littera turca
Journal of Turkish Language and Literature
e-ISSN: 2149-892X
http://dergipark.ulakbim.gov.tr/littera

EDEBİYAT ARAŞTIRMACILARI İÇİN BİLGİSAYAR DİLLERİ VE METİN MADENCİLİĞİ

Computer Languages and Text Mıning for Literary Researchers

Ayşe TARHAN[1]

[1] *Doç. Dr., Uluslararası Kıbrıs Üniversitesi Eğitim Fakültesi Türkçe Öğretmenliği Bölümü, aysedtarhan@gmail.com, orcid.org/0000-0003-0493-7928.*

Araştırma Makalesi/Research Article

Makale Bilgisi
Geliş/Received: 02.05.2024
Kabul/Accepted: 26.07.2024

DOI: 10.20322/littera.1477535

Anahtar Kelimeler
dijital beşerî bilimler, edebiyat, metin madenciliği, bilgisayar dili r ve python.

ÖZ

Bilim, antik Yunanda felsefe demekti. On dokuzuncu yüzyıla gelindiğinde felsefenin çok kapsayıcı olduğu, içinde bulunan edebiyat, tarih, psikoloji gibi sosyal ve beşerî bilimlerin ayrılması ve özerk bir bilim dalları olarak algılanması gerektiği anlaşıldı. Bilgisayar da ilk olarak matematik biliminin bir aracı olarak görüldü. Zamanla teknolojinin var olduğu her alanını kapsayarak büyüdü ve tıpkı felsefenin farklı bilim dallarına ayrılması gibi bilgisayar da farklı bilim dallarına ayrıldı. Şimdi bilgisayarın içinde bulunmadığı hemen hemen hiçbir bilim dalı bulunmamaktadır. Bunlardan biri de Dijital Beşerî Bilimlerdir. Dijital Beşerî Bilimlerin ilgi alanlarından biri edebiyattır. Edebiyatın malzemesi de kelimelerdir. Bir araştırmacının aylarca yapacağı araştırmayı bilgisayar programlama dilleriyle oluşturulan yazılımlar sayesinde kısa sürede yapabilmekte ve verilerini nesnel bir biçimde ortaya koyabilmektedir. Bu çalışmada edebiyat araştırmacıları için metin madenciliğinin temel basamakları anlatılacaktır. Metin madenciliği için oluşturulan araçlardan Stylometry ve The Metrical Tool Hylas edebiyat incelemeleri için model olması nedeniyle araçlar hakkında bilgi verilecektir. Bu araçlar bilgisayar programlama dillerinden Python ve R ile oluşturulmuş ve bilim dünyasına kazandırılmıştır. Bu bağlamda çalışmada bahsedilecek ilk bilgisayar programlama dili R'dır. Diğeri de Python'dur. Çalışmanın son aşamasında metin madenciliğinde dijital araçlarla oluşturulan görselleştirme modellerinden bahsedilecek, görselleştirmeler Latîfî Tezkiresi, Bâkî ve Fuzûlî divanlarından oluşan derlemler aracılığıyla örneklendirilecektir.

Keywords
digital humanities, literature, text mining, r, python.

ABSTRACT

Whereas in Ancient Greece, the study of philosophy included all sciences, by the nineteenth century, many social and human sciences, such as literature, history and psychology, had become to be perceived, not as a part of philosophy, but as separate subjects, distinct from philosophy. Similarly, when computers were first developed, they were seen as a tool of mathematical science. However, over time, computers have become to be considered an integral tool in every field of study and this has brought about the development of separate computer programs for use in distinct subject areas. One such area of study in the field of Digital Humanities, is literature. The material of literature is words and thanks to software created with computer programming languages, research that would otherwise take months can be completed in a very short time and the data can be presented objectively. In this study, the basic steps of text mining for literary researchers will be explained. Information will be given about Stylometry and The Metrical Tool Hylas, as these tools are models for Ottoman literary studies. Since these tools were created with the computer programming languages R and Python, in this study, these two computer programming languages will first be explained. In the second stage of the study, visualisation models created with digital tools in text mining will be examined, and visualisations will be

exemplified through corpora consisting of Latîfî Tezkires items, Bâkî, and Fuzûlî divans from Ottoman literature.

Atıf/Citation: Tarhan, A. (2024), "Edebiyat Araştırmacıları İçin Bilgisayar Dilleri ve Metin Madenciliği", *Littera Turca, Littera Turca Journal of Turkish Language and Literature*, 10/3, 444-457.

Sorumlu yazar/Corresponding author: Ayşe TARHAN, aysedtarhan@gmail.com

GİRİŞ

Teknolojinin kullanım alanı olan her nokta aslında bilgisayarın kullanım alanı olduğunu göstermektedir. Dolayısıyla, bilgisayarın aşağıda verilen 5 ayrı sınıflandırmadaki bilim dallarıyla yakından ilgi ve etkiletişimi vardır. Bunlar şöyledir:

1. Fizik bilimleri (bilgisayar bilimi, matematik ve fizik);
2. Beşerî bilimler (edebiyat, yabancı diller, dilbilim, felsefe, tarih ve güzel sanatlar);
3. Sosyal bilimler (psikoloji, bilişsel bilim, sosyoloji, antropoloji ve ekonomi);
4. Meslekî programlar (kütüphane ve bilgi bilimi, eğitim, gazetecilik, medya);
5. Disiplinlerarası alanlar (teknoloji ve toplum, bilim tarihi, kadın çalışmaları, üçüncü dünya çalışmaları, çevre çalışmaları vb.) (Ehrlich, 1991, s. 316).

Metin madenciliği, yapılandırılmamış metinden bilgi veya öngörü elde etme süreci olarak değerlendirilmektedir (Atan, 2020, s. 221). Metin madenciliği, doğal dil metinlerinde yer alan işlenmemiş verilerin çeşitli yöntem, araç ve tekniklerin kullanılmasıyla analiz edilmesine dayanmaktadır (Çelik, 2020, s. 1343). Metin madenciliğinin araştırma alanlarından biri, metinlerdeki konular arasındaki ilişkileri belirlemektir. Bir diğer konusu, metindeki gizli kalıp ve kalıpların ortaya çıkarılmasıdır. Metin madenciliğinin araştırma malzemeleri arasında belgeler, veri tabanlarında bulunan kayıtlar, kişilerin yazışmaları, sosyal medya girdileri, dil, edebiyat ve tarih gibi bilimlere ait metinler bulunmaktadır (Berry, 2012, s. 1-20).

