Research Article
Journal of Smart Systems Research (JOINSSR) 3(1), 1-10, 2022
Recieved: 03-January-2022 Accepted: 14-January-2022

Osmanlıcadan Türkçeye Uçtan Uca Aktarım

İshak Dölek[1*], Atakan Kurt[2]

[1]Bilgisayar Mühendisliği, Mühendislik Fakültesi, İstanbul Ün-Cerrahpaşa, Türkiye, ishakdolek54@gmail.com

[2]Bilgisayar Mühendisliği, Mühendislik Fakültesi, İstanbul Ün-Cerrahpaşa, Türkiye, atakan.kurt@istanbul.edu.tr

ÖZET

Bu makalede Osmanlıca Dokümanların Modern Türkçeye Uçtan Uca aktarım çalışması sunulmuştur. Devlet arşivleri, kütüphaneleri ve özel koleksiyonlarda milyonlarca Osmanlıca doküman bulunmaktadır. Bunların Modern Türkçeye elle aktarımı mümkün değildir. Osmanlica.com adresinde kullanıma açılan bu çalışmada Osmanlıca dokümanların Modern Türkçeye 3 adımda aktarımı yapılmaktadır: (i) Osmanlıca karakter tanıma (OCR) (ii) Osmanlıca-Türkçe Alfabe Çevrisi (iii) Osmanlıca-Türkçe Dil Çevirisi. Bildiğimiz kadarıyla, bu çalışma Osmanlıca-Türkçe aktarım sürecinin üç adımını da çözmeyi hedefleyen ilk çalışmadır. Bu adımların her biri NLP ve Derin Öğrenmede teknik ve bilimsel olarak karmaşık ve kaynak gerektiren problemlerdir. Birinci adımda doküman görüntüleri OCR ile Osmanlı alfabesinde düz metine dönüştürülür. İkinci adımda Arap-tabanlı Osmanlı alfabesindeki bu metin bir alfabe çevrisi sistemiyle Latin-tabanlı Türk alfabesine dönüştürülür. Türk alfabesindeki metin her ne kadar okunabilir olsa da çok sayıda Arapça ve Farsça kelime ve yapılar barındırdığı için henüz anlaşılabilir değildir. Üçüncü adım bu metin makine çevirisi ile Modern Türkçeye aktarılır. Birinci adımda geliştirilen CRNN tabanlı OCR modeli 21 sayfalık bir veri setinde test edilmiş ve %96 karakter tanıma doğruluk oranı üretmiştir. İkinci adımda geliştirilen alfabe çeviri sistemi 7500 kelimelik bir veri setiyle test edilmiş ve %98 kelime çeviri doğruluk oranı üretmiştir. Üçüncü adım için kelime grubu tabanlı bir makine çeviri sistemi geliştirilmiş ve testlerine başlanmıştır. Bu çalışmanın tarihi, sosyal ve kültürel bir ihtiyaca katkı sağladığı için önemli ve değerli olduğunu düşünüyoruz.

Anahtar Kelimeler: Osmanlıca OCR, Osmanlıca-Türkçe alfabe çevirisi, Osmanlıca-Türkçe harfçevrim, Osmanlıca-Türkçe dil çevirisi

Ottoman-Turkish End to End Conversion

ABSTRACT

In this paper, a study titled End-To-End Conversion Ottoman Documents to Contemporary Turkish is presented. The state archives, libraries, and private collections contain millions of documents written in Ottoman. It is practically impossible to convert all these documents to Modern Turkish manually. In this work which is available at Osmanlica.com Ottoman documents are converted to Modern Turkish in three steps: (i) Ottoman OCR (Optical Character Recognition), (ii) Ottoman-Turkish transliteration, and (iii) Ottoman-Turkish translation. To our knowledge this is the only study to set out to solve all three steps of this conversion to date. Each one of these three steps are technically complex and resource-demanding problems in NLP and deep learning. OCR converts image files to editable text in Ottoman alphabet in the first step. Transliteration tranforms that Ottoman text in Arabic-based Ottoman alphabet to the Latin-based Turkish alphabet making it readable but not yet understandable because of Arabic and Persian words and structures in the second step. In the last step, this Ottoman text in Turkish alphabet is translated to Modern Turkish via machine translation. The CRNN based on model developed in the first step produced %96 OCR accuracy with a 21 pages test document set. The Ottoman-Turkish transliteration system

* Corresponding Author's email: ishakdolek54@gmail.com

developed yielded %98 accuracy with a test set of 7500 words in the second step. The phase-based Ottoman-Turkish machine translation system developed in the third step is being tested presently. We believe that the contribution of this study is significant because it addresses an important social, cultural and scientific problem.

