

Osmanlıca Metinlerin Dijital/Otomatik Transkripsiyonu: Mevcut Durum ve Öneriler

Sümeyye Akça* – Yunus Uğur** – Murat Esringü***

Giriş

Dijital insanî/beşeri bilimler çalışmalarının tarihi 1946'da Roberto Busa'nın Thomisticus İndeksi üzerinde yaptığı leksik metin analizi (lexical text analysis) çalışmasına dayandırılmaktadır.¹ 1970'lerin sonlarından itibaren depolama ve işleme kapasiteleri arttıkça, yapılandırılmış elektronik metin ve multimedya arşivleri alana hakim olmuş, bunu 1990'larda İnternet destekli hiper metinler (hypertext), dijital kütüphaneler ve işbirliğine dayalı (collaborative) çalışmalar takip etmiştir.² Dijital insanî bilimler teriminin yaygınlaşmasının ve alanın kurumsallaşmasının önemli bir göstergesi olarak ise ABD Ulusal İnsanî Bilimler Enstitüsü'nün (Office of Digital Humanities by the National Endowment for the Humanities) 2008'de Dijital İnsanî Bilimler Ofisi'ni kurması kabul edilmektedir. Literatürde pek çok tanımı bulunan bu alan genel olarak bilgisayar teknolojilerinin insanî bilimlere uygulanmasıdır.³ Dijital teknolojilerin geliştirilmesi ve yorumlanması için geniş

* Marmara Üniversitesi.

** Marmara Üniversitesi.

*** Marmara Üniversitesi.

1 Edward Vanhoutte, "The Gates of Hell: History and Definition of Digital Humanities", *Defining Digital Humanities: A Reader*, ed. Melissa Terras, Julianne Nyhan, Edward Vanhoutte (London: Routledge, 2013), s. 119-156.

2 Chris Alen Sula, and Heather V. Hill, "The Early History of Digital humanities: An Analysis of *Computers and the Humanities* (1966–2004) and *Literary and Linguistic Computing* (1986–2004)", *Digital Scholarship in the Humanities*, 34/Supplement_1 (2019), s. 191.

3 Willard McCarty, "What is Humanities Computing? Toward a Definition of the Field", (1998). http://www.dighum.kcl.ac.uk/legacy/teaching/dtrt/class1/mccarty_humanities_computing.pdf.

bir uygulama sahası sunan dijital insanî bilimler⁴ dijitalleştirilmiş tarihi ve kültürel miras üzerinde içeriği daha iyi anlamaya ve çözümlemeye yardımcı bilgisayar araçlarının kullanılmasını teşvik eder.⁵ Son 10 yılda tüm dünyada bu alandaki çalışmalar hız kazanmış, öğrenen bilişim teknolojileri (yapay zeka) ile disiplinlerarası bir çalışma sahası ortaya çıkmıştır.⁶ Dil, coğrafya gibi erişim bariyerlerini en aza indirmeyi hedefleyen bu çalışmalar geniş kitlelerin insanlığın tarihi ve kültürel mirasıyla etkileşimini kolaylaştırmayı hedefler.

Bu çalışmaların Osmanlıca çerçevesindeki gelişmelerini ortaya koymak üzere Marmara Üniversitesi Dijital Beşerî Bilimler Uygulama ve Araştırma Merkezi tarafından 20 Aralık 2023 tarihinde “Osmanlıca Metinlerin Dijital Transkripsiyonu Çalıştayı” düzenlenmiştir. Çalıştaya çeşitli üniversitelerden ve farklı disiplinlerden konu üzerine çalışan akademisyen, kütüphaneci ve proje yöneticisi olmak üzere toplamda 10 kişi davetli konuşmacı ve 20 kadar kişi de müzakereci olarak katılmıştır. Beş oturum şeklinde gerçekleştirilen çalıştay davetli müzakereci ve dinleyicilerle gün boyu devam etmiştir. Bu yazıda, oturumlardaki sunumlar ve sunumlar sırasında gerçekleştirilen tartışmalar literatürden de katkılar sunularak değerlendirilmiştir. El yazısı metin tanıma teknolojileri, dijital Osmanlıca metin külliyatı oluşturma çalışmaları, dil işleme meseleleri, Osmanlıca metinlerin makinelerce yapılan otomatik transkripsiyonuna dair mevcut projeler ve uygulamalar, segmentasyon, çözünürlük, sentetik veri, veri artırma, büyük dil modelleri, süreçlerin dokümantasyonu, oluşturulabilecek disiplinlerarası iş birlikleri gibi konulara temas edilerek geleceğe dair önerilere yer verilmiştir. Bu değerlendirme ile Osmanlıca metinler başta olmak üzere Arap harfleri kullanan diller üzerine çalışmalar yürüten farklı disiplinlerden uzmanlar için sahadaki son durum ortaya konulmuş ve kayda geçirilmiş olacaktır.

Bu çalıştayın ana amacı tarihi, sosyolojik, ekonomik ve kültürel değeri bulunan Osmanlıca metinlerin (belge, kitap gibi çeşitli formatlardaki kaynakların) makinece okunabilmesi yönündeki çalışmalar, sorunlar ve çözüm önerilerini

4 Todd Presner, and Chris Johanson, *The Promise of Digital Humanities: A Whitepaper*, March 1, 2009-Final version. <http://www.itpb.ucla.edu/documents/2009/PromiseofDigitalHumanities.pdf>.

5 Sümeyye Akça, “A Conceptual Model to Increase the Visibility and Usage of Cultural Heritage Objects: The Case of UNESCO’s Memory of the World List”, *Digital Scholarship in the Humanities*, 36/1 (2021), s. 16-31.

