ULUSLARARASI SOSYAL BİLİMLER DERGİSİ
HUMANITAS
INTERNATIONAL JOURNAL OF SOCIAL SCIENCES

ISSN: 2645-8837
HUMANITAS - Uluslararası Sosyal Bilimler Dergisi
HUMANITAS – International Journal of Social Sciences
2021, *9*(18), 273-288

Dede Korkut Kitabı'nın Dresden Nüshası ile Türkistan/Türkmen Sahra Yazmasının Bilgisayar Destekli Benzerlik Karşılaştırması

B. Tahir TAHİROĞLU[1], Şükrü Halûk AKALIN[2]

Öz

Dede Korkut Kitabı tarihsel dönemleri bakımından Türk yazı dilinin en önemli eserleri arasında yer almaktadır. Eski Anadolu Türkçesinin yazı dili özellikleri başta olmak üzere döneme ait kültürel birçok ögenin de yer aldığı eserin Dresden ve Vatikan nüshaları üzerinde dilsel özellikleri bakımından birçok çalışma yapılmış ve önemli bilgiler elde edilmiştir. 2019'da bulunan ve Türkistan/Türkmen Sahra yazması olarak adlandırılan eser, bilim dünyasında heyecan yaratmış ve bu metnin çeviri yazıları yayımlanarak metinle ilgili çalışmalar başlamıştır. Bu çalışmada hesaplamalı yöntemler kullanılarak iki metin arasındaki benzerlik oranının çıkarılması amaçlanmıştır. Çalışmanın amacı doğrultusunda Dresden nüshası temel alınarak yeni bulunan yazma arasında elde edilen benzerlik oranları kosinüs için %39, TF-IDF için %28 ve Jaccard içinse %65, %44, %3 ve %1 biçiminde hesaplanmıştır. Bulunan bu oranlara göre iki nüsha arasında biçimsel olarak benzerliğin düşük olduğu gözlenmiştir. Bu bulgular ışığında yeni bulunan yazmanın Dresden nüshasından farklı söz varlığı özellikleri gösterdiği söylenebilir.

Anahtar Sözcükler

Dede Korkut Kitabı

metin benzerliği

word2vec

dilbilim

Makale Hakkında

Geliş Tarihi: 10.01.2021

Kabul Tarihi: 24.05.2021

Doi:
10.20304/humanitas.857621

The Computational Similarity Comparision Between The Dresden and The Turkestan/Turkmen Sahara Manuscripts of The Book of Dede Qorqut

Abstract

The Book of Dede Qorqut is one of the most important works of Turkish writing language in terms of historical periods. Many studies have been carried out and important information has been obtained in terms of linguistic features on Dresden and Vatican manuscripts in Old Anatolian Turkish. In this study, it is aimed to find the similarity ratio between the two texts using computational methods. For the purpose of the study, the similarity rates obtained between the newly found manuscript based on the dresden copy were calculated as 39% for cosine, 28% for TF-IDF and 65%, 44%, 3% and 1% for Jaccard. According to these ratios, it was observed that the formal similarity between the two copies was low. In the light of these findings, it can be said that the recently founded manuscript has different vocabulary characteristics than the Dresden manuscript.

Keywords

The Book of Dede Qorqut

text similarity

word2vec

linguistics

About Article

Received: 10.01.2021

Accepted: 24.05.2021

Doi:
10.20304/humanitas.857621

[1] Dr. Öğr. Üyesi, Çukurova Üniversitesi, Fen-Edebiyat Fakültesi, Türk Dili ve Edebiyatı Bölümü, Adana/Türkiye, btahir@cu.edu.tr, ORCID: 0000-0002-7956-3257

[2]Prof. Dr., Hacettepe Üniversitesi, Edebiyat Fakültesi, Türk Dili ve Edebiyatı Bölümü, Ankara/Türkiye, sukruhaluk.akalin@hacettepe.edu.tr, ORCID: 0000-0002-5313-1763

Giriş

Dede Korkut Kitabı Türk dili, kültürü ve tarihi açısından son derece önemli verileri barındıran bir eserdir. Eser üzerinde birçok çalışma yapılmış ve eserin çeşitli yönlerinin ele alındığı çalışmaların yapılmaya devam edildiği bilinmektedir. Eserde, Türklerin yaşayış biçimlerine dair özellikler, Eski Anadolu Türkçesinin yazı diline ait çeşitli yönlerini eserde bulmak mümkündür. Bu yönüyle Dede Korkut Kitabı kültürel bir sözlük, tarihsel olarak dilbilimsel bir veri ortamı olarak düşünülebilir. 2019'da bulunan ve Türkistan/Türkmen Sahra yazması olarak adlandırılan eser, bilim dünyasında heyecan yaratmış ve bu metnin çeviri yazıları yayımlanarak metinle ilgili çalışmalar başlamıştır. Yeni bulunan tarihsel öneme sahip bir metnin önceki metinler arasındaki yerinin belirlenmesi, karşılaştırmaların yapılması ve varsa benzerliklerin ortaya çıkarılması gerekmektedir.

