T.C.
İSTANBUL ÜNİVERSİTESİ-CERRAHPAŞA
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

DOKTORA TEZİ

OSMANLICADAN MODERN TÜRKÇEYE UÇTAN UCA AKTARIM SİSTEMİ

İshak DÖLEK

DANIŞMAN
Doç. Dr. Atakan KURT

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

İSTANBUL-2022

Bu çalışma, 27.01.2022 tarihinde aşağıdaki jüri tarafından Bilgisayar Mühendisliği Anabilim Dalı, Bilgisayar Mühendisliği Programında Doktora tezi olarak kabul edilmiştir.

Tez Jürisi

Doç. Dr. Atakan KURT (Danışman)
İstanbul Üniversitesi-Cerrahpaşa
Mühendislik Fakültesi

Prof. Dr. Rüya ŞAMLI
İstanbul Üniversitesi-Cerrahpaşa
Mühendislik Fakültesi

Doç. Dr. Fatih KOÇAN
Atlas Üniversitesi
Mühendislik ve Doğa Bilimleri Fakültesi

Prof. Dr. Ahmet SERTBAŞ
İstanbul Üniversitesi-Cerrahpaşa
Mühendislik Fakültesi

Doç. Dr. Taner ÇEVİK
Nişantaşı Üniversitesi
Mühendislik Mimarlık Fakültesi

20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, İstanbul Üniversitesi-Cerrahpaşa'nın abonesi olduğu intihal yazılım programı kullanılarak Lisansüstü Eğitim Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

Bu tez, 2190252 proje numaralı TUBİTAK 1512 programı kapsamında desteklenmiştir.

ÖNSÖZ

Doktora tezimde destek veren danışman hocam Doç. Dr. Atakan KURT'a, tez izleme süreçlerinden vermiş oldukları desteklerinden ötürü çok değerli tez izleme komitesi üye hocalarım Prof. Dr. Rüya Şamlı ve Doç. Dr. Fatih Koçan hocalarıma teşekkür ederim.

Hayatımın her evresinde yanımda olan başta eşim, annem ve babam olmak üzere tüm kardeşlerime teşekkür ederim.

Gerek doktora sürecinde gerek sosyal hayatımda yanımda olan ve adını burada yazamadığım tüm arkadaşlarıma teşekkür ederim.

Ocak 2022 İshak DÖLEK

İÇİNDEKİLER

Sayfa No

ŞEKİL LİSTESİ

Sayfa No

TABLO LİSTESİ

Sayfa No

SİMGE VE KISALTMA LİSTESİ

Kısaltmalar **Açıklama**

OCR : Optical Characters Recognition (Optik Karakter Tanıma)

DL : Deep Learning (Derin Öğrenme)

CNN : Convolutional Neural Network (Evrişimsel Sinir Ağları)

RNN : Recurrent Neural Network (Yinelemeli Sinir Ağları)

LSTM : Long-Short Term Memory (Uzun Kısa Süreli Bellek)

CTC : Connectionist Temporal Classification (Bağlantıcı Zamansal Sınıflandırma)

DDİ : Doğal Dil İşleme

YSA : Yapay Sinir Ağları

POS : Part of Speech (Sözcük Türü)

GT : Ground Truth (Referans Metin)

NLTK : Natural Language ToolKit (Doğal Dil Araç Kiti)

RTL : Rigth to Left (Sağdan Sola)

LTR : Left to Rigth (Soldan Sağa)

TR : Türkçe

OS : Osmanlıca

ÖZET

OSMANLICADAN MODERN TÜRKÇEYE UÇTAN UCA AKTARIM SİSTEMİ

DOKTORA TEZİ

İshak DÖLEK

İstanbul Üniversitesi-Cerrahpaşa

Lisansüstü Eğitim Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Danışman : Doç. Dr. Atakan KURT

Osmanlıca kaynakların büyük bir kısmı yurt içi ve yurt dışındaki devlet arşivlerinde, kütüphanelerde ve özel koleksiyonlarda muhafaza edilmektedir. Bu kaynakların bilimsel, tarihsel ve kültürel olarak ne kadar değerli olduğu herkesin malumudur. Bu kaynaklardaki milyonlarca belge, yüzbinlerce basılı kitap, dergi, vb. dokümanların manuel olarak yani elle teker teker okunup Günümüz Türkçe'sine aktarılması pratikte mümkün olmayan bir iştir. Bu çalışmada Osmanlıca matbu (basılı) dokümanları Günümüz Türkçe'sine Derin Öğrenme modelleriyle aktaran bir sistem geliştirilmiştir. Bu sistemde Osmanlıca dokümanların Günümüz Türkçe'sine Uçtan Uca Aktarımı üç aşamada gerçekleştirilir: (i) Osmanlıca OCR- optik karakter tanıma ya da görüntü-metin dönüşümü, (ii) Osmanlıca-Türkçe alfabe çevirisi ve (iii) Osmanlıca-Türkçe dil çevirisi. Osmanlıca OCR aşamasında Osmanlıca bir resim ya da

dokümandan Osmanlıca düzenlenebilir metin elde edilmiştir. Osmanlıca OCR aşamasında, Osmanlica.com diye isimlendirdiğimiz derin öğrenme modelinin Google Docs, Fine Reader, Tesseract Arapça, Tesseract Farsça ve Miletos araçları ile deneysel olarak karşılaştırılması 21 sayfalık Osmanlıca test dokümanları kullanılarak gerçekleştirilmiştir. Osmanlica.com modeli diğerlerinden daha iyi performans göstererek %96 karakter tanıma doğruluk oranı elde etmiştir. Bu aşamada ilk defa Osmanlıca Sıklık Analizi yapılmıştır. Bu bağlamda Osmanlıca'nın karakter, katar ve kelime sıklıklarını bulmak için yeni bir çalışma yapılmış, sıklıklar elde edilmiş ve sonuçları paylaşılmıştır. Osmanlıca harfler ayırt edici özelliklerine göre gruplanmış ve literatürde ilk defa bu harf gruplarının sıklık dağılımları verilmiştir. İlk defa Osmanlıca OCR modellerin karakter tanıma ve kelime tanıma doğruluk oranlarına ek olarak bağlı karakter katarı tanıma doğruluk oranları da hesaplanmış ve OCR modellerinin performanslarını karşılaştırmada kullanılmıştır. İkinci aşama olan alfabe çevirisi aşamasında, Osmanlı metinlerin Arapça tabanlı Osmanlı alfabesinden Latin tabanlı Modern Türk alfabesine dönüşümü yapılır. Osmanlıca fonetik bir dil olmadığı için bu dönüşüm güç ve karmaşık bir süreçtir. Osmanlı alfabesi ile Türk alfabesindeki harfler arasında bire bir eşleşme yoktur. Alfabe çevirisi aşaması çok adımlı ve karmaşık olması, bu adımların sözlük, imla kılavuzu, derlem gibi dil kaynakları gerektirmesi, süreçte kısmen de olsa bazı fonetik, morfolojik, gramatik ve semantik problemlerin çözülmesi gerektiğinden dolayı zor bir problem olarak karşımıza çıkmaktadır. Alfabe çevirisi (i) Ortografik Alfabe çevirisi, (ii) kelime bölütleme (iii) yazım düzeltme (iv) seslendirme (v) kelime tahmini (n-grams) (vi) tamlama ve birleşik isimler olmak üzere her biri başlı başına büyük bir problem olan alt aşamalardan meydana gelir. Bu alt adımlar için sözlükler oluşturulmuş ve düzenlemiş, derlem için eserler toplanmış ve algoritmalar geliştirilmiştir. Bu alt adımları kapsayan bir alfabe çevirisi aracı geliştirilmiştir. Alfabe çevirisi aracı - kendi türünde tek uygulama- 7500 kelimelik bir test veri setinde %98 kelime doğruluk oranı elde etmiştir. Üçüncü aşaması olan dil çevirisi ise, Türk alfabesindeki Osmanlıca metnin bilgisayarlı çeviriyle Günümüz Türkçe'sine çevirisidir. Sıradan insanlar için alfabe çevirisi aşamasında üretilen Osmanlıca metin okunabilir olmasına rağmen çoğu zaman anlaşılabilir değildir. Çünkü Osmanlıca'da çok fazla Arapça ve Farsça kelime ve kelime öbeği bulunmaktadır. Günümüz insanı bu öğelerin çoğunun hem anlamına hem de yapılarına yabancıdır. Bu aşamada Osmanlıca'dan Türkçe'ye kural tabanlı bire bir doğrudan dil çeviri aracı geliştirilmiştir. Dil çevirisinin başarı oranını test etmek için derlem ve sözlükler gibi dil kaynaklarının hazırlanması gerekmektedir. Bu üç aşama ile Osmanlıca'dan Türkçe'ye Uçtan Uca aktarımı bütüncül bir sistem olarak ele alınmıştır.

Ocak 2022, 111 sayfa.

Anahtar kelimeler: Osmanlıca, Optik karakter tanıma, Matbu nesih hattı, OCR, Derin öğrenme, Makine Çevirisi, Alfabe Çevirisi, Dil Çevirisi

SUMMARY

END-TO-END CONVERSION SYSTEM FROM OTTOMAN TO MODERN TURKİSH

Ph.D. THESIS

Ishak DÖLEK

Istanbul University-Cerrahpasa

Institute of Graduate Studies

Department of Computer Engineering

Supervisor : Assoc. Prof. Dr. Atakan KURT

Most of the Ottoman sources are kept in state archives, libraries, and private collections at domestic and abroad. Everyone knows how valuable these resources are scientifically, historically, and culturally. There are millions of documents, hundreds of thousands of printed books, magazines, etc. in these sources. It is practically impossible to read these documents manually, one by one, and convert them to the contemporary alphabet. In this study, a system has been developed that transfers Ottoman printed texts to Contemporary Turkish with Deep learning models. In the process of End-to-End Conversion from Ottoman to Contemporary Turkish, it carries out the transfer of Ottoman documents to Contemporary Turkish in three

steps: (i) Ottoman OCR- optical character recognition or image-text conversion, (ii) Ottoman-Turkish alphabet translation (transliteration), and (iii) Ottoman-Turkish language translation. During the Ottoman OCR step, an Ottoman editable text has obtained from an Ottoman image or document. In the Ottoman OCR step, the experimental comparison of the deep learning model, which we call Osmanlica.com, with Google Docs, Fine Reader, Tesseract Arabic, Tesseract Persian, and Miletos tools or models has carried out using a test dataset of 21 pages of original documents. The Osmanlica.com model outperformed the others, achieving 96% character recognition accuracy rate. At Ottoman OCR step, Frequency Analysis in Ottoman has been made for the first time. In this context, a new study has been carried out to find the character, ligature (connected component) and word frequencies of Ottoman, the frequencies have been obtained and the results have been shared. Ottoman letters are grouped according to their distinctive features and the frequency distributions of these letter groups are given for the first time in the literature. For the first time, in addition to the character recognition and word recognition accuracy rates of the Ottoman OCR models, the ligature recognition accuracy rates have also been calculated and used to compare the performances of the OCR models. In the second step, the alphabet translation (transliteration) step, transliteration is the conversion of the Ottoman text from the Arabic-based Ottoman alphabet to the Latin-based Modern Turkish alphabet. Since Ottoman is not a phonetic language, this conversion is a difficult and complex task. There is no one-to-one matching between the letters in the Ottoman alphabet and the Turkish alphabet. Alphabet translation step is a difficult problem because it is multi-step and complex, these steps require language resources such as dictionaries, spelling guides, corpus, and some phonetic, morphological, grammatical, and semantic problems have to be solved in the process. Alphabet translation consists of (i) Orthographic Alphabet translation, (ii) word segmentation (iii) spelling correction (iv) vowelization (v) word prediction (n-grams) (vi) phrase and compound nouns. Each of these sub-steps is a big problem on its own. Dictionaries have been created and organized for these sub-steps, the books have been collected for the corpus, and algorithms have been developed. An alphabet translation tool has been developed that covers these sub-steps. An alphabet translation tool – the only app of its kind – has achieved a 98% word accuracy rate on a 7500-word test dataset. The third step, language translation, is the translation of the Ottoman text in the Turkish alphabet into Contemporary Turkish with computerized translation. Although the Ottoman text produced during the alphabet translation step is readable for ordinary people, it is often not understandable. Because there are too many Arabic and Persian words and phrases in Ottoman. These words and phrases are unfamiliar not

only with their meanings but also their structures foreign to people today. At this step, a rule-based one-to-one direct language-translation tool from Ottoman to Turkish has been developed. In order to test the success rate of this module, it is necessary to prepare language resources such as corpus and dictionaries. With these three modules, the end-to-end conversion from Ottoman to Turkish is presented as a holistic approach.

January 2022, 111 pages.

Keywords: Ottoman, Optical character recognition, Printed naskh font, OCR, Deep learning, Machine Translation, Transliteration, Language Translation

1. GİRİŞ

Eğitim ve Bilimde Sayısal Dönüşüm kapsamında Osmanlıca matbu (basılı) dokümanları Günümüz Türkçe'sine Derin Öğrenme modelleriyle aktaran bir sistem geliştirilmiştir. Osmanlıca kaynakların büyük bir kısmı yurt içi ve dışındaki devlet arşivlerinde, kütüphanelerde ve özel koleksiyonlarda muhafaza edilmektedir. Bu kaynakların bilimsel, tarihsel ve kültürel olarak ne kadar değerli olduğu herkesin malumudur. Bu kaynaklardaki milyonlarca belge, yüzbinlerce basılı kitap, dergi, vb. dokümanların manuel olarak yani elle teker teker okunup Günümüz Türkçe'sine aktarılması pratikte mümkün olmayan bir iştir. Her ne kadar bu konuda ülkemizde yapılmış birkaç Optik Karakter Tanıma (Optical Character Recognition - OCR) çalışması bulunmaktaysa da pratik olarak kullanılabilir ve doğruluk oranı yüksek bir çalışma henüz ortaya konulamamıştır. Son 5-10 yılda Doğal Dil İşleme (DDİ) ve Derin Öğrenme (Deep Learning - DL) konularında meydana gelen yeni gelişmeler daha önce Yapay Sinir Ağları (YSA) ile çözümlemesi mümkün olamayan ya da yüksek başarım sağlanamamış problemlerin artık yüksek başarı oranlarıyla çözülmesine imkân sağlamıştır.

Osmanlıca yaklaşık olarak 13. yüzyıldan 20. yüzyıla kadar Osmanlı İmparatorluğunda kullanılan bir yazı dilidir [1]. Günümüzde Latin (Roman) alfabesine geçiş yapıldığı ve kelimelerin çoğu kullanımdan kalktığı için Osmanlıca'yı hem okumak hem de anlamak güçtür. Millî kültürümüzün temelini oluşturan eserlerin büyük bir kısmı Osmanlıca'da yazılmıştır. Osmanlı arşiv ve kütüphanelerindeki kitap, dergi, gazete, defter, kayıt ve belgeler yüzlerce yıllık kültür, sanat ve tarih mirası içinde önemli bir yer tutmaktadır. Bu kaynaklarda saklı bilgiye hızlı, etkin ve doğru bir şekilde erişilmesi için başta OCR olmak üzere teknolojinin yardımına ihtiyaç vardır. TÜBİTAK destekli Osmanlıca'dan Günümüz Türkçe'sine Yapay Zekâ Destekli Uçtan Uca Aktarım projesi bu amaçla başladığımız bir projedir. Projenin çıktıları Osmanlica.com sitesinden kullanıcılara sunulmuştur. Projede Tablo 1.1'de görüldüğü gibi bütüncül bir yaklaşımla Osmanlıca dokümanların 4 adımda bir uçtan diğer uca yani dokümandan Türkçe metne otomatik olarak çevrilmesi amaçlanmaktadır:

1. Doküman-görüntü dönüşümü (document-image conversion, digitization): Dokümanın tarayıcı veya fotoğraf makinasıyla taranıp, sayısallaştırılması ve ardından JPG, PNG vb. görüntü dosya formatına çevrilmesidir. Bu adım sıradan bir işlem olduğundan projeye dâhil değildir.
2. Görüntü-metin dönüşümü (image-text conversion, OCR): Görüntü dosyasının görüntü işleme ve makine öğrenmesi ile Osmanlı alfabesindeki Osmanlıca metne (editable text) dönüştürülmesi işlemidir. Osmanlıca Optik Karakter Tanıma (Osmanlıca OCR); Bu adım sonunda Osmanlıca bir resim ya da dokümandan Osmanlıca düzenlenebilir metin elde edilmiştir. Osmanlıca OCR yapılırken resimdeki bazı iyileştirmelerin yapılması gerekmektedir. Görüntü ön işlemler ile görüntünün normalleştirilmesi, gürültü temizleme, görüntü iyileştirme, görüntü bölümleme vb. algoritmalar kullanılmıştır. Osmanlıca OCR'da karakter tanıma için derin öğrenme modelleri tercih edilmiştir. Osmanlıca OCR tamamlandıktan sonra yazım kontrolü, kelime bölütleme kontrolü, karakter kontrolü vb. işlem sonrası (post-processing) işlemleri uygulanmıştır.

3. Alfabe (harf) çevirisi (harf çevrim, machine transliteration): Osmanlı alfabesindeki Osmanlıca metnin Türk (Latin Harfli Modern Türkçe) alfabesine metni değiştirmeden yani birebir bilgisayarlı ortografik alfabe çevirisidir. Burada her ne kadar birebir çeviri terimi kullanılsa da, ebced (abjad) tabanlı Arapça'da olduğu gibi Osmanlıca'da genellikle sadece sessizler ve uzun sesliler yazıldığından bilgisayarlı alfabe çevirisi karmaşık bir işlem olup bilindiği üzere şimdiye kadar kullanıma açık web tabanlı uygulaması olan sadece bir adet prototip çalışma mevcuttur [2]. Tablo 2.2'de görüldüğü gibi Osmanlı ve Türk alfabesindeki sesli ve sessiz harfler arasında çoktan çoğa bir eşleşme söz konusudur. Alfabe çevirisinin Türkçe'deki imla ve gramer kuralları (orthographic and syntactic rules) uygun şekilde yapılması gerekmektedir. Türkçe'de uzun seslilere ek olarak kısa sesliler de yazıldığından Osmanlıca ve Türkçe kelimeler arasında da çoktan çoğa eşleşme söz konusudur. Bu yüzden alfabe çevirisinde yüzeysel yapıbilimsel çözümleme (shallow morphological parsing), yapıbilimsel üretim (morphological generation), belirsizlik giderme (disambiguation), seslendirme (vowelization) vb. işlemlerin yapılması gerekmektedir.

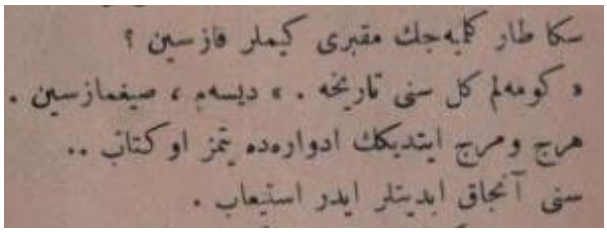
   Alfabe çevirisi (transliteration) zaman zaman Osmanlıca'dan Türkçe'ye fonetik alfabe çevirisi (yazıçevrim, transcription, çeviri yazı, transkripsiyon) ile karıştırılmaktadır.

Terim olarak transkripsiyon “Bir yazıyı kolayca okunacak şekilde belli bir işaret sistemi ile yazma; bir dildeki kelimelerin başka bir dilin alfabesi ile belirli işaretler kullanılarak yazılması; kelimelerin telaffuz edildiği şekilde yazıya geçirilmesi; çevriyazı.” şeklinde verilir [3]. Fonetik alfabe çevirisinde kullanılan alfabe edebiyat ve tarihçiler tarafından transkripsiyon alfabesi denilen IPA (International Phonetic Alphabet) benzeri bir alfabedir. Bu alfabe standart Türk alfabesiyle ifade edilemeyen seslerin ifadesi için oluşturulan bir alfabedir. Bu alfabe genellikle akademik çalışmalarda edebi ve tarihi metinlerin Türkçe’ye fonetik olarak düzgün aktarılarak metnin kolay okunmasına yardımcı olmak amacıyla kullanılmaktadır. Aşağıda transkripsiyonlu bir metin örneği verilmiştir:

> Çatma , ḳurbān olayım , çehreñi ey nāzlı hilāl!
> Ḳahramān ᶜırḳıma bir gül. Ne bu şiddet, bu celāl?
> Saña olmaz dökülen ḳanlarımız ṣonra ḥelāl;
> Ḥaḳḳıdır, ḥaḳḳa ṭapan, milletimiñ istiḳlāl.

4. Dil çevirisi (intra-language machine translation): Türk alfabesindeki Osmanlıca metnin bilgisayarlı çeviriyle Modern yani Günümüz Türkçe’sine çevirisidir. Çeviri sonucu ortaya Türk Alfabesinde Modern Türkçe metin çıkar. Dil çevirisi doğal dil işlemenin en önemli ve en zor problemlerinden bir tanesidir. Ama Osmanlıca - Türkçe Dil çevirisi, aynı dilin farklı özellikler taşıyan iki farklı zaman dilimindeki versiyonları arası çeviri olduğundan farklı diller arası çeviriye göre daha basittir. Fakat bu çeviri söz dağarcığı, söz dizim ve anlamsal anormalliklerden dolayı kolay bir problem değildir. Bu adımda kelime-grubu tabanlı birebir çeviri yaklaşımıyla ilerlenebileceği gibi, derin öğrenme kütüphaneleri kullanılarak cümle tabanlı makine çevirisi yaklaşımı da kullanılabilir. Çeviri öncesi (pre processing), sırası ve sonrasında (post processing) yüzeysel/tam yapıbilimsel ve/veya sözdizimsel çözümleme, üretim, belirsizlik giderme (shallow/full morphological & grammatical parsing, generation, disambiguation), kelime anlamı belirsizliği giderme (word sense disambiguation), varlık ismi tanıma (named entity recognition) vb. işlemlerin yapılması gerekmektedir.

Tablo 1.1: Osmanlıca-Türkçe uçtan uca çeviri (Çanakkale şehitlerine, Mehmet Akif)

*Adım (girdi → **işlem** → çıktı)*	*Çıktı*
1. Doküman-görüntü dönüşümü: *doküman → **sayısallaştırma** → görüntü dosyası*	
2. Görüntü-metin dönüşümü: *görüntü dosyası → **OCR** → Osmanlıca metin (Arap alf)*	سکا طار کلمیه جك مقبری کیملر قازسین؟ کومه لم کل سنی تاریخه دیسه م، صیغمازسین. هرج و مرج ایتدیکك ادواره ده یتمز او کتاب. . سنی آنجاق ابدیتلر ایدر استیعاب
3. Alfabe çevirisi *Osmanlıca (Arap alf) → **çeviri** → Osmanlıca (Latin alf)*	Sana dar gelmeyecek makberi kimler kazsın? "Gömelim gel seni târîhe" desem, sığmazsın. Herc ü merc ettiğin edvâra da yetmez o kitâb... Seni ancak ebediyyetler eder istîâb
4. Dil çevirisi: *Osmanlıca (Latin alf) → **çeviri** → Türkçe (Latin alf)*	Sana dar gelmeyecek mezarı kimler kazsın Gömelim gel seni tarihe desem sığmazsın Alt üst ettiğin çağlara da yetmez o kitab Seni ancak sonsuzluklar eder zapt

Görüldüğü gibi bu tezdeki yaklaşım bütüncül bir yaklaşımdır. Problemi bir bütün olarak ele alıp (i) Osmanlıca OCR (ii) Alfabe çevirisi (iii) Dil çevirisi olmak üzere 3 aşamada çözümü verilmiştir. Bu çalışma; web tabanlı (Osmanlica.com) olarak erişime açılmıştır. Daha önce bu yaklaşımla tasarlanan ve geliştirilen bir tezin veya çalışmanın bulunmaması kendi türünde ilk çalışma olmuştur.

2. GENEL KISIMLAR

2.1. OSMANLICA OCR

Bu bölümde; Osmanlı Alfabesi, Benzer Çalışmalar, Optik Karakter Tanıma ve Osmanlıca OCR problemlerinden bahsedilecektir.

2.1.1. Osmanlı Alfabesi

Osmanlı alfabesi Arapça'dan 28 (ابتثجحخدذرزسشصضطظعغفقكلمنهوى), Farsça'dan 4 (پچژگ), Türkçe'den 3 (ۆۋە) ses alınarak oluşturulmuştur. Osmanlıca Arapça tabanlı olduğu için harfler -rakamlar hariç- sağdan sola (RTL) ve bitişik (joined) yazılır. Sadece 8 harf (ادذرزژوه) kendinden sonra gelen harfle bitişmez. Bitişmeyen (non-joiner) harflerin münferit (tek başına) ve sonda (önceki harfle bitişen), bitişen (joiner) harflerin münferit, başta (sonraki harfle bitişen), ortada (önceki ve sonraki harfle bitişen) ve sonda farklı yazım şekilleri vardır.

Osmanlıca'daki harflerin 16 adedi (احدرسصطعكلمهوگۀ) noktasızdır. Tek noktalı 10 (بجخذزضظغفن), iki noktalı 3 (تقي), ve üç noktalı 6 (ثشپچژڭ) olmak üzere 19 adet noktalı harf vardır. Noktalı harflerin 14 tanesinin noktası üstte (تثخذزشضظغفقنژڭ) beşinin noktası alttadır (بجيپچ). Harflerin bazıları (ورزژ vb.) satır çizgisinin (baseline) altına taşmakta (descenders) diğerleri (ڭگگکا١ل vb.) çizginin üstünde (ascender) yazılmaktadır. Noktalar yok sayılıp harfler gruplandığında tekil harf şekli (ana harfler, skeleton script) sayısı 35'ten 11'e düşmektedir:

ء, ھ, ٥٥,)ل١١, ڭگگک, عغفقمو, صضطظ, رزژ, جحخچ, سشـ, بتثپنيذند

Benzerlik gösterdiğinden bazı gruplara noktalama işaretleri ve rakamlar eklenmiştir. Geri kalan diğer rakam ve noktalama işaretleri de eklenince grup sayısı artmaktadır. Bu gruplamada şekil yönünden benzer harfler bir araya toplanmıştır. Harfler şekil yönünden delikli (looped) (صضطظ ,٥٥ھ ,عغفقمو), dişli (toothed) (سشـ, بتثپنيذند), düz çizgili (line segment) (ڭگگک ,(ل١١) ve diğerleri olarak dört ana gruba ayrılabilir. Görüldüğü gibi gruplama sübjektif bir gruplamadır. Grupların sayısı arttırılabilir veya azaltılabilir. Örneğin صضطظ grubu صضـ ve طظ şeklinde iki gruba bölünebilir. Bitiştiğinde dişli yazılan deliksiz harfler iki grupta toplanmıştır: بتثپنيذند سشـ. بتثپنيذند grubu بتثپنـ ve ـذند şeklinde ikiye bölünebilir ya da سشـ grubuyla birleştirilebilir. ه hariç delikli harfler iki grupta toplanmıştır: عغفقمو ve صضطظ. Vav "و" harfi delikli olmasına

rağmen kendi başına ayrı bir grup olabilir. Osmanlı alfabesi Arapça'daki gruplama esas alınarak gruplandığında 15 farklı grup oluşmaktadır. Tablo 2.1'de gösterilmiştir.

Tablo 2.1: Osmanlıca alfabesi şekillerin gruplanması

Tipi	Harf				Ana Harf				Kod
dişli	بـ تـ	بـ تـ	بـ تـ	ب ت	ـبـ	ـبـ	ـب	ب	066E
	ثـ پـ	ثـ پـ	ثـ پـ	ث پ					
	نـ	نـ	نـ	ن	ـنـ	ـنـ	ـن	ن	06BA
	ی	یـ	یـ	ی	ـی	ـی	ـی	ی	0649
	سـ شـ	سـ شـ	سـ شـ	س ش	ـسـ	ـسـ	ـس	س	0633
diğer	ـذ ـذ			ذ د	ـذ			د	062F
	ـز ـر ـژ			ز ر ژ	ـر			ر	0631
	ـجـ ـحـ ـخـ ـچـ	ـجـ ـحـ ـخـ ـچـ	ـجـ ـحـ ـخـ ـچـ	ج ح خ چ	ـحـ	ـحـ	ـح	ح	062D
delikli	ـصـ ـضـ	ـصـ ـضـ	ـصـ ـضـ	ص ض	ـصـ	ـصـ	ـص	ص	0635
	ـطـ ـظـ	ـطـ ـظـ	ـطـ ـظـ	ط ظ	ـطـ	ـطـ	ـط	ط	0637
	ـعـ ـغـ	ـعـ ـغـ	ـعـ ـغـ	ع غ	ـعـ	ـعـ	ـع	ع	0639

	ق ڧ	ڤ ڤ	ق ڤ	ق ڧ	ى	ـى	ـه	ى	066F
	و			و	و			و	0648
	م	ـم	م	م	م	ـم	ـم	م	0645
	ه ه	ھ	ھ	ه	ه ه	ـھ	ـھ	ه	0647
düz	ك ڭ گ	ک ڪ گ	ک ڪ گ	ك ڭ گ	ك ك ک	ک	ک	ك ک	0643
	ل	ـل	ـل	ل	ل	ـل	ـل	ل	0644
	ا			ا	ا			ا	0627
	ء	ء	ء	ء					0621

Yukarıdaki harf analizi OCR çalışmalarında yönlendirici rol oynamaktadır. Örneğin bazı çalışmalarda OCR işlemi harfler yukarıdaki gibi gruplandırılarak üç aşamalı yapılmaktadır [4]. İlk aşamada metindeki nokta ve harekeler (diacritics) temizlenerek metin ana harflere dönüştürülmekte ve ikinci adımda bu ana harflerin tanıması yapılmaktadır. Üçüncü adımda ana harflerden noktalı yani normal harflere dönüşüm yapılır. Bu yaklaşımın avantajı tanınacak ana harf sayısının normal karakterlere göre çok daha az olması, dezavantajı ise ön işlemede noktaların temizlenmesi ve son aşamada noktalı harflere dönüşümde yaşanan zorluklardır.

