

Türkçe Kısa Metinlerde Dilsel Değişke İncelemesine Çok Boyutlu Bir Yaklaşım*

Hülya Mısır¹

ORCID: ¹0000-0003-4103-682X

¹Birmingham Üniversitesi, Dilbilim ve İletişim Bölümü,
Birmingham, Birleşik Krallık

¹ hulyamsr@gmail.com

(Gönderilme tarihi 21 Nisan 2025; Kabul edilme tarihi 18 Eylül 2025)

ÖZ: Bu çalışmada, nötr, saldırgan ve nefret içerikli tweetlerden oluşan büyük ölçekli bir Türkçe sosyal medya derlemi kullanılarak Türkçedeki dilsel değişkeler incelenmiştir. Sözcük türleri ve dilbilgisel yapılar açısından etiketlenmiş veri setiyle, dilsel değişke türlerinin altında yatan temel boyutlar, Çok Boyutlu Analiz (MDA) kapsamında Çoklu Uyum Analizi (MCA) yöntemiyle belirlenmiştir. Kısa ve bağlamsal olarak sınırlı sosyal medya metinlerine uygunluğu sayesinde MCA'nın dilbilimsel analizlerdeki yeri açıklanmakta, Türkçe kısa metinlerde dilsel değişkeyi ortaya koymadaki avantajları uygulamalı biçimde gösterilmektedir. Analizde, *FactoMineR* paketi ve yaygın olarak kullanılan görselleştirme aracı *ggplot2* birlikte kullanılmaktadır. Bu uygulamalı anlatım, MDA boyutlarının yorumlanması ve veri görselleştirme teknikleriyle ilişkilendirilmesi konusunda rehberlik etmektedir. Ayrıca, tarih etiketi ve konuşma kategorileriyle etiketlenmiş veriler üzerinden dilsel örüntülerdeki zamansal değişim grafikler ve ısı haritalarıyla sunulmaktadır. Bu çalışma, kısa metinlerden oluşan derlemler ve kategorik verilerle çok boyutlu dilsel analiz yapmak isteyen araştırmacılar için olduğu kadar, veri görselleştirme konusunda bilgi edinmek isteyen herkes için faydalı bir kaynak olmayı hedeflemektedir.

Anahtar sözcükler: dilsel değişke, çok boyutlu analiz, Türkçe, nefret söylemi derlemi, kısa metin analizi

* Analizler konusunda danışmanlık sağlayan Jack Grieve'in katkıları bu çalışmanın ilerlemesinde büyük önem taşımıştır. Bu analizin yapılabilmesini mümkün kılan veri setini cömertçe paylaşan Çağrı Toraman'a ve maddi destekleri için TÜBİTAK'a teşekkür ederim.

<https://doi.org/10.18492/dad.1675004>

Dilbilim Araştırmaları Dergisi, 2025/2, 133–157.

© 2025 Dilbilim Derneği, Ankara.

Creative Commons Atıntı-GayriTicari-Türetilemez 4.0 Uluslararası

(CC BY-NC-ND 4.0) lisansı ile lisanslanmıştır.



A Multidimensional Approach to Linguistic Variation in Short Turkish Texts

ABSTRACT: This study investigates linguistic variation in Turkish using a large-scale social media corpus consisting of neutral, offensive, and hate speech tweets. Drawing on a dataset annotated for parts of speech and grammatical structures, the study identifies the main dimensions of linguistic variation through the framework of Multidimensional Analysis (MDA), using Multiple Correspondence Analysis (MCA). The paper presents the use of MCA method in Turkish, which fills a notable gap in Turkish linguistic analysis due to its suitability for short and contextually limited texts such as those found on social media. The analysis is conducted using the *FactoMineR* package in R, along with the widely used visualization tool *ggplot2*. This practical guide helps interpret the dimensions generated by MDA and demonstrates how results can be presented through different data visualization techniques. Additionally, the study presents temporal shifts in linguistic patterns using time-stamped and category-labeled data, presented through various plots and heatmaps. The article is intended as a practical resource for researchers applying MDA to short-text corpora, and for those interested in the use of data visualization in linguistic analysis.

Keywords: linguistic variation, multidimensional analysis, Turkish, hate speech corpus, short-text analysis

1 Giriř

Dil, baęlama baęlı olarak sistematik biçimde farklılık gösterir. Bu *dilsel deęişke* (linguistic variation); yazılı ya da sözlü iletişimde, konuşucunun amacı, hedef kitlesi ve iletişim bağlamı gibi çeşitli etkenlere göre dilsel tercihlerde ortaya çıkan farklılıklarla şekillenir. Biber'in (1988) *Çok Boyutlu Analiz* (Multidimensional Analysis, MDA) yaklaşımı, metinler arasındaki biçimbilimsel, sözdizimsel ve biçimsel örüntülerin sistematik olarak incelenmesini mümkün kılar. Bu sayede *tür ve üslup* (register, style) temelli analizler yapılabilir ve dilsel örüntüler, dolaylı olarak metin işlevleriyle ilişkilendirilerek yorumlanabilir. Biber'in geliřtirdięi ve İngilizce başta olmak üzere farklı dillerde başarıyla uygulanmış MDA, metinler arası dilsel deęişkenin nicel olarak ortaya koyan en köklü yaklaşımlardan biridir. Ancak bu yaklaşım, temelde uzun metinlerin incelenmesinde kullanılmış ve özellikle sözlü ve yazılı dildeki sistematik deęişimleri ortaya koyma amacıyla geliřtirilmiştir. MDA, bir dildeki belirli söylem türleri veya metin tipleri arasındaki çok boyutlu dilsel deęişkeyi ve biçimsel farklılıkları incelemek amacıyla da uygulanmıştır. Ancak, sosyal medya platformlarından sadece biri olan Twitter/X verilerinin MDA ile analiz edilmesi bazı sorunlar barındırmaktadır; çünkü tweetler oldukça kısa metinlerdir ve bu durum, örneğin tek bir tweette gözlemlenen dilsel özelliklerin göreceli sıklığının, söz konusu tweetin ait olduęu genel kümedeki gerçek sıklığı

güvenilir biçimde yansıtamamasına yol açabilir (Clarke ve Grieve, 2019). Bu bağlamda, kısa metinlerin analizi için MDA'nın bazı sınırlılıklarının göz önünde bulundurulması ve bu sorunları aşmaya yönelik uygun yöntemlerin değerlendirilmesi önemlidir. Değişke analizini bu çerçevede düşünmek sosyal medya verisiyle çalışan araştırmacılar için önem kazanmaktadır.

Günümüzde, sosyal medya gibi dijital platformlarda üretilen kısa, bağlamdan kopuk ve çoğu zaman yüksek duygusal yüke sahip metinlerde bu değişkeyi daha da belirgin hale gelmektedir. Bu tür kısa metinler, geleneksel metin analiz yöntemleriyle tam anlamıyla yakalanamayan; ancak dilin biçimsel ve işlevsel çeşitliliğini yansıtan önemli dilsel ipuçları barındırır. Özellikle Türkçe gibi, sosyal medya diline özgü değişkelerin henüz yeterince haritalanmadığı dillerde, bu alan hem yöntemsel hem de betimleyici katkılara açıktır.

Çalışma, Toraman vd. (2022) tarafından derlenen ve Türkçe Twitter verilerini içeren geniş kapsamlı bir derleme kullanılmaktadır. 60.068 tweetten oluşan bu derlemede, her bir tweet manuel olarak üç sınıf altında etiketlenmiştir: nötr, saldırgan ve nefret söylemi. Bu veri seti, Türkçe sosyal medyada dilsel değişkenin farklı yapısal, biçimsel ve işlevsel boyutlarını incelemek için sağlam bir temel sunmaktadır.

Türkçede kısa sosyal medya metinlerinin incelenmesi amacıyla, Biber'in (1988) MDA çerçevesi izlenmiş ve faktör analizi yerine, kategorik verilere dayalı *Çoklu Uyum Analizi* (Multiple Correspondence Analysis, MCA) kullanılarak dilsel değişkenin temel boyutlarının ortaya konması hedeflenmiştir (Clarke ve Grieve, 2017). Analizde *FactoMineR* paketi (Husson vd., 2017) kullanılmış; elde edilen boyutlar ise *ggplot2* kütüphanesi aracılığıyla grafiklerle görselleştirilmiştir. Bu tercihte, *FactoMineR*'in çok boyutlu veri setlerine yönelik MCA analizinde sunduğu kapsamlı istatistiksel fonksiyonlar ile *factoextra* paketinin R ortamında sağladığı özelleştirilebilir ve tematik görselleştirme seçenekleri etkili olmuştur. Bunun yanında, tweetlerin atıldığı tarihler kullanılarak kısa metinlerdeki dilsel örüntülerin zaman içinde nasıl değiştiğinin gözlemlenmesi mümkün hale gelmiştir.

Özetle, bu çalışma, Türkçede dilsel değişke analizinin, özellikle kısa metinler üzerinden yapılmamış olması nedeniyle alana özgün bir katkı sunmaktadır. Sosyal medya verisiyle yapılan dilsel değişke analizi, Türkçe literatüre önemli bir yenilik getirmektedir. Temelde MDA yaklaşımı ve MCA yöntemi kullanılarak dilsel değişkelerin farklı boyutları daha kapsamlı bir şekilde incelenmiş ve bu sayede sosyal medya dilinin *üslup/biçimsel* (stylistics) özellikleri derinlemesine analiz edilmiştir.

Bu çalışma, sosyal medya dilini *tür* (register) analizi çerçevesinde inceleyerek, nefret söylemi, saldırgan dil ve nötr dil gibi farklı dilsel boyutları ayırmada güçlü bir analiz metodu ortaya koymaktadır. Türkçe ve Türkiye'de nefret söylemi çalışmalarına bakıldığında Erdoğan-Öztürk ve Işık Güler (2020) ile Özdüzen ve diğerleri (2021) gibi küçük ölçekli sosyal medya vaka çalışmaları Suriyeli mültecilerle ilgili olarak mülteci meselesinin sosyal medyada, özellikle Twitter'da, yaygın olan ırkçı söylemin başlıca kaynağı olduğunu vurgulamıştır.