Son yıllarda gelişen bilgisayar teknolojisi veya en geniş ifadesiyle bilişim uygulamaları, pek çok alanda olduğu gibi dil bilgisi ve dil bilimi araştırmalarında da yeni ufuklar açmaktadır. Dilin yapısını, özelliklerini, belirli zaman dilimlerindeki değişim ve gelişimini bilgisayar destekli olarak araştırmak, dil bilgisinin ve dil biliminin çeşitli alanlarına ve konularına yönelik çalışmaları bilişim uygulamalarıyla yapmak daha ayrıntılı, geniş kapsamlı ve oylumlu sonuçlara ulaşmamızı sağlamaktadır. Bilişim alanındaki ilerlemeler ve günümüzde dil verilerinin bilgisayarda işlenmesi sonucunda yeni bir bilim dalı olarak Doğal Dil İşleme (DDİ) gelişmeye başladı. Dilbilimi de bilgisayarlı dilbilimi (*computational linguistics*), derlem dilbilimi (*corpus linguistics*) gibi üst alanların yanı sıra sözlüklerin bilgisayarda hazırlanmasıyla başlayan bilgisayarlı sözlük bilimi (*computational lexicography*) veya daha yaygın adlandırmasıyla elektronik sözlük bilimi (*electronic lexicography*), söz birimleştirme (lemmatisation), anlam ve örnek çıkarımı, kullanım sıklığı belirleme, özetleme vb. çalışmalarla hızla gelişen alanlar olarak kendisini gösteriyor. Bu alanlardan genel olarak metin karşılaştırması diye adlandırabileceğimiz, niteliği bakımından ise eleştirili metin (*édition critique*) hazırlama, müellifi bilinmeyen bir yazma eserin müellifini belirleme gibi edebiyat ve dil araştırmalarına yardımcı olacak uygulamaların yanı sıra intihal saptama, imzasız tehdit mektubu, suç kanıtı belge vb.nin sahibini ortaya çıkarma gibi adli dilbilim (*forensic linguistics*) alanlarında kullanılabilen bilişim uygulamalarıdır.

Türk dili ve edebiyatı alanında eleştirili metin yayımı, müellif hattı bulunmuyorsa veya müellifin yaşadığı dönemde istinsah edilip de bizzat gördüğü nüsha ele geçmemiş ise bir gereklilik olarak görülmektedir. Özgün biçimine en yakın ve tam olduğu öngörülen bir nüshanın esas alınarak diğer nüshalarla karşılaştırması fişleme, dipnotlama ve nüsha farklarını

belirleme biçiminde gerçekleştirilen eleştirili metin yayımına Reşid Rahmeti Arat'ın *Kutadgu Bilig*, Ali Nihat Tarlan'ın *Şeyhî Divanı*, Halûk İpekten'in *Nailî-i Kadim Divanı*, Mertol Tulum'un *Tarih-i Ebü'l-Feth*, Şükrü Halûk Akalın'ın *Saltuk-nâme* adlı çalışmaları örnek verilebilir. Eleştirili metin yayımının yanı sıra eserlerin konu, içerik, dil, anlatım, nazım biçimleri, edebî sanatlar, söz varlığı vb. yönlerden karşılaştırması da insan bilgisi, sezgisi, kavrayışı, zekâsı ve belirlenmiş yöntemlerle yapılmaktadır. Dünyada bu türden çalışmaların bilgisayar destekli olarak yapıldığı alanla ilgili araştırmacılar tarafından gözlemlenmektedir. Türk dili ve edebiyatı alanında bu tür çalışmalara örnek oluşturmak amacıyla metinler arasında bilgisayar destekli benzerlik çalışması yapmaya karar verdiğimizde karşılaştıracağımız eser konusunda hiç tereddüt etmeden Dede Korkut Kitabı'nı seçtik. Bunun birkaç nedeni var:

Her şeyden önce Dede Korkut Kitabı, Türk dili, tarihi ve kültürü açısından önemli bir eserdir. Fuat Köprülü "Bütün Türk edebiyatını terazinin bir gözüne, Dede Korkut Kitabı'nı da öbür gözüne koysanız yine Dede Korkut Kitabı ağır basar," diyerek eserin önemini unutulmayacak sözlerle ortaya koymuştur. Öte yandan Dede Korkut Kitabı, gerek söz varlığı gerekse dilbilgisi özellikleriyle Oğuz grubu lehçeleri arasında metin bütünlüğü bakımından eksiksiz bir görünüm sunmaktadır. Eser, yer adlarıyla, kahramanlarıyla, anlatılarıyla, söz varlığıyla Anadolu Türklüğünü kadim ve çağdaş Türk dünyasına bağlayan köprü niteliğindedir. Bu ve benzeri nedenlerin yanı sıra eserin yeni bir yazmasının bulunması ve bu yazmanın Dede Korkut Kitabı'nın bilinen nüshaları ile olan ilgisi, yakınlığı veya uzaklığı konusunda edebiyat çevrelerinde yeni bir tartışmanın başlamasıdır. Yeni bulunan yazmanın kısa sürede çevri yazılı yayımı ve günümüz Türkçesine aktarılmış biçimi birkaç araştırmacı tarafından yapıldı. Bu yazmanın mevcut nüshalarla benzerlikleri, farklılıkları dile getirildi. İşte bu nedenlerle bilgisayarlı benzerlik çalışmasını Dede Korkut Kitabı üzerine yapmaya karar verdik.

Bu çalışmamızın bir yöntem uygulama makalesi olduğunu belirtmemiz gerekir. Dede Korkut Kitabı'nın anlamsal ve biçimsel yönleriyle ilgili çalışmaların artması yanında yeni yöntemlerle de esere ait özelliklerin ve özellikler arası ilişkilerin çıkarılması çalışmalarından Sarı (2020) tarafından yapılan Dede Korkut Kitabı'nda Söylem Belirleyiciler başlıklı çalışma eserdeki söylem yapılarının modern dilbilimsel yönleriyle ayrıntılı olarak ele alındığı hem metne hem de metnin tarihsel dönem içindeki yerine bakış açımızı genişleten çalışmalardan biridir.

Derlem dilbilimi ve doğal dil işlemede (DDİ) gelişen yöntemlerle birlikte son yıllarda metinlerin karşılaştırılmasına dayanan çalışmaların arttığı gözlenmektedir. Karşılaştırmalar iki belge arasında olabileceği gibi çoklu belgeler arasında da yapılabilmektedir.

Metinlerin birbirleriyle karşılaştırılmasında matematiksel ve olasılıksal hesaplamalara dayalı yöntemler kullanılmaktadır.