2.1.2. Benzer çalışmalar

Arapça, Farsça, Urduca, Uygurca, Malay (Jawi), Kürtçe (Sorani), Peştu, Sindi, Arvi, gibi el yazısı (kavisli ve bitişik) alfabeye sahip diller (cursive languages) Osmanlıca'yla çok sayıda ortak ve benzer özelliklere sahiptir. Osmanlıca'nın aksine günümüzde bu dil ve alfabeler dünyada yüz milyonlarca kişi tarafından kullanılmaktadır. Bu dillerde genellikle Şekil 2.1'de örnekleri verilen nesih, rika (riq'ah, ruqʿah رقعة), nastalik, sülüs, kûfi ve divani hatları kullanılır [5]. Nesih ve nastalik diğerlerinden daha yaygındır. Nesih Arapça'da kitap, dergi gibi basılı yayınlarda ve bilgisayarda kullanılan standart fonttur. Arapça ve Farsça genellikle nesih hattıyla Urdu ise nastalikle yazılır. Bu yüzden Arapça [6], Farsça, Urdu, hatta Latin el yazısında yapılan OCR çalışmaları Osmanlıca için önemli bir referans oluşturmaktadır.

Şekil 2.1: Arapça hatlar (Anonim)

Osmanlıca dokümanlarda satır bölümleme ve OCR problemine odaklanan ilk çalışmalar 90'lı yıllarda yapılmıştır [7, 8]. Bu çaba yakın zamanlara kadar başka gruplarca da devam ettirilmiştir [9-11]. Satır bölümleme probleminin yanı sıra [9] ve [12, 13] sözcük içindeki karakterlerin verimli bir temsilini oluşturmayı amaçlar. Arapça karakterler için bir veri tabanı [14] numaralı referansta sunulmuştur. Bunların dışında, genel olarak Arapça ve benzeri diller için faydalı yaklaşımların derlendiği çerçeve oluşturma niteliğindeki çalışmalara örnek olarak [15-18] verilebilir. [16]'da yazarlar OCR çıktısı ölçme ve değerlendirme metrikleri sunmaktadırlar. Osmanlı alfabesi ile yazılmış metinlerin optik karakter tanımasında farklı yöntemler denenmiştir. [12-13]'te yazarlar Doğrusal Diskriminant Analizi yöntemi, [19]'te tek katmanlı bir Yapay Sinir Ağı ile harfleri birer birer tanımayı amaçlar. [20]'deki tanıma sistemi Osmanlıca basılı dokümandaki sözcükleri tanıdıktan sonra sözcükleri harflere bölümleyip Destek Vektör Makinesiyle harfleri tanımak suretiyle sözcükleri tanırlar. [21]'deki tanıma sistemi sözcükleri harflere böldükten sonra her bir harfi tek katmanlı bir Yapay Sinir Ağı ile [20] tanımak suretiyle sözcükleri tanır. Harflerin bitişik yazıldığı Osmanlıca gibi alfabelerde harfleri birbirinden ayırmak zordur. Harfleri tanımada başarı bile elde edilse bölümleme adımındaki yanlışlar toplam sistem başarısını olumsuz etkiler. Sözcükleri harflere bölmeksizin tanıma yaklaşımı [7]'de ve [22]'de Gizli Markov Modelleri kullanılarak yapılmıştır. Bunlara ek olarak [23]'te arama ve sorgulama amacıyla OCR yapan bir sistem geliştirmiştir. Çalışmada karakter bölümlemesinden sonra farklı alternatif yöntemler izlenerek karakterler tanınmaya çalışılmıştır. Bu yöntemler [24] çalışmasıyla benzerlik gösteren görsel karşılaştırma, [19] ve [20] ile benzerlik gösteren Yapay Sinir Ağı ve çizge temelli bir karşılaştırmadır.

İçerik tabanlı erişim (content based retrieval) yaklaşımını kullanan çalışmalarda amaç OCR yapmadan Osmanlıca dokümanlar içinde harf parçası, harf, bağlı karakter katarı, kelime arama

ve sorgulaması sağlamaktır; içerik tabanlı sorgulama görüntü içinde verilen bir örüntünün araması (image search), eşleşmesi (image matching) ve en iyi eşleşmelerin görüntülenmesine dayanır. İçerik tabanlı erişim yöntemlerinde iki farklı yaklaşım mümkündür. Bu yaklaşımların ilki görüntü içerisindeki bağlı şekiller kümesinde, metinlerdeki sözcükleri ve alt sözcükleri bulan yaklaşım (codebook yaklaşımı); diğeri ise doğrudan örüntü araması yapan sözcük eşleme (word spotting, word matching) yaklaşımıdır. İlk yaklaşıma [24-26] çalışmaları örnek verilebilir. Sözcük eşleme yaklaşımına ise [27, 28] ve [29, 30] örnek verilebilir. Bununla beraber son yıllarda birçok alanda çeşitli problemlere çok başarılı bir şekilde uygulanan derin öğrenme OCR'a da başarıyla uygulanmıştır ve Arapça için oldukça iyi sonuçlar alınmıştır [31, 32].

Akademik çalışmalar dışında Osmanlıca OCR için özel olarak geliştirilen araçlara örnek olarak Miletos'un geliştirdiği web tabanlı (miletos.com) araç verilebilir. Bildiğimiz kadarıyla bu araç IRCICA'nın kuruma özel OCR aracı (library.ircica.org) gibi kapalı sistemler hariç tutulduğunda, kendi türünde ilk ticari Osmanlıca OCR aracıdır.

2.1.3. Optik Karakter Tanıma

OCR bir resim dosyasında saklanan görüntüdeki karakterlerinin görüntü işleme ve makine öğrenmesi teknikleri kullanılarak metne (string) dönüştürülmesi işlemidir [33]. Son zamanlarda doküman resimleri yerine rastgele resimden metin tespiti ve tanıma (scene text detection and recognition) popüler bir konu olarak öne çıkmaktadır [34]. OCR temelde bir karakter sınıflandırma problemidir. OCR problemi bazen bağlı-karakter-katarı (connected component) tanıma problemi veya kelime tanıma problemi olarak ta ele alınmaktadır. OCR'da işlemler genellikle 5 adımda gerçekleştirilir:

1. *Görüntü ön işleme (image preprocessing)*: Bu adım görüntü dosya formatının ve görüntü boyutunun normalleştirilmesi, gerekliyse metin fontunun ve boyutunun belirlenmesi, görüntünün filtrelenmesi, eşikleme (tresholding), siyah-beyaza dönüştürme (binarization), gürültü temizleme (noise reduction), görüntü zenginleştirme (image enhancement) vb. ön işlemleri içerir. Amaç bir sonraki adım için en iyi görüntünün oluşturulmasıdır.
2. *Görüntü bölümleme (image segmentation)*: Tanıma işleminin yapılabilmesi için görüntünün tanınacak birimlere bölümlenmesi gerekmektedir. Bu adımda sayfadaki

metinlerin tespiti (text detection) için sayfa düzeni çözümleme (layout analysis) [35] ve sayfa bölümleme (page segmentation) [36], satır bölümleme (line segmentation), kelime bölümleme (word segmentation), bağlı-karakter-kataı (connected component, ligature) ya da alt-kelime (sub word) bölümleme, karakter bölümleme (character segmentation) ya da karakter parçası olarak eğri (stroke, sub character) bölümleme vb. işlemler görüntü işleme teknikleriyle yapılır. Amaç bir sonraki adım için görüntünün içinden tanıması yapılacak doğru bölgenin (minimum bounding box) tespit edilmesidir. Görüntü bölümleme sırasında yapılan işlemler arasında dil tanıma (language identification), alfabe tanıma (script identification), yazı tipi ve stili (font and style recognition) tanıma gibi birtakım çözümlemelerin yapılması gerekebilir [37] .

3. *Özellik çıkarma (feature extraction)*: Tanıma işlemi bir sınıflandırıcı modelle yapıldığından, modeli eğitmek için kullanılacak özelliklerin baştan bir defa belirlenmesi gerekmektedir. Özellik kümesi en basit şekliyle 2 boyutlu görüntüdeki noktaların kümesi (genişlik x yükseklik) olabileceği gibi, kullanılan yaklaşıma bağlı olarak seçilen veya hesaplanan çok farklı parametreler, geometrik özellikler veya karmaşık görüntü verisi olabilir [38].
4. *Tanıma (recognition, OCR)*: Tanınmak istenen birimin (eğri, karakter, alt-kelime, kelime, vb.) bir sınıflandırıcı model kullanılarak tanınması (recognition, labeling, classification) işlemidir. Bu adımda çoğunlukla karakter tanıması yapıldığından işlem optik karakter tanıma OCR olarak adlandırılır. Son zamanlarda tanıma işlemi daha yüksek doğruluk oranları sağladığından derin öğrenme sınıflandırıcı modeller kullanılarak yapılan bir işlemdir [39].
5. *Düzenleme ve hata düzeltme (post processing & error correction)*: Tanıma adımından sonra OCR çıktısındaki karakterlerin normalizasyonu ve OCR'da tanınmamış karakter veya kelimelerin tahmini için yapılan düzenleme ve hata düzeltme işlemleridir. Bu adımda sözlükte arama (dictionary lookup), yazım düzeltme (spelling correction), yüzeysel morfolojik analiz (shallow morphology), dil modelleri (n-grams) vb. teknikler kullanılır. Düzenleme aşamasında OCR iş akışı işlemlerinin ve OCR çıktısının yapılandırılması, saklanması, paylaşılması ve işlenmesi için geliştirilen hOCR - OCR Workflow and Output embedded in HTML standardı [40] kullanılarak çıktılar saklanabilir ve github.com/ocropus/hocr-tools, pythonrepo.com/repo/tmbdev-hocr-tools-python-computer -vision vb. hOCR araçları kullanılarak işlenebilir. Bu aşamadan

sonra Arapça'da yapıldığı gibi [41] Osmanlıca'yı kolay öğrenme ve okuma amacıyla metnin tamamının otomatik harekelendirilmesi (automatic diacritization, vocalization) faydalı bir uygulama olarak geliştirilebilir.

2.1.4. Osmanlıca OCR'da Problemler

Bu bölümde OCR yapılırken Osmanlı alfabesine özgü problemlere değinilecektir. Bu problemlerin çoğu Arap, Fars, Urdu vb. alfabeler için de söz konusudur.

1. *Kavislilik, el yazısı (cursiveness, hand writing)*: Latin alfabesinde matbu yani düz yazı (printed) ve el yazısı (hand writing) olmak üzere iki farklı temel yazı tipi bulunmaktadır. Latindekinin aksine Osmanlıca'da hem düzyazı hem de el yazısı kavisli yazı tipleridir. Osmanlıca'da daha okunaklı olan nesih hattı düz yazıya, okuması daha zor olan rika hattı el yazısına karşılık gelmektedir. Diğer bazı hatlar daha karmaşık yazıma sahiptir. Osmanlıca'daki matbu yazı tipi - karakterler düz çizgilerden ziyade eğri çizgiler kullanılarak yazıldığından - Latindeki el yazısına benzerlik gösterir. Osmanlıca'daki el yazısının (Rika) tanınması matbu nesih hattına göre daha zor olduğundan bu hatla yazılı dokümanlarda OCR yerine görüntü arama ve eşleme (image search and matching) teknikleri kullanan içerik tabanlı arama yöntemleri tercih edilmiştir.
2. *Bitişme (connectedness)*: Latin alfabesinde karakterler birbirinden ayrık (block letters) yazılmaktadır ve karakter bölümleme kolaydır. Osmanlıca'da harflerin çoğu bitişik yazıldığından karakter bölümleme ve karakter tanıma Latine göre daha zordur.
3. *Nokta (dot):* سم الله الرحمن الرحيم (بِسْمِ ٱللَّٰهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ) örneğinde olduğu gibi Arap alfabesinde eskiden harfler noktasız (*resm Ar.*) yazılırken, okumayı kolaylaştırmak için nokta (*icam Ar.*) alfabeye sonradan eklenmiştir. Harekeler ise (kısa) seslileri göstermek için daha sonra eklenmiştir. Arapça'da noktalar göz ardı edilip harflerin münferit ve sonda yazılışları incelendiğinde 18 ana harf gözlenir [42]. Harflerin başta ve ortadaki yazımları göz önüne alındığında ana harf sayısı 18'den 15'e düşer: ا بتثنيـ حخجـ دذ رز سشـ صضـ طظ عغـ فقـ ک لـ مـ هـ و . Osmanlıca'da Arapça gibi nokta sayısı ve konumunun harfi tanımada kullanıldığı noktalı bir alfabedir. Osmanlıca'da bir, iki veya üç noktalı harfler varken, Arapça tabanlı bazı alfabelerde nokta sayısı dörde kadar çıkmakta ve harflerin birçoğunun üstüne veya altına farklı sayıda noktalar getirilerek ڇڃچ ځڅخ ڀٻپ ڎڐڏ vb. yeni harfler elde edilmiştir. Osmanlıca'da noktalar göz ardı edilip benzer harfler gruplandırıldığında 35 harf 15 ana harfe dönüşmektedir

[4]. Bu durum Arapça tabanlı alfabelerin OCR işleminde alternatif çözümlerin geliştirilmesine yol açmıştır [4]. Noktalarla ilgili OCR'ı zorlaştıran bazı durumlar söz konusudur: (i) *Konum*: Noktalar harfin gövdesine göreceli olarak belirli bir konumda tanımlanırlar. Fakat yazı içinde noktaların farklı konumlara yerleştirildiği çokça gözlenen bir durumdur, (ii) *Adet*: Osmanlıca'da harfler bir, iki veya üç noktalıdır. Yazım sırasında noktalar birbiriyle bitiştiğinde iki nokta bir nokta olarak veya üç nokta iki nokta olarak algılanabilir. Ayrıca nokta(lar) yerine سۤ سٓ örneğinde görüldüğü gibi U+065A, ٚ Arabic Vowel Sign Inverted Small V Above, U+065B) kullanımı (alāmātu-l-ihmāl Ar.) yaygın bir uygulamadır. (iii) *Bitişme*: Mürekkep dağılması, yazı boyutunun küçüklüğü, düşük görüntü çözünürlüğü veya kalitesi vb. sebeplerden bazen noktalar harfin gövdesiyle bitişmekte ve bu durum tanımayı güçleştirmektedir. Bazı durumlarda noktalar kendi aralarında değil harfin gövdesiyle bitişmektedir. (iv) *Uzaklaşma*: Bazen noktalar harften uzakta yazıldığında algılanmaz veya başka bir harfin noktası gibi algılanabilir. (v) *Atlama*: ڭ yerine yaygın olarak ك kullanımında olduğu gibi okuyucunun metni doğru okuyabileceği varsayımıyla noktalar ve özel işaretler bazen bilinçli olarak yazılmaz, yani atlanır.

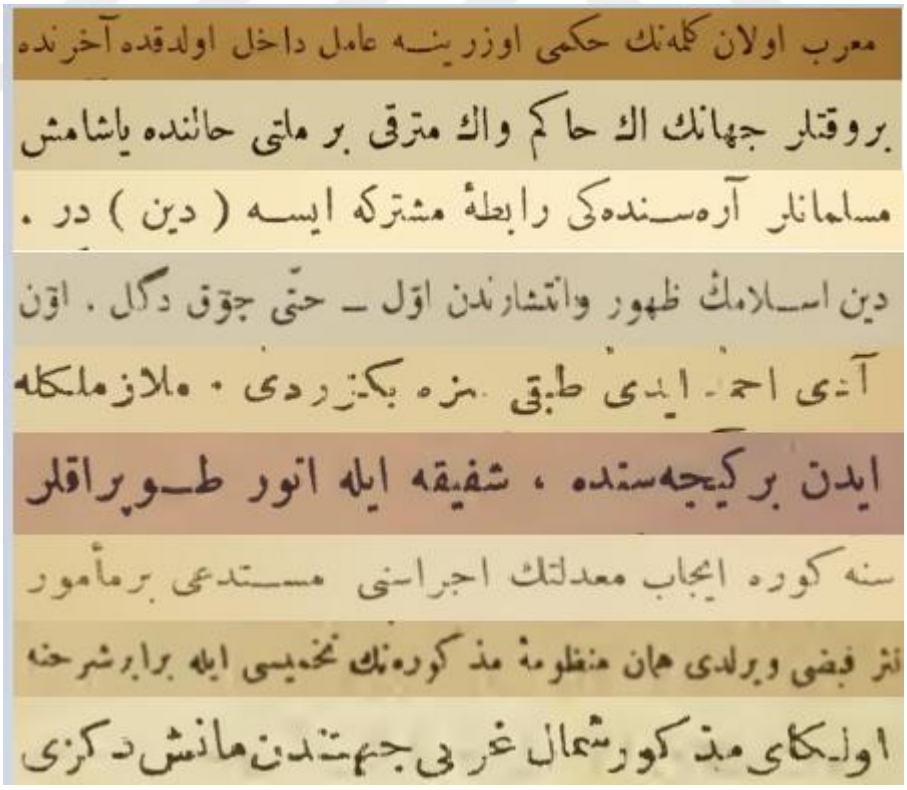
4. *Bağlamsal harf şekilleri (contextual letterforms)*: Osmanlıca'da çoğu harfin başta, ortada, sonda ve münferit olmak üzere dört farklı yazımı bulunmaktadır. Bazı harflerin ھ(ھـ ـھـ ـھ ہ ـہ ـہ) örneğinde olduğu gibi farklı dil ve kullanımdan kaynaklı dörtten fazla yazımı bulunmaktadır. ى ىٖ ي örneğinde görüldüğü gibi bazı harflerin konuma bağlı noktalı ve noktasız yazımları, آ أ إ ؤ ئ ۂ örneklerinde olduğu gibi bazı harflerin hemze veya med gibi harekeyle yazılan şekilleri bulunmaktadır. Bu durum Osmanlıca'da karakter tanımayı - büyük ve küçük olmak üzere sadece iki farklı yazım şekline sahip - Latin Alfabesine göre daha güç hale getirmektedir.
5. *Harf benzerliği (letter similarity)*: Osmanlıca'da çoğu harfler şekilce birbirine benzer olduklarından ve ٥٥ بتثپنیند سسشـ ججحخچ رززژ صضطظ عغفقمـ ککگگالـ örneklerinde olduğu gibi bazı harfler sadece nokta sayısı veya konumu bakımından birbirinden farklılaştığından Osmanlıca'da OCR Latindekine göre daha güçtür.
6. *Yığma yazı (letter stacking, overlapping letters)*: قی یی örneklerinde olduğu gibi Osmanlıca'da bitişen harfler dikeyde tamamen veya kısmen üst üste yani yığılarak

yazılmaktadır. Bu durum tanınacak bağlı karakter katarı çeşidini arttırmaktadır. Osmanlıca'da divani, muhakkak gibi bazı hatlar yığma stilinde yazılmaktadır.

7. *Kelime bölümleme (word segmentation)*: Osmanlıca'da bitişmeyen karakterlerden dolayı kelime içi boşlukla (zero length non space, intra word space) kelime arası boşluklar (inter word space) birbirine karışabilmekte ve bu durum kelime bölümlemeyi zorlaştırmaktadır. Oysa Latinde böyle bir problem söz konusu değildir.
8. *Çoklu karakter kodlaması (multiple character encoding):* Osmanlıca metinlerin OCR çıktısında sık sık karakter yazım ve kodlama hataları gözlenmektedir. Bunun nedenlerinden birisi bir harfin Unicode'da birden fazla alternatifin (allograph) bulunmasıdır. Osmanlıca'ya özel olarak, okutucu he (Unicode u06d5 ە ARABIC LETTER AE) ile normal he (Unicode u0647 ه ARABIC LETTER HEH) harfleri belirli durumlarda aynı şekilde (ه) yazılmakla birlikte farklı karakter kodlarıyla gösterilirler. Kullanıcıların veri seti hazırlarken farkında olmadan bazen standart dışı karakterleri seçmesinden dolayı, veri setlerindeki harflerin karşılaştırması ve sayımı güçleşmektedir. Buna ek olarak bazı harflerin münferit ve bitişik yazımlarında noktaların yeri dahi değişebilmektedir. Sonuç olarak karakterleri doğru şekilde işlemek ve sayabilmek için Osmanlıca karakterlerin normalizasyonuna ihtiyaç vardır.
9. *Hareke (diacritics)*: Osmanlıca'da harekesiz yazılmakla birlikte ihtiyaca binaen ا و ى gibi bazı harfler آ أ إ örneklerinde olduğu gibi sadece hemze ve med harekeleriyle birlikte yazılır. Harekeler tanınacak harf sayısını arttırdığı için OCR zorlaşmaktadır. Osmanlıca'da OCR çıktısı metindeki harekeler normalizasyon kapsamında genellikle temizlenir.

OCR'da dilden bağımsız olarak doğruluk oranına etki eden bazı önemli faktörler aşağıda verilmiştir. Bu faktörlerden en önemlileri görüntünün çözünürlüğü, kalitesi, tanınacak birim, yazı hattı, eğitim veri seti ve sözlüğün kalitesi olarak öne çıkmaktadır:

1. *Görüntü çözünürlüğü (image resolution)*: Çözünürlük azaldıkça veri kaybı söz konusu olduğundan OCR başarısı için mümkün olduğunca yüksek çözünürlük önerilmektedir. Karakter boyu (font size) ya da satır yüksekliği (line height) arttıkça OCR doğruluk oranı da artmaktadır. Osmanlıca OCR için en düşük karakter boyu genellikle 12 nokta (pixel), en düşük çözünürlük 300 dpi (inçteki nokta sayısı) olarak önerilmektedir.
2. *Görüntü kalitesi (image quality)*: Görüntüdeki lekeler (smudge), gürültü, silik, soluk veya okunması imkânsız hale gelmiş şekillerle zamanla özelliğini kaybeden kâğıtta ortaya çıkan görsel olgular (degradation), sayfa kenar çizgileri, sayfa süslemeleri, tarama kaynaklı yapay olgular (artifacts) ve görüntüdeki eğrilik (skew) görüntü kalitesini kötüleştiren etmenlerdendir. Şekil 2.2'de birkaç farklı Osmanlıca doküman görüntü örneği paylaşılmıştır. Görüntü bozuldukça OCR hata oranı yükseldiğinden, ön işlem aşamasında görüntü işleme teknikleriyle görüntü kalitesi arttırılmadan önce, bir seçenek olarak doküman görüntü kalitesi belirleme (document image quality assessment) işlemi yapılır [43]. Bu işlem sırasında doküman kalitesi ölçümü için farklı ölçek (metric) ya da skorlar kullanılır [44].



Şekil 2.2: Osmanlıca nesih hat örnekleri

3. *Tanınacak birim (unit recognized)*: Tanıması yapılacak olan birim, uygulama özelinde harf parçası (eğri), harf, bağlı harf katarı, kelime, hatta kelime grubu veya satır olarak seçilebilir.
4. *Yazı hattı (font, typography)*: Osmanlıca'da başta nesih olmak üzere ondan fazla hat çeşidi vardır [45]. Hattat Hamit Aytaç'ın Osmanlıca'daki hat çeşitlerini bir tablo olarak resmeden *Hutut-u Mütenevvia* adlı eser Şekil 2.3'te verilmiştir. Eserdeki 18 satır sırasıyla kûfi, sülüs, celî sülüs, nesih, muhakkak, sülüs, tevkiî, rikâa, rikâa, celî talik, talik, talik, dîvânî, celî dîvânî, rika, rikâa, rika, rika hatlarıyla yazılmıştır. OCR doğruluk oranı tanınacak hattın karmaşıklığına göre değişmektedir. Nesih gibi bazı hatlar daha basit, okunaklı ve anlaşılır olmasına karşın, bazıları çok karmaşık olduğundan Osmanlıca bilenlerce dahi okunması güç olabilmektedir.
5. *Eğitim seti (training set)*: Bilindiği üzere makine öğrenmesinde sınıflandırma modeli oluşturmak için kullanılan eğitim veri setinin kalitesi ve büyüklüğü oluşturulacak OCR modeli için kritik öneme sahiptir.
6. *Sözlük büyüklüğü (lexicon size)*: Sözlük kullanımı OCR ve kelime tanıma doğruluk oranını arttıran bir olgudur. Sözlük temelde tanınacak olan sınıfları içeren bir sınıf listesi olarak düşünülebilir ve OCR'da tanınacak birime göre farklı sözlükler kullanılır.

Şekil 2.3: Hutut-u Mütenevvia adlı eser

2.2. ALFABE ÇEVİRİSİ

Alfabe çevirisi, kaynak alfabedeki bir metnin hedef alfabeye aktarımı olarak tanımlanabilir. Bu aktarımda amaç hedef dilde yazma veya hedef dilde okuma olabilir. Birbirine karıştırmamak için, hedef dilde yazım amacıyla yapılan aktarıma *ortografik alfabe çevirisi*, okuma amacıyla yapılan aktarıma *fonetik alfabe çevirisi* olarak adlandırıyoruz. Ortografik alfabe çevirisi dilbilimde harf çevrim, harf çevirisi veya transliterasyon olarak, fonetik alfabe çevirisi ise yazı çevrim, çeviri yazı veya transkripsiyon olarak bilinir.

Alfabe çevirisinde kaynak ve hedef alfabeler daima farklıdır. Tablo 2.3'te alfabe çevirisinde kullanılan alfabeler verilmiştir. 1. sütunda kaynak alfabe olarak Osmanlı alfabesi, 2. sütunda harflerin bilgisayarda gösteriminde kullanılan uluslararası bir standart olan Unicode içindeki kodu, 3. sütunda harf çevrimde hedef alfabe olarak Türk alfabesi, 4. 5. ve 6. sütunlarda yazı çevrimde hedef alfabe olarak IPA, ALA-JC ve JIMES alfabeleri verilmiştir. Tablonun her satırında Osmanlıca bir harf ve hedef alfabelerde o harfe karşılık gelen harfler verilmiştir.

Ortografik ve fonetik alfabe çevirisine ait bazı temel özellikler Şekil 2.4'te gösterilmiştir. Ortografik çeviride çıktı, harfleri hedef dilin imla kurallarına uygun olarak dizilmiş bir yazı(m)dır. Fonetik çevirinin çıktısı ise, kaynak dildeki metnin hedef dilde doğru okunuşunu sesler yani fonemler kullanarak özel bir alfabede vermektir. Ortografik ve fonetik çeviride çıktının birimleri farklı olsalar da, çevirinin kendisi kelime kelime yapılır, yani genel anlamda çevirinin birimi kelimedir. İsim tamlamaları ve bileşik kelimelerin çevirisinde ise, zorunlu olarak kelime yerine kelime grupları kullanılır.

Alfabeler kaynak ve hedef alfabedeki harf veya seslerin eşleşmesi bakımından incelendiğinde, ortografik çeviri için hedef alfabenin *örten* olması yani kaynak alfabedeki tüm sesleri ifade etmesi yeterli olurken, fonetik çeviride buna ek olarak, hedef alfabedeki her harfin kaynak alfabede sadece bir sese karşılık gelmesi yani *birebir* eşleşmesi beklenir.

Geri dönüşlülük kaynak alfabeden hedef alfabeye çevrilmiş metin çıktısını tersine işletilen bir süreçle hedef alfabeden kaynak alfabeye geri dönüştürülmesi için girdi olarak verilerek orijinal kaynak metnin elde edilebilir olması sağlayan iki yönlü bir çeviri sürecidir. Geri dönüşlülük kabaca 3 adımda test edilir: Kaynaktan hedefe ileri yönlü çeviri, hedeften kaynağa geri yönlü çeviri, ileti yönlü çevirinin kaynak metniyle, geri yönlü çevirinin hedef metinlerinin karşılaştırması. Ortografik çeviride aranmayan bir özellik olmakla birlikte, iyi tasarlanmışsa

fonetik çevirinin geri dönüşlü olması arzu edilen bir özelliktir. Geri dönüşlü bir Osmanlıca-Türkçe fonetik alfabe ve çeviri, transkripsiyonlu metinden orijinal Osmanlıca metni geri döndürebilir.

Tablo 2.2'de görüldüğü gibi Osmanlı ve Türk alfabesindeki harfler arasında çoktan çoğa eşleşme vardır. Benzer şekilde Osmanlıca ve Türkçe kelimeler arasında da çoktan çoğa eşleşme söz konusudur. Osmanlıca'daki ses sayısı Türkçe'deki ses sayısından fazla, Osmanlı alfabesindeki harf sayısı da Türk alfabesindekinden fazladır. Sonuç olarak Türk alfabesinde Osmanlıca'daki tüm sesleri ifade edecek sayıda harf bulunmamaktadır.