Edebiyat ile bilgisayar arasında bilgisayarın, edebiyatın anlaşılmasında ve araştırılmasında yardımcı olması açısından yakın bir ilişkisi bulunmaktadır. Edebiyat araştırmalarına bilgisayarın pek çok programlama dilleri yardım etmesine rağmen en çok yararlanılan programlama dilleri belki de R ve Python'dur (Tarhan, 2024, s. 17). Bu diller bilgisayar teknolojilerine dayalı bölümlere dair altyapıya sahip olmayanlar tarafından öğrenilmesi de kolay olan dillerdendir. Bugün Avrupa[2] ve Amerika'da pek çok üniversite, Türkiye'de ise Koç Üniversitesi gibi birkaç yüksek öğretim kurumlarında sosyal bilimler kökenli araştırmacılar için R ve Python programlama dilleri öğretilmektedir.

[2] Avrupa'da ve Amerika'da hemen hemen her üniversitede Dijital Beşeri Bilimler bölümü bulunmaktadır. Bu bölümler sosyal bilimler kökenli öğrencileri yetiştirdikleri gibi, bölüm dışından araştırmacılara da yaz ve kış okullarıyla bilgisayar dillerini öğretmekte, araştırmacıların kendilerini geliştirmelerine imkanlar sağlamaktadır. Avrupa için bir örnek Hollanda'da bulunan Leiden Üniversitesinin her yıl sunduğu yaz okullarıdır: *Summer School Literary Studies & Digital Humanities.* Bu yaz okullarına iki defa katılarak ben de kendimi bu iki dil üzerinde geliştirme fırsatı buldum. Amerika'da bu alanda pek çok yaz ve kış okulu bulunmaktadır. Bunlardan en kapsamlılarından biri *Compute Canada Federation*'un düzenlediği *Humanities and Social Sciences Series*'dir. Ben bunlardan 2021'de "*An introductory digital research series for humanities and social science researchers*" adlı kış okuluna katıldım. Bu yaz ve kış okullarında R ve Python ile ilgili pek çok eğitimler yapılmakta ve farklı araçlar ve projeler

Python (http://www.python.org/) bir scripting dilidir, 1980'de oluşturulmuş ve 1991'de ilk sürümünü vermiştir. Python adı, İngiliz komedi topluluğu Monty Python'un The National Research Institute for Mathematics and Computer Science'ta (CWI) çalışan Guido van Rossum'u etkilemesi üzerine verilmiştir. Python anlaşılır ve etkili olduğu için yazılımcılar tarafından en çok tercih edilen dildir. Python yaptığı her şeyde İngilizce ifade kalıplarını ve cümlelerini kullanmaktadır. Python, dinamik object-oriented programlama dilidir ve kullanışlı bir yapısı vardır. Zengin bir standart kitaplık sunar ve çalışma yapısı nedeniyle popüler kitaplıkların çoğuna bağlantılar geliştirmiştir. Python'un en büyük avantajı, çok anlamlı ve okunabilir sözdiziminin olmasıdır. Dolayısıyla, çok kısa bir kodla ve daha az hatayla çok şey başarabilmektedir. (Grabar, 2009, s. 80-145)

Dilde üslup analizi yapmak için R gibi gelişmiş istatistik programlama dili kullanılmaktadır, R programlama dili bu işi çok daha kolay hâle getirmektedir ve çok büyük metinlerin analizini yapabilmektedir. R, istatistik ve veri analizi için üretilmiş açık erişimli, hesaplama ve grafikler için kullanılan ücretsiz bir programlama dilidir. UNIX platformlarında, Windows ve MacOS'ta derlenir ve çalışır. Programlama dilini herkes çalıştırabilir ve geliştirebilir. Dolayısıyla veri işleme ile ilgili her türlü uygulama için R programlama dili sürekli güncellenmekte ve yeni özellikler eklenmektedir. Ortalama, korelasyon ve sıklık analizi gibi basit istatistiksel hesaplamalardan, çok seviyeli modelleme ve yapay zekâ uygulamalarına kadar uzanan geniş bir alan için geliştirilmiştir. İstenilen R sürümü, bilgisayardaki işletim sistemine uygun olarak seçilebilmekte ve indirilebilmektedir (Eder vd., 2015).

Çalışma üç bölümden oluşacaktır. Üç bölümde metin madenciliğinin aşamaları ve metin madenciliğinde en çok kullanılan R ve Python bilgisayar dilleriyle oluşturulan iki araç Türkçe ve Osmanlıca çalışmalarına örnek olması açısından değerlendirilecektir.

1. Metin Madenciliğinde Önişleme-Normalleştirme

Hangi dil olursa olsun bilgisayar incelemelerinde metinler üzerinde ön temizlik yapılması ve bilgisayar tarafından okuyabilir biçime dönüştürülmesi gerekmektedir. Bu aşamaya metin madenciliğinde *normalleştirme* aşaması denilmektedir. Araçları kullanmadan önce metinin derleme dönüştürülmesi sağlanacaktır. Bu aşama metnin diline bağlı olarak kendi içinde aşamaları bulunmaktadır. Burada Türkçe ve Osmanlıca üzerinden hareket edileceği için dört adımla bu aşamanın temel özellikleri sunulacaktır.

a. Metinde Transkripsiyon Standartlaştırılması

Bu adımda, kelimeler üzerinde düzeltme, metinde aynı transkripsiyon yazım şeklini oluşturma işlemleri yapılmaktadır. Dolayısıyla kelimeler arasında standart bir yazım şekline ulaşılacaktır. Bu da bilgisayarın farklı yazılan ancak aynı olan kelimeleri farklı kelimeler olarak değerlendirmesini engelleyecektir.

tanıtılmaktadır. Türkiye'de bu alanda Koç Üniversitesinde bir Yüksek Lisans programı ve Marmara Üniversitesinde de bir araştırma merkezi bulunmaktadır. Ben bunlardan Koç Üniversitesinin *Social ComQuant Project* altında oluşturulan iki yaz okuluna katıldım. Her iki yaz okulunda da programlama dilleri temel seviyeden itibaren farklı seviyelerde okula katılan sosyal bilimler uzmanlarına öğretildi.

