistemlerinde maksimum kelime uzunluğu, pencere boyutu ve girdi temsili gibi mimari parametrelerin daha bilinçli biçimde belirlenmesine olanak tanımaktadır. Bu yönüyle kelime uzunluk dağılımı, yalnızca betimleyici bir istatistik değil, aynı zamanda Osmanlıcaya özgü DDİ sistemlerinin tasarımı için yol gösterici bir bulgudur.

Tablo 5. Derlemdeki en çok geçen 40 kelime (The 40 most frequent words in the corpus)

Kelime	Sıklık	%	Kelime	Sıklık	%
(ve) و	131785	5.64	(olduğu) اولديغى	3797	0.16
(bir) بر	63795	2.73	(büyük) بيوك	3786	0.16
(bu) بو	23860	1.02	(yüz) يوز	3750	0.16
(o) او	18602	0.79	(Allah) الله	3471	0.15
(ile) ايله	16499	0.71	(fakat) فقط	3372	0.14
(ki) كه	14029	0.60	(öyle) اويله	3312	0.14
(de) ده	13229	0.57	(nur) نور	3302	0.14
(olan) اولان	12831	0.55	(sene) سنه	3297	0.14
(gibi) گبى	10098	0.43	(sonra) صوڭره	3292	0.14
(her) هر	10069	0.43	(ol) اول	3258	0.14
(için) ايچون	9793	0.42	(pek) پك	3207	0.14
(hem) هم	8533	0.37	(edip) ايدوب	3177	0.14
(kadar) قدر	6943	0.30	(üç) اوچ	3171	0.14
(bütün) بتون	6879	0.29	(ben) بن	3156	0.13
(eden) ايدن	6516	0.28	(karşı) قارشي	3135	0.13
(iki) ايكى	6127	0.26	(daha) داها	3049	0.13
(çok) چوق	5586	0.24	(var) وار	3049	0.13
(dahi) دخى	5526	0.24	(üzere) اوزره	3001	0.13
(ise) ايسه	5128	0.22	(içinde) ايچنده	2912	0.12
(en) اڭ	4965	0.21	(işte) ايشته	2873	0.12
(ne) نه	4725	0.20	(halde) حالده	2819	0.12
(risale) رساله	4718	0.20	(ehli) اهل	2816	0.12
(eder) ايدر	4657	0.20	(on) اون	3908	0.16
(şu) شو	4327	0.18	(etmek) ايتمك	3804	0.16
(hiç) هيچ	4005	0.17			

Tablo 6. Kelime uzunlukları dağılımı (Distribution of word lengths)

Uzunluk	Sıklık	%	Örnekler
1	140778	6.02	و
2	225130	9.62	هر, اڭ, بر, شو, سر, كه
3	262546	11.22	چوق, برر, ابن, متا, على, مهم
4	407579	17.41	قسمى, زمان, مغرب, عشاء, اوچر
5	467521	19.97	مولفى, اوتوز, برنجى, خصوصا, مذكور, كلمات
6	291745	12.46	, مكتوبڭ, نوريني, اتيلمش, قراڭلق
7	238217	10.18	فضيلتلى, بولونان, مناجاتى, السلامڭ, خلاصهسى, دغدغهلى
8	138386	5.91	مباركهنڭ, انوارندن, گوسترهجك, والسلامڭ, مناجاتدر, فورطنهلى
9	88084	3.76	اوقونماسى, مشهورهسنڭ, گچهبيلسين, الاسبابدن, گورديگندن, گوستردير
10	42619	1.82	اولامديغنى, بولونديردى, دگيشمهسيله, منظرهلرينى, مناجاتمزده, ايليشدكلرى
11	20543	0.88	ادارهسندهدر, كيفلنديرهرك, اهتزازاتندن, دوشونجهسيله, خستهلقلرينڭ
12	9755	0.42	اعتراضكارانه, ايشلتديرييور, بيلديريلمشدر, بولوندقلرينى
13	4337	0.19	استمدادكارانه, ايكيلشديرييور, ديگلهمدكلرندن
14	1806	0.08	ايشيقلانديرسين, اوزاقلاشديرهرق, قاتيلاشديرييور
15	923	0.04	يارادبلماسندهكى, زينتلنديرييورلر
16	325	0.02	قوزموغرافيالرينڭ
17	167	0.07	اوهاملانديرييورلر