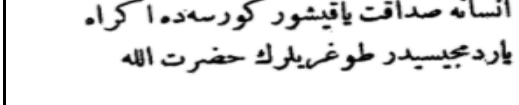
Keywords: Ottoman OCR, Ottoman-Turkish Transliteration, Ottoman-Turkish Translation

1 Giriş

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 Osmanlıcayı hem okumak hem de anlamak güçtür. Millî kültürümüzün temelini oluşturan eserlerin büyük bir kısmı Osmanlıcada yazılmıştır. Osmanlı arşiv ve kütüphanelerindeki kitap, dergi, gazete, defter, kayıt ve belgeler yüzlerce yıllık kültür, sanat ve tarih mirası içinde önemli bir yer tutar. Bu kaynaklarda saklı bilgiye hızlı, etkin ve doğru bir şekilde erişilmesi için başta OCR olmak üzere teknolojinin yardımına ihtiyaç vardır. TÜBİTAK destekli *Osmanlıcadan Günümüz Türkçesine Yapay Zekâ Destekli Uçtan Uca Aktarım Projesi* bu amaçla yapılan bir çalışmadır. Çalışmanın çıktıları Osmanlica.com sitesinden kullanıcılara sunulmuştur. Çalışma Tablo 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 çalışmaya dâhil değildir.
2. *Görüntü-metin dönüşümü (image-text conversion, OCR)*: Görüntü dosyasının görüntü işleme ve makina öğrenmesi ile Osmanlı Alfabesindeki Osmanlıca metne (editable text) dönüştürülmesi işlemidir.
3. *Osmanlıca-Türkçe Alfabe Dönüşümü (harfçevrim, machine transliteration)*: Osmanlı alfabesindeki Osmanlıca metnin Türk alfabesine, yani Latin Harfli Modern Türkçe alfabesine aktarımıdır.
4. *Osmanlıca-Türkçe Dil Çevirisi (intra-language machine translation)*: Türk alfabesindeki Osmanlıca metnin bilgisayarlı çeviriyle Modern yani günümüz Türkçesine çevirisidir. Çeviri sonucu ortaya Türk Alfabesinde Modern Türkçe metin çıkar.

Tablo 1: *Osmanlıca-Türkçe Uçtan-Uca Aktarım Örneği*

Adım (girdi → işlem → çıktı)	*Çıktı*
1. Doküman-görüntü dönüşümü: Doküman → sayısallaştırma → görüntü dosyası	
2. Osmanlıca OCR: Görüntü dosyası → OCR → Osmanlıca (Arap alf)	انسانه صداقت یاقیشر کورسه‌ده اکراه یاردمجیسیدر طوغریلرك حضرت الله
3. Alfabe dönüşüm: Osmanlıca (Arap alf) → dönüşüm → Osmanlıca (Latin alf)	İnsana sadâkat yakışır görse de ikrah Yardımcısıdır doğruların hazreti Allah
4. Dil çevirisi: Osmanlıca (Latin alf) → çeviri → Türkçe (Latin alf)	İnsana doğruluk yakışır görse de kötülük Yardımcısıdır doğruların Hazreti Allah

Bu makalede sunulan çalışma Osmanlıca-Türkçe aktarım problemini bütün olarak ele alan kendi türündeki ilk çalışma olup *Osmanlıca OCR, alfabe çevirisi* ve *dil çevirisi* olmak üzere üç adımdan ve her adım için geliştirilen 3 ayrı araçtan oluşmaktadır. Osmanlıca OCR için akademik çalışmalara [2] [3] ve ticari çalışmalara [4] [5] örnek verilebilir. Çalışmamızda geliştirdiğimiz Osmanlıca OCR modeli, Google Docs, Fine Reader, Tesseract Arapça ve Farsça, Miletos araçları ile karşılaştırılmıştır [6]. Model %96 karakter tanıma doğruluk oranıyla diğer araçlardan belirgin bir şekilde daha yüksek bir performans sağlamıştır. Kısıtlı sayıdaki Osmanlıca-Türkçe alfabe çevirisi çalışmalarına [7] [8] [9] örnek verilebilir. Çalışmamızda geliştirdiğimiz alfabe çevirisi aracıyla yapılan testlerde %98 doğruluk oranına ulaşılmıştır. Osmanlıca-Türkçe dil çevirisi alanında yapılan çalışmalara ise [10] örnek verilebilir. Çalışmamızda geliştirdiğimiz dil çeviri aracı aşağıda Bölüm 4'te özetlenmiş olup testlerine devam edilmektedir.