Özellik	**Harf çevrim**	**Yazı çevrim**
İsimlendirme	Transliteration	Transcription
Çeviri türü	Ortografik alfabe çevirisi	Fonetik alfabe çevirisi
Amaç	Yazmak	Okumak
Çıktı	Yazım (ortografi, imla)	Okunuş (telaffuz)
Çıktı birimi	Harf (grafem)	Ses (fonem)
Çeviri birimi	Kelime, kelime grubu	Kelime, kelime grubu
Kaynak Alfabe	Osmanlı alfabesi	Osmanlı alfabesi
Hedef alfabe	Türk alfabesi	IPA, ALA-JC, JIMES
Hedef alfabe fonetik mi?	Fontetik olmayabilir	Fonetik
Kaynak-hedef eşleşmesi	Örten	Örten ve Birebir
Geri dönüşlü	Değil	Olabilir

Şekil 2.4: Harf çevrim ve yazı çevrim özellikleri

Tablo 2.2: Ortografik ve fonetik çeviri alfabeleri

Osmanlıca Harf	Ortografik çeviri	Fonetik çeviri		
		IPA	ALA-LC	JIMES
ا	a|e	æ|e	ā|'	a|e
ء	—|'|’|ʾ	ʔ	ʾ	ʾ
ب	b|p	b|p	b	b
پ	p	p	p	p
ت	t	t	t	t
ث	s	s	ṣ	ṣ
ج	c	d͡ʒ	c	c
چ	ç	t͡ʃ	ç	ç
ح	h	h	ḥ	ḥ
خ	h	x	ḫ	ḫ
د	d	d	d	d
ذ	z	z	ẕ	ẕ
ر	r	r	r	r
ز	z	z	z	z
ژ	j	ʒ	j	j
س	s	s	s	s
ش	ş	ʃ	ṣ	ṣ
ص	s	s	ṣ	ṣ
ض	z|d	z|d	ż	ż
ط	t|d	t|d	ṭ	ṭ
ظ	z	z	ẓ	ẓ
ع	—|'|‘|ʿ	ʔ	ʿa	ʿa
غ	g|ğ	g|j	ġ	g|ğ
ف	f	f	f	f
ق	k	k	q	ḳ
ك	k	k	k	k|g|ğ|y
گ	g|ğ	g|j	g	g
ڭ	n	ŋ	ñ	ñ
ل	l	l	l	l
م	m	m	m	m
ن	n	n	n	n
و	v|o|ö|u|ü	v|o|œ|u|y	v|ū|aw|avv|ūv	v
ه	h|a|e	h|æ|e	h	h
ى	y|ı|i	j|ɯ|i	y|ī|ay|á|īy	y
34	31	31	43	34

Türkçe fonetik, Osmanlıca ise fonetik olmayan bir dildir. Kabaca, fonetik dil okunduğu gibi yazılan ya da yazıldığı gibi okunan dil demektir. Başka bir tabirle fonetik dillerde alfabedeki harflerle dildeki sesler arasında birebir eşleşme vardır. Daha açık bir ifadeyle, yazıda bir harf dildeki sadece bir sesi ifade eder ve dildeki bir ses yazıda sadece bir harfle ifade edilir. Osmanlıca'nın fonetik olmamasının iki ana sebebi (i) ünlülerin çoğunlukla yazılmaması ve (ii) bir harfin birden fazla sese karşılık gelebilmesidir. Osmanlıca fonetik olmadığı için *şeker, şükür* örneğinde olduğu gibi farklı kelimeler aynı yazılışa sahip (homographs) olabilmektedir. Bundan dolayı Osmanlıca metinler okunurken bu tür kelimelerde belirsizlik (ambiguity) ortaya çıkmakta ve bu durum hem Osmanlıca okumayı hem de alfabe çevrimini güçleştirmektedir. Bu yönden ele alındığında, alfabe çevirisini (harf çevrimi ve yazı çevrimi) *kaynak dilde okuma* ve hedef dilde yazma şeklinde iki temel adımda ifa edebiliriz.

Osmanlıca bir metnin Arap tabanlı Osmanlı alfabesinden Latin tabanlı Türk alfabesine bilgisayarlı aktarım işlemi temelde kelime düzeyinde yapılır. Bilgisayarlı harf çevirisi karmaşık bir işlem olduğundan bildiğimiz kadarıyla şimdiye kadar kullanıma açık web tabanlı uygulaması olan sadece bir adet prototip çalışma mevcuttur. Ortografik alfabe çevirisi alt adımları aşağıda verilen dört adımda gerçekleştirilir:

1. Ön işlemler
    a. Metin dönüştürme ve temizleme
    b. Metin normalizasyonu
    c. Cümle ve kelime ayrıştırma
    d. Varlık ismi tanıma
2. *Ortografik Alfabe Çevirisi*
    a. Osmanlıca'da gövdeleme
    b. Osmanlıca-Türkçe sözlüklü aktarım
    c. Türkçede bitiştirme
3. Hata Düzeltme

   a. *Kelime Bölümleme*

   b. *Yazım Düzeltme*

   c. *Seslendirme*

4. Son işlemler

   a. *Kelime tahmini*

   b. *İsim Tamlaması ve Birleşik kelimeler*

   c. Varlık ismi dönüşümü

Bu alt adımlar arasında en önemlileri sırasıyla (i) Ortografik Alfabe Çevirisi (Transliteration), (ii) Kelime Bölümleme (Word segmentation), (iii) Yazım Düzeltme (Spelling correction), (iv) Seslendirme (Vowelization), (v) Kelime tahmini (Word prediction) ve (vi) İsim Tamlaması (Noun phrases) ve Birleşik kelimeler (Compund words) adımlarıdır.

2.2.1. Osmanlı Türk Alfabeleri

Osmanlıca 13.yy başından 20.yy'a kadar kullanılan Türkçe'nin yazı dildir. Osmanlıca alfabesi 28 harfli Arap alfabesinin genişletilmiş bir çeşididir. Arap alfabesinden 28 harf (ص, ض, ط, ظ ا, ب, ت, ث, ج, خ, ح, د, ذ, ر, ز, س, ش,,ع, غ, ف, ق, ك, ل, م, ن, و, ه, ى) Farsça alfabesinden 4 harf (گ ز چ پ) ve Türkçe alfabesinden 3 harf (ه ڭ ء) alınarak oluşturulmuştu. Osmanlı alfabesi 35 harften oluşur.

Rakamlar hariç olmak üzere harfler sağdan sola ve bitişik yazılır. Sadece 8 harf (ادذرزژوه (ه harfi okutucu olduğunda birleşmez)) kendinden sonra gelen harfle bitişmez. Bitişmeyen (non-joiner) harflerin münferit ve kelime sonunda, bitişen (joiner) harflerin ise münferit, kelime başı, ortası ve sonunda farklı yazım şekilleri vardır. Tablo 2.3'te gösterilmiştir. Bitişmeyen harfler kelime içi boşluğa (zero width non-joiner) sebep olur. Bundan dolayı OCR ve alfabe çevirisi sırasında kelime içi boşluk ile kelime arası boşluk birbirine karıştırıldığı zaman, ortaya kelime bölümleme (word segmentation) problemi çıkar. Bu problem alfabe çevirisi sırasında sıkça karşılaşılan bir problemdir.

Tablo 2.3: Osmanlıca harflerin yazımı

Münferit	Başta	Ortada	Sonda	Bitişik
ب	بـ	ـبـ	ـب	ببب
ص	صـ	ـصـ	ـص	صصص
ه	هـ	ـهـ	ـه	هههه
غ	غـ	ـغـ	ـغ	غغغ
ط	طـ	ـطـ	ـط	طططط
ر	ر	ـر	ـر	ررر

4 tane okutucu harf (ا ی ه و) bulunmaktadır. Bu dört harf Türkçe'deki 8 ünlü sese karşılık gelmektedir. Şekil 2.5'te okutucu harfler verilmiştir.

Türkçe	Osmanlıca		
	Başta	Ortada	Sonda
a	آ	ا	اه
e	اأ	ه	ه
ı	ای	ی	ی
i	ای	ی	ی
o	او	و	و
ö	او	و	و
u	او	و	و
ü	او	و	و

Şekil 2.5: Osmanlıca'da okutucu harfler

Osmanlıca; Arapça ve Farsça'dan pek çok kelime almıştır. Bu kelimelerin orijinal yapıları muhafaza edilmiştir. Arapça ve Farsça kelimelerin asli yapısı her durumda muhafaza edilir [46]. Arapça kelimeler de kısa sesler terkedilirken uzun sesler terk edilmez. Farsça'da ise uzun sesler gösterilirken kısa sesler bazen terk edilebilir. Bazı örnekler Tablo 2.4'te verilmiştir. Osmanlıca'da, Arapça ve Farsça asıllı kelimelerde sesli harflerin yazılmaması alfabe çevirisi için problem oluşturmaktadır.

Tablo 2.4: Osmanlıca Arapça ve Farsça kelime örnekleri

Orijinal	Osmanlıca	Türkçe	Açıklama
Arapça	شكر	şükür	ü sesleri atlanmış
Arapça	مبارك	mübarek	ü ve e sesleri atlanmış
Farsça	ابدست	abdest	e sesi atlanmış
Farsça	اسكر	asker	e sesi atlanmış

Osmanlıca'da Türkçe asıllı kelimelerde bazen okutucu (sesli) harfler terk edilmiştir. Bu durum eklerde daha sık görülmektedir. Bazı örnekler Tablo 2.5'te verilmiştir.

Tablo 2.5: Osmanlıca Türkçe kelime örnekleri

Osmanlıca	Türkçe	Açıklama
بابا	baba	tüm sesli harfler var
كوچوك	küçük	tüm sesli harfler var
صقال	sakal	a sesi atlanmış
تمل	temel	e sesleri atlanmış
لر ملی ر	-ler , -meli -er	sesli harfler atlanmış

Osmanlıca sessiz harfleri bir veya birden fazla karşılığı olabilir. Sessiz harfler ile ilgili bazı durumları şöyle sıralayabiliriz;

ء ع ظ ض ح ث sadece Arapça asıllı kelimelerde

ژ Farsça veya Fransızca asıllı kelimelerde

خ ذ arapça ve farsça kelimelerde

ڭ Türkçe kelimelerde bulunur.

Ayrıca sessiz harfleri kalın ve ince sesli olarak gruplandırılabilir.

Türk alfabesi Latin alfabesinin genişletilmiş halidir. Türk alfabesi 29 harften oluşmaktadır. 8 ünlü harf (a, e, ı, i, o, ö, u, ü) ve 21 ünsüz harften (b, c, ç, d, f, g, ğ, h, j, k, l, m, n, p, r, s, ş, t, v, y, z) oluşur. Sesli harfleri kalın (a, ı, o, u) ve ince (e, i, ö, ü) olmak üzere ikiye ayrılır. Soldan sağa doğru yazılmaktadır. Boşluk (space) kelime arasında boşluk olarak kullanılır. Büyük küçük harf kullanılır.

Osmanlıca – Türkçe harf eşleşmeleri Tablo 2.2'te gösterilmiştir. Görüldüğü üzere Osmanlıca'da bir harfin karşılığı Türkçe'de birden fazla harf olabilir. Bu harfler ile ilgili bazı durumları aşağıda listelenmiştir.

1. Ayın (ع) harfi: Arapça asıllı harflerde kullanılır. Türkçe asıllı kelimelerde ses karşılığı yoktur. Sadece kesme işareti olarak kullanılır. Fakat Osmanlıca kelimedeki yerine göre farklı sesleri temsil edebilir. Başta gelmesi ünlü "e" sesi hariç diğer tüm ünlü sesleri temsil edebilir. Kelime ortasında gelmesi bazen bir sesli harf olarak okunurken bazen atlanan bir ses olabilir. Bazı örnekler Tablo 2.6'da verilmiştir.
2. Hemze (ء) harfi: Genellikle Arapça asıllı kelimelerde kullanılır. Bazen Türkçe'de okumada kolaylık olması için kullanılabilir. Genellikle (ا،ى،و) harfleri ile birlikte kullanılır. Elif (أ) harfi ile kullanıldığı zaman e sesi (أمل → emel) ile okunur. Vav (ؤ) harfi olduğu zaman ö,ü (ؤلوم → ölüm, مؤمن → mümin) sesi ile okunur. Ye (ئ) harfi ile olduğu zaman i (مائل → mail) sesi ile okunur. Bazen de kesme işareti olabilir. Örneğin مسئله → mes'ele şeklinde okunur.
3. Kef (ك), Gef (گ), Nef (ڭ) harfleri: Türkçe'de ك → k, گ→ g, ڭ → n sesine karşılık gelir. Bazı durumlarda ك harfi Türkçe'de k, g, ğ, n seslerine karşılık gelebilir. Tablo 2.7'de bazı örnekler gösterilmiştir.
4. He (ه) harfi: Osmanlıca okutucu he harfi Türkçe'de e sesine bazen a sesine karşılık gelir. Osmanlıca okutucu he harfi kapalı hecelerde ve ilk hecelerde terkedilebilir. Bazı örnekler Tablo 2.8'de verilmiştir.

Tablo 2.6: Ayın harfi için örnekler

Konumu	**Osmanlıca**	**Türkçe**	**Açıklama**
Başta	عارف عالم عدد علم	arif alim aded ilim	Sesli harf gibi okunur.
Ortada	معلوم اعدام شعور ساعت شاعر	malum idam şuur saat şair	Sesli harf gibi okunur.
Sonda	ضايع منع	zayi' men'	Kesme işareti

Tablo 2.7: Kef, gef, nef örnekleri

Harf	Osmanlıca	Türkçe	Açıklama
ك	كمال	kemal	k sesi ile
ك	كولكه	gölge	g sesi ile
ك	اوغلك	oğlun	n sesi ile
ك	كلديكى	geldiği	g ve ğ sesi ile
گ	گل	gül	g sesi ile
ك	اغا بك	ağabey (beğ)	y sesi
ڭ	اوڭجو	öncü	n sesi ile

Tablo 2.8: Okutucu he harfi örnekleri

Konumu	Osmanlıca	Türkçe
Başta	او	ev
	ال	el
	أركك	erkek
Ortada	تمل	temel
	كلەبك	kelebek
	بسين	besin
Sonda	سرچه	serçe
	سويله	söyle
	فيرچه	fırça

Osmanlıca-Türkçe alfabe çevirisi görüldüğü üzere birebir (1↔1), bire çok (1→N) ve çoka çok (N↔N) eşleşme vardır. Yukarıdaki durumlarda göz önüne alındığında Osmanlıca – Türkçe alfabe çevirisi basit değildir. Ortografik alfabe çevirisi, kelime bölütleme, seslendirme, yazım düzeltme, dil modeli ve tamlama işlemlerini barındıran bütüncül bir yaklaşımla Osmanlıca-Türkçe alfabe çevirisi yapılmıştır.

2.2.2. Benzer Çalışmalar

Makine çevirisi (Machine Translation); kaynak dilden hedef dile bilgisayar destekli çeviri işlemin gerçekleştirilmesi olarak tanımlanabilir. Makine çevirisinde içeriğin anlamı hedef dile aktarılması amaçlanır. Kaynak dildeki kelime veya terimler birebire çevrilmeyebilir. Makine Alfabe çevirisi (Machine Transliteration), kaynak alfabedeki bir metnin hedef alfabeye aktarımı olarak tanımlanabilir. Bu aktarımda amaç hedef dilde yazma veya hedef dilde okuma olabilir.

Makine alfabe çevirisi uygulamaların çoğu, İngilizce'den Arapça, Korece, Çince ve Japonca gibi Roman olmayan alfabelerle yazılmış diller arasındaki harf çevirisine odaklanır. Roman olmayan diller arasındaki harf çevirisi [47] çalışmada yedi Avrupa dili bir dereceye kadar incelenmiştir. [48]'de kapsamlı bir makine çevirisi üzerine temel çalışmaları ayrıntılı olarak incelemektedir.

Osmanlıca – Türkçe Alfabe çevirisi alanında yapılan bazı çalışmalar mevcuttur. Sözlük tabanlı bire bir çeviriye Rahmi Tura tarafında yapılan çalışma örnek verilebilir. Bu çalışma sözlükler yardımıyla Osmanlıca - Türkçe çevirisi yapılmıştır. Bu çalışma kişisel amaç için geliştirilmiştir.

Osmanlıca Türkçe Alfabe çevirisi alanında yapılan diğer çalışmaya [2] örnek verilebilir. Doğal Dil İşleme teknikleri kullanarak bir çeviri sistemi önerilmiştir. Önerilen alfabe çevirisi Morfolojik analiz, sözlük araması, morfolojik sentez, kelime bölütleme, yazım düzeltme, seslendirme, dil modelleri ve tamlama olmak üzere birçok işlem yapılmaktadır. Morfolojik analiz ve sözlük yardımıyla Osmanlıca kelimenin kök ve ekleri bulunmaktadır. Morfolojik sentezde ortografik kurallara göre kelime türetilmektedir. Osmanlıca kelime kök bulunmazsa sırasıyla kelime bölütleme, yazım düzeltme ve seslendirme aşamaları ile kelime türetilmektedir. Osmanlıca metin çevirisi tamamlandıktan sonra dil modelleri (n-gram) Osmanlıca bir kelime için Türkçe'de birden fazla çözüm varsa bu çözümler sıralanmaktadır. Son olarak metinde tamlamalar geçiyorsa tamlamalar düzetilip çıktı oluşturulmaktadır. Sözlük olarak Kamusi Türki sözlüğü kullanılmıştır. Bu sözlük yaklaşık 30 bin kelimeden oluşmaktadır.

Bu alanda diğer bir çalışma da Dervaze tarafından yapılmıştır. Morfolojik çeviri aracı tanıtılmıştır. Morfolojik çeviri aracı Türkçe'den Osmanlıca'ya veya Osmanlıca'dan Türkçe'ye çeviri yapmaktadır. 60 bin çekimli kelimelik veri üzerinde test edilmiştir. Başarı oranı %83 olduğu belirtilmiştir.

Diğer bir çalışma da [49] örnek verilebilir. Kural Tabanlı Osmanlıca Türkçe çeviri sistemi tanıtılmıştır. Bu çalışmada kısmı morfolojik analiz yaparak Osmanlıca kelimenin köküne ulaşılmaktadır. Sözlükler yardımıyla kök ve ekler bulunduktan sonra Türkçe'ye çevrilmektedir. Modern Türkçe ve Arapça-Farsça kelimeleri barındıran iki tür sözlük kullanılmıştır.

Diğer bir çalışma da Transkribus örnek verilebilir. Herhangi bir dilde el yazısı tanıyan (Handwriting Text Recognition – HTR) motor geliştirilmiştir. Yapay sinir tabanlı olarak

çalışmaktadır. Veri seti hazırladıktan sonra Osmanlıca için eğitim yapılmıştır. Osmanlıca'dan (Arap alfabesinde) Türkçe'ye (Latin alfabesi) çeviri yapılmıştır.

Osmanlıca Türkçe alfabe çevirisi birçok problemi bünyesine barındıran bir çalışma alanıdır. Daha önce Doç.Dr. Atakan Kurt Danışmanlığında yapılan çalışma esas alınmıştır [2]. Bu çalışma ortografik alfabe çevirisi, kelime bölütleme, yazım düzeltme, seslendirme, kelime tahmini (n-grams), tamlamalar ve birleşik kelimeler olmak üzere altı aşamadan oluşmaktadır.

2.2.3. Önişlemler

2.2.3.1.Metin dönüştürme ve temizleme

Bu adımda kaynak metnin saklandığı dosyanın türünün düz metin (plain text) formatına dönüştürülmesi, dosya karakter kodlamasının UTF-8 gibi bir standart bir kodlamaya dönüştürülmesi, dönüşüm sonrası bozulan veya kaybolan harflerin düzeltilmesi, kaynak metinden Osmanlıca haricindeki bölümlerin, karakterlerin, işaretlerin (markup) temizliği ve gerekliyse metnin yeniden düzenlemesi yapılır. Bu adımdaki işlemlerin büyük çoğunluğu programla otomatik olarak yapılabilmektedir.

2.2.3.2.Metin Normalizasyonu

Osmanlı alfabesindeki metin, kullanılan program, kullanılan klavye, kullanıcı hatası gibi farklı sebeplerden dolayı standart Osmanlı alfabesinde olmayan harfler, rakamlar, noktalama işaretleri içerebilir. Bu standart dışı karakterlerin alfabedeki doğru karakterlerle değiştirilerek düzeltilmesine metin normalizasyonu denir. Metindeki bir harf bilgisayarda yazılırken görünüşü aynı olmasına rağmen kodu farklı bir karakterle yazılırsa, çevirinin sonraki adımlarında probleme sebep olabilir. Bu yüzden metin bilgisayarla kontrol edilerek standart karakterlere dönüştürülür. Metin içindeki her çeşit boşluğun (whitespace) standart boşluk karakteriyle bu adımda değiştirilir. Bu adım programla otomatik olarak yapılır.

2.2.3.3.Cümle ve kelime ayrıştırma

Normalize edilmiş metin önce cümlelere sonra kelimelere ayrıştırılır. Metnin cümlelere ayrıştırmasında nokta, soru işareti, iki nokta, ünlem işareti gibi noktalama işaretleri, kelimelere ayrılmasında ise boşluk (space) karakteri kullanılır. Eğer metinde OCR ya da kullanıcı kaynaklı *bölünmüş* veya *bitişmiş* kelimeler varsa, o hatalar bu adımda değil, *kelime bölümleme* adımında ele alınıp düzeltilir. Bu adım programla otomatik olarak yapılır.

2.2.3.4.Varlık ismi tanıma

Metin içindeki sayı, tarih, isim, noktalama işareti, formül, kısaltma, vb. ögelere varlık, metindeki bu ögeleri bulmaya varlık ismi tanıma denir. Varlıklar hedef ve kaynak dilde farklı biçim veya şekillerde yazılabilecekleri için özel olarak ele alınmaları gerekmektedir. Tarih yazarken kullanılan ayıraç, gün, ay ve yıl ögelerinin sırası, sayılar yazılırken ondalık nokta işareti ile binlik ayıraç işaretleri, özel isimlerde ilk harfin büyük yazılması, vb. durumlar dilden dile farklılık gösterebilmektedir. Metindeki varlıkların çoğu gövdeleme adımından önce bulunabilir. Bu adım programla otomatik olarak yapılabilir.

2.2.4. Ortografik Alfabe Çevirisi

Osmanlıca-Türkçe alfabe çevirisi üç adımda gerçekleşir: Osmanlıca kelime gövdelenerek gövde ve ek katarına ayrılır; Osmanlıca gövde ve ek katarı çeviri sözlükleri kullanarak Türkçe gövde ve ek katarlarına ulaşılır, Türkçe gövde ve ek katarları bitiştirilerek Türkçe kelime elde edilir. Tablo 2.9'de bazı örnekler verilmiştir.

Osmanlıca gövdeleme algoritması: Gövdeleme kelime sonundan başlayıp, harflerin teker teker silinerek, elde edilen kelimeciğin Osmanlıca gövde sözlüğünde aranması, gövde bulunduğunda kelimenin geri kalanının (ek katarı olduğu kabul edilerek) Osmanlıca ek sözlüğünde aranıp bulunması ile tamamlanır. Bu yöntemle algoritma verilen kelimenin tüm olası gövde ve ek katarlarını bulur.

Bu adım aslında göründüğünden daha karmaşıktır. Şöyle ki bir kelime Türkçe'de morfolojik analiz yapılarak birçok farklı şekilde gövdelenebilir. Örneğin "alınmış" kelimesinin 4 farklı gövdesi 7 farklı morfolojik çözümü vardır [50]. Bu yüzden gövdeleme bir çeşit yüzeysel morfolojik çözümleme gerektirmektedir ve birçok çok çözüm vermektedir. Gövdelemeden sonra Osmanlıca-Türkçe Aktarım sözlükleriyle Osmanlıca gövdelerin ve eklerin Türkçe karşılıkları bulunur. Tablo 2.9'de bazı örnek gösterilmiştir. Bu aşamada da hem Osmanlıca kelimenin hem de Osmanlıca eklerin sözlükte birden fazla Türkçe karşılığı olabilmektedir. Bu da kelime çözümlerinin sayısını ciddi oranda arttırmaktadır. Bunun sayısal artışın sebebi Osmanlıca fonetik olmadığından ve Osmanlıca harfler birden fazla sesi ifade edebildiklerinden dolayıdır. Örneğin Osmanlıca'da شكر → {şeker, şükür} veya كل→ {gil, gül, kül, gel, kel, kil} kelimelerinin yazılışı aynıdır.

Kelime gövdelendikten sonra ek katarlarının çözümlenmesi için sözlük tabanlı ve özyinelemeli bir ek bölümleme algoritmasıyla yapılmaktadır.

Son adımda Osmanlıca kelimenin Türkçe karşılıkları, bulunan tüm Türkçe gövdelerle tüm Türkçe ek katarlarının Türkçe ortografik kurallar kullanarak birleştirilerek kelimenin Türkçe yüzey (surface) biçimlerinin elde edilmesini ve kurala aykırı geçersiz yüzeylerin elimine edilmesiyle tamamlanır.

Tablo 2.9: Osmanlıca-Türkçe Harfçevrim Örnekleri

Kelime	**Gövde+Ek katarı**	*Osmanlıca-Türkçe Kelime ve Ek Sözlükleri*	**Gövde+Ek katarı**	**Kelime**
باشلا گلمك بردر عثمانليلاشديرمق چركينلكلري الگزده كي لياقتسزلگمزي	باش+ لا گل+ مك بر+ در عثمان + ليلاشديرمق چركين +لكلري ال +گزده كي لياقت + سزلگمزي		baş+lA gül/gel+mAk bir+DHr osman+lH+lAş+DHr+mAk çirkin+lHk+lAr+sH el+HnHz+DAki liyakat+sHz+lHğH+mHzH	başla gülmek, gelmek birdir osmanlılaştırmak çirkinlikleri elinizdeki liyakatsizliğimizi
عثمانليلاشديرامايابيله جكلريمزدنمشسكزجه سنه			osmanlılaştıramayabileceklerimizdenmişsinizcesine	

2.2.5. Kelime Bölümleme

Osmanlıca'da kelimeler arasındaki boşluk Latindeki gibi net değildir. Osmanlıca'da bitişmeyen karakterlerden dolayı kelime içi boşlukla (zero length non space, intra-word space) kelime arası boşluklar (inter-word space) birbirine karışabilmekte ve bu durum kelime bölümlemeyi zorlaştırmaktadır. Şekil 2.6'da dokümanların OCR yapılması sonucu ortaya çıkan bazı gerçek kelime bölümleme örnekleri verilmiştir. Özellikle okutucu he harfinden sonra bu problem ortaya çıkmaktadır. Bu örneklerde dikkat çeken şu durumlar söz konusudur:

Kelime genellikle iki bazen üç parçaya bölünmektedir. Bir parça geçerli bir ek veya kelime olabilir. Parça bir kelime, gövde ya da ek olduğunda, kelime gövdeleme hiçbir zaman düzeltilmemiş olabilir ya da yanlış bir düzeltme de söz konusu olabilir. Bazı örnekler Şekil 2.7'de verilmiştir.

Osmanlıca	Çıktı	Doğrusu	Bölümleme Türü
سويلەرسه م	söylerse m	söylersem	kelime+ek → kelime
حكايه لر	hikaye ler	hikayeler	kelime+ek → kelime
محزو نلغنى	mahzu nluğunu	mahzunluğunu	parça+parça → kelime
ايتد يگي	etti giy	ettiği	kelime+kelime → kelime
بولوندير يبور	bulundur yiyor	bulunduruyor	kelime+kelime → kelime

Şekil 2.6: Kelime bölütleme problemleri

Osmanlıca Kelime	Türkçe	Osmanlıca Kelime	Türkçe
سويلەرسه م	söylerse m	چله لرك	çile lerin
حكايه لر	hikaye ler	گوره ليم	göre lim
محزو نلغنى	mahzu nluğunu	چرچوه لرى	çerçeve leri
چشمه لرى	çeşme leri	پارچه لانمش	parça lanmış
نه ك	ne n	پرده لرك	perde lerin
سنه لرجه	sene lerce	جومبه لرك	cumba ların
ايده مزسكز	ede mezsiniz	بكله يبشلرى	bekle yişleri
ديسه م	dese m	اوطه لرينك	oda larının
صورسه م	sorsa m		

Şekil 2.7: Kelime bölütleme örnekleri

Bölümlenmiş kelime parçaları sırayla önceki kelimeyle veya sonraki kelimeyle birleştirilerek düzeltilmeye çalışılabilir. Düzeltilme mümkün değilse, çözüm bir sonraki aşamada bulunmaya çalışılabilir.

2.2.6. Yazım Düzeltme

Metindeki hatalı kelimeleri düzeltme işlemidir. Yazım kontrolü ve yazım düzeltme olmak üzere iki adımdan oluşmaktadır. Yazım kontrolünde kelimelerin imla kurallarına göre yazılıp yazılmadığı kontrol edilir. Eğer hata var ise yazım düzeltme ile hatalar giderilir. Osmanlıca metinlerde OCR ya da kullanıcı kaynaklı yazım hataları olabilmektedir. Bazı hatalı kelimeler

Şekil 2.8’de gösterilmiştir. Yazım düzeltme aşamasında yazım hataları sözlük yardımıyla düzeltilir.

Hatalı	Doğru	Yazım düzeltme	Türkçe
نچاق	بچاق	چاق	bıçak
لوردقجه	اوردقجه	طوردقجه	(v)urdukça
مخموعه‌سندن	مجموعه‌سندن	مخموعه‌سندن	mecmuasından
اقیزم	قیزم	قیزم	kızım
بربنك	بربنك	برینك	birinin
مهاجزتلری	مهاجرتلری	مهاجرتلری	mühaceretleri
وفت	وقت	وقت	vakit
مثقسم	منقسم	منقسم	münkasım
ولسه	اولسه	اولسه	olsa
خیسالات	خیالات	خیالات	hayalat
اشکمنجه	اشکنجه	اشکنجه	işkence

Şekil 2.8: Yazım düzeltme örnekleri

2.2.7. Seslendirme

Osmanlıca her harfin Türkçe’de bir veya birden fazla ses karşılığı vardır. Osmanlıca Türkçe harf eşleşmeleri Tablo 2.10’da gösterilmiştir. Seslendirme aşamasında özellikle sözlük dışı kelimeler çözümlenir. Seslendirme aşamasında Osmanlıca bir kelimenin olası tüm çözümleri seslendirme algoritmasıyla bulunur. Bu aşamada bulunan kelimeler sözlük araması ile eşleşen veya benzer çözümler kabul edilip işleme devam edilir. Seslendirme algoritması 4 adımdan oluşmaktadır.

1. Latin Harfleri Üret
2. Meta kümeleri çözümle
3. Ünsüzler arasında ünlü yerleştirme
4. Üretilen kelimeler sözlükte arayıp eşleşen ya da benzer kelimeleri getir.