Şekil 1. Kelime uzunluğu dağılımları (Word-length distributions)

Osmanlıca Arapça tabanlı bir alfabe olduğu için bitişik yazılan bir dildir. Bununla birlikte Osmanlıca kendisinden sonra birleşmeyen harfler vardır. Bu özellik doğal dil işlemede kelime bölütleme (word segmentation) problemini beraberinde getirmektedir. Daha açık bir ifadeyle kelime içi (intra-word) ile kelime arası (inter-word) boşluk net olmamakla birlikte kelimenin parçalanmasına veya birleşmesine sebeb olmaktadır. Tablo 7'de derlemde en sık geçen harf grupları verilmiştir.

Tablo 7. En sık harf grupları (Most frequently ligatures)

Harf grubu	Sıklık	%	Harf grubu	Sıklık	%
ا	719816	12.66	ت	42610	0.75
و	402220	7.08	یا	39544	0.70
ر	345713	6.08	ند	36651	0.64
د	234245	4.12	ما	35428	0.62
ه	160556	2.82	لر	32246	0.57
ن	156529	2.75	یله	29854	0.52
ی	100359	1.77	ل	29201	0.51
بر	89124	1.57	با	28982	0.51
ز	70847	1.25	ب	28393	0.50
ید	59904	1.05	قو	28218	0.50
بو	56293	0.99	قا	27927	0.49
یر	44040	0.77	یو	27447	0.48
لا	43985	0.77	ینه	16805	0.30
م	43198	0.76	قد	16785	0.29
ق	26134	0.46	مر	16561	0.29
ك	24927	0.44	هر	16470	0.29
لد	24479	0.43	نلر	16116	0.28
گو	23584	0.42	که	15868	0.28
لو	23100	0.41	عا	15853	0.28
ك	22172	0.39	حا	15802	0.28
نو	20257	0.36	صو	15777	0.28
نا	16899	0.29	کو	15447	0.27
سا	18813	0.33	تا	14177	0.25
مو	17999	0.32	له	13932	0.24

Tablo 8'de Osmanlıca metinlerinde kelime ve harf gruplarının kullanım sıklıklarını karşılaştırarak dilin yazım özelliklerini ve sözcük frekanslarını analiz edilmiştir. En sık kullanılan kelimeler arasında "و" (ve), "بر (bir)", ve "بو (bu)" gibi bağlaç ve zamirler öne çıkmakla birlikte aynı zamanda birer harf grubudurlar. Arap ve Fars kökenli kelimelerin yanı sıra, Türkçe'ye özgü sözcüklerin de yüksek frekansa sahip olduğu görülmektedir. Harf grubu kullanımı bazı durumlarda kelime sıklığından daha fazla olup yazıda harf birleşimlerinin yoğun kullanımını göstermektedir. Tabloda dikkat çeken bir diğer önemli husus, Osmanlıca metinlerde kelime aralarındaki

boşlukların belirsiz olmasıdır. Bu durum, kelimelerin yanlış bölünmesine ve başka kelimelere dönüşmesine neden olabilmektedir. Bu tür yapısal belirsizlikler, Osmanlıca kelime bölütleme problemini ortaya çıkarmakta ve otomatik işleme süreçlerini daha da karmaşık hale getirmektedir.