Osmanlıca-Türkçe aktarımın her adımı karmaşık ve güç bir problem olduğundan önceki çalışmalar bu adımlardan sadece birine veya belirli bir alt probleme yoğunlaşabilmiş durumdadır. Literatürde aktarım sürecinin tamamı üzerine yapılmış herhangi bir çalışma bulunmamaktadır. Osmanlıca OCR üzerinde yapılan çalışmalarda yukarıda belirtildiği üzere ticari araçlardan veya açık kodlu araçlardan daha yüksek bir doğruluk oranına henüz ulaşamamıştır. Osmanlıca-Türkçe alfabe çevirisi konusunda literatürde birkaç çalışma olsa da alfabe çevirisi OCR'dan daha güç bir problem olduğundan henüz ticari veya açık kodlu bir çeviri sistemi geliştirilememiştir. Osmanlıca-Türkçe dil çevirisi konusunda ise yapılmış çalışmalar küçük metin kümeleri üzerinde derin sinir ağlarının eğitilmesiyle üretilen modeller ile sınırlı kalmış ve Türkçe sondan eklemeli bir dil olduğundan testlerde henüz tatmin edici seviyede bir sonuç elde edilememiştir.

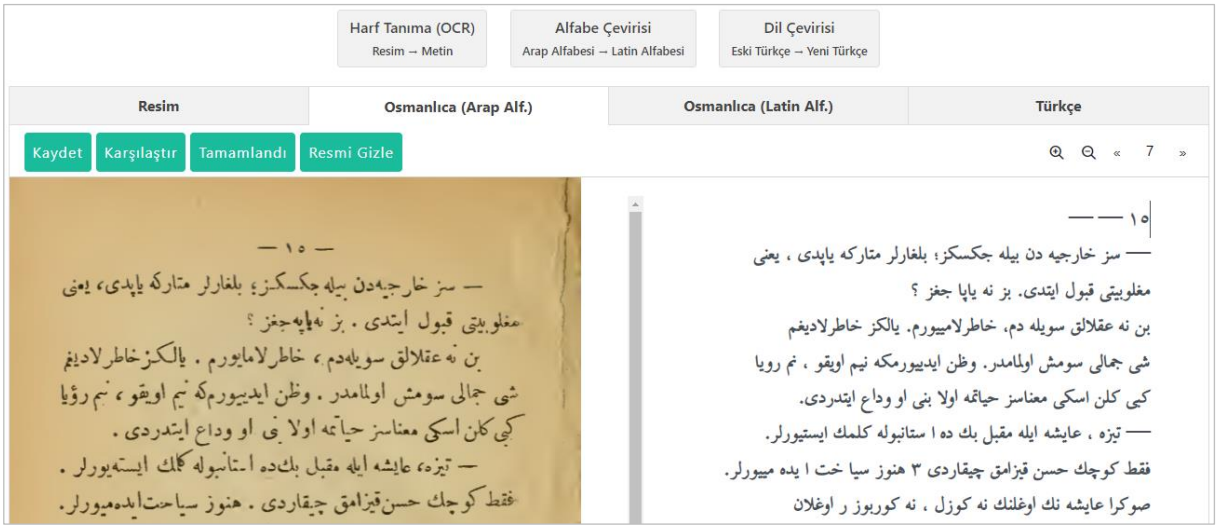
Makale şu şekilde organize edilmiştir: Bölüm 2'de Osmanlıca OCR, Bölüm 3'te Osmanlıca Türkçe Alfabe Çevirisi ve Bölüm 4'te Osmanlıca Türkçe Dil Çevirisi kısaca sunulmuştur. Sonuçlar ve özet Bölüm 5'te verilmiştir.

2 Osmanlıca OCR

Osmanlıca OCR bir resim dosyasında saklanan görüntüdeki karakterlerinin görüntü işleme ve makina öğrenmesi teknikleri kullanılarak metne dönüştürülmesi işlemidir [11]. Osmanlıca OCR 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.
2. *Görüntü bölümleme (image segmentation)*: Tanıma işleminin yapılabilmesi için görüntünün tanınacak birimlere bölümlenmesi gerekmektedir. Görüntü öncelikle satırlara sonrasında kelime karakterlere bölümlenir.
3. *Özellik çıkarma (feature extraction)*: Tanıma işlemi bir sınıflandırıcı modelle yapıldığından, modeli eğitmek için kullanılacak özelliklerin baştan bir defa belirlenmesi gerekmektedir. Çok katmanlı CNN tabanlı derin sinir ağlarında özellik çıkarma model eğitiminin içinde gerçekleşmektedir.
4. *Tanıma (recognition, OCR)*: Tanınmak istenen birimin (eğri, karakter, karakter katarı, kelime, vb.) bir sınıflandırıcı model kullanılarak tanınması işlemidir. Tanima için CNN ve RNN sinir ağlarını birleştiren bir derin öğrenme modeli kullanılmıştır.
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.

