Pamukkale Üniversitesi Eğitim Fakültesi Dergisi
66, 329-347, 2026

Türkçe Klinik Okuma Değerlendirmesi için R Tabanlı Nicel Hesaplamalı Metin Analizi Aracı

Esmehan ÖZER[1*,2,3,4], Sema ACAR-ÜNALGAN[2,3,4,5], Hazal ARTUVAN-KORKMAZ[2,3,4,6], Rahime Duygu TEMELTÜRK[3,4,7,8,9], Özgür AYDIN[2,3,4,8,10]

[1] Gazi Üniversitesi, Gazi Eğitim Fakültesi, Özel Eğitim Bölümü, Türkiye
[2] Nörobilim ve Nöroteknoloji Mükemmeliyet Ortak Uygulama ve Araştırma Merkezi (NÖROM), Türkiye
[3] Ankara Üniversitesi Beyin Araştırmaları Uygulama ve Araştırma Merkezi (AÜBAUM), Türkiye
[4] Ankara Üniversitesi Dil ve Tarih-Coğrafya Fakültesi Dilbilim Laboratuvarı (DiLab), Türkiye
[5] İzmir Bakırçay Üniversitesi, Sağlık Bilimleri Fakültesi, Dil ve Konuşma Terapisi Bölümü, Türkiye
[6] Ankara Üniversitesi, Tıp Fakültesi, Temel Tıp Bilimleri Bölümü, Türkiye
[7] Ankara Üniversitesi, Tıp Fakültesi, Dahili Tıp Bilimleri Bölümü, Türkiye
[8] Ankara Üniversitesi Sağlık Bilimleri Enstitüsü, Disiplinlerarası Sinir Bilimleri Anabilim Dalı, Türkiye
[9] Ankara Üniversitesi, Otizm Uygulama ve Araştırma Merkezi, Türkiye
[10] Ankara Üniversitesi, Dil ve Tarih-Coğrafya Fakültesi, Dilbilim Bölümü, Türkiye

Özet: *Bu çalışma, okuma becerilerinin klinik değerlendirmesinde kullanılan metinlerin nicel dilbilimsel özelliklerini değerlendirmek üzere R programlama dili temelli bilgisayar destekli bir metin analiz aracı geliştirmek ve bu araçla yeni oluşturulan metinleri test etmeyi amaçlamaktadır. Bu bağlamda, araştırmacılar tarafından anlatı metni "Çifte Sürpriz" ve bilgilendirici metin "Deniz Yıldızları", dilbilimsel ve yapısal karmaşıklık açısından üç farklı düzeyde yazılmıştır. Ortografik benzerlik (OLD-20), bigram sıklığı, sözcük sıklığı, morfolojik karmaşıklık ve okunabilirlik gibi nicel parametreler, ilgili R paketleri ("tidytext", "vwr" ve "strngram" paketleri) kullanılarak toplam altı testte hesaplanmıştır. M2a ve M2b metinlerinin, dilbilimsel parametreler açısından birbirine en yakın metinler olduğu belirlenmiştir. Böylelikle, okuma becerilerinin değerlendirilmesinde kullanılacak metinlere ilişkin nicel kriterler belirlenmiş ve Türkçede ilk kez R dilinde bir metin analiz aracı geliştirilmiş; ayrıca, üretilen metinler bu geliştirilen araç kullanılarak test edilmiştir. Çalışma, nicel metin değerlendirmesi için veriye dayalı ve tekrarlanabilir bir yöntem sunarak araştırma ve tanılama uygulamaları için Türkçe okuma materyallerine katkı sağlamaktadır.*

Makale Bilgileri
Araştırma Makalesi
Gönderim Tarihi
14/01/2025
Kabul Tarihi
29/10/2025
Anahtar Kelimeler
Okuma, anlatı metni, bilgilendirici metin, metin benzerliği, metin okunabilirliği.

1. Giriş

Okuma, hedefe yönelik amaçlı bir beceridir (Kucer, 2014). Amaç ise kodu çözüp yazılı dili anlamadır (Hoover ve Gough, 1990). Bu kapsamda, okuma harf (letter), yazıbirim (grapheme), diyakritik (diacritic) ya da diğer yazılı işaretlerin beyin-okuma ağları tarafından işlemlenerek yorumlanması ve anlamlandırılması süreci olarak tanımlanmaktadır (Çetinkaya, 2010). Bu

* *Sorumlu Yazar: Esmehan ÖZER E-mail:esmehann@gmail.com Adres: Gazi Üniversitesi, Gazi Eğitim Fakültesi, Türkiye*

ISSN: 1301-0085 **e-ISSN:** 1309-0275 http://doi.org/10.9779/pauefd.1619827

süreç ilk olarak göz ve görsel sistemle başlayıp hızlı ve başarılı bir şekilde dil işlemleme sistemince devralınmaktadır (Perfetti, 1999, 2012). Bir başka deyişle okuma, okura görsel bilgi sağlanmasıyla başlamaktadır. Ardından okumaya özgü süreçlerde görsel girdinin dil girdisine dönüştürülmesi sağlanıp yazılı dili anlama amacıyla dilbilgisi ve anlam aktifleştirilmektedir (Perfetti, 1999). Böylece okumanın nihai amacı olan anlama için yazı sistemi ve bilişsel işlemlemeler etkin bir şekilde kullanılmaktadır. Günlük hayatta ve iş yaşamında ön koşul olan okuma, akademik başarıda çok güçlü bir role sahiptir (Yıldız, 2013). Bununla beraber okuma, eş zamanlı ve/veya birbirinin ardı sıra gerçekleşen duyusal, nörolojik ve bilişsel-dilsel işlemlerin eşgüdümünü gerektiren karmaşık bir beceridir (Rayner ve Reichle, 2010). Bu çok bileşenli karmaşık becerinin gerçekleşmesinde okunan metin ve sözcüklerin özellikleri de rol oynamaktadır. Alanyazın incelendiğinde okumayı etkileyen metin ve sözcük düzeyinde niceliksel ölçütler olduğu görülmektedir. Bunlar: (i) metnin okunabilirliği (readability) (Çetinkaya, 2010), (ii) metin türü (Baştuğ, 2012), (iii) sözcüklerin sıklığı (frequency) (Rayner, 1998), (iv) sözcüklerin uzunlukları (length) (Rayner vd., 2011), (v) ortografik komşuluk/yakınlık (Andrews, 1997), (vi) morfolojik/biçimbirimsel karmaşıklık (morphological complexity) (Levesque vd., 2021; Verhoeven vd., 2011) gibi çeşitli değişkenlerdir.

Metnin okunabilirliği, metnin okur tarafından ne kadar kolay ya da zor anlaşılır olduğudur (Ateşman, 1997). Metinlerin okunabilirliğinin hesaplanmasında kullanılan okunabilirlik formülleri, metnin yapısal özelliklerinden yola çıkarak metinleri okuma zorluklarına ya da kolaylıklarına göre aşamalı bir biçimde sınıflandırmayı hedefleyen kestirim araçları olarak ifade edilmektedir (Çetinkaya, 2010). Türkçe okuma metinlerinde Ateşman (1997), Bezirci ve Yılmaz (2010) ve Çetinkaya ve Uzun'un (2010) önerdiği farklı formüller mevcuttur [Milli Eğitim Bakanlığı (MEB), 2021]. Öte yandan bu formüllere ilişkin aynı metinler üzerinde farklı sonuçlar vermeleri, formüllerde hece, sözcük ve tümce sayılarının esas alınması gibi çeşitli eleştiriler yer almaktadır (MEB, 2021).

Okuma becerisini etkileyen bir başka değişken ise metnin türüne ilişkin yapısı olduğundan okura metin yapısının öğretiminin kritik olduğu belirtilmiştir (Hamzadayı ve Batmaz, 2022; Hebert vd., 2016). Bilgilendirici ve öyküleyici metin türleri ise metinlerin önemli ana biçimlerindendir (MEB, 2019). Öyküleyici metin, birbirleriyle ilişkili unsurların bazen açıkça bazen ise metne sindirilerek oluşturulduğu metindir. Bilgi verici metin ise belli bir konuda bilgi verip, düşünce aktaran ve öneri sunan yazı olarak tanımlanmaktadır (Başaran ve Akyol, 2009).

Bir sözcüğün dilde sıkça geçmesi, bu sözcüğün tanınmasını hızlandırmaktadır (Zevin, 2009). Örneğin, milyonda 50 kez görülen bir sözcük, milyonda 10 kez görülen bir sözcükten daha hızlı bir biçimde okunmaktadır. Dolayısıyla okurlar yüksek frekanslı sözcüklere düşük frekanslı sözcüklerden daha hızlı ve doğru tepki vererek okumalarını gerçekleştirmektedirler (Yap ve Balota, 2015). Erten ve diğerleri (2014), Türkçe sözcük okumada sözcük sıklığının düşük olmasının yaklaşık 92 milisaniyelik (ms) bir gecikmeye neden olduğunu tespit etmişlerdir.

Okumayı etkileyen bir başka değişken ise sözcük uzunluğu yani sözcüklerin harf sayısı olduğu görülmektedir. Uzun sözcüklerin daha uzun okuma sürelerine sahip oldukları bilinmektedir (New vd., 2006). Bir başka deyişle bir sözcüğün harf sayısı 8-13 harf arasındaysa bu durum sözcüğün okunmasını etkilemektedir. Dolayısıyla sözcük uzunluğu, okuma sırasında okurun sabitleme (fixation) sürelerini ve sekme (saccade) genliklerini etkileyen en önemli değişkenlerdendir. Okuma esnasında okurun gözleri uzun bir sözcük ile karşılaştığında, daha uzun süre sabitleme geçekleştirirken bu sözcükleri okumadan atlama olasılığı daha düşüktür. Dillerde, "için", "ile", "ve", "ama" gibi işlev sözcüklerine göre uzun sözcüklerde daha uzun

sabitleme yapılmakta, bu durum da doğal olarak okuma süresini arttırmaktadır (Rayner, 1998, 2009).

Okurların okuma sürecinde etkisi olan bir başka değişken ise ortografik komşuluk boyutudur (orthographic neighborhood size). Coltheart ve diğerleri tarafından ilk kez 1977'de ifade edilmiştir (Yarkoni vd., 2008). Ortografik komşuluk boyutu, bir sözcüğün tek bir harfi değiştirilerek aynı uzunlukta oluşturulan yeni sözcüklerin sayısı olarak tanımlanmaktadır. Örneğin "süt" sözcüğünden "süz", "sür", "süs", "sat", "set", "tüt" ve "küt" gibi yedi ortografik komşusu olan sözcük elde edilebilir. Birçok dilde okuma esnasında sözcüğün birçok ortografik komşusu olması durumunda sözcüklerin daha hızlı bir şekilde okunduğu raporlanmıştır (Andrews, 1997). Türkçede ise ortografik komşuluk boyutunun okuma üzerine herhangi bir etkisi olmadığı bulgulanmıştır (Erten vd., 2014). Öte yandan ortografik komşuluk boyutunun sadece tek bir harf değişikliği ile oluşturulması, bu durumun dışında ise sözcükteki harflerin yer değiştirmesi, sözcükteki bir harfin silinmesi ya da sözcüğe yeni bir harf eklenmesi gibi farklı işlemleri kapsamaması gibi sınırlılıklarından dolayı Yarkoni ve diğerleri (2008) yeni bir ortografik benzerlik ölçütü geliştirmişlerdir. "Ortografik Levenshtein uzaklığı-20" [orthographic Levenshtein distance-20 (OLD-20)] bir sözcükte en az sayıda yapılan harf değişikliği (yer değiştirme, ekleme, çıkarma) ile 20 yeni sözcük oluştururken yapılan değişiklik sayısının ortalamasını ifade etmektedir. Bir sözcüğün OLD-20 değeri ne kadar düşükse o kadar az harf değişikliği ile yeni sözcükler türetilebilmektedir. Çalışmalarda düşük OLD-20 değeri olan sözcükleri okumanın daha kolay olduğu belirtilmekle birlikte (Yarkoni vd., 2008), Türkçede bu durumun tersi gösterilmiştir (Erten vd., 2014). Dolayısıyla OLD-20'nin Türkçe okuma üzerinde kolaylaştırıcı bir etki yerine engelleyici bir etkiye sahip olduğu düşünülmektedir.

İki birimden oluşan bir dizi "bigram" olarak ifade edilmektedir. Söz gelimi, Türkçede masa sözcüğü "ma", "as" ve "sa" bigramlarını içermektedir. Bigram sıklığı (bigram frequency) ise bir ortografideki bir bigramın sıklığının ölçümüdür (Santangelo, 2023). Örneğin İngilizcede "th", "he" ve "in" bigramlarının "qu" ve "nk" bigramları ile kıyaslandığında metinlerde daha sık kullanıldığı görülmektedir. Bigram sıklığının, sözcüksel karar verme (lexical decision) (Muncer vd., 2014; Schmalz ve Mulatti, 2017) ve sözcükleri isimlendirme (word naming) (Muncer vd., 2014) görevlerinde etkilerinin olduğu ifade edilmektedir. Bigram sıklığının okumaya olan etkileri tartışılmaya devam ederken sesli okumayı hızlandırıcı etkisine ilişkin güçlü kanıtlar olduğu da raporlanmıştır (Schmalz ve Mulatti, 2017).

Morfolojik farkındalık (morphological awerness) ve morfolojik olarak karmaşık sözcüklerin tanımlanması arasında anlamlı bir ilişki bulunmaktadır (Carlisle, 2000). Dolayısıyla dilin en küçük birimi olan morfemlerin farkındalığının okumanın nihai amacı olan anlama becerisi ile ilişkili olup bu ilişki morfolojik olarak karmaşık olan sözcüklerden etkilendiği bulgulanmıştır (Liu vd., 2024). Örneğin İngilizce "work" sözcüğü "works", "workers" ve "working" gibi çekimlenebileceği gibi ön ve son ekler alarak "rework" ve "worker" olabilir. Aynı zamanda "work" sözcüğü, birleşik sözcük olup "workplace" olarak da karşımıza çakabilir (Borleffs vd., 2019). Bu sözcüklerin morfolojik karmaşıklıklarının çözümlenerek analiz edilebilmesi çocukların okuma ve anlama becerilerine yansımaktadır (Levesque vd., 2021; Verhoeven vd., 2011). Bu bağlamada ise metinlerde yer alan sözcüklerin morfolojik karmaşıklığa ilişkin özelliklerinin okumada önemli bir role sahip olduğu ifade edilebilir.

Okuma becerisinin değerlendirmesinde ve kazandırılmasında metinlerden yararlanılır (Milli Eğitim Bakanlığı [MEB], 2019). Metin, anlam açısından bütünlük taşıyan dil parçasıdır (Pilav ve Oğuz, 2013). Bilgi, duygu ve düşüncelerin çeşitli biçim, anlatım ve noktalama özelliklerine göre yerleştirildiği yapılar olarak da ifade edilmektedir (Güneş, 2013). Metinler, okumanın en önemli kaynak ve materyalleridir. MEB'e bağlı Talim ve Terbiye Kurulu Başkanlığının "Taslak

Ders Kitabı ve Eğitim Araçları ile Bunlara Ait Elektronik İçeriklerin İncelenmesinde Değerlendirmeye Esas Olacak Kriterler ve Açıklamaları" (2023) kitapçığında ders kitaplarında yer alacak metinlere ilişkin dil ve üslup, yazı dili standartlarına uygunluk ile anlam ve anlatım başlıklarında sunulan ölçütler yer almaktadır. Bu bağlamda metinlerde yabancı sözcüklerin kullanılmaması, sınıf seviyesine uygun ve sözvarlığını zenginleştirmeye yönelik dil kullanımı, anlatımın basit ve yalın olması, okumayı yavaşlatmaması ve kolay okunabilir nitelikte olması gerektiğine ilişkin açıklamalar bulunmaktadır. Zira alan yazında okuma becerisini etkileyen sözcük sıklığı, uzunluğu, ortografisi ve metnin okunabilirliği ile ilgili birçok değişken ifade edilmesine rağmen Türkçe metin yazarlarının öğrencinin düzeyine uygun metinleri bu değişkenlere ilişkin nasıl oluşturduklarına yönelik net bilgilere ulaşılamamaktadır. Metinlerin daha çok yazarların tecrübe ve hissiyatlarına göre yazıldığı ifade edilebilir. Öte yandan alan yazında metin zorluğu ve metin türlerinin yazımı konuları en çok çalışılması ihtiyaç olunan konular arasında yer almaktadır (Duran ve Kargın, 2022). Aynı zamanda alan yazında okumayı etkileyen faktörlerin metin bağlamında niceliksel olarak bir bütün halinde ele alınmadığı görülmektedir.