Bu alıřmalar, etkileřim niyetlerini ve nefret sylemi ni ok kipli kaynaklar aracılıęıyla destekleyen ierikleri incelemeye odaklanmıřtır. Toraman vd. (2022) ve ltekin (2020) ise Trke nefret sylemi ve saldırgan dilin etiketlendięi kaynaklar retmiř ve *makine ęrenimi* (machine learning) yntemleriyle otomatik tespit alıřmalarını destekleyen sonular sunmuřtur. Ancak, Trkede bu tr bir dilsel deęiřke analizi yapılmamıř olması nedeniyle de bu alıřma nemli bir kaynak olacaktır. Ayrıca, elde edilen bulguların grselleřtirilmesi, verilerin daha anlaşılır hale gelmesini saęlamakta ve analiz srecinin řeffaflıęını artırmaktadır.

2 Tr Analizi ve Dilsel Deęiřke

Dilin yapısının, iinde bulunulan duruma yani iletiřim baęlamına gre deęiřiklik gstermesi dřncesi, sylem *tr* (*register*) deęiřkelerinin arařtırılmasının temelini oluřturur (Biber ve Conrad 2019; Grieve, 2023). Bu arařtırmalar, bireylerin kullandıęı dilbilgisel yapıların yalnızca hedef kitleye deęil; konuřma biimine, iletiřim aracına (rneęin yazılı ya da sztl oluřuna), konuya ve ortama gre sistemli olarak deęiřtięini ortaya koymuřtur. Ayrıca bu alıřmalar, dildeki bu farklılıkların, iletiřim ortamlarının sunduęu imkanlar ve sınırlılıklarla, aynı zamanda insanların bu ortamlardaki iletiřim amalarıyla yakından iliřkili olduęunu gstermektedir.

Tm dillerin en temel zelliklerinden biri, kullanıma baęlı olarak deęiřkenlik gstermeleridir. Sylem tr kuramının amaı, dildeki bu deęiřkenlięi yneten genel ilkeleri ortaya koymaktır. Bylece, hangi durumsal faktrlerin hangi dilsel zellikleri etkiledięini anlayabiliriz. Halliday'ın (1978) sylem trleriyle ilgili yaptıęı tanımdan bu yana, zellikle İngilizce zerine yapılan alıřmalar ve farklı diller arasında yapılan karřılařtırmalı sylem tr analizleri, dildeki deęiřkenlięi etkileyen faktrleri belirleme konusunda nemli ilerlemeler kaydetmiřtir. Bu alıřmaların Trke iin de yapılması ve karřılařtırmalı alan iin bilgi retilmesi noktasında bu alıřmanın nemli bir rnek olması hedeflenmektedir.

Dil yapısının iletiřimsel amaca gre nasıl deęiřtięini anlamının yollarından biri, ok boyutlu analiz– MDA– yntemidir (Biber, 1988, 1989). Uzun bir arařtırma geleneęine sahip olan bu yntem, genellikle belirli bir dil deęiřkeyi temsil eden bir metin derlemi zerindeki ok sayıda szcksel ve dilbilgisel zellięin *greli sıklıklarına* dayanır. Bu verilerden, dildeki temel deęiřke boyutları faktr analizi yoluyla ıkarılır. Ardından, her bir boyutla en gtl řekilde iliřkili olan dilsel zellikler ve metinler temel alınarak bu boyutlar *biemsel* ve *iřlevsel* olarak yorumlanır (Clarke ve Grieve, 2017).

MDA yaklařımı, istatistiksel faktr analizini kullanarak ok sayıda dilbilgisel deęiřkeni, dilsel deęiřkelerin temel *boyutlarına* (dimensions) indirger. Biber (2015, s. 10) her bir boyutu  aıdan incelemektedir. Her boyutun, bu  temel aıdan kendine zg nitelikler tařıdıęını belirtir. (i) Bir takım birlikte ortaya ıkan dilsel zellikler kmesiyle tanımlanır. (ii) Belirli iletiřimsel iřlevlerle iliřkilidir. (iii) Her boyut, farklı sylem tr deęiřke kalıpları ile iliřkilidir. Bu

yaklaşımına göre her bir boyut, pozitif ve negatif olmak üzere iki kutuptan oluşur. Buradaki pozitif ve negatif kutuplar, değer yargısı ifade etmez. Aksine, her kutup, sıklıkla birlikte görülen dilbilgisel özelliklerin oluşturduğu belirli bir “özellik setini” temsil eder. Bu özellikler genellikle birbirini tamamlayan bir dağılım gösterir. Bir metin grubunda bir özellik seti sık kullanılıyorsa, diğer setin nadiren kullanıldığı gözlemlenir çünkü bu iki set iki farklı kutupta konumlanır.

Biber (1988, 1995, 2015), İngilizcede beş farklı boyutta derlenen özellik setleri ile ilgili çalışmasında bu boyutların iki kutupta yer alan dilsel değişke örnekleri sunduğunu belirtmiştir. Örneğin, *anlatı* (narrative) vs *anlatı dışı* (non-narrative) söylem boyutunda, pozitif kutupta geçmiş zaman, geçmiş zaman eylemleri, üçüncü kişi adılı, bitmemişlik görünüşü ve *iletişim eylemleri* (communication verbs) gibi özellikler bir arada yer alırken, negatif kutupta şimdiki zaman eylemleri ve nitelendirme sıfatları gibi özellikler bir arada görülmektedir. Benzer şekilde, *katılımcı* (involved) vs *bilgi odaklı* (informational) üretim boyutunda, pozitif kutupta şimdiki zaman eylemleri, soru sözcüklü yapılar, birinci ve ikinci kişi adıları, 'o' (it) adılı, belirsiz adıl, 'yap-' eylemi, gösterim adıları, *vurgulayıcılar* (emphatics), *kaçınmalar* (hedges), *anlam güçlendiriciler* (amplifiers) bir arada görülebilirken, negatif kutupta adlar, uzun sözcükler, belirteçler, *tür-belirteç oranı* (TTR- type-token ratio) ve nitelendirme sıfatları gibi özellikler yer almaktadır. Bu örnekler İngilizceden alınmıştır. Ancak, farklı dillerde de benzer biçimsel değişkeleri inceleyen çalışmalar yapılmıştır (Kim ve Biber, 1994; Biber ve Hared, 1994; Biber vd., 2006). Her dilin kendi yapısal özellikleri, sözcük türleri ve kullanım biçimleri, bu tür analizlerin farklı boyutlar oluşturmasına yol açar. Bu sayede, sadece dilin yapısal özellikleri değil, aynı zamanda dilin iletişimdeki işlevsel rollerine de dair önemli bilgiler elde edilebilir. Örneğin, farklı kültürel bağlamlarda ve iletişim türlerinde, dilin nasıl biçimlendiği ve hangi dilsel özelliklerin daha baskın olduğu anlaşılabilir. Böylece, dilsel değişkenin, sadece dilin kendisine değil, aynı zamanda iletişimsel hedefler ve toplumsal bağlam ile nasıl şekillendiği de netleşmiş olur.

Türkçe bağlamında dilbilimsel yapılar, farklı disiplinlerin bakış açıları doğrultusunda çeşitli yönleriyle ele alınmıştır. Örneğin, Balcı (2020) dilsel değişkeyi toplumsal ve bağlamsal değişkenler üzerinden kuramsal bir düzlemde ele alırken, Yüceol Özezen (2021) ise Türkçeyi evrensel tipoloji bağlamında biçimbilgisel ve sözdizimsel açılardan değerlendirmiştir. Ancak, özellikle kısa ve bağlamsal olarak sınırlı sosyal medya metinlerinde, bu yapısal özelliklerin dilsel değişke örüntüleri üzerine ampirik çalışmalar henüz sınırlı düzeydedir. Bu bağlamda, bu çalışma, sosyal medya gibi kısa metin türlerinde dilsel öğelerin birlikte kullanımı ve işlevsel kümelenmesini hem biçimsel hem de işlevsel değişke ekseninde inceleyerek alandaki yöntemsel boşluğu doldurmayı amaçlamaktadır. Böylece, Türkçede işlevsel ve biçimsel değişkelerin ampirik, nicel ve bağlama duyarlı biçimde haritalanması hedeflenmiştir.

3 Kısa Metinlerde Çok Boyutlu Analiz ve Çoklu Uyum Analizinin Rolü

Tweetler gibi kısa metinlerden oluřan veri setlerinin analizinde, MDA ile birlikte kullanılan faktör analizi uygun deęildir; çünkü dilsel özelliklerin göreceli frekans dağılımları kısa metinlerde güvenilir olmayabilir. MDA genellikle daha uzun metinleri gerektirir. Kısa metinleri birleřtirerek daha uzun bir veri seti oluřturmak bir çözüm olabilir, ancak bu yaklařım bireysel tweet düzeyindeki deęiřkeleri inceleme imkânı sunmaz. Bir tweetteki göreceli sözcük frekansları, daha geniř bir veri kümesindeki gerçek dağılımı yansıtmayabilir. Örneęin, 10 sözcüklük bir tweette her sözcüğün frekansı yalnızca o tweet temelinde hesaplanır; ancak daha büyük bir veri kümesinde dağılım farklılık gösterebilir, çünkü bazı yaygın kullanılan sözcükler tweet içinde yalnızca birkaç kez geçebilirken, bazı sözcükler hiç yer almayabilir. Bu nedenle kısa metinlerde frekans tahminleri, daha büyük veri setlerindeki gerçek dağılımlarla uyumsuzluk gösterebilir ve MDA'nın doęruluęunu olumsuz etkileyebilir. Buna karřın, MCA, kategorik verilere dayalı bir yöntem olarak frekans tahminine baęımlı deęildir ve bu nedenle tweetler gibi kısa metinlerin analizinde MDA yaklařımı içinde faktör analizinin yerine kullanılabilecek uygun bir yöntem sunar. MCA, MDA'nın kendisini deęiřtirmez; yalnızca MDA içinde kullanılan faktör analizine alternatif bir teknik olarak iřlev görür.