Osmanlıca dokümanlarda OCR çalışması Arapça ve Farsçaya göre daha zor bir problemdir. Osmanlıca OCR için bu zorlukları ikiye ayırabiliriz. (i) dil bağımsız ve (ii) dil bağımlı zorluklar. OCR probleminde dilden bağımsız olarak başarı oranını görüntü çözünürlüğü, görüntü kalitesi, yazı fontu, veri seti, sözlük vb. birçok etken söz konusudur. Dil-bağımlı zorluklar arasında Osmanlıca alfabesindeki harflerin kavisli olması, harflerin bitiştirilmesi, harflerin konuma göre farklı şekillere sahip olması, harfler arasında benzerliğin yüksek olması, harflerin yığılarak yazılabilmesi vb. durumlar söz konusudur. Osmanlıca OCR için geliştirilen sistemin ara yüzü Şekil 1'de verilmiştir. Ara yüzde OCR yapılan dokümanın resim ve OCR çıktısı yan yana gösterilmektedir.



Şekil 1: Osmanlıca OCR aracı

OCR işleminden önce dokümandaki süslü kenarlıklar, çerçeveler, şekiller vb. kaldırılması gerekmektedir. OCR'dan sonra elde edilen metin üzerinde metin normalizasyonu, bölütlenmiş kelimelerin düzeltilmesi, yazım hatalarının düzeltme gibi işlemleri uygulanarak hatalar düzeltilir. Tablo 2'de OCR'daki tanıma hatalarına bir örnek verilmiştir. Bu örnekte Osmanlıca dokümandan bir satır resim, referans yani doğru metin (ground truth) ve OCR'da hesaplanan metin paralel olarak verilmiştir. İşlem satırında ↕ değişim, ↓ ekleme, ↑ silme anlamına gelmektedir. Osmanlıca resim OCR yapıldıktan sonra noktalama işaretleri, kelime arası boşluklar, kelime içi boşluklar vb. hataları normalize edilmesi gerekmektedir. Özellikle Osmanlıca metinlerde 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 karıştırıldığında ortaya kelime bölütleme problemi çıkar. Bu problem alfabe çevirisi sırasında sıkça karşılaşılan bir problemdir.

Tablo 2: *OCR hata örnekleri*

دوردق . بلی اینجه ، چیزمه‌لری دار وپارلاق ، قالپاغی چارپیق ،
دوردق. بلی اینجه ، چیزمه‌لری طار و پارلاق ، قالپاغی چارپ*یق ،
↕ ↑ ↓
طوردق. بلی ا*نجه ، چیزمه‌لری طار و پارلاق ، قالپاغی چارپییق ،

3 Osmanlıca-Türkçe Alfabe Çevirisi

Alfabe çevirisi Osmanlı alfabesindeki Osmanlıca metnin Türk alfabesine metni değiştirmeden yani birebir bilgisayarlı ortografik alfabe çevirisidir. Burada her ne kadar birebir çeviri terimi kullanılsa da, Osmanlıcada genellikle sadece sessizler ve uzun sesliler yazıldığından bilgisayarlı alfabe çevirisi karmaşık bir işlem olup bildiğimiz kadarıyla şimdiye kadar kullanıma açık web tabanlı uygulaması olan sadece bir adet prototip çalışma mevcuttur [3]. Osmanlı ve Türk alfabesindeki sesli ve sessiz harfler

arasında çoktan çoğa bir eşleşme söz konusudur. Ayrıca alfabe çevirisinin Türkçedeki imla ve gramer kuralları uygun şekilde yapılması gerekmektedir. Türkçede 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, yapıbilimsel üretim, belirsizlik giderme, seslendirme (vowelization) vb. işlemlerin yapılması gerekmektedir.

Osmanlıca alfabesi 28 harfli Arap alfabesinin genişletilmiş bir çeşididir. 4 tane okutucu harf bulunmaktadır. Bu dört harf Türkçedeki 8 ünlü sese karşılık gelmektedir. Türkçe alfabesi 29 harften oluşmaktadır. 8 ünlü ve 21 ünsüz harften oluşur.

Osmanlıca-Türkçe Alfabe Çevirisi (i) Ortografik Alfabe Çevirisi, (ii) Kelime Bölütleme (iii) Yazım Düzeltme, (iv) Seslendirme, (v) Kelime tahmini ve (vi) İsim Tamlaması ve Birleşik kelimeler olmak üzere kabaca altı aşamadan gerçekleştirilir. Geliştirilen alfabe çevirisi sistemi bir örnek ile Şekil 2'de gösterilmiştir.