Bu çalışmada Dede Korkut hikâyelerinin yeni bulunan Türkmen Sahra yazması ile var olan nüsha arasındaki benzerliğin hesaplamalı olarak bulunması amaçlanmıştır. İki metnin benzerlikleri hesaplanırken matematiksel yöntemler kullanıldığından anlama dayalı bir benzerlik çözümlemesi amaçlanmamıştır. Çalışmanın amacı doğrultusunda biçimsel olarak sözcüklerin yazılım destekli görselleştirilmesi de göz önüne alınmıştır.

Hesaplamalı Metin Benzerliği

Bu çalışmada temel olarak doğal dil işlemede kullanılan metin benzerliği yöntemlerinden kosinüs benzerliği, Jaccard benzerlik katsayısı ve TF-IDF yöntemleri kullanılmıştır. Bu iki benzerlik yöntemi dışında son yıllarda önemi artan sözcüklerin vektörlerle temsil edilerek modellendiği word2vec yöntemiyle iki metin karşılaştırılmıştır. Bu yöntemler için Python programlama diliyle yazılmış kütüphanelerden yararlanılmıştır. Bilişim uzmanlarının sıkça kullandıkları Github açık kaynak kodlu uygulamalar veritabanı taranarak burada yayımlanan metin benzerliği (*document similarity*) betikler (script) araştırılmıştır.[3], [4] Bu betiklerden kosinüs ve jaccard benzerlik hesaplama yöntemlerini kullanan iki betik iki metin girdisini sözcükler ve karakterler düzeyinde karşılaştırarak metinlerin benzerliklerini oran cinsinden vermektedirler.

Metinlerin karşılaştırılması ve metinler arasındaki farklılıklara dayalı çalışmaların yapıldığı adli dilbilimde (forensic linguistics), yazarlık belirleme (*authorship attribution*) çalışmalarında da metinlerin yapısal ya da biçimsel özellikleri temelde sözcük davranışları hesaplanarak bir bakıma yazarın adeta parmak izi tespit edilmeye çalışılmaktadır. Üslup çözümlemesi, yazar profili çalışmaları yazarı bilinmeyen metinlerdeki benzerliklerin ortaya çıkarılması konularını da ele alan adli dilbilimin metinle ilgili çeşitli düzeylerde çalışma teknikleri bulunmaktadır (Karaman, 2019, s. 217)**.**

Kaya ve Özel (2014), Türkçe belgelerde benzerliklerin bulunmasıyla ilgili yazılımları karşılaştırdıkları çalışmada, internet ve teknolojiyle birlikte artan veri ve bilgi miktarının yanında intihal olaylarının tespit edilmesine yönelik teknolojilerin de arttığını belirtmişlerdir.

[3] https://github.com/machine-learning-projects/document-similarity
[4] https://github.com/TarunSunkaraneni/Document-Similarity

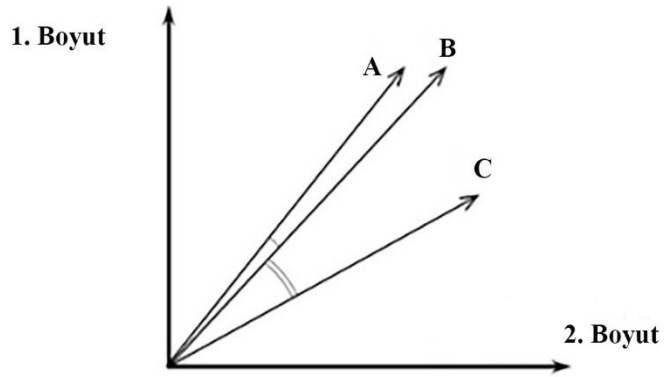
Metin benzerliklerinin hesaplamadaki amaçlardan biri olan intihalin yanında, metinlerin kümelenmesi amacıyla metinler arası benzerliklerden de yararlanıldığı belirtilmiştir.

Metin koleksiyonlarının ve sayılarının artması metin sınıflandırma yöntemlerinin kullanılması gerektirmektedir. İnternetin sağladığı devasa metin yığınlarının otomatik olarak sınıflandırılmasıyla bilgiye erişim daha hızlı biçimde sağlanmaktadır. Metin sınıflama temel olarak verili bir sınıfa eldeki metinlerin atanması olarak tanımlanmaktadır. Bunun için kullanılan yöntemler arasında Naive Bayes, Karar ağaçları ve makine öğrenmesine dayalı yöntemler bulunmaktadır (Tantuğ, 2016, s.9). Metinlerin sınıflandırılması yanında sınıflamadan önce daha kaba bir uygulama olarak kümeleme de (*document clustering*) yapılabilmektedir. Kümeleme için önceden verili bir etiket ya da hedef sınıf adı bulunmamaktadır. Bilgisayar bilimlerinde gözetimsiz öğrenme (*unsupervised learning*) adı verilen bir yöntemler bütünü kullanılarak veride keşfedici bir çözümlemeyle örüntüler bulunmaya çalışılmaktadır.

Belgelerin otomatik kümelenmesi ve sınıflandırılması için sayısal düzlemde ya da uzayda temsil edilmeleri (*document representation*) gerekmektedir (Huang, 2008, s.50). Belgelerin temsil edilmesinde sözcük torbası (*bag of words*) olarak bilinen yöntemin bilgiye erişim ve metin madenciliği alanlarında sıklıkla kullanıldığını belirtmektedir. Belge temsili için kullanılan bu yöntemde metindeki sözcükler bağlamlarından bağımsız olarak matematiksel bir vektörle sıklıklarına göre temsil edilmektedir (Tantuğ, 2016, s.6). Belgelerin vektör düzeyinde temsil edilmeleri başta karşılaştırma olmak üzere metinler üzerinde sözcüklerin dağılım modellerinin yapılmasını da sağlamaktadır. Sidorov'a (2019) göre belgelerin temsil edilmesinde vektör uzayı (*vector space*) modelinin kullanılmasının basitçe amacı karşılaştırmaların daha formel bir düzeyde yapılabilmesinin sağlanmasıdır. Belgeler n boyutlu bir vektör uzayında özellikleri (*features*) bakımından yer alırlar. Bu özellikler belgelerin sayfa numaraları olabileceği gibi sözcüklerin sıklık değerleri de olabilir. Vektörler belgelerin özellikleri kadar boyutlandırılabilir.