6 Yunus Uğur (ed.), *Dijital Beşerî Bilimler ve Osmanlı Çalışmaları* (İstanbul: Vakıfbank Yayınları, 2023).

tartışmaktır. Bu kaynakların pek çok disiplinde veri olarak kullanılabilmesi ve üzerinde algoritmaların ve dijital insanî bilimler araçlarının kullanılabilmesi için kaynakların makinece okunabilir formatla dijitalleştirilmesi gerekmektedir. Osmanlıca matbu eserlerin dijital transkripsiyonu üzerinde pek çok girişim, çalışma ve platform bulunmakla beraber el yazması metinlerin transkripsiyonu üzerine çalışmalar oldukça kısıtlıdır. Sunumlar, alanda çalışmalar yapan araştırmacılar ve akademisyenler tarafından gerçekleştirilmiş olup çalıştay sonucunda mevcut durum değerlendirmesi yapılmıştır.

El Yazısı Metin Tanıma Teknolojisi (Handwritten Text Recognition-HTR)

Dr. Öğr. Üyesi Sümeyye Akça “Dijital Metin Transkripsiyonunda Son Gelişmeler” başlıklı açış konuşmasında el yazısı metinlerin transkripsiyonlarında kullanılan çevrimiçi ve çevrimdışı teknolojiye, yöntemlere, karşılaşılan zorluklara ve literatürde farklı alfabeler üzerinde gerçekleştirilen çalışmalara değinmiştir. Açış konuşmasında üzerinde durulan başlıklar şu şekildedir:

El yazısı metinlerin otomatik transkript edilmesi ve makinece okunabilir, üzerinde işlem yapılabilir birer dijital metne dönüştürülmesi bilgiye erişimin kolaylaşması ve yeni bilgilerin keşfedilebilmesi gibi durumlar açısından oldukça önemli bir süreçtir. Son zamanlarda bilgi teknolojilerinde gelişmelerle paralel olarak pek çok alanda el yazısı tanıma teknolojisi kullanılmakta ve sistem teknolojiye paralel düzeyde geliştirilmektedir. Dünyada alandaki uygulamalara baktığımızda çevrimiçi ve çevrim dışı olmak üzere iki çeşit el yazısı tanıma çalışmalarından bahsedilebilir. Çevrimiçi teknolojiye özel bir sayısallaştırıcı üzerindeki kalemlerin hareketinden algılama, çevrim dışında ise fiziki materyallerden (kağıt, resim gibi) tarama söz konusudur.⁷ Bu uygulamalarda derin öğrenme mimarileri ile (Evrişimli Sinir Ağları (Convolutional Neural Networks), Derin Evrişimli Sinir Ağları (Deep Convolutional Neural Networks), Çok Boyutlu Tekrarlayan Sinir Ağları (Multi-Dimensional Recurrent Neural Networks-RNNs), Kodlayıcı-Kod Çözücü (Encoder-Decoder), Başka bir RNN türü olan Dikkat Ağları (Attention Networks), Graph Neural Networks, Örüntü Tanıma (Pattern Recognition), Dönüştürücü Modeller (Transformer Models), Hidden Markov Models gibi) bağlamın dikkate alınması transkripsiyonun başarımlarını artırmaktadır. Bu gelişmelerle el yazısı tanıma teknolojisi günümüzde finans, sağlık, kamu, güvenlik, kütüphane, lojistik (posta ve kargo) gibi pek çok alanda kullanılmaktadır.

7 Pragati Baheti, “Handwriting Recognition: Definition, Techniques & Uses”, <https://www.v7labs.com/blog/handwriting-recognition-guide>. (2 Aralık, 2022).

El yazısı tanıma teknolojisinin genellikle Latin alfabesi için başarımları oranın daha yüksek ve uygulama sahası daha geniştir. Yine de literatürde farklı alfabeler için gerçekleştirilmiş çalışmalar da bulunmaktadır. Çince, Norveççe, Arapça, Bengal dili ve Hintçe için yapılan çalışmalar olduğu bilinmektedir.⁸ Her ne kadar başarımları gelişen teknolojiyle beraber artmış olsa da el yazması metinlerin dijital metne dönüştürülmesi ve transkript edilmesinde birtakım zorluklarla karşılaşmaktadır. Bu zorluklar şu şekilde ifade edilebilir:

- Metinlerde zaman içerisinde bozulmalar ve desenlerinde değişiklikler meydana gelebilmektedir. Bu durum ilgili algoritmalar tarafından yazının doğru bir şekilde anlaşılmasını olumsuz etkilemekte dolayısıyla başarılı bir transkripsiyon ve dijital metin dönüşümü zorlaşmaktadır.
- Metinlerin farklı yazı stillerinde yazılması, bu yazıları yazan kişilerin değişebiliyor olması ve metni kaleme alan kişinin zaman içerisinde el yazısının değişmesi de karşılaşılan zorluklardandır.
- El yazısı metinlerin belirli bir satır düzenine bağlı kalınmadan oluşturulduğu durumlar da mevcuttur. Ayrıca sayfanın kenarlarına not ya da farklı ek metinler de yazılabilmektedir. Bu sebeple metinler algoritmalar tarafından tam olarak anlaşılammamaktadır ve başarılı bir transkripsiyon işlemi elde edilememektedir.