MCA temelde kategorik verilerle çalıřır ve çok deęiřkenli dilsel bir veri setini birkaç ana boyuta indirger (Le Roux ve Rouanet, 2010). Bařka bir deyiřle, bu yöntem yüksek boyutlu veriyi, deęiřkenler arasındaki en anlamlı iliřkileri koruyarak daha düşük boyutlu bir hale dönüřtürür. Bu boyut indirgeme, veri setlerinde en anlamlı deęiřkeleri ortaya koyar ve biçimsel ve iřlevsel düzeyde dil kullanımı hakkında öngörüler sunar. Elde edilen boyutlar, her bir boyuta hangi dilsel özelliklerin en güçlü řekilde katkıda bulunduęunu belirleyerek yorumlanır.

MCA yönteminin kısa metinlerde uygulanabilirlięi, yalnızca teknik avantajlarından deęil, aynı zamanda bu metinlerin içerdii biçimsel ve iřlevsel dil çeřitlilięinden kaynaklanmaktadır. Özellikle sosyal medya bağlamında üretilen kısa metinler, geleneksel yazılı metinlere kıyasla birçok dilbilimsel farklılık gösterir. Türkçede kısa metinlerde bağlamdan baęımsız ifadeler, eksiltili yapılar, emir kipleri, doğrudan hitap biçimleri ve yer yer sözdizimsel eksiklikler gözlemlenmektedir. Örneęin, “kahrolsun PKK” gibi doğrudan hitap içeren, eksiltili ya da emir kipli ifadeler analiz düzleminde genellikle kısa metinlerle birlikte kümelenmektedir. Buna karřılık, nötr bir söylem örneęi olan “vergi oranları %18 olarak belirlenmiřtir” ifadesi, daha yapılı, edilgen ve resmi özellikleriyle uzun metin kümesiyle birlikte konumlanmaktadır. Bu konumlanmalar, kısa metinlerin bağlamdan baęımsız, doğrudan ve sözdizimsel olarak sade yapılara sahip olduęunu; uzun metinlerin ise yapısal bütünlük, bilgi yoğunluęu ve dolaylı anlatım içerdiiğini göstermektedir. Bu dilsel örüntüler, kısa metinlerin hem biçimsel hem de iřlevsel açıdan ayrı düzlemlerde analiz edilmesini gerekli kılmaktadır. Bu nedenle, kısa ve bağlamsız Türkçe metinlerde

çok boyutlu istatistiksel analizlerin tercih edilmesi, metin içi deęişiklerin yüzeysel deęil, yapısal örüntüler temelinde incelenmesini mümkün kılar. Ancak kısa metinlerin içerdii dilsel özelliklerin metin uzunluęuna baęlı olarak deęişebileceęi göz önünde bulundurulmalıdır. Dolayısıyla MCA, yalnızca çok deęişkenli veriyi indirgemekle kalmaz; aynı zamanda metin uzunluęu ve dilsel özellikler arasındaki örüntüleri açık ve görselleştirilebilir biçimde ortaya koyar.

MCA, aynı zamanda ek nicel deęişkenlerin de dahil edilmesine olanak tanır (Husson vd., 2017), bu da metin uzunluęunu kontrol etmeye yarar. Örneęin, Clarke ve Grieve (2019) çalışmalarında her tweetin sözcük temelinde ölçülen uzunluęunu hesaplayarak metin uzunluęunu MCA'da ek bir nicel deęişken olarak dahil etmiş ve dilsel özelliklerin kategorik varlıkları (Presence ve Absence) ile tweet uzunluęu arasındaki ilişkileri incelemiştir. Bu yaklaşım, metin uzunluęunun ana analizde kontrol edilmemesi durumunda, uzun metinlerin daha fazla özellik içermeye eğiliminde olduęu ve kısa metinlerin daha az özellik barındırdıęı için metin uzunluęunun temel analizle karışabileceęi düşüncesine dayanır. Örneęin, Tablo 1'deki tweete bakıldığında, satırlar bu metindeki dilsel yapıların varlık/yokluk bilgisini gösterirken sütunlar bu dilsel özellikleri listelemektedir (bu örnekte 61 dilsel özellik mevcut). Bu aşamada bir özellięin kullanım sıklıęı değerlendirilmezken, mevcut olup olmadıęı değerlendirilmektedir. Ancak bu durumda, bir metinde mevcut dilsel özellikler metin uzunluęuna baęlı olarak deęişkenlik gösterecektir. Bu sebeple metin uzunluęu ile MCA boyutları arasındaki korelasyonu ana analizi etkilemeden hesaplamak önemlidir.

Tablo 1. Dilsel özelliklerin kategorik matrisi

| Tweet ID | Text | Verb | Aor | A1sg | Adj | ... |
|----------|-----------------------------------|------|-----|------|-----|-----|
| 1 | umarım (Verb) + (Aor) + (A1sg)... | P | P | P | A | ... |

Verb=eylem, Aor=geniş zaman, A1sg= 1. tekil kişi, Adj= sıfat

Kısaca MCA, birkaç kategorik deęişkenin birlikte nasıl yapılandıęını analiz eder. Özellikle anket verileri, dilsel veri kümeleri ya da metin madencilięi gibi alanlarda kullanılır. Gözlemler ve deęişkenler aynı grafik üzerinde gösterilebilir ve kategoriler arasında benzerlikler, kümeleşmeler ve temel boyutlar görselleştirilir. Makalenin devamında, MCA yönteminin uygulanışı adım adım ayrıntılı biçimde ele alınmıştır. Bu anlatımda ilgili kavramlara ve kararlara dair açıklamalar sunulmuştur. Analiz sürecinde tam şeffaflık ve tekrarlanabilirlięin sağlanması amacıyla, her bölümde ilgili R kodlarına ve görselleştirmelere yer verilmiştir. Bu yönüyle çalışma, öğretici bir uygulama sunmayı amaçlamakta ve MCA'nın uygulamalı dilbilim alanına sağlayabileceęi katkıları ortaya koymayı hedeflemektedir.

4 Veri Hazırlama ve Dilsel Özelliklerin Belirlenmesi

Bu çalışma, Toraman ve ark. (2022) tarafından derlenen, kamuya açık ve Türkçe dilinde Twitter etkileşimlerini içeren oldukça kapsamlı ve güvenilir bir nefret söylemi derlemine kullanmıştır. 2006–2021 yıllarını kapsayan ve toplam 60.310 Türkçe tweet içeren bu veri seti, gönderi düzeyinde üç sınıflı bir sınıflandırma görevi için manuel olarak etiketlenmiştir: Normal (32.555), Saldırgan (14.934) ve Nefret (12.821). Analiz kapsamında, birbirine çok benzeyen tweetler dahil edilmemiş ve 1 Ocak 2009'dan önce atılan tweetler, bu döneme ait örneklerin sayıca yetersiz olması (yalnızca 242 tweet) nedeniyle dışarıda bırakılmıştır. Bu işlemler sonucunda, çalışmada kullanılan nihai veri kümesi 60.068 tweetten oluşmaktadır: Normal (32.326), Saldırgan (14.926) ve Nefret (12.816).

Bu kısımda verinin hazırlanması ve bir dizi dilsel özelliğın belirlenmesi sürecinde izlenen adımlar açıklanmaktadır. Öncelikle veri setlerini küçük *parçalara ayırmak* (tokenize) ve biçimsel etiketleme yapmak için, Türkçe için *doğal dil işleme* (Natural Language Processing- NLP) kütüphanelerinden biri kullanılmalıdır¹. Bu noktada Türkçe için mevcut sözcük türü etiketleme araçları değerlendirilerek uygun olanı seçilmiştir². Bu makalede etiketlemeye dair kritik faktörler ile sınırlayıcı ve destekleyici unsurlar ele alınmamıştır, ancak genel olarak Türkçe ile ilgili doğal dil işleme kütüphanelerinin Java ve Python ekosisteminde geliştirildiğini söylemek doğru olacaktır³. Çalışmada kullanılan veri, Python tabanlı Zemberek Kütüphanesi (Akın, 2023) aracılığıyla etiketlenmiştir⁴. Zemberek, Türkçenin dilbilgisel yapısına özgü ihtiyaçlara uygun olarak geliştirilmiş, açık kaynaklı ve güvenilir bir NLP kütüphanesidir.

¹ Zemberek, TRMorpheme, Stanford POS Tagger, ya da spaCy Türkçe modelleri bunlara örnektir.

² Türkçe için POS etiketleme yaparken, *kural tabanlı, makine öğrenimi tabanlı ve derin öğrenme tabanlı* yöntemler arasında tercihler değişebilir. Bu tercihler, veri setinin büyüklüğüne, kullanılan modele ve hedeflenen doğruluğaa göre şekillenebilir. Transformer tabanlı modeller (örneğin BERT, LSTM) gibi modern derin öğrenme yöntemleri, dilin karmaşık yapılarında ve bağlamlarında daha etkili sonuçlar verebilir.

³ Türkçe dil işleme için yaygın olarak kullanılan çoğuu model, genellikle Python ile geliştirilmiş ya da Java ile geliştirilip Python bağlayıcılarıyla kullanılabilen kütüphaneler olarak düzenlenmiştir. R dilinde ise, bu Python kütüphanelerine bağlanabilen bazı araçlar ve paketler (örneğin, stanfordnlp, spacyr veya text) mevcuttur, ancak bunlar doğrudan R tabanlı bağımsız kütüphaneler değildir. Bu yüzden, Türkçe dil işleme projelerinde çoğuu zaman Python kullanılması daha yaygındır.

⁴ Zemberek kural tabanlı (rule-based) bir dil işleme aracıdır. Bu yaklaşımda, dilin yapısı hakkında belirli kurallar ve dilbilimsel bilgilere dayanarak etiketleme yapılır. Yani, Zemberek, biçimsel çözümleme ve sözcük türü (Part-of-Speech) etiketleme gibi işlemleri gerçekleştirmek için dilbilimsel kurallar kullanır. Ancak, bu yaklaşımın avantajları olduđu gibi, sınırlamaları da vardır. Özellikle dilin dinamik yapısı ve çok anlamlı sözcükler gibi karmaşık durumlarda bazen hatalı etiketlemelere yol açabilir.