Yukarıda sıralanan okuma becerisinin hızını ve süresini etkileyen metin ve sözcük düzeyindeki niceliksel ölçütler bağlamında metinlerin oluşturulması okuma becerisinin klinik değerlendirilmesi açısından kritiktir. Bu ölçütler bütünselliğinde oluşturulmayan metinlerin okuma becerisinin klinik değerlendirilmesinde sorunlara yol açtığı düşünülmektedir. Bu ölçütler okuma sürecini doğrudan etkilemektedir. Bu nedenle okuma becerisi açısından problemler yaşayan popülasyonlar için bu ölçütler bağlamında oluşturulan metinlerin klinik değerlendirmede kullanılması önem arz etmektedir. Bu ölçütler gözetilmeden sadece bireyin sınıf düzeyi ve/veya yaşına uygun rastgele metinlerin kullanılması disleksi ve diğer okuma güçlüklerinde gözlemlenen halihazırda okuma becerisinin çok bileşenli karmaşık yapılanmasından kaynaklı heterojenite problemini artırıcı bir faktör olarak rol oynamaktadır. Bilgimiz dâhilince okuma becerisinin klinik değerlendirilmesinde yukarıda sıralanan niceliksel ölçütlerin tamamını kontrol altına alarak oluşturulan ve yaygınlıkla kullanılan Türkçe metinler mevcut değildir. Bu çalışmanın amaçlarından biri alanyazın ve klinik uygulamadaki bu eksikliği giderecek okumayı etkileyen bu niceliksel ölçütler bağlamında geliştirilen bilgilendirici ve öyküleyici olmak üzere iki metini alan yazına kazandırmaktır. Ancak önerilen bu iki metnin salt olarak klinik değerlendirmelerde kullanımının bireyin yaşı, sınıfı düzeyi gibi özellikleri değiştikçe ve okuma becerisinin tekrarlı aralıklarla-değerlendirilmesi gerektiğinde sınırlayıcı olacağı düşünülmektedir. Bu yüzden yukarıda sıralanan ölçütler bağlamında kullanılmak istenen metinleri bütünsel olarak değerlendirmeye olanak sağlayan araçlar ile bu sınırlılığın önüne geçmek mümkün olacaktır.

Kırkgöz ve Ünaldı (2012) anadili Türkçe olup yabancı dil olarak İngilizce öğrenen öğrenciler ile anadili İngilizce olan öğrencilerin, yazdıkları İngilizce metinler üzerinden derlem tabanlı, okunabilirlik, sözdizimi gibi çeşitli ölçütlere sahip çevrim-içi bir veritabanı olan coh-metrix kullanarak sözcüksel (lexical) ağlarını karşılaştırmışlardır. Çalışmada, coh-metrix'in geçerliliği de test edilmiştir. Çalışma sonucunda anadili Türkçe olup yabancı dil olarak İngilizce öğrenen öğrencilerin metinlerindeki cümlelerin sözcüksel olarak bağdaşıklığında sınırlılıklar olduğu tespit edilmiştir. Ayrıca coh-metrix'de kullanılan bazı göstergeler, anadili İngilizce olan öğrenciler ile sonradan İngilizce öğrenen öğrencilerin yazdıkları metinleri birbirinden ayırt edebilmiştir. Arnold ve diğerleri (2019) ise R dilinde farklı metin analiz paketlerini bir araya getirerek nasıl kullanılabileceklerini açıklayıp R dilinde farklı metin analiz paketlerinin metin analiz aracı olabilme potansiyelini vurgulamaktadırlar. Benzer şekilde Aziz ve diğerleri (2010) de okuma materyallerinin sadece genel okunabilirliğini belirlememiş, cümle ve sözcük zorluğuna ilişkin de bilgi sağlayacak daha kapsamlı bir analiz yaklaşımı ortaya koymuşlardır. Araştırmacılar, metin, cümle ve sözcük zorluğunu objektif ve güvenilir bir biçimde belirlemek

için bileşik hesaplama araçlarını önermektedirler. Dilbilim, eğitim bilimleri gibi alanlarda R dilinde oluşturulan farklı metin analiz paketlerinin objektif bir metin analiz aracı olarak kullanılması özel eğitim, dil ve konuşma terapisi, çocuk ve ergen ruh sağlığı ve hastalıkları, klinik psikoloji gibi okumanın klinik değerlendirilmesi ile ilgili alanlarda da kullanılabilecek potansiyelde olabileceğine esin olmaktadır.

Alan yazında okuma becerisinin klinik değerlendirmesinde kullanılan Türkçe metinlerin niceliksel ölçütlerle bütünsel olarak değerlendirilebilineceği bir araca henüz rastlanmamıştır. Ulusal alan yazında hem eğitsel hem de klinik amaçlı okuma performansının değerlendirilmesinde kullanılan, geçerliliği ve güvenilirliği kanıtlanmış standardize edilmiş birçok batarya ve/veya ölçek mevcuttur. Bu bataryalar okuma performansının çeşitli klinik popülasyonlarında değerlendirilmesi açısından oldukça kıymetli ve tanılamada yol gösterici kılavuz niteliğindedir. Bu çalışmada amacımız bu tarz bir ölçek geliştirmekten ziyade okuma değerlendirmesinde kullanılan metinlerin dilbilimsel özellikleri açısından uzmana bilgi verecek analizler yapan R program temelli bir analiz aracı (tool) geliştirmektir. Böylece uzman kullanmak istediği metinleri geliştirdiğimiz R paketi sayesinde analiz edebilecek (https://github.com/ozguraydin66/PreTXT/) ve ilgili dilbilimsel ölçütler hakkında bilgi alabilecektir. Dileyen uzmanlar bu çalışmada örnek ürün olarak sunduğumuz metinleri değerlendirmelerinde kullanabilecek ya da bu çalışmada sunduğumuz metin analizi aracı kodlarını (bkz. PreTXT paketi) kullanarak kendisinin kullanmak istediği herhangi bir metnin dilbilimsel özellikleri hakkında bilgi alabilecektir. Bu bağlamda, çalışmada geliştirilmesi planlanan araç bir klinik değerlendirme aracından ziyade kliniklerde kullanılan metinlerin dilbilimsel özellikleri hakkında uzmana/değerlendirmeciye rehber olabilecek bir araç olacaktır. Alan yazındaki standardize edilmiş ölçeklerden farklı olarak bu çalışmada geliştireceğimiz metin analizi aracı sayesinde, uzman/değerlendirmeci farklı dilbilimsel özelliklere sahip metinler oluşturabilecek ve eğitsel müdahalesi sırasında dilbilimsel özellikleri açısından farklı zorluk ve karmaşıklığa sahip metinleri kullanabilecektir. Dolayısıyla bu çalışmada, okuma becerisinin klinik değerlendirmesinde kullanılacak metinlere ilişkin niceliksel ölçütlere ihtiyaç olduğundan R dili yardımıyla metin analizi aracı geliştirilmesi ve üretilen metinlerin bu araç ile test edilmesi amaçlanmaktadır. Bu kapsamda aşağıdaki sorulara cevaplar aranmıştır:

(a) Metinleri niceliksel ve bütünsel olarak değerlendirilmesine yönelik kendini yenileyen bir metin analiz aracı geliştirilebilecek mi?

(b) Geliştirilen bu araç ile araştırmacılar tarafından yazılan metinler test edilebilinecek mi?

(c) Araç, okuma becerisinin klinik değerlendirmesi için, metin yazacak ya da kullanacağı metnin düzeyine ilişkin bilgi sahibi olmak isteyecek uzmanlara metne ilişkin niceliksel özellikler sunulabilecek mi?

(d) Araç, metinlerde sözlüksel (ortografik, sözcük sıklığı, harf-bigram sıklığı, biçimbirimsel karmaşıklık, sözcük uzunluğu) ve okunabilirlik ölçütlerine bağlı hesaplamalar yapılabilecek mi?

(e) Araştırmacılar tarafından geliştirilen metinlerin benzerlikleri nedir?

2. Yöntem

2.1. R Programlama Dili ve Metin Analizi

Hesaplamalı metin analizi (computational text analysis), özellikle uzun metinlerle ilgili olarak, manuel olarak gerçekleştirilmesi zaman alacak olan işlemlerin hem daha kısa zamanda gerçekleştirildiği hem de hata olasılığının düşürüldüğü bir ortam sağlamaktadır. Okuma becerisinin klinik değerlendirmesinde kullanılacak metinlere ilişkin olarak ortografik ve sözlüksel özelliklerle ilgili türlü niceliksel ölçütlerin hesaplanması, metinlerin okunabilirliklerinin hesaplanması, metinler arasındaki benzerliklerin ortaya konması manuel

olarak emek yoğun bir işlemi gerektirir. Bu çalışmada, R yazılımı kullanılarak hesaplamalı metin analiziyle ilgili genel adımlar ve işlemler hakkında genel bir bakış sunuyoruz.

Diğer programlama dillerinin tersine, R özellikle istatistiksel analiz için oluşturulduğu için R programlama dili veri bilimi uygulamaları için son derece uygun bir ortam sunmaktadır. Açık kaynak platformu olan R, çok çeşitli metin analizi paketleri içeren (Kwartler, 2017; Kumar ve Paul, 2016) ve oldukça geniş bir kullanıcı topluluğu olan bir programlama dilidir. Bu çalışmada söz konusu metin analiz paketlerini ve geliştirdiğimiz kodları kullanarak bir metin analizi örneği sunuyoruz. R paketleri, R'nin geniş kullanıcı topluluğu tarafından üretilen yazılım kütüphaneleri koleksiyonudur. Her paket, temel R dili ve çekirdek paketlerin işlevselliğini genişletir ve kullanıcılara kullanımı kolay birtakım fonksiyonlar sunar. Bu çalışmada, metin düzenlemelerinde tidytext paketi (Silge ve David, 2016) ve textTinyR paketi (Mouselimis, 2021) kullanılırken, ortografik hesaplamalarda vwr paketi (Keuleers, 2013) ve strngram (van Heuven, 2024) paketi kullanılacaktır.

Ayrıca istatistiksel analizler için stats paketi ve rstatix paketi (Kassambara, 2023) gibi paketlerden yararlanılacaktır. En iyi bilinen paket deposu olan Comprehensive R Archive Network (CRAN), şu anda 10.000'den fazla yayınlanmış pakete sahiptir ve stats paketi dışında yukarıda sıraladığımız tüm paketler CRAN kapsamındadır. R ortamından kolayca ve güvenli bir şekilde yüklenebilen bu paketler yeni analiz araçlarının geliştiricileri ve kullanıcıları arasında sağlam bir köprü oluşturarak, bilimsel iş birliği için çok uygun bir programlama ortamı sunmaktadır.

2.2. Metinlerin Oluşturulması

2.2.1. Metinlerin Yazımı

Çalışma yer alan anlatı türündeki "Çifte Sürpriz" adlı metin araştırmanın ilk ve üçüncü yazarı tarafından metin olay veya bilgi örüntüsü bozulmadan sözcük uzunluğu, morfolojik karmaşıklık, ortografik komşuluk değişkenleri bağlamında değiştirilerek üç farklı düzeyde yazılmıştır. Benzer biçimde bilgilendirici türündeki "Deniz Yıldızları" metni araştırmanın ikinci ve dördüncü yazarı tarafından üç farklı düzeyde yazılmıştır. Sonuç olarak çalışmada üçü anlatı türünde, üçü de bilgilendirici türünde olmak üzere toplam altı farklı düzeyde metin yer almaktadır. Metin temaları MEB temalarına uygun olarak belirlenmiştir.

2.2.2. Metinlerin Değerlendirilmesi (Uzman Görüşleri)

Metinler oluşturulduktan sonra Özel Eğitim alanında okuma becerilerine ilişkin çalışmalar yürüten 2 öğretim üyesi, Dilbilim Alanında görev yapan 1 öğretim üyesi, Türk Dili ve Edebiyatı alanında görev yapan 1 öğretim üyesi, Türkçe Eğitimi Alanında görev yapan 1 öğretim üyesi, Yabancı Dil Olarak Türkçenin Öğretimi Alanında görev yapan 1 öğretim üyesi ile 2 sınıf öğretmeni olmak üzere 8 ayrı uzmandan görüş alınmıştır. Uzmanların mesleki tecrübeleri 10 ile 28 yıl arasında değişmektedir. Araştırmacılar tarafından geliştirilen uzman görüş formları elektronik olarak uzmanlara iletilmiştir. Bu formda "metnin tema ile uyumu, metnin tutarlılığı, metnin olay akışı ve metnin anlatım akıcılığının 2. ve 3. sınıf düzeyine uygunluğu, metinlerde yer alan sözcüklerin, cümlelerin 2. ve 3. sınıf düzeyine uygunluğu" gibi çeşitli ölçütler yar almaktadır. Uzmanlar bu formlara her bir metni belirlenen ölçütlere uygunluğu bakımından sırasıyla inceleyerek 1 (hiç uygun değil) ile 5 puan (çok uygun) arasında değerlendirerek görüşlerini belirtmişlerdir. Ayrıca metinler üzerinde uygun gördükleri düzeltme önerilerinde bulunmuşlardır. Uzman görüşlerine göre metinler yeniden revize edilmiştir. Uzman görüşü tekrarlanmış ve metinler analiz öncesi son haline getirilmiştir. Metinlere ilişkin genel bilgiler Tablo 1'de görülmektedir. Metinlere metin düzenleme ve analiz kodlarının bulunduğu github (https://github.com/ozguraydin66/PreTXT/tree/main/documents) sayfasından ulaşılabilmektedir.

Tablo 1. *Metinlere İlişkin Genel Bilgiler*

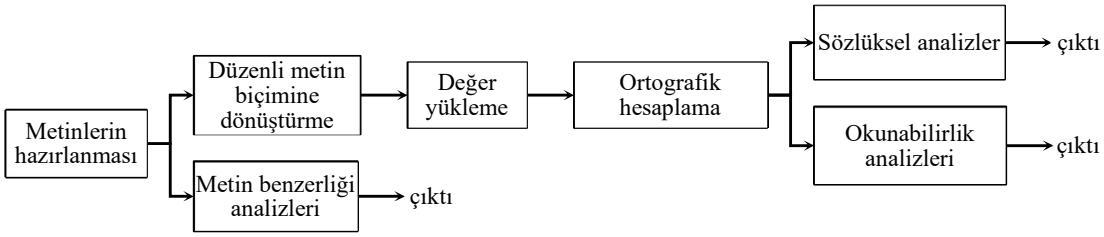
Metin kodu	Metin türü	Metin adı	Sözcük biçim sayısı*	Sözlükbirim sayısı*	Tümce Sayısı
M1*a*	anlatı	Çifte Sürpriz	171/200	137/194	26
M1*b*	bilgilendirici	Deniz Yıldızları	134/176	112/170	23
M2*a*	anlatı	Deniz Yıldızları	120/149	93/145	18
M2*b*	bilgilendirici	Çifte Sürpriz	113/148	96/145	18
M3*a*	anlatı	Deniz Yıldızları	163/183	125/179	30
M3*b*	bilgilendirici	Çifte Sürpriz	145/172	122/170	22

* Sözcük biçim ve sözlükbirim sayısına ilişkin sayılar Farklı Sözcük/Toplam Sözcük Sayısı olarak sunulmaktadır.