8 Shi Yan, Jin-Wen Wu, Fei Yin, and Cheng-Lin Liu, “Recognizing Handwritten Chinese Texts with Insertion and Swapping Using a Structural Attention Network”, *Document Analysis and Recognition—ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings*, Part IV, ed. Josep Lladós, Daniel Lopresti, Seiichi Uchida (Cham: Springer International Publishing, 2021), s. 557-571; Md. Mahbubar Rahman, M. A. H. Akhand, Shahidul Islam, Pintu Chandra Shill, and M. M. Hafizur Rahman, “Bangla Handwritten Character Recognition Using Convolutional Neural Network”, *International Journal of Image, Graphics and Signal Processing*, 8 (2015), s. 52-59; Martin Maarand, Yngvil Beyer, Andre Kåsen, Knut T. Fosseide, Christopher Kermorvant, “A Comprehensive Comparison of Open-Source Libraries for Handwritten Text Recognition in Norwegian”, *Document Analysis Systems: 15th IAPR International Workshop, DAS 2022, La Rochelle, France, May 22–25, 2022, Proceedings*, (Cham: Springer International Publishing, 2022), s. 399-413; A. Bharath, and Sriganesh Madhvanath, “Online Handwriting Recognition for Indic Scripts”, *Guide to OCR for Indic Scripts: Document Recognition and Retrieval*, ed. Venu Govindaraju, Srirangaraj Setlur (Springer Science & Business Media, 2010), s. 209-234; Minhaz Kamal, Fairuz Shaiara, Chowdhury Mohammad Abdullah, Sabir Ahmed, Tasnim Ahmed, and Md. Hasanul Kabir, “Huruf: An Application for Arabic Handwritten Character Recognition Using Deep Learning”, *25th International Conference on Computer and Information Technology (ICCIT)*, (2022), s. 1131-1136.

- Metinlerin çeşitli süslemelere ve kendine has yazım özelliklerine sahip olması, algoritmaların metinleri algılayamamasına neden olmaktadır. Farklı yüzyıllara ait metinlerin farklı şekillerde kaleme alınması da karşılaşılan bir diğer zorluktur.
- Makine öğrenimi esnasında kullanılan veri setinin homojen dağılımı öğrenmeyi zorlaştırmaktadır. Algoritmalar ya da transkripsiyon araçlarıyla az sayıda ve birbirine benzer metin verileri üzerinde çalışılması yapılan transkripsiyon işleminin ezberlenmesine sebep olmaktadır ve bu durum ezberleme (overfitting) olarak ifade edilmektedir. Böyle bir işlem sonrasında doğruluk oranı yüksek bir transkript elde edilebilir fakat bu işlemin başarı oranı, sadece üzerinde çalışılan kısıtlı sayıdaki veri seti ve benzer metin verileri için geçerli olmaktadır. Çalışmaya farklı yeni metin verileri eklendiğinde başarılı transkript etme oranının düşmesi oldukça muhtemel bir durumdur. Dolayısıyla farklı tür ve yapıdaki metinleri çalışmalara dahil etmek gerekmektedir.
- Arapça metinlerin sağdan sola yazılmaları ve sesli harflerde hareke kullanılması da Arap alfabesiyle yazılan metinlerin dijital transkripsiyonunda önemli zorluklardandır.

Tüm bu zorluklarla beraber gelişen teknolojinin kullanılmasıyla dünyada pek çok kültürel bellek kurumu yaptıkları projelerle koleksiyonlarındaki el yazısı metinlerin dijital kütüphanelerine eklemişlerdir. Örneğin Europeana öncülüğündeki Transcribathon projesi ile pek çok el yazması eser ve belge gönüllüler tarafından transkribe edilmiş ve projenin devamında sistem, Read-Coop şirketi tarafından geliştirilen Transkribus ile entegre edilmiştir.⁹ Yine Norveç Milli Kütüphanesi (National Library of Norway) Transkribus aracı kullanarak geliştirdikleri Norveççe el yazısı tanıma modeli kullanarak pek çok el yazması eseri dijital kütüphanelerinde kullanıcıya açmışlardır.¹⁰

Osmanlıca Metin Külliyyatı Oluşturma, Erişim ve Dil İşlemede Problemler

Osmanlıca olarak yazılmış standart kaynakların yanı sıra standardın dışında notlar, yemek tarifleri, şiirler gibi pek çok tarihi doküman da mevcuttur. Bu kısa

9 “Europeana”, *Transcribathon*. <https://europeana.transcribathon.eu/about-us/>. (11 Ocak, 2024).

10 Yngvil Beyer, “Handwritten Text Recognition at the National Library of Norway”, <https://cerlblog.wordpress.com/2021/06/24/handwritten-text-recognition-at-the-national-library-of-norway/>. (24 Haziran, 2021).

kısa not şeklinde oluşturulmuş kaynakları temiz dijital Osmanlıca metin külliyyatı oluşturmak ve erişim noktasında değerlendirmek önemlidir. Bununla ilgili çalıştayın ikinci oturumunda Viyana Üniversitesi'nden Dr. Ercan Akyol Viyana Üniversitesi tarafından desteklenen ve hakemli bir dergi olan *Keshif* dergisi çalışmalarından bahsetmiştir.