Adım 1’de Osmanlıca - Türkçe eşleşmeleri yapılmaktadır. Adım 2’de Osmanlıca bir harfin Türkçe’de birden fazla karşılığı olan harfler üretilmektedir. Osmanlıca kelimelerde okutucu harfler gösterilmeyebilir. Okutucu harf kullanılmayan durumları şöyle sıralayabiliriz;

1. Arapça ve Farsçadan alınan ve asıllarında okutucu kullanılmayan kelimeler
2. Okutucusuz yazımı yaygınlaşmış kelimeler
3. Geleneksel olarak okutucular terk edildiği ekler
4. Okutucu harfler bazı durumlarda terk edildiği kelimeler

Bu gibi durumlardan dolayı Adım 3’te ünsüzler arasında ünlü yerleştirme yapılarak olası kelimeler üretilmektedir. Adım 4’te ise oluşan tüm kelimeler sözlükte aranır. Eşleşen ya da benzerlik oranı yüksek olan kelimeler çözüm olarak kabul edilmektedir. Bazı örnekler Tablo 2.11’da verilmiştir.

Tablo 2.10: Osmanlıca Türkçe harf eşleşmesi

Türkçe	Osmanlıca	Türkçe	Osmanlıca	Türkçe	Osmanlıca
A	ع ◌ه ◌ا	I	ع ◌ی	R	ر
B	ب	İ	◌ی	S	ث ص س
C	ج	J	ژ	Ş	ش
Ç	چ	K	ق ك	T	د ت ط
D	ط ض د	L	ل	U	ع ◌و
E	ه ◌ا	M	م	Ü	◌و
F	ف	N	ڭ ن	V	و
G	ك غ گ	O	◌و	Y	ی
Ğ	ك غ گ	Ö	◌و	Z	ض ظ ذ ز
H	خ ح ه	P	پ	‘	ء
				’	ع

Tablo 2.11: Seslendirme algoritması aşama örnekleri

Kelime	Adım 1	Adım 2	Adım 3	Adım 4
وار	var	var	var	var
نهال	nhal	nhal	nehal, nahal, nühal, nuhal, nihal, nıhal, nohal, nöhal	nihal
خدمتكارم	h{d,t}mt{k,g}arm	hdmtkarm, hdmtgarm, htmtkarm, htmtgarm	… (çok sayıda seçenek olduğundan verilmedi)	hidmetkarım

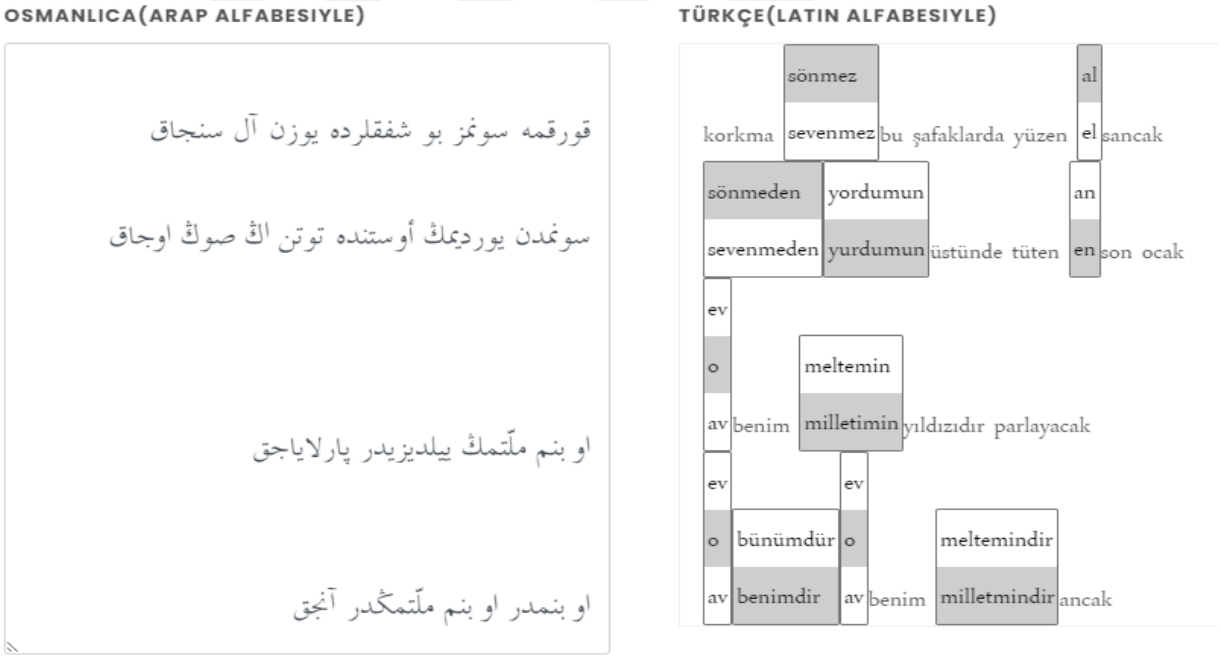
2.2.8. Kelime Tahmini

Osmanlıca bir metin Türkçe'ye çevrildiği zaman Osmanlıca bir kelime için birden fazla Türkçe çözüm olabilmektedir. Şekil 2.9'da bazı örnekler gösterilmiştir. Osmanlıca bir kelime için birden fazla çözüm varsa n-gram ile bu çözümler sıralanır.

N-grams; Hesaplamalı dilbilimde bir n-gram, belirli bir metin veya konuşma örneğinden alınan n öğeden oluşan bir dizidir. Öğeler uygulamaya göre harf, hece, kelime vb. olabilir. Bu çalışmada n-gram hesaplamaları kelime bazında yapılmıştır. unigram; 1 kelimelik grupları, bigram 2 kelimelik grupları ve trigram 3 kelimelik grupları temsil etmektedir. n-gramlar tipik olarak bir metin derleminden hesaplanır. Osmanlıca bir metin Türkçe'ye çevrildiği zaman Osmanlıca bir kelime için birden fazla Türkçe çözüm olabilmektedir. Şekil 2.4'te geliştirdiğimiz uygulama (osmanlica.com) ile örnek bir metnin Osmanlıca-Türkçe alfabe çevirisi yapılmıştır. Şekilde görüldüğü gibi birden fazla çözüm olduğu durumda n-grams ile olasılığı en yüksek çözüm en başta gösterilmektedir.

Osmanlıca	Türkçe	Unigram	Bigram
كلمەنك	gelmenin, kelimenin	gelmenin	(bir) kelimenin
اولنور	olunur, evlenir	olunur	(ıtlak) olunur
اسم	asım, isim, esim	asım	(için) isim
مرور	mürur, mürver	olan	(dahil) olan
ايسه	ise, aysa, ısa	ise	(olur) ise
لكن	lakin, leğen	lakin	(denilir) lakin
فرق	fark, ferak, fırak	fark	(bir) fark
كورنلر	görenler, gevirenler, körenler	görenler	(dinini) görenler

Şekil 2.9: Kelime tahmini örnekleri



Şekil 2.10: Osmanlıca Türkçe çeviri n-gram sıralaması

Dil modelleri (n-grams) için bir derleme ihtiyaç vardır. Üstünde çalıştığımız Osmanlıca eserlerin büyük bir kısmı 18. 19. ve 20. yüzyıllara ait olduğundan, bu yüzyıllarda yazılmış eserlerden bir derlem oluşturulmuştur. Farklı alanlardan (edebi, dini vb.) 100'e yakın eser

seçilmiştir. Bazı eserler Tablo 2.12'de verilmiştir. Derlem 470046 cümle 4941189 kelime 567984 tekil kelime oluşmaktadır.

Tablo 2.12: Derlem için bazı eserler

Eser	**Yazar**	**Eser**	**Yazar**
Akif Bey, Cezmi, Vatan veyahut Silistre	Namık Kemal	Tılsımlar, Lemalar, Sözler, Şualar, Mektubat	Bediüzzaman Said Nursi
Araba Sevdası	Recaizade Mahmut Ekrem	Felatun Bey ile Rakim Bey	Ahmet Mithat Efendi
Taaşşuk-ı Talat ve Fitnat	Şemsettin Sami	Karabibik	Nabizade Nazım
Türkçüğün Esasları	Ziya Gökalp	Hak dini Kuran Dili(9 cilt)	Elmalı Hamdi Yazır Efendi
Ruhul Beyan Tefsiri(23 Cilt)	İsmâil Hakkı Bursevî	Sıratul Müstakim (4 cilt)	Gazete (Mehmet Akif)

Derlem önce bir ayrıştırıcıyla cümlelere ayrılmıştır. Peşi sıra cümledeki her kelimenin gövdesi bulunmuştur. Son olarak derlemden n-gram (unigram, bigram ve trigram) hesaplanır. Vatan yahut Silistre (Namık Kemal) eserinde alınan bir paragraf Tablo 2.13'da gösterilmiştir.

Alfabe çevirisinde Osmanlıca metin Türkçe'ye çevrilirken bir kelime için birden fazla çözüm varsa bu aşama bu çözümler sıralanır.

Tablo 2.13: Dil modeli hesaplama aşamaları

Ham	*"gönlüme niçin bu kadar rikkat verdin? fikrimi niçin bu kadar açtın? sen de şimdi kızını görsen okuttuğuna pişman olurdun... benim gönlüm öyle büyük büyük hayata nasıl dayansın?"*
Cümle	*"gönlüme niçin bu kadar rikkat verdin?* *fikrimi niçin bu kadar açtın?* *sen de şimdi kızını görsen okuttuğuna pişman olurdun...* *benim gönlüm öyle büyük büyük hayata nasıl dayansın?"*
Gövdeleme	*gönl-üme niçin- bu- kadar- rikkat- ver-din* *fikr-imi niçin- bu- kadar- aç-tın* *sen- de- şimdi- kız-ını gör-sen oku-ttuğuna pişman- ol-urdun* *ben-im gönl-üm öyle- büyük- büyük- hayat-a nasıl- daya-nsın*
N-gram hesaplanması	*unigram, bigram, trigram hesaplanır. Bu işlem tüm derlem için yapılır. Hesaplama işlemi bittikten sonra dil modeli oluşturulmaktadır.*

2.2.9. Tamlamalar

Osmanlıca'da isim tamlamaları Arapça, Farsça, Türkçe ve hibrit isim tamlamaları olmak üzere dört farklı şekilde oluşturulur.

Arapça tamlamalar lam-ı tarif denilen ال (el-) takısıyla yapılır. ال takısı önüne geldiği harfin türüne bağlı olarak ses değişikliğine sebep olabilir: عبد الملك → abd elmelik → abdulmelik, عبد السلام → abd esselam → abdusselam. Arapça tamlamalar Osmanlıca'da ayrık yazılırken, Türkçe'de bitişik ve konuşma dilindeki gibi yazılır. Bu tamlamaların Türkçe'ye çevirisi tamlama sözlüğü kullanılarak yapılır.

Farsça isim tamlamalarında yardımcı sesler yazılmaz. Fakat bu tamlamalar Türkçe'ye çevrildiğinde, yardımcı sesler yazılır: دولت علیه عثمانییەنك → devlet aliyye osmaniyyenin → devlet-i aliyye-i osmaniyyenin, شاه ولایت → şah velayet → şah-ı velayet, دیوان همایون → divan hümayun → divan-ı hümayun.

Türkçe isim tamlamalarında tamlayan -in eki alır. Osmanlıca'da bu ekler yazıldığı için çeviride problem olmaz.

Hibrit isim tamlamaları farklı dillerden gelen kelimelerden oluşturulur. Bu tamlamalar hangi tür (Arapça, Farsça, Türkçe) tamlama yapısına sahip ise, algoritmada o şekilde ele alınır. İsim tamlamalarını Türkçe'ye aktarmak için Osmanlıca-Türkçe tamlamalar sözlüğü kullanılır.

Osmanlıca'da isim tamlamaları tespit etmek bazen oldukça zor olabilmektedir. Bazı durumlarda isim tamlamasını tespit için anlamsal bağlama bakmak gerekebilir. Örneğin bu şehir istanbul… (بو شهر استانبول) cümlesinde bu *şehr-i istanbul*'un tamlama olup olmadığı bağlamdan anlaşılır.

2.3. DİL ÇEVİRİSİ

Dil çevirisi (intra-language machine translation); Türk alfabesindeki Osmanlıca metnin bilgisayarlı çeviriyle Modern Türkçe'ye yani Günümüz Türkçe'sine çevirisidir. Çeviri sonucu ortaya Türk Alfabesinde Modern Türkçe metin çıkar. Dil çevirisi doğal dil işlemenin en önemli ve en zor problemlerinden bir tanesidir. Ama Osmanlıca-Türkçe çeviri, aynı dilin farklı özellikler taşıyan iki farklı zaman dilimindeki sürümleri arası çeviri olduğundan farklı diller arası çeviriye göre daha basittir. Fakat bu çeviri iki dil arasındaki söz dağarcığı, söz dizim ve anlamsal bazı farklılıklardan dolayı basit bir problem değildir. Dil çevirisinde kelime-grubu tabanlı birebir çeviri yaklaşımıyla ilerlenebileceği gibi, derin öğrenme kütüphaneleri kullanılarak cümle tabanlı makine çevirisi yaklaşımı da kullanılabilir. Çeviri öncesi (pre processing), sırası ve sonrasında (post processing) yüzeysel veya tam yapı bilimsel ve/veya söz dizimsel çözümleme, üretim, belirsizlik giderme (shallow/full morphological & grammatical parsing, generation, disambiguation), kelime anlamı belirsizliği giderme (word sense disambiguation), varlık ismi tanıma (named entity recognition) vb. işlemlerin yapılması gerekmektedir.

2.3.1. Osmanlıca – Türkçe Dil Çevirisi

Osmanlıca - Türkçe dil çevirisi için kelime grubu tabanlı basit bir çeviri sistemi sunulmuştur. Bu yöntem dil içi çeviri içi kullanılabilecek en basit yöntemlerden bir tanesidir. Yöntem kelime öbeği tabanlı ve mekanik olarak işleyen ve bir cümleyi öbek öbek soldan sağa doğru tarayarak Osmanlıca'dan Modern Türkçe'ye aktaran bir yöntemdir. Bu yöntemin basit olmasına karşın

çok karmaşık olmayan cümleler için etkin bir yöntem olduğu düşünülmektedir ve elde edilen sonuçların daha ileri yöntemlerin geliştirilmesinde ve performans analizinde bir temel (baseline) olarak ta kullanılabileceği hedeflenmektedir.

Çeviride kullanılan bu yöntem kısaca şu şekilde özetlenebilir: Bir kelime öbeği iki ya da daha fazla kelimeden oluşan ve bir arada anlamlı bir birim oluşturan kelime gruplarıdır. Bu yöntemde kelime grubu veya öbeği olarak isimlendirilen yapılar, söz dizimsel çözümlemede kullanılan kelime öbeklerine (constituent) benzerlik göstermektedir. Bu yüzden kelime öbeklerinin belirlenmesinde söz dizimsel çözümleme kullanılması önerilmekle birlikte, bu çalışmanın bir temel (baseline) çalışma olması ve zaman kısıtından dolayı söz dizimsel çözümleme yapmaya vakit olmaması bizi kelime gruplarını çeviri sözlüğü kullanarak belirlemeye yöneltmiştir. Bu yöntemde Şekil 2.5'de verilen sözde (pseudo) kodda görüldüğü gibi Osmanlıca cümledeki kelimeler gövdelendikten sonra cümle soldan sağa doğru sözlük kullanarak öbek öbek taranır. Tarama sırasında önce en uzun kelime öbeği sözcükte bulunmaya çalışılır, bulunamazsa uzundan kısaya doğru daha kısa kelime öbekleri sırayla sözlükte taranır. Bu şekilde 5 kelimelik kelime öbeğinden tek kelimelik kelime öbeğine doğru tarama yapılır. Sözlükte bir öbek bulununca öbeğin Türkçe karşılıkları kullanılarak Türkçe cümlenin gövdeleri bir dizide saklanır. Bir sonraki aşamada Osmanlıca kelime ekleri morfolojik sentezleme ile Türkçe gövdelere bitiştirilerek Osmanlıca cümleler oluşturulur. Osmanlıca bazı kelimelerin sözlükte Türkçe için birden fazla karşılığı ya da anlamı olan kelime olacağından Türkçe birden fazla cümle ortaya çıkacaktır. Çevirinin son ana adımı Türkçe dil modeli (n-grams) kullanarak birden fazla Türkçe karşılığı olan Osmanlıca kelimelerin Türkçe karşılıklarının doğru sıralanması amaçlanmaktadır.

```
function aktar(wordsOT) wordsTR
  input: wordsOT[]  // osmanlıca (OT) kelime listesi olarak bir cümle
  output: wordsTR[]  // turkce (TR) kelime listesi olarak bir cümle
begin
  string stemOT, suffixOT, stemsTR = [];
  // Osmanlıca kelimeleri gövdele (en uzun gövde + en kısa ek katarı)
  for(i=0; i < len(wordsOT); i++)
    (stemOT[i], suffixOT[i]) = morpParsing(wordsOT[i])
  // Osmanlıca öbekleri (en uzundan en kısaya doğru) sözlükte ara
  for(i=0; i < len(stemOT); i++)
    for(j=5; j > 0; j--){  // en uzun 5 kelimeli öbek olabilir
      stemsTR = wordLookUp(stemOT[i..i+j])
      if(stemsTR)  // öbek sözlükte bulundu
        i = i+j
        break
    }
    // Osmanlıca ekleri Türkçe gövdelere bitiştir
    for(k=0; k < len(stemsTR); k++)
      wordsTR[i+j] += morhGen(stemsTR[k],  suffixOT[i+j])
  }
end
```

Şekil 2.11: Dil çevirisi morfolojik kök ve ek bulma algoritması

Osmanlıca - Türkçe Dil Çevirisi Osmanlıca'da morfolojik çözümleme, çeviri sözlüğünde arama ve Türkçe'de morfolojik sentez olmak üzere üç adımdan oluşmaktadır. Osmanlıca kelime gövdelenerek gövde ve ek katarına ayrılır; Osmanlıca gövde çeviri sözlükleri kullanarak Türkçe gövdesine ulaşılır, Türkçe gövde ve ek katarları bitiştirilerek Türkçe kelime elde edilir.

2.3.2. Benzer Çalışmalar

Osmanlıca - Türkçe Dil çevirisi aynı dilin farklı zaman dilimlerindeki sürümleri arasında çeviri olarak karşımıza çıkmaktadır. Literatürde bu tür çeviri problemlerine benzer bir problem olan yeniden yazım (paraphrasing) problemidir. Yeniden yazım; bir ifadenin, cümlenin veya metnin benzer mana içeren alternatif ifadelerle yeniden düzenlenme işlemidir. Yeniden yazım alanında yapılmış çok fazla çalışma olmamakla birlikte derin öğrenme modelleri bu alanda da kullanılmaya başlanmıştır. Yeniden yazım probleminde bir modelin eğitilebilmesi için paralel veri kümelerine ihtiyaç vardır. Bu veri kümelerinin diğer çeviri veri kümelerinden ana farkı, çeviride iki dilin farklı olması, yeniden yazımda ise dillerin sadece zaman dilimi yönünden farklı olmasıdır. Türkçe yeniden yazım problemi üzerinde [51] bir çalışma yapılmıştır. Bu çalışmada farklı kaynaklardan (roman, gazete vb.) metinler toplanmıştır. Bu metinlerden dil

uzmanları ile birlikte paralel bir veri seti hazırlanmıştır. Ayrıca yine derlem oluşturma noktasında [52]'de bir çalışma yapmıştır. Bu çalışmada bir web crawler ile 2015 yılında bazı haber sitelerinde haber metinleri toplanmıştır. Microsoft Araştırma Yorumlama Derlemi (Microsoft Research Paraphrase Corpus - MSRPC) ve Twitter Yorumlama Derlemi (Twitter Paraphrase Corpus - TPC) metodolojileri kullanarak Türkçe bir paraphrase corpus oluşturulmuştur. Bu çalışmada hazırlanan veri setinde paralel cümlelerin anlamsal olarak mümkün olduğunca benzer olması sağlanmaya çalışılmıştır. Yeniden yazım problemine benzer bir problem olan metin sadeleştirme probleminde Türkçe kapsamında [53] eski Türkçe metinlerin modernleştirilerek Günümüz Türkçe'siyle yeniden oluşturulabilmesi için istatistiksel makine çevirisi yöntemini tatbik etmişlerdir. Öbek (kelime grubu, kelime öbeği) tabanlı istatistiksel makine çevirisi alanındaki öncü başarılı çalışmalardan olan [54] gerçekleştirdiği çalışmada kelime tabanlı modellerden ziyade öbek tabanlı yaklaşımların çeviri sistemlerinde daha iyi sonuçlar üretildiği sonucuna ulaşılmıştır. Bir yeniden yazım (paraphrasing) tekniği olarak dil içi makine çevirisi yöntemini kullanan [55], kaynak cümlelerin hedef dile göre yeniden yorumlanması ile makine çevirisi değerlendirmesinin iyileştirilmesine yönelik bir yöntem sunmuşlardır. Çalışmalarında hem kural tabanlı hem de öbek tabanlı istatistiksel makine çevirisi yöntemlerini denemişlerdir. Yine dil içi çeviri yöntemini uygulayarak metin üretme (paraphrase generation) probleminde çalışan [56] bu araştırmalarında metinlerin otomatik şekilde oluşturulması ve değerlendirilmesi için öbek tabanlı bir istatistiksel makine çevirisi sistemi kurgulamışlar ve değerlendirme sürecinde insan değerlendiricilerin yargılarıyla göreceli olarak benzer BLEU ölçüm değerlerine ulaşmışlardır. Diğer bir çalışma Doç Dr. Atakan KURT'un danışmanlığındaki İsmail ÖZSOY'un "*Türkçe İçin Yeniden Yazım Sistemi*" adlı yüksek lisans tezidir. Bu tezde birçok farklı yöntemle kurgulanabilecek yeniden yazım/yorumlama (paraphrasing) problemi, öbek tabanlı istatistiksel makine çevirisi yöntemi kullanılarak gerçeklenmeye çalışılmıştır. Oluşturulan sistem ile çeşitli çeviri deneyleri ve testleri yapılmıştır. Altyapısal düzeyde Moses istatistiksel makine çevirisi aracından yararlanılmış olup, test sonuçları BLEU ölçütüyle değerlendirilmiştir. Bu çalışmada çeviri sisteminin eğitim ve test aşamalarında kullanılmak üzere Türkçe için az bulunan kaynaklardan olan 2 adet Hizalanmış Türkçe Paralel Derlem oluşturulmuştur. Bunlardan birisi Hayvan Çiftliği romanının cumhuriyetin ilk dönemlerinde Osmanlıca Türkçe'sine yakın bir dil ile Türkçe'ye tercüme edilmiş sürümü, diğeri ise yakın zamanda Modern Türkçe'yle tercüme edilmiş yeni bir versiyonunun paralelleştirilmesiyle elde edilmiştir. Derin öğrenme tabanlı

Moses makine çevirisi framework'u kullanılarak çeviri modeli oluşturulmuş ve oluşturulan model ile test veri seti kullanılarak testler yapılmış ve sonuçlar değerlendirmiştir. Türkçe'nin eklemeli bir dil olması, eğitim veri setinin boyutunun küçük olması gibi problemlerden dolayı elle tutulur bir başarı elde edilmemiş olsa da bulgular bu konuda daha fazla araştırma ile daha iyi sonuçların elde edilebileceği yönünde olmuştur. Buna benzer bir diğer akademik çalışmada bir yüksek lisans tezi olarak, Nutuk'un Osmanlıca'ya yakın orijinal diliyle basımı ve Modern Türkçe'ye tercümesinden elde edilen hizalanmış paralel derlem kullanılarak yapılan dil-içi makine çeviri çalışmasıdır [53]. Yukarıdaki iki çalışmada önce paralel derlemle bir derin öğrenme modeli eğitilmekte ve peşi sıra test verisi kullanılarak modelin performansının ölçülmesine dayanmaktadır.

2.3.3. Osmanlıca Morfolojik Çözümleme

Morfolojik çözümlemede etkinlik bakımından tam çözümleme yerine (full morphological parsing) "ekten gövdeye ulaşma yöntemi" kullanarak gövde ve ek katarlarına ulaşılmıştır. Bu işlem sırasında kelimenin sonundan başına doğru harfler taranarak ve her adımda Osmanlıca gövde sözlüğü ve Osmanlıca ek katarı sözlüğü kullanılarak bir çeşit gövdeleme yapılmakta ve olası tüm gövde ve ek katarları elde edilebilmektedir. Bazı durumlarda bir kelime birden çok morfolojik çözümü olmaktadır. Örneğin "alınmış" kelimesinin 4 farklı gövdesi 7 farklı morfolojik çözümü vardır [50]. Bu durumda 3 farklı yol izlenebilir:

1. Çözümlerin (gövde + ek-katarı) tamamı doğru kabul edilir bir sonraki aşamaya (çeviri sözlüğü kullanılarak Türkçe karşılığa ulaşma) gönderilir.
2. Çözümler n-grams kullanılarak kendi içinde sıralanır ve olasılığı en yüksek çözüm doğru çözüm kabul edilir, diğerleri göz önüne alınmaz.
3. Ekten gövdeye ulaşma yöntemiyle ilk ulaşılan çözüm yani "en uzun gövde + en kısa ek katarı" doğru çözüm kabul edilir, sonraki çözümler göz önüne alınmaz.

Bu tezimizde 3. yöntem ile ilerledik. Gövdelemeden sonra Osmanlıca-Türkçe Aktarım sözlüğüyle Osmanlıca gövdelerin Türkçe karşılıkları bulunur.

2.3.4. Osmanlıca - Türkçe Aktarım Sözlüğü

Osmanlıca- Türkçe Dil çevirisi için kelime öbeği bazında bir çeviri işlemi uygulanmıştır. Çeviri için Kamus-i Türki isimli yaygın olarak bilinen kaynaklardan olan bir aktarım sözlüğü kullanılmıştır. Zaman kısıtından dolayı bu sözlükten sadece 10 bin kelime alınmış ve Günümüz

Türkçe'sindeki güncel karşılıkları manuel olarak sisteme girilmiştir. Birden fazla Türkçe karşılık olması durumunda en yaygın anlamdan en seyrek anlama olmak üzere Türkçe karşılıklar sıralanarak sözlüğe girilmiştir. Sözlük hazırlama kendi başına zaman alıcı, yorucu ve dikkat isteyen bir iştir. Aktarım sözlüğü bu aşamanın önemli bir birleşenini oluşturmaktadır. Sözlüğün seçimi, sözlüğü taratıp OCR yapmak, OCR hatalarını düzeltmek, Türkçe karşılıkları bulmak, sözlük formatına uygun olarak dosyaya girmek, kullanım ve anlam sıklığına göre karşılıkları sıralamak, hataları bulmak ve temizlemek emek yoğun bir çalışma gerektirmektedir. Bu yüzden sözlük çalışması bu aşamada kısmen yapılmış olup ileride uzmanlar tarafından hem tamamlanması hem de kontrolden geçirilerek düzeltilmesi gerekmektedir. Sözlükten bazı örnekleri Tablo 2.14'de verilmiştir. Tabloda Osmanlıca kelimenin Osmanlıca yazılışı hem Arap ve Latin alfabesiyle hem de Türkçe karşılıkları tutulmaktadır. Türkçe karşılıkları birden fazla olabilmektedir.

Tablo 2.14: Dil Çevirisi aktarım sözlüğü

Osmanlıca	**Osmanlıca**	**Türkçe**
اناء	ânâ'	anlar, zamanlar
ابخره	ebhire	buharlar, dumanlar, buğular
رستوراسیون	restorasyon	yenileme
زلال	zülâl	berrak, saf
زمام	zimâm	yular
ذوق بخش	zevk bahş	zevk veren
مشعر	meş'ar	hasse, duygu, ziyaret yerleri
ظلم	zulm	cefa, eziyet, zulüm
ظهور	zuhûr	ortaya çıkma, görünme, çıkma

3. MALZEME VE YÖNTEM

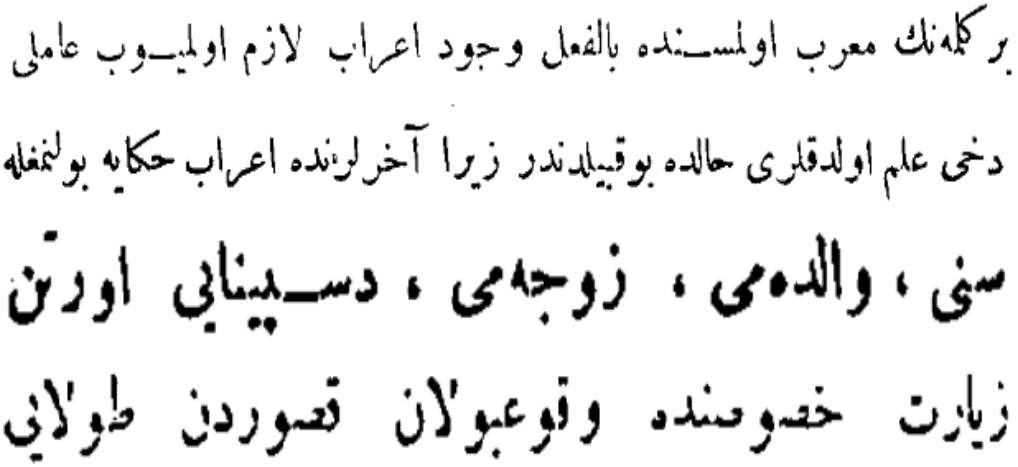
3.1. OSMANLICA OCR

Bu bölümde görüntü ön işlemlerden, veri setlerinden, derin öğrenme modelinden ve OCR araç/model'lerinden bahsedilecektir.