Python üzerinden erişilebilirliđi ve güçlü biçimbilimsel çözümlene kapasitesi nedeniyle bu araç kullanılmıřtır. Veri setinde etiketlenmiř bir tümce Tablo 2'deki gibi görünür.

Tablo 2. Veri örneđi

| Tarih | Metin | Etiket | Tweet Uzunluđu |
|------------|---|--------|----------------|
| 2019-04-26 | umarım (Verb) + (Aor) + (A1sg) ... | 0 | 12 |
| 2020-10-29 | kadın (Noun) + (A3sg) cinayetleri (Noun) + (A3pl) + (P3sg) katliama (Noun) + (A3sg) + (Dat) | 0 | 16 |
| 2012-04-10 | vatan (Noun) + (A3sg) hainliđini (Adj) + (Ness) + (Noun) + (A3sg) + (P2sg) + (Acc) | 2 | 33 |

Verb=Eylem, Aor=geniř zaman, A1sg=1. tekil kiři, Noun=ad, A3sg=3. tekil kiři, A3pl=3. çođul kiři, P3sg=3. tekil kiři iyelik, Dat= yönelme hali, Adj=sıfat, Ness=adlařtırma eki, P2sg=2. tekil kiři iyelik, Acc= belirtme hali

Zemberek kütüphanesinin sözcük türü etiketlerine ek olarak, Twitter verilerinde sıkça karřılařılan Hashtag, URL, Mention ve Sym (emoji & emoticon) gibi özel etiketler de analize dâhil edilmiřtir. Ardından, her bir etiketin cümlelerdeki toplam görölme sıklıđı hesaplanmıř; verinin %5'inden az tümcede yer alan etiketler analizden çıkarılmıřtır. Bu filtreleme, nadiren görölen etiketleri eleyerek daha yaygın ve anlamlı örüntülere odaklanmayı amaçlamaktadır. Son olarak, kalan etiketlerin her bir tümcede yer alıp almadıđına göre, her bir hücrelerinde etiketin varlıđını (P) ya da yokluđunu (A) gösteren bir varlık-yokluk matrisi oluřturulmuřtur (Tablo 1).

Bu bölümde, MCA'ya dair yalnızca temel adımlara odaklanılmıř; ilgili kod parçaları paylařılmıř ve veri iřleme sürecinde izlenen kararların gerekçelerine yer verilmiřtir. Bu řekilde, analizin uygulanıř biçimine dair genel bir çerçeve sunulması hedeflenmiřtir. Tüm kodlara ve ayrıntılı açıklamalara, ek materyale yönlendiren bađlantı aracılıđıyla erişilebilmektedir.

Uygulama kapsamında öncelikle gerekli R kütüphaneleri yüklenmiřtir. MCA analizi için temel araç olan *FactoMineR* (Husson vd., 2017) paketi, kategorik verilerin analiz edilmesi, boyut indirgeme ve görselleřtirme iřlemlerini gerçekteřtirmek üzere kullanılmıřtır. *FactoMineR* çıktılarının görsel olarak analiz edilmesine olanak tanıyan yardımcı paket *factoextra* da sürece dâhil edilmiřtir. Ayrıca, MCA sonuçlarının daha zengin ve açıklayıcı grafiklerle sunulabilmesi amacıyla *ggplot2* kütüphanesinden de yararlanılmıřtır (Şekil 1). R dili, istatistiksel analiz ve görselleřtirme ekosistemindeki zengin paket desteđi nedeniyle tercih edilmiřtir. Python'da MCA analizleri için *mca* ve *prince* gibi paketler bulunmakla birlikte, bu paketlerin görselleřtirme iřlevleri genellikle *matplotlib* ve *seaborn* gibi kütüphaneler aracılıđıyla manuel olarak

gerçekleřtirilmektedir. Buna karřılık, *FactoMineR* paketinin tamamlayıcı birey ve deęişkenlerin ana boyutlara göre konumlandırılmasını saęlaması, yeni gözlemlerin bileřen düzlemine yansıtılmasına olanak tanınması ve tamamlayıcı niceliksel deęişkenleri (*quanti.sup*) desteklemesi; ayrıca *factoextra* paketinin sunduęu gelişmiş görselleřtirme olanakları ve *ggplot2*'nin esnek yapısı, bu araçların tercih edilmesinde belirleyici olmuřtur.

```
library(readxl)
library(stringr)
library(dplyr)
library(FactoMineR)
library(ggplot2)
library(factoextra)
library(zoo)
library(tidyr)
```

Şekil 1. MCA analizinde kullanılan R kütüphaneleri

Metinlerde yer alan sözcük türü etiketlerini (örneğin (Verb), (Aor), (A1sg)) ayıklamak için *düzenli ifadeler* (regular expressions, regex) kullanılmıştır. Regex, metin içerisinde belirli desenleri bulma, eşleřtirme veya dönüřtürme amacıyla kullanılan bir araçtır. Bu çalışmada, parantez içerisindeki etiketlerin tespit edilmesi için `\([^\)]+\)` regex ifadesi kullanılmıştır (Şekil 2). Örneęin, “umarım (Verb) + (Aor) + (A1sg)” şeklindeki bir örnekten Verb, Aor, A1sg etiketleri çıkarılır.

```
extract_pos_tags <- function(tweet)
{unique(gsub("[()]", "",
stringr::str_extract_all(tweet,
"\([^\)]+\)"))[[1]])}
```

Şekil 2. Sözcük türü etiketlerinin çıkarılması

Etiketlerin her bir tweette varlık ve yokluk durumunu gösteren bir *varlık-yokluk* (presence-absence) matrisi oluşturulmuřtur. Şekil 3'teki kod bloęunda bu işlem gösterilmektedir.

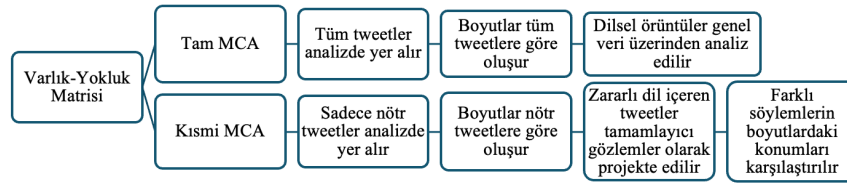
```
binary_matrix <- data.frame(matrix("A", nrow =
nrow(data), ncol = length(pos_tags)))
colnames(binary_matrix) <- pos_tags
for (i in seq_len(nrow(data))) {
  tags <- extract_pos_tags(data$text[i])
  for (tag in tags) {
    if (tag %in% pos_tags) {
      binary_matrix[i, tag] <- "P"}}}
```

Şekil 3. Varlık (P) ve yokluk (A) matrisinin oluşturulması

Uygulanan işlemler sonucunda, 61 POS etiketinden oluşan ve 60.068 tweet içeren bir kategorik matris oluşturulmuştur. Bu matris, her tweet için hangi sözcük türü etiketlerinin yer aldığı kategorik bir gösterimle sunar (Tablo 1).

5 MCA Analizine Giriş ve Veri Hazırlığı

Varlık-yokluk matrisi oluşturulduktan sonra tweetler, nötr, saldırgan ve nefret dili etiketlerine göre ayrılmıştır. Bu ayrım sayesinde, farklı etiketler arasında gözlemlenen dilsel örüntülerin analiz edilmesi mümkün hale gelmiştir. Analiz sürecinin başlangıcında iki yöntem uygulanmıştır: Tam MCA ve Kısmi MCA (Şekil 4). Tam modelde tüm veriler boyut oluşturma sürecine dâhil edilmiştir. Ancak nihai analizde Kısmi MCA tercih edilmiştir. Bu modelde yalnızca nötr tweetler analizde aktif olarak yer almış; zararlı dil içeren tweetler ise *tamamlayıcı gözlemler* (supplementary individuals) olarak yerleştirilmiştir. Bu yöntem ile nötr veriler temel alınarak boyutlar yapılandırılmış ve zararlı içerikler bu yapıya göre konumlandırılmıştır. Bu yaklaşım, zararlı dilin genel yapıyı bozmadan analiz edilmesini sağlamış; dilsel desenlerin daha belirgin biçimde incelenmesine olanak tanımıştır. Böylece, farklı söylem türleri arasındaki yapısal benzerlikler ve ayrışmalar net bir biçimde gözlemlenebilmiştir.



Şekil 4. Tam ve Kısmi MCA çalışma prensibi

Analizden önce veri seti, uygun formata dönüştürülmüştür (Şekil 5). İlk adımda, `cbind` fonksiyonu kullanılarak kategorik matris, etiketler ve tweet uzunluğu tek bir veri çerçevesi içinde birleştirilmiştir (`combined`). Bu birleşik

yapı, hem analiz sürecinin daha düzenli ilerlemesini saęlamıř hem de iřlem adımlarının aık biimde izlenmesine imkân tanımlıřtır.

```
combined <- cbind(binary_matrix, label = data$label,
tweet_length = data$tweet_length)
```

řekil 5. Veri setinin bileřik yapısı

Veri, ardından etiketlerine gre u gruba ayrılmıřtır: ntr, nefret ve saldırgan. Bu sınıflandırma, řekil 6’da da grlebileceęi zere, etiketler arasındaki iliřkilerin daha ayrıntılı řekilde incelenmesine zemin hazırlamıřtır.

```
neutral <- combined %>% filter(label == 0)
offense <- combined %>% filter(label == 1)
hate <- combined %>% filter(label == 2)
```

řekil 6. Ntr, nefret ve saldırgan kategorilerinin ayrılması

6 Boyutlar ve Veri Grselleřtirme

İlk MCA analizi, temel boyutların belirlenmesi ve bu boyutlarla etiketler arasındaki iliřkinin ortaya konulması amacıyla gerekleřtirilmiřtir (řekil 7). Nefret ve saldırgan dil grupları, sıralı řekilde ntr grupla aynı veri erevesine eklenmiř; bu iřlem, satırların alt alta eklenmesini saęlayan rbind fonksiyonu ile gerekleřtirilmiřtir. Ardından, bu gruplar ind.sup parametresi kullanılarak ana analiz dıřında tamamlayıcı gzlem olarak tanımlanmıřtır.

```
mca_SUP <- rbind(hate, offense)
mca_MAIN <- rbind(neutral, mca_SUP)
mca_results <- MCA(mca_MAIN %>% select(-tweet_length,
-label), ncp = 5, ind.sup = (nrow(neutral) +
1):nrow(mca_MAIN), graph = FALSE)
```