Dergide yayımlanan metinler organiktir. Bu metinler tek başına bir anlam ifade etmeyen bilgiler-kayıtlar olup bir araya geldiklerinde bir konu, durum ya da kişi hakkında önemli kaynak oluşturabilecek metinlerdir. Örneğin metnin bağlamından ayrı olarak not alınan bir yemek tarifi bu tarz metinlere örnektir. *Keshif* dergisi bu metinleri bir araya getirerek benzer durumlar için geniş bir veri tabanı hizmeti sunmayı hedeflemektedir. *Keshif* dergisi temelinde oluşturulacak veri tabanında örneğin belirli bir yüzyıl için yemek tariflerine, anahtar kelimelerle ya da bağlama göre erişim hedeflenmektedir. El yazmalarındaki notların tüm külliyyatla birlikte ele alınması; yani notun elde edildiği kaynağın künye bilgisinin verilmesi erişim açısından önemli görülmektedir. Dergiye gönderilen kaynaklar üzerinde analiz yapılabilmesi açısından metinlerin yapısal formatlarda indirilebilmesi (xml, html gibi) yapılması gerekenler arasına not edilmiştir.

Yılda iki kere yayımlanan *Keshif* dergisi standart dışı Osmanlıca kaynaklar için önemli bir veri tabanı olma yolundadır. Dergi filoloji esaslarında hassas olduğu için yazılar klasik transkripsiyon kurallarıyla çıkmaktadır. Diğer yandan, derginin yurtdışında İngilizce olarak yayımlanması, yapılan tercüme çalışmalarının (Osmanlıca külliyyatın) dil bariyeri olmaksızın daha geniş çevrelerce ulaşılabilir olmasını mümkün kılması açısından önemlidir. Ayrıca dergiye gönderilen yazılarda okunamayan kısımlar için gönüllülerin de yardım edebileceği bir yapı (crowdsourcing-kitle kaynak) oluşturulmasının da hedeflendiğini ifade edilmiştir. Fakat derginin bir veri tabanı haline dönüşmesinin uzun zaman alacağı bu sebeple ilerleyen süreçte bir proje olarak değerlendirilmesinin düşünüldüğü belirtilmiştir.

Osmanlıca metinleri okumayı kolaylaştıran ve 2017 yılından beri Osmanlıca metinlerin transkripsiyonu için önemli bir girişim olan LexiQamus ve metin neşri için yeni geliştirdikleri projesi BuildLQ ile ilgili çalışmalar çalıştayın ilk oturumunda Dr. Ahmet Abdullah Saçmalı tarafından paylaşılmıştır. Saçmalılı "Osmanlıca Dil İşlemede Karşılaşılan Problemler ve Çözümler" başlıklı sunumuna kendi girişiminde karşılaştığı problemlere atıfta bulunarak "kusursuz dijitalleştirme"nin pek de mümkün olmadığı yönünde görüş bildirmiştir. LexiQamus projesinin mutfak kısmında hangi iş ve işlemlerin yapıldığı yönünde

örneklerle, makineye öğretilen transkripsiyonların da yanlış yapılabileceği ve makinenin de yanlış öğreneceği gibi argümanları desteklemek için Osmanlıca metin okumalarda bir harfin dahi yanlış okunduğu durumların hangi sonuçları doğurabileceği noktasında tereddütlerden bahsedilmiştir (Örneğin, İraniye kelimesinin ünlü bir tarihçi tarafından İbraniye olarak okunmasının metnin Yahudilik kapsamında yorumlanmasına sebep olması). Bu ve benzeri hataların en aza indirilmesi için LexiQamus'da her bir kelimenin kaynakta yazılı olan asıl halinin kullanıcıya gösterildiği, ayrıca bir kelimenin okunuşunun üç-dört ayrı kişi tarafından onaylandıktan sonra sisteme eklendiği (öğretildiği) bilgisi paylaşılmıştır. Bu durumun aslında el yazısı tanıma programları için uygulanan bir yöntem olduğu, Google'ın ReCAPTHCAs uygulaması ile kitle kaynaktan yararlanarak sisteminin optimize edildiği belirtilmiştir. ReCAPTHCAs uygulamasıyla, dijital görüntüleri alınan basılı malzemenin Optik Karakter Tanıma (Optical Character Recognition) ile okunamayan kelimeleri internet kullanıcılarına sorulmuş böylelikle de pek çok kitabın transkribe edilmesi sağlanmıştır (Von Ahn ve diğerleri, 2008). BuildLQ'da ayrıca üzerinde işlem yapılan metnin beş ayrı kısma ve ardından sütunlara bölünerek dilimlendiği, böylece kelimelerin hem bağlam içerisinde değerlendirilmesinin hem de tek tek incelenmesinin mümkün kılındığı belirtilmiştir. Bir kelimenin varsa farklı yazım şekillerinin de okuyucuya sunulduğu ve bu farklı şekillerin de aranabilir olarak veri tabanında yer aldığı ifade edilmiştir (Örneğin kesüp-kesüb-kesip kelimeleri farklı yazıldığı halde Türkçe'deki kesip kelimesidir).

Osmanlıca Metinlerin Makinelerce Transkripsiyonu

Çalıştay'ın ikinci oturumunda hem matbu Osmanlıca eserlerin dijital transkripsiyonu hem de el yazması eserlerin dijital transkripsiyonu noktasında uzun yıllardır çalışmaları bulunan İstanbul Üniversitesi-Cerrahpaşa Öğretim Üyesi Sayın Doç. Dr. Atakan Kurt el yazması eserlerin dijital transkripsiyonu üzerine yeni başladıkları çalışma hakkında bilgiler paylaşmıştır. Osmanlıca matbu ve el yazısı tanıma projesi olan Osmanlıca.com sitesinin ara yüzünü ve programda henüz kullanıma sunulmayan deneme aşamasındaki el yazması metinlerin transkripsiyon işlemlerinde elde edilen başarı oranlarını metin örnekleriyle birlikte paylaşan Kurt, projenin ana amacının Osmanlıca belgelerin modern Türkçeye çevrilmesini sağlamak olduğunu ifade etmiştir.