3.1.1. Görüntü Ön İşlemeler

1. *Eşikleme:* Görüntüyü siyah beyaza dönüştürme işlemidir. Bu işlemde Osmanlıca dokümanları eşikleme uygulanarak siyah beyaz görüntüyü dönüştürülmüştür. Normalde bir gri görüntüyü ikili biçime çevirmek için izlenecek yöntem oldukça basittir. Bir eşik değeri belirlenir ve bu eşik değerin üzerindeki renkler beyaza, altındaki renkler siyaha dönüştürülür. Ancak tüm görüntüler aynı niteliklere sahip değildir. Sabit bir eşik değeri tüm görüntüler üzerinde kabul edilebilir sonuçlar üretemeyebilir. Dolayısıyla eşik değerin, resmin renk dağılımına uygun olarak belirlenmesini sağlayacak bir yönteme ihtiyaç duyulur. Bu yüzden eşikleme için Otsu eşikleme tercih edilmiştir. Otsu metodu, gri seviye görüntüler üzerinde uygulanabilen bir eşik tespit yöntemidir. Bu metot kullanılırken görüntünün arka plan ve ön plan olmak üzere iki renk sınıfından oluştuğu varsayımı yapılır. Daha sonra tüm eşik değerleri için bu iki renk sınıfının sınıf içi varyans değeri hesaplanır. Bu değerin en küçük olmasını sağlayan eşik değeri, optimum eşik değeridir.

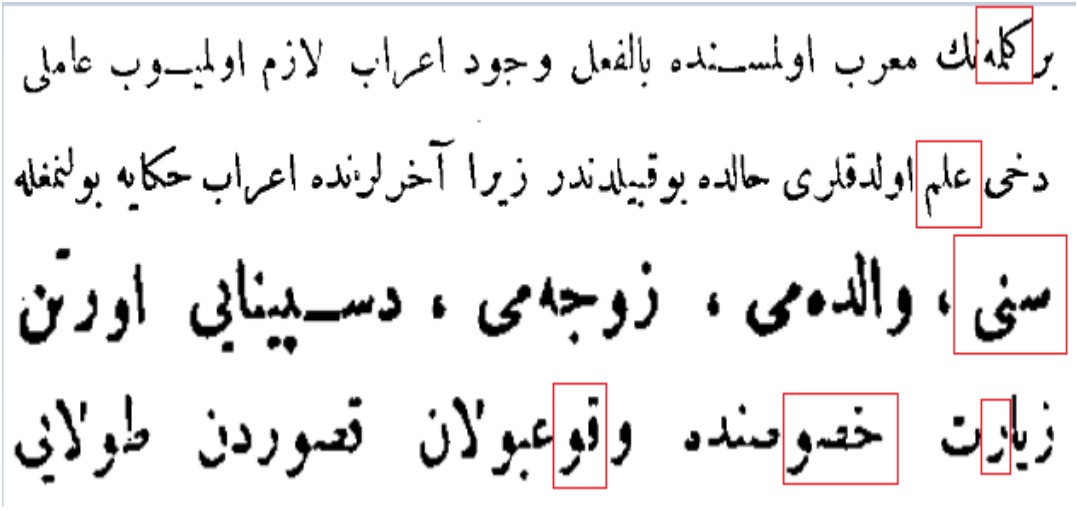
   Osmanlıca dokümanlarda kullanılan eşikleme metodu Otsu metodudur. Otsu eşikleme kullanılarak resimler ikili resme dönüştürülmüştür. Resimler ikili görüntü formatına dönüştürülürken beyaz arka plan metinler siyah renk olarak yapılmaktadır. Osmanlıca dokümanlardaki bütün resimler beyaz arka plan ve siyah metin olacak şekilde eşikleme uygulanmıştır. Farklı eserlerden birkaç satır Şekil 3.1'de gösterilmiştir.



Şekil 3.1: Osmanlıca resme eşikleme uygulama

2. *Resim Boyut:* Osmanlıca dokümanları incelediğimizde ortalama 22 satır bulunmaktadır. Her satır yüksekliği ortalama 45px olmaktadır.
3. *Gürültü Temizleme*: Gürültü, görüntünün metninin okunmasını zorlaştırabilecek bir görüntüdeki parlaklık veya rengin rastgele çeşitlemesidir. İkili ayrılma adımında bazı gürültü türleri doğruluk oranlarının düşmesine neden olabilir. Bu sebepten dolayı Osmanlıca dokümanlar ikili formatta dönüştürdükten sonra resimdeki gürültüleri silmek için bazı filtreler uygulanmıştır. Bu filtreler Osmanlıca'daki olası gürültüleri temizlemiştir.

Osmanlıca görüntü ön işlemlerden geçirildiği zaman harf gruplarında bazı kopmaların meydana geldiği gözlemlenmiştir. Şekil 3.2'de gösterilmiştir. Bu kopmaların gidermek için görüntü ön işlemlerinden genişleme, aşındırma, kapama ve açma vb. ön işlemler uygulanması gerekmektedir.



Şekil 3.2: Önişlemlerden sonra resimde bazı harf kopuklukların oluşması

3.1.2. Veri Seti

Bu bölümde derin öğrenme modelini eğitmek için kullanılan eğitim veri seti ve modelleri karşılaştırma için kullanılan test veri seti tanıtılacaktır.

3.1.2.1.Eğitim veri seti

Etkin sınıflandırma modelleri oluşturmak için mümkün olduğunca büyük veri modelleriyle eğitim yapılması gereklidir. Eğitim verisi orijinal ve sentetik olmak üzere iki alt veri kümesinden oluşmaktadır:

1. *Orijinal veri*: Farklı Osmanlıca eserlerden yaklaşık 1000 sayfalık görüntü dosyası toplanmış ve bunlar yarı otomatik yöntemlerle metin dosyasına dönüştürülmüştür. Tablo 3.1'de özetlenen veri seti yaklaşık 1000 sayfa, 18 bin satır, 35 bin kelime, 252 bin karakterden oluşmaktadır. Orijinal eğitim ve test veri kümelerinde sayfalar yaklaşık 1400 x 2000 piksel boyutlarında ve 300 dpi çözünürlüktedir. Bir sayfada ortalama 20 satır bulunmaktadır. Font boyutu yaklaşık 12 piksel ve satır yüksekliği 48 pikseldir. Bazı sayfalar Şekil 3.3'ta gösterilmiştir.

Tablo 3.1: Orijinal eğitim veri seti

Eser	Yazar	Sayfa
Ateşten Gömlek	Halide E. Adıvar	210
Ay Peşinde	Refik H. Karay	120
+50 eser	...	600

Şekil 3.3: Osmanlıca orijinal veri seti için örnek sayfalar

2. *Sentetik veri*: Orijinal veri hazırlamak uzun zaman aldığı için metin-görüntü dönüşüm araçları kullanarak sentetik bir veri hazırlanmıştır. Bu veri setindeki metinler 70 farklı Arapça fontuyla görüntü dosyalarına dönüştürülmüştür. Bu veri seti için Tablo 3.2'deki kaynaklar kullanılmıştır. Bu veri seti yaklaşık 26 bin sayfa, 1.3 milyon satır, 263 bin kelime, 78 milyon karakterden oluşmaktadır. Sentetik veri setindeki dokümanlar 2500x4800 boyutunda ve 300 dpi çözünürlükte olup, font boyutu 12 noktadır. Her sayfada ortalama 42 satır olup satır yüksekliği 48 noktadır. Bazı örnekler Şekil 3.4'de gösterilmiştir.

Şekil 3.4: Sentetik veri örnekleri

Tablo 3.2: Sentetik eğitim veri seti

Eser	Yazar	Sayfa
Çözümlemeli Osm. Metinler	Mustafa Özkan	552
Resimli Kitaptan Osm. Metinler	Ahmet Z. İzgöer	200
Altın Çiftlik	Mustafa Rahmi	128
Kar	Orhan Pamuk	464

Veri kümlerinin sayfa, kelime, satır ve karakter sıklıkları Tablo 3.3'te verilmiştir. Tabloda sentetik, orijinal, hibrit ve test veri kümeleri verilmiştir. Hibrit veri kümesi sentetik ve orijinal veri kümeleri birleştirilerek elde edilmiştir. Osmanlıca OCR eğitim veri seti karakter sıklıkları Tablo 3.5'te, en çok geçen katar (ligature) Tablo 3.4'te oranları verilmiştir. Osmanlıca harflerin ayırt edici bazı temel özelliklere göre gruplanması Şekil 3.5'te verilmiştir.

Tablo 3.3: Veri seti sıklıkları

Veri seti	Sayfa	Satır	Kelime	Karakter
Sentetik	26B	1.3M	263B	78M
Orijinal	1B	18B	35B	252B
Hibrit	27B	1.3M	298B	78M
Test	21	420	3B	23B

Tablo 3.4: Eğitim seti katar sıklıkları

Katar	Sıklık	Oran %	Katar	Sıklık	Oran %	Katar	Sıklık	Oran %
ا	147947	11,700	يد	10412	0,823	يو	6928	0,548
ر	79248	6,267	لر	10156	0,803	ل	6834	0,540
و	73595	5,820	يا	9818	0,776	گو	6688	0,529
د	59348	4,693	م	9281	0,734	ما	6594	0,521
ه	39707	3,140	پا	9153	0,724	ت	6298	0,498
ن	34605	2,737	لا	8716	0,689	ك	6254	0,495
ى	29955	2,369	قا	7844	0,620	ب	6228	0,493
ز	17913	1,417	ند	7621	0,603	ق	6155	0,487
بر	16148	1,277	ير	7508	0,594	يله	6009	0,475
بو	13778	1,090	با	7186	0,568	قو	5917	0,468

Tablo 3.5: Eğitim seti karakter sıklıkları

	Karakter	Sıklık	Oran %		Karakter	Sıklık	Oran %		Karakter	Sıklık	Oran %
0	ا	299674	9.773	**21**	ف	25297	0,825	**42**	؛	1124	0,037
1	ی	281385	9,177	**22**	،	21246	0,693	**43**	١	1116	0,036
2	ر	194499	6,343	**23**	ص	19163	0,625	**44**	٢	800	0,026
3	و	180434	5,885	**24**	ك	18400	0,600	**45**	!	765	0,025
4	ل	160079	5,221	**25**	چ	17626	0,575	**46**	:	749	0,024
5	ن	158129	5,157	**26**	(	17417	0,568	**47**	«	596	0,019
6	د	138488	4,517	**27**	)	17313	0,565	**48**	٠	585	0,019
7	ه	133503	4,354	**28**	ط	17179	0,560	**49**	٣	561	0,018
8	م	120282	3,923	**29**	خ	16489	0,538	**50**	»	519	0,017
9	ب	86833	2,832	**30**	غ	16375	0,534	**51**	٥	513	0,017
10	ت	82384	2,687	**31**	.	10048	0,328	**52**	٤	494	0,016
11	ك	75714	2,469	**32**	ض	6727	0,219	**53**	٩	435	0,014
12	س	66018	2,153	**33**	ث	6291	0,205	**54**	ء	409	0,013
13	ق	63454	2,069	**34**	ه	6182	0,202	**55**	٦	377	0,012
14	ش	43261	1,411	**35**	أ	4830	0,158	**56**	٨	359	0,012
15	پ	41171	1,343	**36**	ئ	4723	0,154	**57**	٧	341	0,011
16	ع	35294	1,151	**37**	ظ	4164	0,136	**58**	[	292	0,010
17	ز	34049	1,110	**38**	ذ	4099	0,134	**59**	]	263	0,009
18	ج	29035	0,947	**39**	؟	3900	0,127	**60**	ؤ	155	0,005
19	ح	25921	0,845	**40**	ژ	3468	0,113	**61**	إ	70	0,002
20	گ	25524	0,832	**41**	—	2562	0,084				

Grup	Harf	Sayı
Bağlı	بتثجحخسشصضطظعغفقكلمنهەیپچگڤ	28
Bağlı Olmayan	ادذرزوۆ	7
Noktalı	بتثجخذزشضظغفقنیپچژڤ	20
Noktasız	احدرسصطعكلمهەوگ	15
Yükselen	الكگڤظطلذ	9
Alçalan	یونمقغعضصشسزژرخحجج	19
Ortada	هەفثتبپ	7
Yuvarlak	صضطظعغفقمەهو	13
Dişli	بتثسشنپی	8
Düz	اكلگ	5
Diğer	جچحخدذرزژ	9
Noktalı-Bitişik	بتثجخشضظغفقنیپكچ	17
Noktasız-Bitişik	حسصطعكلمهگ	10
Noktasız-Bitişmeyen	ادرەو	5
Noktasız-Bitişmeyen	ذزژ	3

Şekil 3.5: Harf grupları

3.1.2.2.Test veri seti

Test için 8 farklı eserden 21 orijinal sayfa görüntüsü kullanılarak bir veri seti hazırlanmıştır. Test veri seti seçilen eserler Tablo 3.6'de verilmiştir. Test veri seti eğitim setinde kullanılmamış verileri içermektedir. Test veri seti hazırlanırken eserlerden silik harfli, mürekkebi dağılmış ve farklı tür kâğıda basılı sayfalar seçilmeye çalışılmıştır. Bazı sayfalardan tek satır alınarak Şekil 2.2'de gösterilmiştir. Ortalamada her sayfa 20 satırdan, her satır 7 kelime ve 55 karakterden oluşmaktadır. Test veri seti osmanlica.com/test adresinde paylaşılmıştır. Paylaşım 21 adet görüntü dosyası, 6 adet OCR test çıktısı dosyası, doğru metni içeren bir adet Osmanlıca metin dosyası ve difflib kullanılarak doğruluk oranını hesaplayan bir adet Python dosyası içermektedir.

Tablo 3.6: Orijinal test veri seti

Yazar	**Eser**	**Sayfa**
Ahmet Mithat	Zeyli Kâinat	3
Esad Halil	Hulasatul Şuruh	2
Mehmet Celal	Venüs	3
Mehmet E. Yurdakul	Türkçe Çeviriler	2
Mehmet E. Köprülü	Yeni Eserlerim	3
Namık Kemal	Vatan Yahut Silistre	3
Şemseddin Günaltay	Hurufattan Hakikate	3
Ziya Paşa	Zafername	2

Test veri setinde karakterlerin sıklık oranları Tablo 3.7'de, sık geçen katarlar (ligature) Tablo 3.8'de ve kelime sıklıkları da Tablo 3.9'da verilmiştir. Tablolar incelendiği zaman bazı karakterlerin sıklık oranları çok az olduğu görülecektir. Bunun sebeblerinde bir tanesi de Osmanlıca'da karakteri barındıran kelime sayısının nadir olmasında kaynaklanmakadır. Örneğin ژ (je harfi) Osmanlıca'da diğer harflere kıyasen çok az kelimede bulunmaktadır.

Tablo 3.7: Test veri seti karakter sıklıkları

Karakter	Sıklık	Oran %	Karakter	Sıklık	Oran %	Karakter	Sıklık	Oran %
ا	1937	9.478	ج	171	0,837	؟	17	0,083
ى	1740	8,514	.	157	0,768	؛	14	0,069
ر	1377	6,738	ص	130	0,636	:	14	0,069
ل	1210	5,921	،	110	0,538	ؤ	10	0,049
و	1171	5,730	غ	109	0,533	١	9	0,044
ن	1127	5,515	ط	103	0,504	٢	7	0,034
م	917	4,487	خ	102	0,499	٥	7	0,034
د	909	4,448	چ	86	0,421	—	7	0,034
ه	707	3,460	ض	59	0,289	٨	6	0,029
ب	640	3,132	ك	52	0,254	٠	5	0,024
ك	630	3,083	(	51	0,250	٣	5	0,024
ت	528	2,584	)	50	0,245	٦	5	0,024
ق	452	2,212	ظ	49	0,240	٧	4	0,020
س	413	2,021	پ	41	0,201	ء	3	0,015
ه	301	1,473	أ	41	0,201	٤	3	0,015
ع	300	1,468	ذ	37	0,181	٩	2	0,010
ش	261	1,277	!	37	0,181	ژ	1	0,005
ف	234	1,145	ث	35	0,171			
ز	225	1,101	ئ	30	0,147			
ح	189	0,925	گ	24	0,117			

Tablo 3.8: Test veri seti katar sıklıkları

Katar	Sıklık	Oran %	Katar	Sıklık	Oran %	Katar	Sıklık	Oran %
ا	1067	12,656	يد	89	1,056	ل	56	0,664
و	546	6,476	لر	87	1,032	ق	55	0,652
ر	503	5,966	بو	84	0,996	يا	55	0,652
د	366	4,341	ه	78	0,925	ند	53	0,629
ن	251	2,977	ب	72	0,854	لد	49	0,581
ه	226	2,681	كو	68	0,807	نلر	46	0,546
ى	176	2,088	ك	60	0,712	ما	42	0,498
بر	139	1,649	لا	59	0,700	قو	42	0,498
ز	96	1,139	ير	58	0,688	م	41	0,486

Tablo 3.9: Test veri seti kelime sıklıkları

Kelime	Sıklık	Oran %	Kelime	Sıklık	Oran %
بر	76	2,126	کبی	15	0,420
و	71	1,986	اعراب	12	0,336
او	38	1,063	ایچون	12	0,336
ایله	21	0,587	معرب	10	0,280
ایسه	19	0,531	کوره	9	0,252
اولان	17	0,476	ایدن	9	0,252
دخی	16	0,448	اولوب	8	0,224
بو	15	0,420	اکر	8	0,224
اول	15	0,420	قدر	8	0,224
اولور	7	0,196	اعرابه	7	0,196
ایدی	7	0,196	در	7	0,196
دیمك	7	0,196	اك	7	0,196
مبنی	7	0,196	ایدی	7	0,196
زیرا	7	0,196	حالده	6	0,168

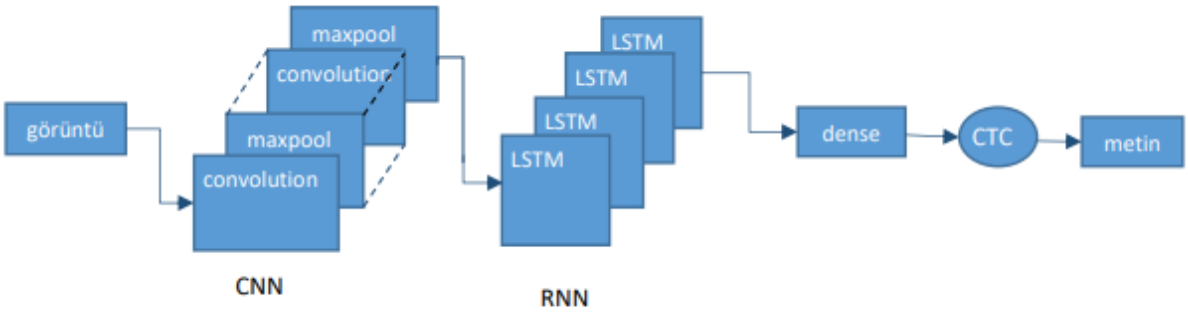
3.1.3. Osmanlıca OCR Derin Öğrenme Modeli

Derin öğrenme (aynı zamanda derin yapılandırılmış öğrenme, hiyerarşik öğrenme ya da derin makine öğrenmesi) bir veya daha fazla gizli katman içeren Yapay Sinir Ağları ve benzeri makine öğrenme algoritmalarını kapsayan çalışma alanıdır [57]. Derin öğrenmede; Evrişimsel Sinir Ağları (Convolutional Neural Network - CNN), Yinelemeli Sinir Ağları (Recurrent Neural Network - RNN) vb. çalışma alanları bulunur. CNN daha çok görüntü işlemede (resimde nesne tanıma, tibbi görüntü analizi, resim sınıflandırma vb.), RNN ise daha çok doğal dil işlemede (metin sınıflandırma, makine çevirisi, duygu analizi vb.) kullanılır. Derin öğrenme modelleri, OCR dâhil olmak üzere birçok zor problemde geleneksel Yapay Sinir Ağlardan daha iyi performans göstermiştir [58]. Derin öğrenme modelleri yüksek doğruluk oranına sahip sınıflandırma yapar ve önemli ölçüde hatayı minimize eder [59]. Derin öğrenme modelleri, Arapça temelli yazılara da uygulanmış ve başarılı oldukları bir kez daha kanıtlanmıştır [60-62].

Evrişimsel Sinir Ağları genelde Evrişim katmanı (convolution layers), Ortaklama (pool) ve Tam Bağlantı (fully connected) katmanlarından oluşur. Evrişim katmanında resme filtreler uygulayarak bir resmin öznitelikleri çıkartılır. Ortaklama katmanı (pooling), tipik olarak bir miktar uzamsal değişkenlik gösteren bir evrişim katmanından sonra uygulanan bir örnekleme işlemidir [63]. Maksimum ve Ortalama Ortaklama olmak üzere 2 yöntem bulunur. Bu çalışmada Maksimum Ortaklama (Max-pooling) tercih edilmiştir. Maksimum Ortaklama (Max Pooling) evrişim katmanına uygulanan filtrelerden maksimum değeri alarak örnekleme yapar. Ortaklama (Max pooling) katmanı hesaplamayı azaltır ve özelliklerin algılanmasını ölçekleme veya yönlendirme işlemlerinde kararlı sonuçları evrişimli katmanlardan tahmin eder.

Yinelemeli Sinir Ağları (Recurrent Neural Network- RNN) katman, temelde bir önceki çıktıyı girdi olarak kullanılmasına olanak veren sinir ağlardır. Uzun kısa süreli bellek (Long Short Term Memory - LSTM) mimarisi ise RNN genişletilmiş halidir.

RNN katmanı dört alt katmana organize edilmiş LSTM'lerden oluşur. Çift yönlü LSTM'ler, birçok dizi sınıflandırma probleminde model performansını artıran geleneksel LSTM'lerin daha iyi bir sürümüdür. Çift yönlü LSTM'ler, giriş dizisinde iki LSTM'yi eğitir. İlk LSTM, giriş dizisi üzerinde eğitilir ve ikinci LSTM, ters giriş dizisi üzerinde eğitilir, böylece ek verilerle daha iyi ağlar kurar ve girişi tek yönlü LSTM'lerden daha iyi öğrenir. Son olarak, transkripsiyon veya çıktı katmanı, karakter sınıflarının çıktı dizisini üretir. Çıktı katmanındaki sınıf etiketlerinin kodunu çözmek için bir Connectionist Temporal Classification (Bağlantıcı Zamansal Sınıflandırma – CTC) kayıp işlevi kullanılır. Bu çalışmada kullanılan model, değişken boyutlu görüntüleri kabul eder. Hem genişlik hem de yükseklik değişebilir. Görüntüler önce ikiliye dönüştürülür ve ardından görüntüler satırlara ve karakterlere ayrılmadan önce görüntüleri temizlemek ve normalleştirmek için görüntü işlemleri uygulanır. Eğitim veri setindeki her satır için bir görüntü ve buna karşılık gelen gerçek metni (Ground Truth – GT) üretilir.



Şekil 3.6: Osmanlıca OCR derin öğrenme modeli

Deneylerde kullanılan CRNN mimarisi bir CNN, bir RNN, bir yoğun katman ve bir CTC işlevinden oluşmaktadır. Mimarinin CNN bölümü 16 filtreli 3x3 convolution içeren bir katman ve 3x3'lük bir maxpool katmanından oluşmaktadır. Mimarinin RNN bölümü 4 katmanlıdır ve her katman bir adet LSTM içermektedir. Birinci katman y boyutunda 64 çıktılı bir LSTM, ikinci ve üçüncü katman x boyutunda 128 çıktı ile özetleme yapan biri ileri yönlü diğeri geri yönlü olmak üzere iki yönlü (bidirectional) bir LSTM, ve dördüncü katman x boyutunda 256 çıktı üreten ileri yönlü bir LSTM'den oluşmaktadır. RNN bölümünden sonra sınıflandırmanın yapıldığı bir yoğun (dense) katman ve bu katman çıktısını yorumlayan bir CTC işlevi kullanılmaktadır. Eğitim sırasında doküman satırlara bölümlendikten sonra satır görüntüsü CRNN'e girdi olarak, satır görüntüsündeki metin ise çıktı olarak kullanılır.

3.1.4. Karşılaştırma: Deney ve Sonuçlar

Geliştirilen OCR modelinin performansını ölçmek için test veri seti kullanılarak deneysel bir karşılaştırmalı yapılmıştır. Bu bölümde önce karşılaştırma deneylerinde kullanılan OCR araçları/modeller tanıtılacaktır. Sonrasında karşılaştırmada kullanılan doğruluk oranları açıklanacak ve son olarak hatalar detaylı olarak incelenecektir.

3.1.5. OCR Araçları

OCR araçları ticari, açık kodlu ve web/API tabanlı olmak üzere üç başlıkta incelenebilir. Piyasada Google Docs [64] , Microsoft Office 360 [65], Adobe Acrobat Pro [66], Abby Finereader [67], Omnipage [68], ReadIRIS [69], Sakhr [70] vb. ticari OCR araçlara ek olarak Tesseract [71], OCRopus [72], Ocrad [73], GOCR [74], Cuneiform [75] benzeri açık kaynak kodlu OCR araçları mevcuttur. Özellikle Google tarafından desteklenen Tesseract OCR projesi FreeOCR [76] vb. birçok açık kodlu OCR uygulamasında kaynak olarak kullanılmaktadır.

Bundan başka son zamanlarda uygulamalara entegre edilmesi amacıyla Google Cloud Vision, Microsoft Computer Vision, Amazon Rekognition, Tesseract.js vb. web servisleri OCR hizmeti sağlamaktadır.

Bu çalışmada geliştirilen OCR modeli Tesseract'ın Arapça ve Farsça için eğitilmiş (pretrained) OCR modelleri, ABBY FineReader OCR, Google Docs OCR ve Miletos OCR araçlarıyla karşılaştırılmıştır ve sonuçlar aşağıda verilmiştir. Karşılaştırmada kullanılan test veri setinin OCR araçlarının model eğitiminde kullanılmadığı var sayılmıştır. Test sonuçları bu varsayımı destekler mahiyettedir.

1. *Tesseract Arapça ve Farsça*: Tesseract; açık kaynak kodlu geliştirilen bir optik karakter tanıma motorudur. Tesseract OCR motoru ilk olarak 1985-1994 yılları arasında Bristol, İngiltere ve Greeley, Colorado'da Hewlett Packard laboratuvarlarında C dilinde geliştirilmiştir. 1998'de C'den C ++'ya bazı geçişler gerçekleşti. Daha sonra tüm kod C++ diline çevrilmiştir. Takip eden on yılda Tesseract'da çok az çalışma yapılmıştır. Daha sonra 2005 yılında Hewlett Packard ve Nevada Üniversitesi, Las Vegas (UNLV) tarafından açık kaynak olarak piyasaya sürülmüştür. Tesseract, 2006'dan beri Google tarafından desteklenmektedir. Birçok sürümü vardır. Sürüm 4'ten sonra derin öğrenme modellerini destekler. Arapça ve Farsça için sürüm 4 ile başarılı sonuçlar elde edilmiştir. Arapça ve Farsça modeller oluşturulmuştur. Modeller yaklaşık 400000 satır ve 4500 farklı fontla eğitilmiştir.
2. *ABBY FineReader* (sürüm 15): ABBYY tarafından geliştirilen bir optik karakter tanıma aracıdır. Taranan görüntüleri, belgeleri, fotoğrafları ve PDF dosyalarını Microsoft Word, Microsoft Excel, Microsoft PowerPoint, Rich Text Format, HTML, PDF/A, aranabilir PDF, CSV ve text dosyaları gibi düzenlenebilir ve aranabilir formatlara çevirebilmek için tasarlanan profesyonel düzeyde bir uygulamadır. Finereader piyasada yaygın olarak kullanılan ticari OCR araçlarından bir tanesi olup Arapça ve Farsça dâhil çok sayıda dil için OCR hizmeti sağlamaktadır.
3. *Google Docs*: Google'ın Google Documents kapsamında kullanıcılarına sunduğu çevrimiçi OCR servisi yaygın olarak kullanılan OCR araçlarından bir tanesidir. Google docs'a yüklenen jpg, gif, png ve pdf formatındaki dosyaları OCR yaparak düzenlenebilir metin haline getirmektedir.

4. *Miletos*: Miletos adlı firma tarafından Osmanlıca için geliştirilen ve çevrimiçi OCR hizmeti sağlayan ticari bir OCR aracıdır. Miletos özellikle Osmanlıca için geliştirildiğinden karşılaştırmaya dâhil edilmiştir.

3.2. ALFABE ÇEVİRİSİ

Osmanlıca - Türkçe Alfabe çevirisi 6 aşamadan oluşur. Alfabe çevirisinin başarısı bu 6 aşamanın başarısına bağlıdır. Alfabe çevirisi süreci kendi içinde genel olarak ortografik alfabe çevirisi, kelime bölütleme, sözcük tahmini, yazım düzeltme, seslendirme, tamlama ve birleşik kelimeler gibi her biri başlı başına ayrı bir problem olan alt adımlardan meydana gelir. Her alt adımın etkinliği ve doğruluğu sürecin toplam başarısına doğrudan etki yapmaktadır. Alfabe çevirisi sürecin çok adımlı ve karmaşık olması, bu adımların sözlük, imla kılavuzu, derlem gibi dil kaynakları gerektirmesi, süreçte kısmen de olsa bazı fonetik, morfolojik, gramatik ve semantik problemlerin çözülmesi gerektiğinden dolayı Osmanlıca OCR probleminden daha güç bir problem olarak karşımıza çıkmaktadır.

