Türk Kütüphaneciliği, 32, 4 (2018), 251-286
Doi: 10.24146/tkd.2018.41

Hakemli Yazılar / *Refereed Papers*
Araştırma Makaleleri / Research Articles

Türkçe Metinler Üzerine Yapılan Sayısal Üslup Araştırmalarını İnceleyen ve *Benim Adım Kırmızı* Çevirilerinin Aslına Olan Sadakatini Ölçen Bir Çalışma

A Survey of Stylometry Research on Turkish Texts and A Study on Quantification of Loyalty for Translations of My Name is Red

Sevil Çalışkan[*] ve Fazlı Can[**]

Öz

Bu makalede bilişimin beşerî bilimlerdeki önemli bir uygulaması olan sayısal üslup analizi yönteminin tanıtılması hedeflenmiş ve çevirilerin aslına sadakatini ölçen özgün bir araştırma sunulmuştur. Sayısal üslup analizi, bilgi ve belge yönetiminde çeşitli sınıflama işlemlerini gerçekleştiren ve edebiyat araştırmalarında yakın okuma sırasında görülmesi mümkün olmayan gözlemleri sağlayan yaklaşımlardan oluşmaktadır. Makalede, öncelikle Türkçe metinler üzerinde çalışmak isteyen araştırmacılar için, üslup analizinin Türkçeye nasıl uyarlanacağı anlatılmış ve bu konuda Türkçe metinler üzerinde yapılan çalışmaları inceleyen kapsamlı bir kaynak taraması sunulmuştur. Üslup analizinin uygulama amaçları örneklerle incelenmiş, ön işleme ve öznitelik çıkarımı, sınıflandırma yaklaşımları, başarı düzeyi değerlendirmesi ve yardımcı bilişim araçları konularına yer verilmiştir. Orhan Pamuk'un Benim Adım Kırmızı isimli romanı ve çevirilerindeki üslup uyumuna ilişkin sunulan özgün araştırma, roman kahramanlarının temel bileşenler düzlemindeki dağılımlarını inceleyen yeni bir yaklaşım kullanmaktadır. İstatistiksel olarak kayda değer olan gözlemler yazar üslubunun çevirilerde korunduğunu gösteren niteliktedir.

***Anahtar Sözcükle**r: Üslup analizi; metin madenciliği; yazar doğrulama; yazar ataması; metin sınıflandırma.*

Abstract

In this article an important problem of digital humanities, stylometry, is introduced and a novel study on quantification of translation loyalty is presented. Stylometry involves approaches that perform various classification tasks in information and document management and provides observations in literary analyses that cannot be obtained by close reading. A comprehensive

[*] Yüksek Lisans Öğrencisi. Bilgisayar Mühendisliği Bölümü, Bilkent Üniversitesi. e-posta: sevil.caliskan@bilkent.edu.tr
Graduate Student. Computer Engineering Department, Bilkent University, Turkey.
[**] Prof. Dr. Bilgisayar Mühendisliği Bölümü, Bilkent Üniversitesi. e-posta: canf@cs.bilkent.edu.tr
Prof. Dr. Computer Engineering Department, Bilkent University, Turkey.

Geliş Tarihi - ***Received***: 07.09.2018
Kabul Tarihi - ***Accepted***: 13.12.2018

survey of related studies and ways of adapting them to Turkish are presented for researchers who want to work on Turkish texts. In this context, the purpose of stylistic analysis, pre-processing, feature extraction and classification approaches, performance measures and available software tools are provided. Our new study on Orhan Pamuk's novel My Name is Red quantifies the consistency of translations with the original work and uses a new approach that examines the distributions of novel protagonists on the principal components analysis plane. Statistically significant observations show that the writer style is preserved in translations.

Keywords: *Stylometry; text mining; authorship verification; authorship attribution; text categorization.*

Giriş

Eskiler "Ars longa, vita brevis", "sanat uzun hayatsa kısadır" demişler (Schulz, 2011). Peki, günümüzde okunacak ya da incelenecek metinler hayali bile olanaksız bir hızla birikirken, bunların hepsini okumak mümkün müdür? Moretti (2013), bunun imkânsız olduğunun farkında olan bir edebiyat araştırmacısı olarak, uzaktan okumayı (distant reading) önerir. Tek bir çalışmayı (ya da bir grup çalışmayı) dikkatlice okumak ve analiz etmek yerine, uzaktan okuma, binlerce eserin analizi için bilgisayara güvenir. Başka bir deyişle, uzaktan okuma ancak bilgisayar destekli veri çözümleme teknikleri ile mümkün olacaktır. Bu tekniklerden en sık kullanılanlardan birisi de stylometry, yani üslup analizidir (Holmes, 1998).

Üslup analizi ya da stil analizi, edebi üslubun istatistiki veri madenciliği ve benzeri yöntemlerle incelenmesi şeklinde tanımlanabilir. Edebi üslup ise yazarlara özgü olabilen ve metinlerinde düzenli olarak görülebilen bir takım biçimsel ölçütlerdir (Tweedie, Singh ve Holmes, 1996). Daha geniş bir tanım yapacak olursak; bir metin üzerinde çeşitli sayısal ölçütler (öznitelikler) ile inceleme yapılarak, belirli bir yazarın edebi üslubundan izler aramak uğraşısına üslup analizi diyebiliriz (Oakes, 2009). Bu bağlamda, üslup analizi genellikle metinlerin içeriğinden bağımsız hareket eder.

Üslup analizinin tarihi Augustus de Morgan'ın, 1851 yılında yazdığı bir mektupta, yazarları kesin olarak bilinemeyen metinlerin kimin olduğunu, metinlerde geçen kelimelerin uzunluk sıklıklarına (metinde geçen aynı uzunluktaki kelimelerin sayılarına) bakarak çözülebileceğini önermesine uzanır (de Morgan, 1882). 1901 yılında Thomas C. Mendenhall; Bacon, Marlowe ve Shakespeare'in metinlerinin kelime uzunluğu dağılımlarını inceler. Shakespeare oyunlarının gerçek yazarını belirlemek amacıyla yapılan bu çalışma, bilgisayarlardan önce, yani el ile hesaplanarak yapılan ilk sayısal üslup analizi çalışmasıdır (Neal ve diğerleri, 2017; Tweedie ve diğerleri, 1996). Bilgisayar yardımı ile yapılan ilk çalışma ise Mosteller ve Wallace'ın, 1960'lı yılların başında, yazarları uzun süredir tartışmalı olan *The Federalist Papers* üzerine yaptıkları çalışmadır (Mosteller ve Wallace, 1964).

Türkçede metnin sayısal yaklaşımlarla incelenmesine ilişkin ilk çalışma olarak Mustafa İnan'ın 1963 yılında "Dil ve Matematik" konferansında sunduğu araştırması gösterilebilir: Çalışma "Kelime teşkilinde hece malzemesi ne oranda ekonomik olarak kullanılmaktadır?" sorusunu ele alır. Oğuz Atay *Bir Bilim Adamının Romanı Mustafa İnan* adlı eserinde bu çalışmaya ayrıntılı olarak değinmektedir (2001, s.155-56).

Üslup analizinin çözmeye çalıştığı problemler doğrudan bilgi ve belge yönetimi ile ilgilidir. Öte yandan literatürde Türkçe üzerine yapılmış çok sayıda çalışma bulunmakta ancak sayısal üslup analizini Türkçe metinler genelinde inceleyen bir araştırma olmadığı görülmektedir. Bu nedenlerle, bu makale ile üslup analizi kavram ve yöntemlerinin, öncelikle edebiyat araştırmacılarına ve kütüphanecilere, tanıtılması hedeflenmiştir. Aynı zamanda, bu alanda çalışan ya da çalışacak veri madenciliği ile ilgilenen ve farklı disiplinlerden gelen araştırmacılar için üslup analizinin Türkçeye uygulanabilirliğini ve yapılmış çalışmaları

inceleyen kapsamlı bir Türkçe kaynak taraması ile literatüre katkı yapılması amaçlanmıştır. Tarama sonrasında, Türkçe ile yapılan çalışmalardan farklı olarak Orhan Pamuk'un *Benim Adım Kırmızı* adlı romanın çevirileri ile ilgili özgün bir çalışma sunularak, üslup analizinin geniş uygulama alanları hakkında okuyuculara fikir verilmek istenmiştir.

Makalenin devamında, kısaca açıklamasını yaptığımız üslup analizi uygulamalarını ayrıntılandırarak, ön işleme, öznitelik elde edilmesi, sınıflandırma algoritmaları ve sonuçların değerlendirilmesinden bahsedeceğiz. Bu başlıklar altında istatistiksel doğrulama uygulamalarından da söz ederken, Türkçede yer bulmuş çalışmaları inceleyeceğiz. Makaleyi Orhan Pamuk'la ilgili olan çalışmamızı sunarak bitireceğiz. Üslup analizini daha ayrıntılı öğrenmek isteyen okuyucular için, Koppel, Schler ve Argamon'un 2009 yılında, Stamatatos'un yine 2009 yılında, Joula'nın 2008 yılında ve Neal ve diğerlerinin 2017 yılında yayımlanan makalelerini incelemeleri de yararlı olacaktır.

Uygulama Amaçları

Üslup analizi metinlere yazar ataması için yapılmaya başlanmış olsa da, gelişen teknoloji ve değişen zaman ile farklı amaçlar ve çözümler için de uygulanmıştır. Bu uygulamalar genellikle yazar ataması, yazar doğrulaması, yazar profilleme ve tarih ataması (stylochronometry) başlıkları ile literatürde yer bulmuştur.

Yazar atamasının amacı metinlerin belirli bir yazar tarafından yazılmış olması ihtimalini bulmaktır. Bu çalışmalarda altta yatan varsayım, yazarların üsluplarını bilinçli olarak değiştirebilmelerine rağmen, çalışmalarında her zaman kendi üslup özelliklerinden bir kısmını farkında olmadan tutarlı bir şekilde kullanacak olmalarıdır (Holmes, 1997). Yazar ataması, yazarları bilinen örnek metinlerin üslup analizi sonuçlarının, atama yapılacak metnin analiz sonuçlarıyla karşılaştırılması ile yapılır. Metne, örnek metinler arasından en benzer olan metnin yazarı atanabileceği gibi benzerlik için bir alt sınır konulduğu takdirde, örnek metin yazarlarından herhangi biri atanamayabilir. Benzer şekilde, belli bir benzerlik üst sınırını geçen metinlerin yazarlarının tümü farklı ihtimaller ile incelenen metne çoklu yazar olarak atanabilir (Afroz, Caliskan, Stolerman, Greenstadt ve McCoy, 2014). Bir başka yazar ataması uygulaması da bir yazarın belli bir alanda yazdığı metnin incelenmesi ile farklı alanda yazılan başka bir metnin yazarı olup olmadığının tespitidir. Bu problem "Bu romanın yazarı, bilinen köşe yazarlarından hangisidir?" sorusu ile doğrudan ilişkilendirilebilir. Türkçe metinler ile yapılan çalışmalar genellikle farklı yazarların köşe yazıları kullanılarak yeni bir metnin yazarının tespit edilebilmesini amaçlar (Diri ve Amasyalı, 2003; Şirin, Amghar, Levrat ve Acarman, 2017; Taş ve Görür, 2007; Taşçı ve Ekinci, 2012). Bunun yanında birkaç alanda birden yazan köşe yazarlarının metinlerinin kullanılmasıyla, belli bir alandaki yeni bir metnin yazarının bulunması üzerine de deneyler yapılmıştır (Aslantürk, Sezer, Sever ve Raghavan, 2010; Yavanoglu, 2016).

Yazar doğrulaması, Canbay, Sezer ve Sever'in (2018) de Türkçe metinler kullanarak yaptığı gibi iki metnin aynı yazar tarafından yazıldığına dair kanıt arar. Karşılaştırılan metinlerin aynı yazar tarafından yazılmadığı durumlarda, yazar doğrulama problemi açık-küme problemine dönüşür, başka bir deyişle metnin yazarı karşılaştırılan örnekler arasında olmayabilir (Koppel ve Winter, 2014).

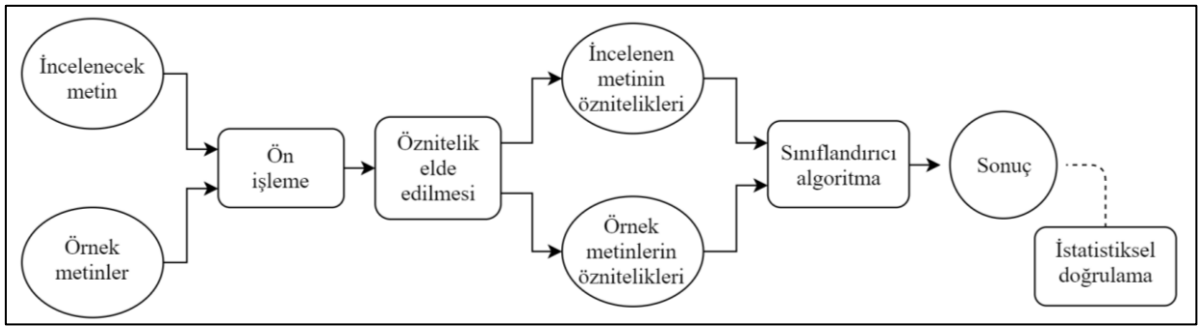
Yazar profilleme, yazar atamasının uygulanabilir olmadığı durumlarda, yazarın cinsiyeti, yaşı gibi demografik özelliklerinin belirlenmesi ile arama uzayını daraltmayı amaçlar (Reddy, Vardhan ve Reddy, 2016; Verhoeven, Skrjanec ve Pollak, 2017). Dil kullanımının yazarların yaşına ya da cinsiyetine göre değişim gösterebildiği daha önceki araştırmalar sayesinde bilinmektedir (Küçükyılmaz, Cambazoğlu, Aykanat ve Can, 2008; Peersman, Daelemans ve Vaerenbergh, 2011). Amasyalı ve Diri (2006), Türkçe köşe yazıları kullanarak yazarların cinsiyetlerinin ve yazının konusunun tahmini üzerine çalışmışlardır. Başka bir çalışmada, yazarların kişiliğinin profillenmesi için dışadönüklük, duygusal denge, uyumluluk,

vicdanlılık ve deneyime açıklık kişisel özelliklerinin metinler yardımıyla incelenmesi amaçlanmıştır (Verhoeven, Company ve Daelemans, 2014).

Tarih ataması, tarihi bilinmeyen metinleri, tarihi bilinen metinler ile karşılaştırarak tarih atamayı amaçlar. Üslup analizi dilin zaman içindeki değişimi göstermek için de kullanılmıştır (Altıntaş, Can ve Patton, 2007). Ayrıca, Can ve Patton'ın (2004, 2010) Yaşar Kemal'in bazı romanlarını ve Çetin Altan'ın gazete yazılarını kullanarak yaptıkları çalışmalarında olduğu gibi, yazarların zaman içindeki üslup değişimini incelerken dillerin zamanla değişimi ile de ilgilenir. Vurgulayacak olursak, dillerin zamanla değişimi verilen bir metne tarih atamasını mümkün kılar.

Geleneksel kullanımlarının dışında, üslup analizi edebi amaç gütmeden ve günümüzün zaman içinde durmadan değişerek gelen dijital metinlerine de uygulanabilir. Bilimsel makalelere yazar ataması uygulaması (Bergsma, Post ve Yarowsky, 2012) edebi amaç dışındaki kullanıma örnek gösterilebilir. Meyer ve Stein (2006), bilimsel makalelerdeki üslup değişimini gözlemleyerek intihal vakalarını tespit etmeyi amaçlamışlardır. Abbasi, Chen ve Salem (2008), internet üzerindeki film değerlendirmelerinin üslup analizini yaparak kullanıcıları fikirlerine göre sınıflandırmışlardır. Zheng, Qin, Huang ve Chen (2003), üslup analizi ile internet ortamındaki illegal mesajların ve e-postaların yazarlarını tanımlamaya çalışmışlardır. Bir başka ilginç çalışma da internet üzerinden yapılan Türkçe sohbet mesajlarının yazar tahmini üzerine yapılmış olan uygulamadır (Küçükyılmaz, Cambazoğlu, Aykanat ve Can, 2008). Tennyson'ın (2013), kaynak kodlara (program kodlarına) yazar ataması çalışması da dikkat çeker. Üslup analizi metinlerin üslubunu değiştirmek ya da başka bir metnin üslubunu kopyalamak için de kullanılabilir. Metinler bir yazarın tarzında yazılmak istenildiğinde üslup analizi yardımı ile yeni metin belli bir yazarın üslubuna benzetilerek yazılabilir. Bir metinden ya da kaynak kodundan yazarı ya da yazım tarihi hakkında bilgi edinilmesinin tercih edilmediği durumlarda, üslup analizi yardımı ile anlamın korunması amaçlanarak metnin üslubu değiştirilebilir (Kacmarcik ve Gamon, 2006; Nguyen, 2014).

Kısaca özetleyecek olursak, üslup analizi genel anlamda metinlerin biçimsel özellikleri bakımından karşılaştırılması ve karşılaştırma sonuçlarının istatistiksel olarak anlamlı olduğunun doğrulanması ile uygulanır. İstatistiksel olarak anlamlı sonuçlar, iki metnin benzerliğinin (ya da farklılığının), tesadüf sonucu değil, pek çok kere ve istikrarlı bir şekilde gözlemlendiğini ve yeni durumlarda da gözlemlenebileceğini ifade eder (Can, 2018). Şekil 1'de üslup analizi akış şeması görülebilir. İstatistiksel doğrulama, diskriminant çalışmalarında olduğu üzere sınıflandırıcı algoritmanın bir parçası olabileceği gibi, bazı algoritmalar için sınıflandırma sonrasında da uygulanabilir. İstatistiksel doğrulama çoğu çalışmada yer almasa da araştırmalarda sonuçların güvenilirliğini artıran önemli bir adımdır ve uygulanmalıdır.