Bu çalışmada ilk olarak bir dokümanın fotoğrafının Optik Karakter Tanıma (OCR) ya da El Yazısı Metin Tanıma (HTR) vasıtasıyla makinelerce anlaşılabilir

hale getirilmesi sağlanmaktadır. Fakat bilhassa el yazısı Osmanlıca metinlerin çeşitliliği gibi noktalarda farklı yazı çeşitleri ya da belge türleri için farklı doğruluk yüzdeleri ile sistem öğrenmeye devam etmektedir. Sistemin öğrenmesi aşamasında veri olarak daha önce elle transkripsiyon çalışması yapılan tezlerin (şeriiye sicillerinde olduğu gibi) indirilip makineye tanıtıldığı belirtilmiştir. Sistemde ayrıca Osmanlıca bir metin Türk alfabesine çevrilebilmekte olup; projenin ilerleyen süreçlerinde ise metnin Türkçeye tercüme edilmesi de hedeflenmektedir. Osmanlıca.com projesinin iş paketlerinin üç adıma sahip olduğuna değinen Kurt bu adımları şu şekilde açıklamıştır:

1. Veri setini hazırlanması (10.000 sayfa ve 250.000 satır),
2. Modellerin eğitilmesi,
3. Osmanlıca.com'da yayımlanması (Şu an sadece matbu OCR çalışması mevcut).

Osmanlıca el yazısı metin tanıma (HTR) işlemi için çalışmalarının yaklaşık altı ay önce başladığı ve henüz geliştirmiş oldukları modelin yeterli sayılabilecek miktarda bir veri setiyle test edilmediği belirtilmiştir. Bu süreçte paragrafların düz bir şekilde yazılmamış olmasının çalışmada karşılaşılan önemli bir problem olduğu ifade edilmiştir. Kurt, bu çalışmadaki hedeflerini minimum bin sayfalık çeşitli (varyasyonlu) veri setiyle yaklaşık %70'lik bir başarı oranının elde edilmesi şeklinde belirtmiştir.

Osmanlıca metinlerin dijital transkripsiyonu çalışmalarını ilk başlatan kurumlardan biri de IRCICA'dır. IRCICA'da 2016 yılında başlayan çalışmalar iyi bir seviyeye geldikten sonra zamanla yavaşlamıştır. IRCICA'nın matbu eserlerin dijitalleştirilmesi çalışmaları dışarıdan erişime kapalı olup yapılan çalışmalar hakkında IRCICA kütüphanesi Müdür Yardımcısı Neslihan Sandal üçüncü oturumun ilk sunumunda konuyla ilgili çalışmayı paylaşmıştır. Kütüphane dermesi üzerinde yaptıkları OCR işlemlerinin süreçleri hakkında bilgiler paylaşan Sandal, bu süreçler sonunda elde edilen başarı oranlarının yüksek olduğuna dikkat çekmiştir. Ayrıca her bir sayfaya ayrı şekilde OCR yapıldığı ve belgenin üst verilerinin de sisteme girildiği belirtilmiştir. Kullanıcılardan gelen talep üzerine IRCICA'da OCR ile dijital metin haline dönüştürülen Osmanlıca eserlerin Latin harflerine çevrilmesi projesine de başlanmıştır. El yazması belgeler için ise nesih, rika, talik ve nestalik gibi yazım şekilleriyle kaleme alınan el yazmaları üzerinde çalışmalara başladıkları ve bu işlemler sonucunda yaklaşık %60 oranında bir başarı oranı elde ettikleri belirtilmiştir. Oluşturdıkları veri tabanında Arap ya da Latin harfleriyle

arama yapılabildiği, kaynaklarını dünyadaki tüm araştırmacılara açık erişim olarak sundukları ve Osmanlıca bilmeyen kullanıcılarının da kaynaklardan istifade edebilmesini sağlamayı amaçladıkları sunumda ifade edilmiştir.

Tüm bu girişimlere rağmen Osmanlıca el yazmalarının dijital transkripsiyonunda pek çok zorlukla karşılaşmaktadır. El yazısı tanıma programlarında sıklıkla karşılaşılan önemli bir problem bilhassa Osmanlıca el yazmaları için segmentasyon (bölütleme) yani satırların düzgünce ayrılabilmesi problemidir. Bununla ilgili literatürde düz satır şeklinde yazılmayan el yazmalarından nasıl bilgi çekilebileceği ve makineye tanıtılabileceği üzerine çalışmalar bulunmaktadır. Örneğin tarihi el yazması metinlerde yer alan tablolardan bilgilerin çekilerek otomatik olarak transkripsiyonu, ya da satırın kenarına üstüne ya da altına alınan notlar, eklemeler olduğu zaman nasıl bir segmentasyon uygulanacağı üzerine çalışmalar bulunmaktadır.¹¹ Bu bağlamda İstanbul Teknik Üniversitesinde yapılan bir yüksek lisans tezi tez danışmanı Prof. Dr. Behçet Uğur Töreyin tarafından anlatılmıştır. Octavio Ernesto Romo Rodriguez'in yaptığı tezin amacının, bir derin öğrenme modeliyle (Inception) Osmanlıca *Ameli Elektrik* dergisinin kapak sayfalarının dökümü ve sayfa düzeninin bölütlenerek metni ayrı, metin dışı unsurları (resim gibi) ayrı bir bölüt olarak çıkarmak olduğu belirtilmiştir. Kapak sayfalarındaki metnin, imaj ya da fotoğrafların ayrı ayrı çıkarılması üzerine yapılan çalışmada modelin daha iyi işleyebilmesi açısından ön işlemede birbiri üzerine gelen elemanlar tek tek çıkarılmıştır. 20 belge üzerinde 300 adet her bir sınıfa ilişkin elemanla transfer öğrenme yapıldığında makinenin metin ve metin dışı şeklinde ayrı iki bölüt oluşturabildiği belirtilmiştir.¹²