řekil 7. MCA ana analizinin yrtlmesi

İkinci ařamada, tweet uzunluęunun dilsel rntler zerindeki etkisi incelenmiřtir. Uzun tweetlerin daha fazla szck tr etiketi iermesi nedeniyle, tweet uzunluęu anlamlı bir deęiřken olarak deęerlendirilmiř ve quanti.sup parametresiyle analizde tamamlayıcı niceliksel deęiřken olarak kullanılmıřtır (řekil 8). Bu amala, yalnızca ntr tweetlerin yer aldıęı veriyle ikinci bir MCA analizi gerekleřtirilmiřtir. Bu analizde, etiket stunu ıkarılmıř; tweet uzunluęu ise ayrı bir niceliksel deęiřken olarak analiz edilmiřtir.

```
tweet_length_cor <- MCA(neutral %>% select(-label),
ncp = 5, quanti.sup = ncol(neutral) - 1, graph =
FALSE)
```

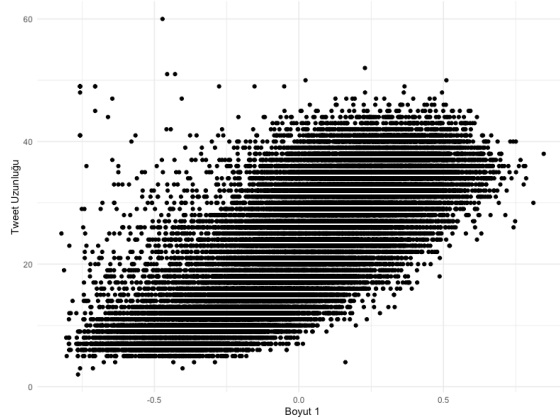
řekil 8. Tweet uzunluęu ve MCA boyutlarının korelasyonu

Elde edilen sonuçlar, tweet uzunluęu ile dilsel özellikler arasında belirgin bir ilişki olduğunu göstermektedir (Şekil 9 ve 10). Özellikle Boyut 1 ile tweet uzunluęu arasında yüksek bir korelasyon ($r = 0.76$) saptanmıştır. Bu bulgu, Clarke ve Grieve'in (2019) kısa sosyal medya metinlerinde dil özellikleri ile metin uzunluęu arasında gözlemledikleri pozitif ilişkiyi destekler niteliktedir. Korelasyon matrisinde, Boyut 1 ile güçlü bir ilişki; diğer boyutlarla ise daha zayıf ilişkiler gözlenmiştir (Boyut 2: -0.14, Boyut 3: 0.03, Boyut 4: 0.07, Boyut 5: -0.00).

```
ggplot(plot_data, aes(x = Dimension1, y =
tweet_length)) +
+ geom_point() +
+ labs(x = "Boyut 1", y = "Tweet Uzunluęu") +
+ theme_minimal()
```

Şekil 9. Boyut 1 ile tweet uzunluęu arasındaki ilişki

MCA analizinde tweet uzunluęu, niceliksel tamamlayıcı bir deęişken olarak kullanıldığında, daha uzun tweetlerin daha fazla dilsel özellik taşıdığı, kısa tweetlerin ise daha az özellięe sahip olduğu belirlenmiştir. Bu nedenle, diğer boyutların tweet uzunluęunun etkisinden bağımsız olarak incelenebilmesi amacıyla analizlerde Boyut 1 hariç tutulmuş ve odak Boyut 2'den Boyut 5'e kadar olan boyutlara yönlendirilmiştir. Boyut 1, tweet uzunluęunu temsil eden bir biçimsel özellik olarak önemli bir bulgu oluşturmakta; diğer boyutların dilsel deęişkeleri ile karışmadığı için ayrı deęerlendirilmiştir.



Şekil 10. Boyut 1 ile tweet uzunluęu (sözcük sayısı) arasındaki korelasyon

MCA sonuçlarının daha anlamlı biçimde yorumlanabilmesi amacıyla, temel boyutlara ait özdeęerlerin (eigenvalues) analizi gerçekleştirilmiştir (Şekil 11).

Özdeğerler, her bir boyutun açıklayıcı gücünü ve analizdeki önemini belirlemektedir. Boyutların taşıdığı bilginin büyüklüğünü değerlendirebilmek adına, özdeğerlerin incelenmesi gerekli görülmüştür. Bu çalışmada, ilk beş boyutun analizde öne çıktığı gözlemlenmiştir. Söz konusu boyutların, toplam varyansın %82,8'ini açıkladığı ve beşinci boyuttan sonraki boyutların bilgi katkısının anlamlı düzeyde azaldığı belirlenmiştir. Bu nedenle yorumlamalar, ilk beş boyutla sınırlandırılmıştır.

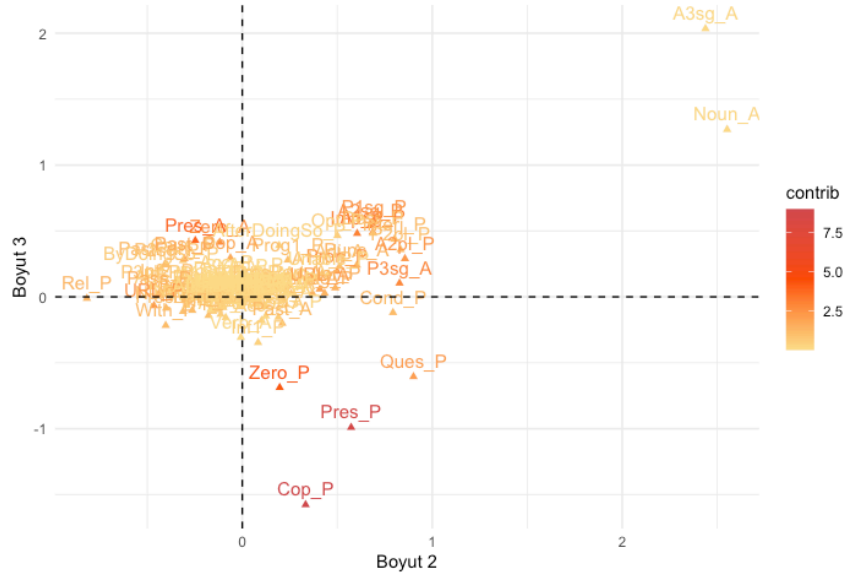
```
eig_values <- mca_results$eig[, 1]
eig_values
```

Şekil 11. MCA analizinde boyutlara ait özdeğerler

Her bir boyutun yorumlanabilirliğini artırmak amacıyla, ortalama katkı değerinin üzerinde kalan dilsel özellikler dikkate alınarak bu özelliklerin koordinat düzlemindeki konumları analiz edilmiştir (Şekil 12). Benzer koordinatlara sahip tweetlerin ortak dilsel örüntüler taşıdığı tespit edilmiştir. Bu durum, ilgili boyutların altında yatan iletişim işlevlerinin belirlenmesinde yol gösterici olmuştur. Örneğin, Boyut 2 ve Boyut 3 düzleminde yer alan dilsel özelliklerin dağılımı, Şekil 13'te görselleştirilmiştir. Bu görselleştirme, *ggplot2* ve *factoextra* paketleri aracılığıyla gerçekleştirilmiştir. Renk yoğunluğu yardımıyla, her bir değişkenin bu iki boyut üzerindeki açıklayıcı gücü vurgulanmıştır.

```
fviz_mca_var(mca_results, axes = c(2, 3), geom =
c("point", "text"), col.var = "contrib",
gradient.cols = c("#FFDD89", "#FC4E07", "#D34C4E"),
repel = FALSE, arrow = NULL) +
labs(x = "Boyut 2", y = "Boyut 3") +
theme_minimal()
```

Şekil 12. Boyut 2 ve 3'te sözcük türü katkılarının görselleştirilmesi



Şekil 13. 2. ve 3. boyutlar arasındaki değişken katkıları

Tablo 3'te, Boyut 2'ye katkıda bulunan dilbilimsel özelliklerin, her iki koordinat eksenini üzerindeki katkı değerleri ve bu katkıların azalan sıralamaları sunulmuştur. Elde edilen bulgulara göre, bazı dilbilimsel özellikler boyutun yapısına *varlıkları* (presence) yoluyla katkı sağlamış, bazıları ise *yoklukları* (absence) üzerinden bu katkıyı gerçekleştirmiştir. Yapılan analizler neticesinde, Boyut 2'nin pozitif koordinatlarında, etkileşimsel ve konuşma odaklı dilbilimsel özelliklerin yoğunlaştığı belirlenmiştir.

Tablo 3. Boyut 2'deki pozitif koordinatlardaki özellikler

| | |
|--------|---|
| B2 (+) | 3rd person singular possessive (4.5), Present tense (3.6), 2nd person plural (3.4), <i>absence</i> of URL (3.2), 1st person singular (3.2), Pronoun (3.2), Imperative (3), 2nd person singular (2.6), Question (2.4), Aorist (2.4), Negation (2.3), <i>absence</i> of Locative (2.2), 1st person singular possessive (2.2), <i>absence</i> of Punctuation (1.9), Conditional (1.9), <i>absence</i> of Adjective (1.4), Interjection (1.4), 1st person plural (1.3), <i>absence</i> of Genitive case (1.3), 2nd person plural possessive (1.3), <i>absence</i> of Passive (1.1), <i>absence</i> of Past participle (1), <i>absence</i> of Present participle (.9), Desiderative conditional (.9) |
|--------|---|

Negatif koordinatlarda konumlanan dilbilimsel özelliklerin, daha bilgilendirici ve haber metinlerine benzer bir anlatım biçimini yansıttığı gözlemlenmiştir. Bu

özelliklerin, kişisel bir üsluptan ziyade, olaylara ve eylemlere dair açıklayıcı ve gerçekçi bir raporlama diliyle ilişkilendiđi tespit edilmiştir. Yapılan bu çözümlene dođrultusunda, Boyut 2'nin, *etkileşimsel* ve *bilgilendirici* (interactive vs. informational) kutuplar arasında yapılandığı deđerlendirilmiştir (Tablo 4).