Bu çalışmada iki metin arasındaki benzerlik oranını hesaplamak için kullanılan hesaplama yöntemleri; kosinüs benzerliği, Jaccard katsayısı ve TF-IDF yöntemleridir. Bu üç yöntemden elde edilen puanlar iki metin arasındaki benzerlik oranını göstermektedir. Kosinüs benzerliğinde puanlar 0 ve 1 arasında yer alır. Sözcüklerinin vektör düzeyinde temsil edildiği iki belgede terimler (sözcükler) boyutlandırılarak her boyutta terimlerin belgedeki ağırlıkları temsil edilir. Vektörler arasındaki korelasyonların 0 ve 1 arasında 1'e yakın olması benzerliğin çok, 1 olması iki belgenin aynı olduğunu, 0 olması da iki belgenin farklı belgeler

olduğunu göstermektedir. (Huang, 2008, s. 51). Kosinüs benzerliğinde ayrıca vektörlerdeki açı da önemlidir. Vektörlerdeki açının keskinliği benzerliğin ölçüsüdür. Açı keskinleştikçe benzerlik artmaktadır (Sidorov, 2019, s. 8). Aşağıdaki şekilde iki boyutlu bir uzayda A, B ve C vektörlerinin açıları gösterilmiştir.



Şekil 1. İki boyutlu uzayda benzerlik açıları (Sidorov, 2019, s. 8)

Jaccard katsayısında iki metinde bakılan özellikler açısından ortak olan özelliklerin iki metindeki tüm özelliklere bölünmesiyle elde edilen bir indekse dayanarak benzerlik sonucuna ulaşılır (Karaman, 2019, s. 217). 0 ve 1 arasında değer alan bu sonuçlar kosinüs benzerliğinde elde edilen sonuçlar gibi yorumlanır. Bu çalışmada kullanılan Jaccard katsayısı hesaplamasında her iki Dede Korkut metnindeki ikili ve üçlü karakter dizileri ile ikili sözcük birliktelikleri özellik olarak çıkarılarak karşılaştırılmıştır. Jaccard benzerlik katsayısını hesaplayan betikte[5], ikili ve üçlü karakter dizileri ile ikili sözcük dizileri her iki metin açısından temel alınarak karşılaştırılmıştır.

TF-IDF terimi (*term frequency-inverse document frequency*) metinlerde geçen terimlerin sıklıkları ve terim sıklıklarının bir derlemdeki tüm metinlere dağılımının hesaplanmasıyla elde edilen değeri ifade etmektedir. Metinlerdeki terim sıklıkları her bir metin göz önüne alınarak her metindeki terim sıklıkları hesaplanır (sözcüğün metindeki sıklığının o metindeki tüm sözcüklere bölünmesiyle elde edilir) ve bu terim sıklığı (*term frequency*) olarak adlandırılır. Bir sözcüğün tüm metin derlemindeki sıklık dağılımı ise o sözcüğün ters belge sıklığını (*inverse document frequency*) oluşturur. Bir sözcük tüm belgelerde dolayısıyla metinlerde görülüyorsa ve yüksek sıklıktaysa bu sözcük genel bir sözcük olarak kabul edilir ve önemli bir birim olarak kabul edilmez. Tersine, sözcük tüm belgelerde yalnızca bir kez geçmişse bu sözcüğün belgeleri ayırt etme gücü vardır denilebilir

[5] https://github.com/TarunSunkaraneni/Document-Similarity/blob/master/Document%20Similarity.ipynb

ve anahtar bir birim olarak görülür. Dolayısıyla sözcük sıklıklarının belgelerdeki dağılımıyla her belge başına düşen sıklıklar tf-idf belge benzerliklerinin hesaplanmasında da kullanılır. Terimler ve terimlerin geçtiği belgeler bir matrise dönüştürülerek belgeler arasında karşılaştırmalar sözcüğün iki belgede bulunup bulunmamasına göre ağırlıklandırılır. Tf-idf değeri, terim sıklığının ters belge sıklığından elde edilen puanın çarpımından elde edilir. Sözcük ve belge tablosu biçiminde yer alan matrislerde tf-idf skorları tutularak bu skorlar üzerinden karşılaştırma yapmak mümkün hâle gelir (Sidorov, 2019, s. 11-13).

Sözcük gömme (*word embedding*), sözcük temsili yöntemi olarak adlandırılan bir metindeki ya da derlemdeki sözcüklerin matematiksel olarak bir uzay vektöründe gerçek sayılarla temsil edilmesidir. Bu şekilde metinler yapay sinir ağlarına girdi olarak aktarılabilir bir yapıya dönüştürülmekte, n-gram temelli dil modellerine göre sözcüklerin dağılımları daha genellenebilir duruma gelmektedir. Sözcükler arasındaki anlamsal ilişkiler ve yakınlıklar böylelikle otomatik olarak oluşmaktadır (Mikolov vd., 2013, s. 2). Word2vec yöntemi sözcük gömme yöntemlerinden biri olarak GloVe ve FastText gibi en çok kullanılan yöntemler arasında yer almaktadır.

Sözcüklerin vektör uzayında temsili aynı zamanda doğal dil işlemede uygulanan makine öğrenmesi yöntemlerinden biridir. Sözcüklerin birbirleriyle olan bağlamsal ilişkileri bir vektör uzayında modellendikten sonra özellikle LSTM (*long short term memory*) adı verilen derin öğrenme (*deep learning*) uygulamalarına girdi olarak verilmektedir. Böylelikle sözcüklere dayalı otomatik sınıflandırma, etiketleme uygulamalarında bağlam duyarlılık artırılarak başarım da buna koşut artmaktadır ((Eisenstein, 2019, s. 325-328).