2.3. Metinlerin Düzenlenmesi

Bu çalışmada metinlerin analizi sürecinde Şekil 1'deki aşamalar izlenmiştir. Söz konusu aşamalar tek tek sözcüklerden oluşan metinleri veri çerçevelerine dönüştürerek metnin çeşitli özelliklerinin analiz edilmesine ve görselleştirilmesine olanak tanımaktadır.

Şekil 1. *Metin Analizi Aşamaları*



İlk aşamada metinler *.txt* formatında hazırlanarak metindeki birimler hem sözcük biçimler (word form) hem de sözlükbirimler (lexeme) olarak düzenlenmiştir. Metinlerdeki sözcüklerin lematizasyon süreci otomasyon süreci dışında tutulmuştur. Metinlerin söz konusu iki biçimdeki düzenlenişinin ardından metinlerin düzenli metin biçimine dönüştürülmesi aşamasına geçilmiştir. Düzenli metin oluşturması, metinlerin veri çerçevesine dönüştürülmesi anlamına gelmektedir. Düzenli verilerde her değişkenin bir sütunda sunulması, her satırda tek bir gözlemin sunulması gerekmektedir (Wickham, 2014). Metinlerin hazırlanması aşamasında R'de (R Core Team, 2023) *tidytext* paketi (Silge ve David, 2016) yardımıyla, aşağıda görüldüğü gibi, önce *split_sentence()* fonksiyonu kullanılarak (satır 2) her bir satırda tek bir tümcenin yer aldığı veri çerçevesi oluşturulmaktadır.

```
1  textfile = readLines('M1.txt')
2  TextSentence=split_sentence(textfile[2])%>%
3    as.data.frame() %>%
4    purrr::pluck(1) %>%
5    tibble() %>%
6    purrr::set_names('text') %>%
7    mutate(line=row_number(),
8           WordCountinSent=stri_count_words(text),
9           SentCount=length(text),
10          TextType= 'M1')
```

Daha sonra da *unnest_tokens()* fonksiyonu kullanılarak (satır 4) her bir satırda tek bir sözcüğün yer aldığı veri çerçevesi oluşturulmaktadır:

```
Word <-list()
for(z in TextSentence$line){
  Word[[z]] = TextSentence[z,]  %>%
    unnest_tokens(word, text) %>%
    dplyr::count(word, sort = TRUE)
  Word[[z]]=as.data.frame(Word[[z]])
  Word[[z]]$LineNum = z}
WordList=dplyr::bind_rows(Word)
```

Değer yükleme aşamasında ise düzenli metin biçimine dönüştürülmüş veri çerçevelerindeki her bir sözcük biçimine ilişkin biçimbirim sayılarının bulunduğu listeler de verilere aktarılmaktadır.

Ortografik hesaplama aşamasında her bir sözcük biçimi ve sözlükbirimi için *vwr* paketinde (Keuleers, 2013) *old20()* fonksiyonu yardımıyla Türkçe sözcük listesindeki (Erten vd., 2014) 20 komşu sözcüğe ortalama OLD-20 değeri hesaplanmıştır (satır 1). Bu aşamada ayrıca, *strngram* (van Heuven, 2024) paketi yardımıyla sözcüklerin bigram değerleri hesaplanmıştır. Öncelikle *get_ngram_frequencies()* fonksiyonu kullanılarak sözcük listesindeki sözcüklerden elde edilen bigramların farklı biçimlerinin sıklıklarıyla (type frequency) ile toplam sıklıkları (token frequency) belirlenmiştir (satır 2-3). Sonrasında elde edilen bu değerler, *ngram_frequency()* fonksiyonu içinde kullanılarak metinlerdeki sözcük biçimlerdeki ve sözlükbirimlerdeki bigramların ortalamaları hesaplanmıştır (satır 5-7):

```
df$OLD20 <- old20(df$word, lexicon[,1])
z =get_ngram_frequencies(lexicon$V1, lexicon$V3,
                         type = "bigram", position_specific = TRUE)
newcol = ncol(df) +1
df[,newcol] <- ngram_frequency(df$word, z, type = "bigram",
                               position_specific = TRUE,frequency = "token",
                               func = "mean", progressbar = TRUE)
```

2.4. Metin Analizleri

Metinlerin analizine ilişkin ilk aşama metinlerin benzerliğine yönelik analizleri içermektedir. Burada sözü edilen benzerlik, metinler arasındaki sözdizimsel ya da anlambilimsel bir benzerlik değil, biçimsel (sözcüksel) boyuttaki bir benzerliktir. Dolayısıyla metinlerin benzerliğinin ölçümünde yakın sözcük komşulukları ve sözcük yinelemeleri temel alınmaktadır. Bu analizlerden biri olan Jaccard katsayısı iki küme arasındaki örtüşme derecesini ölçer ve B_u ve B_v'nin (u metninin benzerliği ve v metnini benzerliği) paylaşılan niteliklerinin (sözcüklerinin) sayısının B_u veya B_v birimleri tarafından sahip olunan sayıya oranı olarak hesaplanır (bkz. 1). Örneğin, iki kümenin ikili gösterge vektörleri B_u = {*canlı*, *yıldız, kol, beş*} ve B_v = {*su, yıldız, eklem, eski*} verildiğinde, kesişimlerinin kardinalitesi 1 ve birleşimlerinin kardinalitesi 3'tür ve Jaccard katsayısı 1/3 olur. Bu çalışmada kullandığımız diğer bir benzerlik hesaplaması olan kosinüs benzerliği daha dar açının daha büyük benzerliği ve daha geniş açının daha az benzerliği gösterdiği iki derecelendirilmiş vektör arasındaki açıyı ölçmektedir. Formülde *R*(*u, i*), *u* metnini sunduğu *i* öğesinin derecelendirmesini ve *B*(*u, v*), *u* ve *v* metinlerinin ortak derecelendirilmiş öğelerinin sayısını göstermektedir (bkz. 2). Her iki hesaplama için R'de *textTinyR* paketi içindeki (Mouselimis, 2021) *JACCARD_DICE()* ve *cosine_distance()* fonksiyonları kullanılmıştır.

$$benzerlik(u,v)^{Jaccard} = \frac{|B_u \cap B_v|}{|B_u \cup B_v|} \quad (1)$$

$$benzerlik(u,v)^{Kosinus} = \frac{\Sigma_{i \in B(u,v)} - R(u,i) \cdot R(v,i)}{\sqrt{\Sigma_{i \in B(u,v)}} R(u,i)^2 \cdot \sqrt{\Sigma_{i \in B(u,v)}} R(v,i)^2} \quad (2)$$

Aşağıda, metinlerdeki noktalama işaretleri temizlendikten sonra (3-5. satılar), *JACCARD_DICE()* ve *cosine_distance()* fonksiyonlarının (7-11. satırlar) nasıl kullanıldığı görülmektedir.

```
1  text.path <- list.files(path=file.path(rootdir, paste0("texts/",ListType)),
2                          pattern = ".txt", full.names = TRUE)
3  text.file <-list()
4  for(i in 1:length(text.path)){
5    text.file[[i]] <- removePunctuation(readLines(text.path[i]))}
6
7  JScore = JACCARD_DICE(strsplit(text.file[[1]][2], "\\s+"),
8                        strsplit(text.file[[2]][2], "\\s+"),
9                        method = 'jaccard', threads = 1)
10 CScore = cosine_distance(text.file[[1]][2],
11                          text.file[[2]][2], split_separator= "")
```

Sözcüksel analiz aşamasında, verilerin birleştirilmesi aşamasında elde edilen sözcük sıklığı, hece sayısı, biçimbirim sayısı, sözcük uzunluğu verilerinin yanı sıra ortografik hesaplamaya ilişkin olarak OLD-20 ve bigram verilerinin analizleri gerçekleştirilmektedir. Söz konusu analizlerde yukarıda sıraladığımız parametrelere ilişkin olarak verilerin normal dağılım sergileyip sergilemediğini sınamak için R'de *Shapiro–Wilk* testi kullanılmış, ardından da parametreler bakımından metinler arasında anlamlı bir fark olup olmadığını görmek için *Wilcoxon Signed Rank* testi kullanılmıştır.

Diğer yandan, herhangi bir dile ait metinlerin okuyucular tarafından ne derece okunabilir olduğunu belirleyen okunabilirlik analizlerinde Türkçe için geliştirilmiş üç farklı okunabilirlik analizi temel alınmıştır. Türkçe metinlerin okunabilirliğini ölçmek için Ateşman (1997) tarafından geliştirilen formül (bkz. 3), FRES (Fresch Reading Ease Score; Flesch, 1948) formülüne dayanmaktadır. Bu formülde seslem sayısı (H), sözcük sayısı (K) ve tümce sayısı (C) temel alınarak ortalama sözcük sayısı ve tümce uzunluğu formüle eklenmiştir. Formül sonucunda 90-100 arası değerler çok kolay metinleri, 70-89 arası değerler kolay metinleri, 50-69 arası değerler orta zorluktaki metinleri, 30-49 arası değerler zor metinleri, 1-29 arası değerler de çok zor metinleri göstermektedir. Bu çalışmada okunabilirlik düzeyini ölçmek için kullanılan diğer bir formül, Çetinkaya ve Uzun (2010) tarafından Ateşman'a benzer şekilde ortalama sözcük sayısı ve tümce uzunluğu temel alarak geliştirilen formüldür (bkz. 4). Bu formülde 0-34 arası değerler engelli düzeyi (10., 11. ve 12. sınıflar), 35-50 arası değerler eğitsel düzeyi (8. ve 9.sınıf), 51< değerleri bağımsız okumayı (5., 6. ve 7. sınıf) göstermektedir. Türkçe için geliştirilmiş son okunabilirlik formülü, tümcelerdeki ortalama sözcük sayısı ve 3, 4, 5 ve 6< seslemli ortalama sözcük sayısını temel alan Bezirci ve Yılmaz'ın (2010) formülüdür (bkz. 5). Bu formüle göre, metinlerin okunabilirlik düzeylerini belirli bir sınıf düzeyine karşılık gelecek şekilde açıklamaktadır: 1-8: ilköğretim; 9-12: lise; 13-16: lisans ve ≥16: akademik.

$$OS = 198.825 - \left(40.175\ \frac{H}{K}\right) - \left(2.610\ \frac{K}{C}\right) \quad (3)$$

$$OS = 118.823 - \left(25.987\ \frac{H}{K}\right) - \left(0.971\ \frac{K}{C}\right) \quad (4)$$

$$OS = \sqrt{\frac{K}{C}\ ((H3\ \ 0.84) + (H4\ \ 1.5) + (H5\ \ 3.5) + (H6\ \ 26.25))} \quad (5)$$

Yukarıdaki formüllerin R'de hesaplanabilmesi için öncelikle her bir metne ilişkin değişkenlerin hesaplandığı 'df' adında bir veritabanı oluşturulmuştur. Bu veritabanında her bir metnin seslem sayısı (H), sözcük sayısı (K) ve tümce sayısı, seslemlere göre ortalama sözcük sayısı gibi özellikleri sütunlarda tek tek hesaplanıp işlenmiştir. Ardından her üç formül de R'de (R Core

Team, 2023), *dplyr* paketinde (Wickham vd., 2023) *mutate()* fonksiyonu kullanılarak hesaplanmıştır:

```
Calculate <- df %>%
  group_by(TextType) %>%
  mutate(
    Atesman   = 198.825-(40.175*(H/K))-(2.610*(K/C)),
    Cetinkaya = 118.823-(25.987*(H/K))-(0.971*(K/C)),
    Bezirci   = sqrt((K/C)*((H3*0.84)+(H4 *1.50)+(H5 *3.50)+(H6 *26.25)))) %>%
  select(c(21:23)
  )
```

Metinlere ilişkin metin benzerliği analizleri, sözcüksel analizler ve okunabilirlik analizleri yapıldıktan sonra elde edilen bulgulardan yararlanarak iki farklı metin türünde birbirine en çok benzeyen ve 2. ve 3. sınıf düzeyindeki öğrencilere en uygun biri bilgilendirici biri öyküleyici olmak üzere iki metin seçilmiştir. Bu iki metin sözcük ve cümle sayısı açısından kontrol edilmiştir. Her iki metindeki sözcük sayısı 150, cümle sayısı 22 olup her iki metindeki tüm cümleler kurallı cümledir. Ek olarak metinlerdeki basit ve karmaşık, olumlu ve olumsuz, eylem ve isim cümle türü sayıları birbirine yakın olarak düzenlenmiştir. Metinlerde yer alan cümlelerdeki minimum ve maksimum sözcük sayıları da birbirine yakın olarak ayarlanmıştır.

Bu aşamada yukarıdaki analizlerin yanı sıra R'de stats paketi (R Core Team, 2023) içindeki hclust() fonksiyonu yardımıyla hiyerarşik kümeleme de yapılmıştır. Bunun için öncelikle aşağıda 1-2. satırlarda veritabanından ortalama seslem sayısı (M), sözcük sayısı (K) ve tümce sayısı (C) seçilmiş, ardından da 4-6. satırlarda veri normalize edilmiştir. Normalize edilen verinin (nor), aralık matrisleri hesaplanmıştır (satır 7) ve hclust() fonksiyonu yardımıyla (satır 8) hiyerarşik küme hesaplanmış ve çizdirmiştir (satır 9):

```
Clust <- df %>%
  select(TextType, M, K, C)
z <- Clust[,-c(1,1)]
means <- apply(z,2,mean)
sds <- apply(z,2,sd)
nor <- scale(z,center=means,scale=sds)
distance = dist(nor)
df.hclust = stats::hclust(distance)
plot(df.hclust,labels=Clust$TextType, main="")
```

3. Bulgular

3.1. Metin Benzerliği Bulguları

Altı metnin birbirlerine olan benzerliklerinin incelenmesinde metinler, sözlükbirim olarak düzenlenen ve sözcük biçim olarak düzenlenen metinler olmak üzere kendi içlerinde ayrı ayrı karşılaştırılmıştır. Bu aşamada aynı türdeki metinler birbirine ne kadar benzer ya da ne kadar farklı olduğunu ortaya koymak amaçlandığından aynı metin türleri içinde benzerlik karşılaştırmaları yapılmıştır. Bu nedenle, Jaccard katsayıları ve kosinüs katsayıları farklı metin çiftlerinde (M1, M2 ve M3) ama aynı metin türlerinde (a ve b) karşılaştırılmıştır (yani, M1a krş. M2a, M2a krş. M3a, M1a krş. M3a vb.). Tablo 2'de görüldüğü gibi, Jaccard katsayıları ve kosinüs katsayıları, anlatı metinlerinde, M1 ile M3'ün birbirine daha yakın olduğunu, M2'nin ise M1 ve M3'e neredeyse eşit uzaklıkta olduğunu göstermektedir. Bilgilendirici metinlerde ise M1 ile M2'nin birbirlerine daha yakın olduğu, M3'ün ise M1 ve M2'ye neredeyse eşit uzaklıkta olduğu görülmektedir. Genel olarak metin benzerliklerinin sözcüklerin dilbilgisel biçimbirimlerle sunulduğu sözcük biçim formlarında düştüğü görülmektedir. Farklı Sözcüklerin Toplam Sözcük Sayısına oranının anlatı türü metinlerde daha yüksek, bilgilendirici türündeki metinlerde ise daha düşük olduğu görülmektedir. Bu da bilgilendirici türü metinlerde sözcük çeşitliliğinin daha fazla olduğu anlamına gelmektedir.