Osmanlıca el yazmasının otomatik transkripsiyonunu zorlaştıran bir diğer etmen ise metinlerdeki çözünürlük düzeylerinin düşük olması, gürültü ve metin üzerindeki metni okumayı zorlaştıran bozulmaların olmasıdır. Bu durum özellikle, çeşitli olumsuz deformasyon, aşınma ve bozulmalardan muzdarip olan düşük kaliteli tarihi belgeler için geçerlidir. Tarama cihazı özellikleri ve çevresel aydınlatma seviyeleri gibi faktörler tarihi belgelerin çözünürlüğünü ve kalitesini doğrudan etkileyerek içeriklerinin okunmasını, işlenmesini ve analiz edilmesini

11 Yan, Wu, Yin, and Liu, "Recognizing Handwritten Chinese Texts"; Jose Ramón Prieto, José Andrés, Emilio Granell, Joan Andreu Sánchez, and Enrique Vidal, "Information Extraction in Handwritten Historical Logbooks", *Pattern Recognition Letters*, 172 (2023), s. 128-136.

12 Octavio Ernesto Romo Rodriguez, "Computer Vision Based Segmentation of Printed Documents from the Late Ottoman Empire and Early Turkish Republic Eras" (yüksek lisans tezi), İstanbul Teknik Üniversitesi, 2023.

zorlaştırmaktadır. Artvin Çoruh Üniversitesinden Dr. Öğr. Üyesi Hakan Temiz ikinci oturumun son sunumunda “Derin Öğrenmeye Dayalı Süper Çözünürlük Teknikleri Kullanılarak Tarihi Osmanlı Metinlerinin Çözünürlüğünün Arttırılması” başlıklı çalışmasının verilerini paylaşmıştır. Çalışmasında düşük çözünürlüğe sahip bir belge görüntüsünün “süper çözünürlük” olarak ifade edilen yüksek çözünürlüklü bir belgeye dönüştürebildiğini belirten Temiz, bu işlemlerde kullanılabilen yapılardan bahsetmiştir. Çözünürlüğün derin sinir ağları teknikleriyle bir problem olmaktan çıktığı literatürde pek çok çalışmayla ortaya konulmuştur.

Süper çözünürlük, görüntüye özgü bir öncül bilgiden yararlanarak çözünürlüğü artıran algoritmik bir çözümdür. Günlük fotoğraflar, havadan görüntüleme ve tıbbi görüntüleme dahil olmak üzere çeşitli görüntüleme yöntemlerinde yaygın olarak uygulanır.¹³ Bu yöntemler ya aynı imajın iç benzerliklerinden yararlanır ya da harici düşük ve yüksek çözünürlüklü örnek çiftlerden eşleme işlevlerini öğrenir.¹⁴ Osmanlıca belgelerin çözünürlüklerinin artırılması üzerine yapılan söz konusu çalışmada ise dört farklı süper çözünürlük modeli (VDSR, SRCNN, DECUSR, RED-Net) uygulanmış ve Wikilala’dan 1166 adet tarihi metin görseli (imaj) örnek olarak kullanılmıştır. Bu verilerin 966 tanesi eğitim, 100 tanesi test ve 100’ü ise doğrulama verisidir. Temiz, dört farklı teknikle görüntülerin çözünürlüklerinin artırıldığı örnekler göstermiştir. Çalışmanın neticesinde, öğrenme oranı $1E-4$ ve görüntü boyutu 16×16 olan imajlarda en iyi öğrenmeyi modifiye DECUSR modeli gerçekleştirebilmiştir. Temiz, bu çalışmanın amacını tarihi bir görüntüdeki detayların belirgin hale getirilerek iyileştirilmesi, daha iyi görüntüler elde edilerek belge üzerinde başarılı OCR ya da HTR işlemlerinin yapılabilmesi olduğunu ifade etmiştir.

El yazması tanıma uygulamalarında bilhassa Latin alfabesi dışındaki diller için makineye öğretilecek verinin az olması ciddi bir problem olarak adreslenmiştir.¹⁵ Bu sorunu gidermek için veri setleri geliştirilmekte ya da farklı uygulamalar denenmektedir. Dr. Öğr. Üyesi Esmâ F. Bilgin Taşdemir son oturumda

13 Hakan Temiz, “Enhancing the Resolution of Historical Ottoman Texts Using Deep Learning-Based Super-Resolution Techniques”, *Traitement du Signal*, 40/3 (2023), s. 1075.

14 Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, “Learning a Deep Convolutional Network for Image Super-Resolution”, *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*, ed. David Fleet, Tomas Pajdla, Bernt Schiele, Tinne Tuytelaars (Cham: Springer International Publishing, 2014), s. 184.