Tablo 4. Boyut 2'deki Negatif Koordinatlardaki Özellikler

| | |
|--------|---|
| B2 (-) | URL (3.7), Passive (2.3), Relative (2.1), Locative (1.8), <i>absence</i> of Pronoun (1.7), Number (1.6), <i>absence</i> of Present tense (1.6), Genitive case (1.5), Infinitive 2 (1.5), With (1.4), Past participle (1.4), Past tense (1.4), Present participle (1.4), <i>absence</i> of Negation (1.1), 3rd person plural possessive (1.1), 2nd person singular possessive (1.1), Instrumental (1), 3rd person singular possessive (1), <i>absence</i> of Imperative (.9), <i>absence</i> of Aorist (.9), Agentive (.8) |
|--------|---|

Bu boyut örnek olarak verilmiştir ve diđer boyutların da analizi yapıldığında Boyut 1: *metin uzunluğu* (text length), Boyut 3: *zamana odaklı* vs. *zamansız anlatım* (time-angled vs. timeless narration), Boyut 4: *yönlendirici* vs. *içgözlemsel konuşma* (directive vs. introspective speech), Boyut 5: *grup içi* vs. *grup dışı yönelim* (in-group vs. out-group orientation) olarak belirlenmiştir. Bu adlandırmalar, nitel analiz yoluyla yapılmış; boyutlara yüksek katkı sağlayan dilsel özelliklerin kümelenme biçimlerine dayanılarak belirlenmiştir.

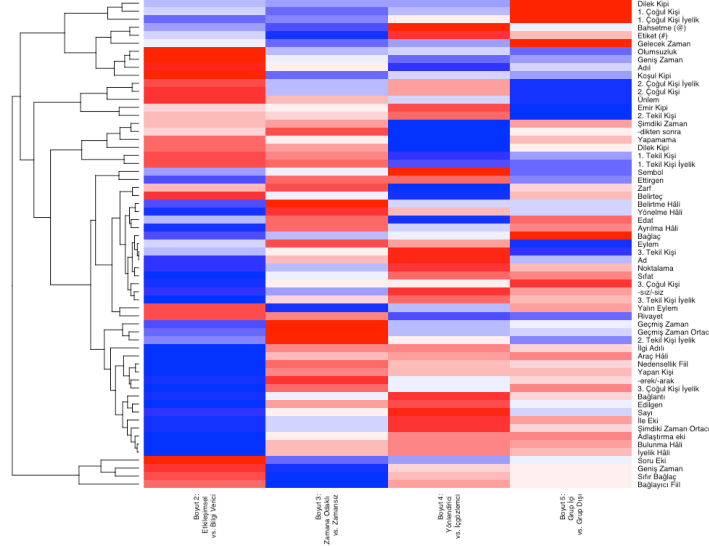
Sözcüksel etiketlerin dört temel boyut üzerindeki dağılımının görselleştirilmesi için, MCA sonuçlarından elde edilen deđişkenlerin (*var*) koordinatları (*coord*) kullanılmıştır. Anlamalı bir görselleştirme sağlayabilmek adına, yalnızca varlık (presence) bilgisi içeren ve "_P" etiketiyle biten sözcüksel ögeler bu analiz kapsamına alınmıştır (Şekil 14).

```
heatmap_data <- mca_results$var$coord
heatmap_data <- heatmap_data[grep("_P$",
rownames(heatmap_data)), 2:5]
rownames(heatmap_data) <- pos_tag_tr[gsub("_P$", "",
rownames(heatmap_data))]
heatmap_data <- as.matrix(heatmap_data)
```

Şekil 14. Dört boyutta dilsel özelliklerin ısı haritası (*heatmap*)

Şekil 15, sağda listelenen dilsel özelliklerin dört boyut üzerinde nasıl kümelendiđini ve bu kümelerin nasıl yapısal örüntüler oluşturduđunu gösteren bir *hiyerarşik kümeleme* (hierarchical clustering) ve *ısı haritası* (heatmap) sunmaktadır. Her bir boyut, altında yatan bir dilsel karşıtlığı temsil etmektedir (örneğin "zamana odaklı" vs. "zamansız" anlatım). Bu haritada satırlar dilsel özellikleri ve sütunlar ana boyutlarını temsil eder. Renk yoğunluğu ise her özelliğin ilgili boyuttaki katkı düzeyini göstermektedir. Bu görselleştirme aracı, onlarca dilsel özelliđi daha etkili bir biçimde anlamlandırmayı sağlar. Bu araç, özellikler arası benzerlikleri ve gizli yapıların keşfini kolaylaştırırken dođal dil

işleme, metin analizi ve söylem çözümlemelerinde etkili bir araç olarak değerlendirilebilir.



Şekil 15. Dört boyutta dilsel özelliklerin hiyerarşik kümeleme ile gösterimi

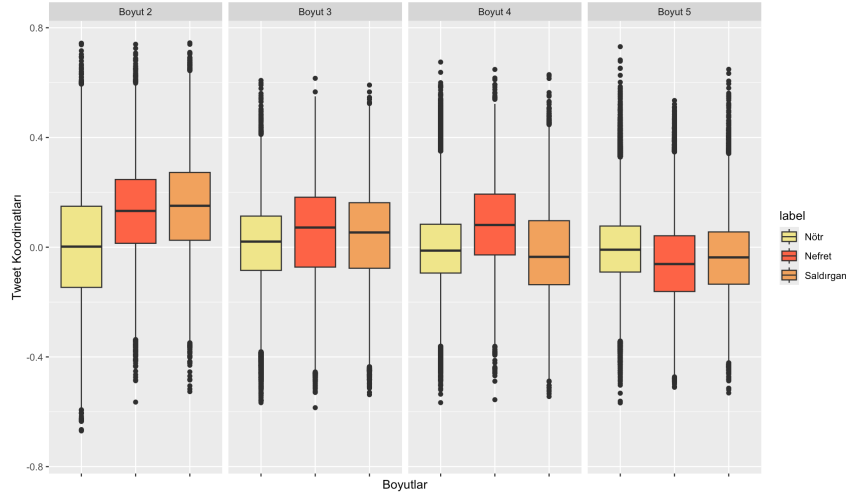
MCA boyutları ile etiketler (nötr, saldırgan, nefret) arasındaki ilişkilerin incelenmesi amacıyla, boyut koordinatlarının *kuşu grafikleriyle* (boxplot) görselleştirildiği bir diğer analiz gerçekleştirilmiştir (Şekil 16). Bu analizde, tamamlayıcı gözlemler olan nefret ve saldırgan tweetler ile nötr tweetler arasındaki dağılım farkları boyutlar üzerinden değerlendirilmiştir.

```
ggplot(mca_scores_long, aes(x = label, y =
MCA_Scores, fill = label)) +
  geom_boxplot() +
  labs(x = "Boyutlar", y = "Tweet Koordinatları") +
  scale_fill_manual(values = group_colors, labels =
group_labels) +
  scale_y_continuous(limits = c(-0.75, 0.75)) +
  theme_gray() +
  theme(axis.text.x = element_blank(),
legend.position = "right") +
  facet_wrap(~ Dimension, ncol = 4)
```

Şekil 16. Boyut değerlerinin kuşu grafiği

Bu analizde, tamamlayıcı gözlemler olan nefret ve saldırgan kategoriler ile MCA’da oluşturulan ana boyutlar arasındaki farklar, her bir boyutun koordinat deęerleri üzerinden deęerlendirilmiřtir. Elde edilen bu koordinatlar, kutu grafikleri aracılıęıyla görselleřtirilmiř ve farklı etiket gruplarının boyutlardaki daęılımları karřılařtırılmal olarak sunulmuřtur. Böylece, tamamlayıcı gözlemler ile ana boyut yapısı arasındaki uyum ve ayrıřmalar açık biçimde ortaya konmuřtur. Bu yaklařım, nefret söylemi ve saldırgan dilin tařıdığı dilsel örüntülerin, nötr dil yapıları ile nasıl bir karřıtlık oluřturduęunu göstermek açısından anlamlı bir çerçeve sunmaktadır. Ayrıca, verinin çok boyutlu yapısı içerisinde yer alan iliřkisel desenlerin daha derinlemesine kavranmasına katkı saęlamaktadır.

Görselleřtirme sürecinde kullanılan `geom_boxplot()` her bir etiket grubunun medyanını, çeyrek deęerlerini (IQR) ve uç deęerlerini (outliers) içeren kutu grafiklerini üretmiř; bu sayede gruplar arası daęılımlar ayrıntılı biçimde gösterilmiřtir. `facet_wrap()` fonksiyonu ise analizdeki boyutlara göre veriyi alt gruplara ayırarak her boyut için ayrı grafik panelleri oluřturulmasına olanak tanımuřtur. Böylece, örneęin Boyut 2 ve Boyut 3 gibi farklı boyutlara iliřkin örüntüler, kategoriler düzeyinde daha açık řekilde karřılařtırılabilmiřtir. řekil 17’de görüldüęü üzere, nötr tweetlerin daha çok bilgilendirici dilsel özelliklerle iliřkilendięi; buna karřılık, nefret ve saldırgan dil içeren tweetlerin ise etkileřimsel özellikler tařıdığı gözlemlenmiřtir. Bu farklılıklar özellikle Boyut 2 üzerinde belirgin bir biçimde ortaya çıkmaktadır.



řekil 17. Boyutlara göre nötr, saldırgan ve nefret içeren tweetlerin kutu grafikleri

Veri setinde gözlemlenen biçimsel değişkelerin zaman içindeki seyrini daha sağlıklı bir biçimde analiz edebilmek amacıyla, dilsel boyutların zamansal evrimi incelenmiştir. Bu analizde hem ana MCA hem de tamamlayıcı gözlemlerden elde edilen boyutsal konum değerleri temel alınarak, her bir tweetin çok boyutlu MCA uzayındaki konumu hesaplanmıştır. Elde edilen bu koordinat değerleri kullanılarak 2009–2021 yılları arasındaki dönem boyunca, her bir MCA boyutunda meydana gelen değişimler görselleştirilmiştir.