Şekil 2: Osmanlıca-Türkçe alfabe çeviri aracı

Birinci aşamada ortografik alfabe çevirisi aşaması olup bu aşamada Osmanlıca kelime gövdelenerek gövdeler ve ek-katarlarına ayrılır. Sonrasında Osmanlıca gövdeler ve ek-katarları gövde ve katar çeviri sözlükleri kullanılarak Türkçe gövdeler ve ek- katarlarına ulaşılır. Bir sonraki adımda Türkçe gövdeler ve ek katarları bitiştirilerek Osmanlıca kelimenin Türkçe karşılığı elde edilir. Bu şekilde çözümlenemeyen bir kelime olursa, bu kelimeler bir sonraki aşama olan kelime bölütleme aşamasına aktarılır. Kelime bölütleme aslında kelime bitişmesi (joined words) ve bölünmesi (segemented words) şeklinde iki problemden oluşmaktadır. Kelime bölütleme giderme algoritmasında şu an için pratik sebeplerden dolayı bir kelime bir önceki kelime ile birleştirilerek çözümlenmeye çalışılır. Eğer birleştirme sonucu geçerli bir kelime oluşursa, bu çözüm doğru kabul edilir. Eğer geçerli bir kelime oluşmazsa kelime bir sonraki aşama olan yazım düzeltmeye aşamasına aktarılır. Bu aşamada kelimede bir yazım hatası olduğu kabul edilerek yazım hatası düzeltme (spelling correction) algoritması uygulanır. Yazım düzeltme işlemi sonrası elde edilen kelime tekrar çözümlenmeye çalışılır. Çözümleme sonucu geçerli bir kelime elde edilirse bu çözüm doğru kabul edilir. Eğer kelime çözümlenmezse, söz konusu kelime bir sonraki aşama olan seslendirme (vowelization/vocalization) aşamasına aktarılır. Bu aşamada kelimenin sözlükte bulunmayan bir kelime (out of dictionary word) olduğu varsayımıyla kural tabanlı bir harf çevirisi algoritması uygulanır. Bu aşamada her zaman birçok çözüm elde edilir. Fakat çözümlerin doğrulanması mümkün değildir. Çünkü üretilen kelimeler sözlükte bulunmayan yabancı bir kelime ya da özel bir isim olabilir. Osmanlıca bir kelimenin Türkçede birden fazla karşılığı olabildiği için bir sonraki aşamada n-grams (dil modeli) kullanılarak çözümler arasından olasılığı en yüksek olan kelime seçilerek işlem tamamlanır. Bir sonraki aşamada metindeki isim tamlamaları ve bileşik isimler için geliştirilen algoritma ile isim tamlamaları ve bileşik isimler Türkçeye aktarılır ve böylece Türkçe metin oluşturulur.

Bu aşamaların bir örnek ile Tablo 3'de gösterilmiştir. Tabloda birden fazla çözüm { } sembolleri arasında verilmiştir. Kelime bölütleme aşamasında bölütlenmiş (اوقویه جق okuya+cak) kelimesi doğru şekilde çözümlenmiştir. Yazım düzeltme aşamasında حرایه → haraye kelimesi حرابه → harabe şeklinde düzeltilmiştir. Seslendirme aşamasında معمورهٔ → mamure doğru biçimde elde edilmiştir. Osmanlıca kelimelerin Türkçede 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 اسلامیه→ *bilad islamiyye* kelimesi *bilad-ı islamiyye* şeklinde düzeltilmektedir.