3.2.1. Osmanlıca - Türkçe Alfabe Çevirisi Sistemi

Arapça Tabanlı Osmanlıca Alfabesinden Latin Tabanlı Modern Türkçe'ye Alfabe Çeviri Sistemi (Ottoman-Turkish Transliteration System) tasarlanmış ve geliştirilmiştir. Bu konuda daha önce Esma F Bilgin [2] tarafından geliştirilen sistem esas alınmıştır. Yukarıda anlatıldığı gibi önerilen yöntem temelde 6 aşamadan oluşur. Bu aşamaları kısaca şöyle sıralayabiliriz;

1. Alfabe Çevirisi (Transliteration)
2. Kelime Bölütleme (Word Boundary Correction)
3. Yazım Düzeltme (Spell Correction)
4. Seslendirme (Vowelization)
5. n-grams (Word prediction/Language models)
6. Tamlamalar & birleşik isimler (Nominal phrases & Compound words)

Birinci aşama olan Alfabe çevirisinde; morfolojik analiz ile metinde olan tüm kelimelerin gövdelemesi yapılarak gövde ve ekleri bulunur. Morfolojik analiz kelimelerin dilbilgisi kurallarına göre kök ve eklerine ayrıştırılmasıdır. Morfolojik analiz kökler, gövdeler, ön ekler ve son ekler gibi sözcüklerin yapısını inceler. Morfolojik analizde her Osmanlıca kelimenin olası tüm kök ve ek bulunmaktadır. Kelimenin kökünü bulmak için sözlük kullanılmaktadır. Morfolojik analiz için kullandığımız sözlükler şunlardır;

1. Kelime Kökleri (33800)
2. Fiil Ekleri (11050 ek)
3. İsim Ekleri (313 ek)

Osmanlıca kelimeyi sözlükte ararken ilk önce tamamı aranır. Osmanlıca kelime bulunmaz ise kelime sonunda bir harf çıkartılarak işleme devam edilir. Osmanlıca kelime iki harf kalana kadar bu işlem devam etmektedir. İşlemin iki harf kalana kadar devam edilmesinin sebebi ise Osmanlıca kelimelerin en az iki harften (bazı kelimeler dışında örneğin و → ve) oluşmasıdır. Böylece Osmanlıca kelime için olası tüm morfolojik çözümler çıkartılır. Çözümlerin doğruluğu morfolojik sentezde ortografik kurallara dikkat edilerek karar verilir.

Morfolojik analizde Osmanlıca kelimenin kökü ve ekleri bulunmaktadır. Bir kelime için birden fazla kök veya ek bulunabilir. Bulunan kök ve ekler kök ve ek sözlükleri yardımıyla birleştirmektedir. Sentez sonucu Osmanlıca bir kelime için Türkçe'de birden fazla çözüm bulunabilir.

Osmanlıca kelime kökü ve ekleri veri tabanına kayıt edilirken belli kurallara göre kayıt edilmektedir. Osmanlıca kelimenin isim mi, fiil mi olduğu kayıt edilir. Tablo 3.10'da gösterildiği gibi bir kelimenin hem farklı yazım şekilleri hem de isim veya fiil olduğu kayıt edilir.

Tablo 3.10: Alfabe çeviri kelime kökü sözlük örnekleri

Osmanlıca	Türkçe	Pos
ال	al	isim
آل	al	isim
ال	al	fiil
آل	al	fiil
عمر	amr	isim
عمر	ömer	isim
عمر	ömür	isim

Osmanlıca ekler isimden-isim, isimden-fiil, fiilden-fiil, fiilden-isim türlerine dikkat edilerek kayıt edilmiştir. Kelime çözümlerken bu kurallara dikkat edilerek çözümleme gerçekleşmektedir. Tablo 3.11'de gösterildiği gibi eklerdeki bazı sesler belli bir küme ile gösterilmiştir. D={d,t} H ={i,ı,ü,u} A ={a,e} seslerine karşılık gelmektedir. Osmanlıca kelime ve ek karşılıkları bulunduktan sonra birleştirme ortografik kurallara göre yapılmıştır.

Tablo 3.11: Alfabe çeviri ekler sözlüğü: küme örnekleri

Osmanlıca	Türkçe		Osmanlıca	Türkçe	
Ek	Yüzey	Ek	Ek	Yüzey	Ek
می	mH	mi	دق	DHk	dık
می		mı	دق		duk
می		mu	دق		tuk
می		mü	دق		tık
لر	lAr	lar	دك		dik
لر		ler	دك		dük
ەجك	ecek	ecek	دك		tük
اجق	acak	acak	دك		tik

Türkçe sondan eklemeli bir dildir (agglutinative). Yapım ve çekim ekleriyle yeni kelimeler türetilmekte veya çekilebilmektedir. Bir kelime çok sayıda ek alabilir. Morfolojik analiz yapılırken gövdeler bulunduktan sonra ekler özyinele (rekürsif) bir şekilde çözümlenmiştir. Tablo 3.12'da bazı örnekler gösterildiği gibi;

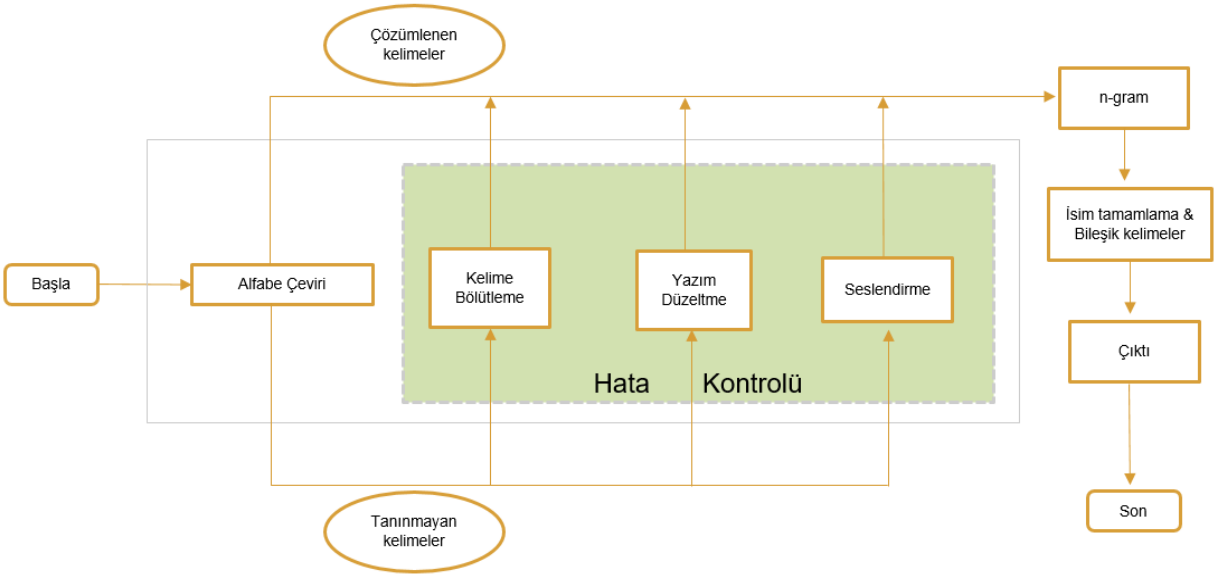
osmanlılaştıramayabileceklerimizdenmişsinizcesine→osman+ lH+IAş+DHr+A+mA+yAbil+ecekler+Hm+Hz+DAn+mHşsHnHz+CA+SHnA kelimesi bir gövde ve ek katarlarından oluşmaktadır. Özyinele algoritması ile bu ekler bulunur. Morfolojik sentez ile ortografik kurallara göre kelime türetilir. Eğer kelime kurallara uyarsa geçerli sayılır kurallara uymazsa geçersiz sayılır.

Tablo 3.12: Morfolojik analiz ve sentez örnekleri

Osmanlıca	Kök	Ekler	Türkçe
چركينلكلري	çirkin	lHk + lAr + H	çirkinlikleri
الكزده كي	el	HnHz + dAki	elinizdeki
لياقتسزلگمزي	liyakat	sHz + lHğH + mHzH	liyakatsizliğimizi
عثمانليلاشديرامايابيله جكلريمزدنمشسكزجه سنه	osman	lH+IAş+DHr+A+mA+yAbil+ecekler+Hm+Hz+DAn+mHşsHnHz+CA+SHnA	osmanlılaştıramayabileceklerimizdenmişsinizcesine

Birinci aşamada çözümlenmeyen kelimeler çıkabilir. Bu kelimeler bir sonraki aşamaya (Aşama 2:Kelime Bölütleme) aktarılır. Kelime Bölütleme aşamasında, kelimeler düzgün bölütlenmediği için bir kelime bir önceki kelimenin parçası gibi durumlar oluşabilir. Bundan dolayı bu aşamada bir kelime bir önceki kelime ile birleştirilerek çözümlenmeye çalışılır. Eğer çözümlenir ise bu çözüm doğru kabul edilir. Bu kelime çözülmezse Aşama 3'e geçilir. Aşama 3: Yazım Düzeltme'de çözümlenemeyen kelimede yazım hatası olduğu kabul edilerek ilerlenir. Yazım hatası düzeltme algoritması uygulanır. Yazım düzeltme algoritması sonrası elde edilen kelime yeniden çözümlenmeye çalışılır. Eğer kelime çözümlenebilirse bu çözüm doğru kabul edilir. Eğer kelime çözümlenmezse, bir sonraki aşamaya Aşama 4'e geçilir. Aşama 4'te yani Seslendirme aşamasında birebir harf eşleşmeleri yapılır ve bir çözüm elde edilir.

Alfabe çevirisi sonucundan Osmanlıca bir kelimenin Türkçe'de birden fazla çözümü bulunabilir. Bundan sonra n-gram işlemi uygulanarak birden fazla çözüm arasından olasılığı en yüksek olanı seçilir. En son aşamada metinde isim tamlamaları ve birleşik isimler varsa düzeltilmeler yapılır ve Türkçe çıktı oluşturulur. Bu altı aşama Şekil 3.7'da gösterilmiştir.



Şekil 3.7: Alfabe çevirisi akış şeması

Tablo 3.13'de bu aşamalar bir örnek ile gösterilmiştir. Tabloda birden fazla çözüm { } sembolleri arasında verilmiştir. Kelime bölütleme aşamasında bölütlenmiş (اوقويه جق okuya+cak) kelimesi doğru şekilde çözümlenmiştir. Yazım düzeltme aşamasında حرايه → haraye kelimesi حرابه → harabe şeklinde düzeltilmiştir. Seslendirme aşamasında معمورة → mamure doğru biçimde elde edilmiştir. Osmanlıca kelimelerin Türkçe'de birden fazla okunuşu söz konusu olabilmektedir: بر →{bir, ber}, اولان →{ölen, olan, avlan} Bu durumda Türkçe karşılıkları n-grams ile doğru sıraya koymak gerekmektedir: بر →{bir, ber}, اولان →{olan, ölen, avlan} şeklinde sıralanmıştır. Tamlama ve Birleşik kelime aşamasında بلاد اسلاميه→ *bilad islamiyye* kelimesi *bilad-ı islamiyye* şeklinde düzeltilmektedir.

Tablo 3.13: Alfabe çevirisi sistem çıktıları

Aşamalar	**Sonuçlar**
Osmanlıca Metin	ماضیده هر بری بر معمورۀ مدنیت اولان بلاد اسلامیه بوکون؛ قسوت انکیز برر خرایه حالنی المش. قوای طبیعیه‌یه میدان اوقویه جق
Aşama 1: Ortografik Çeviri	mazide her {berri, beri, biri} {bir, ber} معمورۀ {müdünüyet, medeniyet} {ölen, olan, avlan} bilâd islâmiyye {bugün, bükün} kasvet {enekyiz, engiz} birer حرایه halını {elmiş, almış}. kuva tabiyeye meydan اوقویه جق
Aşama 2: Kelime bölütleme	mazide her {berri, beri, biri} {bir, ber} معمورۀ {müdünüyet, medeniyet} {ölen, olan, avlan} bilâd islâmiyye {bugün, bükün} kasvet {enekyiz, engiz} birer حرایه halını {elmiş, almış}. kuva tabiyeye meydan okuyacak
Aşama 3: Yazım Kontrolü	mazide her {berri, beri, biri} {bir, ber} معمورۀ {müdünüyet, medeniyet} {ölen, olan, avlan} bilâd islâmiyye {bugün, bükün} kasvet {enekyiz, engiz} birer harabe halını {elmiş, almış}. kuva tabiyeye meydan okuyacak
Aşama 4: Seslendirme	mazide her {berri, beri, biri} {bir, ber} mamure {müdünüyet, medeniyet} {ölen, olan, avlan} bilâd islâmiyye {bugün, bükün} kasvet {enekyiz, engiz} birer harabe halını {elmiş, almış}. Kuva tabiyeye meydan okuyacak
Aşama 5: Kelime Tahmini (n-gram)	mazide her {biri, berri, beri} {bir, ber} mamure { medeniyet, müdünüyet} {olan, ölen, avlan} bilâd islâmiyye {bugün, bükün} kasvet {enekyiz, engiz} birer harabe halını {elmiş, almış}. Kuva tabiyeye meydan okuyacak
Aşama 6: Tamlamalar & Birleşik isimler	mazide her {berri, beri, biri} {bir, ber} mamure {müdünüyet, medeniyet} {ölen, olan, avlan} bilâd - ı islâmiyye {bugün, bükün} kasvet {enekyiz, engiz} birer harabe halını {elmiş, almış}. Kuva- ı tabiyeye meydan okuyacak

3.2.2. Test Veri Seti

Osmanlıca – Türkçe Alfabe çevirisi için test veri seti hazırlanmıştır. Test veri seti farklı eserlerden farklı sayfalar alınmıştır. Tablo 3.15'te eserler verilmiştir. Test veri setindeki sayfalar ayrıca OCR yapılıp çıktıları alınıp referans metinler ile paralel bir şekilde hazırlanmıştır. Test veri setinde bazı örnekler Tablo 3.14'te verilmiştir. Tabloda gösterildiği gibi veri seti Osmanlıca (Ground Truth – GT), Türkçe (Ground Truth – GT) ve OCR çıktısından oluşmaktadır. Test veri seti 48 sayfa 7500'e yakın kelime, yaklaşık 11 bin Osmanlıca karakter ve 13 bin Türkçe karakter içermektedir. Osmanlıca-Türkçe Alfabe çevirisi bu test veri seti üzerinde denenmiştir. Sonuçları Bölüm 4'te verilmiştir.

Tablo 3.14: Alfabe çevirisi test veri seti örnekleri

Osmanlıca (GT)	Türkçe	OCR	Osmanlıca (GT)	Türkçe	OCR
اوچنجی	üçüncü	اوچنجی	قورقەجق	korkacak	قورقاجق
فصل	fasıl	فصل	بر	bir	بر
۸۳	83	۸۳	قادین	kadın	قادین
ایچندەکی	içindeki	ایچندە کی	بر	bir	بر
اتش	ateş	اتش	چوجوق	çocuk	چوجوق
کرەیی	küreyi	کرەیی	بولەمزسن	bulamazsın	بولەمزسن
پاتلاتسه	patlatsa	پانلاتسه	بیله	bile	بیله
انك	anın	اك	کورلدیسندن	gürültüsünden	کورلدیسندن

Tablo 3.15: Osmanlıca Türkçe alfabe çevirisi test veri seti

Eser	Sayfa	Kelime	Harf
OCR test seti	21	3226	15,997 Os 19,255 Tr
Ateşten Gömlek	14	2252	12,284 Os 14,252 Tr
Ay peşinde	13	2182	10,989 Os 12,798 Tr
Toplam	48	7433	39,270 Os 46,305 Tr

3.2.3. Alfabe çevirisi algoritması

Osmanlıca – Türkçe alfabe çevirisinde kullandığımız sözlük arama algoritmasının sözde kodu Şekil 3.8'de verilmiştir. Algoritmaya verilen cümle ilk önce tekilleştirilmektedir. Sonra sondan bir harf atılarak Osmanlıca gövde sözlüğünde arama yapılır. Bu arama kelime 1 harf kalana kadar devam etmektedir. Sözlükteki karşılığı bulunan kelimeler ekler ile birleştirilerek çıktı oluşturulur.

```
func alfabe_ceviri(cumle):
    kelimeler = sozluk_kelime_getir()  # sözlükte bulunan kelimeler
    ekler = sozluk_ek_getir()  # sözlükte bulunan ekler

    # Sözlüğü osmanlıca kelime boyuta göre büyükten küçüğe doğru sırala.
    sozluk = sirala(kelimeler)
    # cumleyi tekilleştir ve kelime boyutuna göre büyükten küçüğe doğru sırala
    tekil_kelimeler  sirala(tekillestir(cumle))

    kelimesayisi  kelime_sayisi(tekil_kelimeler)
    sozluksayisi = kelime_sayisi(sozluk)

    while (kelimesayisi >= 0 && sozluksayisi >= 0):
        osmanlica_kelime = kelimeler[kelimesayisi]
        sozlukdeki_osmanlica_kelime = sozluk[sozluksayisi]
        if (len(sozlukdeki_osmanlica_kelime) > len(osmanlica_kelime)):
            sozluksayisi -= 1
        else:
            i = sozluksayisi
            while (i > 0):
                sozcuk = sozluk[sozluksayisi]
                sozcukUzunluk = len(sozcuk)
                osmanlica_govde = osmanlica_kelime[0:sozcukUzunluk]
                if (osmanlica_govde == sozcuk):
                    osmanlıca_ek = osmanlica_kelime[sozcukUzunluk:]

                    # Osmanlıca kelimenin birden fazla türkçe karşılığı olabilir hepsini getiriyoruz.
                    bulunan_tum_osmanlica_kelime = sozlukten_kelime_getir(osmanlica_govde)

                    # Osmanlıca kelimenin Türkçe ve ek karşılıkları bulunup kelime üretiliyor.
                    result = kelime_ve_ek_birlestir(bulunan_tum_osmanlica_kelime, osmanlıca_ek)

                # Osmanlıca da kelime uzunlukları 1 olan kelime sayısı oldukça azdır.
                if len(sozcuk) <= 2:
                    break

    return result
```

Şekil 3.8: Alfabe çevirisi algoritması

3.3. DİL ÇEVİRİSİ

3.3.1. Osmanlıca Türkçe Dil Çevirisi

Osmanlıca - Türkçe dil çevirisi için kelime grubu tabanlı basit bir çeviri sistemi sunulmuştur. Bu yöntem dil içi çeviri içi kullanılabilecek en basit yöntemlerden bir tanesidir. Yöntem kelime öbeği tabanlı ve mekanik olarak işleyen ve bir cümleyi öbek öbek soldan sağa doğru tarayarak Osmanlıca'dan Günümüz Türkçe'sine aktaran bir yöntemdir. Bu yöntemin basit olmasına karşın çok karmaşık olmayan cümleler için etkin bir yöntem olduğunu düşünülmektedir ve elde edilen sonuçların daha ileri yöntemlerin geliştirilmesinde ve performans analizinde bir temel (baseline) olarak ta kullanılabilecektir.

3.3.2. Test veri seti

Osmanlıca - Türkçe Dil çeviri sistemini test edebilmek için cümle cümle paralel bir veri seti hazırlanmıştır. Tablo 3.16'de veri seti için yararlanılan eserler gösterilmiştir. Veri setinde zorluk ve yabancı kelime yoğunluğu farklı metinlerden yararlanılmıştır. Bu küçük test veri seti yaklaşık 200 cümle, 3.150 kelime, 1.900 tekil kelime, 18.500 karakterden oluşmaktadır. Test veri setinden bazı örnekler Tablo 3.17'de verilmiştir.

Tablo 3.16: Dil çevirisi test veri set kaynakları

Yazar	**Eser Adı**	**Cümle**
Katib Çelebi	Biyografi	37
Koçi Bey	Risale (Devlet işleri hk.)	70
Ali Rıza Bey	Üsküdar ve Boğaziçinde Kır Gezileri	68

Tablo 3.17: Dil çevirisi veri test seti örnekleri

Osmanlıca	**Türkçe**
Abdullah Efendi ulûm-ı akliyye vü nakliyyede ehl idi. (Katib Çelebi)	Abdullah Efendi pozitif ve dini bilimlerde yetenekli idi
İlm ü diyâneti olıcak şâb ise de kayırmaz. (Koçi Bey)	Bilgili ve ahlâklı olacak genç ise de önemli değil.
1284 târîhlerinde o civârda bir de belediye bahçesi tanzîm ü güşâd edildi.	1284 tarihlerinde o yakın çevrede bir de belediye bahçesi düzenlendi ve açıldı

4. BULGULAR

4.1. OSMANLICA OCR

Bu bölümde Osmanlıca OCR test veri seti üzerinde alınan sonuçlardan bahsedilecektir.

4.1.1. Doğruluk Oranı Türleri

OCR performansını değerlendirebilmek için doğru metin ve hesaplanan OCR metinlerini karşılaştırılarak sonuçların yorumlanması gerekmektedir. Bunun için ISRI Analytics Tool, Eval-tools (Tesseract), hOCR vb. OCR ölçüm ve değerlendirme araçları geliştirilmiştir. Bu araçlar genel olarak Latin OCR çıktılarını analiz etmek için geliştirildiğinden boşluk normalizasyonu, özel karakterlerin ele alınması, Arapça harflerin normalizasyonu ve hareke temizleme gibi ön işlemlerden dolayı Osmanlıca için doğrudan kullanılamamaktadır. Bundan dolayı bu çalışmada OCR çıktılarını karşılaştırmak için difflib kullanılmıştır.

OCR sonuçlarını değerlendirmek için Tablo 4.1'da ham, boşluksuz ve normalize olmak üzere 3 farklı doğruluk oranı verilmiştir. Doğruluk oranı paylaşılan test.py Python programıyla difflib kütüphanesinin SequenceMatcher.ratio() fonksiyonu kullanılarak hesaplanmıştır:

*doğruluk = 2.0*M/T* formülünde T doğru ve hesaplanan (OCR çıktısı) metinlerin toplam karakter sayısını, M eşleşen karakterlerin sayısını ifade eder. Eğer metinler birebir aynıysa doğruluk 1.0, eğer eşleşen karakter yoksa doğruluk 0.0 olarak hesaplanır. Örneğin burada doğru metin bu satır "سز خارجیه دن بیله جکسکز؛ بلغارلر متارکه یاپدی ، یعنی" olsun. OCR çıktısıda yani hesaplanan metinde "سز حارجیه دن بیله جکسکز؛ بلغارلر متارکه یاپدی ، یعنی" bu satır olsun. Bu cümlede toplam 43 karakter –boşluklar hariç– bulunmaktadır. Bir karakter (خ→ح dönüşmüştür) hariç diğer karakterlerin tümü eşleşmektedir. Doğruluk oranı hesaplamak için doğruluk oranı=2*42/86 → 0,97 çıkmaktadır. Yüzdelik olarak bu satırın doğruluk oranı % 97 olmaktadır.

Tablo 4.1: Test veri seti karakter doğruluk oranları

Araç/Model	Ham %	Normalize %	Bitişik %
Osmanlica.com Sentetik	73,16	77,64	78,10
Abby FineReader	71,98	80,19	81,05
Tesseract Arapça	76,92	82,37	83,27
Tesseract Farsça	75,30	83,85	84,48
Miletos	75,76	86,46	86,88
Google Docs	83,86	92,02	92,63
Osmanlica.com Orijinal	87,73	94,87	96,16
Osmanlica.com Hibrit	88,86	96,12	97,37
Ortalama	79,20	86,69	86,97

Test veri setinde en sık geçen karakterler, en çok hatalı olan harfler ve en çok hatalı karaterler Tablo 4.2'de verilmiştir.

Tablo 4.2: En sık geçen harfler, En çok hatalı olan harfler, En çok hatalı karakterler

En sık geçen harfler		En çok hatalı olan harfler		En çok hatalı karakterler	
Karakter	Sayı	Karakter	Sayı	Karakter	Hata %
boşluk	3959	boşluk	636	؛ ؛؛؛	53,8
ا ااا	1979	ي+ى	83	، ،،،	29,0
ي+ى	1793	ن نن	53	گ گگگ	25,0
ر رر	1380	ه+ھ	49	ث ثث	20,0
ل لل	1214	ب بب	40	- ---	17,6
و وو	1192	و وو	31	Space	16,1
ن نن	1136	ت تت	31	پ پپ	9,8
ه+ھ	1013	ا ااا	29	ب بب	6,2
م مم	925	- ---	29	) (((	5,9
د دد	911	، ،،،	29	ت تت	5,8

Ham doğruluk oranı OCR çıktısı metnin herhangi bir normalizasyon uygulanmadan doğru veriyle (ground truth) karşılaştırılmasıyla hesaplanan orandır. Ham doğruluk oranları incelendiğinde karakter doğruluk oranlarının %72 ile %89 arasında değiştiği görülmektedir.

Google Docs ve Osmanlica.com'un belirgin şekilde diğerlerinden daha başarılı OCR yaptığı âşikardır.

Normalize doğruluk oranı OCR çıktısındaki (i) çoklu boşluk karakterlerinin (whitespace) tek boşluk karakteriyle değiştirilmesi, (ii) okutucu ve normal he harflerinin beraber sayılması (tekilleştirilmesi) ve parantezlerin düzeltilmesi (parantezler BIDI algoritmasına göre tarafsız/neutral karakterler olduğundan RTL metin içinde düzgün şekilde görüntülenmemekte ve parantez yönü değişmektedir), ve (iii) standart dışı karakterlerin standart karakterle değiştirilerek, başka bir ifadeyle şekilleri (grapheme) aynı fakat Unicode kodları farklı olan karakterlerin (allograph) tekilleştirilip peşi sıra harekelerin (diacritics) temizlenmesiyle normalize edilmesini sonrası hesaplanan değerdir. Bu üç işlem test.py programında *normalizeWhitespace*, *normalizeSpecialCases*, ve *normalizeLetters* değişkenleriyle ifade edilmiştir. test.py'de harflerin normalizasyonu için *normalize_arabic()* fonksiyonu, harakelerin temizlenmesi için *remove_accents()* fonksiyonu kullanılmıştır. Normalizasyonda normal he (ه) ile okutucu he (ہ) harfleri Unicode kodları farklı olmasına karşın herhangi bir harfle birleşmezlerse yazılışları aynı olduğundan birlikte sayılmışlardır (tekilleştirme). Çünkü OCR programları çıktı üretirken bu iki harfi Unicode kodu yönünden ayırt edememektedir. Harf normalizasyona örnek olarak noktalı ye (U+064A, yeh, ي) ve noktasız ye (U+0649, Alef Maksura, ى) harfleriyle, Arapçadaki nokta (U+06D4, RTL, ARABIC FULL STOP, ۔) ve Latindeki nokta (U+002E, LTR, -FULL STOP) karakterleri verilebilir. Görüldüğü gibi nokta gibi birbirleriyle karıştırılan Arapça ve Latin karakterler extra bir problem olarak metin akış yönünü (RTL/LTR) dinamik olarak değiştirmekte, böylece metin editörü içinde metnin değiştirilmesi (editing) ve görüntülenmesinde (display) sıkıntıya sebep olabilmektedir. Tablo 4.1'de verilen normalize doğruluk oranları %80 ile %96 arasında değişmekte olup ham oranlardan yaklaşık %8 daha yüksektir. Yakından incelediğinde normalize oranların ham oranlara göre gerçek oranlara daha yakın olduğu anlaşılmaktadır. Çünkü ham OCR çıktısı yukarıda bahsedilen üç tür hatayı içermektedir ve normalizasyonla bu hatalar düzeltilmektedir.

Boşluksuz doğruluk oranı, metindeki tüm boşlukların silinerek kelimelerin birleştirilmesiyle oluşan boşluksuz uzun bir karakter dizisinin (string) yukarıdaki gibi normalize edilerek hesaplanan orandır. Boşluksuz metin ile doğruluk oranı hesaplanmasının amacı Osmanlıca'daki kelime bölütleme (word segmentation) probleminden bağımsız olarak doğruluk oranını görebilmektir. Çünkü kelime içi ve kelimeler arası boşluklar birbiriyle sıkça

karıştırılmakta ve doğruluk oranı hesaplanırken sonucu olumsuz etkilemektedir. Boşluksuz doğruluk oranı yönünden araçlar arasındaki sıralama değişiklik göstermemektedir: Tesseract Farsça (%84), Finereader (%81), Tesseract Arapça (%83), Miletos (%87), Google Docs (%93), Osmanlica.com (%97). Boşluksuz doğruluk oranıyla %97'nin üzerinde doğruluk oranı hesaplanmaktadır. Boşluksuz doğruluk oranında normalize oranlara göre en yüksek iyileşme yaklaşık %1 ile Google Docs ve Osmanlica.com'da gözlenmiştir.

Sonuç olarak Osmanlica.com'un %88,86 (ham), %96,12 (normalize) ve %97,37 (boşluksuz) doğruluk oranlarıyla en yakın alternatif modelden yaklaşık %4 daha iyi sonuç verdiği görülmüştür.

4.1.2. Karakter Hataları

Karışıklık matrisi ve hata dağılımları bir Python programı kullanılarak hesaplanmıştır. Karışıklık matrisi Şekil 4.1'de verilmiş olup, ayrıca hata miktar ve oranları Tablo 4.3'de özetlenmiştir. Tabloda karakterler (i) Arapça harfler (ابتثجحخدذرزسشصضطظعغفقكلمنهوى), (ii) Farsça+Türkçe harfler (ﮒگچپژۀء), (iii) imla işaretleri (():.؟؛،-) ve sayı (٠١٢٣٤٥٦٧٨٩٢) olarak gruplanmıştır. Tabloda her karakter için karakterin müstakil yazılışı, Unicode değeri, hatalı karakter sayısı ve hata oranı, toplam karakter sayısı ve test veri setinde bulunma oranı ve hatalı tanınan karakterlerin dağılımları ayrı sütunlarda verilmiştir. Hata dağılımlarında hatalı tanınan karakter ve parantez içinde hata sıklığı verilmiştir.