15 Per E. Kummervold, Javier de la Rosa, Freddy Wetjen, and Svein Arne Brygfeldt, “Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model”,

“Sentetik Veri (Synthetic Data) ve Veri Artırma (Data Augmentation) Yöntemi ile Osmanlıca Metinlerin Transkripsiyonu” adlı çalışmasını paylaşmıştır. El yazısı tanıma sistemleri eğitilirken kullanılan veri artırma ve sentetik veri yöntemleri sadece veri girdisini çoğaltmak için değil aynı zamanda sentetik varyasyonlardan (çeşitlemeler) dolayı desenlerdeki soyutlama düzeyini genişletmek için de kullanılmaktadır.¹⁶ Sentetik veri kümesi; mevcut metin görüntülerinden kelime veya karakterleri kırarak kelime ve satırlara dönüştürme, metin görüntülerini farklı arka plan ve görüntülerle harmanlama, belirli bir metnin görüntülerini oluşturma ve görüntüden görüntüye çeviri şeklinde farklı yöntemlerle oluşturulabilir. Bu yöntemlerin her biri farklı sorunlar için farklı uygulanabilirliğe sahip yaygın yaklaşımlardan bazılarıdır.¹⁷ Taşdemir, Osmanlıca üzerinde özellikle nesih yazı stilineki belgelerde veri artırma yoluyla zenginleştirilen veri setinin makineye tanıtıldığı çalışmanın sonuçlarını paylaşmıştır. Verilerin etiketlenmesinin oldukça zorlu ve fazlaca iş gücü gerektiren bir işlem olduğunu ve sentetik veriyle eldeki etiketli verinin artırılabilirdiğini ifade eden Taşdemir, ayrıca eldeki verinin çeşitli varyasyonlarının da oluşturulmasının bu yolla mümkün olduğunu belirtmiştir. Çalışmalarda kelimelerin varyasyonun artmasının Osmanlıca gibi farklı el yazısı kullanılan, ya da sürekli katibi değişen durumlar için başarı oranında artış sağlayacağı özellikle belirtilmiştir.

Osmanlıca Metinlerin Transkripsiyonunda Kullanılan Araçlar

Dünyada el yazması eserlerin otomatik olarak transkripsiyonu amacına hizmet eden uygulamalar bulunmaktadır ve bu uygulamalar pek çok önde gelen kütüphane tarafından koleksiyonlarında bulunan el yazması eserlerin dijital kütüphanelerine eklenmesi amacıyla kullanılmaktadır (Transcribathon, Connecticut Dijital Arşivi, Norveç Milli Kütüphanesi gibi). Bu uygulamalardan halihazırda en yaygın olarak kullanılanı ise Transkribus programıdır. Transkribus, herhangi bir yerden, herhangi bir zamanda ve herhangi bir dilde metin tanıma,

Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), (2021), s. 20–29. <https://aclanthology.org/2021.nodalida-main.3/>

16 Riaz Ahmad, Saeeda Naz, Muhammad Afzal, Sheikh Rashid, Marcus Liwicki, and Andreas Dengel, “A Deep Learning based Arabic Script Recognition System: Benchmark on KHAT”, *The International Arab Journal of Information Technology*, XVII/3 (2020), s. 299-305.

17 Esma F. Bilgin Tasdemir, “Printed Ottoman Text Recognition Using Synthetic Data and Data Augmentation”, *International Journal on Document Analysis and Recognition (IJDAR)*, (2023), s. 1-15.

transkripsiyon ve tarihi belgelerin aranması için yapay zeka destekli bir platformdur.¹⁸ Osmanlıca el yazısı metinlerin bu platform aracılığıyla transkripsiyonu üzerine yapılan bir çalışmada Osmanlıca metinlerin transkripsiyonundaki öncül zorluğun karakterin tanımlanması değil karakterin modern Türkçedeki formunun bulunması olarak gösterilmiştir. Osmanlı Türkçesi ve Modern Türkçe alfabeleri arasında bire bir örtüşme olmaması, birden fazla olası transkripsiyon çıktısına yol açmaktadır. El Yazısı Tanıma programı bir Osmanlı Türkçesi karakterini doğru bir şekilde tanımlasa bile, doğru okuma yine de hassas seslendirmeye ve çok sesli harflerin bağlama uygun telaffuzuna bağlıdır. Literatürde konu üzerine yapılan bir çalışmada, hiçbir derin öğrenme aracının Osmanlıca üzerinde kusursuz bir transkripsiyon gerçekleştirmesinin mümkün olmadığı belirtilmiştir.¹⁹ Bu bağlamda Elif Can Derin ve Fatma Aladağ, “Osmanlı Türkçesi Metinlerin Transkribus ile Otomatik Transkripsiyonu ve Dijital Edisyon” adlı sunumlarında Transkribus uygulamasının Osmanlıca eserlerin dijital transkripsiyonu açısından nasıl kullanılması gerektiğini ve uygulamanın metinler üzerinde yapılabilecek analizler açısından nasıl ek faydaları olduğu gibi bilgileri paylaşmışlardır. Örneğin metin kodlama (Text Encoding Initiative-TEI) ile birlikte tematik indeksleme ve kelime aramasının, ayrıca yer ve mekân haritalandırmasının da yapılabileceği belirtilmiştir. Sunumda bu gibi programları kullanmanın en büyük dezavantajlarından birinin makineye öğretilen verilerin standart bir yapıda transkripsiyonunun olmaması olarak belirtilmiştir. Transkripsiyonda belirli bir standardın olmaması (farklı alanlar ya da ekollerin farklı transkripsiyon standardı kullanması) ve fontların standartlara uygun olmaması dijital transkripsiyonu zorlaştırmaktadır.