Tablo 3: *Osmanlıca-Türkçe alfabe çevirisi adımları*

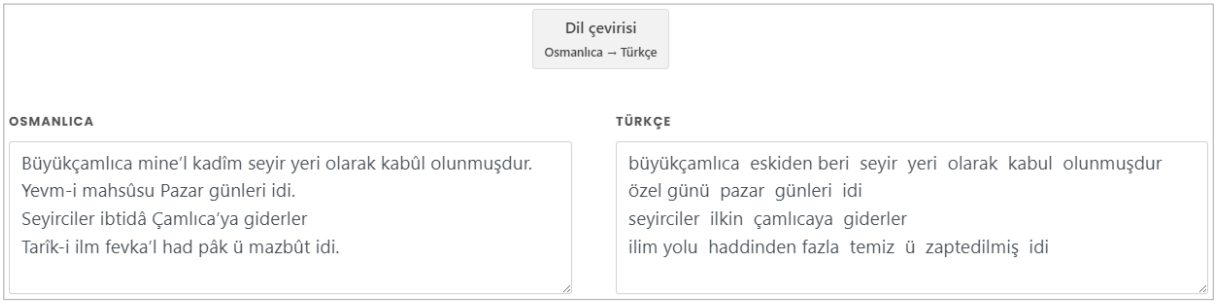
Adım	*Sonuç*
Osmanlıca metin	ماضیده هر بری بر معمورهٔ مدنیت اولان بلاد اسلامیه بوکون؛ قسوت انکیز بررر خرایه حالنی المش. قوای طبیعیه‌یه میدان اوقویه جق
Alfabe çevirisi	mazide her {berri, beri, biri} {bir, ber} معمورهٔ {müdünüyet, medeniyet} {ölen, olan, avlan} bilâd islâmiyye {bugün, bükün} kasvet{enekyiz, engiz} birer حرایه halını {elmiş,almış}. kuva tabiyeye meydan اوقویه جق
Kelime bölütleme	mazide her {berri, beri, biri} {bir, ber} معمورهٔ {müdünüyet, medeniyet} {ölen, olan, avlan} bilâd islâmiyye {bugün, bükün} kasvet {enekyiz, engiz} birer حرایه halını elmiş, almış}. kuva tabiyeye meydan okuyacak
Yazım düzeltme	mazide her {berri, beri, biri} {bir, ber} معمورهٔ {müdünüyet, medeniyet} {ölen, olan, avlan} bilâd islâmiyye {bugün, bükün} kasvet {enekyiz, engiz} birer harabe halını {elmiş, almış}. kuva tabiyeye meydan okuyacak
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
n-grams	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
Tamlama & Birleşik kelimeler	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

4 Osmanlıca-Türkçe Dil Çevirisi

Dil çevirisi Türk alfabesindeki Osmanlıca metnin bilgisayarlı çeviriyle Modern yani günümüz Türkçesine 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 söz dağarcığı, sözdizim ve anlamsal anormalliklerden dolayı yine de kolay bir problem değildir. Bu adımda kelime-grubu tabanlı birebir çeviri yaklaşımıyla ilerlenebileceği gibi, derin öğrenme kütüphaneleri kullanılarak cümle tabanlı makina çevirisi yaklaşımı da kullanılabilir. Çeviri öncesi (pre processing), sırası ve sonrasında (post processing) yüzeysel/tam

yapıbilimsel ve/veya sözdizimsel çözümleme, üretim, belirsizlik giderme, kelime anlamı belirsizliği giderme, varlık ismi tanıma vb. işlemlerin yapılması gerekmektedir.

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 Osmanlıcadan Modern Türkçeye aktaran bir yöntemdir. Bu yöntemin basit olmasına karşın çok karmaşık olmayan cümleler için etkin bir yöntem olduğunu düşünüyoruz ve elde edilen sonuçların daha ileri yöntemlerin geliştirilmesinde ve performans analizinde bir temel (baseline) olarak ta kullanılabileceği kanaatindeyiz. Osmanlıca Türkçe bir çeviri örneği **Şekil 3**'te verilmiştir.



Şekil 3: Osmanlıca-Türkçe dil çeviri aracı

Problemin zorluğuna dikkat çekmek amacıyla Osmanlıca Türkçe dil çevirisinde bazı problem ve zorlukları örnekleriyle aşağıda paylaşılmıştır. Bu örnekler hem veri setinin hem de çeviri algoritmasının geliştirilmesi ve iyileştirilmesinde yönlendirici değere sahiptir.

1. *Yazılışı aynı kelimeler*: Osmanlıca ve Modern Türkçedeki ortak ve yaygın kelimeler çeviriye sokulmadan doğrudan aktarılmalıdır. "bir" gibi Osmanlıcada *kuyu* anlamına da gelen bazı kelimeler yanlış çeviriye sebep olabilmektedir.
2. Osmanlıcada bazı metinlerde anlatım ağdalı 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 ifade bozukluğuna sebep olabilmektedir. Örneğin Türkçe geçmiş zaman -dH eki Osmanlıcada "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 bazen korunmakta bazen de tekilleştirilerek yazılmaktadır. Örneğin *hilâfiyyeyi* kelimesi *hilâfiyeyi* şeklinde de yazılabilmektedir.
6. *Küçük büyük harf ayırımı*: Osmanlıcada küçük büyük harf ayrımı olmadığından özel isimlerin bulunması ve çevriminde problem yaşamaktadır. Örneğin *Selimiye* kelimesinin *sağlamıye* diye çevrilmesi yada *murad rabi*'nin *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 şu an için kullandığımız *en uzun gövde en kısa ek katarı* yöntemi morfolojide yetersiz kalmaktadır. Örneğin Osmanlıca *al* kelimesi *almak* ya da *aile* olarak Türkçeye aktarılabilir. Burada doğru karşılığı bulmak için çeviri sonrası dil modelleri (n-grams) kullanılarak doğru çözüme ulaşılabilir. Benzer örnekler:
    - şehr → şehir (isim), ay (isim)
    - bak → bak (fiil), korku (isim)
    - azm → gayret (isim), kemik (isim)
8. *Çoklu yüzeye sahip kelimeler*: Osmanlıcadaki birçok kelimenin birden fazla imlası (yazım şekli, yüzey, surface) 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. Osmanlıcanın Türk alfabesiyle yazımında ayırıcı ya da düzeltme işaretleri denen diakritiklerin metinde doğru ve standart bir şekilde kullanımı çeviride çok önemlidir. *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. Örneğin *bû* kelimesinin Türkçedeki karşılığı *koku* kelimesidir. Çeviri uygulamasına verilen bir metinde bu kelimenin şapkasız yazılması durumunda Türkçeye işaret sıfatı ya da zamiri olarak yanlış çevrilmesi söz konusu olacaktır.
9. *Farklı dilden 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.