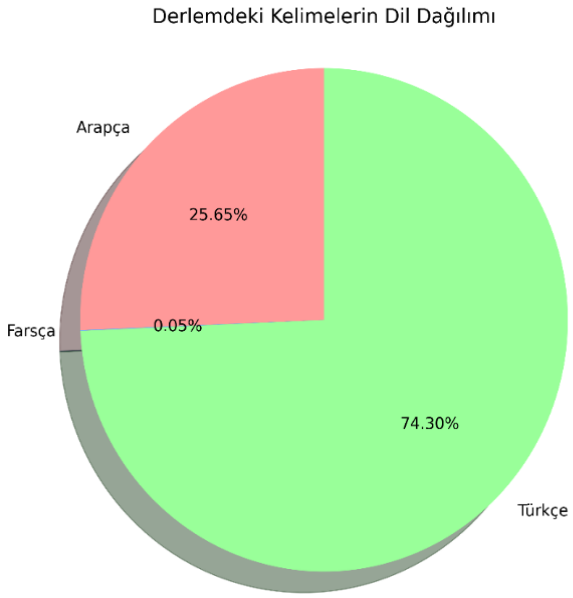
Tablo 8. Kelime olan harf grupların sıklıkları (Frequencies of 'words' ligature)

Kelime	Türkçe	Kelime Sıklık	Harf Grubu Sıklığı	Kelime	Türkçe	Kelime Sıklık	Harf Grubu Sıklığı
و	ve	131787	402231	حقيقت	hakikat	1646	1647
بر	bir	63796	89125	عين	aynı	1591	1601
بو	bu	23860	56293	كبي	gibi	1534	1547
كه	ki	14029	15868	شی	şey	1451	6071
گبی	gibi	10099	10099	فی	fi	1449	3251
هر	her	10069	16471	گلن	gelen	1425	1426
هم	hem	8534	9431	بز	biz	1371	5058
نه	ne	4725	13866	سنك	senin	1366	1841
شو	şu	4328	9563	يا	ya	1350	39544
هيچ	hiç	4005	4009	حسن	hasan	1350	1401
فقط	fakat	3372	3373	خبر	haber	1304	1858
سنه	sene	3301	5168	مى	mi	1287	8442
پك	pek	3207	3233	قطعى	kati	1211	1211
بن	ben	3156	3830	ينه	yine	1202	16805
بلكه	belki	2734	2734	تك	tek	1176	2248
عليه	aleyh	2395	2657	علم	ilim	1172	1456
بش	beş	2216	2218	سن	sen	1134	1270
بيك	bin	1993	1996	حقيقى	hakiki	1106	1107
على	ali	1963	2053	بيگلر	binler	1082	1288
بنم	benim	1946	1991	قسم	kısım	1082	1083
حق	hak	1926	1968	تا	ta	1078	14178
يعنى	yani	1795	1837	مثلا	mesela	1076	1147
بعض	bazı	1788	1795	مهم	mühim	1058	1079
بكا	bana	1746	1752	حتى	hatta	1643	1712

5. Derlem Üzerine Sayısal Analizler (Statistical Analysis on the Corpus)

Derlem üzerinde gerçekleştirilen nicel dilsel analizler, Osmanlıcanın söz varlığı, yapısal özellikleri ve istatistiksel dağılımları hakkında önemli bilgiler sunmaktadır. Bu çalışmada kullanılan Osmanlıca metinlerin dilsel analizinde, kelimelerin köken dağılımları belirli harf kümeleri üzerinden sınıflandırılarak değerlendirilmiştir. Şekil 1'de sunulan verilere göre, incelenen kelimelerin yaklaşık %25.65'i Arapça, %0.05'i Farsça ve %74.30'ü Türkçe kökenlidir. Köken tespiti sürecinde, Osmanlıcada sıkça kullanılan Arapça ve Farsça kökenli sözcükleri ayırt edebilmek amacıyla karakteristik harf yapıları dikkate alınmıştır. Arapça kökenli kelimeler, Arap alfabesine özgü ve genellikle yalnızca Arapça kelimelerde rastlanan "ء, ث, ص, ض, ط, ظ, ع, خ, ذ, ح" harflerinden en az birini içermeleri durumunda bu kategoriye dahil edilmiştir (Tulum, 2014). Farsça kökenli kelimelerin tespitinde ise, yalnızca Farsça yazımda yer alan "ژ" harfi ayırt edici unsur olarak kullanılmıştır (Tulum, 2014). Geriye kalan kelimeler, bu özel karakterleri içermemeleri durumunda Türkçe kökenli olarak sınıflandırılmıştır. Bu yöntem, kelime kökenlerinin otomatik olarak belirlenmesine olanak tanıyan kural tabanlı bir yaklaşım sunmakta ve Osmanlıca metinlerinin dil içi çeşitliliğinin nicel olarak analiz edilmesine katkı sağlamaktadır. Ancak yöntemin sınırlılıkları arasında, bazı alıntı kelimelerin bu özel karakterleri içermemesi nedeniyle yanlış sınıflandırılabilme ihtimali yer almaktadır. Örneğin, şükür (شكر), fikir (فكر) ve devlet (دولت) gibi Arapça kökenli kelimeler, ilgili harfleri içermedikleri için harf tabanlı sınıflama yönteminde yanlışlıkla Türkçe kökenli olarak değerlendirilmiştir. Bu durum, gelecekte morfolojik analiz ve etimolojik sözlük tabanlı hibrit yöntemlerin geliştirilmesini gerekli kılmaktadır.