Bir diğer standartlaştırma adımında yazar ya da şair tarafından yapılan yanlış yazımların düzeltilmesi gelmektedir. Burada amaç yine bilgisayar tarafından kelimelerin farklı yazımlarından dolayı yanlış sayısal veriye ulaşmasını engellemektir.

Türkçede ve Osmanlıcada yer alan bazı transkripsiyon harflerinin (örneğin ç ve û gibi) aracın özelliklerine bağlı olarak bilgisayar tarafından okunur hâle dönüştürülmesi gerekmektedir. Eğer araç UTF-8 Unikodunu[3] okuyabilen bir araçsa pek çok transkripsiyon harfinin okunmasında sorun yaşanmamaktadır. Bu nedenle metin, UTF-8 Unikodlu formata sahip olması gerekmektedir. Bu, aşağıda ç maddesinde sunulan bir adımdır. Bu bölümde yapılacak bir diğer standartlaştırma Osmanlıcada hemze ve ayn harfine karşılık kesme işaretine benzeyen transkripsiyon işaretini bilgisayarın kesme işareti olarak görmesi problemidir. Bilgisayar me'mur gibi yazılan kelimelerdeki işaretleri kesme işareti olarak sayabilmekte ve kelimeyi ortadan ikiye ayırmaktadır. Bu da sayısal verinin yanlış sonuçlanmasına neden olmaktadır. Dolayısıyla bu problem de bu adımda çözülmelidir.

b. Noktalamaların Silinmesi

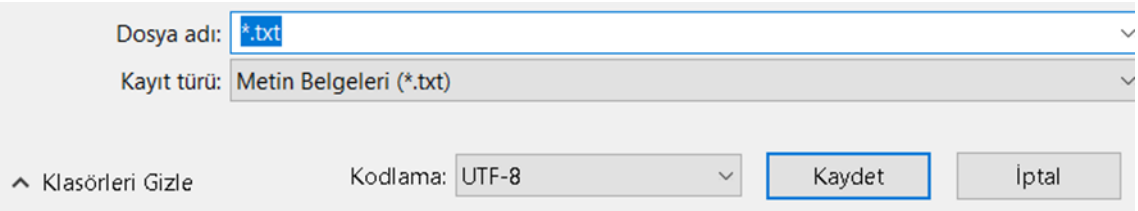
Bu adımda bilgisayarın metni okumasında problem oluşturmaması amacıyla metinde noktalama işaretlerinin silinmesi gerekmektedir.

c. Bazı Eklerin Düzeltilmesi

Türkçe eklemeli bir dil olduğu için soru eki olan mı/mi/mu/mu formları bilgisayar tarafından kelime olarak değerlendirilmektedir. Osmanlıcada buna örnek izafet kesreleridir. Bu ekler araştırmacıya bağlı olarak metinden silinebileceği gibi silinmeden elde edilen veriden bu ekler için tespit edilen verilerin çıkarılmasıyla da doğru sayısal verilere ulaşılabilmektedir.

ç. İşlenilen Metnin Utf-8 Unikodlu Txt File Formatına Dönüştürülmesi

Metinlerin, bütün araçlar tarafından okunabilmesi amacıyla, txt dosya formatına dönüştürülmesi gerekmektedir. UTF-8 Unikodu, bir metin karakter kodlama yöntemidir. Metinler txt formatında kaydedilirken txt dosyasının bu özelliği ile kaydedilmesi beklenmektedir. Metin txt dosyası olarak aşağıda görüldüğü gibi UTF-8 Unikodu ile kaydedilmelidir:



(Görsel 1: Word metninin UTF-8 Unikodlu txt file metnine dönüştürülmesi)

[3] Transkripsiyon harflerinin bilgisayarın okuyabileceği 8-bitlik bir Unicode sistemine dönüşüm biçimidir.

2. Metin Madenciliğinde Bilgisayar Dillerinin ve Dijital Beşerî Bilimler Araçlarının Kullanımı

a. Metin Analizlerinde R Program Dilinin Kullanımı: Hesaplamalı Metin Analizi İçin R Paketi ile Stylo

Stylometry, İngilizce "style" ve "metry" kelimelerinin birleşmesiyle "stil ölçümü" gibi bir anlamla yeni bir kelime oluşturmuş bir araçtır. İngilizce style, "stil, tarz, tip, biçim, format, moda, şıklık, çeşit, mil, madde, teknik, kalem, kalem ucu" gibi anlamlara gelmektedir. Aynı şekilde "metry" İngilizce "ölçü, ölçüm" anlamında kullanılmaktadır. Bu araç yazı stilinin yüksek seviye analizini sağlayabilmek için bir programlama dili olan R'den yararlanarak oluşmaktadır. Bu nedenle adı "Stylometry with R"dır. Stylometry, yazı stilinin nicel çalışmasıyla ilgilenmektedir. Araç, elde bulunan nüshalarda yazarlık doğrulaması yaparak tarihsel araştırmalarda kullanılabildiği gibi, adlî bağlamlarda delillerin incelenmesi ya da intihal yapıldığının kanıtı olması açısından önemli bir potansiyele sahip bir uygulamadır. *Stylometry with R*'ı 2015'te inşa edenler Maciej Eder, Jan Rybicki, Mike Kestemont'tur (Eder vd. 2015).

Aracın amacı, yazarı belli olmayan ya da başka birine atfedilen eserlerin yazarlarını bu yazılım aracılığıyla tespit etmeye çalışmaktır. Böyle bir çalışma Türkçeye ya da Osmanlıcaya uyarlanabilirse mahlasları ortak olan ve birbirine atfedilen şiirlerin asıl yazarları tespit edilebilecektir. Aracı oluşturanlar da böyle bir mantıkla takma adla yayınlanan bir çalışmanın ünlü Harry Potter romancısı Rowling'e atfedilmesi ve Harper Lee'nin *To Kill a Mockingbird* adlı eserinin orijinal versiyonunun yayınlanması ve editörünün burada rol oynamış olabileceği hakkında tartışmalara cevap aramaya çalışmışlardır (Eder, Rybicki, & Kestemont 2016). Türkçe üzerine Prof. Dr. Fazlı Can (2015), "Yazı Üslubunun Zaman İçinde Değişimi" adlı çalışmasında bu aracı kullanmış ve orada aktadığı çalışmalarından biri "A short non-quantitative presentation in the remembrance meeting of Ali Teoman 'Yazmadıklarını Yazan Yazar' ("The Writer who Writes his Unwritten Writings")"dır. Prof. Dr. Can, burada bir romanın yazarı konusundaki iddialara bu aracı kullanarak cevap vermeye çalışmıştır.