Sözcük vektörleri mühendislik uygulamaları yanında dilbilimsel amaçlar için de kullanılabilir. Sözcük vektör uzayları büyük hacimli metin verisinde anlamca ilgili sözcükleri yüksek doğruluk/başarım oranında bir araya getirebilmektedir. Böylelikle sözcüklerin dolayısıyla kavramların anlamsal dağılım örüntüleri görselleştirilerek metnin kavramsal haritası ortaya konabilmektedir. Bu şekilde, metinle ilgili daha önce görülemeyen ya da farkına varılmayan gizil ilişkilerin ve özelliklerin çıkarılması için de bir yöntem olarak büyük boyutlu derlemlerde de kullanılması söz konusu olmaktadır.

Yöntem

Çalışmada yukarıda sözü edilen hesaplama yöntemleri uygulanırken Prof. Dr. Muharrem Ergin'in Dede Korkut hikâyelerinin Vatikan nüshası ile eleştirili yayımını gerçekleştirdiği Dresden nüshasının çeviriyazılı metni esas alınmış öncelikle bu metin sayısallaştırılarak oluşan optik karakter okuma hataları elle düzeltilmiştir. Prof. Dr. Metin

Ekici tarafından çeviri yazısı yapılan Türkistan/Türkmen Sahra yazması sayısallaştırılmış ve her iki Dede Korkut metni .txt formatında kaydedilmiştir. İki metinde de karakter hataları düzeltildikten sonra metinlerdeki çeviriyazıya özgü karakterler standartlaştırılmıştır. Bu yapılırken uzunluk işaretleri iki ünlü (aa, ii vb.) ile gösterilmiş, damaksı n ise ng olarak değiştirilmiştir. Burada mümkün olduğunca karşılaştırma esnasında karakter farklılıklarından doğacak yanlılıkların (hataların sebep olacağı eşitsizliklerin) giderilmesi amaçlanmıştır. Metinlerde büyük küçük harf duyarlılıkları tüm harfler küçük olacak biçimde değiştirilmiştir. Böylelikle büyük-küçük harflerin hesaplamalarda farklılık oluşturmaması sağlanmıştır.

Çalışmada Dede Korkut metinleri Python programlama diline dayalı bir uygulama olan word2vec[6] uygulamasıyla eğitilmiştir. Bu eğitim denetimsiz ya da öğretmensiz bir eğitim biçimidir. Bu biçime göre verideki sözcükler arasındaki ilişkiler bulunmaya çalışılarak önceden belirlenmiş herhangi bir hedef etiket ya da sınıf adı verilmez. Uygulamada parametre olarak 20000 iterasyon (yineleme), 100 boyut, 16 pencere büyüklüğü ve skip-gram algoritması kullanılmıştır. word2vec ile elde edilen .vec uzantılı dosyalar görselleştirmede kullanılacak biçim için Gensim kütüphanesiyle[7] TSV formatlı dosyalara dönüştürülmüştür.

```
3180 100
kara -0.532441 -0.412085 -0.128039 0.3260
bir -0.719668 0.454828 -0.182439 0.474694
ol -0.051520 1.061253 -0.071219 1.013630
kazan -0.074769 0.542641 -1.069681 -1.396
er -0.068077 0.092032 -0.556815 -0.123887
ala 0.208775 -0.170098 0.174384 0.024883
güni -0.702941 0.398150 0.202345 -0.27531
neye -0.070142 0.101361 0.429036 -0.03103
yarar -0.022745 0.095063 0.286411 0.29993
kimi -0.273701 -0.202001 -1.048918 -0.247
yer -0.690650 0.046638 -0.209920 0.276098
eli 0.976638 -0.038664 0.240539 0.620454
gerek -0.396782 1.164801 0.081147 0.57226
```

Şekil 2. Dede Korkut Kitabı Türkistan/Türkmen Sahra yazmasında vektörlere dönüştürülen sözcükleri temsil eden değerler.

Yukarıdaki şekilde 3180 sayısı metinde geçen tüm sözcüklerin sayısını, 100 ise boyutu temsil etmektedir. Aynı şekilde Dresden nüshasında da 7730 sözcük ve 100 boyutlu bir vektör

[6] https://github.com/danielfrg/word2vec
[7] https://medium.com/@aakashchotrani/visualizing-your-own-word-embeddings-using-tensorflow-688b3a7750ee

uzayında temsil edilmiştir. Vektörün görselleştirilmesi için Tensorflow kütüphanesinin görselleştirme web sayfası[8] kullanılmıştır.

Bulgular

Kosinüs Benzerliği ve TF-IDF Skoru

İki metin için kosinüs benzerliğini hesaplayan betik[9] çalıştırıldığında benzerlik oranı 0.39 (%39) olarak bulunmuştur. Bu oran iki metin arasındaki benzerliğin düşük olduğunu göstermektedir.

Aynı betiğin TF-IDF hesaplaması iki metnin birbirine göre oranını 0.28 (%28) olarak hesaplamıştır. Bu oran da iki metin arasındaki benzerliğin düşük olduğunu göstermektedir.

Jaccard Benzerlik Katsayısı

Bu katsayıyı hesaplayan betiğe[10] göre D1 olarak temsil edilen Dresden nüshasının D2 olarak temsil edilen Türkistan/Türkmen Sahra yazmasıyla iki karakter dizisi (gram) benzerliği %65.25, üç karakter dizisine benzerliği %44.63 olarak hesaplanmıştır. İkili ve üçlü sözcük dizisi hesaplamamasına göre sırayla benzerlikler %3.11 ve %1.00'dir. Bu sonuçlar karakter dizilerinde görece yüksek benzerliklerin olduğunu ancak sözcük dizileri bakımından aynı ortaklıklardan söz edilemeyeceğini göstermektedir.

Sözcük Temsili Çözümlemesi

Çalışmada her iki metinde geçen ortak sözcüklerin sıklık listesi çıkarılmıştır. Aşağıdaki tabloda çıkarılan liste yer almakta, tablodan da görüleceği üzere *bir, kara* ve *kazan* sözcükleri her iki metinde de yüksek sıklıkta geçmektedir.