Şekil 1. Üslup analizi akış şeması

Ön İşleme

Üslup analizi, metinlerin bazı sayısal özellikler ya da özniteliklere ayrıştırılarak incelenmesi ile mümkündür. Bu öznitelikler, metinlerin özelliklerini yansıtan ve sayısal olarak karşılaştırılabilmesine olanak veren, metinlerin yeni bir formu olarak düşünülebilir. Metnin bütün

özelliklerinin sayısal formlara dönüştürülmesi, hesaplama süresini artırırken sınıflandırma performansını düşürebilir. Sadece sınıflandırma açısından önem taşıyabilecek özelliklerin kullanılması doğru sınıflandırmalar için yeterli olabilir, aynı zamanda üzerinde işlem yapılacak öznitelik sayısı azaldığı için sonuç alma süresi de azalır. Bu sebeple, metinler öznitelik elde edilmesinden önce bazı ön işlemlere tabi tutulurlar. Ön işlemler genellikle karşılaştırılmada kullanılmayacak özelliklerin metinden çıkarılması şeklinde olur. Örnek verecek olursak, metin içinde geçen büyük harf ve küçük harflerin sayısı yazarların ayrıştırılmasında etkili olacak bir özellik değil ise, metin içindeki büyük harfler küçük harflere dönüştürülebilir. Bu ön işlem sonrası büyük harf - küçük harf sayısı öznitelik olarak elde edilemez fakat kelimelerin sayılmasında büyük harf - küçük harf ayrımı kalmadığı için bu işlemi kolaylaştırır.

Ön işlemler yazının türü ile doğrudan ilişkilidir. Bu sebeple ön işleme teknikleri yazı türleri kadar çeşitli olacaktır. Bunun yanında en çok kullanılan ön işleme tekniklerine birkaç örnek verebiliriz. Yazıları cümle, kelime gibi belli birim parçalarına ayırmak (*tokenization*); kelimelerin eklerini kaldırarak kökleri ile değiştirmek (*stemming*); kelimeleri sözcük türleri ile değiştirmek (*tagging*); harf olmayan karakterleri ve boşlukları kaldırmak; büyük harfleri küçük harflere çevirmek; dilde çok fazla kullanılan kelimelerin kaldırılması (*stopword removal*) ve benzeri biçimsel değişiklikler bazı önişleme teknikleri arasındadır (Çakır ve Güldamlasıoğlu, 2016; Neal ve diğerleri, 2017).

Literatürde ön işleme tekniklerinin incelendiği çalışmalar çok sayıda bulunmaktadır. Srividhya ve Anitha (2010), gereksiz kelimeleri kaldırarak ve metindeki kelimelerin yalnızca köklerini kullanarak sınıflandırma performanslarının geliştiğini belirtmişlerdir. Gonçalves ve Quaresma (2007), İngilizce metinler ile yaptıkları deneylerde kelimelerin köklerinin kullanılmasının daha iyi sonuç verdiğinden bahsederken, Portekizce metinler için gereksiz kelimelerin kaldırılmasının daha iyi sınıflandırma sonuçları verdiğini yazmışlardır. Torunoğlu, Çakırman, Ganiz, Akyokuş ve Gürbüz (2011), Türkçe metinlerle yaptıkları çalışmada gereksiz kelimeleri kaldırarak ve kelime köklerini kullanarak metinleri sınıflandırmaya çalışmışlardır. Deneyler sonucunda, bu iki ön işleme tekniğinin sınıflandırma performansına etkisinin göz ardı edilebilecek kadar az olduğunu raporlamışlardır. Tunalı ve Bilgin (2012), kelime köklerinin kullanılmasının Türkçe metinleri kümelemede performansı geliştirdiğine dair kanıt bulamamışlar fakat köklerin kullanımının performansı düşürmeden öznitelik sayısını azalttığını belirtmişlerdir.

Öznitelikler ve Özniteliklerin Seçimi

Öznitelik elde edilmesi, işlenmemiş veriden ya da bizim durumumuzda metinden, temsili veri çıkarma işlemidir. Bu temsili veri bahsedilen özniteliklerin birleşiminden oluşur. Üslup analizi için araştırmacılar tarafından çeşitli öznitelikler kullanılmıştır ve araştırmacılar öznitelikleri çeşitli şekillerde gruplandırmışlardır. İşi, Çemrek ve Yıldız (2013) öznitelikleri biçim bilgisi, sözcük bilgisi ve cümle bilgisi başlıklarında incelemiştir. Hurtado, Taweewitchakreeya ve Zhu (2014), sözcüksel özellikler, sözdizimsel özellikler, üslup özellikleri ve metne ya da yazara ilişkin özellikler başlıklarına yer vermiştir. Reddy ve diğerleri (2016) özenitelikleri karakter tabanlı özellikler, sözcüksel özellikleri, sözdizimsel özellikler, yapısal özellikler, içeriğe ilişkin özellikler, okunabilirlik özellikleri ve bilgi alma özellikleri olarak gruplamıştır. Stamatatos (2009) ise sözcüksel, karakter tabanlı, sözdizimsel, anlamsal ve uygulamaya ilişkin özelliklerden bahsetmiştir. Farklı gruplar olarak isimlendirilseler de aslında gruplanan özniteliklerin çoğu aynıdır.

Öznitelikler

Sözcüksel Özellikler

Bu özellikler sözcüklere dayanan ve sözcükler kullanılarak elde edilen özelliklerdir ve limitsiz sayıda oluşturulabilir. Sözcüklerin tanımlanabildiği dillerde kolayca uygulanabildiği için

hemen hemen her dile adapte edilebilir (Stamatatos, 2009). Metinler içinde iki boşluk arasında geçen her bir kelime, sayı ya da yapıya üslup analizinde *token* denilmektedir. *Token* olarak özellikler; metindeki toplam kelime sayısı, kelimelerin harf olarak ortalama uzunluğu, kelimelerin sesli harf olarak ortalama uzunluğu, toplam cümle sayısı ve cümlelerin kelime olarak ortalama uzunluğudur.

Kelime zenginliği ölçüsü olarak genellikle farklı kelime sayısının toplam kelime sayına oranı olarak hesaplanır (*type - token ratio*). Metinde yalnızca bir kere kullanılan kelimeler ve kelimelerin sayısı, yani hapax legomana, ve hapax dislegomana, yani iki kere kullanılan kelimeler ve bunların sayısı, da kelime zenginliği ölçüsü olarak kullanılabilir (Holmes, 1992). Kelime zenginliği için Zipf'in Yasası, Yule'un K Ölçüsü, Yule'un I Ölçüsü gibi çeşitli ölçüler de öne sürülmüştür (Neal ve diğerleri, 2017).

Bir başka yaklaşım kelime sıklıkları vektörü oluşturulmasıdır. Vektör uzunluğu metinlerdeki farklı kelime sayısı kadar olabileceği gibi (bag-of-words / kelimeler çantası), belirlenmiş bir uzunlukta en sık kullanılan kelimeler vektörleri de oluşturulabilir. Bu durumda vektörün her bir elemanı belirli bir kelimenin metinde kaç defa geçtiğini belirtir. Aynı işlemler sözcükler ile n-gramlar oluşturularak da yapılabilir. Sözcük n-gramları, metnin n tane kelime art arda gelen parçalara bölünmesini ifade eder. Bahsedilen işlemleri Orhan Pamuk'un *Benim Adım Kırmızı* isimli romanın ilk cümlesi ile kısaca örnekleyelim (1998, s.1).

"Şimdi bir ölüyüm ben, bir ceset, bir kuyunun dibinde."

Bu cümledeki toplam kelime sayısı dokuz ve kelime olarak cümle uzunluğu da dokuzdur. Ortalama kelime uzunluğu 4,67'dir (42/9). Dokuz kelimenin sesli harf sayıları sıra ile 2, 1, 3, 1, 1, 2, 1, 3 ve 3'dür. Ortalamaları ise 17/9, yani 1,89'dur. Kelime zenginliğine bakacak olursak, cümlenin toplam kelime sayısı dokuzdur fakat farklı kelime sayısı yedi olduğundan, 7/9, yani 0,78 olur. Kelime sıklığı vektörünü, cümledeki kelimelerin sırasını takip ederek oluşturduğumuzu kabul edersek vektör <1,3,1,1,1,1,1> olacaktır. Sıralama alfabetik ya da sıklığa dayalı olarak da yapılabilir. Son olarak cümleyi 2-gramlara bölecek olursak {şimdi bir, bir ölüyüm, ölüyüm ben, ben bir, bir ceset, ceset bir, bir kuyunun, kuyunun dibinde} parçalarını elde ederiz. Bu parçalar ile örnek verdiğimiz işlemler tekrarlanabilir. N-gramlarla en çok tercih edilen işlem sıklık vektörü oluşturulmasıdır.

Karakter Tabanlı Özellikler

Karakter tabanlı özellikler metinleri karakter dizileri olarak inceler ve metnin özelliklerinin karakter olarak sayısal formlara dönüştürülmesi ile yapılır. Alfabetik karakter sayısı, büyük harf - küçük harf sayısı, noktalama işaretleri sayısı, rakam sayısı, boşluk karakterlerinin sayısı ya da karakter sıklığı, noktalama işaretleri sıklığı vektörleri gibi özellikler karakter tabanlı özelliklere örnek olarak gösterilebilir. Karakter n-gramlar da sözcükler de olduğu gibi elde edildikten sonra özellik çıkarımı için kullanılabilirler. Karakter tabanlı özellikler de sözcüksel özellikler gibi pek çok dile uygulanabilir. Bir başka avantajı da yanlış yazımlardan ya da yanlış noktalama işareti kullanımlarından diğer özellikler kadar etkilenmemesidir (Stamatatos, 2009). Öte yandan n sayısının ne olacağının tespiti için genelgeçer bir yöntem olmadığından, pek çok deney yapılmasını gerektirebilir.

Sözdizimsel Özellikler

Sözdizimsel özellikler yazarların bilinçli olmadan benzer sözdizimlerini kullanacağı varsayımına dayanır. Sözdizimsel özellikler genellikle kelime türlerinin incelenmesi ile elde edilirler. Bunun için cümle ve sözcük öbeği yapılarının incelenmesi gerekir. Bu işlem daha önce bahsettiğimiz kelimeleri sözcük türleri ile değiştirme ön işleme metodunun uygulanmasını gerektirebilir. Sözcük türlerinin sıklığı (isim, fiil, sıfat gibi), sözcük öbeği ve cümle yapılarının sıklığı (isim tamlaması, ikileme, soru cümlesi, devrik cümle gibi) ve yardımcı eylem, bağlaç ya da edat gibi diğer

sözcüklerin yapısal ilişkileri için gereken işlevsel sözcüklerin (function words) sıklığı örnek verilebilir. Bu özellikler için kelime sıklığında olduğu gibi vektörler oluşturulabilir. Vektörlerin her elemanı bir sözcük türünün, sözcük öbeği ve cümle yapısının ya da işlevsel sözcüğün sıklığını temsil eder. Sözdizimsel özelliklerin dezavantajı olarak, kelime türlerini ya da metin içindeki yapıyı incelemek için başarılı doğal dil işleme metotları gerekmesini gösterebiliriz. Bununla birlikte Türkçedeki ek ve kök çeşitliliği, farklı ve karmaşık yapıların oluşmasına imkân tanığı için, doğal dil işleme uygulamaları Türkçe için zorlu bir görev olacaktır (Oflazer, 2014).

Anlamsal Özellikler

Anlamsal özellikler kelimelerin ve cümlelerin anlamlarından metnin özelliklerini yansıtmayı hedefler. Bunun için kelimeler arasındaki anlamsal bağlantılar tespit edilebilir, eş anlamlı ya da zıt anlamlı kelimeler kullanılabilir. Dil bilgisi kullanılarak kelimeleri anlamsal olarak bağlayan yapılar incelenebilir. Anlamsal özelliklerin hatasız şekilde oluşturulması zordur ve elde edilmeleri için farklı araçların kullanılması gerekebilir (Stamatatos, 2009).

Uygulamaya İlişkin Özellikler

Metinlerin genel üslubu hakkında bilgi veren tüm sayısal veriler öznitelik ya da özellik olarak kullanılabilirler. Bu nedenle, üslup analizi uygulamasına, bağlı olarak çok çeşitli özellikler oluşturulabilir. Bu özellikler, analiz edilecek metinlerin türü, analizin amacı ve hatta sınıflandırma için kullanılacak algoritma ile doğrudan ilintili olduğundan bunlar için avantaj sağlayacak özelliklerin kullanılması ya da oluşturulması beklenir. Örnek vermek gerekirse paragraf uzunluğu, tırnak işareti varlığı ya da yokluğu, font büyüklüğü ve rengi gibi yapısal özellikler e-posta, blog yazıları gibi internet üzerindeki metinlerin analizi için uygun olabilirken bunlara müdahale edilebilecek daha resmi yazılar için uygun olmayacaktır. Bazı anahtar sözcüklerin sayısı ya da sıklığı gibi içeriğe ya da alana ilişkin özellikler de oluşturulabilir (Zheng, Li, Chen ve Huang, 2006). Giriş, öz, dergi kelimeleri gibi anahtar sözcükler incelenen yazıların makale olabileceği yönünde bilgi sağlar. Yazım hataları ya da yazar ve yazı hakkında fikir verebilecek herhangi bir yazım farklılığı bile özellik olarak kullanılabilir. Çekirdek - çiğdem kelimeleri bu duruma örnek verilebilir. Bir metinde çekirdek yerine çiğdem kelimesinin kullanılması yazar ve metin hakkında İzmir ili ile ilintili olabileceği yönünde fikir verir.

Öznitelik Seçimi

Öznitelikler seçilirken ya da temsili veri elde edilirken bilgi kaybı olmamasına dikkat edilmelidir. Öte yandan metnin her özelliği her zaman kullanışlı olmayabilir. Guyon ve Elisseeff'in (2006) makalesindeki örneği verecek olursak, bir doktor hastalık teşhisi için kan basıncı, kan şekeri, ateş, boy ve kilo gibi değişkenlere bakabilir. Bir başka doktor bu değişkenlere yeme alışkanlıkları, ailede görülen hastalıklar hatta hastanın yaşadığı bölgenin iklimini bile ekleyebilir. Fakat değişken sayısı arttıkça, gereksiz ya da birbiriyle ilintili değişkenlerin inceleniyor olması ihtimali artar. Bu sebeple öznitelikler, en çok bilgi veren ya da katkı sağlayanlar arasından seçilmelidir. Bunun için bilgi kazanımı (information gain), kazanma oranı (gain ratio), simetrik belirsizlik (symmetrical uncertainty), korelasyon (correlation) gibi öznitelik seçimi metotları uygulanabilir (Jovic, Brkic ve Bogunovic, 2015).

Türkçe metinlerde yapılan çalışmalarda genellikle öznitelik seçimi yoğun olarak kullanılmamış, farklı özellik setleri ile deneyler yapılmıştır (Aslantürk, 2014). Bay ve Çelebi (2016), köşe yazıları ile yazar ataması için yaptıkları çalışmada ki-kare (chi-square) metodu ile öznitelik sayısını 20'den 17'ye düşürmüşler, sonrasında atama performansının yükseldiğini gözlemlemişlerdir. Türkoğlu, Diri ve Amasyalı (2007), çok sayıda sözcüksel ve sözdizimsel özellik kullanarak deneyler yapmış, 2.000'i aşkın özellik arasından korelasyon temelli özellik seçimi (CFS) metodunu kullanarak çeşitli özellik setleri elde etmişler ve bu özellik setleri ile atama başarısının daha yüksek olduğunu gözlemlemişlerdir. Aynı sonuçlar Türkoğlu'nun (2006)

yüksek lisans tezinde de görülebilir. Grieve'nin (2007) makalesinde de aynı veri seti kullanılarak pek çok özellik ile yapılan deneylerin sonucu sunulmuştur. Deneyler Türkçe metinler ile yürütülmemiştir fakat özellikler ve kullanımları açısından fikir verebilecek detaylı bir çalışmadır.

Sınıflandırma Yaklaşımları

Öznitelik elde edilmesi ya da seçiminden sonra yazarları ya da kategorileri bilinen metinler, bilinmeyen metinler ile karşılaştırılabilir hale gelir. Yazarları ya da kategorileri bilinen metinler sınıflandırma algoritmalarına karar vermede yardımcı olacağı için eğitici veri (training data) ya da metinler olarak isimlendirilirler. Yazarı ya da kategorisi araştırılan metinler de deney verisi (test data) olarak isimlendirileceklerdir. Sınıflandırma yaklaşımları çok çeşitlidir. Bu makalede bahsedeceğimiz yöntemler makine ile öğrenme, uzaklık temelli ve olasılık ya da istatistik temelli yaklaşımlar olacaktır. Çoğu uzaklık temelli ve olasılık temelli yaklaşım da makine ile öğrenme yaklaşımlarının içinde de değerlendirebilir, bu makaledeyse ayrı ayrı inceleneceklerdir. Tablo 2'de, Türkçe metinler kullanılarak yapılmış üslup analizi çalışmalarının bir kısmı incelenmektedir. Tablonun yöntem sütunu, çalışmalarda kullanılan sınıflandırma yaklaşımlarını belirtir. Tablo-1'de ise bu sütun için kullanılan kısaltmalar görülebilir.

Makine ile öğrenme, yazar ataması ve metin sınıflandırması çalışmalarında sıkça kullanılan yöntemlerdendir. Makine ile öğrenme yöntemleri sınıflandırıcı ve kümelendirici algoritmalar olarak ikiye ayrılabilir. Sınıflandırıcı algoritmalar, kategorileri bilinen veriyi kullanarak sınıflar arasındaki sınırları çizmek için eğitim yapar. Eğitim sonrası, sınıflandırılacak veriyi inceleyerek hangi sınıf sınırları içinde kaldığını hesaplar ve bir sınıfa atar. Bu algoritmalar denetimli öğrenme algoritmaları olarak isimlendirilirler. Kümelendirme algoritmalarında ise sınıflar ya da kategoriler önceden belirli değildir, bilinmemektedir ya da dikkate alınmazlar. Bu durumda kullanılan özellikler bağlamında birbirine benzeyen veriler aynı kümelere atanır. Bu kümeler bilinen sınıflar ile örtüşebileceği gibi bunlarla bağlantısız da olabilir. Kümelendirme algoritmalarında küme sayısının ne olacağı önemli sorulardan biridir. Üslup analizi çalışmalarında kullanılan makine ile öğrenme algoritmalarına naïve Bayes, karar ağacı (decision tree), destek vektör makinesi (support vector machine, SVM), yapay sinir ağları (artificial neural networks, ANN) ve derin öğrenme (deep learning) algoritmaları örnek olarak verilebilir.