Edebî eser analizinde Stylometry metni okumadan yola çıkmaz; bunun yerine, hesaplama teknikleri ve görselleştirmeleri kullanarak büyük metin koleksiyonlarını anlamaya çalışmaktadır. Genellikle Stylometry analizleri, ön işleme, özellik çıkarma, istatistiksel veriyi belirleme ve son olarak sonuçlarını görselleştirme yoluyla sunulmasından oluşan karmaşık, çok aşamalı bir sistemdir.

Aracın kurulumu için R programlama dilinde yazılan kaynak dosyaların bilgisayara indirilmesi gerekmektedir. Bu dosyalar, *Comprehensive R Archive Network*'ten ücretsiz olarak indirilebilmektedir. Aracı geliştirenlerin paylaştığı kodlar kolayca uyarlanabilir ve genişletilebilir. Stylometry, kısaca Stylo'nun kodu açık erişime sahiptir ve lisanslıdır. Bu kodlara GitHub[4] üzerinden ulaşılabilmektedir. Araç bilgisayara aşağıdaki dizimle kaydedilmektedir.

C:\Users\Masaüstü\R_Workshop\corpus

Stylo (Stylometry), metinlerde üslup analizini yapmaktadır. Stylo'yu kullanmak için "R_Workshop" adında bir klasörün masaüstünde oluşturulması gerekmektedir. "R_Workshop" klasörünün içine "Corpus" adında yeni bir klasör

[4] https://github.com/computationalstylistics/stylo

oluşturulmalıdır. Daha sonra analiz için bir derlem meydana getirilmelidir. Derlemin içinde karşılaştırılacak en az iki metin bulunmalıdır. Daha sonra, metin dosyası bağlantısına sağ tıklayarak ve "bağlantıyı farklı kaydet" aracılığıyla dosya başlıkları için "Yazaradı_metinnumarası" şeklinde formatı kullanarak metin adlandırılmalıdır. Dosyalar "corpus" klasörüne kaydedilmelidir. Sonra analiz için kullanılacak Stylo paketi indirilmelidir. Stylo paketi https://github.com/computationalstylistics/stylo adresinden indirilebilmektedir.

Bu çalışma için "corpus" dosyasının içine Latîfî Projesi kapsamında Hâlet Efendi ve Râşid Efendi el yazmalarından alınan hâtime bölümü ile Ahmed Paşa, Dâ'î, Fânî, Fazlî, Mihrî Hatun, Şeyhî ve Zeyneb Hatun maddelerinin bulunduğu iki dosya txt formatıyla "corpus" dosyasına yerleştirildi.

Corpus dosyasından sonra internetten indirilen komut dosyası R_Workshop klasörüne taşınmalıdır. Dolayısıyla dosya böyle bir görüntüye sahip olacaktır:

Ad	Değiştirme tarihi	Tür	Boyut
corpus	23.12.2020 15:14	Dosya klasörü	
.RData	20.03.2018 23:49	R Workspace	3 KB
.Rhistory	20.03.2018 23:49	RHISTORY Dosyası	1 KB
R_Workshop_CA_100_MFWs_Culled_0_C...	20.03.2018 23:36	Microsoft Excel C...	1 KB
stylo_config	20.03.2018 23:36	Metin Belgesi	2 KB
table_with_frequencies	20.03.2018 23:36	Metin Belgesi	144 KB
wordlist	20.03.2018 23:36	Metin Belgesi	36 KB

(Görsel 2: R_Workshop klasörü)

Stylo'da derlem hazırlama, metin verilerinin ya R dosyasında ya da özel bir klasörde depolanan derlem dosyalarından doğrudan yüklenmesine izin vermesiyle oluşturulur. Derlem oluştururken verileri tanımlayan ve onlarla ilgili bilgi veren bir veri kümesi olan metedataya bir ad verilmesi gerekmektedir. Dosya adı, büyük ve küçük harfe duyarlı herhangi bir karakter dizisinin -bu yazar adı da olabilir-, ardından bir alt çizgi eklenmelidir. Oluşturulması hedeflenen grafiklerde (scatterplots ve dendrograms), örneklerin renkleri bu kurala göre atanır; ortak dosya uzantıları atılır. Örneklerin yazar sınıflarına göre renklendirilmesi için dosyalar şu şekilde adlandırılabilir:

REfendi_Fani.txt

HEfendi_Fani.txt

Baki_Divani.txt

Fuzuli_Divani.txt

Aşağıdaki tüm örnekler, yazarların proje web sitesinden indirilebilen veri setlerinde kullanıcı tarafından çoğaltılabilir. Örneğin proje kapsamında araştırmacılar veri kümesi (dataset) olarak Jane Austen ve Brontë kardeşlerin dokuz romanını kullanmışlar ve bu derlemi "data(novels), data(galbraith), data(lee)" gibi dosyalar şeklinde isimlendirmişler.

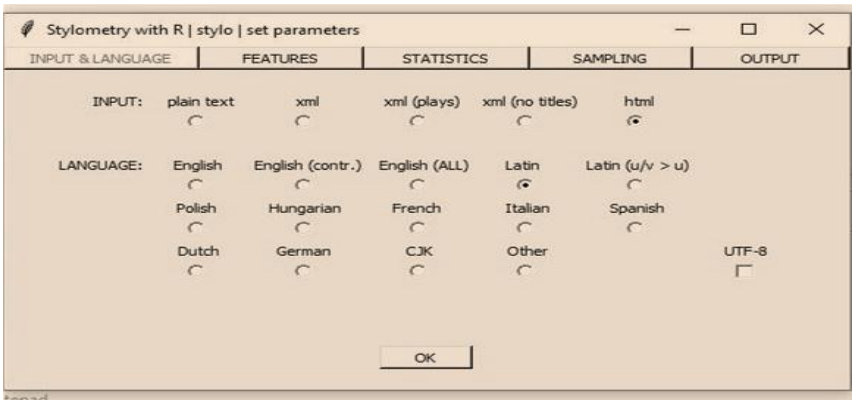
Bir derlem olarak kaydedilen tüm dosyalar "Corpus_files" adı altında toplanmalıdır ve bu dosyanın yüklenmesi için aşağıdaki komut yazılmalıdır:

raw.corpus <- load.corpus(files = "all", corpus.dir = "corpus_files", encoding = "UTF-8")