Zamana bağlı biçimsel eğilimlerin izlenebilmesi için tweetler, etiket gruplarına göre ayrıştırılmış ve her kategoriye ait değişim eğrileri ayrı ayrı incelenmiştir. Bu yöntem, farklı söylem biçimlerinin birbirine karışmadan izlenebilmesini sağlamak ve her kategoriye özgü biçimsel örüntülerin korunmasına olanak tanımaktadır.

Zamansal değişimlerin analizinde dikkat edilmesi gereken önemli bir husus, kısa vadeli dalgalanmaların etkisinin azaltılması ve daha belirgin uzun dönemli eğilimlerin ortaya çıkarılmasıdır. Bu doğrultuda, analizde *kayan pencere düzleştirme* (rolling window smoothing) yöntemi uygulanmıştır. Bu yöntem, zaman serisinde yer alan rastlantısal oynaklıkların etkisini azaltarak daha istikrarlı ve yorumlanabilir örüntülerin ortaya konmasını mümkün kılar. Kullanılan pencere boyutu, düzleştirmenin derecesini belirleyici bir faktör olduğundan analiz bağlamında dikkatle seçilmiştir. Örneğin, 3 günlük küçük pencereler, kısa vadeli değişimlere duyarlı bir yapı sunarken; 30 günlük veya haftalık pencereler, daha uzun dönemli yapısal eğilimlerin gözlemlenmesine olanak tanımaktadır. Bu çalışmada, veri sıklığının görece yüksek olması nedeniyle, her MCA boyutu için 30 günlük *kayan ortalama* (moving average, win = 30) hesaplanmıştır. Böylelikle, her döneme ait boyutsal konum değerleri ortalanarak verideki dalgalanmalar azaltılmış ve ortaya çıkan desenlerin daha pürüzsüz ve anlamlı hale gelmesi sağlanmıştır (bkz. Şekil 18).

```
mca_dimension_over_time <- function(mca_scores_main,
mca_scores_sup, dim_num, win=30) {
  col_name <- paste("Dim", dim_num, sep = " ")
```

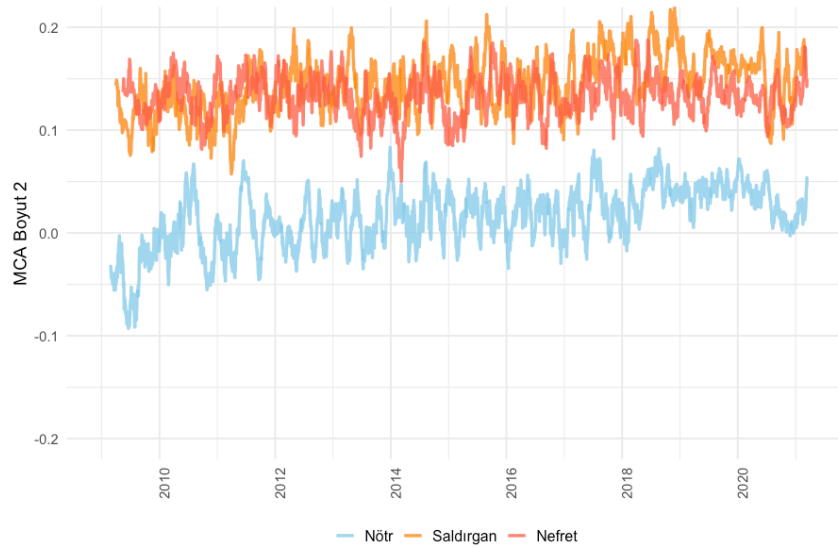
Şekil 18. Zamana göre MCA boyutlarının değişimini hesaplayan fonksiyon

Ortalamanın hesaplanmasında kullanılan bir diğer kritik parametre, kayan pencere fonksiyonundaki hizalama biçimidir. Bu bağlamda, align = "center" ayarı tercih edilmiştir. Bu parametre, her veri noktasının hem öncesindeki hem de sonrasındaki gözlemlerle birlikte değerlendirilmesini sağlayarak, pencerenin ortalanarak uygulanmasına olanak tanımaktadır (Şekil 19). Bu sayede, her veri noktasının etrafındaki çevresel eğilimlerin daha dengeli ve temsil gücü yüksek bir biçimde yansıtılması mümkün olmuştur.

```
mutate(y_smoothed = zoo::rollmean(y, k = win, fill =
NA, align = "center")) %>%
ungroup()
```

Şekil 19. Zamana baęlı deęerlerin hareketli ortalama ile dengelenmesi

Analiz bulguları, nötr dile ait biçemsel örüntülerin, problemlili dil türlerine (nefret söylemi ve saldırgan dil) kıyasla daha belirgin bir biçimde ayrıştığını ortaya koymuştur. Bununla birlikte, zaman içinde gözlemlenen genel biçemsel eğilimlerin görece durağan kaldığı ve yalnızca dar bir deęişim aralığında dalgalandığı belirlenmiştir (Şekil 20). Bu durum, dilsel yönelimlerde belirgin bir yapısal dönüşümün gerçekleşmediğine işaret etmektedir. Her bir etiket grubu, dönemsel olarak kısa vadeli deęişkeler sergilemiş olsa da genel biçemsel desenlerin büyük ölçüde tutarlılığını koruduğu ve güçlü yönlü bir eğilim göstermediği gözlemlenmiştir.



Şekil 20. Boyut 2'ye göre nötr, saldırgan ve nefret söylemlerinin zaman içindeki eğilimleri

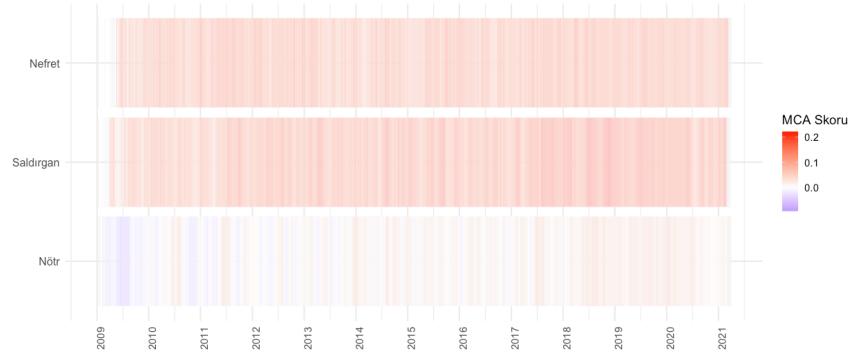
Benzer bir analiz, ısı haritası kullanılarak da görselleştirilmiştir. Çizgi grafikler, zaman içindeki eğilimleri daha ayrıntılı biçimde izlemeye olanak tanırken; ısı haritaları, büyük veri setlerinin daha yoğun ve etkili bir biçimde görselleştirilmesi açısından avantaj sağlamaktadır. Her iki görselleştirme türü de benzer temel parametrelerle oluşturulmakla birlikte, ısı haritasına özgü bazı görsel ve teknik tercihler, anlamlı sonuçlar elde edebilmek açısından önem arz etmektedir. Bu bağlamda, renk skalasının seçimi, veri yorumlanabilirliğini

doğrudan etkileyen faktörlerden biri olarak öne çıkmaktadır. Şekil 21'de kullanılan `scale_fill_gradient2()` fonksiyonu aracılığıyla hem sezgisel hem de erişilebilir bir renk paleti oluşturularak verinin anlamlandırılması desteklenmiştir.

```
scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0, na.value = "grey90", name = "MCA Skoru") +
scale_x_date(date_breaks = "1 year", date_labels = "%Y") +
```

Şekil 21. Analiz sonuçlarının zamansal renk kodlamasıyla gösterimi

Aynı şekilde, görselin yatay ekseninde kullanılan zaman çözünürlüğü (`scale_x_date`) de, veri setinin frekansına uygun biçimde belirlenmiştir. Bu sayede, ekran yoğunluğu azaltılmış ve okuyucunun eğilimleri daha rahat takip edebilmesi sağlanmıştır. Örneğin, Şekil 20'de iki yıllık aralıklarla sunulan zaman dilimleri tercih edilirken; Şekil 22'de tüm yıllar aynı görselde bütüncül bir biçimde gösterilmiştir. Bu tür teknik tercihler, ısı haritalarının analitik değerini ve yorumlanabilirliğini doğrudan etkilemektedir.



Şekil 22. MCA Boyut 2'ye göre nötr, saldırgan ve nefret söylemlerinin zaman içindeki eğilimlerinin ısı haritası

Yapılan bu analiz adımları ve kullanılan görselleştirme teknikleri, biçimsel örüntülerin zaman içinde ve kategoriler arasında karşılaştırmalı olarak izlenmesine olanak tanıyarak, çok boyutlu dilsel verilerin sistematik ve anlaşılır biçimde çözümlenmesini mümkün kılmıştır.

7 Tartıřma ve Sonu

Bu alıřma, MDA erevesinde MCA yntemini kullanarak Trke sosyal medya metinlerindeki dilsel deėiřkelerin temel boyutlarını ve nefret sylemi ile saldırgan dil kullanımına iliřkin belirgin rntleri incelemiřtir. Bulgular, nefret dili, saldırgan dil ve ntr sylemde dilbilgisel yapıların eř-oluřum rntlerini ve dilsel deėiřke rntlerini ortaya koymuřtur. Yntembilimsel bir ğretici ara retmeyi ve MCA ve grselleřtirme aralarını tanıtmayı amalayan bu alıřmada, yntemin uygulanması sırasında alınan kararlar ve analiz srecine zg tercihlerin, ortaya ıkan rntlerin biimlenmesinde nasıl etkili olabileceėi gsterilmiřtir. Bu baėlamda, bulguların yorumlanması srecinde veri hazırlıėı, belirlenen parametreler, arařtırmacının yaptıėı tercihler ve grselleřtirme aralarının oluřturulması ve sunumuna dair alınan kararlar, yntem ve kullanılan R paketlerinin iřleyiřinden ayrı dřnlemez bir btn teřkil etmektedir.