Tablo 1 *Dresden Nüshası ve Sahra Yazması İlk 25 Ortak Sözcük Listesi*

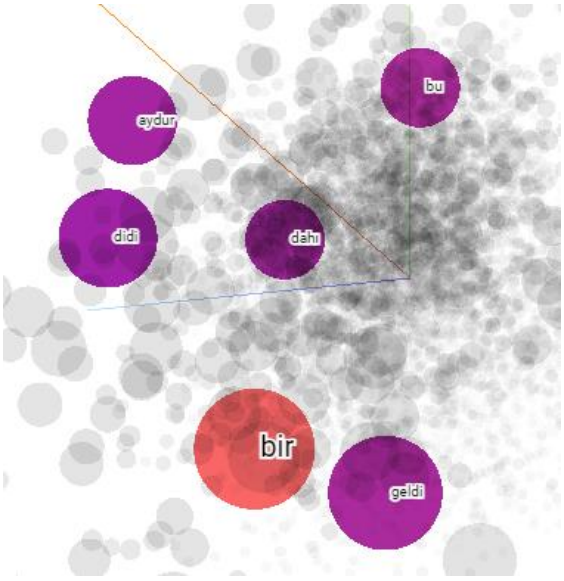
Sözcük	Dresden Nüshası Sıklığı	Sahra Yazması Sıklığı
bir	392	62
kara	349	78
kazan	235	37
ne	233	12
geldi	205	22

[8] https://projector.tensorflow.org/
[9] https://github.com/machine-learning-projects/document-similarity/blob/master/src/cos_dist.py
[10] https://github.com/TarunSunkaraneni/Document-Similarity/blob/master/Document%20Similarity.ipynb

bu	183	13
oldı	162	5
ol	155	46
dahı	149	1
manga	129	5
oğuz	122	5
kaafir	120	5
menüm	117	6
kan	108	17
delü	105	3
ala	104	32
sanga	99	1
han	96	1
kız	94	3
kim	92	6
at	91	15
men	90	9
aldı	87	1
sen	83	2
senüng	76	3

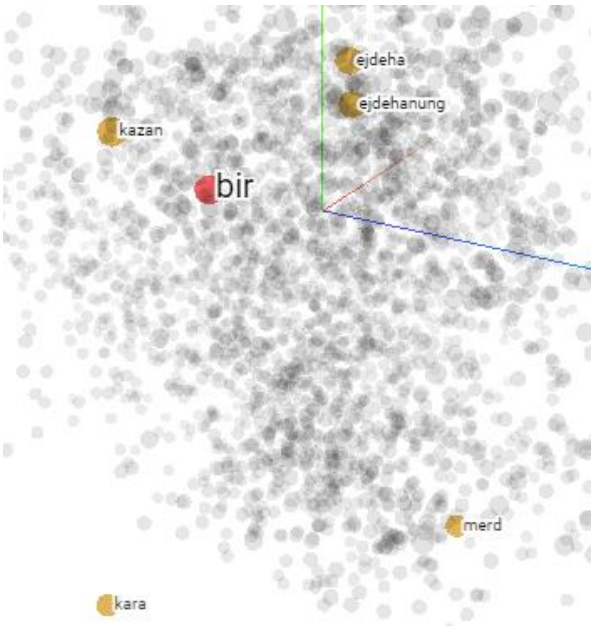
Sözcük temsili çözümlemesi metinlerin birbirlerine benzerlik oranlarını veren yöntemlerden farkı sözcüklerdeki komşuluk dolayısıyla kavramsal açıdan karşılaştırma yapmaya ve bu yönden farklılıkları ve benzerlikleri çıkarmaya yardımcı olmasıdır. Çalışmada bu yöntem, önceki yöntemlerden farklı bir bakış açısı getirmesi ve yeni bir teknoloji olması nedeniyle uygulanmıştır. Word2vec çözümlemesi için hazırlanan vektör dosyalarının görselleştirilmesi https://projector.tensorflow.org/ adresinde gerçekleştirilmiştir. İki metnin

karşılaştırılması için iki metinde de sıklığı yüksek ortak sözcüklerden "bir", "kara", "kazan" sözcüklerinin en yakın komşuluk ilişkilerinin üç boyutlu görünümü elde edilmiştir. Bu üç sözcüğün sıklıkları bakımından iki metni temsil eden birimler olduğu düşünülmüştür. Sırasıyla Dresden nüshasında *bir* 392, *kara* 349, *kazan* 235 sıklık değerine sahiptir. Türkistan/Türkmen Sahra yazmasındaysa *bir* 62, *kara* 78 ve *kazan* 37 sıklık değerleri gözlenmiştir. Aşağıdaki şekillerde komşuluk ilişkileri en yakın beş birimle temsil edilen bu görünümler verilmiştir. Komşuluk ilişkilerinin bulunmasında vektörler arasındaki uzaklıklar kosinüs uzaklık hesaplamasıyla bulunmuştur.



Şekil 3. Dresden nüshasında "bir" sözcüğünün en yakın beş sözcükle ilişkisi.

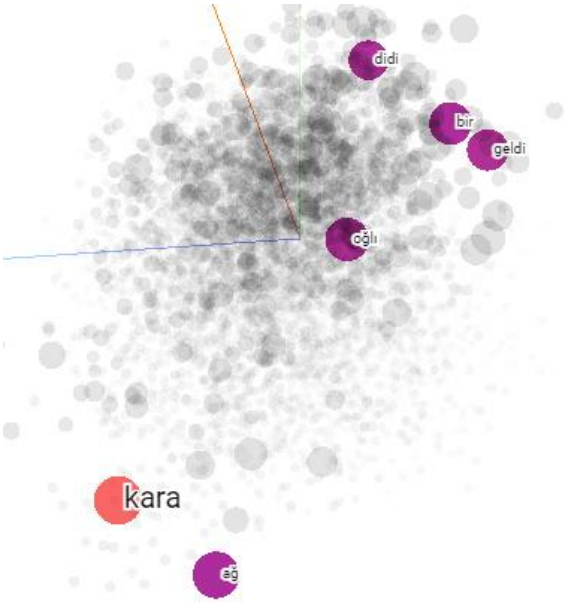
Şekle göre "bir" sözcüğünün *geldi*, *didi*, *dahı*, *aydur*, *bu* sözcükleriyle yakın ilişkisi görülmektedir. Dresden nüshasında sözcük sayısı dolayısıyla vektör uzayında temsil edilen nokta sayısı fazla olduğundan simgeler sıklıklara göre daha büyük temsil edilmektedir.