Tablo 4.3: Osmanlica.com test veri seti hata dağılımı

Harf	Kod	Hata %	Sıklık %	Hata Dağılımları
ا	0627	0,006 (12)	0,134 (1984)	ل(1) ن(1) ه(2) و(1) -(3) (1) (3)١(
ب	0628	0,015 (9)	0,041 (610)	د(1) ن(1) پ(7)
ت	062a	0,027 (14)	0,035 (518)	ب(1) د(1) ش(1) ق(2) ل(1) ن(7) ه(1)
ث	062b	0,235 (8)	0,002 (34)	ت(3) ك(1) ن(4)
ج	062c	0,024 (4)	0,011 (168)	ح(1) چ(3)
ح	062d	0,021 (4)	0,013 (189)	خ(4)
خ	062e	0,030 (3)	0,007 (101)	ح(3)
د	062f	0,006 (5)	0,061 (906)	س(2) ط(3)
ذ	0630	0,000 (0)	0,002 (36)	
ر	0631	0,001 (2)	0,093 (1373)	ز(2)
ز	0632	0,022 (5)	0,015 (225)	ث(1) ر(2) ك(1) ن(1)
س	0633	0,000 (0)	0,027 (403)	
ش	0634	0,008 (2)	0,018 (260)	ث(1) س(1)
ص	0635	0,008 (1)	0,009 (130)	س(1)
ض	0636	0,017 (1)	0,004 (59)	ص(1)
ط	0637	0,000 (0)	0,007 (103)	
ظ	0638	0,000 (0)	0,003 (49)	
ع	0639	0,007 (2)	0,02 (299)	ص(1) ق(1)
غ	063a	0,018 (2)	0,007 (110)	ف(1) ن(1)
ف	0641	0,026 (6)	0,016 (235)	ق(5) ن(1)
ق	0642	0,013 (6)	0,03 (445)	ت(1) ج(1) ف(4)
ك	0643	0,003 (2)	0,041 (609)	ل(2)
ل	0644	0,007 (9)	0,081 (1203)	ا(5) ب(1) د(1) ن(2)
م	0645	0,003 (3)	0,061 (907)	ص(1) ع(1)
ن	0646	0,025 (28)	0,074 (1102)	ه(1) ب(2)ت(14)ث(1)ح(1)د(1)ر(1)س(1)ش(3)ك(1)ل(1)م(1
هھ	0647/d5	0,022 (22)	0,066 (982)	و(1)(
و	0648	0,005 (6)	0,079 (1167)	ا(17) ت(1) ن(1) -(2) ،(1))
ي‌ى	0649/A4	0,000 (0)	0,000 (0)	د(1) ر(4) ق(1)
ء	0621	0,000 (0)	0,000 (3)	
ڭ	06ad	0,258 (17)	0,004 (66)	ك(17)
گ	06af	0,217 (5)	0,002 (23)	ك(4) ڭ(1)
چ	0686	0,048 (4)	0,006 (84)	ج(3) ح(1)
ژ	0698	0,000 (0)	0,000 (1)	
پ	067e	0,026 (1)	0,003 (38)	ب(1)

- .	06d4/46	0,049 (7)	(142) 0,010	(5)ق (1)٠ (1)٣
،	060c	0,217(20)	(92) 0,006	(19)- (1)چ
؛	061b	0,400 (4)	(10) 0,001	(2): (1)؟ (1)،
؟	061f	0,000 (0)	(17) 0,001	
:	003a	0,000 (0)	(14) 0,001	
-	002d	0,000 (0)	(0) 0,000	
)	0028	0,000 (0)	(0) 0,000	
(	0029	0,000 (0)	(94) 0,006	RTL metinde nötr parantezin türü(yönü) yanlış
٠	0660	0,400 (2)	(5) 0,000	belirleniyor.
١	0661	0,111 (1)	(9) 0,001	(2)-
٢	0662	0,000 (0)	(0) 0,000	(1)٦
٣	0663	0,000 (0)	(5) 0,000	
٤	0664	0,000 (0)	(3) 0,000	
٥	0665	0,143 (1)	(7) 0,000	
٦	0666	0,000 (0)	(3) 0,000	(1)ط
٧	0667	0,000 (0)	(0) 0,000	
٨	0668	0,000 (0)	(5) 0,000	
٩	0669	0,000 (0)	(0) 0,000	

	ا	ب	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق	ك	ل	م	ن	ه	ھ	و	ی	ء	ك	گ	چ	ژ	پ	D	I	Count	Corr	Errs	Err%
ا	1948	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	0	2	1	0	0	0	0	0	0	0	0	16	5	1979	1948	31	2
ب	1	605	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	6	7	6	644	605	39	6
ت	0	0	502	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	2	0	1	0	8	0	1	0	3	0	0	0	0	0	0	5	4	530	502	28	5
ث	0	0	2	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	1	0	0	0	0	0	0	0	0	0	0	35	28	7	20
ج	0	0	0	0	165	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	4	0	0	0	0	171	165	6	4
ح	0	0	0	0	0	185	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	189	185	4	2
خ	0	0	0	0	0	2	97	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	101	97	4	4
د	0	0	0	0	0	0	0	901	0	0	0	2	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	2	911	901	10	1
ذ	0	0	0	0	0	0	0	0	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	37	36	1	3
ر	0	0	0	0	0	0	0	0	0	1370	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	4	4	1381	1370	11	1
ز	0	0	0	1	0	0	0	0	0	1	221	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	225	221	4	2
س	0	0	0	0	0	0	0	0	0	0	0	403	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	7	7	418	403	15	4
ش	0	0	0	1	0	0	1	0	0	0	0	1	257	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	263	257	6	2
ص	0	0	0	0	0	0	0	0	0	0	0	1	0	129	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	130	129	1	1
ض	0	0	0	0	0	0	0	0	0	0	0	0	0	1	58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	59	58	1	2
ط	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	103	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	103	103	0	0
ظ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	49	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	50	49	1	2
ع	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	298	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	300	298	2	1
غ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	107	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	110	107	3	3
ف	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	228	4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	2	236	228	8	3
ق	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	440	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	2	453	440	13	3
ك	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	623	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	627	623	4	1
ل	5	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	2	0	0	1190	0	2	0	1	0	1	0	0	0	0	0	0	5	7	1216	1190	26	2
م	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	904	0	0	0	0	2	0	0	0	0	0	0	9	7	924	904	20	2
ن	0	2	13	1	0	1	1	1	0	1	0	1	2	0	0	0	0	0	0	0	0	1	1	1	1082	0	0	1	8	0	0	0	0	0	0	7	10	1135	1082	53	5
ه	14	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	2	257	0	2	0	0	0	0	0	0	14	0	301	2	299	99
ھ	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	698	0	1	0	0	0	0	0	0	6	8	714	698	16	2
و	0	0	0	0	0	0	0	1	0	4	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1161	3	0	0	0	0	0	0	9	11	1190	1161	29	2
ی	1	5	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	1	1	618	0	0	0	0	0	0	12	31	679	618	61	9
ء	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	1	4	3	1	25
ك	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	49	0	0	0	0	1	0	51	49	2	4
گ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	1	18	0	0	0	1	0	24	18	6	25
چ	0	0	0	0	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	81	0	0	0	0	86	81	5	6
ژ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0
پ	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	37	0	1	41	37	4	10

Şekil 4.1: Karışıklık matrisi

Tablo 4.4'te karakter hata oranlarını model/araçlar bazında karşılaştırılması verilmiştir.

Tablo 4.4: Karakter hata oranları model/araçlar bazında karşılaştırılması (%)

Karakter	Orijinal	Hibrit	Go.Docs	AbbyFine	Tess.Arab	Tess.Pers	Miletos	En kötü	En iyi
ا	2,7	1,5	3,4	9,3	6,5	4,5	4,8	Abby	Hibrit
ب	7,1	6,2	13,7	46,0	40,4	43,9	21,3	Tess.Pers	Hibrit
ت	7,5	5,8	12,6	30,8	25,7	34,8	21,9	Tess.Pers	Hibrit
ث	35,1	20,0	34,3	76,3	38,9	62,5	83,3	Abby	Hibrit
ج	9,9	2,9	4,7	15,0	33,3	32,8	65,7	Miletos	Hibrit
ح	3,6	2,1	4,3	15,8	13,7	14,9	25,0	Miletos	Hibrit
خ	2,0	2,0	7,7	18,9	26,9	18,4	16,0	Tess.Arab	Orijinal
د	2,1	1,1	2,9	8,0	6,5	3,8	9,3	Miletos	Hibrit
ذ	10,8	2,7	0,0	31,1	24,4	18,8	75,0	Miletos	Hibrit
ر	2,2	0,7	3,4	14,1	8,8	7,0	4,7	Abby	Hibrit
ز	7,1	1,8	16,0	24,0	28,3	30,6	14,7	Tess.Pers	Hibrit
س	4,8	3,6	8,0	30,8	19,1	13,4	25,8	Abby	Hibrit
ش	1,9	1,9	3,8	31,4	20,6	16,5	20,0	Abby	Hibrit
ص	3,0	0,8	3,8	40,2	9,8	7,3	39,1	Abby	Hibrit
ض	3,4	1,7	6,8	11,9	8,5	13,0	33,3	Miletos	Hibrit
ط	2,9	0,0	5,8	18,3	5,8	7,7	13,0	Abby	Hibrit
ظ	0,0	2,0	4,1	16,7	10,2	13,7	35,7	Miletos	Orijinal
ع	5,4	0,7	3,7	29,1	15,6	19,1	22,4	Abby	Hibrit
غ	6,4	2,7	23,9	32,1	43,9	48,4	51,9	Tess.Pers	Hibrit
ف	5,5	2,6	3,4	23,4	19,1	17,5	27,9	Miletos	Hibrit
ق	2,0	3,1	8,5	27,5	18,1	19,5	28,7	Miletos	Orijinal
ك	1,1	0,8	5,6	14,8	12,9	14,7	9,2	Abby	Hibrit
ل	3,5	1,7	10,0	21,1	30,3	24,8	14,9	Tess.Arab	Hibrit
م	5,3	2,3	9,0	33,3	22,9	24,9	25,1	Abby	Hibrit
ن	5,5	4,7	10,0	28,4	34,4	22,0	14,8	Tess.Arab	Hibrit
ه+ھ	8,1	4,8	5,2	26,1	21,1	9,3	12,0	Abby	Hibrit
و	3,8	2,6	3,2	10,3	5,1	3,9	5,4	Abby	Hibrit
ي+ى	5,8	4,6	15,9	30,5	34,6	39,8	19,3	Tess.Pers	Hibrit
ء	66,7	50,0	25,0	75,0	57,1	33,3	-	Abby	Go.Docs
ڭ	36,7	3,9	100,0	100,0	100,0	100,0	100,0	Çoğu	Hibrit
گ	54,2	25,0	29,2	100,0	100,0	39,3	-	Çoğu	Hibrit
چ	7,2	5,8	50,6	100,0	100,0	71,6	20,0	Çoğu	Hibrit

ژ	100,0	0,0	0,0	100,0	100,0	50,0	-	Çoğu	Çoğu
پ	19,5	9,8	54,8	100,0	100,0	89,2	84,6	Çoğu	Hibrit
-+.	18,8	17,6	18,3	35,0	31,4	41,4	18,9	Tess.Pers	Hibrit
،	29,4	29,0	23,5	41,1	98,9	100,0	12,1	Tess.Pers	Miletos
؛	50,0	53,8	28,6	55,0	37,5	35,7	-	Abby	Tess.Pers
؟	29,4	0,0	5,9	35,3	14,3	0,0	100,0	Miletos	Çoğu
:	21,4	0,0	0,0	50,0	29,4	18,8	0,0	Abby	Çoğu
-	100,0	100,0	100,0	100,0	100,0	100,0	100,0	Çoğu	Çoğu
Space	16,8	16,1	18,1	24,2	16,4	16,8	22,9	Miletos	Hibrit
(+)	4,0	5,9	6,9	16,0	9,1	11,7	15,2	Miletos	Original
٠	60,0	40,0	66,7	87,0	70,0	75,0	-	Tess.Pers	Hibrit
١	0,0	11,1	33,3	76,5	83,3	50,0	100,0	Miletos	Orijinal
٢	22,2	25,0	16,7	55,6	100,0	0,0	100,0	Çoğu	Tess.Pers
٣	40,0	0,0	40,0	40,0	100,0	0,0	-	Tess.Arab	Çoğu
٤	25,0	0,0	100,0	66,7	100,0	100,0	-	Çoğu	Hibrit
٥	37,5	28,6	100,0	60,0	100,0	75,0	-	Çoğu	Hibrit
٦	20,0	40,0	100,0	71,4	100,0	100,0	-	Çoğu	Orijinal
٧	0,0	0,0	25,0	100,0	100,0	20,0	0,0	Çoğu	Çoğu
٨	28,6	16,7	33,3	83,3	100,0	20,0	-	Tess.Arab	Hibrit
٩	0,0	0,0	0,0	100,0	0,0	50,0	100,0	Çoğu	Çoğu

Aşağıdaki OCR'daki tanıma hatalarına test veri setinden bir örnek verilmiştir. Bu örnekte doğru ve OCR'da hesaplanan metin müstakil, bitişik ve biçimlendirilmiş olarak paralel verilmiştir. İşlem satırında ↕ değişim, ↓ ekleme, ↑ silme ve * silme/ekleme konumu anlamına gelmektedir.

```
قابينك اوكنده ، سرى ايله قارشى قارشو*يه
  ↓↕          ↑↑          ↕
قاپينك اوكنده** سرى ايله قارشى يه
```

En büyük 10 OCR hatası ا ve ه harflerinin 23'er kez tanınmaması (deletion), 19 kez ،'in .(nokta) olarak tanınması (substitution), 17 kez ا'in ه olarak tanınması (substitution), 17 kez ڭ'in ك olarak tanınması (substitution), 17 kez ن harfi oluşturulması (insertion), 16 kez ،'ın tanınmaması (deletion), 14 kez ن'un ت olarak tanınması (substitution), 14 kez . (nokta)'nın tanınmaması (deletion), 13 kez ا harfi oluşturulması (insertion) olarak bulunmuştur.

En hatalı 10 karakter olarak ن 28 kez, ه 22 kez, ، 20 kez, ڭ 17 kez, ت 14 kez, ا 12 kez, ب 9 kez, ث 8 kez hata ile başta gelmektedir. Yakından incelendiğinde ن harfinin 14 kez ت harfiyle, ه harfinin 17 kez ا ile, ،'ün 19 kez .(nokta) ile, ڭ'in 17 kez ك ile, ت'nin 7 kez ن ile, ا'in 3 kez) ile, ب'nin 7 kez پ ile, ل'ın 5 kez ا ile, ث'nin 4 kez ن ile değiştirildiği görülmüştür.

Silme (deletion): En çok tanınmayan 10 karakter (dokümanda geçtiği halde OCR çıktısında olmayan) olarak 23 kez ا, 23 kez ه (okutucu he), 16 kez ،, 14 kez . (nokta), 10 kez ب, 10 kez م, 9 kez ن, 9 kez و, 8 kez ت, 10 kez ل görülmüştür.

Ekleme (insertion): Metinde geçmediği halde OCR'da en çok oluşturulan (insertion) 10 karakter olarak 17 kez ن, 13 kez ا, 13 kez و, 10 kez . (nokta), 8 kez ب, 8 kez م, 7 kez س, 7 kez ه (okutucu he), 5 kez ت, 4 kez ل gözlenmiştir.

Yer değiştirme (değişim, substitution): Birbirleriyle en çok karıştırılan 10 karakter olarak ، 19 kez . (nokta), ه 17 kez ا, ڭ 17 kez ك, ن 14 kez ت, ب 7 kez پ, ت 7 kez ن, ف 5 kez ق, ل 5 kez ا, . (nokta) 5 kez ق, ث 4 kez ن, ح 4 kez خ olarak tanınmıştır. Görüldüğü gibi şekil bakımından benzer ve noktasız halleri birbiriyle aynı harfler OCR'da birbirine karışmaktadır. Diğer taraftan şekil olarak benzemeseler de birbirleriyle karıştırılan karakterlerin karıştırılma sebeplerini anlayabilmek için söz konusu işlemin oluştuğu yerin görsel olarak yakından incelenmesi gereklidir.

Yukarıda sıklıkları en çok karşılaşılan OCR hataları, en hatalı 10 karakter ve en çok karşılaşılan ekleme, silme ve yer değiştirme hataları verilmiştir. OCR doğruluk oranını iyileştirmek için yapılacak çalışmalarda yukardaki hataların doküman içindeki konumunda (yerinde) analizinin yapılarak hata nedenlerinin araştırılması gereklidir. Küçük bir veri setindeki doküman görüntülerinde hatanın yerini bulmak çok zaman almasa da, büyük bir veri setinde yüzlerce hatta binlerce hatayı dokümanlarda manuel olarak bulmak güç ve zaman alıcı bir iştir. Bu yüzden doğrudan karmaşıklık matrisindeki bir hücrede gösterilen hata sayısının üzerine tıklayarak söz konusu hatanın oluştuğu dokümanlardaki tüm hataları listeleyip görüntüleyecek bir OCR hata görüntüleme aracına ihtiyaç vardır.

Karakter türüne göre normalizasyon sonrası karakter hata oranları Tablo 4.4'da verilmiştir. Her ne kadar sayı ve oran olarak en yüksek hata Arapça'da karakterlerde görülse de, Farsça ve Türkçe hata oranlarının göreceli olarak yüksek olduğu anlaşılmaktadır. Tablo 4.5'te harf

kelimedeki konumuna göre sıklık ve hata oranı verilmiştir. Tablo 4.6'da harf katar konumuna göre sıklık ve hata oranı verilmiştir. Tablo 4.7'de normalize metin üzerinde harf türüne göre grup içi ve genel karakter hataları verilmiştir.

Tablo 4.5: Harf kelimedeki konumuna göre hata oranı

Konum	Sıklık	Oran %
1 harf	30	1,12
2 harf	189	7,08
3 harf	470	17,6
4 harf	407	15,24
5 harf	459	17,19
6 harf	377	14,12
7 harf	288	10,79
8 harf	205	7,68
9 harf	120	4,49
10 harf	78	2,92
11 harf	26	0,97
12 harf	10	0,37
13 harf	6	0,22
14 harf	2	0,07
15 harf	1	0,03
16 harf	1	0,03

Tablo 4.6: Katarların konumuna göre hata oranları

Konum	Sıklık	Oran %
1 katar	29	1,43
2 katar	789	39,5
3 katar	558	27,6
4 katar	358	17,7
5 katar	174	8,61
6 katar	66	3,26
7 katar	29	1,43
8 katar	5	0,24
9 katar	2	0,09
10 katar	1	0,04

Tablo 4.7: Harf türüne göre grup içi ve genel karakter hataları (normalize metin, %)

	Orijinal		Sentetik		Hibrit		Google D.		Abby F.R.		T. Arapça		T. Farsça		Miletos	
Türü	**Grup**	**Genel**	**Grup**	**Genel**	**Grup**	**Genel**	**Grup**	**Genel**	**Grup**	**Genel**	**Grup**	**Genel**	**Grup**	**Genel**	**Grup**	**Genel**
Arapça	6,0	60	27,6	78	4,7	59	9,8	67	25,2	76	28,3	78	21,7	79	18,5	77
Diğer	22,8	3	44,4	2	8,4	1	57,8	5	100	3	100	3	74,9	3	29,0	1
Noktalı	6,7	22	38,4	35	5,0	20	14,3	34	40,4	38	40,7	36	34,5	41	29,6	36
Noktasız	6,0	42	22,7	44	4,6	39	8,2	37	19,5	41	23,8	45	16,6	41	14,0	41
Nokta üstte	5,7	12	34,1	21	4,0	11	11,0	16	28,4	18	34,9	20	26,7	21	21,5	19
Nokta altta	8,6	9	46,9	14	7,1	10	19,2	18	66,4	20	52	15	49,8	20	50,7	17
1 nokta	6,0	11	37,4	19	4,3	10	9,7	12	28,8	16	38,1	19	30,5	20	20,4	15
2 nokta	6,7	7	39,3	12	6,1	8	17,3	17	55,6	17	41,2	13	38,4	16	48,5	18
3 nokta	10,2	3	41,2	3	5,0	2	27,2	6	58,8	5	54,2	4	43,5	5	25,7	3
Harf	6,2	63	27,7	79	4,7	60	10,3	71	26,0	79	29,1	81	22,3	83	18,6	78
Diğer	19,3	4	44,4	3	16,2	4	22,4	3	46,6	3	59,9	3	33,1	2	25,6	3

4.1.3. Kelime Hataları

Veri setindeki kelimelerin karakter sayısına göre normalizasyon sonrası dağılımları Tablo 4.10'de verilmiştir. Metinde en çok 4 harfli (%18), 5 harfli (%18), 6 harfli (%13), 3 harfli (%13) kelimeler geçmektedir. Bir harfli kelimeler genellikle noktalama işaretleri ve yanlışlıkla ayrı yazılmış harflerden oluşmaktadır. Burada doğru kabul edilen metinde -ground truth- az sayıda da olsa hatalar olduğu anlaşılmaktadır. Fakat bu hatalar deney sonuçlarında gözle görülür bir etkiye sahip değildir. Osmanlıca eklemeli bir dil (agglutinative) olduğu için metinde on harften daha uzun kelimelere rastlanmaktadır.

Kelime sıklıkları incelendiğinde -noktalama işaretleri hariç- metinde en sık geçen kelimeler arasında 17 kez (بر), 10 kez (او), 5 kez (اسه), 5 kez (الله), 4 kez (اولان), 4 kez (مسلمانلرك), 4 kez (اىدن), 4 kez (دىن), 4 kez (بو) kelimelerinin geçtiği görülmektedir. Beklendiği gibi Türkçe'nin çok sık geçen kelimelerinin (stopwords) Osmanlıca'da da aynı şekilde sıkça geçtiği gözlenmektedir.

Kelime tanıma oranları OCR çıktıları ile doğru metin karşılaştırılarak Kelime tanıma oranları OCR çıktıları ile doğru metin karşılaştırılarak holianh.github.io/portfolio/Cach-tinh-WER adresindeki python programıyla hesaplanmış ve sonuçlar Tablo 4.8'de özetlenmiştir. Tabloda OCR işleminde doğru tanınan kelime sayısı, değiştirilen (substitution) kelime sayısı, eklenen

(insertion) kelime sayısı ve silinen (deletion) kelime sayıları ve hata oranı ayrı sütunlar şeklinde verilmiştir. Kelime tanıma hata oranları, ham çıktı ve harf normalizasyonlu çıktı kullanılarak iki defa hesaplanmıştır. Kelime hata oranları genelde karakter hata oranlarına göre çok daha yüksek çıkmaktadır. Bunun sebeplerinden birisi kelime sıklığının harf sıklığına göre ortalama 5 kat daha küçük olmasıdır. Tablo 4.9'da kelime uzunluğa göre hata oranları verilmiştir.

Tablo 4.8: Kelime hata oranları

Araç/Model	Ham					Normalize				
	Hata %	Doğru	Değiştirme	Ekleme	Silme	Hata %	Doğru	Değiştirme	Ekleme	Silme
Tesseract Arapça	85	712	2723	184	136	78	979	2433	204	158
FineReader	84	665	2542	105	364	80	799	2396	117	375
Tesseract Farsça	87	555	2540	116	476	80	856	2207	147	507
Miletos	86	115	523	2	170	72	224	412	172	4
Google Docs	78	1320	2196	538	55	62	1892	1602	560	76
Osmanlica Hibrit	64	1561	1896	298	114	42	2369	1066	320	135

Tablo 4.9: Kelime uzunlukları bazında hata oranları

Harfler	Sıklık	Oran %	Örnekler
1	307	08,59	: ! د ه ك ى س (ـ ؛ ؟) ، و
2	298	08,34	...حس بو كه عا هر او بر بش سن بك قز پك
3	463	12,96	... لكى هيچ بوق بنه لمى افر محل در۔ فقط دىن فرق
4	653	18,28	...شبهه موثر ارضك اىدن خسته لغك، منشا مرضك متحد سفىل اسه
5	649	18,17	...اولان عمومى بقاده اوقيا خلاصه اقطار مسكون اولان محكوم افراد
6	469	13,13	...عمومى واردر۔ اساده عمومنى استلا مشابهت شبههيه دنيانك طرفنده:
7	324	09,07	...بوسقوطك نوساده مادامكه اكلهن انسانيه دركهلره ارهسنده ايتمزمى
8	194	05,43	...اوروپاده اسلاملرى مسلمانلر ارهسندكى انقراضدن اضمحلاله اسلاملرك
9	129	03,61	...مسلمانلرك اولدىغنده براقمسورا! سوروكلهن اقلىملرده اولمشلردى مسلمانلرك
10	50	01,40	...مختلفهسنده اعراضندهكى مسلمانلردر چوقورلرينه ىووارلانان ترغيبلريله
11	19	00,53	...,مانفعتلريتى افوروزلرينه اوغرايهجقدر محاربهلرينه اىدىتوردى۔ بولنمدىغندن
12	8	00,22	اوچوروملرينه برلشدىرلمشدى ايلكلهمهكى،
13	7	00,19	

Ham (normalizasyonsuz) çıktı kullanılarak hesaplanan kelime hata oranlarının tamamı %50'den daha yüksek olduğu için kayda değer bir sonuç görülmemiştir. Bununla beraber en düşük hata oranları %78 ile Google Docs ve %64 ile Osmanlica.com modeli tarafından elde edilmiştir. Normalize edilmiş çıktı üzerinde hesaplanan hata oranlarında sadece Osmanlica.com %42 ile %50'nin altında hata oranına sahiptir. Bir sonraki en iyi oranla (Google Docs %62) bu oran arasında %20'lik büyük bir fark gözlenmiştir. Hatalar kendi içinde incelendiğinde en fazla değişim (%60-75) hatası, sonra ekleme (~%1-15) ve silme (~%1-15) hataları oluşmaktadır. Google Docs ve Osmanlica.com'da (~400, ~%10) ekleme hatasına diğer 4 araca (~200, ~%5) göre daha yüksek miktarda rastlanmıştır. Silme hatasında ise Google Docs ve Osmanlica.com'da (~75-100, ~%2-3) diğer 4 araca (~250, ~%7) göre daha düşük miktarda rastlanmıştır. Miletos'un ham kelime ekleme sıklığı (2) ve normalize kelime silme sıklığı (4) aşırı düşük çıkmıştır. Normalizasyonla silme hataları sıklık bakımından ekleme hatalarına (2 → 172, 170 → 4) dönüşmektedir. Ekleme ve silmede gerçek hata değerlerinin normalize metinde sonuçlara yansıdığı düşünülmektedir.

Tablo 4.10: Harf sayılarına göre kelime sıklıkları

Uzunluk	Sıklık	Oran %	Örnek
1 harf	307	08,59	: ! د ه ك ى س (- ؛ ؟) ، و
2 harf	298	08,34	...حس بو كه عا هر او بر بش سن بك قز پك
3 harf	463	12,96	... لكى هنچ بوق بنه لمى افر محل در- فقط دين فرق
4 harf	653	18,28	...شبهه موثر ارضك ايدن خسته لغك، منشا مرضك متحد سفيل ايسه
5 harf	649	18,17	...اولان عمومى بقاده اوقيا خلاصه اقطار مسكون اولان محكوم افراد
6 harf	469	13,13	...عمومى واردر- اساده عمومنى استيلا مشابهت شبهه دنيانك طرفنده:
7 harf	324	09,07	...بوسقوطك نوساده مادامكه ايكلهين انسانيه دركهلره ارهسنده ايتمزمى
8 harf	194	05,43	...اوروپاده اسلاملرى مسلمانلر ارهسندكى انقراضدن اضمحلاله اسلاملرك
9 harf	129	03,61	...مسلمانلرك اولدبغنده براقمسورا! سوروكلهين اقليملرده اولمشلردى مسلمانلرك
10 harf	50	01,40	...مختلفهسنده اعراضندهكى مسلمانلردر چوقورلرينه ووارلانان ترغيبلريله
11 harf	19	00,53	...,مانفعتلريتى افوروزلرينه اوغرايهجقدر محاربهلرينه ايدبستوردى- بولنمدبغندن
12 harf	8	00,22	اوچورومليرينه برليشدبرلمشدى ايليكلهمهكى،
13 harf	7	00,19	

4.2. ALFABE ÇEVİRİSİ

Osmanlıca Türkçe Alfabe çevirisi Şekil 3.7'de önerilen sisteme göre geliştirilmiştir. Tablo 3.15'te verilen test veri seti ile 2 farklı deney yapılmıştır. Birincisi sözlükte herhangi bir değişiklik yapılmadan alınan sonuçlardır. Şekil 4.2'de gösterilmiştir. Tabloda normal metinde geçen kelime sayısı ve doğruluk oranları, OCR metinde kelime sıklıkları ve doğruluk oranları verilmiştir. İkincisi ise veri setinde olan fakat sözlüğümüzde olmayan kelimelerin sözlüğe eklenerek elde edilen sonuçlardır. Şekil 4.3'te gösterilmiştir.

	Referans Metin		**OCR Metin**	
Problem	**Sıklık**	**Oran %**	**Sıklık**	**Oran %**
Alfabe Çevirisi	6465/7033	91,92	5964/6795	87,77
Kelime bölütleme	33/38	86,84	54/64	84,37
Yazım düzeltme	31/104	29,80	102/296	34,45
Seslendirme	120/257	46,69	90/276	32,60
Toplam	6649/7432	89,46	6210/7432	83,56

Şekil 4.2: Alfabe çevirisi ilk sonuçlar

	Referans Metin		**OCR Metin**	
Problem	**Sıklık**	**Oran %**	**Sıklık**	**Oran %**
Alfabe Çevirisi	7137/7261	98,29	6401/6905	92,70
Kelime bölümleme	0/17	100	58/63	90,47
Yazım düzeltme	37/43	86,04	122/276	44,20
Seslendirme	99/111	89,18	82/187	43,85
Toplam	7292/7432	98,08	6646/7432	89,42

Şekil 4.3: Sözlük düzenlemesinden sonraki sonuçlar

Osmanlıca - Türkçe Alfabe Çevirisi 6 aşamadan oluşuyor. Bu modülün başarısı 6 aşamanın başarısına bağlıdır. Bu modülde her aşamanın başarısı test edilmiştir. Toplam modülün ilk deneyde Osmanlıca gerçek metin üzerinde % 89.46 ve OCR metin üzerinde % 83.56 başarı elde edilmiştir. Yapılan 2.deneyde ise Osmanlıca gerçek metin üzerinde % 98.08 ve OCR metin üzerinde % 89.42 başarı elde edilmiştir.