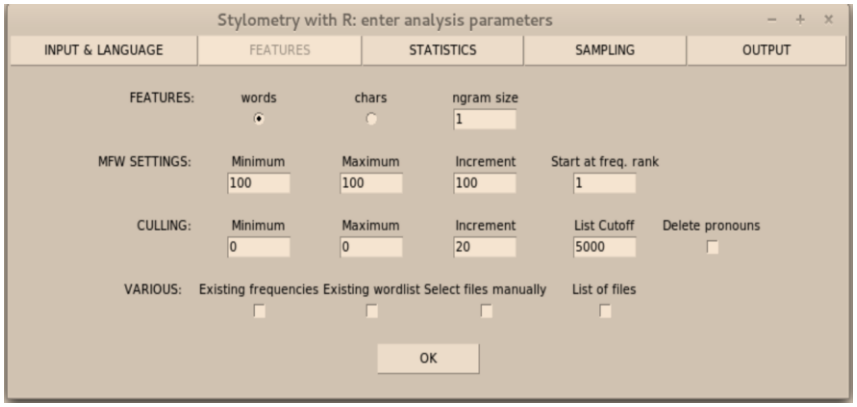
Stylo, metnin yükleme işleminde metin dosyasını pek çok araçta olduğu gibi UTF-8 ile kodlanmış txt dosya biçimini kullanmaktadır.

Görsel 3'te görüldüğü gibi Stylo, verileri önceden işlemek için İngilizce, Latince, Almanca, Fransızca, İspanyolca, Felemenkçe, Lehçe, Macarca, Korece, Çince, Japonca, İbranice, Arapça, Kıpti ve Yunanca gibi dillerin alfabelerini desteklemektedir. Görsel 4'te olduğu gibi, bu programlama dilinde simgeleştirme, bir dizi girdi metnini sözcük simgeleri gibi sayılabilir birimlere bölme sürecini ifade eder. İngilizce metinleri belirtmek için, örneğin öğeleri 'don't' kelimesini 'do' ve 'n't' olarak ayırıp tüm kelimeleri küçük harfle ayırarak bir sonraki komut kullanılabilmektedir.

Stylo'da yapılabilecek bir özellik, etkisiz kelimeleri veya ekleri (stop words) silme işlemidir. Bunun için delete.stop.words() kodunun kullanılması gerekmektedir. Metin madenciliği ile incelenecek metinde öncelikle metin kelime sayılarına göre kümelere ayrılması gerekmektedir. Sonra ayrılan kelime kümelerinde sık tekrarlanan kelimeler, harf sayısı, harf çiftleri benzerliği, kelime ve cümle uzunluğu gibi hesaplamalarla metinlerin üslup açısından değerlendirilmesi yapılabilmektedir. Stylo kurulduğunda açılan *set parameters* ayar sayfası Görsel 3'te olduğu gibidir.

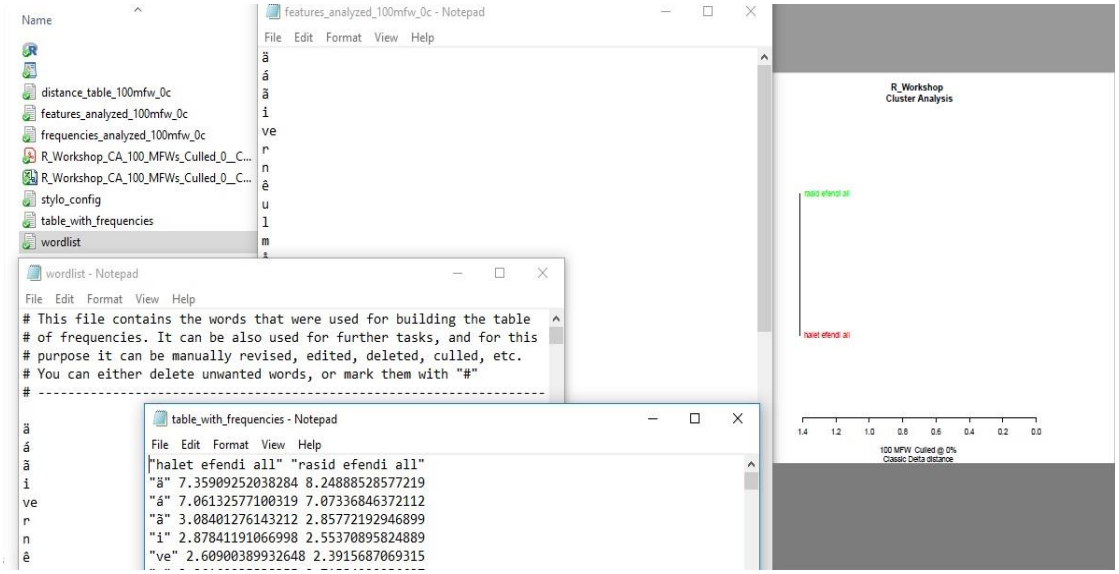


(Görsel 3: Stylo parametreleri)



(Görsel 4: Stylo analiz parametreleri)

Bu çalışma kapsamında Görsel 5'te görüldüğü gibi, Latîfî Tezkiresi'nin 1546 tarihli Râşid Efendi ve 1575 tarihli Hâlet Efendi el yazmalarından birer derlem oluşturuldu ve karşılaştırma yapıldı:



(Görsel 5: Stylo aracında Latîfî Tezkiresi'nin 1546 tarihli Râşid Efendi ve 1575 tarihli Hâlet Efendi elyazmalarından oluşan derlemin görüntüsü)

Bu program ile yapılabilen örnek inceleme Prof. Dr. Fazlı Can'ın Türk romanı üzerine yaptığı araştırmasıdır (Can, 2015). Romanlarda cümle uzunluklarının, en çok kullanılan kelimelerin de bulunduğu farklı açılardan karşılaştırmaların sonuçları ilgili çalışmada yer almaktadır (Can ve Patton, s. 2004).

b. Metin Analizlerinde Python Program Dilinin Kullanımı: Yunan ve Latin Şiiri İçin Ölçüm Aracı