Şekil 4. Türkistan/Türkmen Sahra yazmasında "bir" sözcüğünün en yakın beş sözcükle ilişkisi.

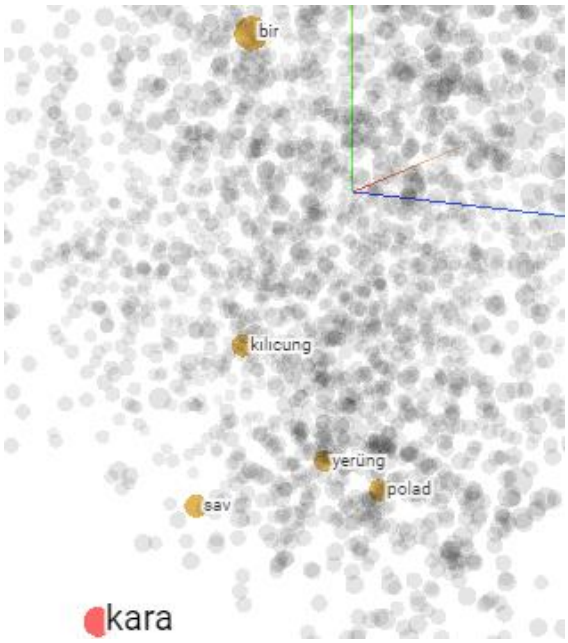
Şekle göre "bir" sözcüğünün *kazan, ejdeha, ejdehanun, merd* ve *kara* sözcükleriyle yakın ilişki kurduğu görülmektedir.

" bir" sözcüğü Dresden nüshası ile ortak komşuluk ilişkisine sahip değildir.



Şekil 5. Dresden nüshasında "kara" sözcüğünün en yakın beş sözcükle ilişkisi.

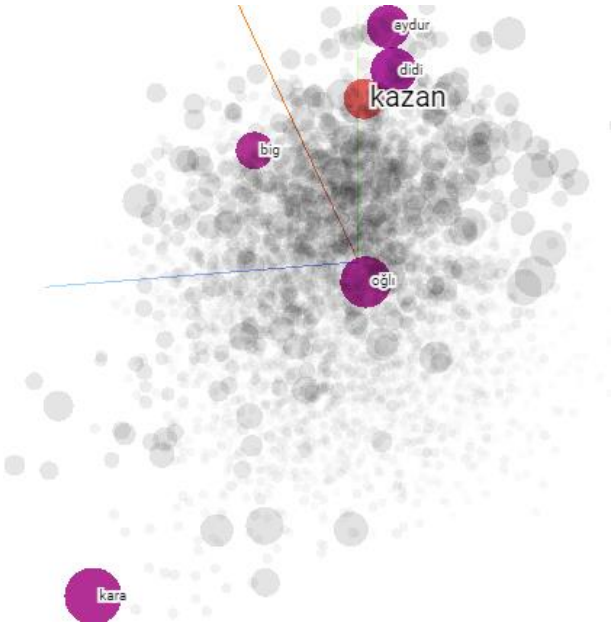
Şekle göre "kara" sözcüğünün *ağ, oğlı, geldi, bir* ve *didi* sözcükleriyle yakın ilişki kurduğu görülmektedir.



Şekil 6. Türkistan/Türkmen Sahra yazmasında "kara" sözcüğünün en yakın beş sözcükle ilişkisi.

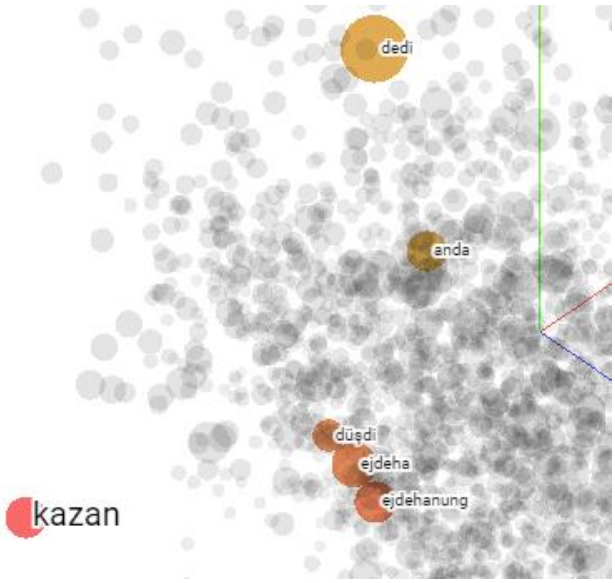
Şekle göre "kara" sözcüğünün *sav*, *polad*, *yerüng*, *kılıcung* ve *bir* sözcükleriyle yakın ilişki kurduğu görülmektedir.

Burada "kara" sözcüğünün Dresden nüshasındaki *bir* sözcüğü ile ortak komşuluk yaptığı görülmektedir.



Şekil 7. Dresden nüshasında "kazan" sözcüğünün en yakın beş sözcükle ilişkisi.

Şekle göre "kazan" sözcüğünün *didi*, *aydur*, *big*, *oğlı* ve *kara* sözcükleriyle yakın ilişki kurduğu görülmektedir.



Şekil 8. Türkistan/Türkmen Sahra yazmasında "kazan" sözcüğünün en yakın beş sözcükle ilişkisi.