Çalıştay’ın son oturumunda dijitalleştirilen metinler üzerinde büyük dil modellerinin nasıl bir etki oluşturacağı ya da dijital transkripsiyonda büyük dil modeli kullanımı yönünde ne gibi beklentiler olduğu tartışılmıştır. Dr. Öğr. Üyesi Müge Akbulut “Büyük Dil Modelleri ve Bilişimsel Erişim” başlıklı sunumunda genel olarak büyük dil modellerinin yapısını oluşturan Transformers mimarisini tanıtarak, büyük dil modellerinin günümüzde nerelerde kullanıldığına değinmiştir. Metin içerisindeki her bir kelimeyi bağlama göre değerlendirebilen büyük

18 Transkribus. <https://readcoop.eu/transkribus/>. (1 Şubat, 2024).

19 Suphan Kirmizialtin, and David J. Wisley, “Automated Transcription of Non-Latin Script Periodicals: A Case Study in the Ottoman Turkish Print Archive”, *Digital Humanities Quarterly*, 16/2 (2022). <http://www.digitalhumanities.org/dhq/vol/16/2/000577/000577.html>.

dil modellerinden²⁰ metin sınıflama ve özetleme gibi pek çok alanda faydalana-
bileceğimiz gibi farklı dillerdeki dijitalleştirilmiş materyal üzerinden (Gürültülü
OCR’ların dahi iyi sonuçlar verdiği görülmüştür) model eğitimi ve yeni model ge-
liştirme gibi amaçlarla da çalışmalar yapılmaktadır.²¹ Büyük dil modellerinin test
edilmesi ve analizlerin yapılabilmesi için Hugging Face platformunun önemine
değinen Akbulut, yakında kullanıma açılacak olan Google’ın Gemini dil mode-
line de değinmiştir. Gemini, sıfırdan çok modlu olacak şekilde inşa edilmiş olup;
metin, kod, ses, görüntü ve video dahil olmak üzere farklı bilgi türlerini anlayarak
genelleştirebilen bir modeldir.²²

Sonuç ve Öneriler

Osmanlıca Metinlerin Dijital Transkripsiyonu Çalıştayı’nda gün boyunca ya-
pılan sunumların sonunda genel bir değerlendirme yapılmıştır. Değerlendirme ve
öneriler şöyle ifade edilebilir:

- 1- Osmanlıca metinlerin dijital transkripsiyonu için tek bir çözümün olma-
dığı daha bütüncül bir yaklaşım sergilenmesi gerektiği görüşü öne çık-
mıştır. Otomatik transkripsiyon çalışmaları şu anda iki kulvardan devam
etmektedir. Meseleyi bütüncül olarak ele alan ve metinlerin dijitalleştiril-
mesini, OCR’lanmasını, transkripsiyonu, makine eğitimi ve bugünkü
dillere aktarılmasını hedefleyen ve kendi yazılım ve editör programlarına
sahip projeler (IRCICA OCR, BuildQ, Osmanlıca.com gibi) ilk kulvar-
dır. Hazır yazılımlar ya da altyapılar kullanıp burada doğrudan Osmanlıca
metinleri eğiten ve OCR, HTR, otomatik transkripsiyon ve diğer metin
analiz işlemlerini yapmaya çalışan projeler (Transkribus gibi) de ikinci
kulvar olarak değerlendirilebilir. Bu her iki kulvarda çalışanların iş birliği
yapması, mümkünse ülkemizdeki büyük bilimsel kurumların desteği ile
tek çatı altında bu çalışmaların toplanması ve proje süreçlerinin dokü-
manite edilip yayınlaması son derece önemli görülmektedir.

20 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is All you Need”, *Neural Information Processing Systems* (2017).

21 Martin Frické, *Artificial Intelligence and Librarianship*, (2023), <https://softoption.us/sites/default/files/AlinLibrariesNotesForTeaching.pdf>; Kummervold, Rosa, Wetjen, and Brygfjeld, “Operationalizing a National Digital Library”.

22 Google DeepMind, <https://deepmind.google/technologies/gemini/#introduction>. (12 Şubat, 2024).

- 2- Otomatik transkripsiyon için gerekli makine eğitimi için temiz ve çeşitlik barındıran metinlerin elde olması önemlidir. Bununla birlikte sentetik veri ve veri artırma teknikleri de kullanılarak makine eğitimine veri sağlanabilir.
- 3- Makine eğitimi sırasında karşılaşılan segmentasyon (bölütleme) meselesi otomatik transkripsiyon sürecinin çok önemli bir boyutudur. Bununla ilgili metin ve görseller için çözüm sunan bir kısım çalışma mevcuttur bunların geliştirilmesi gerekir.
- 4- Osmanlıca metinlerde ve özellikle görsellerde karşılaşılan “gürültü”lerin ve diğer çözünlülük sorunlarının süper çözünlülük yöntemleri ile giderilmesi mümkündür.
- 5- Makine eğitimi ve Osmanlıca dil modelinin oluşmasında büyük dil modelleri ve transformers alanındaki gelişmeler dikkatle takip edilmeli ve dijital transkripsiyon ve neşir süreçlerine entegre edilmelidir.

Özetle çalıştayda bilimsel çalışmaların birikimli olması yasasından hareketle çeşitli kurum ya da farklı disiplinlerden kişiler tarafından yapılan çalışma ve girişimlerin muhakkak bir raporla literatüre kazandırılması önemli görülmektedir. Sosyal ve beşeri bilimciler ile bilişim alanındaki uzmanların etkileşiminin sağlanması ve ihtiyaçların doğru konumlandırılması için hem bu raporların hem de iki tarafı anlayabilecek kişilerin varlığı hayatidir. Gerçekleştirilen bu tür çalıştayların oluşabilecek iş birliklerine zemin hazırlaması açısından mühim olduğu ve dolayısıyla konu üzerinde çalıştay/konferansların düzenlenmesinin devam ettirilmesinin gerekliliği vurgulanmıştır.