Metinleri ya da şiirleri analiz etmek için sosyal bilimlere uygulanan dinamik bir programlama dili gereklidir. Python, en gelişmiş programlama dillerinden birini temsil eder. Birçok programlama ihtiyacını karşılayan kitaplıklarıyla çok yönlü uygulamalardan oluşur, ayrıca birçok programlama gereksinimini kapsayan ve destekleyen, çok basit ve tutarlı bir sözdizimi sunmaktadır. Sosyal bilimler için veri işlemeyi mümkün kılan kütüphanelere sahiptir. Bu tür

kütüphaneler, büyük ve çok boyutlu sayısal verilerin hesaplanmasını işlemek için NumPy, Jupiter, Matplotlib, Pygame, PySAL, Rpy ve Python 3 gibi kütüphaneleriyle son yıllarda önemli bir büyüme göstermiştir. Classical Language Toolkit (CLTK), modern öncesi Avrasya dilleri için doğal dil işleme (NLP) sunan bir Python kitaplığıdır ve 19 dil üzerine geliştirilmiştir (Brooker, 2019, s. 150).

The Metrical Tool Hylas, Antik Yunan şiirinde kelimelerin vurgusunu, hecelerin uzunluklarını analiz etmek için Python 3 kütüphanesini kullanmaktadır. Bu araç, Python'da geliştirilmiş, Yunan ve Latin şiirini algoritmik olarak tarayabilen ve kullanıcının 200.000'den fazla satırlık bir şiirde ölçü kalıpları aramasını yapmaktadır. The Metrical Tool Hylas, araştırmacıların verilen herhangi bir ölçülü şiir kalıbını bulmalarını ve ayrıca ölçümün yapısını tespit etmesini sağlamaktadır. The Metrical Tool Hylas, kurallara dayalı bir algoritma kullanmaktadır. Bilgisayara ölçüm birimlerinin öğretilmesiyle %98 başarı sağladığı söylenmektedir. Algoritma, Yunanca ve Latince şiirdeki hecelere yüzdeler atamaktadır: birincisi, iki hecenin birleştiği zaman ve ikincisi, belirli bir hecenin uzun mu kısa mı sayıldığı şeklinde kurulmuştur. Bu yaklaşımla araç %100 doğruluğa ulaşabilmektedir (Tueller, 2022, s. 26-29)

Aracın uygulama yöntemleri şöyledir:

- Şiirde bulunan uzun heceleri "–" işareti temsil etmektedir. Kısa heceleri de "ᵛ" işareti karşılamaktadır.
- Şiirde yaygın bulunan kalıplar vardır ve araç araştırmacının şiirine uygun olabilecek kalıpları vermektedir: Bunlara bir örnek bu şiir ölçüsü kalıbıdır: –ᵛ ᵛ |– –
- Heceler arasında ara vermemek için " ‿ " ve ara vermek için "," işareti seçilmelidir.

Osmanlıca yazılan şiirler de Arap geleneğinden gelen aruz ölçü birimine sahiptir. Şiirin en küçük birimi beyittir ve iki mısradan oluşur. Mısralar, tıpkı Yunan şiirinde olduğu gibi kısa ve uzun hecelerin birleşmesiyle oluşan ölçü birimleriyle yazılırlar. En çok kullanılan kalıplardan biri "Fâ'ilâtün / Fâ'ilâtün / Fâ'ilâtün / Fâ'ilün"dür ve kısalık ve uzunluk açısından aşağıdaki işaretler kalıbı temsil etmektedir:

— ᵛ — — | — ᵛ — — | — ᵛ — — | — ᵛ —

3. Bilgisayar Programlama Dilleriyle Oluşan Bu Araçlar ile Görseller Oluşturma

Bilgisayarlı dil ve edebiyat incelemelerinde metin madenciliği yapılacaksa kullanılacak aracın amacına dayalı farklı görselleştirmeler seçilebilmektedir. Burada bütün araçların ortak özelliklerinden birkaçı örnek olarak sunulacaktır.

a. Kelime Bulutu Oluşturma

Burada metinde yer alan kelimeler tespit edilip görselleştirilmektedir. Görsel içinde bulunan kelimelerin sıklıklarına göre büyüklükleri değişmektedir. En büyük size ile yazılmış kelimeler ya da yapılan metinde en sık tekrar edenlerdir. Aşağıda Latîfî Tezkiresi'nin kelime bulutu yer almaktadır. Görüldüğü gibi ve, bir gibi kelimeler ile izafet kesrelerini temsil eden –i ve –ı yapıları en sık tekrar edenlerdendir.

(Görsel 6: Latîfî Tezkiresi kelime bulutu)

b. Kelime Sıklıklarının Tespiti

Burada metinde geçen en çok tekrar eden kelimeler tespit edilerek görselleştirilmektedir. Dolayısıyla yazarın veya şairin üslubunu ele veren "tercih edilen kelime listesine" ve her kelimenin kaç kez tekrar edildiği bilgisine ulaşılabilmektedir. Aşağıda Görsel 7'de, Bâkî Dîvânı'nın kelime sıklık raporu bulunmaktadır. Raporun üstünde Bâkî Dîvânı'nda 12000 kelime çeşidi olduğu, bu kelimelerin de 49303 defa tekrar ettiği verisi bulunmaktadır. Altında en sık tekrar edilen ilk 10 kelimenin verisi sunulmaktadır. Araç, Farsça tamlamaları (ayrı yazıldığı için) birer kelime olarak saymaktadır. Bu yanlışlığı önlemek için yukarıda (c. Bazı Eklerin Düzeltilmesi maddesinde) verilen metin madenciliğinin ön işleme aşamasının yapılması gerekmektedir. Buradaki iki yanlış veri aracın tespit ettiği sayıdan çıkarıldığında Bâkî Dîvânındaki kelime çeşidinin 11998, kelime tekrarlama sayısının ise 43794 olduğu söylenebilmektedir.