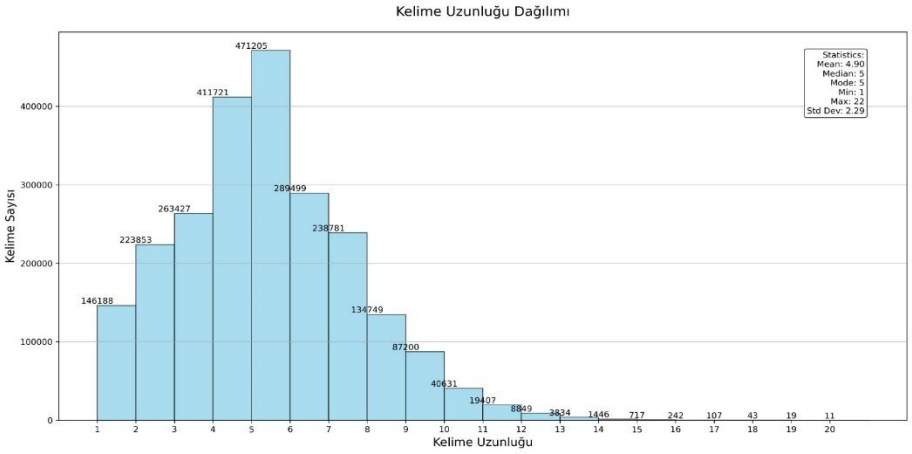
Kelime uzunluğu dağılımına ilişkin yapılan istatistiksel analiz sonuçları Tablo 9'da ve histogramı Şekil 2'de sunulmuştur. Derlemdeki kelimelerin ortalama uzunluğu 4.90 harf olup, hem medyan (5) hem de mod (5) değerleriyle büyük ölçüde örtüşmektedir. Bu durum, kelime uzunluklarının dengeli ve simetrik bir dağılım sergilediğini göstermektedir. En kısa kelimenin yalnızca 1 (و - ve) harf, en uzun kelimenin ise 22 (قوشقولاندیرمایاجقلرینی - kuşkulandırmayacaklarını) harf olması, Osmanlıcadaki biçimbilimsel zenginliği ve yapısal çeşitliliği yansıtmaktadır. Özellikle birleşik fiiller, ekli yapılar ve ad tamlamaları gibi uzun kelimelerin varlığı, doğal dil işleme sistemlerinde özel ele alınması gereken istisnai yapılardır. Standart sapmanın 2.29 olarak hesaplanması, kelime uzunluklarının genellikle 3 ila 7 harf arasında yoğunlaştığını ve aşırı uç değerlerin sınırlı olduğunu göstermektedir.