Destek vektör makinesi Türkçeye uygulanan üslup analizi çalışmalarında yaygın olarak kullanılan algoritmalardan biridir. Üslup analizi çalışmalarında SVM'nin ayırt edici avantajı binlerce farklı özelliği işleyebilme yeteneğidir (Diederich, Kindermann, Leopold ve Paass, 2003). Türkoğlu ve diğerlerinin (2007) 2.000'i aşkın özellik vektörü ile yaptığı çalışmada SVM; naïve Bayes, rastgele orman ve çok katmanlı algılayıcı algoritmaları arasında en yüksek başarı düzeyine ulaşan makine ile öğrenme algoritması olmuştur. Yine Bozkurt, Bağlıoğlu, ve Uyar'ın (2012) çalışmalarında sözcüksel özellikler ve işlevsel kelimelerin sıklığı kullanılarak uygulanan histogram metodu, k-en yakın komşuluk gibi algoritmalar karşısında kelime çantası kullanılarak uygulanan SVM daha başarılı sonuçlar vermiştir. Yapay sinir ağları da metin sınıflandırmada ve yazar atamada kullanılan bir diğer algoritmadır. Tablo-2'ye bakıldığında, yapay sinir ağlarının Türkçede de SVM ya da naïve Bayes kadar yoğun olmamak ile beraber kullanım alanı bulduğu görülebilir.

Uzaklık temelli yaklaşımlar, sınıflandırılmak istenilen metnin, farklı sınıflara üye metinlere olan uzaklığının ölçülmesi temeline dayanır. Metin hangi sınıf üyelerine daha yakınsa, o sınıftan olması ihtimali artar. Bu yaklaşıma verilebilecek en temel örnekler k-en yakın komşuluk (k-nearest neighbor, KNN) ve k-ortalamalar (k-means) algoritmalarıdır.

Uzaklık temelli yaklaşımlarda kullanılacak uzaklık ölçüsü önemlidir. Uzaklığın hesaplanış biçimi, sınıflandırma başarısını yakından etkileyebilir. Bu sebeple araştırmacılar pek çok farklı uzaklık ölçüsü ortaya koymuştur. Metinlerin sınıflandırılması özelinde ise Burrows ve Stamatatos iki farklı uzaklık hesaplama yöntemi sunar. Burrows (2002), metinlerde geçen en sık kelimelerin, her metin için z-skorlarını hesaplar ve skorların farklarını uzaklık olarak

kabul eder. Stamatatos (2007) n-gram tabanlı uzaklık formülünde, her bir n-gramın yazarı bilinmeyen metindeki sıklığını ve karşılaştırılan yazarın metinlerinde aynı n-gramın geçme sıklığını kullanarak bir oran hesaplar ve her bir n-gram için hesapladığı oranları toplar. Bu toplam metin ile karşılaştırılan yazar arasındaki uzaklıktır. Bunlar dışında, ki-kare uzaklığı, kosinüs uzaklığı gibi yaygın kullanılan uzaklık hesapları da vardır.

Tablo 2'de sınıflandırma yöntemi olarak uzaklık temelli yaklaşımları kullanan çalışmalar görülebilir. Örnek verecek olursak, Canbay ve diğerleri (2018), yazar doğrulama amacıyla, sözcüksel ve sözdizimsel özellikleri kullanarak döküman vektörleri oluşturmuşlar ve şüpheli metnin vektörü ile doğrulanacak yazarın dökümalarının doğrudan kosinüs uzaklığı kullanarak hesaplamışlardır. Yazarların kendi metinleri arasındaki uzaklıkları incelemişler ve bir metni bir yazara atayabilmek için benzerliğin %100 - 75 arasında olması gerektiği sonucuna varmışlardır. Can, Can ve Karbeyaz (2010) ise Shakespeare'in sonatlarını ve Türkçe çevirilerini sık kullanılan sözcük öbekleri ve K-ortalamalar algoritması kullanarak kümelemişler ve farklı dildeki kümelerin benzerliğini araştırmışlardır. Küme benzerliğinin rastlantısal benzerlikten daha fazla olduğu ve çevirinin kaynak metnin anlamını koruduğu sonucuna varmışlardır.

Olasılık ya da istatistik temelli yaklaşımlar, genellikle söz konusu metnin bir yazara ait olma olasılığı ile ilgilenir. Bu olasılık $P(x|a)$ şeklinde, koşullu olasılık olarak belirtilir. Olasılık temelli yaklaşımlar aynı zamanda yazarların çoklu metinleri kullanılarak t-testi, varyans analizi (ANOVA), diskriminant analizi gibi yöntemleri de içerir. Bu yöntemler, ayrıca makine ile öğrenme algoritmalarının ve uzaklık temelli yaklaşımların güvenilirliğini ölçmek için de kullanılabilir. Yine Tablo-2'ye göz atıldığında istatistik temelli yöntemlerin tek başına kullanıldığı çalışmalar olduğu gibi, diğer yöntemler ile beraber kullanıldığı çalışmalar da görülebilir.

Makalede değinilmeyen pek çok sınıflandırma yöntemi ve uzaklık hesaplama seçenekleri bulunur. Hepsinin avantajları ve dezavantajları olduğu gibi, farklı durumlar için biri diğerinden daha uygun olacaktır. Bu sebeple, sınıflandırıcı algoritma ya da uzaklık formülü seçilirken, veri ya da metin yakından incelenmeli, aynı şekilde algoritma seçenekleri de araştırıldıktan sonra verinin özelliklerine göre en uygun olan algoritmalar seçilmelidir. Makalede bahsedilen yaklaşımlar çalışmalarda genellikle karşılaştırma amaçlı birlikte kullanılırlar. Aynı amaçla, sınıflandırıcı algoritma seçilirken, karşılaştırma yapmak ve beklenen sonuçların alındığını denetlemek için birkaç tane algoritma seçmek mantıklı olacaktır.

Tablo 1
Algoritma adları için kullanılan kısaltmalar

Kısaltma	**Algoritma**
NB	Naïve Bayes
MNB	Çok Değikenli Naïve Bayes
SVM	Destek Vektör Makinesi
LR	Lojistik Regresyon
DA	Diskriminant Analizi
KNN	K-En Yakın Komşuluk
KM	K-Ortalamalar Algoritması
ANOVA	Varyans Analizi
DECORATE	Yapay Eğitim Örneklerinin Karşıt Olarak Yeniden Etiketlenmesi ile Farklı Sınıflandırıcı Yaratma

Tablo 2
Türkçe metinler üzerinde yapılan üslup analizi çalışmaları

Referans	Amaç	Veri	Özellikler	Yöntem	Başarı Düzeyi
Agün, H. V., Yılmazel, S. ve Yılmazel, O. (2017)	Yazar ataması	En az 1.000 karakterli köşe yazıları, her yazar için 60 adet.	Sözcüksel ve sözdizimsel özellikler	Makine ile öğrenme (LR, çok değişkenli NB ve çok katmanlı sinir ağı)	F-skoru 0,37 - 0,95 (10 kat çapraz doğrulama)
Altıntaş, K., Can, F. ve Patton, J. M. (2007)	Dilin değişiminin sayısal tespiti	Dört yazar, yedi eserin Türkçe çevirileri	Sözcüksel özellikler ve sözdizimsel özellikler	İstatistiksel yöntemler (ANOVA, DA, LR, olasılık oranı)	Varyans analizlerinde 0,05'den az p-değerleri, diskriminant analizi için %80 doğruluk oranı
Amasyalı, M. F., ve Diri, B. (2006)	Yazar ataması, yazar profillemesi, metinlerin türlerine göre sınıflandırılması	Dört kadın, 14 erkek yazardan politika, spor ve genel kültür üzerine, 35'er köşe yazısı	Sözcüksel özellikler (2-grams ve 3-grams)	Makine ile öğrenme (NB, SVM, C4.5 ağacı, rastgele orman)	Yazar doğrulaması için %59 - 83, türlere göre sınıflandırma için %79 - 93, cinsiyet doğrulaması için %83- 96 doğruluk oranı (beş kez çapraz doğrulama)
Aslantürk, O., Sezer, E. A., Sever, H., ve Raghavan, V. (2010)	Yazar ataması	Dokuz yazardan, politika ve yaşam konularında toplam 513 köşe yazısı	Sözcüksel ve sözdizimsel özellikler	Makine ile öğrenme (kaba küme tabanlı sınıflandırma)	%70 doğruluk oranı
Aslantürk, O. (2014)	Yazar ataması	Sekiz yazarın 12.115 adet yaşam ve siyaseti konu alan köşe yazıları	Sözcüksel ve sözdizimsel özellikler	Makine ile öğrenme (kaba küme tabanlı sınıflandırma)	Toplam 1.134 deneyden 498 tanesi için %70 üzerinde doğrulukla oranı
Bay, Y., ve Çelebi, E. (2016)	Yazar ataması	17 farklı yazardan toplam 850 köşe yazısı	Sözcüksel özelikler	Makine ile öğrenme (NB, SVM ve Karar ağacı) ve uzaklık (KNN)	%96-100 arası doğruluk oranı (10 kat çapraz doğrulama)
Bozkurt, I. N., Bağlıoğlu, O., ve Uyar, E. (2012)	Yazar ataması	18 farklı yazardan her biri için 500 köşe yazısı	Sözcüksel ve sözdizimsel özellikler, İşlevsel sözcük sıklığı	Makine ile öğrenme (Histogram metodu, KNN, Bayes sınıflandırma, KM, bu algoritmaların kombinasyonu ve SVM)	En yüksek SVM ile %95,7 (10-kat çapraz doğrulama)
Can, E. F., Can, F., Duygulu, P., ve Kalpaklı, M. (2011)	Yazar ataması, tarih ataması	15-19. yüzyıla kadar beş farklı yüzyıldan on şairinin toplanmış divan eserleri	Sözcüksel özellikler	Makine ile öğrenme (SVM ve NB)	Yazar ataması için en yüksek %93, tarih ataması için en yüksek %95 doğruluk oranı (Çapraz doğrulama)
Can, F., Can, E. F., ve Karbeyaz, C. (2010)	Çeviri benzerliği ölçümü	Shakespeare'in sonatları ve Türkçe çevirileri	Sık kullanılan sözcük öbekleri	Uzaklık (KM ve Yao'nun formülü) ve istatistiksel yöntemler	Benzerliğin rastgele benzerlikten daha fazla olduğunu belirten düşük p-değerleri (<0,05)
Can, F., ve Patton, J. M. (2010)	Dilin değişiminin sayısal tespiti, tarih ataması, yazar profilleme	40 farklı yazardan faklı 10 yıllar için 40 roman	Sözcüksel özellikler	İstatistiksel yöntemler (temel bileşenler analizi, DA, doğrusal regresyon)	Cinsiyete göre sınıflandırmada %94,1, tarihe göre sınıflandırmada %57,27, sözcüklerin yıllar içinde uzadığını gösteren düşük p-değerleri (Çapraz doğrulama)
Can, F., ve Patton, J. M. (2004)	Yazar tarzı değişimi tespiti	Çetin Altan ve Yaşar Kemal'in eski ve yeni eserleri	Sözcüksel özellikler	İstatistiksel yöntemler (t-test, LR, DA, regresyon analizi)	Yazar tarzının yıllar içinde değiştiğini gösteren düşük p-değerleri (Çapraz doğrulama)

Canbay, P., Sezer, E. A. ve Sever, H. (2018)	Yazar doğrulama	12 farklı yazardan her biri için 100 köşe yazısı	Sözcüksel ve sözdizimsel özellikler	Uzaklık (Kosinüs uzaklığı)	En yüksek %92 doğruluk oranı
Canbay, P., Sever, H., ve Sezer, E. A. (2018)	Yazar ataması	10 farklı blog yazarından her biri için 50 blog yazısı	Sözcüksel özellikler (Noktalama işaretleri ve kelime çantası)	Makine ile öğrenme (SVM ve yapay sinir ağı)	Ortalama %25-75 doğruluk oranı (10 - kat çapraz doğrulama)
Dalkılıç, G., ve Çebi, Y. (2003)	Türkçede ortalama kelime uzunluğu hesaplanması	Farklı konulardaki web sitelerinden hem konuşma hem de yazı dilini temsil eden metinler	Kelime uzunluğu	İstatistiksel yöntemler (Olasılık hesaplama)	Ortalama kelime uzunluğunun 6,241 harf olduğu belirlenmiş, 7 harfe kadar olan kelimelerin külliyatın %69,11'ini oluşturduğu görülmüştür.
Demirci, S. (2014)	Duygu analizi	Toplam 6000 tweet	Sözcüksel özellikler, 1-gram, 2-gram ve 3-gramlar, dijital ifadeler	Makine ile öğrenme (NB, SVM) ve ve uzaklık (KNN)	En yüksek %70 doğruluk oranı
Diri, B., ve Amasyalı, M. F. (2003)	Yazar ataması	18 farklı yazardan her biri için 20 metin	Sözcüksel ve sözdizimsel özellikler	Skor tabanlı metot	En yüksek %84 doğruluk oranı
Karbeyaz, C. (2011)	İntihal tespiti	PAN'09 intihal veri kümesi, Leylâ ve Mecnun	Kelimeler çantası (bag-of-words)	Uzaklık (kapsama katsayısına dayalı kümeleme yöntemi)	En yüksek %30 doğruluk oranı
Küçükyılmaz, T., Cambazoğlu, B. B., Aykanat, C., ve Can, F. (2008)	Yazar ataması, yazar profillemesi, metin sınıflandırması	Sohbet mesajları koleksiyonu	Sözcüksel özellikler, karakter özellikleri, dijital ifadeler	Makine ile öğrenme (Patient Rule Induction Method, SVM, NB) ve uzaklık (KNN)	Yazar ataması %100 - 97, yazarların internet alanı tahmini %91 - 67, cinsiyet tahmini %81 - 71, yazarların okul tahmini %68 - 29, mesajların yazıldığı gün periyodu %71 - 41 arası doğruluk oranı (10 - kat çapraz doğrulama)
Patton, J. M., ve Can, F. (2014)	Yazar profillemesi, yazar tarzı değişimi tespiti	*İnce Memed* tetralojisi	Sözcüksel ve sözdizimsel özellikler	İstatistiksel (ANOVA, çoklu varyans analizi, DA)	Ciltler arasında üslup farkı olduğunu gösteren düşük p-değerleri, cilt sınıflandırmasında %87 doğruluk oranı (Çapraz doğrulama)
Patton, J. M., ve Can, F. (2012)	Çeviride değişen ve değişmeden kalabilen özelliklerin tespiti	James Joyce'un *Dubliners* hikayelerinin çevirileri ve orijinal metinleri	Sözcüksel özellikler	İstatiksel yöntemler ve DA	Farklı özellikler ile İngilizce ve Türkçe metinleri ayırmada %100 doruluk oranı (Çapraz doğrulama)
Saygılı, Ş. N., Amghar, T., Levrat, B. ve Acarman, T. (2017)	Yazar ataması	Dokuz yazardan her biri için 50, yedi yazardan her biri için 250 köşe yazısı	İsim-fiil, sıfat-fiil ve zarf-fiil sıklığı	Makine ile öğrenme (SVM)	İlk veri seti için 0,78 doğruluk, duyarlılık oranı ve F1 değeri, ikinci veri seti için 0,63 doğruluk oranı, 0,61 duyarlılık ve F1 değeri

Taş, T., ve Görür, A. K. (2007)	Yazar ataması	20 yazardan her biri için 25 köşe yazısı	Sözcüksel ve sözdizimsel özellikler, farklı kelime zenginliği özellikleri	Makine ile öğrenme (Bayes ağı, NB, MNB, NB güncellenebilir lojistik regresyon, Çok katmanlı öğrenme, Radyal temel fonksiyon ağı, Basit lojistik, Regresyon, DECORATE, Çok sınıflı sınıflandırıcı)	Özellik seçimi sonrası %80-57 doğruluk oranı (10-kat çapraz doğrulama)
Taşçı, H. ve Ekinci, E. (2012)	Yazar ataması	10 farklı yazardan 10 ayrı köşe yazısı	Karakter özellikleri ve işlevsel sözcükler	Uzaklık (Kosinüs uzaklığı)	Karakter özellikleri ile ortalama %86, işlevsel sözcükler ile ortalama %53 doğruluk oranı
Toraman, C., Can, F. ve Koçberber, S. (2011)	Metinlerin sınıflandırılması	Bilkent çevrimiçi portalından alınan haber yazıları	Kelime çantası (bag-of-words)	Makine ile öğrenme (C4.5, NB, SVM) ve uzaklık (KNN)	En yüksek %83 ve %87,5 doğruluk oranı
Torunoğlu, D., Çakırman, E., Ganiz, M. C., Akyokuş, S. ve Gürbüz, M. Z. (2011)	Metinlerin sınıflandırılması	Cafe, dünya, ege, ekonomi, güncel, siyaset, spor, Türkiye, yaşam kategorilerinde gazetelerden 2.230 doküman	Kelimeler çantası (bag-of-words)	Makine ile öğrenme (NB, MNB, SVM) ve uzaklık (KNN)	Eğitim için %50 üzeri veri kullanılan bütün deneylerde %70 üzeri doğruluk oranı
Tunalı, V. ve Bilgin, T. T. (2012)	Metinlerin gruplandırılması	Ekonomi, siyaset, spor, bilim, dünya, sanat, sağlık, Türkiye, yaşam ve yeşil haberler kategorilerinde haber yazıları	Kelimeler çantası (bag-of-words)	Uzaklık (Küresel K-ortalamalar)	k = 20 iken 0,966 saflık, 0,066 entropi ve 0,587 normalize edilmiş bilgi katsayısı (normalized mutual information)
Türkoğlu, F. (2006)	Yazar atama	18 yazara ait, 35 adet doküman alınarak 630 metin	Sözcüksel ve sözdizimsel özellik, n-gramlar, işlevsel kelimeler	Makine ile öğrenme (NB, SVM, Rastgele Orman, Çok Katmanlı Algılayıcı ve Öz Düzenleyici Özellik Haritası) ve uzaklık (KNN)	Farklı veri seti kombinasyonları için en iyi sonuçlar %82,1, %85 ve %89,2 doğruluk oranları
Türkoğlu, F., Diri, B., ve Amasyalı, M. F. (2007)	Yazar ataması	18 farklı yazardan her biri için 35 köşe yazısı	Çok sayıda sözcüksel ve sözdizimsel özellik, n-gramlar, işlevsel kelimeler	Makine ile öğrenme (NB, SVM, Rastgele Orman ve Çok Katmanlı Algılayıcı) ve uzaklık (KNN)	SVM ile ortalama %88,9 doğruluk oranı
Yavanoğlu, O. (2016)	Yazar ataması	Dokuz yazardan ekonomi, yaşam ve politika kategorilerinde 20000'i aşkın köşe yazısı	Sözcüksel ve sözdizimsel özellikler	Makine ile öğrenme (Yapay sinir ağları)	Ekonomi için %98, politika için %97, yaşam için %81 ve kategoriler arası %80 doğruluk oranları (10-kat çapraz doğrulama)

Başarı Düzeyi Değerlendirmesi

Başarı Ölçütleri

Sınıflandırma başarısının değerlendirilebilmesi için pek çok ölçüt kullanılmaktadır. Bunlardan en yaygın olanları arasında, doğruluk oranı (accuracy), doğru pozitif oranı ya da anma (recall), duyarlılık (precision), F-skoru ve ROC (reciever operator characteristics curve) eğrisi altındaki alan yer alır. Tablo 3, örnek metinlerin herhangi bir yazar A'nın metni olarak sınıflandırılması durumunda olabilecek durumları belirtir. Tek yazarlı bu durum üzerinden başarı düzeyi değerlerini örneklendirebiliriz. Aynı hesaplamalar, çok yazarlı durumlar için de yapılabilir. Bu durumda tablo, doğru sınıflandırılanlar ve yanlış sınıflandırılanlar şeklinde değişecektir.