Tablo 2. *Metin Benzerlikleri ve Yoğunlukları*

	Jaccard benzerliği				Kosinus benzerliği				Yoğ.*
	M1*a*	M2*a*	M1*b*	M2*b*	M1*a*	M2*a*	M1*b*	M2*b*	
				Sözlükbirim					
M2*a*	.38				.65				.64
M3*a*	**.70**	.40			**.86**	.74			.70
M2*b*			**.60**				**.84**		.66
M3*b*			.39	.35			.73	.71	.72
	.71		.66						
				Sözcük biçim					
M2*a*	.28				.50				.81
M3*a*	**.61**	.34			**.76**	.59			.89
M2*b*			**.54**				**.75**		.76
M3*b*			.20	.21			.58	.52	.84
Yoğ.*	.86	–	.76	–	–	–	–	–	–

*Yoğ.= Yoğunluk

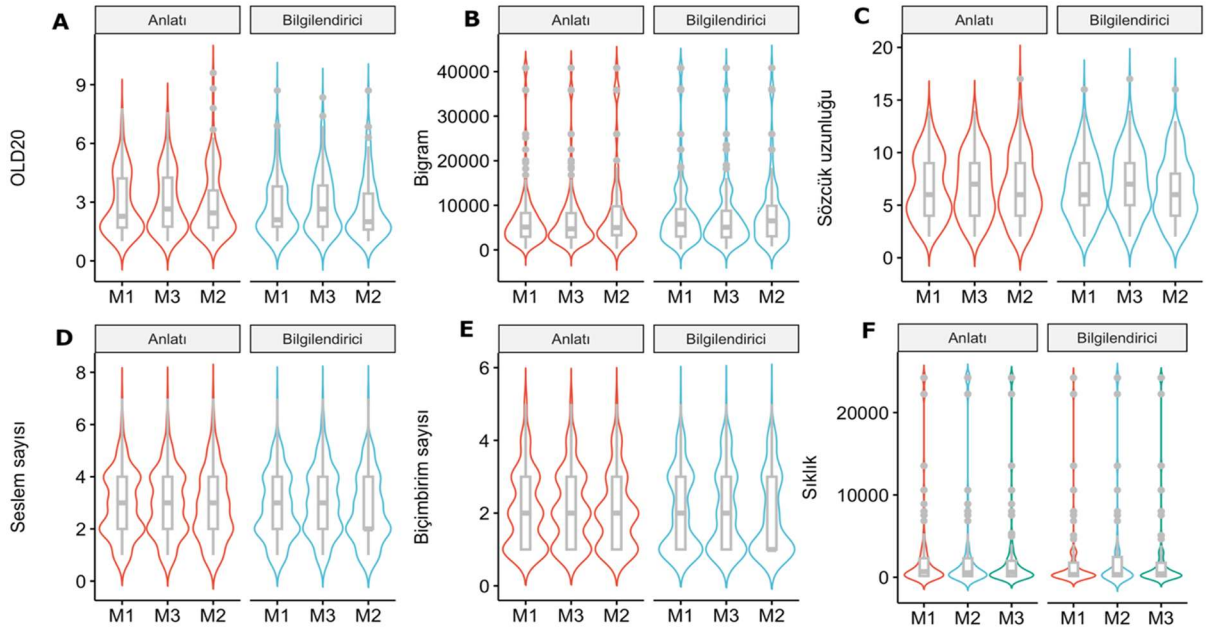
3.2. Sözlüksel Analiz Bulguları

Tablo 3'te ve Şekil 2'de (A-B panelleri) görüldüğü gibi, ortografik özellikler bakımından OLD-20 ve bigram parametrelerinde aynı metin çiftleri (yani M1, M2 ve M3) için metin türleri bakımından anlamlı bir farklılık bulunmamaktadır. Benzer biçimde sözlüksel özelliklere ilişkin sözcük uzunluğu, seslem sayısı ve biçimbirim sayısı parametrelerinde metin türleri bakımından anlamlı bir farklılık bulunmamaktadır (bkz. Tablo 4 ve Şekil 2C- 2E). Sözlükbirimlerin incelenmesinde de ortografi bakımından benzer şekilde OLD-20 ve bigram parametreleri gruplar arasında anlamlı bir farklılığa yol açmamaktadır. Kullanılan sözcüklerin genel olarak derlemde görülme sıklıkları yönüyle de gruplar birbirine benzer bir görünüm sergilemektedir.

Tablo 3. *Sözlükbirimlere İlişkin Sözcük Sıklığı, OLD-20 ve Bigram Parametrelerinin Metinler Arasındaki Farklılıkları Gösteren Wilcoxon Sıra Toplamı Analizi*

Metin	Metin türü	*N*	Sözcük sıklığı			OLD-20			Bigram		
			Ort (SH)	*Z*	*p*	*Ort (SH)*	*Z*	*p*	*Ort (SH)*	*Z*	*p*
M1*a*	anlatı	171	2918 (410)	-1.68	.09	1.61 (.04)	-.01	.99	10939 (667)	-.19	.84
M1*b*	bilgilendirici	134	2263 (371)			1.58 (.04)			10409 (650)		
M2*a*	anlatı	120	3399 (552)	-1.29	.19	1.70 (.05)	-.97	.33	11228 (834)	-.06	.94
M2*b*	bilgilendirici	113	2871 (494)			1.60 (.05)			10808 (746)		
M3*a*	anlatı	163	2725 (423)	-.98	.32	1.62 (.04)	-1.21	.22	10328 (663)	-.27	.78
M3*b*	bilgilendirici	145	2580 (435)			1.69 (.05)			10389 (648)		

Şekil 2. *Sözcüksel Parametrelerin Karşılaştırılması*



Not: A, B, C, D, E Panelleri sırasıyla sözcük biçimde OLD-20, bigram, sözcük uzunluğu, seslem sayısı ve biçimbirim sayısı parametrelerini göstermektedir. F paneli ise sözlükbirimlerde sözcük sıklığı parametresini göstermektedir.

Tablo 4. *Sözcük Biçimlere İlişkin OLD-20, Bigram, Sözcük Uzunluğu, Seslem Sayısı ve Biçimbirim Sayısı Parametrelerinin Metinler Arasındaki Farklılıkları Gösteren Wilcoxon Sıra Toplamı Analizi*

Metin	Metin türü	*N*	OLD-20			Bigram			Sözcük uzunluğu			Seslem sayısı			Biçimbirim sayısı		
			Ort (SH)	*Z*	*p*	*Ort (SH)*	*Z*	*p*	*Ort (SH)*	*Z*	*p*	*Ort (SH)*	*Z*	*p*	*Ort (SH)*	*Z*	*p*
M1*a*	anlatı	200	2.87 (.11)			7405 (560)			6.67 (.21)			2.89 (.08)			2.08 (.07)		
				-.72	.46		-1.01	.31		-.06	.94		-.74	.45		-.67	.50
M1*b*	bilgilendirici	176	2.72 (.11)			8032 (606)			6.63 (.22)			2.80 (.09)			2.01 (.08)		
M2*a*	anlatı	149	2.86 (.13)			8763 (780)			6.55 (.26)			2.84 (.10)			2.04 (.08)		
				-1.05	.29		-.83	.40		-.06	.95		-.89	.36		-1.20	.22
M2*b*	bilgilendirici	148	2.64 (.12)			8595 (682)			6.47 (.24)			2.70 (.10)			1.94 (.09)		
M3*a*	anlatı	183	2.96 (.11)			7058 (559)			6.78 (.22)			2.93 (.09)			2.13 (.07)		
				-.17	.86		-.75	.45		-.79	.42		-.41	.67		-.97	.33
M3*b*	bilgilendirici	172	2.92 (.11)			7405 (551)			7.05 (.23)			3.00 (.09)			2.05 (.08)		

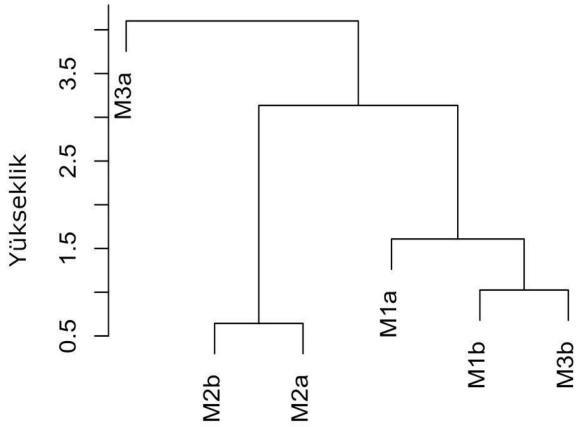
3.3. Metin Okunabilirliği Bulguları

Metinlerin okunabilirliğinin belirlenmesi için Türkçe için geliştirilen üç okunabilirlik formülü altı ayrı metne uygulanmıştır (bkz. Ateşman, 1997; Bezirci ve Yılmaz, 2010; Çetinkaya ve Uzun, 2010). Söz konusu analizlerin gerçekleştirilebilmesi için metinlerdeki seslem sayıları, sözcük sayıları ve tümce sayıları belirlenmiştir (bkz Tablo 5). Tablo 5'te de görüldüğü gibi metinlerin okunabilirlik düzeyleri, M1*a* anlatı metni Ateşman'a (1997) göre orta düzeyde, Çetinkaya ve Uzun'a (2010) göre engelli düzeyinde ve Bezirci ve Yılmaz'a (2010) göre lise düzeyindedir. M1*b* bilgilendirici metni Ateşman'a (1997) göre orta düzeyde, Çetinkaya ve Uzun'a (2010) göre engelli düzeyinde ve Bezirci ve Yılmaz'a (2010) göre ilköğretim düzeyindedir. M2*a* anlatı metni Ateşman'a (1997) göre orta düzeyde, Çetinkaya ve Uzun'a (2010) göre engelli düzeyinde ve Bezirci ve Yılmaz'a (2010) göre lise düzeyindedir. M2*b* bilgilendirici metni Ateşman'a (1997) göre orta düzeyde, Çetinkaya ve Uzun'a (2010) göre engelli düzeyinde ve Bezirci ve Yılmaz'a (2010) göre ilköğretim düzeyindedir. M3*a* anlatı metni Ateşman'a (1997) göre orta düzeyde, Çetinkaya ve Uzun'a (2010) göre engelli düzeyinde

ve Bezirci ve Yılmaz'a (2010) göre ilköğretim düzeyindedir. M3*b* bilgilendirici metni Ateşman'a (1997) göre orta düzeyde, Çetinkaya ve Uzun'a (2010) göre engelli düzeyinde ve Bezirci ve Yılmaz'a (2010) göre lise düzeyindedir.

Tablo 4'teki parametreler üzerinden hiyerarşik kümeleme ile hangi metin türü çiftlerinin birbirlerine daha yakın olduğu belirlenmiştir. Hiyerarşik kümeleme yöntemi R'de stat paketi içinde *hclust()* paketi yardımıyla elde edilen hiyerarşi Şekil 3'te yer almaktadır. M2a ve M2b metinleri Şekil 3'te görüldüğü gibi en uygun kümelenmiş metinler görünümü sergilemektedir. Bu belirlemeye ve daha önceki analizlere dayanarak en yakın çift tekrar uzman görüşüne sunulmuştur. İkinci uzman görüşü sonrası uzmanlar görüş birliği ile metinlerin tema, konu ve bilgi akışı açısından 2 ve 3. sınıf düzeyine uygun olduğunu bildirmişlerdir.

Şekil 3. *Metinlerin Hiyerarşik Kümelemesi (Distance, Hclust (*,'Complate')*



Tablo 5. *Metin Okunabilirliği Bulguları*

	M1*a* *anlatı*	M1*b* *bilgilendirici*	M2*a* *anlatı*	M2*b* *bilgilendirici*	M3*a* *anlatı*	M3*b* *bilgilendirici*
Seslem	24.04	22.91	24.83	24.56	19.47	24.27
Çok seslemli sözcük	7.12	7.00	7.28	7.50	5.70	7.18
İki seslemli sözcük	2.5	2.70	2.5	2.94	1.83	2.36
Üç seslemli sözcük	1.54	1.74	2	1.94	1.4	1.86
Dört seslemli sözcük	2.19	1.74	1.72	1.72	1.7	1.91
Beş seslemli sözcük	0.69	0.74	0.78	0.83	0.6	0.82
6 < seslemli sözcük	0.19	0.09	0.28	0.06	0.17	0.23
Toplam sözcük sayısı	200	178	150	150	183	176
Toplam tümce sayısı	26	23	18	18	30	22
Ateşman (1997)	53.2	59.68	57.35	58.69	54.7	56.05
Çetinkaya (2010)	30.14	34.37	33.29	34.16	29.97	32.21
Bezirci ve Yılmaz (2010)	9.63	8.32	10.91	8.46	7.89	10.3

4. Sonuç, Tartışma ve Öneriler

Bu çalışmada, okuma becerisinin değerlendirmesinde kullanılacak metinler ile ilgili niceliksel ölçütlerle R dili kullanılıp ilk kez Türkçede bir metin analiz aracı geliştirilerek araştırmacılar tarafından yazılan metinlerin, geliştirilen bu araç ile test edilmesi amaçlanmıştır. Ayrıca araştırmacılar tarafından geliştirilen metinlerin benzerlik analizleri yapılarak karşılaştırılmıştır. Bu çalışma kapsamında geliştirilen R paketi https://github.com/ozguraydin66/PreTXT/ adresinde kullanıma açıktır. Ayrıca, bu paketin web uygulamasına dönüştürülmüş biçimine https://oaydin.shinyapps.io/pretxt/ sayfasından ulaşılabilmektedir.

Araştırmanın metin benzerlik bulguları sözlükbirim ve sözcük biçim olarak sunulmaktadır. Bu kapsamda anlatı metinlerinde, M1 ile M3'ün birbirine daha yakın olduğunu, M2'nin ise M1 ve M3'e neredeyse eşit uzaklıkta olduğu belirlenmiştir. Bilgilendirici metinlerde ise M1 ile M2'nin birbirlerine daha yakın olduğu, M3'ün ise M1 ve M2'ye neredeyse eşit uzaklıkta olduğu tespit edilmiştir. Nitekim sözlükbirim olarak anlatı ve bilgilendirici metinlerin oldukça benzer oldukları ifade edilebilir. Bununla beraber metin benzerlikleri sözcük biçim olarak incelendiğinde bilgilendirici türdeki metinlerin sözcük çeşitliliğinin daha fazla olduğu görülmüştür.

Metinlerin sözcüksel analiz bulguları sözcük biçim bakımından ele alındığında ortografik özelliklere ilişkin olarak OLD-20 ve bigram, sözcük uzunluğu, seslem sayısı ve biçimbirim sayısı parametrelerinde anlamlı farklılık göstermemektedir. Metinlerin sözlükbirimlerinin de ortografi bakımından benzer şekilde OLD-20 ve bigram parametreleri anlamlı bir farklılık sergilememiştir. Dolayısıyla OLD-20'nin Türkçe okuma üzerinde engelleyici bir etkiye sahip olduğu bilindiğinden (Erten vd., 2014) hem sözcükbiçim hem de sözlükbirim bakımından OLD-20 değerlerinin metinlerde anlamlı farklılık göstermemesi önemli olmuştur. Bunun yanı sıra metinlerde birbirlerinden daha uzun sözcüklerin olması halinde bu durumun okurların okuma sürelerine yansıyacağı bilindiğinden (New vd., 2006; Rayner, 1998, 2009) sözcük uzunluğu parametresine ilişkin metinlerin farklılaşmadığı ortaya konulmuştur. Ayrıca metinlerde kullanılan sözcüklerin genel olarak derlemde görülme sıklıklarına ilişkin de birbirine benzer görünümde oldukları belirlenmiştir. Dolayısıyla bir dilde bir sözcüğün sıkça geçmesi bu sözcüğün tanınmasını hızlandıracağından (Erten vd., 2014; Zevin, 2009) metinlerde yer alan sözcüklerin derlemde görülme sıklıklarının benzerliği sağlanmıştır.