Şekil 2. Kökenlerine göre kelime dağılımı (Word origin Distribution)

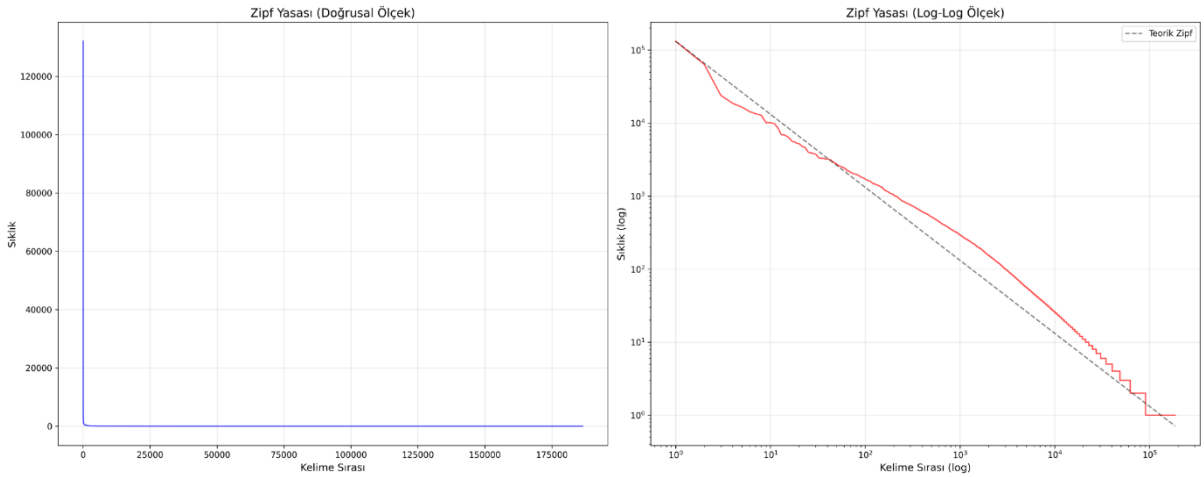
Tablo 9. Kelime Uzunluklarının İstatistiksel Özeti (Statistical Summary of Word Lengths)

Ölçüt	Değer
Ortalama (Mean)	4.90
Medyan (Median)	5
Mod (Mode)	5
Minimum	1
Maksimum	22
Standart sapma	2.29



Şekil 3. Kelime uzunluklarının histogramı (Histogram of word lengths)

Zipf Yasası, doğal dillerdeki kelime kullanım sıklıklarının belirli bir matematiksel düzene göre dağıldığını ifade eden istatistiksel bir ilkedir (Dalkılıç ve Çebi, 2004; Ilgen ve Karaoglan, 2007). Bu yasa, bir dildeki kelimelerin frekanslarının, kelimenin sıralamasıyla ters orantılı olduğunu öne sürer; yani en sık kullanılan kelime, ikinci en sık kullanılan kelimenin yaklaşık iki katı, üçüncü en sık kullanılan kelimenin ise üç katı kadar kullanılır. Osmanlıca derlemdeki kelime sıklıkları üzerinde gerçekleştirilen analizler, dilin doğal yapısının Zipf yasasına büyük ölçüde uyduğunu ortaya koymuştur. Şekil 3'te sunulan grafiklerde, kelimelerin frekanslarına göre sıralandığı doğrusal (sol) ve log-log (sağ) ölçekli dağılımlar yer almaktadır. Doğrusal ölçekte elde edilen eğri, az sayıda kelimenin (örneğin bağlaçlar, zamirler ve fiiller) oldukça yüksek sıklıkta kullanıldığını, büyük çoğunluğun ise nadiren geçtiğini göstermektedir. Log-log ölçekteki grafikte, kelime frekanslarının sıralarıyla logaritmik ilişkisi incelenmiş ve bu ilişkinin teorik Zipf eğrisi ile büyük ölçüde örtüştüğü gözlemlenmiştir. Bu durum, Osmanlıcanın da modern dillerde olduğu gibi az sayıda yüksek frekanslı kelime ve uzun kuyruklu düşük frekanslı kelime dağılımı karakteristiği taşıdığını göstermektedir. Zipf analizinden elde edilen bu bulgular, dil modeli eğitimi, otomatik sözcük tahmini ve nadir kelime işleme gibi doğal dil işleme uygulamaları için önemli bir temel sunmaktadır.