Tablo 3
Hata matrisi (Confusion matrix)

	Yazar A olarak sınıflandırılan örnekler	**Yazar A olarak sınıflandırılmayan örnekler**
Yazar A'ya ait örnekler	Gerçek pozitif (True positive, TP)	Yanlış negatif (False negative, FN)
Yazar A'ya ait olmayan örnekler	Yanlış pozitif (False positive, FP)	Doğru negatif (True negative, TN)

Doğruluk oranı, bütün örnekler içinde doğru sınıflandırılan örneklerin oranını hesaplar.

$$Doğruluk\ oranı = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{(Formül 1)}$$

Anma doğru sınıflandırılan pozitif örneklerin oranıdır. Örnek dururumuzda, Yazar A'nın metinleri arasından gerçekten de yazar A olarak sınıflandırılan metinlerin oranıdır.

$$Anma = \frac{TP}{TP + FN} \qquad \text{(Formül 2)}$$

Duyarlılık doğru pozitif olarak sınıflandırılan örneklerin pozitif olarak sınıflandırılan örneklere oranıdır. Örneğimize bakacak olursak, doğru şekilde yazar A olarak sınıflandırılan metinlerin, doğru ve yanlış şekilde yazar A olarak sınıflandırılan metinlere oranlanması ile bulunur.

$$Duyarlılık = \frac{TP}{TP + FP} \qquad \text{(Formül 3)}$$

F-skoru duyarlılık ve anma arasındaki dengeyi ölçer. Sınıflandırmada doğru sınıflandırılan pozitif örneklerin çok olması istenirken aynı zamanda yanlış sınıflandırılan pozitif örneklerin düşük olması da istenir. Bunun sebebi bütün örneklerin pozitif sınıflandırılarak, 1 anma oranına erişilebilecek olunmasıdır. Ya da az bir miktar doğru pozitif sınıflandırma yapılarak 1 duyarlılık oranı elde edilebilir. Bu iki durum da gerçek başarı göstergesi olmayabilir. Böyle durumlarda iki ölçüt, yeni bir ölçüt hesaplaması için kullanılarak ikisi arasındaki denge görülebilir. F-skoru yükseldikçe, iki ölçü birbiri ile daha dengelidir.

$$F - skoru = 2 \times \frac{Duyarlılık \times Anma}{Duyarlılık + Anma} \qquad \text{(Formül 4)}$$

ROC eğrisi bir eksenin yanlış pozitif oranı ($\frac{FP}{FP+TN}$), diğerinin doğru pozitif oranı olduğu bir eğridir. ROC eğrisi altındaki alana AUC (Area Under the ROC Curve) denilmektedir. AUC yükseldikçe sınıflandırma doğruluğu artar.

Bahsedilen değerlendirme ölçütleri dışında, probleme bağlı farklı ölçütler de kullanılabilir. İstatistiksel yaklaşımlarda kullanılan p-değeri gibi ölçütler bunlara örnek olarak gösterilebilir. Başarı düzeyi değerlendirmesi için diğer ölçütler ve ayrıntılı bilgi Sokolova ve Lapalme'nin 2009 yılında yayımlanan makalelerinden elde edilebilir.

Başarı Ölçümü Yaklaşımları

Çapraz doğrulama başarı düzeyi değerlendirmesini daha güvenilir yapan bir yöntemdir. Çapraz doğrulamada, başarı düzeyi değerlendirmesi hiçbir zaman tek bir eğitim ve test seti ile yapılmaz. Değerlendirme için çok kere yapılan deneylerin ortalaması alınır. Sonuçlar eğitim ve test setlerine bağımlı olabileceği için, farklı setler ile deneyleri yinelemek güveni artırır. K-kat çapraz doğrulama için, eldeki veri seti k parçaya bölünür ve sınıflandırma k defa tekrarlanır. K deneyin her birinde farklı bir veri seti parçası test seti olurken kalan k-1 parça veri seti eğitim seti olarak kullanılır.

Başarı değerlendirmeleri karşılaştırmalı olmalıdır. Önerilen yönteme benzer bir çalışma literatürde varsa karşılaştırma bu çalışmanın yöntemi ile olmalıdır. Böyle bir yöntem yoksa en yüksek başarıyı elde etmek amacıyla karşılaştırma yöntem içinde parametreler değiştirilerek yapılabilir. Karşılaştırma sonuçları istatistiksel olarak test edilmeli ve sonuçların kayda değer olduğu gösterilmelidir. Eğer sonuçlar rakip yaklaşımlardan daha iyi değilse, önerilen yaklaşımın hangi koşullarda daha iyi sonuç vereceğinin araştırılması da ilginç sonuçlar sağlayabilir. İstatistiksel testler, sayısal üslup araştırmaları için hazır yazılımlar başlığında tanıtılan yazılımların bir kısmında dâhil edilmiştir. Dahil edilmeyenler ya da bu yazılımların kullanılmaması halinde R ya da uzun adıyla The R Project for Statistical Computing (R Core Team, 2014) ve SPSS (IBM Corp., 2017) yazılımları bu amaç için kullanılabilir. Son olarak, karşılaştırma sırasında rakip olarak literatürdeki güçlü yaklaşımların kullanılması karşılaştırmanın anlamlı olması açısından önemlidir.

Sayısal Üslup Araştırmalarına Yönelik Hazır Yazılımlar

Tablo 4'de üslup analizi yapılırken yardımcı olabilecek açık kaynak kodlu programlar ya da uygulamalar verilmiştir. Seçim sırasında, Türkçeye uygulanabilir olmalarına önem verilmiş ve başlangıç düzeyi programlama bilgisi ile kullanılabilecek olanlara da yer verilmeye çalışılmıştır. Uygulama açıklamaları tabloda görülebilir. Tablodaki araçlar dışında araç kullanımı halinde, seçilirken Türkçe karakter destekleyenlerin seçilmesine önem verilmelidir. Kelime türü belirleyen ve morfolojik analiz yapan araçlar dillere özeldir. Bu sebeple bu araçlardan Türkçe için tasarlanmış olanlar kullanılmalıdır.

Tablo 4
Üslup analizi için kullanılabilecek yardımcı araçlar

İsim	Açıklama	Bağlantı adresi
İTÜ Türkçe Doğal Dil İşleme Yazılım Zinciri	İstanbul Teknik Üniversitesi Doğal Dil İşleme Grubu tarafından geliştirilen Türk doğal dil işleme araçları bağlantı adresindeki internet sitesinde sağlanmaktadır (Eryiğit, 2014).	http://tools.nlp.itu.edu.tr/
JGAAP (Java Graphical Authorship Attribution Program)	Duquesne Üniversitesi tarafından geliştirilmiştir. Bazı ön işleme ve özellik elde etme teknikleri Türkçeye uygundur. Başlangıç düzeyi programlama bilgisi yeterlidir.	https://github.com/evllabs/JGAAP
JSAN	Metinlerin çeşitli özelliklerinin çıkarımı için seçenekler sunarak yazar tespitini amaçlar. Aynı zamanda metinlerin anonimliğini korumak için metinleri değiştirme seçeneği de vardır. Başlangıç düzeyi programlama bilgisi yeterlidir.	https://psal.cs.drexel.edu/index.php/Main_Page
Online Authorship Attribution Tool	Yazar ataması deneyleri için internet üzerinden kullanılabilen bir araç. Sayısal sonuçlar vermediği için bilimsel amaçlı kullanım mümkün olmayacaktır. Programlama bilgisi gerektirmez.	http://www.aicbt.com/authorship-attribution/online-software/
PRETO (Türkçe Metinleri Ön İşleme için Yüksek Performanslı Bir Metin Madenciliği Aracı)	Türkçe için kök bulma, gereksiz kelimeleri filtreleme ve n-gram üretimi gibi çok çeşitli önişleme seçenekleri sağlayan araçtır (Tunalı ve Bilgin, 2012).	*Bağlantı sağlanmamıştır. Program için yazarlar ile iletişime geçiniz.*
Signature	Harf / kelime uzunlukları ve sıklıklarının grafiksel gösterimlerini sağlayan yazılım. Kelime listelerini, cümleler, n-gramları destekler ve çoklu dil desteği vardır. Başlangıç düzeyi programlama bilgisi yeterlidir.	http://www.philocomp.net/humanities/signature.htm
StyleTool	Basit, kelime sıklığı tabanlı bir üslup analizi aracıdır. Başlangıç düzeyi programlama bilgisi yeterlidir.	https://github.com/lnmaurer/StyleTool
Stylometry with R: a Suite of Tools	Bir yandan, gelişmiş kullanıcılar için istatistik uygulamalarını sıfırdan oluşturma fırsatı sağlarken diğer yandan, daha az gelişmiş araştırmacıların hazır senaryo ve kütüphaneleri kullanmalarına olanak tanır (Eder, Rybicki ve Kestemont, 2016).	https://github.com/computationalstylistics/stylo
Trmorph	Türkçe için hazırlanmış morfolojik analiz aracıdır. Bu araçla üretilebilecek özellikler üslup analizinde kullanılabilir (Çöltekin, 2014).	http://coltekin.net/cagri/trmorph/index.php
Yıldız Teknik Üniversitesi Kemik Doğal Dil İşleme Yazılımları	Yıldız Teknik Üniversitesi Kemik Doğal Dil İşleme grubu tarafından hazırlanmış, Türkçe metinleri için çeşitli önişleme ve sınıflandırma yazılımları bağlantı adresinde sağlanmaktadır.	http://www.kemik.yildiz.edu.tr/?id=29
Zemberek	Türk dili işleme kütüphanesidir. Sözcük düzeyindeki özelliklerin istatistiksel bilgisini oluşturmak için kullanılır (Akın ve Akın, 2007).	https://github.com/ahmetaa/zemberek-nlp

Benim Adım Kırmızı Romanı ve Çevirileri Arasındaki Üslup Uyumunun Nicel Olarak Değerlendirilmesi

Çeviri metinlerde özgün metinin anlamı değiştirilmeden özgün metinin başka bir dilde ifade edilmesi amaçlanır. Peki, anlam korunurken üslup da korunabilir mi? Bu soruya cevap arayan yani, çeviri metinleri üslup açısından inceleyen araştırma örnekleri literatürde bulunur. Can ve diğerleri (2011), Shakespeare soneleri ile Türkçeye çevirileri arasındaki üslup ilişkisini sayısal olarak incelemeyi amaçlar. Patton ve diğerleri, James Joyce'un (2012) *Dubliners* hikâyeleri ile Türkçeye çevirileri arasında değişmeyen özellikleri üslup analizi ile belirlemeye çalışır. Baker (2000), üslup analizi kullanarak aynı metnin çevirilerini inceler ve farklı çevirmenlerin izlerini arar. El-fıqı, Petraki ve Abbass (2016) üslup analizini çevirmen tespiti için kullanmışlardır.

Bu çalışmada, Orhan Pamuk'un *Benim Adım Kırmızı* romanı ve romanın İngilizce, Fransızca ve İspanyolca çevirileri arasındaki üslup sadakatinin sayısal olarak değerlendirilmesi amaçlanmıştır. Bunun için romanda her biri farklı bölüm olarak yer alan karakterler kullanılmıştır. Karakterlerinin üsluplarının birbirlerine benzerliklerinin, üslubun değişmediği

çevirilerde korunacağı varsayımı test edilmiştir. Öncelikle özgün metin ve çeviriler için karakterlerin öznitelik vektörleri oluşturulmuş, bu öznitelik vektörleri arasındaki uzaklıkları hesaplanarak özgün ve çeviri metinler arasındaki uzaklıklar korelasyonu hesaplanmıştır.

Yöntem

Benim Adım Kırmızı, 59 bölümden oluşur ve her bölüm romandaki 20 farklı karakterden birinin sesinden yazılmıştır. Karakterlerin özgün metin ve çeviriler için listesi Tablo 5'de görülebilir. Tablodaki karakter sırası romanda ilk görünme sırası ile aynıdır. Çeviri metinlerde, Türkçeden İngilizceye çeviri (*My Name is Red)* Erdağ M. Göknar tarafından, Fransızcaya (*Mon Nom est Rouge*) Gilles Authier tarafından ve İspanyolcaya (*Me Ilamo Rojo*) Rafael Carpintero tarafından yapılmıştır.

Karakterlerin konuşmalarının/seslerinin romanda bölümler olarak ayrılmış olarak yer alması, karakterlerin üsluplarını ayrı ayrı inceleme olanağı sunar. Özgün metin ve çeviriler arasındaki üslup benzerliğin sayısal olarak değerlendirilmesi için öncelikle metinler bölümlerine ayrılmış, aynı karakterlere ait bölümler birleştirilmiştir. Bu birleştirme ile karakterlerin roman içinde dağılan metinleri/sesleri bir araya getirilerek üslup analizi için karakter özelinde en geniş veri setinin oluşturulması amaçlanmıştır.

Tablo 5
Karakter Tablosu

Türkçe	**İngilizce**	**Fransızca**	**İspanyolca**
Ben Ölüyüm	I am a corpse	Je suis mon cadavre	Estoy muerto
Benim Adım Kara	I am called Black	Mon nom est Le Noir	Me llamo Negro
Ben, Köpek	I am a dog	Moi, le chien	Yo, el perro
Katil Diyecekler Bana	I wıll be called a Murderer	On m'appellera l'Assassin	Me llamarán Asesino
Ben Eniştėnizim	I am your beloved Uncle	Je suis votre Oncle	Soy vuestro Tío
Ben, Orhan	I am Orhan	Moi, je m'appelle Orhan	Yo, Orhan
Benim Adım Ester	I am Esther	Mon nom est Esther	Me llamo Ester
Ben, Şeküre	I, Shekure	Moi, Shékuré	Yo, Seküre
Ben Bir Ağacım	I am a tree	Je suis l'arbre	Soy un árbol
Bana Kelebek Derler	I am called "Butterfly"	On m'appelle Papillon	Me llaman Mariposa
Bana Leylek Derler	I am called "Stork"	On m'appelle Cigogne	Me llaman Cigüeña
Bana Zeytin Derler	I am called "Olive"	On m'appelle Olive	Me llaman Aceituna
Ben, Para	I am a gold coin	Moi, l'Argent	Yo, el Dinero
Benim Adım Ölüm	I am Death	Mon nom est la Mort	Me llamo Muerte
Benim Adım Kırmızı	I am Red	Mon nom est Rouge	Me llamo Rojo
Ben, At	I am a horse	Moi, le Cheval	Yo, el caballo
Üstat Osman, Ben	It is I, Master Osman	Moi, Maître Osman	Yo, el Maestro Osman
Ben, Şeytan	I, Satan	Moi, le Diable	Yo, el Diablo
Biz, İki Abdal	We two dervishes	Nous, les deux Errants	Nosotros, dos derviches errantes
Ben, Kadın	I am a woman	Moi, la Femme	Yo, la mujer

Roman karakterlerinin üsluplarının karşılaştırılabilmesi amacıyla her bir karakter için bir öznitelik vektörü oluşturularak metinler sayısal formlara dönüştürülmüştür. Bu çalışmada aşağıdaki sözcüksel özellikler öznitelik olarak kullanılmıştır.

- Karakterlerin metinlerindeki kelime sayısı (*token* sayısı),
- Farklı kelime sayısı (*type* sayısı),
- Ortalama kelime uzunluğu (harf olarak),
- Ortalama farklı kelime uzunluğu (harf olarak),
- Ortalama cümle uzunluğu (kelime olarak),
- Ortalama kelime başına düşen sesli harf sayısı,
- Ortalama farklı kelime başına düşen sesli harf sayısı ve
- En sık kullanılan kelimelerin geçiş sayısıdır.