5 Sonuçlar

Bu çalışmada Osmanlıca dokümanların *OCR, alfabe çevirisi* ve *dil çevirisi* olmak üzere bir uçtan diğer uca yani Osmanlıca dokümandan Modern Türkçe metne üç adımda otomatik olarak çevrilmesi amaçlanmaktadır. OCR ile Osmanlıca dokümanların görüntü dosyaları düzenlenebilir metine dönüştürülmekte, alfabe çevirisiyle Osmanlıca metin Arap-tabanlı Osmanlı alfabesinden Latin-tabanlı Türk alfabesine aktarılmakta ve dil çevirisiyle Türk alfabesindeki Osmanlıca metin Modern Türkçeye çevrilmektedir.

Çalışmanın birinci amacı Osmanlıcadan Türkçeye aktarım sürecini tüm yönleriyle ortaya koymak ve bu süreçte karşılaşılan problemlerin tanımlayıp onlara yönelik olası çözümleri yakından incelemektir. Çalışmanın ikinci amacı aktarımın her üç adımına yönelik çevrimiçi araçlar geliştirmek, bu araçları benzer bilimsel ve ticari araçlarla karşılaştırarak test etmek ve sonrasında kullanıma açmaktır. Çalışmanın üçüncü amacı geleceğe yönelik bir hedeftir. Gelecekte her adım için ayrı ayrı geliştirilen araçların bir arada daha etkin çalışabilmesi ve aktarım sürecinin toplam performansının iyileştirilmesine yönelik çalışmalar yapılacaktır. Derin sinir ağları büyük veri setleri eğitilerek daha önce çözümü zor olan problemlere yönelik başarılı çalışmalar yapılabilmektedir. Örneğin bir adımda elde edilen veriler bir önceki adıma geri beslenerek o adımın iyileştirilmesi söz konusu olabilmektedir. Ya da ardışık iki adım birleştirilerek bir arada çözümlenebilecektir. Örneğin ORC adımı atlanarak -diğer bir deyişle bir sonraki adımla birleştirilerek- doküman görüntüsünden doğrudan Türk alfabesine dönüşüm yapılabilir. Veya alfabe çevirisi atlanarak, Osmanlıca metinden günümüz Türkçesine doğrudan çeviri yapılabilir. Hatta Osmanlıca doküman görüntüsünden günümüz Türkçesine direkt çeviri söz konusu olabilir.

Osmanlıca dokümanların Modern Türkçeye aktarımı toplumumuzun teknoloji ile çözülmesi gereken önemli kültürel ve bilimsel problemleri arasında yer almaktadır. Bilimsel veya ticari, şimdiye kadar yapılan çalışmalarda geliştirilen OCR araçları sayıca az olup elde edilen doğruluk oranları düşük olduğu için ihtiyaca cevap verecek durumda değildir. Üzerinde en çok ve en uzun süre çalışılan problemlerden biri olan Osmanlıca OCR konusunda geliştirilen araçlar, Abby FineRader veya Google Docs gibi kapalı kodlu ticari OCR araçlarının veya Tesseract gibi açık kodlu ürünlerin performansının altında kalmaktadır. Bu çalışmada geliştirdiğimiz OCR aracı hem Google Docs hem FineReader hem de Tesseract'tan daha yüksek doğruluk oranı (%96) vermektedir.