Deney sonuçlarına bakıldığı zaman çeviri metnin büyük bir kısmı ortografik alfabe çevirisi aşamasında çözümlenmektedir. Birinci deneyde Alfabe Çevirisi aşamasında 7033 kelime çözümlenmiş 6465 kelimesi doğru çözümlenmiştir. Doğruluk oranı ise %91,92 olmaktadır. İkinci deneyde ise bu aşamada 7261 kelime çözümlenmiş ve 7137 kelime doğru çözüm elde edilmiştir. Doğruluk oranı ise %98,29 çıkmaktadır. Metinde bulunan kelimeler sözlükte varsa çeviri işlemi %98 üzerinde bir başarı elde edilir. Ayrıca yazım düzeltme ve seslendirme aşamaların iyileştirme yapılırsa başarı oranı artacaktır.

Osmanlıca – Türkçe Alfabe çevirisinde bazı özel durumlar aşağıda listelenmiştir. Bu özel durumlar başarı oranını doğrudan etkilemektedir.

1. Osmanlıca'da sonu (ة) ile biten kelimeler Türkçe (en) olarak okunabilir: حقيقة → hakikaten, مادةً →maddeten vb.
2. Osmanlıca da sonu (ا) ile biten kelimeler Türkçe (en) olarak okunabilir.حكما → hükmen, غالبًا→ galiben vb.
3. Osmanlıca'da bir kelimenin yazılış ve okunuş birbirinden farklı olabilir.قانبور → kanbur (kambur), پنجشنبه → pencşenbe (perşembe), هينجغيرەرق→ hıncğırarak (hıçkırarak) vb.
4. Osmanlıca iki kelime birleşip farklı kelime oluşabilir. *او قدر* → *o kadar* kelimesi birleştiği zaman *اوقدر*→ *oktur* şeklinde çeviri yapılmaktadır.
5. Osmanlıca da yabancı kelimelerde ek aldığında büyük ünlü uyumuna uymamaktadır. كوفيلره → kufilara (kufilere) vb.
6. Osmanlıca da bir kelimenin kökü muhafaza edilir. Fakat ek aldığında ses düşmesi meydana gelebilir. Osmanlıca kelimelerde kök her ne kadar muhafaza edilse de bazı dokümanlarda ek aldığında terk edilmiştir. ايلر (doğrusu:ايلەر) → eyler,سويلر (سويلەر)→ söyler vb.
7. Osmanlıca'da bir kelime parçalanıp anlamlı iki kelimeye dönüşebilir.كلمه سنك → kelime senin (kelimesinin), كلمه سندن → gelme senden (gelmesinden), مختلفه سنده → muhtelife sende (muhtelifesinde)
8. Okutucu ye harfi ek alımında kendisinden sonra gelen ye harfi ile birleşmez. قاپی يی → kapı ye (kapıya), صوجی يی → sucu ye (sucuyu) vb. Bu bir kelimenin iki kelime şekilde çözümlenmesine sebep olmaktadır.

Osmanlıca – Türkçe Alfabe çevirisinde yukarıda özel durumlar çözüldüğünde önerilen sistemin başarı oranı yükselecektir.

4.3. DİL ÇEVİRİSİ

Bu çalışmada ilk deneyler Tablo 3.16'de verilen 200 cümlelik test veri seti ile yapılmış olup sonuçlar incelenmiştir. Bu aşamada bir doğruluk oranı vermek için veri setinin yeterince büyük olmadığı düşünülmektedir. Daha büyük veri setleri ile test yaparak güvenilirliği ve doğruluk oranlarının hesaplanması hedeflenmektedir.

Test veri seti hazırlarken ve bu test setiyle çeviri deneyleri yapılırken karşılaştığımız problem ve zorlukları örnekleriyle paylaşılmıştır. Buradaki tartışmalar hem veri setinin hem de çeviri algoritmasının geliştirilmesi ve iyileştirilmesinde yönlendirici değere sahiptir.

1. *Yazılışı aynı kelimeler*: Osmanlıca ve Modern Türkçe'deki ortak ve yaygın kelimeler çeviriye sokulmadan doğrudan aktarılmalıdır. "*bir"* gibi bazı kelimelerin Osmanlıca'da bir karşılığı (*kuyu*) olduğu için yanlış çeviriye sebep olabilmektedir.
2. Osmanlıca'da bazı metinlerde anlatım ağdaki ve karmaşık olduğunda çeviri zorlaşmaktadır. Aşağıdaki örnekler bu tür örneklerdir:

   makūle mesâil-i hilâfiyyeyi pâ-bend-i ahmakān
   Niçe kīl ü kāl ve bahs ü cidâle müeddî oldu.
   Fenn-i kitâbet ü hesâb ü siyâkate müteallik müşkil
   Reca'nâ mine'l-cihâdi'l-asgar ile'l cihâdi'l-ekber (Arapça)
3. *Ayrı yazılan ekler:* Osmanlıca ve Türkçe bazı anlatım farklılıkları birebir çeviride anlatım veya ifade bozukluğuna sebep olabilmektedir. Örneğin Türkçe geçmiş zaman -dH eki Osmanlıca'da "idi" olarak ayrı yazılmaktadır. Örneğin *ehl idi* metni *yetenekli idi* yerine *yetenekliydi* diye çevrilmelidir.
4. *Bitişik yazılan ekler*: Normalde ayrı yazılması gerekirken birleşik yazılan ekler sözlükte arama ve eşleştirme yapmayı güçleştirmektedir. Örneğin *makūle mesâil-i hilâfiyyeyi pâ-bend-i ahmakā* ifadesindeki *hilâfiyyeyi* kelimesindeki *-i* ekinin *hilâfiyye-i* şeklinde ayrı yazılması gerekmektedir.
5. *Harf ikilemesi*: Türkçe olmayan bazı kelimelerdeki harf ikilemeleri Türk alfabesine aktarılırken bu ikilemeler bazen korunmakta bazen de tekilleştirilerek yazılmaktadır. Örneğin *hilâfiyyeyi* kelimesi *hilâfiyeyi* şeklinde de yazılabilmektedir.

6. *Küçük büyük harf ayırımı*: Osmanlıca'da küçük büyük harf ayrımı olmadığından özel isimlerin bulunması ve çevriminde problem yaşamaktadır. Örneğin *Selimiye* kelimesinin *sağlamıye* diye çevrilmesi ya da *murad rabi*'nin (*murad-ı râbi, 4. Murat*) *istek dört* gibi çevrimi söz konusu olabilir.
7. *Çoklu morfolojik çözüme sahip kelimeler*: Bir kelimenin birden fazla morfolojik çözümü yani birden fazla gövdesi söz konusu olduğunda en uzun gövde en kısa ek katarı yöntemi yetersiz kalmaktadır. Örneğin Osmanlıca *al* kelimesi *almak* ya da *aile* olarak Türkçe'ye aktarılabilir. Burada doğru karşılığı bulmak için çeviri sonrası dil modelleri (n-grams) kullanılarak doğru çözüme ulaşmaya çalışıyoruz. Benzer örnekler:
    - şehr → şehir (isim), ay (isim)
    - bak → bak (fiil), korku (isim)
    - azm → gayret (isim), kemik (isim)
8. *Çoklu yüzeye sahip kelimeler*: Osmanlıca'daki birçok kelimenin birden fazla imlası (yazım şekli, yüzeyi) olması sözlüğün hazırlanmasını ve çeviriyi güçleştirmektedir. Kelimelerin farklı yüzeylerinin sözlüğe işlenmesi ya da sözlükte arama yaparken kelime benzerliği üzerinde arama yapmak gerekebilir. Örneğin yukarıda *azm* kelimesinin *azim* olarak *elh* kelimesinin *elih* olarak yazımları daha yaygın kullanılmaktadır. Osmanlıca'nın Türk alfabesiyle yazımında ayırıcı ya da düzeltme işaretleri denen diakritiklerin doğru ve standart bir şekilde kullanımı çeviride çok önemli hale gelmektedir. *Reca'nâ mine'l-cihâdi'l-asgar ile'lcihâdi'l-ekber* örneğinde olduğu gibi kesme işareti (apostrof), kısa çizgi (tire) ve şapka gibi işaretler metin ön işlemesi ve normalizasyonunda ve sözlük geliştirmede dikkat edilmesi gereken konulardır. *bû* kelimesinin Türkçe'deki karşılığı *koku* kelimesidir. Çeviri uygulamasına verilen bir metinde bu kelimenin şapkasız yazılması durumunda Türkçe'ye işaret sıfatı ya da zamiri olarak yanlış çevrilmesi söz konusu olacaktır.
9. *Farklı dildeki alıntılar:* Metin içerisinde özellikle Arapça ve Farsça alıntılar Osmanlıca ile karıştırılabilmekte ve Osmanlıca zannedilebilmektedir. *Reca'nâ mine'l-cihâdi'l-asgar ile'l cihâdi'l-ekber* örneğinde olduğu gibi Arapça, Farsça ve Osmanlıca çok sayıda ortak kelimeye sahip olduğu için, yabancı dildeki alıntıları ana metinden ayırt etmek için yapılacak bir dil tanıma (language detection/recognition) işlemi zorlaşmaktadır. Bu örnekteki metin tamamen Arapça olmasına rağmen *cihad, ekber* ve *asgar* kelimeleri aynı zamanda Osmanlıca kelimelerdir.

Osmanlıca – Türkçe Dil çevirisi için temel bir çalışma yapılmıştır. Bununla ilgili çalışmalara devam edilecektir.

5. TARTIŞMA VE SONUÇ

Bu çalışmamızda Osmanlıca dokümanların (i) Osmanlıca OCR (ii) Alfabe çevirisi (iii) Dil çevirisi olmak üzere 3 adımda bir uçtan diğer uca yani dokümandan Türkçe metne otomatik olarak çevrilmesi gerçekleştirilmiştir. Bu çalışma Osmanlica.com sitesinde yayına alınmıştır.

Osmanlıca OCR modülünde Osmanlıca matbu nesih doküman resimlerini derin öğrenme modelleriyle metne dönüştüren web tabanlı bir optik karakter tanıma sistemi sunulmuştur. Geliştirilen OCR aracı Osmanlica.com adresinde kullanıma açılmıştır. Sistem sentetik ve orijinal verilerden oluşan bir veri kümesiyle eğitilip 21 sayfalık farklı bir veri setiyle test edilmiştir. Geliştirilen OCR modeli Tesseract'ın Arapça ve Farsça, Google Docs'ın Arapça, Finereader'ın Arapça ve Miletos'un Osmanlıca OCR modeliyle test veri seti kullanılarak karşılaştırılmıştır. Kelime tanımada Osmanlica.com OCR modeli %64 ham ve %42 normalize değerleriyle diğer araçlardan çok daha düşük bir hata yüzdesi vermektedir. Karakter tanımada ise Osmanlica.com OCR modeli %88,64 ham, %96,12 normalize ve %97,37 boşluksuz doğruluk oranıyla benzerlerinden belirgin şekilde daha yüksek bir performans sağlamıştır. Test veri seti ve test sonuçları bağımsız doğrulama yapılabilmesi için osmanlica.com/test adresinde paylaşılmıştır.

Karşılaştırma sonucunda OCR'da ortaya çıkan değişim, ekleme ve silme hataları özetlenerek paylaşılmasına rağmen sonuçların daha detaylı analizine ihtiyaç vardır. Bu amaçla geliştirdiğimiz yeni hata görüntüleme aracı ile söz konusu hataların doküman resimleri içinde oluştukları konumları görsel olarak analiz edilecek ve hata nedenleri hakkında daha anlamlı ve gerçekçi yorumlar yapılabilecektir. Amacımız bu yorumların sistemin doğruluk oranını iyileştirecek somut adımlara dönüştürülmesidir. Bu analizlerden elde edilecek bilgi ve tecrübenin rika, talik, divani gibi diğer Osmanlıca fontlarının OCR'ında değerli bir kaynak olacağı düşünülmektedir.

Arapça Tabanlı Osmanlıca Alfabesinden Latin Tabanlı Modern Türkçe'ye Alfabe Çeviri Sistemi (Ottoman-Turkish Transliteration System) tasarlanmış ve geliştirilmiştir. Alfabe çevirisi (i) Ortografik Alfabe çevirisi, (ii) kelime bölütleme (iii) yazım düzeltme (iv) seslendirme (v) kelime tahmini (n-grams) (vi) tamlama ve birleşik isimler olmak üzere 6 aşamadan oluşur. Alfabe çevirisi bu altı aşamanın başarısına bağlıdır. Alfabe çevirisi için 7500

kelimelik bir paralel test veri seti hazırlanmıştır. Test veri seti Osmanlıca, Türkçe ve OCR çıktısı olmak üzere paralel olacak şekilde hazırlanmıştır. Alfabe çevirisi test veri seti üzerindeki başarı oranı Osmanlıca kelime için % 98.08 OCR çıktısı için % 89 başarı elde edilmiştir. Alfabe çevirisi seslendirme ve yazım düzeltme aşamalarında doğruluk oranı düşük olduğu gözlenmiştir. Seslendirme ve Yazım düzeltme aşamaları iyileştirme yapılırsa başarı oranı daha da artacaktır.

Osmanlıca Türkçe Alfabe Çevirisi modülünde üretilen diğer bazı önemli çıktılar şöyle sıralayabiliriz; (i) Osmanlıca-Türkçe Kelime Çeviri Sözlüğü (ii) Osmanlıca-Türkçe Ek Çeviri Sözlüğü (iii) Osmanlıca derlem (n-gramlar için) (iv) Osmanlıca-Türkçe Kelime Çevirisi Test Veri Seti (v) Osmanlıca İsim Tamlamaları Sözlüğü çıktıları hazırlanmıştır. Bu çıktıların bazıların düzenlenmesine devam edilmektedir.

Osmanlıca - Türkçe dil çevirisi için kelime grubu tabanlı basit bir çeviri sistemi geliştirilmiştir. Bu yöntem dil içi çeviri için kullanılabilecek en basit yöntemlerden bir tanesidir. Yöntem kelime öbeği tabanlı ve mekanik olarak işleyen ve bir cümleyi öbek öbek soldan sağa doğru tarayarak Osmanlıca'dan Modern Türkçe'ye aktaran bir yöntemdir. Osmanlıca kelimeler morfolojik analizle gövde ve ek katarlarına ayrıştırılmakta, sözlük kullanılarak gövdelerin Türkçe'deki karşılıkları bulunmakta ve Türkçe gövdelere Osmanlıca kelimelerin ekleri bitiştirilerek çeviri tamamlanmaktadır. Bu yöntemde Osmanlıca morfolojik analizde birden fazla çözüm elde edilebilmekte, bulunan Osmanlıca gövdelerin sözlükte Türkçe karşılıkları da birden fazla olabilmektedir. Birinci durum morfolojik belirsizlik (morphological ambiguity) problemine ikinci durum ise kelime anlamı belirsizliği (word sense ambiguity) problemine yol açmaktadır. Zamanımız kısıtlı olduğundan morfolojik belirsizlik problemini en uzun gövde + en kısa ek katarını seçerek aşmayı, kelime anlamını belirsizliği problemini de dil modelleri (n-grams) kullanarak çözmeyi denenmiştir. Çeviri sözlüğü zaman kısıtından dolayı sadece 10 bin kelimelik bir Osmanlıca-Türkçe kelime hazırlanmıştır. Fakat sözlüğümüzün çok sayıda hata içerdiği testlerde gözlemlenmiştir. Sözlük üzerinde ileride kapsamlı ve detaylı bir çalışma yapılması ihtiyacı olduğu tespit edilmiştir.

Osmanlıca Türkçe Dil çevirisi ilk deneyler 200 cümlelik veri seti ile yapılmış olup sonuçlar incelenmektedir. Bu aşamada bir doğruluk oranı vermek için veri setinin yeterince büyük olmadığı düşünülmektedir. Daha büyük veri setleri ile test yaparak güvenilirliği iyi doğruluk oranlarının hesaplanması planlamaktadır. Bu testlerde elde ettiğimiz bazı bulguları yukarıda

paylaşılmıştır. Testlerde elde edilen bulgular bize çeviride hangi çeşit hatalarla karşılaşıldığı, bu hataların hangi sıklıkta ortaya çıktığı, bu hataların giderilmesi için izlenmesi gereken yol/yöntemin ne olması ve hataların giderilmesi için gereksinim duyulan sözlük ve derlem gibi kaynaklar hakkında bilgilendirici ve aydınlatıcı olmuştur. Burada elde ettiğimiz bilgi ve deneyimle ileride daha etkin ve daha doğru çeviri için yapılması gerekenleri bize kısmen göstermiş bulunmaktadır. Yapılan testlerin yeterli olmadığı açıktır. Daha fazla kaynak, daha fazla insan gücü ve daha fazla teste ihtiyaç olduğu açıktır. Buradaki tartışmalar hem veri setinin hem de çeviri algoritmasının geliştirilmesi ve iyileştirilmesinde yönlendirici değere sahiptir.

Geliştirilen çeviri sisteminin testler daha kaliteli ve daha büyük dil kaynağıyla testi ve iyileştirilmesi kendi başına ayrı bir projenin konusudur. Bu alanda derin öğrenme modelleri ve istatistiksel tabanlı çeviri yöntemleriyle yapılan bazı çalışmalara yukarıda referans verilmiştir.

Kelime öbekleri genellikle Osmanlıca tamlamalara karşılık gelmektedir. Osmanlıca Türkçe, Arapça, Farsça ve hibrit olmak üzere 4 çeşit tamlama bulunmaktadır. Bu tamlama tiplerini kendine özgü farklı yapı ve özellikleri söz konusudur. Sözlük tasarımı özellikle kelime öbeği ve tamlamalar yönünden baştan ele alınmalı ve çevirinin ihtiyacını karşılayacak yapı ve özellikleri barındıran sözlüklerin geliştirilmesi gerekmektedir. Çeviriye yönelik sözlük tasarımı ve geliştirimi için özel çalışma yapılması gereklidir.

KAYNAKÇA

[1] M. ERGİN, Osmanlıca Dersleri*, İstanbul: Boğaziçi yayınları*., 2020.

[2] E. F. Bilgin, "Machine transliteration of Ottoman Turkish texts to modern Turkish", *Fatih Üniversitesi: Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi,56,* 2012.

[3] Örneklerle Türkçe Sözlük, İstanbul: MEB Yayınları, 2000.

[4] Q. u. A. Akram, S. Hussain, A. Niazi, U. Anjum ve F. Irfan, "Adapting Tesseract for Complex Scripts: An Example for Urdu Nastalique,", *11th IAPR International Workshop on Document Analysis Systems,* 2014.

[5] Z. Urooba, D. N. Hakro, K.-u.-R. Khoumbat, M. A. Zaki ve M. Hameed, "Issues & Challenges in Urdu OCR,", *University of Sindh Journal of Information and Communication Technology (USJICT),* 2019.

[6] A. Mansoor ve W. Teahan, "Printed Arabic Script Recognition: A Survey," *International Journal of Advanced Computer Science and Applications(IJACSA),* 2018.

[7] F. T. Yarman Vural ve A. A. Atici, "A heuristic algorithm for optical character recognition of Arabic script," *Signal Processing,* 1997.

[8] E. Öztop, A. Mülayim, V. Atalay ve F. T. Yarman Vural, "Repulsive attractive network for baseline extraction on document images," *IEEE,* 1998.

[9] E. F. Can ve P. Duygulu, "A line-based representation for matching words in historical manuscripts," *Pattern Recognition Letters,* 2011.

[10] H. Adiguzel, E. Sahin ve P. Duygulu, "A Hybrid for Line Segmentation in Handwritten Documents," *International Conference on Frontiers in Handwriting Recognition,* 2012.

[11] H. Adigüzel, P. D. Şahin ve M. Kalpaklı, "Line Segmentation of Ottoman Documents," *Signal Processing and Communications Applications Conference*, 2012.

[12] Z. Kurt, H. I. Turkmen ve M. E. Karsligil, "Ottoman Alphabet Character Recognition by LDA," *Signal Processing and Communications Applications*, 2007.

[13] Z. Kurt, H. I. Turkmen ve E. Karsligil, "Linear Discriminant Analysis in Ottoman Alphabet Character Recognition," *Proceedings of the European Computing Conference*, 2009.

[14] R. M. Saabni ve J. A. El-Sana, "Comprehensive synthetic Arabic database for on/off-line script recognition research," *International Journal on Document Analysis and Recognition (IJDAR)* , 2013.

[15] P. Sankar, C. V. Jawahar ve R. Manmatha, "Nearest Neighbor based Collection OCR," *Das'10*, 2010.

[16] Z. Yalniz ve R. Manmatha, "A Fast Alignment Scheme for Automatic OCR Evaluation of Books," *IEEE,* 2011.

[17] Y. Z. ve M. R., "An Efficient Framework for Searching Text in Noisy Document Images," *IAPR,* 2012.

[18] L.Chen, S. Kapoor ve R. Bhatia, "Intelligent Systems for Science and Information", SpringerLink, 2014.

[19] S. G. Öztürk ve Y. Özbay, "Multifont Ottoman Character Recognition", *7th IEEE Int. Conf. on Electronics Circuits and System (ICECS),* 2000.

[20] P. Gorgel, N. Kilic, B. Ucan, A. Kala ve O. N. Ucan, "A Backpropagation Neural Network Approach For Ottoman Character Recognition", *Intelligent Automation & Soft Computing* , 2009.

[21] N. Kilic, P. Görgel, O. N. Ucan ve A. Kala, "Multifont Ottoman character recognition using Support Vector Machine", *Communications, Control and Signal Processing*, 2008.

[22] A. Onat, F. Yildiz ve M. Gündüz, "Ottoman Script Recognition Using Hidden Markov Model," *World Academy of Science, Engineering and Technology, Open Science Index 14, International Journal of Computer and Information Engineering, 2(2), 462 - 464.*, 2008.

[23] I. Z. Yalnız, I. S. Altingovde, U. Güdükbay ve Ö. Ulusoy, "Ottoman archives explorer: A retrieval system for digital Ottoman archives," *Journal on Computing and Cultural Heritage,* 2009.

[24] E. Saykol, . A. K. Sinop, U. Gudukbay, O. Ulusoy ve A. E. Cetin, "Content-based retrieval of historical Ottoman documents stored as textual images," *IEEE Transactions on Image Processing,* 2004.

[25] I. Z. Yalniz, I. S. Altingovde, U. Güdükbay ve Ö. Ulusoy, "Integrated segmentation and recognition of connected Ottoman script," *Optical Engineering,* 2009.

[26] I. S. Altingovde, E. Şaykol, Ö. Ulusoy, U. Gudukbay, A. E. Cetin ve Gocmen, "Content-Based Retrieval (CBR) System for Ottoman Archives," *Signal Processing and Communications Applications,* 2006.

[27] E. Ataer ve P. Duygulu, "Retrieval of Ottoman documents," *Proceedings of the 8th ACM SIGMM International Workshop on Multimedia Information Retrieval,* 2006.

[28] E. Ataer ve P. Duygulu, "Matching ottoman words: An image retrieval approach to historical document indexing," *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, 2007.

[29] M. Khayyat, L. Lam ve C. Suen, "Learning-based word spotting system for Arabic handwritten documents," *Pattern Recognition,* 2014.

[30] P. Duygulu, D. Arifoglu ve M. Kalpakli, "Cross-document word matching for segmentation and retrieval of Ottoman divans," *Pattern Analysis and Applications,* 2014.

[31] T. Bluche, J. Louradour, M. Knibbe ve B. Moysset, "The A2iA Arabic Handwritten Text Recognition System at the OpenHaRT2013 Evaluation," *International Workshop on Document Analysis Systems,* 2014.

[32] V. Pham, T. Bluche, C. Kermorvant ve J. Louradour, "Dropout improves Recurrent Neural Networks for Handwriting Recognition," *Computer Vision and Pattern Recognition*, 2013.

[33] J. Memon, M. Sami, R. A. Khan ve M. Uddin, "Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)," *IEEE Access,* 2020.

[34] Y. Zhu, C. Yao ve X. Bai, "Scene text detection and recognition: recent advances and future trends," *Frontiers of Computer Science,* 2015.

[35] M. Baechler, A. Fischer, N. Naji ve R. Ingold, "HisDoc: Historical Document Analysis, Recognition, and Retrieval," *Digital Humanities*, 2012.

[36] F. Shafait, "Geometric Layout Analysis of Scanned Documents," *Technical University of Kaiserslautern*, 2008.

[37] N. S. Dash, "Language, Script anf Font Recognition," 2019.

[38] M. Arif, H. Hassan, D. Nasien ve H. Haron, "A Review on Feature Extraction and Feature Selection for Handwritten Character Recognition," *International Journal of Advanced Computer Science and Applications,* 2015.

[39] N. Subramani, A. Matton, M. Greaves ve A. Lam, "A Survey of Deep Learning Approaches for OCR and Document Understanding," *Computation and Language,* 2021.

[40] T. Breuel ve K. Baierer, " hOCR - OCR Workflow and Output embedded in HTML," *International Conference on Document Analysis and Recognition (ICDAR),* 2020..

[41] H. Abbad ve S. Xiong, "Multi-components System for Automatic Arabic Diacritization," *Advances in Information Retrieval,* Advances in Information Retrieval, 2020, pp. 341-355.

[42] Wikiwand, [Çevrimiçi]. Available: www.wikiwand.com/en/rasm. [Erişildi: 2021].

[43] P. Ye ve D. Doermann, "Document Image Quality Assessment: A Brief Survey," *International Conference on Document Analysis and Recognition,* 2013.

[44] D. Kumar ve A. Ramakrishnan., "QUAD: Quality Assessment of Documents," *4th International Workshop on Camera-based Document Analysis and Recognition,* 2011.

[45] İ. Balcı, "Osmanli İmparatorluğu'nda Kullanilan Hat Çeşitleri, Bunlarin Öğretimi Ve Uygulama Metinleri," 2017.

[46] M. A. Ensari, Osmanlıca İmla Müfredatı, Hayrat Vakfı Yayınları, 2015.

[47] K. Lindén, "Multilingual modeling of cross-lingual spelling variants," *Information Retrieval,* pp. 295-310., 2006.

[48] S. Karimi, F. Scholer ve A. Turpin, "Machine transliteration survey," *ACM Computing Surveys,* 2011.

[49] J. Korkut, "Morphology and lexicon-based machine translation of Ottoman Turkish to Modern Turkish," 2019.

[50] K. Oflazer, "Two-level Description of Turkish Morphology," *Literary and Linguistic Computing,* 1994.

[51] S. Demir, I. D. El-Kahlout, E. Unal ve H. Kaya, "Turkish Paraphrase Corpus," *LREC,* 2012.

[52] A. Eyecioglu ve B. Keller, "Constructing a Turkish corpus for paraphrase identification and semantic similarity," *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 588-599, 2016.

[53] E. Özkan ve G. Ercan, "Eski Türkçe Metinlerin Modernleştirilmesi," *26th Signal Processing and Communications Applications*, 2018.

[54] P. Koehn, F. J. Och ve D. Marcu, "Statistical phrase-based translation," *Unıversity Of Southern California Marina Del Rey Information Sciences Inst*, 2003.

[55] Barancíková, Petra, Rudolf Rosa, and A. Tamchyna. "Improving evaluation of English-Czech MT through paraphrasing." *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* 2014.

[56] S. Wubben, A. V. D. Bosch ve E. Krahmer, "Paraphrase generation as monolingual translation: Data and evaluation," *Proceedings of the 6th International Natural Language Generation Conference*, 2010.

[57] Wikipedia Erişim: https://tr.wikipedia.org/wiki/Derin_öğrenme. [Erişildi: 2021].

[58] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu ve F. E. Alsaadi, "A survey of deep neural network architectures and their applications" *Neurocomputing*, 2017.

[59] N. Subramani, A. Matton, M. Greaves ve A. Lam, "A Survey of Deep Learning Approaches for OCR and Document Understanding," 2020.

[60] M. Pourreza, R. Derakhshan, S. Bibak, M. Fallah, H. Fayyazi ve M. Sabokrou, "Persian OCR with Cascaded Convolutional Neural Networks Supported by Language Model," *10th International Conference on Computer and Knowledge Engineering (ICCKE)*, 2020.

[61] M. Mohd, F. Qamar, I. Al-Sheikh ve R. Salah, "Quranic Optical Text Recognition Using Deep Learning Models," *IEEE Access*, 2021.

[62] S. Naz, A. I. Umar, R. Ahmad ve I. Siddiqi, "Urdu Nastaliq Recognition using Convolutional-Recursive Deep Learning," *Neurocomputing*, 2017.

[63] Derin Öğrenme Çevrimiçi: https://stanford.edu/~shervine/l/tr/teaching/cs-230/cheatsheet-convolutional-neural-networks. [Erişildi: 2021].

[64] Google Docs,Çevrimiçi: https://workspace.google.com/intl/en/products/docs/. [Erişildi: 2021].

[65] Microsoft Office, Çevrimiçi: https://www.office.com/. [Erişildi: 2021].

[66] Adobe Acrobat Pro, Çevrimiçi: https://acrobat.adobe.com/ , 2021

[67] Abby Finereader, Çevrimiçi: https://pdf.abbyy.com, 2021

[68] Omnipage, Çevrimiçi: https://www.kofax.com/products/omnipage, 2021

[69] Readiris OCR, Çevrimiçi: https://www.irislink.com/, 2021

[70] Sahr OCR, Çevrimiçi: http://www.sakhr.com/, 2021

[71] Tesseract OCR, Çevrimiçi: https://github.com/tesseract-ocr/. 2021

[72] OCRopus, Çevrimiçi https://github.com/ocropus. 2021

[73] Ocrad, Çevrimiçi: https://www.gnu.org/software/ocrad/, 2021

[74] GOCR, Çevrimiçi: http://jocr.sourceforge.net/. 2021

[75] Cuneiform, Çevrimiçi: https://github.com/jwilk-mirrors/cuneiform-multilang. 2021

[76] FreeOCR, Çevrimiçi: http://www.paperfile.net/. 2021

EKLER