Şekil 3. Kelimelerin Zipf yasasına göre grafiği (Graph of words according to Zipf's law)

6. Sonuçlar ve Tartışma (Results and Discussion)

Bu çalışmada, Osmanlıca metinlerden oluşan özgün bir derlem oluşturulmuş ve bu derlem üzerinde harf, harf grubu (ligatür) ve kelime düzeyinde istatistiksel analizler gerçekleştirilmiştir. Derlemde yaklaşık 2,34 milyon kelime, 5,7 milyon harf grubu ve 14 milyon harf yer almakta olup, bu yönüyle Osmanlıca üzerine yürütülen doğal dil işleme (DDİ) çalışmalarına nicel bir kaynak sunma amacı taşımaktadır. Analiz sonuçları, Osmanlıca kelimelerin büyük çoğunluğunun 3–7 harf uzunluğunda olduğunu ve elif (ا) ile ye (ى) harflerinin metinlerde en sık karşılaşılan karakterler olduğunu ortaya koymuştur. Ayrıca kelime sıklıklarının Zipf yasası ile büyük oranda uyum göstermesi, Osmanlıcanın doğal dil yapısının istatistiksel açıdan diğer dillerle benzerlik taşıdığını göstermektedir. Harf grubu analizleri, özellikle OCR ve HTR gibi karakter tanıma sistemlerinde karşılaşılan segmentasyon ve tanıma problemlerine çözüm geliştirmek açısından önemli veriler sunmaktadır. Elde edilen bulgular, Osmanlıca metinlerin otomatik işlenmesi, dijitalleştirilmesi ve modern Türkçeye aktarılması gibi uygulamalarda kullanılabilecek dil modeli geliştirme süreçlerine katkı sağlamaktadır. Ayrıca, bu çalışma, Osmanlıcanın tarihsel ve kültürel bağlamda daha geniş kitlelerce erişilebilir hâle getirilmesi açısından da önemli bir adım niteliğindedir. Gelecek çalışmalarda, derlemin farklı yazı türleri ve dönemleri kapsayacak şekilde genişletilmesi; morfolojik çözümleme, anlamsal etiketleme ve makine çevirisi gibi ileri düzey DDİ uygulamalarına yönelik modellerin geliştirilmesi hedeflenmektedir.

Teşekkür (Acknowledgement)

Bu çalışma kapsamında kullanılan Osmanlıca metinlerin temininde destek sağlayan Osmanlica.com ekibine teşekkür ederiz. The authors would like to thank the Osmanlica.com team for their support in providing access to Ottoman texts.

Çıkar Çatışması (Conflict of Interest)

Yazarlar tarafından herhangi bir çıkar çatışması beyan edilmemiştir. No conflict of interest was declared by the authors.

Kaynaklar (References)

Agan, C., & Diri, B., Türkçe Derlemler İçin Söz Dizimsel Görselleştirme ve Sorgulama Aracı, Türkiye Bilişim Vakfı Bilgisayar Bilimleri Ve Mühendisliği Dergisi, cilt 9, no. 1, pp. 1-10, 2016.

Aksan, Y., et al., Construction of the Turkish National Corpus (TNC), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), İstanbul, 2012.

Becker, D. & Riaz, K., A Study in Urdu Corpus Construction, COLING-02: The 3rd Workshop on Asian Language Resources and International Standardization, https://aclanthology.org/W02-1201.pdf, 2002.

Çöltekin, Ç., A Corpus of Turkish Offensive Language on Social Media, Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, 2020.

Dalkılıç, G. & Çebi, Y., Türkçe Külliyat Oluşturulması Ve Türkçe Metinlerde Kullanılan Kelimelerin Uzunluk Dağılımlarının Belirlenmesi, Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, cilt 5, no. 1, pp. 1-7, 2003.

Dalkılıç, G. & Çebi, Y., Zipf's Law and Mandelbrot's Constants for Turkish Language Using Turkish Corpus (TurCo), Advances in Information Systems. ADVIS, Berlin, 2004.