Çevirilerde ortalama kelime başına düşen sesli harf sayısı hesaplanırken kelimelerin yazılışları dikkate alınmış, okunuşları göz önüne alınmamıştır. Bu öznitelikler kullanılarak her bir karakter için öznitelik vektörleri oluşturulmuştur. Kelimelerin sıklık vektörleri oluşturulurken, her karakterin en sık kullandığı *k* kelimenin birleşimi alınmıştır. Bu vektörden tekrar eden kelimeler çıkarılmadan önce boyutu 20**k*'dir. Tekrar eden kelimeler her dil için farklı olacağından, özgün metin ve çeviriler için kelime sıklığı vektörlerinin boyutları farklıdır. Sözcüksel özellikler ve kelime sıklığı vektörleri sonrasında farklı kombinasyonlar ile birleştirilerek her bir karakter için öznitelik vektörleri oluşturulmuştur. Özgün metin ve her bir çeviri için karakter sayısı kadar yani 20 tane öznitelik vektörü vardır. Sık kullanılan kelime vektörleri ile diğer öznitelikler birleştirildiğinde elde edilen vektörlerin boyutu minimum 7+*k* olacaktır.

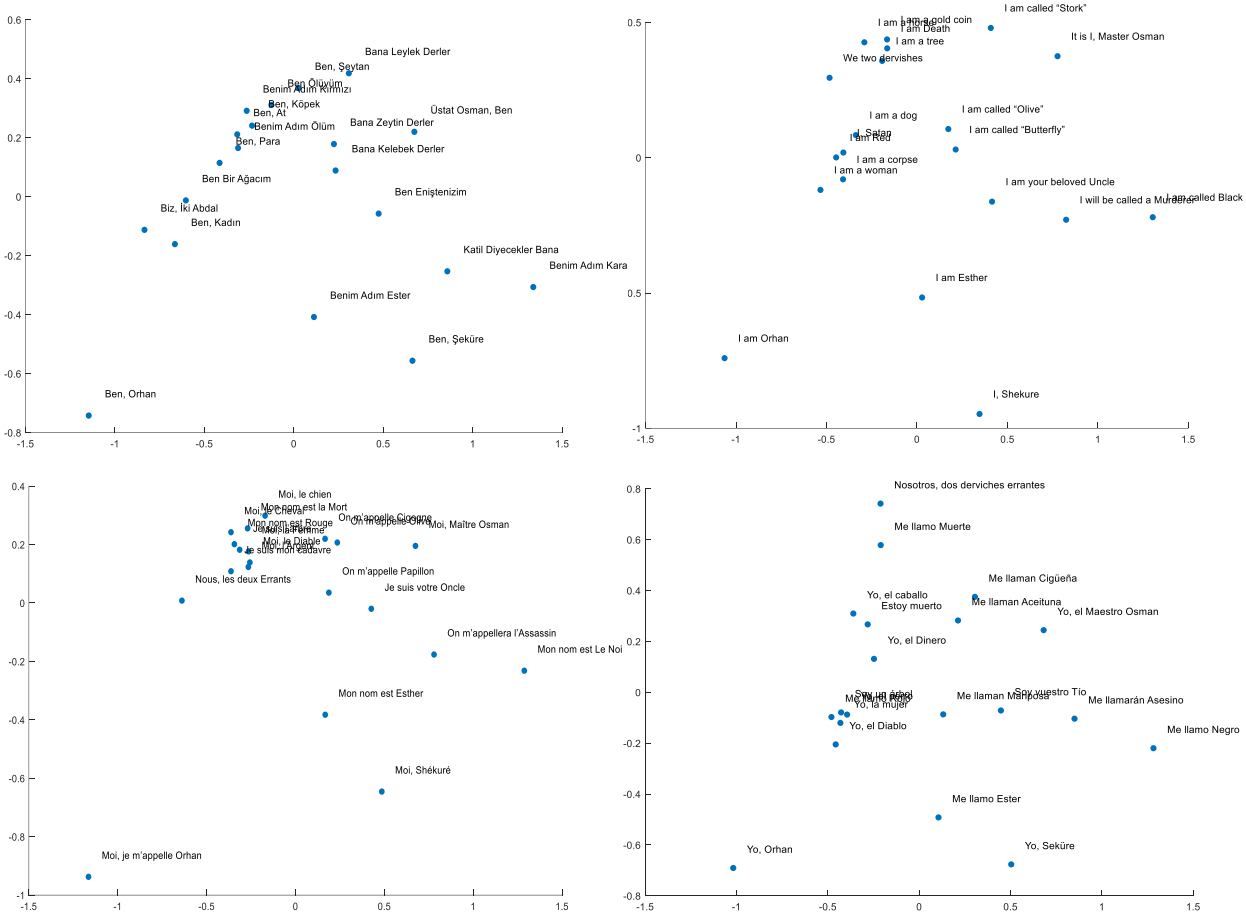
Makalede, listelenen sekiz sözcüksel özellikten ilk yedisi anlatım kolaylığı için *sözcüksel özellikler* olarak nitelendirilmiş, sonuncusu ise *en sık kullanılan kelime vektörü* olarak tanımlanmıştır.

Üslubun korunabildiği çevirilerde, karakterlerinin üsluplarının da korunacağı varsayımından yola çıkılarak, özgün metin ile hesaplanan karakterlerin öznitelikleri arasındaki uzaklıkların çevirilerde de aynı dağılımı gösteriyor olması, başka bir deyişle benzer olması beklenir. Bu durumun sayısal olarak test edilebilmesi için, özgün metin ve her bir çeviri için karakter vektörleri arasındaki uzaklıklar Öklid mesafesi olarak hesaplanmıştır. Bu uzaklıklar arasındaki korelasyon değerleri hesaplanarak özgün metin ve çeviriler arasındaki ilişki gözlemlenmiştir. Korelasyon hesaplaması için Kendall'ın Tau Katsayısı (Kendall's Rank Correlation Coefficient) kullanılmıştır. Rank correlation ya da sıralama korelasyonu, sıralanmış iki değişken arasındaki benzerlik derecesini ölçer ve bu ilişkinin önemini istatistiksel olarak değerlendirir. Kendall'ın Tau Katsayısı iki değişken arasındaki bağlantıyı değişkenlerin dağılımı ile ilgili bilgiye ihtiyaç duymadan ölçebildiği için seçilmiştir. Bu adımda PCA düzlemindeki uzaklıklar da kullanılabilir. Karakterler açısından sıralama aynı olacağından sonuç değişmeyecektir. Üslup benzerliğini ölçmek amacıyla önerdiğimiz bu yaklaşımı *Sıra Uyuşumu-tabanlı Benzerlik (SUB)* olarak adlandırıyoruz.

Deneysel Sonuçlar

Her bir karakter için öznitelik vektörünün çıkarılmasının ardından, karakterlerin üslup benzerliğini gözlemleyebilmek amacıyla sözcüksel özellikler ile oluşturulan vektörlere temel bileşenler analizi (principle component analysis, PCA) uygulanmıştır. Temel bileşenler analizi, büyük bir değişken kümesini kümedeki bilgilerin çoğunu içeren küçük bir kümeye indirgemek için kullanılabilecek bir boyut küçültme aracıdır (Jolliffe, 2002, s.1). PCA kullanılarak öznitelik vektörleri iki boyuta düşürülmüş ve karakterleri dağılım grafikleri Şekil 2'de olduğu gibi elde edilmiştir.

Şekil 2'de, 4 farklı dağılım grafiği görülmektedir. Grafiklerden karakterlerin üslup özellikleri dağılımının, özgün metin ve çeviriler için benzer olduğu gözlemlenebilir. Grafiklerdeki dağılımlarda karakterlerin kelime sayısının etkili olduğu gözlemi yapılabilir; bununla birlikte tek güçlü etken değildir. Örnek vermek gerekirse, Orhan karakteri ile İki Abdal'ın kelime sayısı (851, 755) birbirine yakın ve Ester ve Şeküre'nin kelime sayısı (8.723, 18.404), Orhan'dan daha fazla iken Orhan'ın İki Abdal'a olan uzaklığı ile Şeküre ve Ester'e olan uzaklığı çok farklı olmadığı söylenebilir. Bu sebeple, özgün metin ile çeviriler arasındaki benzerlik ilişkisine özelliklerin etkisini incelemek amacıyla farklı özellikler ile deneyler yapılmıştır.



Şekil 2. Sözcüksel özellikler ile oluşturulan PCA diyagramı (sırası ile Türkçe, İngilizce, Fransızca ve İspanyolca için)

İlk deney, bütün sözcüksel özellikler kullanılarak oluşturulan öznitelik vektörleri ile yapılmıştır. Deneyde çeviriler arasındaki üslup ilişkisi doğrudan gözlemlenmek istenmiştir. En sık kullanılan kelimeler, karakterler için etkili bir belirleyici olabileceği ve öteki üslup özelliklerinin etkilerini gizleyebileceği için öznitelik vektörlerine eklenmemiştir. Özgün metin ve çevirilerin karakterleri arasındaki korelasyon değerleri Tablo 6.a'da görülebilir. Makalede hesaplanan tüm korelasyon katsayıları için p-değerleri 0,001'de küçük çıktığı için tablo yapılmasına gerek duyulmamıştır. p-değeri hesaplanan korelasyon değerlerinin tesadüf olma olasılığını gösterir. Bu durumda küçük bir p-değeri, bu durumun tesadüf olma ihtimalinin düşük olduğunu kanıtlar niteliktedir. Hesaplanan korelasyonun iki değişken arasında gerçekten var olduğunu ve sonuçların istatistiksel olarak kayda değer olduğunu ifade eder.

Tablo 6 (a, b)

Sözcüksel özellikler ile hesaplanan (sol) ve sözcüksel özelliklerden kelime sayısı ve farklı kelime sayısı çıkarıldıktan sonra hesaplanan (sağ) SUB değerleri

	Türk.	İng.	Fran.	İspan.
Türk.	1	0,6493	0,7206	0,6345
İng.		1	0,7414	0,7011
Fran.			1	0,7141
İspan.				1

	Türk.	İng.	Fran.	İspan.
Türk.	1	0,5329	0,6292	0,5135
İng.		1	0,6545	0,5857
Fran.			1	0,6207
İspan.				1

Kelime sayısı ve farklı kelime sayısının özgün metin ve çevirilerde karakterler bağlamında benzer şekillerde değişeceği varsayımı ile korelasyonu artırması beklenir. Tablo 6.b'deki sonuçlar, kelime sayısı ve farklı kelime sayısının özniteliklerden çıkarılması ile korelasyon değişimini gözlemlemek için yapılan deneyindir. Tablo 6.a ile karşılaştırıldığında bütün korelasyon değerlerinde düşüş gözlenmiştir. Sonuçlar varsayımı doğrular niteliktedir.

Daha önce bahsedildiği gibi, en sık kullanılan kelimeler, karakterleri ayırmada etkili olabilirler. Bu durumu gözlemlemek için farklı k sayıları ile en sık kelimeler seçilerek deneyler yapılmıştır. Tablo 7.a'da *k*=10, Tablo 7.b'de *k*=20 ve Tablo 7.c'de *k*=30 değerleri ile hesaplanmış katsayı değerleri görülebilir. Deneylerin sonuçları, Tablo 6.a'daki sonuçlar ile karşılaştırıldığında, sözcüksel özelliklerin benzerliği yakalamakta genellikle daha iyi sonuçlar verdiğini görebiliriz. Daha ilginç olan ise en sık kelimeler vektörleri *k*= 20 değeri için özgün metin ile çeviriler arasındaki en iyi katsayı değerlerini verir. Az sayıda en sık kullanılan kelimelerin ayırt edici kelimeleri yakalayamaması ve çok sayıda olduklarında ayırt edici kelimelerin başka karakterlerde de görülmeye başlanması bu durumun sebebi olabilir.

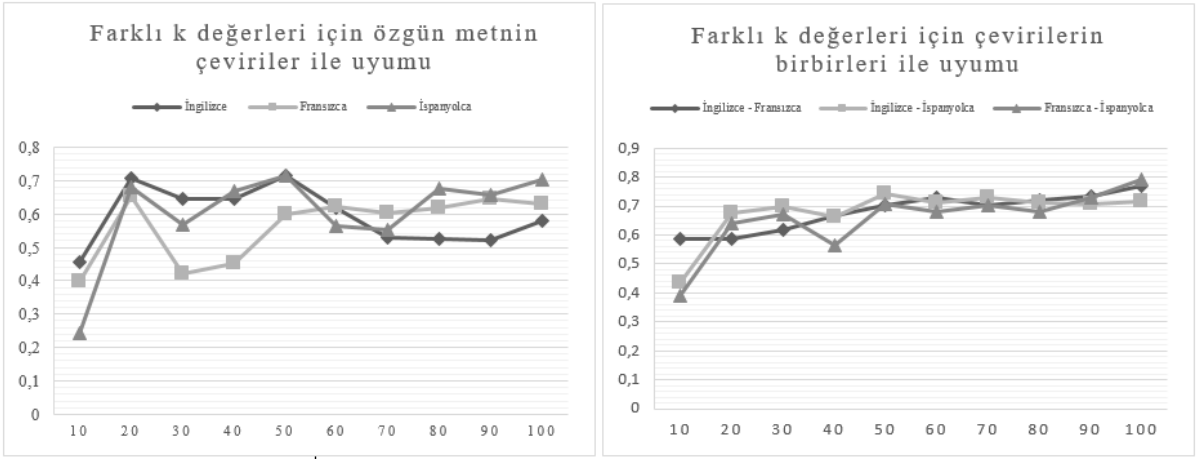
Tablo 7 (a, b, c)
En sık kelimeler vektörleri ile hesaplanan SUB değerleri k=10, k=20 ve k=30

k=10	Türk.	İng.	Fran.	İspan.
Türk.	1	0,4549	0,3995	0,2414
İng.		1	0,5853	0,4364
Fran.			1	0,3913
İspan.				1

k=20	Türk.	İng.	Fran.	İspan.
Türk.	1	0,6722	0,6536	0,6804
İng.		1	0,5884	0,6736
Fran.			1	0,6401
İspan.				1

k=30	Türk.	İng.	Fran.	İspan.
Türk.	1	0,6441	0,4213	0,5683
İng.		1	0,6159	0,6969
Fran.			1	0,6728
İspan.				1

Sık kullanılan kelimelerin etkisini ölçmek amacıyla, *k* sayısı arttıkça metinler arasındaki uyumun gözlemi için yapılan deneylerin sonuçları Şekil 3'deki grafiklerde görülebilir. Özgün metin ile çevirilerin uyumu *k* değerleri yükseldikçe, çevirilerin birbirleri arasındaki uyum bir süre sonra doyuma yaklaşır. İngilizce, Fransızca ve İspanyolcanın Hint-Avrupa dil ailesinin Avrupa kolundan ve Türkçenin farklı bir aile olan Ural- Altay dil ailesinden olması, bu durumun sebebi olabilir.



Şekil 3. Farklı k değerleri için Türkçe (sol) ve çeviri (sağ) metinler arasındaki SUB değeri eğrileri

Sözcüksel özellik ve en sık kelimeler vektörlerinin birleşimi ile yapılan deneyin sonuçları Tablo 8'de görülebilir. Bu deneyin amacı ayrı ayrı incelenen bu vektörlerin, birleştirildiğinde elde edilecek sonuçları gözlemlemektir. Tablo 8, Tablo 6.a ve Tablo 7.b ile karşılaştırıldığında, bu vektörlerin birleştirildiklerinde ayrı ayrı olduklarından daha iyi sonuçlar verdiği görülür.

Tablo 8
Sözcüksel özellikler ve en sık kelimeler vektörleri ile hesaplanan SUB değerleri (k=20)

	Türk.	İng.	Fran.	İspan.
Türk.	1	0,7069	0,6746	0,7053
İng.		1	0,6556	0,7194
Fran.			1	0,6882
İspan.				1

Katsayılar incelendiğinde, bütün deneyler için ortak olarak İngilizce, Fransızca ve İspanyolca çevirilerin uyumlarının genellikle Türkçe olan özgün metin ile olan uyumlarından yüksek olduğu görülebilir. Örneğin, Tablo 8'de özgün Türkçe metin ile çeviri İspanyolca metin arasındaki uyum katsayısı 0,7069 iken İngilizce çeviri ve İspanyolca çeviri arasındaki uyum katsayısı 0,7194'dir. Bu durum yine çeviri dillerinin aynı dil ailesinden gelmesinin bir sonucu olabilir. Türkçenin bu dillerden farklarını incelemek üzere farklı sayısal analizler yapılmıştır.

İlk analiz, karakterler için kelime sayıları oranlarının hesaplanmasıdır. Tablo 9'da satır başlarında yer alan dildeki metinlerin kelime sayıları sütun başlarında yer alan dildeki metnin kelime sayısına oranlanmış ve sonuç bu satır ile sütunun kesiştiği hücreye yazılmıştır. Bu oranlar her karakter için hesaplanmış ve ortalamaları alınarak sonuçlar Tablo 9.a'da verilmiştir. Sonuçlara göre İngilizce, Fransızca ve İspanyolcada aynı durumlar için daha fazla sözcük kullanıldığı yorumu yapılabilir. Türkçede zaman ya da özne sondan eklenerek ifade edilirken, çeviri dillerinde bunlar için farklı kelimeler kullanılır. Kısaca, bu durumun sebebi Türkçenin sondan eklemeli bir dil olması olarak açıklanabilir.

Tablo 9 (a, b)
(a) Karakter için kelime sayılarının ve (b) farklı kelime sayılarının oran ortalamaları

	Türk.	İng.	Fran.	İspan.
Türk.	1	0,6401	0,6033	0,6172
İng.		1	0,9437	0,9657
Fran.			1	1,0254
İspan.				1

	Türk.	İng.	Fran.	İspan.
Türk.	1	1,2228	1,0673	1,1725
İng.		1	0,8816	0,9674
Fran.			1	1,0997
İspan.				1

Karakterler için farklı kelime sayılarının oranları yine aynı şekilde farklı diller ile hesaplanarak, sonrasında oranların ortalaması alınmıştır. Tablo 9.b'den incelenebileceği gibi, toplam kelime sayısının tam aksine Türkçede aynı durumu anlatmak için kullanılan farklı kelime sayısı İngilizce, Fransızca ve İspanyolcaya göre daha fazladır. Bu durum yine Türkçenin eklemeli dil olması ve aynı kökten çok sayıda farklı kelime üretilmesiyle açıklanabilir. Türkçeden sonra Fransızca diğer dillere göre daha çeşitli kelime kullanımı göstermiştir.