Entries 12000 **Total Freq** 49303

	Type	Rank	Freq
1	i	1	3433
2	ı	2	2076
3	bir	3	484
4	dil	4	457
5	ol	5	448
6	gül	6	436
7	u	7	391
8	ile	8	368
9	ey	9	338
10	ü	10	327

(Görsel 7: Bâkî Dîvânı'nın kelime sıklık raporu)

b. N-grams

N-grams, kelimelerin eşdizimlerinin n sayısı kadar birlikte kullanımın tespit edilmesi anlamına gelmektedir. Burada yan yana gelen kelimeler 2 kelime de olabilmekte, 3 veya 4 de olabilmektedir. 2 kelimeden oluşan ve sık tekrar

edenlere 2-grams, 3 kelimeden oluşan ve birden çok tekrar eden kalıplara 3-grams denilmektedir (Tarhan, 2024, s. 230). Bunlar, metinde sık tekrar eden kalıplaşmış yapıların sayısal verisidir. Örneğin "nazar etmek" bir kalıplaşmış yapıdır. Metinlerde, *etmek* farklı ekler alarak kullanılabilmektedir. Bir yazarın bu iki kelimeyi ne kadar sık kullandığı yazarın hem dili hem de eserinin konusu hakkında araştırmacıya bilgi verebilmektedir. Aşağıda Görsel 8'de, *bezm* kelimesinin Fuzûlî Dîvânı'nda oluşturduğu 2-grams listesi yer almaktadır. Bu veriye göre *bezm* kelimesi yanındaki 18 çeşit kelime ile 49 farklı birliktelik aracılığıyla 2-gramları oluşturmuştur.

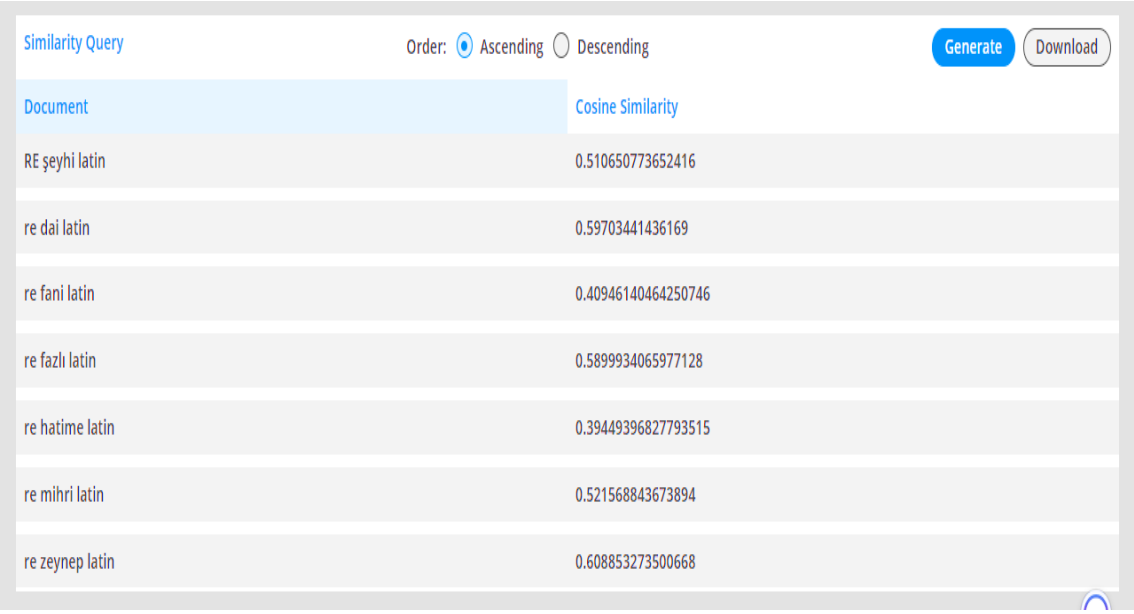
N-Gram Types 18 N-Gram Tokens 49 Pa

	Type	Rank	Freq	Range
1	bezm i	1	29	1
2	bezm gâh	2	2	1
3	bezm içre	2	2	1
4	bezm ârâ	2	2	1
5	bezm bir	5	1	1
6	bezm bî	5	1	1
7	bezm edip	5	1	1
8	bezm efrûz	5	1	1
9	bezm etmiş	5	1	1
10	bezm eyledi	5	1	1
11	bezm eyler	5	1	1
12	bezm gâhı	5	1	1
13	bezm gül	5	1	1
14	bezm için	5	1	1
15	bezm kânûnu	5	1	1
16	bezm olup	5	1	1

(Görsel 8: Bezm kelimesinin Fuzûlî Dîvânı'nda oluşturduğu 2-grams listesi)

c. Eserlerin Benzerlik Raporunu Oluşturma

Stylo aracı başta olmak üzere pek çok araç, eserler arasında benzerlik raporları da oluşturulabilmektedir. Bu aslında mahlas benzerliği olan şiirler ile şairlerin nazireleri üzerine uygulanınca iyi sonuçlara ulaşılabilecek uygulamalar olabilmektedir. Bu çalışma kapsamında Latîfî Tezkiresi'nin maddeleri arasında benzerlik raporunun oluşturulması sağlanmıştır. Görsel 9'da görüldüğü gibi, Latîfî Tezkiresi 1546 tarihli Râşid Efendi nüshasında hâtime bölümü ile Ahmed Paşa, Dâ'î, Fânî, Fazlî, Mihrî Hatun, Şeyhî ve Zeyneb Hatun maddelerinin üslup açısından karşılaştırması yapıldığında kelime farklılıkları açısından hâtime bölümünün 0.3 oranıyla en düşük benzerlik oranına sahip olduğu sonucuna ulaşıldı. Diğer 7 maddenin ise benzerlik oranlarının birbirine çok yakın olduğu ortaya çıktı. İncelemede 0.6 oranıyla Zeyneb Hatun maddesinde en çok benzerlik oranı çıktığı görüldü. Dolayısıyla yazarın Zeyneb Hatun maddesini tezkire geleneğinin klişe ifadeleriyle üzerinde çok düşünmeden yazdığı fikrini uyandırdı.



Similarity Query | Order: Ascending / Descending | Generate | Download

Document	Cosine Similarity
RE şeyhi latin	0.510650773652416
re dai latin	0.59703441436169
re fani latin	0.40946140464250746
re fazlı latin	0.5899934065977128
re hatime latin	0.39449396827793515
re mihri latin	0.521568843673894
re zeynep latin	0.608853273500668

(Görsel 9: Latîfî Tezkiresi 1546 tarihli Râşid Efendi nüshasında Hâtime bölümü ile Ahmed Paşa, Dâ'î, Fânî, Fazlî, Mihrî Hatun, Şeyhî ve Zeyneb Hatun maddelerinde benzerlik raporu)

SONUÇ

Çalışmada dil ve edebiyat incelemelerinde metin madenciliğinin yapılması için aşamalar açıklanmıştır ve örnek çalışmalar sunulmuştur. Dolayısıyla bu çalışma dil ve edebiyat çalışmalarıyla ilgilenenlere kılavuzluk edebilecek temel bilgileri sunmuştur.