Demirhan, U. U. & Aksan, M., Türkçede Eylemlerin Derlem-Temelli Değerlik Sözlüğünün Oluşturulması: Yöntembilimsel Bir Çalışma, Dil Ve Edebiyat Dergisi, cilt 19, no. 1, pp. 61-88, 2023.

Dölek İ. & Kurt A. A deep learning model for Ottoman OCR. Concurrency Computat Pract Exper. 2022; 34(20):e6937. doi:10.1002/cpe.6937

Dölek, İ. & Kurt A., "Ottoman OCR: Printed Naskh Font," 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Kocaeli, Turkey, 2021, pp. 1-5, doi: 10.1109/INISTA52262.2021.9548616.

Dölek, İ., & Kurt A. "Derin sinir ağlarıyla Osmanlıca optik karakter tanıma". Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi 38, sy. 4 (Nisan 2023): 2579-94. https://doi.org/10.17341/gazimmfd.1062596.

Dölek, İ., &Kurt A. "Osmanlıcadan Türkçeye Uçtan Uca Aktarım". Journal of Smart Systems Research 3, no. 1 (June 2022): 1-10.

Dölek İ, & Kurt A., "Ottoman HTR: Recognition of the Ottoman Riqa Font Using Deep Learning Models," 2025 Innovations in Intelligent Systems and Applications Conference (ASYU), Bursa, Turkiye, 2025, pp. 1-6, doi: 10.1109/ASYU67174.2025.11208325.

Hashemi, H. B., Shakery A. & Faili, H., Creating a Persian-English Comparable Corpus, Multilingual and Multimodal Information Access Evaluation. CLEF , https://doi.org/10.1007/978-3-642-15998-5_5, 2010.

Ilgen, B. & Karaoglan, B., Investigation of Zipf's 'law-of-meaning' on Turkish corpora, 22nd international symposium on computer and information sciences, Ankara, 2007.

Özkan, B., Terim Bilim Çalışmalarında Derlem Kullanımı Ve "Kendi Kendine Derlem Oluşturma Platformu" Terim Bilim Modülü, Avrasya Terim Dergisi, cilt 8, no. 1, pp. 8-17, 2020.

Özkan, M., Osmanlı Türkçesi, İstanbul: Türkiye Diyanet Vakfı İslam Ansiklopedisi, 2007.

Saad, M. K. & Ashour, W., OSAC: Open Source Arabic Corpora, 6th International Conference on Electrical and Computer Systems (EECS'10), Lefke, 2010.

Sak, H., Güngör, T. & Saraçlar, M., Resources for Turkish morphological processing, Lang Resources & Evaluation, Cilt 1 /2, https://doi.org/10.1007/s10579-010-9128-6, p. 249–261, 2011.

Sak, H., Güngör, T. & Saraçlar, M., Turkish language resources: Morphological parser, morphological disambiguator and web corpus, Advances in Natural Language Processing: 6th International Conference, GoTAL 2008, Gothenburg, 2008.

Say, B., Zeyrek, D., Oflazer, K. & Özge, U., Development of a corpus and a treebank for present-day written Turkish, Proceedings of The Eleventh International Conference of Turkish Linguistics, 2002.

Schweter, S., BERTurk - BERT models for Turkish, Zenodo, https://doi.org/10.5281/zenodo.3770924, 2020.

Sezer, T. & Sever Sezer, B., TS Corpus: Herkes için Türkçe derlem, The 27th Turkish National Linguistics Conference, 2013.

Tulum, M., Osmanlı Türkçesine Giriş-I, Eskişehir: Anadolu Üniversitesi, 2014.

Uludoğan, G., Yirmibeşoğlu, Z., Akkurt, F., Türker, M., Güngör, O., & Üsküdarlı, S., TURNA: A Turkish encoder-decoder language model for enhanced understanding and generation. arXiv. https://arxiv.org/abs/2401.14373, 2024

Zaghouani, W., Critical Survey of the Freely Available Arabic Corpora, https://arxiv.org/abs/1702.07835, 2017.