Ortalama cümle uzunluğu oranları Tablo 9'da olduğu gibi fakat cümle uzunlukları ile hesaplanmış ve sonuçlar Tablo 10'da verilmiştir. Tablo incelenerek, Türkçede cümlelerin çeviri dillerinden daha kısa olduğu yorumu yapılabilir. Öte yandan çeviri dilleri arasındaki cümle uzunlukları çok farklı değildir. Türkçe - İngilizce için cümle uzunluğu oranı Patton ve Can'ın (2012) araştırmasındaki 0,666 oranı ile örtüşür.

Tablo 10
Ortalama cümle uzunluğu oranları

	Türk.	İng.	Fran.	İspan.
Türk.	1	0,6729	0,6534	0,6347
İng.		1	0,9722	0,9447
Fran.			1	0,9759
İspan.				1

Tartışma

Bu çalışmada *Benim Adım Kırmızı* romanı ve çevirileri arasındaki üslup uyumunun nicel olarak değerlendirilmesi amaçlanmıştır. Bunun için romandaki karakterlerin üslupları sayısal olarak incelenmiş, sözcüksel özellikler ve en sık geçen kelime vektörleri ile deneyler yapılmıştır. Sözcüksel özellikler ve en sık kelime vektörleri ile ayrı ayrı yapılan deneylerde özgün metnin çeviriler ile uyum katsayısı birbirine yakın çıkmış, bu vektörler birleştirilerek yapılan deneylerde daha iyi sonuçlar elde edilmiştir. Farklı özellikler roman karakteriyle ilgili yeni bilgiler taşıyacağından, bu beklenen bir sonuçtur.

En sık kelime vektörleri ile yapılan deneyler, k sayısı genişletilerek devam edilmiş, k sayısı arttıkça çevirilerin birbirleri arasındaki uyumun 0,7 - 0,8 değerleri arasında sabit kaldığı gözlenmiştir. Özgün Türkçe metnin çeviriler ile uyumunda k değerinin artmasıyla sabitleşme gözlemlenmemiştir. Bu durum Türkçenin çeviri dillerinden köken olarak farklı bir dil olması ile açıklanabilir.

Türkçe metin ve çeviriler arasındaki kelime sayıları oranlarına baktığımızda, çeviri dillerde kullanılan kelime sayısının daha fazla olduğu görülür. Türkçenin sondan eklemeli bir dil olduğu için bu sonuç beklenilenden farklı değildir. Metinlerde kullanılan farklı kelime oranları, Türkçede kullanılan farklı kelime sayısının çevirilere göre daha fazla olduğunu gösterir. Ayrıca çeviri dilleri bu konuda birbirlerine Türkçeye olduklarından daha benzerdir. Bu sonuçlar da Türkçenin sondan eklemeli bir dil olmasıyla açıklanabilir. Cümle uzunluğu oranlarında, Türkçenin çeviri dillerine oranları ve çeviri dillerinin kendi içlerindeki oranları birbirlerine yakındır. Çeviri dillerinin kökenlerinin benzerliği yine bu sonuçları açıklayabilir.

Sonuç

Bu makalede, stil analizini (stylometry) Türkçe metinler üzerine yapılmış araştırmalar bağlamında tanıtan kapsamlı bir kaynak sağlanmış ve Orhan Pamuk'un *Benim Adım Kırmızı* romanının çevirilerinin aslına olan sadakati sayısal yöntemlerle, önerdiğimiz yeni bir yaklaşımla, ölçülmüştür. Bu amaç doğrultusunda, üslup analizinin kullanım amaçları incelenmiş, değişik uygulama alanları örneklendirilmiştir. Üslup analizi uygulama adımları anlatılmış ve Türkçe metinler ile yapılmış çalışmalar incelenmiştir. Bu çalışmalar ayrıntılı bir tablo ile okuyuculara sunulmuş ve ayrıca sayısal üslup araştırmaları için hazır yazılımlar tanıtılmıştır.

Benim Adım Kırmızı çalışmamızda özgün metin ve çevirilerin üslup benzerliği araştırılmıştır. Çeviri dillerinin de üsluba etkisinin incelenmesi amacıyla, çeviriler farklı dillerden (Fransızca, İngilizce, İspanyolca) seçilmiştir. Sözcüksel özellikler ve en sık kelimeler vektörleri ile yapılan deneyler sonucunda çeviriler ile özgün metin arasındaki üslup benzerliğinin istatistiksel anlamda kayda değer olduğu doğrulanmıştır. Deneyler, çeviri dillerindeki metinlerin birbirleri ile olan uyumunun Türkçe ile olan uyumlarından daha yüksek olduğunu göstermiştir. Aynı dil ailesinden gelen bu üç dil için bu sonuçlar beklendiği gibidir.

Gelecekteki araştırma hedefi olarak, farklı dil ailelerinden çeviriler ile deneyler tekrarlanıp sonuçlar karşılaştırılabilir. Deneyler, farklı metinler, romanlar ve çevirileriyle tekrarlanabilir. Bu amaçla kaynak metnin tutarlı bölümlere ayrılması; kısa öykü, bir romanın içindeki ilişkili temalar gibi; gerekebilir.

Teşekkür

"İnternet yaşamdır" sloganıyla yaşayan özgürlük savunucusu Mustafa Akgül hocamızı hatırlamamızı sağlayan *Türk Kütüphaneciliği* dergisine; yapıcı eleştirileri için hakemlere ve dergi editörüne teşekkür ederiz.

Kaynakça

Abbasi, A., Chen, H. ve Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems. 26*(3), 1-34. Erişim adresi: https://doi.org/10.1145/1361684.1361685

Afroz, S., Çalışkan İslam, A., Stolerman, A., Greenstadt, R. ve McCoy, D. (2014). Doppelgänger finder: Taking stylometry to the underground. İçinde *Proceedings - IEEE Symposium on Security and Privacy*, 212-226. Erişim adresi: https://doi.org/10.1109/SP.2014.21

Agün, H. V., Yılmazel, S. ve Yılmazel, O. (2017). Effects of language processing in Turkish authorship attribution. *2017 IEEE International Conference on Big Data (Big Data)*, (1), 1876-1881. Erişim adresi: https://doi.org/10.1109/BigData.2017.8258132

Akın, A. A. ve Akın, M. D. (2007). Zemberek, an open source NLP framework for Turkic languages. *Structure, 10*, 1-5. Erişim adresi: https://doi.org/10.1.1.556.69

Altıntaş, K., Can, F. ve Patton, J. M. (2007). Language change quantification using time-separated parallel translations. *Literary and Linguistic Computing*, *22*(4), 375-393. Erişim adresi: https://doi.org/10.1093/llc/fqm026

Amasyalı, M. F. ve Diri, B. (2006). Automatic Turkish text categorization in terms of author, genre and gender. *Natural Language Processing and Information Systems, Proceedings*, *3999*, 221-226. Erişim adresi: https://doi.org/10.1007/11765448_22

Aslantürk, O. (2014). *Tamgacı: artırımsal ve geri beslemeli Türkçe yazar çözümleme* (Doktora tezi). Hacettepe Üniversitesi, Bilgisayar Mühendisliği Bölümü.

Aslantürk, O., Sezer, E. A., Sever, H. ve Raghavan, V. (2010). Application of cascading rough set-based classifiers on authorship attribution. *Proceedings - 2010 IEEE International Conference on Granular Computing, GrC 2010*, 656-660. Erişim adresi: https://doi.org/10.1109/GrC.2010.110

Atay, O. (2001). *Bir Bilim Adamının Romanı Mustafa İnan*. İstanbul: İletişim Yayınları.

Baker, M. (2000). Towards a methodology for investigating the style of a literary translator. *Target, 12*(2), 241-266

Bay, Y. ve Çelebi, E. (2016). Feature selection for enhanced author identification of Turkish text. İçinde *The 30th International Symposium on Computer and Information Sciences, 2015*, 371-379. Erişim adresi: https://doi.org/10.1007/978-3-319-22635-4_34

Bergsma, S., Post, M. ve Yarowsky, D. (2012). Stylometric analysis of scientific articles. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 327-337. Erişim adresi: www.clsp.jhu.edu/

Bozkurt, İ. N., Bağlıoglu, Ö. ve Uyar, E. (2012). Authorship attribution. *22nd International International Symposium on Computer and Information Sciences*, 1-5. Erişim adresi: https://doi.org/10.1109/ISCIS.2007.4456854

Burrows, J. (2002). "Delta": A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, *17*(3), 267-287. Erişim adresi: https://doi.org/10.1093/llc/17.3.267

Can, E. F. Can, F. Şahin, P. D. ve Kalpaklı, M. (2013). Automatic categorization of Ottoman poems. *Glottotheory International Journal of Theoretical Linguistics*, *4*(2), 40-57. Erişim adresi: https://doi.org/10.1524/glot.2013.0014

Can, F. (2018). İnce Memedlerin sayılarında gizlenenler ve Türkçenin değişimi üzerine. Yazarından alınmış, yayımlanacak metin.

Can, F., Can, E. F. ve Karbeyaz, C. (2011). Translation relationship quantification: A cluster-based approach and its application to Shakespeare's sonnets. *Proceedings of the 25th International Symposium on Computer and Information Sciences* (ISCIS'10) içinde, London, UK, September 22-24, 2010. *Lecture Notes in Electrical Engineering* 62, DOI 10.1007/978-90-481-9794-1_25, pp. 117-120

Can, F. ve Patton, J. M. (2004). Change of writing style with time. *Computers and the Humanities*, *38*(1), 61-82. Erişim adresi: https://doi.org/10.1023/B:CHUM.0000009225.28847.77

Can, F. ve Patton, J. M. (2010). Change of word characteristics in 20th-century Turkish literature: A statistical analysis. *Journal of Quantitative Linguistics*. Erişim adresi: https://doi.org/10.1080/09296174.2010.485444

Canbay, P., Sever, H. ve Sezer, E. A. (2018). Determining of discriminative blog size for authorship attribution on the Turkish texts. *2018 6th International Symposium on Digital Forensic and Security (ISDFS)*, (May), 1-5. Erişim adresi: https://doi.org/10.1109/ISDFS.2018.8355373

Canbay, P., Sezer, E. A. ve Sever, H. (2018). Authorship modelling approach for authorship verification on the Turkish texts. *2018 26th Signal Processing and Communications Applications Conference (SIU), İzmir*, 1-4. Erişim adresi: 10.1109/SIU.2018.8404436

Çakır, M. U. ve Güldamlasıoğlu, S. (2016). Text mining analysis in Turkish language using big data tools. *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, 614-618. Erişim adresi: https://doi.org/10.1109/COMPSAC.2016.203

Çöltekin, C. (2014). A set of open source tools for Turkish natural language processing. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1079-1086.

Dalkılıç, G., ve Çebi, Y. (2003). Türkçe külliyat oluşturulması ve Türkçe metinlerde kullanılan kelimelerin uzunluk dağılımlarının belirlenmesi. *DEÜ Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, *5*(1), 1-7.

Demirci, S. (2014). *Emotion analysis on Turkish tweets* (Yüksek lisans tezi). Ortadoğu Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü. Erişim adresi: http://etd.lib.metu.edu.tr/upload/12618821/index.pdf

de Morgan, S. E. (1882). *Memoir of Augustus de Morgan by his wife Sophia Elizabeth de Morgan with selections from his letters*. London: Longmans, Green, and Co.,

Diederich, J., Kindermann, J., Leopold, E. ve Paass, G. (2003). Authorship attribution with support vector machines. *Applied Intelligence*, *19*(1/2), 109-123. Erişim adresi: https://doi.org/10.1023/A:1023824908771

Diri, B. ve Amasyalı, M. F. (2003). Automatic author detection for Turkish texts. *Artificial Neural Networks and Neural Information*, 1. Erişim adresi: http://www.yildiz.edu.tr/~diri/ICANN.pdf

Eder, M., Rybicki, J. ve Kestemont, M. (2016). Stylometry with R: A package for computational text analysis. *R Journal*, 8(1): 107-21. Erişim adresi: https://journal.r-project.org/archive/2016/RJ-2016-007/index.html

Eryiğit, G. (2014). İTÜ Turkish NLP web service. İçinde *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Gothenburg, Sweden: Association for Computational Linguistics. Erişim adresi: http://web.itu.edu.tr/gulsenc/papers/itunlp.pdf

Mosteller, F. ve Wallace, D. (1964). Inference and disputed authorship: *The Federalist*. Reading, Mass: Addison-Wesley.

Gonçalves, T. ve Quaresma, P. (2007). Evaluating preprocessing techniques in a text classification problem. *São Leopoldo, RS, Brasil: SBC-Sociedade Brasileira de Computação*, 841-850. Erişim adresi: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.131.1271verep=rep1vetype=pdf

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, *22*(3), 251-270. Erişim adresi: https://doi.org/10.1093/llc/fqm020

Guyon, I. ve Elisseeff, A. (2006). Feature extraction, foundations and applications: An introduction to feature extraction. *Studies in Fuzziness and Soft Computing*, *207*, 1-25. Erişim adresi: https://doi.org/10.1007/978-3-540-35488-8_1

El-Fiqi, H., Petraki, E. ve Abbass, H. A. (2016). Pairwise comparative classification for translator stylometric analysis. *ACM Trans. Asian Low-Resour. Lang. Inf. Process. 16(1)*, 1-26. Erişim adresi: https://doi.org/10.1145/2898997

Holmes, D. I. (1992). A stylometric analysis of Mormon scripture and related texts. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 155*(1), 91-120.

Holmes, D. I. (1997). Stylometry, its origins, development and aspirations. İçinde *ACH-ALLC '97 Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing. Kingston, Ontario, Canada, June 37, 1997.*

Holmes, D. I. (1998). The evolution of sylometry in humanities scholarship. *Literary and Linguistic Computing*, *13*(3), 111-117. Erişim adresi: https://doi.org/10.1093/llc/13.3.111

Hurtado, J., Taweewitchakreeya, N. ve Zhu, X. (2014). Who wrote this paper? Learning for authorship de-identification using stylometric featuress. *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration, IEEE IRI 2014*, 859-862. Erişim adresi: https://doi.org/10.1109/IRI.2014.7051981

IBM Corp. Released. (2017). IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.

İşi, A., Çemrek, F. ve Yıldız, Z. (2013). İstatistikten edebiyata bir köprü: Stilometri analizi. *Uşak Üniversitesi Sosyal Bilimler Dergisi*, *6*(3), 271-271. Erişim adresi: https://doi.org/10.12780/UUSBD170

İnan, M. (1987). Dil ve Matematik. *Prof. Dr. Mustafa İnan: Konferansları, makaleleri ve konuşmaları.* İstanbul: İTÜ.

Jolliffe, I. T. (2002). *Principal component analysis*. Springer.

Juola, P. (2008). Author attribution, foundations and trends. *Information Retrieval*, 1(3), 233-334.

Jovic, A., Brkic, K. ve Bogunovic, N. (2015). A review of feature selection methods with applications. İçinde *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (ss. 1200-1205). IEEE. Erişim adresi: https://doi.org/10.1109/MIPRO.2015.7160458

Kacmarcik, G. ve Gamon, M. (2006). Obfuscating document stylometry to preserve author anonymity. *Proceedings of the COLING/ACL on conference poster sessions* -, (July), 444-451. Erişim adresi: https://doi.org/10.3115/1273073.1273131

Karbeyaz, C. (2011). *A cluster-based external plagiarism and parallel corpora detection method.* (Yüksek lisans tezi). Bilkent Üniversitesi, Bilgisayar Mühendisliği Bölümü.

Koppel, M., Schler, J. ve Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, *60*(1), 9-26. Erişim adresi: https://doi.org/10.1002/ASI.V60:1

Koppel, M. ve Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the American Society for Information Science and Technology*, *65*(1), 178-187. Erişim adresi: https://doi.org/10.1002/asi.22954

Küçükyılmaz, T., Cambazoğlu, B. B., Aykanat, C. ve Can, F. (2008). Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing and Management*, *44*(4), 1448-1466. Erişim adresi: https://doi.org/10.1016/J.IPM.2007.12.009

Meyer, S. ve Stein, B. (2006). LNCS 3936 - Intrinsic plagiarism detection. *Advances*, 565-569. Erişim adresi: https://doi.org/10.1007/11735106

Moretti, F. (2013). *Distant Reading.* London: Verso.

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y. ve Woodard, D. (2017). Surveying stylometry techniques and applications. *ACM Computing Surveys*, *50*(6), 1-36. Erişim adresi: https://doi.org/10.1145/3132039

Nguyen, D. P. (2014). Obfuscation techniques for Java source code. Erişim adresi: https://dr.ntu.edu.sg/handle/10220/26030

Oakes, M. P. (2009). Corpus linguistics and stylometry. İçinde *Corpus Linguistics: An International Handbook*. Erişim adresi: https://doi.org/10.1515/9783110213881.2.1070

Oflazer, K. (2014). Turkish and its challenges for language processing. *Language Resources and Evaluation*, *48*(4), 639-653. Erişim adresi: https://doi.org/10.1007/s10579-014-9267-2

Pamuk, O. (1998). *Benim Adım Kırmızı*. İstanbul: İletişim Yayınları.

Pamuk, O. (2002). *My Name is Red* (E. M. Göknar, Çeviri). New York : Vintage International.

Pamuk, O. (2002). *Mon nom est Rouge* (G. Authier, Çeviri). Gallimard.

Pamuk, O. (2006). *Me Ilamo Rojo* (R. Carpintero, Çeviri). Buenos Aires: Alfaguara.

Patton, J. M. ve Can. F. (2012). Determining translation invariant characteristics of James Joyce's Dubliners. *Quantitative Methods in CorpusBased Translation Studies* (M. P. Oakes and M. Ji Ed.) içinde. Amsterdam and Philadelphia: John Benjamin Publishing Company

Patton, J. M. ve Can, F. (2004). A stylometric analysis of Yaşar Kemal's *İnce Memed* tetralogy. *Computers and Humanities, 38(4),* 457-467. Erişim adresi: https://doi.org/10.1007/s10579-004-1906-6

Peersman, C., Daelemans, W. ve Vaerenbergh, L. Van. (2011). Predicting age and gender in online social networks. *International Conference on Information and Knowledge Management, Proceedings*, 37-44. Erişim adresi: https://doi.org/10.1145/2065023.2065035

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Erişim adresi: http://www.R-project.org/.