Metin okunabilirliği bulguları incelendiğinde araştırmada yer alan metinlerin Ateşman'a (1997) göre orta düzeyde oldukları, Çetinkaya ve Uzun'a (2010) göre engelli düzeyinde, Bezirci ve Yılmaz'a (2010) göre ise ilköğretim ve lise düzeylerine karşılık gelmektedir. Bu metinlere OLD-20, bigram, sözcük uzunluğu, seslem sayısı ve biçimbirim sayısı parametreleri kullanılarak hiyerarşik kümeleme uygulanıp hangi metin türü çiftinin birbirlerine daha yakın oldukları hesaplanmıştır. Nitekim M2a ve M2b metinlerinin kümelenmiş en yakın metinler oldukları tespit edilmiştir. Böylelikle çalışmada, araştırmacı, öğretmen, dil konuşma terapisti gibi uzmanların, öğrencilerin okuma becerilerini hem değerlendirme hem de bu becerilerine ilişkin uygun olan müdahaleyi yapabilmeleri için metin belirleme sürecinde kullanabilecekleri bir araç ortaya konmuştur. Ayrıca çalışma, MEB'e bağlı Talim ve Terbiye Kurulu Başkanlığının "Taslak Ders Kitabı ve Eğitim Araçları ile Bunlara Ait Elektronik İçeriklerin İncelenmesinde Değerlendirmeye Esas Olacak Kriterler ve Açıklamaları" (2023) kitapçığında ders kitaplarında yer alacak metinlere ilişkin dil ve üslup, yazı dili standartlarına uygunluk ile anlam ve anlatım başlıklarında sunulan ölçütlerin niceliksel olarak bütünsel bir biçimde sağlanmasına aracılık edebilir. Bununla beraber alan yazında diğer dillerdeki metinler için analizler yapılmışken (Arnold vd., 2019; Aziz vd., 2010; Kırkgöz ve Ünaldı, 2014) bu çalışmada Türkçe metinler için ilk kez R dilinde bir araç geliştirilerek metinlerin analizleri gerçekleştirilmiştir.

Bu çalışmada, metin dilbilimsel bir analiz aracı geliştirilerek araştırmacı ve uzmanlar için önerilmektedir. Türkçe metinler ile ilk kez böyle bir çalışma gerçekleştirildiği için referans olabilecek ilk parametreler bu çalışma ile tespit edilmiştir. Böylece Türkçede metin benzerliği, sözcüksel analiz ve metin okunabilirliğine ilişkin ilk değerler bu çalışmada hesaplanmıştır. Sonuç olarak M2a ve M2b metinlerinin sahip olduğu okuma becerisini değerlendirmeye dönük özellikleri sunulmaktadır. Dolayısıyla bu çalışmada geliştirilen R tabanlı metin analiz aracı, okuma becerisinin değerlendirilmesinde yaygın olarak kullanılan diğer ölçeklerden farklı olarak, metinlerin dilbilimsel özelliklerine odaklanmaktadır. Alan yazında kullanılan ölçekler özellikle klinik popülasyonlarda okuma performansının belirlenmesinde oldukça değerli bilgiler sunmakta, ancak kullanılan metinlerin yapısal özellikleri hakkında sınırlı veri sağlamaktadır. Bu araştırma ile geliştirilen araç ise okuma değerlendirmesi yapmayı planlayan araştırmacı ve uzmanlara metin seçiminde sözcüksel analiz, metin okunabilirliği gibi çeşitli metin dilbilimsel özelliklere ilişkin önemli veriler sunabilecektir. Böylece araştırmacı ve uzmanlar değerlendirme sürecinde kullanılacakları metinlerin dilbilimsel özellikleri hakkında ayrıntılı bilgiler edinerek metin seçiminde bilinçli kararlar alabileceklerdir.

Araştırma kapsamında geliştirilen R tabanlı metin analiz aracının bulguları, aracın güvenilir ve pratik bir biçimde çalıştığını göstermektedir. Metinlerin değerlendirilmesinde, tek bir ölçüte bağlı kalmak yerine, farklı dilbilimsel parametrelerin bütüncül bir yaklaşımla incelenebileceği ortaya konmuştur. Nitekim M2a ve M2b metinleri üzerinde yapılan örnek analizler aracın sunduğu ölçülebilir karşılaştırma imkânlarını doğrulamaktadır. Bu yönüyle çalışma, klinik değerlendirme amaçlı doğrudan bir ölçek geliştirmekten ziyade, mevcut ölçeklerde kullanılan ya da kullanılabilecek metinlerin niteliğine dair uzmana ve araştırmacıya bilgi sağlayan, tamamlayıcı bir yaklaşım sunmaktadır. Dolayısıyla geliştirilen aracın, hem eğitsel müdahalelerde kullanılacak metinlerin seçiminde hem de klinik uygulamalarda değerlendirmecilere rehberlik etme potansiyeli bulunduğu ifade edilebilir.

Araştırmada okuma becerisinin değerlendirmesinde kullanılacak metinler ile ilgili niceliksel ölçütlerle R dili kullanılıp ilk kez Türkçede bir metin analiz aracı geliştirilerek önerilmiş ve araştırmacılar tarafından yazılan metinler, geliştirilen bu araç ile test edilmiştir. Ayrıca araştırmacılar tarafından geliştirilen metinlerin benzerlik analizleri de yapılmıştır. Çalışma, bu yönleriyle özgün ve yenilikçidir. Bununla beraber önerilen aracın öğrencilere uygulanıp geçerlilik ve güvenilirlik çalışmaları yapılarak standart bir araç olma yolunda ilerlemesi öngörülmektedir. Sonuç olarak çalışmada, okuma becerisinin klinik değerlendirmesinde kullanılacak metinler için R dilinde, niceliksel ölçütlerle ve bütünsel olarak değerlendirme yapabilmek amacıyla bir metin analiz aracı geliştirilmesi önerilmiştir. Böylece 2. ve 3. sınıfa devam eden öğrencilerin okuma becerilerinin değerlendirilmesi ve ileri tanılama-müdahale araçları için en uygun metinler ortaya konulmaya çalışılmıştır. Öte yandan araştırmanın bazı sınırlıkları mevcuttur. Bunlardan ilki, biçimbirim sayısı belirlemeye yönelik otomatik bir sürece yer verilmemiş, ayrıca dizge (v.01.16, Mutlu vd., 2021) aracı kullanılarak seslem sayıları belirlenmiştir. İkincisi, metinlerin bağdaşıklık ve tutarlılığı incelenmemiştir. Sonuncusu ise metin analiz aracı öğrencilere henüz uygulanmamıştır.

Geliştirilmesi önerilen metin değerlendirme aracı ile 2. ve 3. sınıfa devam eden öğrencilerin okuma becerilerinin değerlendirilmesine yönelik metinler hem okuma becerilerinin incelenmesi yönünde çalışmalar yapan araştırmacılar hem de özel eğitim öğretmeni, sınıf öğretmeni, rehberlik ve psikolojik danışmanlık öğretmeni, dil konuşma terapisti gibi uzmanların kullanımlarına sunulabilir. Bu kapsamda araç ve metinlerin okullarda, danışmanlık merkezlerinde ve Rehberlik Araştırma Merkezlerinde (RAM) kullanılması önerilebilir. Öte yandan önerilen araç, özel eğitim, dil ve konuşma terapisi gibi çeşitli alanlarda çalışmalar yürüten araştırmacıların okuma becerisini değerlendirip tanı koyup ya da müdahale programları uyguladıkları süreçlerde etkin bir biçimde kullanılabilir. Araç, dilbilim ve Türkçe eğitimi gibi

alanlarda sözcük sıklığı, harf-bigram sıklığı gibi parametreler bağlamında yürütülen çalışmalarda da kullanılabilir. Ayrıca göz-izleme gibi ileri teknolojileri kullanan araştırmacılar önerilen araçla kapsamlı incelemeler gerçekleştirebilirler.

Bilgilendirme

Çalışma, "Okul Çağı Disleksi Tanılamasına Nörodilbilimsel Yaklaşım: Fonem Diskriminasyonu, Sözcük ve Metin Okuma Süreçlerinin Psikometrik, Elektrofizyolojik ve Göz İzleme Teknikleri ile İncelenmesi ve Karşılaştırılması" adlı 122k308 nolu TÜBİTAK1001 projesi kapsamında gerçekleştirilmiştir. TÜBİTAK'a destekleri için teşekkür ederiz.

Etik Kurul İzin Bilgisi

Bu araştırma, İzmir Bakırçay Üniversitesi Girişimsel Olmayan Klinik Araştırmalar Etik Kurulunun 28/09/2022 tarihli 717 sayılı kararı ile alınan izinle yürütülmüştür.

Yazar Çıkar Çatışması Bilgisi

Yazarların beyan edeceği bir çıkar çatışması yoktur.

Yazar Katkısı

Esmehan Özer: *Metinlerin yazılması, çalışmanın giriş, tartışma bölümlerinin yazımı ve araştırmanın gözden geçirilerek raporlanması.* ***Sema Acar-Ünalgan:*** *Araştırma izninin alınması, metinlerin yazılması, giriş, yöntem ve tartışma bölümlerinin geliştirilmesi.* ***Hazal Artuvan-Korkmaz:*** *Metinlerin yazılması ve araştırmanın gözden geçirilerek raporlanması.* ***Rahime Duygu Temeltürk:*** *Metinlerin yazılması ve araştırmanın gözden geçirilerek raporlanması.* ***Özgür Aydın:*** *Yöntem bölümünün yazılması, verilerin analiz edilmesi ve bulguların yazılması.*

Orcid

Esmehan Özer https://orcid.org/0000-0001-5919-8072

Sema Acar-Ünalgan https://orcid.org/0000-0001-9129-2229

Hazal Artuvan-Korkmaz https://orcid.org/0000-0001-7739-8676

Rahime Duygu Temeltürk https://orcid.org/0000-0002-9303-5944

Özgür Aydın https://orcid.org/0000-0003-2925-4146

KAYNAKÇA

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review, 4*(4), 439-461. https://doi.org/10.3758/BF03214334

Arnold, T., Ballier, N., Lissón, P., & Tilton, L. (2019). Beyond lexical frequencies: Using R for text analysis in the digital humanities. *Language Resources and Evaluation, 53*(4), 707-733. https://doi.org/10.1007/s10579-019-09456-6

Ateşman, E. (1997). Türkçede okunabilirliğin ölçülmesi. *Ankara Üniversitesi Türkçe ve Yabancı Dil Uygulama ve Araştırma Merkezi Dil Dergisi, 58*, 171-174.

Aziz, A., Fook, C. Y., & Alsree, Z. (2010). Computational text analysis: A more comprehensive approach to determine readability of reading materials. *Advances in Language and Literary Studies, 1*(2), 200-219. https://doi.org/10.7575/aiac.alls.v.1n.2p.200

Başaran, M. ve Ayol, H. (2009). Okuduğunu anlama ve metne karşı geliştirilen tutum üzerinde metnin bilgi verici veya hikâye edici olmasının etkisi. *Uşak Üniversitesi Sosyal Bilimler Dergisi, 2*(1), 11-23. https://doi.org/10.12780/UUSBD41

Baştuğ, M. (2012). İlköğretim I. kademe öğrencilerinin akıcı okuma becerilerinin çeşitli değişkenler açısından incelenmesi [Investigation of primary school first stage students' reading fluency skills in terms of certain variables] (Tez Numarası: 311007) [Doktora tezi, Gazi Üniversitesi]. Yükseköğretim Kurulu Ulusal Tez Merkezi.

Bezirci, B. ve Yılmaz, A. E. (2010). Metinlerin okunabilirliğinin ölçülmesi üzerine bir yazilim kütüphanesi ve Türkçe için yeni bir okunabilirlik ölçütü. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 12*(3), 49-62.

Borleffs, E., Maassen, B. A. M., Lyytinen, H., & Zwarts, F. (2019). Cracking the code: The impact of orthographic transparency and morphological-syllabic complexity on reading and developmental dyslexia. *Frontiers in Psychology, 9*, 1-19. https://doi.org/10.3389/fpsyg.2018.02534

Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing, 12*(3), 169-190. https://doi.org/10.1023/A:1008131926604

Çetinkaya, G. (2010). Türkçe metinlerin okunabilirlik düzeylerinin tanımlanması ve sınıflandırılması [Identifying and classifying the readability levels of the Turkish texts] (Tez Numarası: 265580) [Doktora tezi, Ankara Üniversitesi]. Yükseköğretim Kurulu Ulusal Tez Merkezi.

Çetinkaya, G. ve Uzun, L. (2010). Türkçe ders kitaplarındaki metinlerin okunabilirlik özellikleri. H. Ülper (Ed.), *Türkçe ders kitabı çözümlemeleri* (4, 141-153) içinde. Pegem Akademi.

Duran, E. ve Kargın, T. (2022). Okuma yazma eğitiminde güncel konular: Ne çalışmalı, nasıl çalışmalı?. *Milli Eğitim Dergisi, 51*(235), 1859-1876. https://doi.org/10.37669/milliegitim.886879

Erten, B., Bozşahin, C., & Zeyrek, D. (2014). Turkish resources for visual word recognition. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 2106-2110). European Language Resources Association (ELRA).

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221-233. https://doi.org/10.1037/h0057532

Güneş, F. (2013). Türkçede metin öğretimi yerine metinle öğrenme. *Adıyaman Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *11*, 603-637. https://doi.org/10.14520/adyusbd.454

Hamzadayı, E. ve Batmaz, Ö. (2022). Okuduğunu anlamayı etkileyen etmenlere yönelik bir inceleme. *Ege Eğitim Dergisi, 23*(2), 190-209. https://doi.org/10.12984/egeefd.1105439

Hebert, M., Bohaty, J. J., Nelson, J. R., & Brown, J. (2016). The effects of text structure instruction on expository reading comprehension: A meta-analysis. *Journal of Educational Psychology, 108*(5), 609-629. https://doi.org/10.1037/edu0000082

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2*(2), 127-160. https://doi.org/10.1007/BF00401799

Kassambara, A. (2023). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.2. https://rpkgs.datanovia.com/rstatix/.

Keuleers, E. (2013). _vwr: Useful functions for visual word recognition research_. R package version 0.3.0. https://CRAN.R-project.org/package=vwr

Kirkgöz, Y., & Ünaldi, I. (2012). Coh-metrix: Introduction and validation of an online tool for text analysis. *Çukurova University Faculty of Education Journal, 41*(2), 1-17.

Kucer, S. B. (2014). *Dimensions of literacy: A conceptual base for teaching reading and writing in school settings*. Routledge.

Kumar, A., & Paul, A. (2016). Mastering text mining with R. Packt Publishing Ltd.

Kwartler, T. (2017). Text mining in practice with R. John Wiley & Sons.

Levesque, K. C., Breadmore, H. L., & Deacon, S. H. (2021). How morphology impacts reading and spelling: Advancing the role of morphology in models of literacy development. *Journal of Research in Reading, 44*(1), 10-26. https://doi.org/10.1111/1467-9817.12313

Liu, Y., Groen, M. A., & Cain, K. (2024). The association between morphological awareness and reading comprehension in children: A systematic review and meta-analysis. *Educational Research Review, 42*, 100571. https://doi.org/10.1016/j.edurev.2023.100571

Milli Eğitim Bakanlığı (2019). *Türkçe Dersi Öğretim Programı (ilkokul ve ortaokul 1-8. sınıflar).*

Milli Eğitim Bakanlığı (2021). *Ders kitaplarında okunabilirlik.* Ders Kitapları ve Öğretim Materyalleri Daire Başkanlığı.

Milli Eğitim Bakanlığı (2023). *Taslak ders kitabı ve eğitim araçları ile bunlara ait elektronik içeriklerin incelenmesinde değerlendirmeye esas olacak kriterler ve açıklamaları.* Talim ve Terbiye Kurulu Başkanlığı.

Mouselimis, L. (2021). textTinyR: Text Processing for Small or Big Data Files}, R package version 1.1.8. https://CRAN.R-project.org/package=textTinyR.

Muncer, S. J., Knight, D., & Adams, J. W. (2014). Bigram frequency, number of syllables and morphemes and their effects on lexical decision and word naming. *Journal of Psycholinguist Research, 43*, 241-254. https://doi.org/10.1007/s10936-013-9252-8

Mutlu, M. U., Yetimaslan, N., & Atagün, İ. (2021). Dizge: The grammar analyzer for Turkish. https://github.com/dizge

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review, 13*(1), 45-52. https://doi.org/10.3758/BF03193811

Perfetti, C. A. (1999). Comprehending written language: A blueprint of the reader. In C. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 167-208). Oxford University Press.