Analiz sreci boyunca, her bir boyutun temsil ettiėi dilsel zellikler sistematik bir yaklařımla deėerlendirilmiř; bu deėerlendirmeler, zellikle kısa metinlerdeki dil kullanımına iliřkin daha derinlemesine bir anlayıř geliřtirilmesine katkı saėlamıřtır. Metin uzunluėu, etkileřimsel vs. bilgilendirici slup, zamana odaklı-zamansız anlatım, ynlendirici-igzlemsel konuřma, grup ii-grup dıřı ynelim gibi temel boyutlar ne ıkmıřtır. Bu boyutları oluřturan dilsel zellikler, ntr dil ile saldırgan ve nefret dili arasındaki benzerlik ve farklılıkları ortaya koyma gcne sahiptir. rneėin, nefret syleminin emir kipinde ve doėrudan hedefe ynelik dil ierdiėi, ntr dilin ise daha formal, bilgilendirici bir yapı sergilediėi belirlenmiřtir. Zamansal deėiřim analizi, szck kullanımının zamanla deėiřtiėini ancak dilbilgisel yapıların sabit kaldıėını ortaya koymuřtur. Bulgular, nefret sylemi tespitinde dilbilgisel yapıların dikkate alınmasının nemini vurgulamakta ve MDA temelli dilsel temsillerin makine ėrenimi modellerine katkı saėlayabileceėini nermektedir.

alıřmada uygulanan yntembilimsel yaklařımlar, elde edilen sonuların gvenilirliėini ve yorumlanabilirliėini nemli lde artırmıřtır. MCA, ilgili istatistiksel paketler aracılıėıyla uygulanmıř ve analiz sonucunda ortaya ıkan boyutlar, eřitli aralarla grselleřtirilmiřtir. Bu yazının amalarından biri de grselleřtirme modellerinin analiz sonularının yorumlanmasındaki roln ortaya koymaktır. MCA ve biemsel deėiřke analizi ve grselleřtirilmesi konularında Clarke ve Grieve'in (2017, 2019) ve Grieve'in (2023) alıřmaları nemli birer kaynak teřkil etmektedir. Bu nedenle, her ne kadar bu alıřmalar İngilizce dilinde analizler sunsa da bu eserlere bařvurulması faydalı olacaktır.

Veri grselleřtirme srecinde, ncelikle *ggplot2* olmak zere eřitli aralar kullanılarak analiz ıktıları desteklenmiř ve boyutlar arası iliřkiler ile veri setlerindeki daėılımlar grafiksel gsterimlerle aık ve sistematik bir biimde sunulmuřtur. Uygulanan grselleřtirme yntemleri, hem dilbilgisel eř-oluřum rntlerinin hem de biemsel farklılıkların yorumlanmasını kolaylařtırarak bulguların daha anlaşılır ve anlamlı kılınmasına katkı saėlamıřtır. Roemling,

Winter ve Grieve'in (2025) belirttiđi üzere, tercih edilen görsel form ve tasarım ne olursa olsun, tüm görselleştirmeler belirli bir amaca hizmet etmektedir. Bu amaç doğrultusunda, alanımızda R ve *ggplot2*'nin yaygınlığı ile dilbilimin çeşitli alt alanlarında tür ve biçimsel analizlerin önemi dikkate alınarak, bu makalede dilbilim için R ile dilsel deđişikelerin görselleştirilmesine yönelik uygulamalı ve gerekli bir giriş sunulmaktadır. Bunun yanında, dilbilimsel uygulamalarla R diline yönelik pratik bir başlangıç için Winter (2019) kaynağına başvurulması önerilmektedir.

Ayrıca, bu çalışma MCA yönteminin kısa ve bağlamdan kopuk sosyal medya metinlerinde nasıl etkin bir biçimde uygulanabileceğini ortaya koyarak, geleneksel frekansa dayalı yöntemlerin yetersiz kaldığı durumlarda alternatif bir analiz yaklaşımı sunmaktadır. Bu yönüyle, çalışmanın hem yöntemsel hem de betimleyici olarak Türkçe dilbilim literatürüne özgün bir katkı sağladığı düşünülmektedir.

Yazar Katkıları: Bu metin tek yazar tarafından yazılmıştır. Bu metindeki Türkçe analizler ve yöntem anlatımı ile oluşturulan görseller tek yazar tarafından hazırlanmıştır.

Sunum beyanı ve doğrulama: Bu çalışma daha önce başka bir yerde yayımlanmamıştır. Başka bir dergide değerlendirme sürecinde deđildir. Çalışmanın yayımlanması tüm yazarlar ve çalışmanın yapıldığı üniversitedeki/araştırma merkezindeki sorumlu makamlar tarafından örtük ya da açık olarak onaylanmıştır. Çalışma yayımlanmak için kabul edilirse, Dilbilim Araştırmaları Dergisi'nin yazılı izni olmadan başka bir basılı ya da elektronik ortamda Türkçe ya da başka bir dilde aynı biçimde yayımlanmayacaktır.

Çıkar Çatışması Beyanı: Yazar, kurum, kuruluş ve kişiler ile bu çalışmayı etkileyebilecek mali ve akademik çıkar çatışması bulunmamaktadır.

Veri Kullanımı: Bu araştırmada veri kullanılmıştır.

Etik Onay/Katılımcı Onamı: Çalışmada etik onaya ihtiyaç bulunmamaktadır.

Maddi Destek: Bu araştırma, Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) 2219- Yurt Dışı Doktora Sonrası Araştırma Burs Programı tarafından desteklenmiştir. Bu çalışmada yer alan bulgular ve sonuçlar yalnızca yazara ait olup, TÜBİTAK'ın görüşlerini yansıtmak zorunda deđildir.

Kaynaklar

- Akın, A. A. (2023, 2 Ocak). *zemberek-python* (Sürüm 0.2.3) [Bilgisayar yazılımı]. PyPI. <https://pypi.org/project/zemberek-python/>
- Balcı, H. A. (2020). Varyant, deđişken ve varyasyon dilbilimi. *Littera Turca Journal of Turkish Language and Literature*, 6(3), 301-317. <https://doi.org/10.20322/littera.740237>

- Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.
- Biber, D. (1989). A typology of English texts. *Linguistics*, 27(3), 3–43. <https://doi.org/10.1515/ling.1989.27.3.3>
- Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.
- Biber, D. (2015). *Genre- and register-related discourse features in contrast*. M. Lefer ve S. Vogeleeer (Eds), *Genre- and register-related discourse features in contrast* (ss. 1–20). John Benjamins Publishing Company.
- Biber, D., & Conrad, S. (2019). *Register, genre, and style*. Cambridge University Press.
- Biber, D., & Hared, M. (1994). Linguistic correlates of the transition to literacy in Somali: Language adaptation in six press registers. D. Biber ve E. Finegan (Ed.), *Sociolinguistic perspectives on register* (ss. 182–216). John Benjamins Publishing Company.
- Biber, D., Davies, M., Jones, J. K., & Tracy-Ventura, N. (2006). Spoken and written register variation in Spanish: A multi-dimensional analysis. *Corpora*, 1(1), 1–37. <https://doi.org/10.3366/cor.2006.1.1>
- Clarke, I., & Grieve, J. (2017). Dimensions of abusive language on Twitter. *First Workshop on Abusive Language Online* (ss. 1–10). Association for Computational Linguistics. <https://aclanthology.org/W17-3000>
- Clarke, I., & Grieve, J. (2019). Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. *PLOS ONE*, 14(9), e0222062. <https://doi.org/10.1371/journal.pone.0222062>
- Çöltekin, Ç., (2020). A Corpus of Turkish Offensive Language on Social Media. *12th International Conference on Language Resources and Evaluation*. <https://coltekin.github.io/offensive-turkish/troff.pdf>
- Erdoğan-Öztürk, Y. & Işık Güler, H. (2020). Discourses of exclusion on Twitter in the Turkish Context: #ülkemdesuriyeliistemiyorum (#idontwantsyriansinmycountry). *Discourse, Context & Media*, 36, 100400.
- Grieve, J. (2023). Register variation explains stylometric authorship analysis. *Corpus Linguistics and Linguistic Theory*, 19(1), 47-77. <https://doi.org/10.1515/cllt-2022-0040>
- Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. Edward Arnold.
- Husson, F., Josse, J., Le, S., & Mazet, J. (2017). *FactoMineR: Multivariate exploratory data analysis and data mining* (ss. 1–96). <https://cran.r-project.org/web/packages/FactoMineR/FactoMineR.pdf>
- Kim, Y., & Biber, D. A. (1994). Corpus-based analysis of register variation in Korean. D. Biber ve E. Finegan (Ed.), *Sociolinguistic perspectives on register* (ss. 157–181). John Benjamins Publishing Company.

- Le Roux, B., & Rouanet, H. (2010). *Multiple correspondence analysis*. Sage Publications. <https://doi.org/10.4135/9781412993906>
- Özdüzen, Ö., Korkut, U., & Özdüzen, C. (2021). 'Refugees are not welcome': Digital racism, online place-making and the evolving categorization of Syrians in Turkey. *New Media & Society*, 23(11), 3349-3369. <https://doi.org/10.1177/1461444820956341>
- Roemling, D., Winter, B., & Grieve, J. (2025). Visualizing map data for linguistics using ggplot2: A tutorial with examples from dialectology and typology. *Journal of Linguistic Geography*, 1-15. <https://doi.org/10.1017/jlg.2024.11>
- Yüceol Özezen, M. (2021). Dilbilimsel tipoloji ve Türkçe. *Türklük Bilimi Araştırmaları*, 49, 117-133. <https://doi.org/10.17133/tubar.696950>
- Winter, B. (2019). *Statistics for linguists: An introduction using R*. Routledge. <https://doi.org/10.4324/9781315165547>
- Toraman, Ç., Şahinuç, F., & Yılmaz, E. (2022). Large-Scale Hate Speech Detection with Cross-Domain Transfer. *Thirteenth Language Resources and Evaluation Conference* (ss. 2215-2225). Marseille, France. <https://aclanthology.org/2022.lrec-1.238/>