Reddy, T. R., Vardhan, B. V. ve Reddy, P. V. (2016). A survey on authorship profiling techniques. *International Journal of Applied Engineering Research, 11*(5), 3092-3102.

Saygılı, N. Ş., Amghar, T., Levrat, B. ve Acarman, T. Taking advantage of Turkish characteristic features to achieve authorship attribution problems for Turkish. *2017 25th Signal Processing and Communications Applications Conference (SIU), Antalya*, 1-4. Erişim adresi: 10.1109/SIU.2017.7960438

Schulz, K. (2011). The mechanic muse - what is distant reading? - *The New York Times*. Erişim adresi: https://www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html

Sokolova, M. ve Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*, 427-437. Erişim adresi: https://doi.org/10.1016/j.ipm.2009.03.002

Srividhya, V. ve Anitha, R. (2010). Evaluating preprocessing techniques in text categorization. *International Journal of Computer Science and Application*, *47*(11), 49-51.

Stamatatos, E. (2007). Author identification using imbalanced and limited training texts. *18th International Conference on Database and Expert Systems Applications (DEXA 2007)* içinde (ss. 237-241). IEEE. Erişim adresi: https://doi.org/10.1109/DEXA.2007.5

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, *60*(3), 538-556. Erişim adresi: https://doi.org/10.1002/asi.21001

Taş, T. ve Görür, A. K. (2007). Author identification for Turkish texts. *Çankaya Üniversitesi Fen-Edebiyat Fakültesi, 7*, 151-161. Erişim adresi: http://www.arastirmax.com/system/files/dergiler/64470/makaleler/7/1/arastirmax_11081_pp_151-161.pdf

Taşçı, H. ve Ekinci, E. (2012). Character level authorship attribution for Turkish text documents. *The Online Journal of Science and Technology*, *2*(3), 12-16.

Tennyson, M. F. (2013). A replicated comparative study of source code authorship attribution. *Proceedings - 2013 3rd International Workshop on Replication in Empirical Software Engineering Research, RESER 2013*, içinde (ss. 76-83). Erişim adresi: https://doi.org/10.1109/RESER.2013.12

Toraman, C., Can, F, ve Koçberber, S. (2011). Developing a text categorization template for Turkish news portals. *INISTA 2011 - 2011 International Symposium on INnovations in Intelligent SysTems and Applications*, 379-383. Erişim adresi: https://doi.org/10.1109/INISTA.2011.5946096

Torunoğlu, D., Çakırman, E., Ganiz, M. C., Akyokuş, S. ve Gürbüz, M. Z. (2011). Analysis of preprocessing methods on classification of Turkish texts. *INISTA 2011 - 2011 International Symposium on INnovations in Intelligent SysTems and Applications* içinde (ss. 112-117). Erişim adresi: https://doi.org/10.1109/INISTA.2011.5946084

Tunalı, V. ve Bilgin, T. (2012). PRETO: A high-performance text mining tool for preprocessing Turkish texts. *Proceedings of the 13th International Conference,* 134-140. Erişim adresi: https://doi.org/10.1145/2383276.2383297

Tunalı, V. ve Bilgin, T. T. (2012). Examining the impact of stemming on clustering Turkish texts. *INISTA 2012 - International Symposium on Innovations in Intelligent SysTems and Applications* içinde (ss. 2-5). Erişim adresi: https://doi.org/10.1109/INISTA.2012.6246966

Türkoğlu, F. (2006). *Melez yaklaşımlarla Türkçe dokümanlarda yazar tanıma*. (Yüksek lisans tezi).Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü.

Türkoğlu, F., Diri, B. ve Amasyalı, M. F. (2007). Author attribution of Turkish texts by feature mining. *Advanced Intelligent Computing Theories,* 1086-1093. Erişim adresi: https://doi.org/10.1007/978-3-540-74171-8_110

Tweedie, F. J., Singh, S. ve Holmes, D. I. (1996). Neural network applications in stylometry: *The Federalist Papers*. *Language Resources and Evaluation*, *30*(1), 1-10. Erişim adresi: https://doi.org/10.1007/BF00054024

Verhoeven, B., Company, J. S. ve Daelemans, W. (2014). Evaluating content-independent features for personality recognition. *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition - WCPR '14*, 7-10. Erişim adresi: https://doi.org/10.1145/2659522.2659527

Verhoeven, B., Skrjanec, I. ve Pollak, S. (2017). Gender profiling for Slovene twitter communication: The influence of gender marking, content and style. *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing* içinde (ss. 119-125). Valencia, Spain. Erişim adresi: https://doi.org/10.18653/v1/W17-1418

Yavanoğlu, O. (2016). Intelligent authorship identification with using Turkish newspapers metadata. İçinde *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016* , 1895-1900. https://doi.org/10.1109/BigData.2016.7840809

Zheng, R., Li, J., Chen, H. ve Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology, 57*(3), 378-393. Erişim adresi: https://doi.org/10.1002/asi.20316

Zheng, R., Qin, Y., Huang, Z. ve Chen, H. (2003). Authorship analysis in cybercrime investigation. *Lecture Notes in Computer Science 2665*, 59-73. Erişim adresi: https://doi.org/10.1007/3-540-44853-5_5

Summary

In this article, we introduce stylometry, an important problem of digital humanities, and provide a survey of Turkish studies on this topic. We also present a novel study on quantification of translation loyalty by using a work of Orhan Pamuk: *My Name is Red*.

Stylometry is used for authorship attribution, author verification, author profiling, date attribution, and other related problems. Studies conducted on Turkish texts generally aim to identify the author of a new text by using newspaper columns of different authors (Diri and Amasyalı, 2003; Şirin, Amghar, Levrat and Acarman, 2017; Taş and Görür, 2007; Taşçı and Ekinci, 2012). Moreover, by using the texts of the columnists who wrote in various areas, a new text in a particular area was aimed to be attributed (Aslantürk, Sezer, Sever and Raghavan, 2010; Yavanoglu, 2016). For author verification purposes, Canbay, Sezer and Sever (2018) conducted a study looking for evidence that two Turkish texts are written by the same author. It is known that the use of language may vary according to age or gender of authors (Küçükyılmaz, Cambazoğlu, Aykanat and Can, 2008; Peersman, Daelemans and Vaerenbergh, 2011). Author profiling aims to narrow the search space by determining the author's gender, age, and other demographic characteristics. Date attribution tries to find the date on which a text is written. It can also be used to examine language and writer style change with time. Can and Patton (2004, 2010) studied this problem on Turkish novels and newspaper columns.

Stylometric analysis involves the following steps: Pre-processing, feature extraction, classification, and performance assessment. The analysis is possible by converting a text into some numeric features or attributes. Converting all properties of a text into numeric attributes can increase the computation time; and reduce the classification accuracy. The use of right features usually decreases the number of attributes and computation time. Therefore, a text is subject to some pre-processing prior to obtaining the attributes. Torunoğlu, Çakırman, Ganiz, Akyokuş and Gürbüz (2011) tried to classify Turkish texts by removing stopwords and using roots of words. They reported that the effect of these two pre-processing methods on the classification performance was negligible.

Various features are used by researchers, and they are grouped in different ways. Most common groupings can be found as lexical features, character-based features and syntactic features. Many other domain-specific features and their groupings are proposed (Neal et.al, 2017). When selecting attributes, care should be taken to avoid loss of information. On the other hand, every feature of the text may not always be useful. To evaluate the relevancy, measures like information gain, gain ratio, symmetrical uncertainty, correlation are used (Jori, Brkic and Bogunovic, 2015). In the studies conducted on Turkish, feature selection did not attract much attention. Bay and Çelebi (2016), used chi-square method in their study to reduce the number of attributes from 20 to 17, then observed that the performance is increased. Türkoğlu, Diri and Amasyalı (2007) conducted experiments using a large number of lexical and syntactic features. From more than 2,000 attributes, they selected several sets using the correlation-based feature selection method and observed that the success of the attribution was higher.

Classification approaches are quite diverse. The papers analyzed in our survey use machine learning-based, distance-based, probabilistic and statistical methods. Many distance-based, and probabilistic approaches can be considered as machine learning methods as well. After selecting the classifier, results should be compared with proper baselines. We present a comprehensive survey of related studies in Table 1 for researchers who want to work on Turkish texts.

Table 1
Stylometry studies on Turkish

Reference	Subtask	Data	Features	Method	Performance
Agün, H. V., Yılmazel, S. and Yılmazel, O. (2017)	Author attribution	Newspaper columns with at least 1.000 characters, 60 from each author	Lexical and syntactic features	Machine learning (LR, MNB and multi-layer NN)	F-score 0.37 - 0.95 (10-fold cross validation)
Altıntaş, K., Can, F. and Patton, J. M. (2007)	Language change quantification	From 4 author, translations of 7 different novels	Lexical and syntactic features	Statistical methods (ANOVA, DA, LR, odds ratio)	p-values less than 0,05 in ANOVA, 80% precision for discriminant analysis
Amasyalı, M. F., and Diri, B. (2006)	Author attribution, author profiling, categorizing texts	From 4 female, 14 male author 35 newspaper columns for each, about politics, sport and general knowledge	Lexical features (2-grams and 3-grams)	Machine learning (NB, SVM, C4.5 tree, random forest)	Precision between 59 - 83% for author verification, 79 - 93% for categorization by type, 83- 96% for categorization by gender (5-fold cross validation)
Aslantürk, O., Sezer, E. A., Seandr, H., and Raghavan, V. (2010)	Author attribution	From 9 authors, total of 513 newspaper columns about politics and life	Lexical and syntactic features	Machine learning (rough set-based classifier)	70% precision
Aslantürk, O. (2014)	Author attribution	12.115 newspaper columns from 8 authors about life and politics	Lexical and syntactic features	Machine learning (rough set-based classifier)	Among 1134 experiments precision over 70% for 498
Bay, Y., and Çelebi, E. (2016)	Author attribution	From 17 authors, 850 newspaper columns in total	Lexical features	Machine learning (NB, SVM, decision tree) and distance (KNN)	Precision between 96-100% (10-fold cross validation)
Bozkurt, I. N., Bağlıoğlu, O., and Uyar, E. (2012)	Author attribution	From 18 author, 500 newspaper columns for each	Lexical and syntactic features, frequencies of functional words	Machine learning (Histogram method, KNN, Bayes classifier, KM, combination of these and SVM)	Highest 95.7% precision with SVM (10-fold cross validation)
Can, E. F., Can, F., Duygulu, P., and Kalpaklı, M. (2011)	Author attribution, date attribution	From centuries between 15-19, Ottoman poetry of 10 different poets	Lexical features	Machine learning (SVM and NB)	Precision 93% highest for author attribution, 95% precision highest for date attribution (cross validation)
Can, F., Can, E. F., and Karbeyaz, C. (2010)	Translation relationship quantification	Shakespeare's sonnets and translations to Turkish	Frequently used phrases	distance (KM and Yao's formula) and Statistical methods	Low p-values indicating that similarity is greater than random similarity (<0,05)
Can, F., and Patton, J. M. (2010)	Language change quantification, date attribution, author profiling	From 40 different author, 40 novels written in different decades	Lexical features	Statistical methods (PCA, DA, linear regression)	Precision 94.1% for categorization by gender, 57.27% for categorization by date, low p-values showing that word lengths increase by time (cross validation)
Can, F., and Patton, J. M. (2004)	Writing style change quantification	New and old works of Çetin Altan and Yaşar Kemal	Lexical features	Statistical methods (t-test, LR, DA, regression analysis)	Low p-values showing that writing style changes over time (cross validation)
Canbay, P., Sezer, E. A. and Seandr, H. (2018)	Author verification	From 12 different author, 100 newspaper columns for each	Lexical and syntactic features	Distance (Cosines distance)	92% precision highest

Canbay, P., Seandr, H., and Sezer, E. A. (2018)	Author attribution	From 10 different author, 50 blog post for each	Lexical features (punctuation marks and bag-of-words)	Machine learning (SVM and ANN)	average precision between 25-75% (10 - fold cross validation)
Dalkılıç, G., and Çebi, Y. (2003)	Average word length calculation for Turkish	From different websites with different subjects, texts representing both written and spoken language	Lengths of the words	Statistical methods (Probability calculations)	The average word length was determined to be 6241 letters and it was seen that the words up to 7 letters constituted 69.11% of the total corpus.
Demirci, S. (2014)	Sentiment analysis	6000 tweets in total	Lexical features, 1-grams, 2-grams and 3-grams, digital emojis	Machine learning (NB, SVM) and distance (KNN)	70% precision highest
Diri, B., and Amasyalı, M. F. (2003)	Author attribution	From 18 different author, 20 text for each	Lexical and syntactic features	Score based method	84% precision highest
Karbeyaz, C. (2011)	Plagiarism detection	PAN'09 plagiarism dataset, Layla and Majnun	Bag-of-words	Distance (cover coefficient-based clustering method)	30% precision highest
Küçükyılmaz, T., Cambazoğlu, B. B., Aykanat, C., and Can, F. (2008)	Author attribution, author profiling, text categorization	Chat messages	Lexical features, character level features, digital emojis	Machine learning (Patient Rule Induction Method, SVM, NB) and distance (KNN)	Precision between 100 - 97% for author attribution, 91 - 67% for author's' domain prediction, 81 - 71% for gender prediction, 68 - 29% for authors' school prediction, 71 - 41% for prediction of the period of the day (10 - fold cross validation)
Patton, J. M., and Can, F. (2014)	Author profiling, writing style change quantification	*İnce Memed* tetralogy	Lexical and syntactic features	Statistical methods (ANOVA, Multi variable ANOVA, DA)	Low p-values showing that there is style difference between books, 87% precision for categorization by books (cross validation)
Patton, J. M., and Can, F. (2012)	Determining translation invariant characteristics	*Dubliners* stories of James Joyce and translations to Turkish	Lexical features	Statistical methods and DA	100% precision in categorizing stories as Turkish and English (cross validation)
Saygılı, Ş. N., Amghar, T., Levrat, B. and Acarman, T. (2017)	Author attribution	From 9 different authors, 50 newspaper columns for each, from 7 authors 250 columns for each	Frequencies of verbal nouns, verbal adjectives and verbal adverbs	Machine learning (SVM)	Precision 78% for first dataset, 63% for second dataset, F1 value 0.78 for first dataset and 0.61 for second dataset
Taş, T., and Görür, A. K. (2007)	Author attribution	From 20 authors, 25 newspaper columns for each	Lexical and syntactic features, vocabulary richness	Machine learning (Bayes net, naive Bayes, naive Bayes multinomial, naive Bayes updateable logistic, multilayer perceptron, RBF network, simple logistic, SMO)	After feature selection 80-57% precision (10-fold cross validation)
Taşçı, H. and Ekinci, E. (2012)	Author attribution	From 10 different author, 10 newspaper columns for each	Character level features and functional words	Distance (Cosines distance)	average precision 86% with character level features, 53% with functional words
Toraman, C., Can, F. and Koçberber, S. (2011)	Text categorization	News from the online portal of Bilkent University	Bag-of-words	Machine learning (C4.5, NB, SVM) and distance (KNN)	Precision 83% and 87.5% highest

Torunoğlu, D., Çakırman, E., Ganiz, M. C., Akyokuş, S. and Gürbüz, M. Z. (2011)	Text categorization	2230 texts from the newspapers in the categories cafe, world, Aegean, economics, daily, politics, sport, Turkey, life	Bag-of-words	Machine learning (NB, MNB, SVM) and distance (KNN)	For the experiments in which at least 50% of data was used precision over 70%
Tunalı, V. and Bilgin, T. T. (2012)	Text clustering	News in the categories economics, politics, sport, science, world, art, health, Turkey, life and green news	Bag-of-words	Distance (Spherical K-means)	When k = 20 purity 0.966, entropy 0.066 and 0.587 normalized mutual information
Türkoğlu, F. (2006)	Author attribution	From 18 authors, 35 text from each, 630 texts in total	Lexical and syntactic features, n-grams, functional words	Machine learning (NB, SVM, random forest, multi-layer perceptron and self-organizing map) and distance (KNN)	Precision with different dataset combinations 82.1%, 85% and 89.2% highest
Türkoğlu, F., Diri, B., and Amasyalı, M. F. (2007)	Author attribution	From 18 different author, 35 newspaper columns from each	Many lexical and syntactic features, n-grams, functional words	Machine learning (NB, SVM, random forest and multi-layer perceptron) and distance (KNN)	With SVM average precision 88.9%
Yavanoğlu, O. (2016)	Author attribution	From 9 authors, newspaper columns more than 20000 int he categories economics, life and politics	Lexical and syntactic features	Machine learning (ANN)	Precision for economy 98%, politics 97%, life 81% and cross categories 80% (10-fold cross validation)

Quantification of Loyalty for Translations of *My Name is Red*

In the translation texts, it is aimed to express the original text in another language without changing the meaning of it. So, can the style be preserved while preserving the meaning? Studies are available in the literature which examine the translation texts in terms of style loyalty. Can et al., aims to examine the style relation between Shakespeare sonets and the translations to Turkish numerically (2011). Patton et al, tries to determine the unchanging style features of James Joyce's *Dubliners* stories and the translation to Turkish (2012). Baker examines translations of the same text by using stylistic analysis and looks for traces of different translators (2000).

In this study, it is aimed to evaluate the loyalty of translations of Orhan Pamuk's novel *My Name is Red* to English, French, and Spanish. For this purpose, different chapters in the novel are used, where each chapter is from the voice of a different novel character. The hypothesis was the styles of characters will be preserved to some point in the translations if the style has not changed consciously by translator. To test this hypothesis; firstly, attribute vectors of characters were created for original text and target language translations, and the distances between original and translation texts were calculated by these vectors.