Perfetti, C. (2012). Thru but not wisht: Language, writing, and universal reading theory. *Behavioral and Brain Sciences, 35*(5), 299-300. https://doi.org/10.1017/S0140525X12000234

Pilav, S. ve Oğuz, M. (2013). Türkçe ders kitaplarında yer alan metin türleri üzerine bir araştırma. *Kırıkkale Üniversitesi Sosyal Bilimler Dergisi, 3*(2), 16-30.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372-422. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology, 62*(8), 1457-1506. https://doi.org/10.1080/17470210902816461

Rayner, K., & Reichle, E. D. (2010). Models of the reading process. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(6), 787-799. https://doi.org/10.1002/wcs.68

Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance, 37*(2), 514-528. https://doi.org/10.1037/a0020990

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Santangelo, D. M. (2023). *The influence of bigram frequency on naming speed: A meta-analysis.* [Master's thesis, The Florida State University].

Schmalz, X., & Mulatti, C. (2017). Busting a myth with the Bayes factor: Effects of letter bigram frequency in visual lexical decision do not reflect reading processes. *The Mental Lexicon, 12*(2), 263-282. https://doi.org/10.1075/ml.17009.sch

Silge, J., & David, R. (2016). tidytext: Text mining and analysis using tidy data principles in r. JOSS, 1 (3). https://doi.org/10.21105/joss.00037

van Heuven, W. (2024). _strngrams: ngram functions_. R package version 0.3.6. https://github.com/waltervanheuven/strngrams

Verhoeven, L., & Perfetti, C. A. (2011). Morphological processing in reading acquisition: A cross-linguistic perspective. *Applied Psycholinguistics, 32*(3), 457-466. https://doi.org/10.1017/S014271641 1000154

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). dplyr: A Grammar of Data Manipulation. R package version 1.1.4. https://dplyr.tidyverse.org.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software 59*(1), 1-23. https://doi.org/10.18637/jss.v059.i10.

Yap, M. J., & Balota, D. A. (2015). Visual word recognition. In A. Pollatsek & R. Treiman (Eds.), *The Oxford handbook of reading* (pp. 26-43). Oxford University Press.

Yıldız, M. (2013). Okuma motivasyonu, akıcı okuma ve okuduğunu anlamanın beşinci sınıf öğrencilerinin akademik başarılarındaki rolü. *International Periodical For The Languages, Literature and History of Turkish or Turkic, 8*(4), 1461-1478.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*(5), 971-979. https://doi.org/10.3758/PBR.15.5.971

Zevin, J. (2009). Word recognition. In L. R. Squire (Ed.), *Encyclopedia of neuroscience* (pp. 517-522). Academic Press.

Pamukkale University Journal of Education
66, 329-347, 2026

R-Based Quantitative Computational Text Analysis Tool for Clinical Reading Assessment in Turkish

Esmehan ÖZER[1, 2,3,4*], Sema ACAR-ÜNALGAN[2,3,4,5], Hazal ARTUVAN-KORKMAZ[2,3,4,6], Rahime Duygu TEMELTÜRK[3,4,7,8,9], Özgür AYDIN[2,3,4,8,10]

[1] Gazi University, Gazi Faculty of Education, Department of Special Education, Türkiye
[2] Center of Excellence for Neuroscience and Neurotechnology (NÖROM), Türkiye
[3] Ankara University Brain Research Application and Research Center (AÜBAUM), Türkiye
[4] Ankara University Faculty of Language, History and Geography Linguistics Laboratory (DiLab), Türkiye
[5] İzmir Bakırçay University, Faculty of Health Sciences, Department of Speech and Language Therapy, Türkiye
[6] Ankara University, Faculty of Medicine, Department of Basic Medical Sciences, Türkiye
[7] Ankara University, Faculty of Medicine, Department of Internal Medicine, Türkiye
[8] Ankara University Institute of Health Sciences, Interdisciplinary Neuroscience Department, Türkiye
[9] Ankara University, Autism Intervention and Research Center, Türkiye
[10] Ankara University, Faculty of Language, History and Geography, Department of Linguistics, Türkiye

Abstract: *This study aims to develop a computational text analysis tool based on R programming environment to evaluate the quantitative linguistic characteristics of texts used in the clinical assessment of reading skills, and to test newly created texts using this tool. In this context, the narrative text "Double Surprise" and the informative text "Starfish" were written by researchers at three different levels of linguistic and structural complexity. Quantitative parameters such as orthographic similarity (OLD-20), bigram frequency, word frequency, morphological complexity, and readabilityin total six tests were calculated using relevant R packages ("tidytext," "vwr," and "strngram" packages). Texts M2a and M2b were determined to be clustered closest texts in terms of linguistic parameters. Thus, quantitative criteria related to the texts to be used in the assessment of reading skills were established and a text analysis tool was developed in the R language for the first time in Turkish, in addition; the produced texts were tested using this developed tool. The study provides a data-driven and reproducible method for quantitative text evaluation and contributes to Turkish reading materials for research and diagnostic applications.*

Article Details
Research Article
Received
14/01/2025
Accepted
29/10/2025
Keywords
Reading, narrative text, informative text, text similarity, text readability.

1. Introduction

Reading is a goal-directed and purposeful skill (Kucer, 2014), the primary aim of which is decoding written symbols and comprehending written language (Hoover & Gough, 1990). In this context, reading is defined as the process of perceiving, interpreting, and making sense of letters, graphemes, diacritics, and other written symbols through brain-based reading networks

*Corresponding Author: Esmehan ÖZER E-mail: esmehann@gmail.com Address: Gazi University, Gazi Faculty of Education, Türkiye

ISSN: 1301-0085 **e-ISSN:** 1309-0275 http://doi.org/10.9779/pauefd.1619827

(Çetinkaya, 2010). This process begins with the eyes and the visual system and is then rapidly and efficiently taken over by the linguistic processing system (Perfetti, 1999, 2012). In other words, reading begins when visual information is provided to the reader; visual input is then converted into linguistic input, activating grammatical and semantic processing to achieve comprehension—the ultimate goal of reading (Perfetti, 1999). Thus, both the writing system and cognitive processing mechanisms are actively employed to achieve meaning construction. Reading, a prerequisite skill in everyday life and professional contexts, plays a crucial role in academic success (Yıldız, 2013). Furthermore, it is a complex skill that requires the coordination of simultaneous and/or sequential sensory, neurological, and cognitive-linguistic processes (Rayner & Reichle, 2010). The characteristics of the words and texts being read also play a role in the realization of this multifaceted process. The literature indicates that certain quantitative variables influence reading at both the word and text levels. These include:

(i) text readability (Çetinkaya, 2010),

(ii) text type (Baştuğ, 2012),

(iii) word frequency (Rayner, 1998),

(iv) word length (Rayner et al., 2011),

(v) orthographic neighborhood size (Andrews, 1997), and

(vi) morphological complexity (Levesque et al., 2021; Verhoeven et al., 2011).

Text readability refers to the degree to which a text can be easily understood by a reader (Ateşman, 1997). Readability formulas, which are based on the structural features of texts, serve as predictive tools that classify texts according to reading difficulty or ease (Çetinkaya, 2010). In Turkish, several readability formulas have been proposed, including those by Ateşman (1997), Bezirci and Yılmaz (2010), and Çetinkaya and Uzun (2010) [Ministry of National Education (MoNE), 2021]. However, these formulas have been criticized for yielding inconsistent results across the same texts and for relying solely on syllable, word, and sentence counts (MoNE, 2021).

Another factor influencing reading performance is text type, as the structural properties of a text are crucial for comprehension (Hamzadayı & Batmaz, 2022; Hebert et al., 2016). Informational and narrative texts are considered as two primary text types (MoNE, 2019). A narrative text consists of interrelated elements, either explicitly stated or implicitly embedded in the text, while an informational text conveys knowledge, presents ideas, and offers explanations or suggestions on a particular topic (Başaran & Akyol, 2009).

Word frequency also plays a significant role in reading fluency. Frequently encountered words are recognized more quickly (Zevin, 2009). For instance, a word appearing 50 times per million is read faster than one appearing 10 times per million. Consequently, readers respond to high-frequency words more rapidly and accurately than to low-frequency words (Yap & Balota, 2015). In Turkish, Erten et al. (2014) found that low-frequency words cause an average delay of approximately 92 milliseconds in word recognition.

Word length—measured by the number of letters—also affects reading. Longer words are associated with longer fixation durations during reading (New et al., 2006). When readers encounter long words, they tend to fixate for a longer time and are less likely to skip them. Functional words such as “for,” “and,” or “but” are typically processed more quickly than longer content words, which increases overall reading time (Rayner, 1998, 2009).

Another variable that influences readers' reading process is orthographic neighborhood size. It was first described by Coltheart et al. in 1977 (Yarkoni et al., 2008). Orthographic

neighborhood size is defined as the number of new words of the same length that can be formed by changing a single letter in a word. For example, the word “süt” (milk) has seven orthographic neighbors: “süz” (strain), “sür” (drive), “süs” (decoration), “sat” (sell), “set” (set), ‘tüt’ (smoke), and “küt” (chunk). It has been reported that in many languages, words are read more quickly when they have many orthographic neighbors (Andrews, 1997). In Turkish, however, it has been found that the orthographic neighborhood dimension has no effect on reading (Erten et al., 2014). On the other hand, the orthographic adjacency dimension is created by changing only a single letter and does not cover other operations such as transposing letters within a word, deleting a letter from a word, or adding a new letter to a word. Due to these limitations, Yarkoni et al. (2008) developed a new orthographic similarity measure. “Orthographic Levenshtein distance-20” [OLD-20] refers to the average number of changes made to create 20 new words with the minimum number of letter changes (swapping, adding, deleting) in a word. The lower the OLD-20 value of a word, the fewer letter changes are required to derive new words. While studies indicate that words with low OLD-20 values are easier to read (Yarkoni et al., 2008), the opposite has been shown to be true in Turkish (Erten et al., 2014). Therefore, it is thought that OLD-20 has an inhibiting rather than a facilitating effect on Turkish reading.

A sequence consisting of two units is referred to as a “bigram.” For example, the Turkish word ‘masa’ (table) contains the bigrams “ma,” “as,” and “sa.” Bigram frequency is a measure of the frequency of a bigram in an orthography (Santangelo, 2023). For example, in English, the bigrams “th,” “he,” and “in” are used more frequently in texts than the bigrams ‘qu’ and “nk.” Bigram frequency has been reported to have effects on lexical decisions (Muncer et al., 2014; Schmalz & Mulatti, 2017) and word naming (Muncer et al., 2014) tasks. While the effects of bigram frequency on reading continue to be debated, strong evidence has been reported regarding its speed-enhancing effect on audible reading (Schmalz & Mulatti, 2017).

There is a significant relationship between morphological awareness and the identification of morphologically complex words (Carlisle, 2000). Therefore, it has been found that awareness of morphemes, the smallest meaningful units of language, is related to comprehension skills, which is the ultimate goal of reading, and that this relationship is influenced by morphologically complex words (Liu et al., 2024). For example, the English word “work” can be inflected as “works,” “workers,” and “working,” and can also take prefixes and suffixes to become “rework” and “worker.” At the same time, the word ‘work’ is a compound word and can also appear as “workplace” (Borleffs et al., 2019). The ability to analyze these words by resolving their morphological complexity is reflected in children's reading and comprehension skills (Levesque et al., 2021; Verhoeven et al., 2011). In this context, it can be stated that the morphological complexity characteristics of words in texts play an important role in reading.

Texts are used in the assessment and development of reading skills (MoNE, 2019). A text is a piece of language that carries meaning as a whole (Pilav & Oğuz, 2013). It is also defined as structures in which information, feelings, and thoughts are placed according to various forms, expressions, and punctuation features (Güneş, 2013). Texts are the most important sources and materials for reading. The booklet “Criteria and Explanations to Be Used as a Basis for Evaluation in the Examination of Draft Textbooks and Educational Tools and Their Electronic Content” (2023) contains criteria related to language and style, compliance with written language standards, and meaning and expression for texts to be included in textbooks. In this context, the booklet explains that foreign words should not be used in the texts, that the language used should be appropriate for the grade level and aimed at enriching vocabulary, that the expression should be simple and straightforward, that it should not slow down reading, and that it should be easy to read. Although many variables affecting reading skills in the field literature are mentioned, such as word frequency, length, orthography, and text readability, there is no clear information on how Turkish text writers create texts appropriate to the student's level in

relation to these variables. It can be said that texts are written more according to the writers' experience and feelings. On the other hand, text difficulty and the writing of text types are among the topics most in need of study in the literature (Duran & Kargın, 2022). At the same time, it is seen that the factors affecting reading in the literature are not addressed quantitatively as a whole in the context of the text.

The creation of texts within the context of quantitative criteria at the text and word level that affect the speed and duration of the reading skill listed above is critical for the clinical assessment of reading skills. It is thought that texts not created in accordance with these criteria in their entirety cause problems in the clinical assessment of reading skills. These criteria directly affect the reading process. Therefore, it is important to use texts created within the context of these criteria in clinical assessment for populations with reading difficulties. The use of random texts appropriate only for the individual's grade level and/or age, without considering these criteria, plays a role as a factor that increases the heterogeneity problem observed in dyslexia and other reading difficulties, stemming from the multi-component complex structure of existing reading skills. To our knowledge, there are no Turkish texts that are widely used and developed by controlling all the quantitative criteria listed above in the clinical assessment of reading skills. One of the aims of this study is to contribute two texts to literature, developed in the context of these quantitative criteria affecting reading, which are informative and narrative, to address this gap in the literature and clinical practice. However, it is thought that the use of these two texts solely in clinical assessments may be limiting as the individual's age, grade level, and other characteristics change, and as reading skills need to be reassessed at regular intervals. Therefore, it will be possible to overcome this limitation by using tools that allow for a comprehensive evaluation of the texts to be used in the context of the criteria listed above.

Kırkgöz and Ünaldı (2012) compared the lexical networks of students whose native language is Turkish and who are learning English as a foreign language with those of students whose native language is English. They used coh-metrix, an online database based on corpora and various criteria such as readability and syntax, to analyze the English texts written by these students. The validity of coh-metrix was also tested in the study. The results revealed limitations in the lexical cohesion of sentences in the texts written by students whose native language is Turkish and who are learning English as a foreign language. Furthermore, some indicators used in coh-metrix were able to distinguish between texts written by students whose native language is English and those who learned English later. Arnold et al. (2019) explain how different text analysis packages in R can be combined and emphasize the potential of different text analysis packages in R to serve as text analysis tools. Similarly, Aziz et al. (2010) not only determined the general readability of reading materials but also presented a more comprehensive analysis approach that provides information on sentence and word difficulty. The researchers suggest composite calculation tools to determine text, sentence, and word difficulty in an objective and reliable manner. The use of different text analysis packages created in the R language in fields such as linguistics and educational sciences as an objective text analysis tool may inspire their potential use in areas related to the clinical assessment of reading, such as special education, language and speech therapy, child and adolescent mental health and disorders, and clinical psychology.

No tool has yet been found in the literature that allows Turkish texts used in the clinical assessment of reading skills to be evaluated holistically using quantitative criteria. In the national literature, there are many standardized batteries and/or scales with proven validity and reliability used to assess reading performance for both educational and clinical purposes. These batteries are quite valuable in terms of assessing reading performance in various clinical populations and serve as a guiding manual for diagnosis. The aim of this study is not to develop

such a scale, but rather to develop an R program-based analysis tool that performs analyses providing experts with information about the linguistic characteristics of texts used in reading assessment. Thus, experts will be able to analyze the texts they wish to use in this program thanks to the codes we have developed and obtain information about the relevant linguistic criteria. Experts who wish to do so can use the texts we present as sample products in this study for their evaluations, or they can use the text analysis tool codes we present in this study to obtain information about the linguistic features of any text they wish to use. In this context, the tool planned to be developed in this study will be a tool that can guide the expert/evaluator about the linguistic features of texts used in clinics, rather than a clinical evaluation tool. Unlike standardized scales in the field, the text analysis tool we will develop in this study will enable the expert/evaluator to create texts with different linguistic features and use texts with varying levels of difficulty and complexity in terms of linguistic features during educational interventions. Therefore, since quantitative criteria are needed for texts to be used in the clinical assessment of reading skills, this study aims to develop a text analysis tool using the R language and to test the generated texts with this tool. In this context, answers to the following questions were sought:

(a) Can a self-updating text analysis tool be developed to evaluate texts quantitatively and holistically?