Şekle göre "kazan" sözcüğünün *dedi*, *anda*, *düşdi*, *ejdeha* ve *ejdehanung* sözcükleriyle yakın ilişki kurduğu görülmektedir.

Burada "kazan" sözcüğünün Dresden nüshasındaki *didi* (*dedi*) sözcüğü ile ortak komşuluk yaptığı görülmektedir.

Sözcük temsili ya da word2vec modeline göre incelenen ve ortak sıklıkları en yüksek üç sözcük (*bir, kara, kazan*) her iki metinde yalnızca iki kez ortak komşuluk ilişkisine sahip bulunmuştur. Bunlar "kara" sözcüğünün "bir" ile ve "kazan" sözcüğünün "didi (dedi)" ortaklığı biçimindedir. En yakın beş sözcük komşuluğu ilişkisinde birer kez yer alan bu ortaklıklar düşük görülmekte bu da yukarıda belirtilen sözcükler arası kosinüs benzerliği yüzdesinin düşük bulunmasıyla uyumlu olarak yorumlanmıştır.

Sonuç

Bu çalışmada Dede Korkut Hikâyelerinin yeni bulunan yazması var olan Dresden nüshası ile metin benzerliği açısından hesaplamalı yöntemlerle karşılaştırılmıştır. Elde edilen ilk bulgulara göre iki metin arasında benzerlik oranları kosinüs için %39, TF-IDF için %28 ve Jaccard içinse %65, %44, %3 ve %1 biçiminde hesaplanmıştır. Jaccard katsayısındaki %65 ve %44 ikili ve üçlü karakter dizisi benzerliğinin karakter dizisi sayısı arttıkça düşük bulunacağı söylenebilir. Özellikle sözcük eşdizimselliğini gösteren ikili ve üçlü sözcük dizilerinin (n-gram) ortaklık yüzdesi (n=2 %3, n=3 %1) son derece düşüktür. Sözcük temsili çözümlemesine göre de ortaklık gösteren üç sözcüğün metinlerdeki vektör uzayındaki ortak

komşuluk ilişkileri de sadece iki sözcükle sınırlı bulunmuştur. Bu bulgulardan hareketle ilk bulgular bu iki metin arasında biçimsel benzerliğin düşük düzeyde kaldığını göstermektedir.

Dede Korkut metinlerinin yeni geliştirilen yazılımsal yöntemlerle daha ayrıntılı incelenmesi, bu metinlere ait yeni örüntülerin bulunması elbette mümkündür. Bu çalışmada bilinen benzerlik yöntemleri uygulanmış ve iki metin arasında istatistiksel olarak hesaplanan benzerliklerin düşük olduğu sonucuna varılmıştır. Bu sonuç ileri çalışmalar için bir ön değerlendirme olarak yorumlanmalı bununla birlikte sonuçlar Sahra yazmasının Dresden nüshasından farklı bir söz varlığı özelliği gösterdiğini ortaya koymaktadır.

İki metin arasında daha ileri anlam, konu ve kültürel bileşenlere dayalı karşılaştırmalı çalışmaların farklı boyutlar ve bakış açılarıyla bulunan yeni yazmanın konumunun belirginleşmesine katkıda bulunacağına inanıyoruz. Bu çalışmanın tarihsel metinlerin karşılaştırmalı incelemelerine hesaplamalı yöntem kullanımı açısından da yarar sağlayacağı düşünülmektedir. Hesaplamalı metin karşılaştırma yöntemlerinin hem yöntem bilgisi açısından tartışılması, Türkçenin tarihsel metinlerinin yeni bir bakış açısıyla değerlendirilerek metinlerin özelliklerine ait bilgimizin genişlemesine katkı sağlayacaktır. Türkçenin tarihsel sözlüğü ve tarihsel derlemlerinin çeşitlendirilmesinde bilgisayar destekli karşılaştırmalı analizi veriden yeni keşiflerin yapılmasında önemli roller üstlenecektir.

Kaynakça

Eisenstein, J. (2019). *Introduction to natural language processing*. The Mit Press.

Ekici, M. (2019). *Dede Korkut Kitabı Türkistan/Türkmen Sahra nüshası, soylamalar ve 13. boy*. İstanbul: Ötüken Neşriyat A.Ş.

Ergin, M. (1994). *Dede Korkut kitabı I* (3. baskı). Ankara: Türk Dil Kurumu Yayınları.

https://github.com/danielfrg/word2vec, 09.08.2020

https://github.com/machine-learning-projects/document-similarity, 11.12.2020

https://github.com/machine-learning-projects/document-similarity/blob/master/src/cos_dist.py, 07.11.2020

https://github.com/TarunSunkaraneni/Document-Similarity, 10.09.2020

https://github.com/TarunSunkaraneni/Document-similarity/blob/master/Document%20Similarity.ipynb, 10.09.2020

https://medium.com/@aakashchotrani/visualizing-your-own-word-embeddings-using-tensorflow-688b3a7750ee, 05.10.2020

https://projector.tensorflow.org/, 06.10.2020

Huang, A. (2008). *Similarity measures for text document clustering*, in New Zealand Computer Science Research Student Conference - Proceedings of NZCSRSC, New Zealand.

Karaman, B. İ. (2019). Dilbilimsel otopsi. *The Bulletin of Legal Medicine*, *24*(3), 214-225.

Kaya, M., ve Özel, S. A. (2014). Türkçe dokümanlardaki benzerliklerin tespiti için mevcut yazılımların karşılaştırılması ve Türkçe karakter kullanımı ile kök almanın etkisinin incelenmesi. *Çukurova Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, *29*(2), 115-130.

Sarı, İ. (2020). Dede Korkut Kitabı'nda söylem belirleyiciler. *Bilig*, (93), 29-52.

Sidorov, G. (2019). *Syntactic n-grams in computational linguistics*. Springer International Publishing.

Tantuğ, A. C. (2016). Metin sınıflandırma. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, *5*(2), 1-12.