Osmanlıca-Türkçe alfabe çevirisi konusu şimdiye kadar daha çok dilcilerin ilgilendiği bir problem olarak kalmışsa da son zamanlarda doğal dil ve makine öğrenmesi konularındaki gelişmelere paralel olarak bu konuya yönelik bir elin parmakları kadar az sayıda çalışma yapılmıştır. Bu çalışmalarda kullanıma açık tek araç bizim geliştirdiğimiz alfabe çevirisi aracıdır. Yapılan testlerde bu aracın doğruluk oranı %98 olarak belirlenmiştir.

Osmanlıca-Modern Türkçe dil çevirisi konusunda ise çok daha az sayıda çalışma yapılmış olup bizim geliştirdiğimiz araç hariç henüz ortaya elle tutulur bir model ya da araç geliştirilememiştir. Oysa batı dillerinde çeviri konusunda başarılı birçok çalışma mevcut olup günlük hayata girmiş araçlar ortaya çıkmıştır.

6 Beyanname

6.1 Çalışma Sınırlamaları

Yazarlar, bu çalışmada araştırma sonucunu önemli ölçüde etkileyebilecek herhangi bir sınırlama ile karşılaşmadığını beyan eder.

6.2 Teşekkür

Yazarlar, bu çalışmanın kalitesini artıran yapıcı öneriler için anonim hakemlere teşekkürlerini sunar.

6.3 Finansman Kaynağı

Yazar(lar) herhangi bir fon kaynağı beyan etmemiştir.

6.4 Rakip Çıkarlar

Bu çalışmada herhangi bir çıkar çatışması yoktur.

6.5 Yazarların Katkıları

Sorumlu Yazar; İshak DÖLEK: Makaledeki çalışmayı bir doktora tezi olarak yapmak, sunulan araçlara ait algoritmaları ve örnekleri hazırlamak, yorumlamak ve anlatmak, geliştirilen araçların ara yüzlerini sunmak, test sonuçlarını paylaşmak, makalenin temel çatısını oluşturup Giriş ve Sonuçlar bölümü hariç makaleyi yazmak.
2. Atakan KURT: Makalede sunulan doktora tezini yönetmek, Giriş ve Sonuçlar bölümlerini yazmak, makalenin tamamı üzerinde yorum, düzenleme ve düzeltmeler yapmak.

7 İnsan ve Hayvanlarla İlgili Çalışma

Bu tür bir çalışma için resmi onay gerekli değildir

7.1 Etik Onay

Bu çalışma bir masa başı incelemesi içerdiğinden, yazarlar tüm prosedürlerin ilgili kurumsal komitelerin etik standartlarına uygun olduğunu iddia etmektedir. Bu tür bir çalışma için resmi onay gerekli değildir.

7.2 Bilgilendirilmiş Onay

Çalışmaya dâhil edilen tüm bireysel katılımcılardan bilgilendirilmiş onam alınmıştır

Kaynakça

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

[2] S. Kirmizialtin ve D. Wrisley, "Automated Transcription of Non-Latin Script Periodicals: A Case Study in the Ottoman Turkish Print", *Archive arXiv preprint arXiv:2011.01139*, 2020. https://arxiv.org/abs/2011.01139

[3] M. Mohd, F. Qamar, I. Al-Sheikh and R. Salah, "Quranic Optical Text Recognition Using Deep Learning Models", *IEEE Access*, vol. 9, pp. 38318-38330, 2021, doi: 10.1109/ACCESS.2021.3064019.

[4] Miletos OCR [Çevrimiçi]. Erişim: www.miletos.com

[5] IRCICA [Çevrimiçi]. Erişim: library.ircica.org.

[6] I. Dolek and A. Kurt, "Ottoman OCR: Printed Naskh Font," *2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, 2021, pp. 1-5, doi: 10.1109/INISTA52262.2021.9548616.

[7] A. Kurt, and E. F. Bilgin. "The Outline of an Ottoman-to-Turkish Automatic Machine Transliteration System" *First Workshop on Language Resources and Technologies for Turkic Languages.* 2012.

[8] J. Korkut, "Morphology and Lexicon-Based Machine Translation of Ottoman Turkish to Modern Turkish", *Princeton University, Princeton, NJ, USA*, 2019.

[9] A. A. Jaf, S. K. Kayhan, "Machine-Based Transliterate of Ottoman to Latin-Based Script", *Scientific Programming*, vol. 2021, Article ID 7152935, 8 pages, 2021. https://doi.org/10.1155/2021/7152935

[10] E. Özkan and G. Ercan, "Modernization of old turkish texts,", *26th Signal Processing and Communications Applications Conference (SIU)*, 2018, pp. 1-4, doi: 10.1109/SIU.2018.8404308.

[11] J. Memon, M. Sami, R. A. Khan and M. Uddin, "Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)," *in IEEE Access*, vol. 8, pp. 142642-142668, 2020, doi: 10.1109/ACCESS.2020.3012542.