(b) Can texts written by researchers be tested with this developed tool?

(c) Can the tool provide quantitative characteristics of the text to specialists who wish to know the level of the text they will write or use for the clinical assessment of reading skills?

(d) Can the tool perform calculations based on lexical parameters (orthographic, word frequency, letter-bigram frequency, morphological complexity, word length) and readability criteria in texts?

(e) What are the similarities between texts developed by researchers?

2. Method

2.1. R Programming Language and Text Analysis

Computational text analysis provides an environment where processes that would take time to perform manually, especially with long texts, can be carried out in a shorter time with a reduced likelihood of error. Calculating various quantitative measures related to orthographic and lexical features for texts to be used in the clinical assessment of reading skills, calculating the readability of texts, and revealing similarities between texts require a labor-intensive manual process. In this study, we provide an overview of the general steps and procedures related to computational text analysis using R software.

Unlike other programming languages, R was specifically designed for statistical analysis, making it an extremely suitable environment for data science applications. As an open-source platform, R is a programming language that includes a wide variety of text analysis packages (Kwartler, 2017; Kumar & Paul, 2016) and has a large user community. In this study, we present an example of text analysis using the aforementioned text analysis packages and the code we developed. R packages are a collection of software libraries produced by R's large user community. Each package extends the functionality of the basic R language and core packages and provides users with a set of easy-to-use functions. In this study, the tidytext package (Silge & David, 2016) and the textTinyR package (Mouselimis, 2021) are used for text editing, while the vwr package (Keuleers, 2013) and the strngram package (van Heuven, 2024) are used for orthographic calculations.

Additionally, packages such as the stats package and the rstatix package (Kassambara, 2023) will be utilized for statistical analyses. The Comprehensive R Archive Network (CRAN), the most well-known package repository, currently hosts over 10,000 published packages, and all packages listed above, except for the stats package, are available on CRAN. These packages, which can be easily and securely loaded from the R environment, form a solid bridge between developers and users of new analysis tools, providing a highly suitable programming environment for scientific collaboration.

2.2. Text Creation

2.2.1. Text Writing

The narrative text titled "Double Surprise" included in the study was written by the first and third authors of the study at three different levels by modifying the text in terms of word length, morphological complexity, and orthographic proximity variables without distorting the event or information pattern. Similarly, the informative text "Starfish" was written at three different levels by the second and fourth authors of the study. As a result, the study includes a total of six different levels of text, three of which are narrative and three of which are informative. The text themes were determined in accordance with MoNE themes.

2.2.2. Evaluation of Texts (Expert Opinions)

After the texts were created, opinions were obtained from 8 experts: 2 faculty members working in the field of Special Education on reading skills, 1 faculty member working in the field of Linguistics, 1 faculty member working in the field of Turkish Language and Literature, 1 faculty member working in the field of Turkish Language Education, and 1 faculty member working in the field of Teaching Turkish as a Foreign Language, as well as two classroom teachers. The experts' professional experience ranged from 10 to 28 years. The expert opinion forms developed by the researchers were sent to the experts electronically. This form includes various criteria such as "the text's consistency with the theme, the text's coherence, the text's flow of events, and the text's narrative fluency being appropriate for 2nd and 3rd grade levels, and the words and sentences in the text being appropriate for 2nd and 3rd grade levels." Experts reviewed each text in these forms in turn for compliance with the specified criteria and expressed their opinions by rating them on a scale of 1 (not at all appropriate) to 5 (very appropriate). They also made corrections for the texts as they deemed appropriate. The texts were revised based on the experts' opinions. The expert opinions were refined, and the texts were finalized prior to analysis. General information about the texts is presented in Table 1, and the texts, along with their editing and analysis codes, can be accessed via the GitHub page. The texts and text editing and analysis codes can be accessed from the GitHub page (https://github.com/ozguraydin66/PreTXT/tree/main/documents).

Table 1. *General Information About Texts*

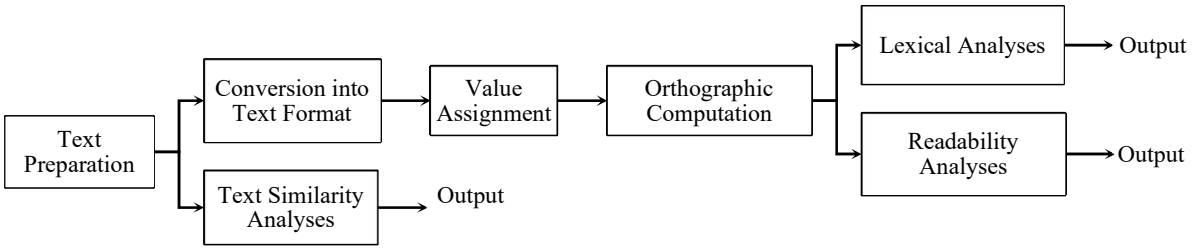
Text Code	Text Type	Text Title	Number of Word Forms*	Number of Lexemes*	Number of Sentences
M1*a*	Narrative	Double Surprise	171/200	137/194	26
M1*b*	Informational	Starfish	134/176	112/170	23
M2*a*	Narrative	Double Surprise	120/149	93/145	18
M2*b*	Informational	Starfish	113/148	96/145	18
M3*a*	Narrative	Double Surprise	163/183	125/179	30
M3*b*	Informational	Starfish	145/172	122/170	22

* Numbers related to word form and lexeme count are presented as Different Words/Total Words.

2.3. Text Organization

In this study, the stages shown in Figure 1 were followed during the text analysis process. These stages enable the analysis and visualization of various characteristics of the text by converting texts consisting of individual words into data frames.

Figure 1. *Text Analysis Stages*



At the first stage, texts were prepared in .txt format, and the units in the text were organized both as word forms and lexemes. The lemmatization process of the words in the texts was excluded from the automation process. After organizing the texts in these two formats, the texts were converted into regular text format. Creating regular text means converting the texts into a data frame. In regular data, each variable must be presented in a column, and each row must present a single observation (Wickham, 2014). At the preparation of the texts, the tidytext package (Silge and David, 2016) in R (R Core Team, 2023) was used to create a data framewith a single sentence in each row, as shown below, using the split_sentence() function (line 2).

```
1  textfile = readLines('M1.txt')
2  TextSentence=split_sentence(textfile[2])%>%
3    as.data.frame() %>%
4    purrr::pluck(1) %>%
5    tibble() %>%
6    purrr::set_names('text') %>%
7    mutate(line=row_number(),
8           WordCountinSent=stri_count_words(text),
9           SentCount=length(text),
10          TextType= 'M1')
```

Then, using the unnest_tokens() function (line 4), a data frame is created where each row contains a single word:

```
1  Word <-list()
2  for(z in TextSentence$line){
3    Word[[z]] = TextSentence[z,]  %>%
4      unnest_tokens(word, text) %>%
5      dplyr::count(word, sort = TRUE)
6    Word[[z]]=as.data.frame(Word[[z]])
7    Word[[z]]$LineNum = z}
8  WordList=dplyr::bind_rows(Word)
```

For the value loading phase, lists containing the number of tokens for each word form in the data frames converted to plain text format are also transferred to the data.

For the orthographic calculation phase, the average OLD-20 value for the 20 neighboring words in the Turkish word list (Erten et al., 2014) was calculated for each word form and lexeme using the old20() function in the vwr package (Keuleers, 2013) (line 1). In this stage, the bigram values of the words were also calculated using the strngram package (van Heuven, 2024). First, the get_ngram_frequencies() function was used to determine the frequencies of different forms

(type frequency) and total frequencies (token frequency) of bigrams obtained from the words in the word list (lines 2-3). Subsequently, these values were used within the ngram_frequency() function to calculate the averages of bigrams in word forms and lexemes in the texts (lines 5-7):

```
1 df$OLD20 <- old20(df$word, lexicon[,1])
2 z =get_ngram_frequencies(lexicon$V1, lexicon$V3,
3                          type = "bigram", position_specific = TRUE)
4 newcol = ncol(df) +1
5 df[,newcol] <- ngram_frequency(df$word, z, type = "bigram",
6                                position_specific = TRUE,frequency = "token",
7                                func = "mean", progressbar = TRUE)
```

2.4. Text Analyses

The first stage of text analysis involves analyzing the similarity between texts. The similarity referred to here is not a syntactic or semantic similarity between texts, but rather a formal (lexical) similarity. Therefore, the measurement of text similarity is based on close word proximity and word repetitions. One such analysis, the Jaccard coefficient, measures the degree of overlap between two sets and is calculated as the ratio of the number of shared attributes (words) between Bu and Bv (the similarity of text u and the similarity of text v) to the number possessed by Bu or Bv units (see 1). For example, given the binary indicator vectors of two sets Bu = {living, star, arm, five} and Bv = {water, star, joint, old}, the cardinality of their intersection is 1 and the cardinality of their union is 3, and the Jaccard coefficient is 1/3. Another similarity measure used in this study, cosine similarity, measures the angle between two ranked vectors, where a narrower angle indicates greater similarity and a wider angle indicates lesser similarity. In the formula, R(u, i) represents the rating of element i in text u, and B(u, v) represents the number of common rated elements between texts u and v (see 2). For both calculations, the JACCARD_DICE() and cosine_distance() functions in the textTinyR package (Mouselimis, 2021) were used in R.

$$similarity(u,v)^{Jaccard} = \frac{|B_u \cap B_v|}{|B_u \cup B_v|} \quad (1)$$

$$similarity(u,v)^{Kosinus} = \frac{\Sigma_{i \in B(u,v)} - R(u,i) \cdot R(v,i)}{\sqrt{\Sigma_{i \in B(u,v)}} R(u,i)^2 \cdot \sqrt{\Sigma_{i \in B(u,v)}} R(v,i)^2} \quad (2)$$

Below, after punctuation marks have been removed from the text (lines 3-5), you can see how the JACCARD_DICE() and cosine_distance() functions (lines 7-11) are used.

```
1  text.path <- list.files(path=file.path(rootdir, paste0("texts/",ListType)),
2                          pattern = ".txt", full.names = TRUE)
3  text.file <-list()
4  for(i in 1:length(text.path)){
5    text.file[[i]] <- removePunctuation(readLines(text.path[i]))}
6
7  JScore = JACCARD_DICE(strsplit(text.file[[1]][2], "\\s+"),
8                        strsplit(text.file[[2]][2], "\\s+"),
9                        method = 'jaccard', threads = 1)
10 CScore = cosine_distance(text.file[[1]][2],
11                         text.file[[2]][2], split_separator= "")
```

At the lexical analysis phase, in addition to the word frequency, syllable count, morphological unit count, and word length data obtained during the data integration phase, analyses of OLD-20 and bigram data related to orthographic calculations are performed. In these analyses, the Shapiro–Wilk test was used in R to test whether the data related to the parameters listed above

exhibited a normal distribution, followed by the Wilcoxon Signed Rank test to see if there was a significant difference between the texts in terms of parameters.

On the other hand, three different readability analyses developed for Turkish have been used as a basis in readability analyses that determine the degree to which texts in any language are readable by readers. The formula developed by Ateşman (1997) to measure the readability of Turkish texts (see 3) is based on the FRES (Fresch Reading Ease Score; Flesch, 1948) formula. In this formula, the average number of words and sentence length are added to the formula based on the number of syllables (H), number of words (K), and number of sentences (C). The formula results in values between 90 and 100 indicating very easy texts, values between 70 and 89 indicating easy texts, values between 50 and 69 indicating moderately difficult texts, values between 30 and 49 indicating difficult texts, and values between 1 and 29 indicating very difficult texts. Another formula used in this study to measure readability is the one developed by Çetinkaya and Uzun (2010), similar to Ateşman's, based on the average number of words and sentence length (see 4). In this formula, values between 0 and 34 indicate the frustration level (10th, 11th, and 12th grades), values between 35 and 50 indicate the educational level (8th and 9th grades), and values greater than 51 indicate independent reading (5th, 6th, and 7th grades). The latest readability formula developed for Turkish is Bezirci and Yılmaz's (2010) formula, which is based on the average number of words in sentences and the average number of words with 3, 4, 5, and 6 syllables (see 5). According to this formula, it explains the readability levels of texts corresponding to a specific grade level: 1-8: elementary school; 9-12: high school; 13-16: undergraduate; and ≥16: academic.

$$OS = 198.825 - \left(40.175\ \frac{H}{K}\right) - \left(2.610\ \frac{K}{C}\right) \quad (3)$$

$$OS = 118.823 - \left(25.987\ \frac{H}{K}\right) - \left(0.971\ \frac{K}{C}\right) \quad (4)$$

$$OS = \sqrt{\frac{K}{C}\ ((H3\ \ 0.84) + (H4\ \ 1.5) + (H5\ \ 3.5) + (H6\ \ 26.25))} \quad (5)$$

To calculate the above formulas in R, a database named ‘df’ was first created, in which the variables for each text were calculated. In this database, the number of syllables (H), number of words (K), and number of sentences for each text, as well as characteristics such as the average number of words per syllable, were calculated and processed individually in columns. Then, all three formulas were calculated in R (R Core Team, 2023) using the mutate() function in the dplyr package (Wickham et al., 2023):

```
Calculate <- df %>%
  group_by(TextType) %>%
  mutate(
    Atesman   = 198.825-(40.175*(H/K))-(2.610*(K/C)),
    Cetinkaya = 118.823-(25.987*(H/K))-(0.971*(K/C)),
    Bezirci   = sqrt((K/C)*((H3*0.84)+(H4 *1.50)+(H5 *3.50)+(H6 *26.25)))) %>%
  select(c(21:23)
  )
```

After conducting text similarity analyses, lexical analyses, and readability analyses on the texts, two texts were selected that were most similar to each other in two different text types and most suitable for 2nd and 3rd grade students, one informative and one narrative. These two texts were checked for word and sentence count. Both texts contain 150 words and 22 sentences, and all sentences in both texts are grammatically correct. Additionally, the number of simple and complex, positive and negative, action and noun sentences in the texts are arranged to be close

to each other. The minimum and maximum number of words in the sentences in the texts are also set to be close to each other.

At this stage, in addition to the above analyses, hierarchical clustering was also performed using the hclust() function in the stats package (R Core Team, 2023) in R. To do this, the average number of utterances (M), number of words (K), and number of sentences (C) were first selected from the database in lines 1-2 below, and then the data was normalized in lines 4-6. The distance matrices were calculated for the normalized data (nor) (line 7), and hierarchical clustering was performed and plotted using the hclust() function (line 8) (line 9):

```
1  Clust <- df %>%
2    select(TextType, M, K, C)
3  z <- Clust[,-c(1,1)]
4  means <- apply(z,2,mean)
5  sds <- apply(z,2,sd)
6  nor <- scale(z,center=means,scale=sds)
7  distance = dist(nor)
8  df.hclust = stats::hclust(distance)
9  plot(df.hclust,labels=Clust$TextType, main="")
```

3. Findings

3.1. Text Similarity Findings

In examining the similarities between the six texts, the texts were compared separately within themselves, organized as lexical units and organized as word forms. At this stage, since the aim was to reveal how similar or how different texts of the same type are to each other, similarity comparisons were made within the same text types. Therefore, Jaccard coefficients and cosine coefficients were compared in different text pairs (M1, M2, and M3) but within the same text types (a and b) (i.e., M1a vs. M2a, M2a vs. M3a, M1a vs. M3a, etc.). As shown in Table 2, the Jaccard coefficients and cosine coefficients indicate that, in narrative texts, M1 and M3 are closer to each other, while M2 is almost equidistant from M1 and M3. In informative texts, M1 and M2 are closer to each other, while M3 is almost equidistant from M1 and M2. In general, text similarities decrease in word form formats where words are presented with morphological inflections. The ratio of different words to the total number of words is higher in narrative texts and lower in informative texts. This means that there is greater word variety in informative texts.