Çalışmanın bir diğer çıktısı da Türkçe ve Osmanlıca için model araçlar sunmasıdır. Bir romanın ya da özellikle mecmualardaki bir şiirin ünlü yazar ya da şairlere ait olduğu iddiaları asırlardır edebiyat araştırmacılarının akıllarını yormuş ve bunun üzerine çalışmalar gerçekleştirilmiştir. Ancak bu çalışmalar nesnel verilere dayanmadığından soru işaretleriyle birlikte bilim dünyasında yerlerini almıştır. Stylo, bir diğer ismiyle Stylometry bilgisayar programlama dillerinden R ile üretilmiş bir yazılımdır ve edebiyatçıların uzun zamandır akıllarını yorduğu bir kitabın ya da şiirin kime ait olduğunu hesaplayabilen ve verileriyle olasılığı gösterebilen bir araçtır. Dolayısıyla R programlama dili edebiyat araştırmacılarının işlemlerini kolaylaştırabilecek yazılımlar üretmede kullanışlıdır.

Edebiyat araştırmacılarından yükselen problemleri, oluşturulan yazılımlarıyla ortadan kaldırabilecek, çok kullanışlı araç Python'dur. Python ile sayısal verilerin ortaya konulması ve görselleştirmeler yapmak daha kolaydır. Bu bağlamda bu çalışmada örnek olarak sunulan Python ile yazılan bir araç Yunanca ve Latince şiirlerin ölçülerini bulabilen The Metrical Tool Hylas'dur. Makalede hem Stylo hem de The Metrical Tool Hylas Hylas örnek olarak verilmiştir. Ulaşılan noktada bu araçların ileride Türkçe ve Osmanlıca üzerine yapılabilecek araç ve programlar için model olmaları önerilmektedir.

KAYNAKÇA

Aksan, Doğan (2013). *Şiir Dili ve Türk Şiir Dili*. Ankara: Bilgi Yayınevi.

"An Interdisciplinary Bibliography for Computers and the Humanities Courses". *Computers and the Humanities.* 25 (5), 1991, 315–26. https://doi.org/10.1007/BF00120968.

Atan, Suat (2020). "Metin Madenciliği: İmkânlar, Yöntemler ve Kısıtlar". *Mehmet Akif Ersoy Üniversitesi SosyalBilimler Enstitüsü Dergisi.* (31): 220-39.

Bâkî Dîvânı, İstanbul Üniversitesi Nadir Eserler Kütüphanesi T 5571 Numara İle Kayıtlı Bâkî Dîvânı Yazması.

Berry, David M., ed. (2012). *Understanding Digital Humanities*. Houndmills, Basingstoke, Hampshire ; New York: Palgrave Macmillan UK.

Brooker, Phillip D. (2019). Programming with Python for Social Scientists. SAGE. https://books.google.com.cy/books?id=Frq9DwAAQBAJ

Can, Fazlı (2015). *Things Hidden Behind The Numbers of İnce Memeds*. http://www.cs.bilkent.edu.tr/~canf/libraryTalk2015.pdf. Bilkent Library Lunchtime Lecture.

Can, Fazli; Jon M. Patton (2004). "Change of Writing Style with Time". *Computers and the Humanities*. 38 (1): 61-82.

Classical Language Toolkit (CLTK) 21.11.2021. Erişim Adresi: http://cltk.org/

Çelİk, Sadullah. (2020). "Metin Madenciliği ile Shakespeare Külliyatının İncelenmesi". *MANAS Sosyal Araştırmalar Dergisi.* 9 (3): 1343-57.

Eder, Maciej, Jan Rybicki, ve Mike Kestemont (2016). "Stylometry with R: A Package for Computational Text Analysis". *The R Journal*. (8) 1: 107. https://doi.org/10.32614/RJ-2016-007.

_____ (2015). *Package 'stylo' Functions for a Variety of Stylometric Analyses*. https://mran.microsoft.com/snapshot/2015-11-17/web/packages/stylo/stylo.pdf.

Fuzulî. (2021). *Fuzûlî Dîvânı. hzl.* Abdulhakim Kılınç. Ankara: Türkçe Yazma Eserler Kurumu Başkanlığı.

Grabar, Darko (2009). "07. 1957-2007: 50 Years of Higher Order Programming Languages". 33, 1: 72.

Laṭīfī. 1575. "Laṭīfī Tezkiresi" Hâlet Efendi Nüshası. Süleymaniye Kütüphanesi, Hâlet Efendi Koleksiyonu, 342 Numaralı Nüsha.

Laṭīfī. 1546. "Laṭīfī Tezkiresi" Kayseri Râşid Efendi Nüshası. Kayseri Râşid Efendi Kütüphanesi, Yazma Eserler Bölümü, 1160 Numaralı Nüsha.

Mailund, Thomas (2017). *Beginning Data Science in R. Berkeley*. *CA: Apress.* https://doi.org/10.1007/978-1-4842-2671-1.

Mason, Winter, Jennifer Wortman Vaughan, ve Hanna Wallach (2014). "Computational Social Science and Social

Computing". *Machine Learning* 95 (3): 257-60. https://doi.org/10.1007/s10994-013-5426-8.

Python. 15.11.2021. Erişim Adresi: http://www.python.org/

Tarhan, Ayşe (2024), *Exploring the Applicability of Digital Humanities Tools to Ottoman and Turkish Languages*, Eastern Mediterranean University, Master Thesis of Technology in Information Technology, Gazimağusa, North Cyprus.

_____ (2021). Digital Ottoman Projects. Geliş tarihi 19 Ocak 2021, gönderen https://digitalottomanprojects.org/

_____ (t.y.). The Latifi Project / Latifi Projesi. Geliş tarihi 19 Ocak 2021, gönderen http://www.thelatifiproject.org/

The Metrical Tool Hylas. 15.11.2021. Erişim Adresi: http://206.207.50.59/about

Tueller, Michael A (2022). "HYLAS: A new metrical search tool for Greek and Latin poetry". *Copyright\copyright 2022 AIUCD Associazione per l'Informatica Umanistica e la Cultura Digitale*.