Method

My Name is Red is composed of 59 chapters and each chapter is written in the voice of one of 20 different characters in the novel. The list of characters for original text and translations can be seen in Table 2. The order of the characters in the table is the same as the order as they first appeared in the novel. Translations were made from Turkish to English by Erdağ M. Göknar, to French by Gilles Authier and to Spanish by Rafael Carpintero.

The voices of characters are divided into chapters in the novel, allowing us to examine the style of characters separately. In order to evaluate the stylistic similarity between original

text and translations in numerical terms, firstly the chapters of the same characters are combined. With this merge, it is aimed to create the widest dataset for characters.

Table 2
The Character Table

Turkish	English	French	Spanish
Ben Ölüyüm	I am a corpse	Je suis mon cadavre	Estoy muerto
Benim Adım Kara	I am called Black	Mon nom est Le Noir	Me llamo Negro
Ben, Köpek	I am a dog	Moi, le chien	Yo, el perro
Katil Diyecekler Bana	I wıll be called a Murderer	On m'appellera l'Assassin	Me llamarán Asesino
Ben Enişte nizim	I am your beloved Uncle	Je suis votre Oncle	Soy vuestro Tío
Ben, Orhan	I am Orhan	Moi, je m'appelle Orhan	Yo, Orhan
Benim Adım Ester	I am Esther	Mon nom est Esther	Me llamo Ester
Ben, Şeküre	I, Shekure	Moi, Shékuré	Yo, Seküre
Ben Bir Ağacım	I am a tree	Je suis l'arbre	Soy un árbol
Bana Kelebek Derler	I am called "Butterfly"	On m'appelle Papillon	Me llaman Mariposa
Bana Leylek Derler	I am called "Stork"	On m'appelle Cigogne	Me llaman Cigüeña
Bana Zeytin Derler	I am called "Olive"	On m'appelle Olive	Me llaman Aceituna
Ben, Para	I am a gold coin	Moi, l'Argent	Yo, el Dinero
Benim Adım Ölüm	I am Death	Mon nom est la Mort	Me llamo Muerte
Benim Adım Kırmızı	I am Red	Mon nom est Rouge	Me llamo Rojo
Ben, At	I am a horse	Moi, le Cheval	Yo, el caballo
Üstat Osman, Ben	It is I, Master Osman	Moi, Maître Osman	Yo, el Maestro Osman
Ben, Şeytan	I, Satan	Moi, le Diable	Yo, el Diablo
Biz, İki Abdal	We two dervishes	Nous, les deux Errants	Nosotros, dos derviches errantes
Ben, Kadın	I am a woman	Moi, la Femme	Yo, la mujer

In order to compare the styles of the characters, texts were transformed into numerical forms, in other words attributes were obtained from the texts. In this study, lexical features were used as attributes, which are number of words (number of tokens), number of different words (number of types), average word length in letters for both types and tokens, average sentence length in words, average count of vowels for both token and types, and frequency of the most frequently used words (vowel count is equal to syllable count in Turkish). While count of vowels are calculated, only the letters are considered, and their sounds are ignored. When creating frequency vectors, the most frequently used *k* words of each character were merged, and repeating words were removed. Since repeating words will be different for each language, size of word frequency vectors differs in original text and in translations. Lexical properties and word frequency vectors were then combined with different combinations to form feature vectors for each character. There are 20 attribute vectors matching the number of characters.

We have listed eight lexical features and first seven of them will be referred as lexical features and the last one will be referred as the most frequent words frequencies, for the ease of expression.

Based on the assumption that the styles of the characters will be preserved in the translations where the style is preserved, it is expected that the distances between the attributes of the characters calculated with the original text show the same distribution in translations. In order to test it numerically, the distances between the character vectors for the original text and each translation are calculated as Euclidean distance on the PCA plane. Correlation values between these distances were calculated and the relationship between original text and translations was observed. We refer to this correlation coefficient values as rank consistency-based similarities (RCS). Kendall's Tau Coefficient (Kendall's Rank Correlation Coefficient) was used for correlation calculation. Rank correlation measures the degree of similarity between the two ordered variables and evaluates the significance of this relationship statistically. Kendall's Tau Coefficient was chosen because it can measure the similarity between the two variables without the need for information on the distribution of variables. In

this step, distances among the translations and original text can be calculated as well, without reflecting them on a PCA plane. Since the rankings will be the same, results would not differ.

Experimental Results

Following the extraction of the attribute vector for each character, the principal component analysis (PCA) was applied to the vectors in order to observe the stylistic similarity of the characters. The principal component analysis is a size reduction tool that can be used to reduce a large set of variables to a small set containing most of the information in the set (Jolliffe, 2002, p.1). Using the PCA, the feature vectors were reduced to two dimensions and the scatter plots of the characters were obtained as in Figure 1.

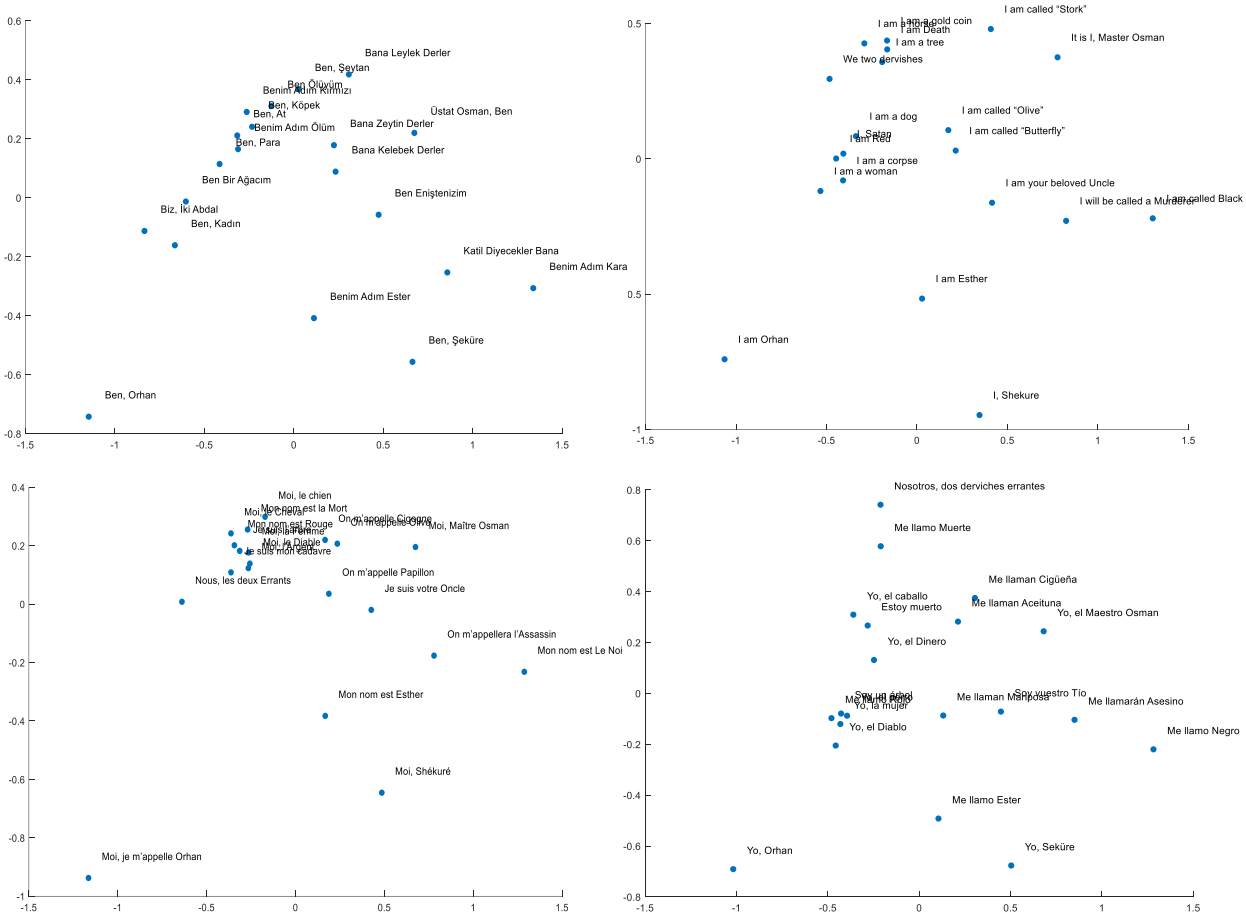


Figure 1. PCA diagram created with lexical features (for Turkish, English, French and Spanish, respectively)

Figure 1 shows 4 different scatter plots. It can be observed that the distribution of the stylistic features of the characters are similar for the original text and translations. It is possible to observe in the plots that number of words in the text segments corresponding to novel characters. is quite effective to differentiate them. However, it is not the only strong factor. For example, the features for the number of words of Orhan character and Two Abdals (851, 755) are close to each other and the number of words of Ester and Shekure (8723, 18404) is quite large compared to of Orhan. However, Orhan's distance from Two Abdals, does not seem very different from the distance from Shekure and Ester. For this reason, experiments have been carried out with different properties in order to examine the effect of attributes on the relationship between original text and translations.

The first experiment was made with feature vectors created using all lexical features but not frequency vectors. In the experiment, the style relationship between translations was aimed to be observed directly. The most commonly used words can be an effective determinant for the characters, so they can hide the effects of other stylistic features. That is why, they are not included in the attribute vector in this experiment. Correlation values among the original text and the translations can be seen in Table 3.a. As the p-values for all the correlation coefficients calculated in the article were smaller than 0.001, it was not required to make a table of them. The p-value indicates the probability that the calculated correlation values are coincidental. In this case, a small p-value indicates that this is unlikely to be a coincidence.

Table 3 (a, b)

RCS values calculated with lexical features (left), and correlations after words counts and different word counts are removed from lexical features (right)

	Turk.	Eng.	Fren.	Span.
Turk.	1	0.6493	0.7206	0.6345
Eng.		1	0.7414	0.7011
Fren.			1	0.7141
Span.				1

	Turk.	Eng.	Fren.	Span.
Turk.	1	0.5329	0.6292	0.5135
Eng.		1	0.6545	0.5857
Fren.			1	0.6207
Span.				1

It is expected that the number of words and the number of different words will increase the correlation with the assumption that they will change in similarly for characters in the original text and translations. In Table 3.b, the results were shown for the experiment conducted with the feature vectors from which the number of words and the number of different words were removed. All correlation values decreased compared to Table 3.a. The results confirm the expectation.

As mentioned, the most commonly used words can be effective in separating characters. In order to confirm this, as mentioned earlier experiments were performed with the most frequent words vectors for different *k* numbers. In Table 4.a results for *k*= 10, in Table 4.b for *k*= 20, and in Table 4.c for *k*= 30 can be seen. When the results of the experiments are compared with Table 3.a, it is seen that lexical features without the frequency of most frequent words give better results for all k values. More interestingly, the most common words vectors give the best coefficient values between the original text and the translations for the *k*= 20 value. This may be due to the fact that a few of the most frequently used words may not capture distinctive words, and when they are too many, distinctive words may appear in other characters as well.

Table 4 (a, b, c)

RCS values calculated with the most frequent words vectors k=10, k=20, and k=30

k=10	Turk.	Eng.	Fren.	Span.
Turk.	1	0.4549	0.3995	0.2414
Eng.		1	0.5853	0.4364
Fren.			1	0.3913
Span.				1

k=20	Turk.	Eng.	Fren.	Span.
Turk.	1	0.6722	0.6536	0.6804
Eng.		1	0.5884	0.6736
Fren.			1	0.6401
Span.				1

k=30	Turk.	Eng.	Fren.	Span.
Turk.	1	0.6441	0.4213	0.5683
Eng.		1	0.6159	0.6969
Fren.			1	0.6728
Span.				1

The results of the experiments performed to observe the similarity between the texts as the k increases can be seen in the plots of Figure 2. While the similarity of the translations with

the original text continues to change as the k values increase, the similarity between the translations reaches a steady state. The fact that English, French and Spanish are from the European branch of the Indo-European family of languages and Turkish is from a different family, the Ural-Altaic language family, may be the reason for these observations.

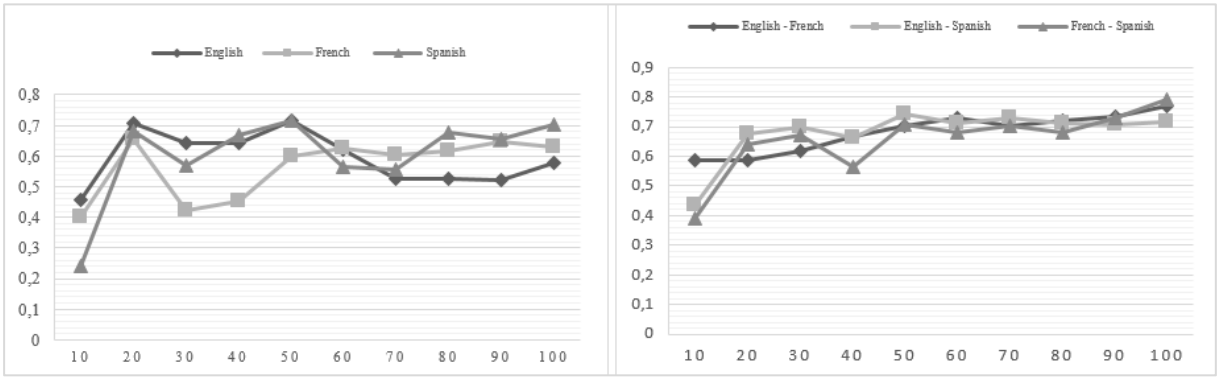


Figure 2. RCS value graphs for different k values between Turkish (left) and translation (right) texts

The results of the experiment with the combination of lexical features and the most frequent words vectors can be seen in Table 5. Compared to Table 3.a and Table 4.b, these vectors appear to yield better results when they are combined.

Table 5

RCS values calculated with lexical features and the most frequent words frequencies vectors (k=20)

	Turk.	Eng.	Fren.	Span.
Turk.	1	0.7069	0.6746	0.7053
Eng.		1	0.6556	0.7194
Fren.			1	0.6882
Span.				1

When the coefficients are examined, it can be seen that the similarity of the English, French and Spanish translations is usually higher than the similarity to the the original text in Turkish. For example, in Table 5, the coefficient between the original Turkish text and the Spanish translation is 0.7069, while the coefficient between the English translation and the Spanish translation is 0.7194. This may be, again, the result of the translation languages being from the same language family. Different numerical analysis were performed to examine the differences between Turkish and these languages.

The first analysis is the calculation of the word count rates for the characters. In Table 6, word counts of the texts in the language of the rows are divided by the number of words in the language of the columns and the result is written in the cell where the row and the column intersect. These ratios are calculated for each character and averaged over the characters. According to the results, it can be interpreted that more words are used to express the same situations in English, French and Spanish. In Turkish, time or subject is expressed by suffixes, while in the translation languages usually different words are used for them. In short, results can be explained by the Turkish language being a agglutinative language.

Table 6 (a, b)
Averages ratios of (a) number of words and (b) number of different words in terms of the characters

	Turk.	Eng.	Fren.	Span.
Turk.	1	0.6401	0.6033	0.6172
Eng.		1	0.9437	0.9657
Fren.			1	1.0254
Span.				1

	Turk.	Eng.	Fren.	Span.
Turk.	1	1.2228	1.0673	1.1725
Eng.		1	0.8816	0.9674
Fren.			1	1.0997
Span.				1

The proportions of the different word numbers for the characters were calculated by using the different languages in the same way and then the averages were taken. As can be seen from Table 6.b, in contrast to the total number of words, the number of different words used in Turkish is higher than in English, French and Spanish. This situation can also be explained by the fact that Turkish is an agglunative language and many different words are produced using the same root. After Turkish, French shows more varied vocabulary than other languages.

The average sentence length ratios were calculated as in Table 6 with sentence lengths and the results were given in Table 7. Examining the table, it can be interpreted that the sentences in Turkish are shorter than the translation languages. On the other hand, sentence lengths in translation languages are not very different. The sentence length ratio for Turkish - English overlaps with the ratio of 0.66 in the study of Patton and Can (2012).

Table 7
Average sentence length ratios

	Turk.	Eng.	Fren.	Span.
Turk.	1	0.6729	0.6534	0.6347
Eng.		1	0.9722	0.9447
Fren.			1	0.9759
Span.				1

Discussion

In this study, it is aimed to evaluate the style similarity between *My Name is Red* novel and its translations. For this purpose, style of the fictional characters in the novel were examined numerically. With lexical features and most frequent word frequencies, experiments were conducted. In the experiments performed with lexical features and the most frequently used words separately, correlation coefficient of the original text with the translations was close to each other. Best result was seen when those features were merged as a single attribute vector. This is an expected result since different features have the potential of bringing new information about the novel character.

Experiments with the most frequently used words were continued by expanding the number of k, and as the k-value increases, the correlation between the translations remained constant between 0.7 and 0.8. No significant improvement was observed with the increase of k value in the similarity between original Turkish text and translations. It can be explained by the fact that Turkish has a different origin from the translation languages.

When we look at the ratio of the number of words between Turkish texts and translations, it is seen that the number of words used in the translation languages is higher. Since the Turkish language is a agglutinative language, this result is not different from expected. The different word ratios used in the texts indicate that the number of different words used in Turkish is higher than the translations. In addition, translation languages are almost identical to each other in terms of different word numbers. In sentence length ratios, the rates of translation languages are also close to each other. The translation languages being the members of the same language family can also explain these results.

Conclusion and Future Work

In this study, we present a comprehensive survey of stylometry in the context of Turkish texts and a new work on the quantification of loyalty for translations. In the survey part, different application areas are explained and the related stylometry studies are presented accompanied by a detailed table. In the full body of the paper another table for open source software tools that may help style researchers is provided.

In the second part, the style similarity between original text of *My Name is Red* and its translations in English, French and Spanish is analyzed by using a rank consistency-based measure that we introduce in this paper. The experiments with lexical features and most frequently used words statistically significantly confirm the similarity. They also show that the pairwise compatibility of translation languages is higher than that of their compatibility with the Turkish original. The observations are as expected for these target languages that are the members of the same language family.

In future research, experiments with different language families can be performed. Furthermore, similar studies can be done with other works by dividing them into cohesive units, such as short stories or parts of novels containing related themes.