Table 2. *Jaccard and Cosine Similarity Values and Density Ratios for Texts*

	Jaccard Similarity				Kosinus Similarity				Density
	M1*a*	M2*a*	M1*b*	M2*b*	M1*a*	M2*a*	M1*b*	M2*b*	
				Lexeme Level					
M2*a*	.38				.65				.64
M3*a*	**.70**	.40			**.86**	.74			.70
M2*b*			**.60**				**.84**		.66
M3*b*			.39	.35			.73	.71	.72
	.71		.66						
				Word Form Level					
M2*a*	.28				.50				.81
M3*a*	**.61**	.34			**.76**	.59			.89
M2*b*			**.54**				**.75**		.76
M3*b*			.20	.21			.58	.52	.84
Density	.86	–	.76	–	–	–	–	–	–

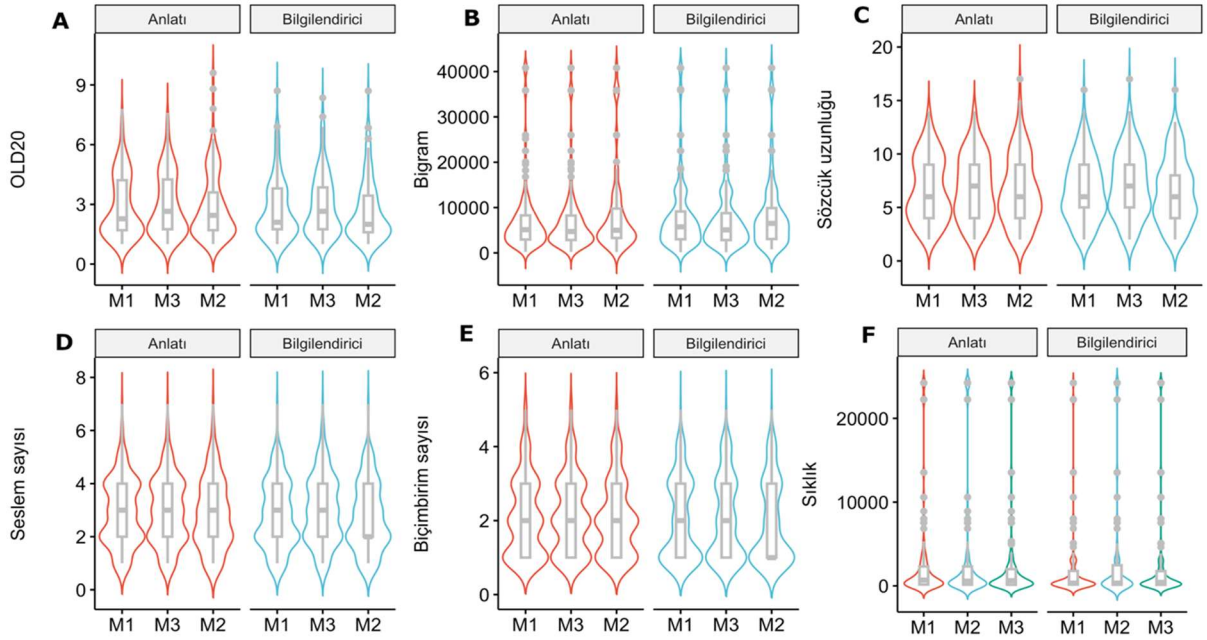
3.2. Lexical Analysis Findings

As shown in Table 3 and Figure 2 (panels A-B), there is no significant difference in text types for the same text pairs (i.e., M1, M2, and M3) in terms of orthographic features in the OLD-20 and bigram parameters. Similarly, there is no significant difference between text types in terms of lexical features such as word length, number of syllables, and number of morphemes (see Table 4 and Figure 2C-2E). In the analysis of lexical units, OLD-20 and bigram parameters also do not lead to significant differences between groups in terms of orthography. The groups also exhibit a similar appearance in terms of the general frequency of occurrence of the words used in the corpus.

Table 3. *Wilcoxon Signed-Rank Test Results for Lexeme-Level Parameters (Word Frequency, OLD-20, and Bigram Frequency)*

Text	Text Types	*N*	Word Frequency			OLD-20			Bigram		
			Mean (SE)	*Z*	*p*	*Mean (SE)*	*Z*	*p*	*Mean (SE)*	*Z*	*p*
M1*a*	Narrative	171	2918 (410)	-1.68	.09	1.61 (.04)	-.01	.99	10939 (667)	-.19	.84
M1*b*	Informative	134	2263 (371)			1.58 (.04)			10409 (650)		
M2*a*	Narrative	120	3399 (552)	-1.29	.19	1.70 (.05)	-.97	.33	11228 (834)	-.06	.94
M2*b*	Informative	113	2871 (494)			1.60 (.05)			10808 (746)		
M3*a*	Narrative	163	2725 (423)	-.98	.32	1.62 (.04)	-1.21	.22	10328 (663)	-.27	.78
M3*b*	Informative	145	2580 (435)			1.69 (.05)			10389 (648)		

Figure 2. *Comparison of Lexical Parameters*



Note: Panels A, B, C, D, and E show the parameters OLD-20, bigram, word length, number of syllables, and number of morphemes, respectively. Panel F shows the word frequency parameter in lexical units."Anlatı": Narratives; "Bilgilendirici": İnformative, "Seslem Sayısı": Number of syllables, "Biçimbirim sayısı": number of morphemes. "Sıklık": frequency, "Sözcük uzunluğu": Word lenght.

Table 4. *Wilcoxon Rank Sum Test Showing Differences in OLD-20, Bigram, Word Length, Number of Syllables, and Number of Morphemes Parameters Between Texts*

Text	Text Type	*N*	OLD-20			Bigram			Word Lenght			Number of Syllables			Number of Morphemes		
			Mean (SE)	*Z*	*p*	*Mean (SE)*	*Z*	*p*	*Mean (SE)*	*Z*	*p*	*Mean (SE)*	*Z*	*p*	*Mean (SE)*	*Z*	*p*
M1*a*	Narrative	200	2.87 (.11)			7405 (560)			6.67 (.21)			2.89 (.08)			2.08 (.07)		
				-.72	.46		-1.01	.31		-.06	.94		-.74	.45		-.67	.50
M1*b*	Informative	176	2.72 (.11)			8032 (606)			6.63 (.22)			2.80 (.09)			2.01 (.08)		
M2*a*	Narrative	149	2.86 (.13)			8763 (780)			6.55 (.26)			2.84 (.10)			2.04 (.08)		
				-1.05	.29		-.83	.40		-.06	.95		-.89	.36		-1.20	.22
M2*b*	Informative	148	2.64 (.12)			8595 (682)			6.47 (.24)			2.70 (.10)			1.94 (.09)		
M3*a*	Narrative	183	2.96 (.11)			7058 (559)			6.78 (.22)			2.93 (.09)			2.13 (.07)		
				-.17	.86		-.75	.45		-.79	.42		-.41	.67		-.97	.33
M3*b*	Informative	172	2.92 (.11)			7405 (551)			7.05 (.23)			3.00 (.09)			2.05 (.08)		

3.3. Text Readability Findings

Three readability formulas developed for Turkish to determine the readability of texts were applied to six separate texts (see Ateşman, 1997; Bezirci & Yılmaz, 2010; Çetinkaya & Uzun, 2010). In order to perform these analyses, the number of syllables, words, and sentences in the texts were determined (see Table 5). As shown in Table 5, the readability levels of the texts are intermediate according to Ateşman (1997), at frustration level according to Çetinkaya and Uzun (2010), and high school level according to Bezirci and Yılmaz (2010). The M1b informative text is at an intermediate level according to Ateşman (1997), at a frustration level according to Çetinkaya and Uzun (2010), and at an elementary school level according to Bezirci and Yılmaz (2010). The narrative text M2a is at an intermediate level according to Ateşman (1997), at a disability level according to Çetinkaya and Uzun (2010), and at a high school level according to Bezirci and Yılmaz (2010). The M2b informative text is at an intermediate level according to Ateşman (1997), at a disability level according to Çetinkaya and Uzun (2010), and at an elementary school level according to Bezirci and Yılmaz (2010). The narrative text M3a is at an intermediate level according to Ateşman (1997), at a disability level according to Çetinkaya and Uzun (2010), and at an elementary school level according to Bezirci and Yılmaz (2010). The informative text of M3b is at an intermediate level according to Ateşman (1997), at a disability level according to Çetinkaya and Uzun (2010), and at a high school level according to Bezirci and Yılmaz (2010).

Using the parameters in Table 4, hierarchical clustering was employed to determine which text type pairs are closer to each other. The hierarchy obtained using the hierarchical clustering method in R via the hclust() package within the stat package is shown in Figure 3. As seen in Figure 3, the M2a and M2b texts exhibit the most appropriately clustered text appearance. Based on this determination and previous analyses, the closest pair was submitted for expert review. Following the second expert review, the experts unanimously reported that the texts were appropriate for the 2nd and 3rd grade levels in terms of theme, subject matter, and information flow.

Figure 3. *Hierarchical Clustering of Texts (Distance, Hclust (*,'Complate')*

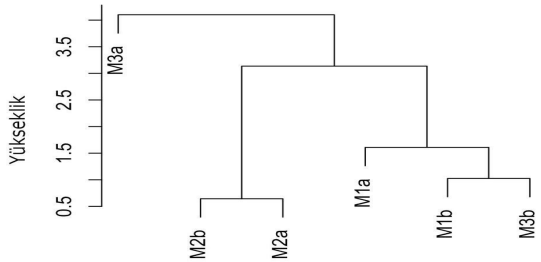


Table 5. *Text Readability Findings*

	M1*a* *Narrative*	M1*b* *Informative*	M2*a* *Narrative*	M2*b* *Informative*	M3*a* *Narrative*	M3*b* *Informative*
Syllable	24.04	22.91	24.83	24.56	19.47	24.27
Polysyllabic word	7.12	7.00	7.28	7.50	5.70	7.18
Bisyllabic word	2.5	2.70	2.5	2.94	1.83	2.36
Trisyllabic word	1.54	1.74	2	1.94	1.4	1.86
Tetrasyllabic word	2.19	1.74	1.72	1.72	1.7	1.91
Pentasyllabic word	0.69	0.74	0.78	0.83	0.6	0.82
Hexasyllabic word	0.19	0.09	0.28	0.06	0.17	0.23
Total number of words	200	178	150	150	183	176
Total num. of sentences	26	23	18	18	30	22
Ateşman (1997)	53.2	59.68	57.35	58.69	54.7	56.05
Çetinkaya (2010)	30.14	34.37	33.29	34.16	29.97	32.21
Bezirci & Yılmaz (2010)	9.63	8.32	10.91	8.46	7.89	10.3

4. Discussion, Conclusion, and Suggestions

The purpose of this study was to develop a computational text analysis tool in the R programming environment for the first time in Turkish to evaluate the quantitative linguistic characteristics of texts used in the clinical assessment of reading skills and to test six texts created by the researchers using this tool. In addition, similarity analyses were performed on the texts and compared. By integrating linguistic variables such as orthographic similarity (OLD-20), bigram frequency, word frequency, morphological complexity, and readability, the tool was shown to offer an analytical framework that allows experts to evaluate text difficulty objectively and reproducibly.

In the study, six texts—three narrative (*Çifte Sürpriz* / Double Surprise) and three informational (*Deniz Yıldızları* / Starfish) — were developed at different levels of linguistic and structural complexity. The quantitative parameters of these texts were calculated using the R packages tidytext, vwr, strngram, and textTinyR, and their orthographic, lexical, and readability characteristics were examined. The R package developed as part of this study is available at https://github.com/ozguraydin66/PreTXT/. Additionally, the web application version of this package can be accessed at https://oaydin.shinyapps.io/pretxt/.

The overall results showed that the texts were similar in terms of orthographic and lexical properties, confirming that the manipulations made for each version were consistent across both genres. According to the three Turkish readability formulas (Ateşman, 1997; Bezirci & Yılmaz, 2010; Çetinkaya & Uzun, 2010), all texts were found to be of moderate difficulty, indicating that they were suitable for use with children at the early primary level. Hierarchical clustering analyses demonstrated that the texts M2a and M2b were the most similar in terms of quantitative parameters. The findings obtained from the computational analyses were supported by the expert evaluations, which revealed that these two texts were considered the most appropriate in terms of linguistic structure, theme, and conceptual clarity for use with students at the second and third grade levels.

The text similarity findings of the study were presented in two dimensions: Lexical units and word forms. In this context, it was determined that M1 and M3 are closer to each other in narrative texts, while M2 is almost equidistant from M1 and M3. In informative texts, it was found that M1 and M2 are closer to each other, while M3 is almost equidistant from M1 and M2. Indeed, it can be stated that narrative and informative texts are quite similar in terms of lexical units. However, when text similarities are examined in terms of word form, it is seen that informative texts have greater word diversity.

The lexical analysis of the texts was examined in terms of word form. It was shown that there were no significant differences between OLD-20 and bigram in terms of orthographic features in the parameters of word length, number of syllables, and number of morphemes. Similarly, the lexical units of the texts did not show any significant differences between OLD-20 and bigram in terms of orthography. Therefore, given that OLD-20 is known to have an inhibitory effect on Turkish reading (Erten et al., 2014), it is important that OLD-20 values do not show significant differences in the texts in terms of both word form and lexical unit. In addition, since it is known that the presence of longer words in texts will be reflected in readers' reading times (New et al., 2006; Rayner, 1998, 2009), it was found that the texts did not differ in terms of the word length parameter. Furthermore, it was determined that the words used in the texts were generally similar in terms of their frequency of occurrence in the corpus. Therefore, since the frequent occurrence of a word in a language speed up its recognition (Erten et al., 2014; Zevin, 2009), the similarity of the frequencies of occurrence of the words in the texts in the corpus has been ensured.

When examining the findings on text readability, the texts included in the study were found to be at an intermediate level according to Ateşman (1997), at the frustration level according to Çetinkaya and Uzun (2010), and at the primary and secondary school levels according to Bezirci and Yılmaz (2010). Hierarchical clustering was applied to these texts using the OLD-20, bigram, word length, number of syllables, and number of morphemes parameters to calculate which text type pairs were closer to each other. Indeed, it was determined that the M2a and M2b texts were the closest clustered texts. Thus, the study presents a tool that experts such as researchers, teachers, and speech therapists can use in the text selection process to both assess students' reading skills and intervene appropriately based on those skills. Additionally, the study examines the criteria and explanations to be used as a basis for evaluation in the review of draft textbooks, educational tools, and their associated electronic content, as outlined in the MoNE Directorate of Education and Training's “Criteria and Explanations for the Evaluation of Draft Textbooks, Educational Tools, and Their Associated Electronic Content” (2023) regarding language and style, compliance with written language standards, and meaning and expression for texts to be included in textbooks. While analyses have been conducted for texts in other languages in the field literature (Arnold et al., 2019; Aziz et al., 2010; Kırkgöz & Ünaldı, 2014), this study is the first to analyze Turkish texts using a computational tool based on the R language enviroment.

In this study, a computational linguistic text analysis tool is developed and recommended for researchers and experts. Since this is the first such study conducted with Turkish texts, the initial parameters that can serve as a reference have been identified in this study. Thus, the first values related to text similarity, lexical analysis, and text readability in Turkish have been calculated in this study. Consequently, the features of M2a and M2b texts that are relevant to assessing reading ability are presented. Therefore, the R-based text analysis tool developed in this study focuses on the linguistic features of texts, unlike other scales commonly used in assessing reading ability. The scales used in the literature provide valuable information, particularly in determining reading performance in clinical populations, but they provide limited data on the structural linguisitc characteristics of the texts used. The tool developed in this research will

provide researchers and experts planning to conduct reading assessments with valuable data on various linguistic features of texts, such as lexical analysis and text readability, for text selection. Thus, researchers and experts will be able to make informed decisions about text selection by obtaining detailed information about the linguistic features of the texts to be used in the assessment process.

The findings of the R-based text analysis tool developed demonstrate that the tool works reliably and practically in evaluating texts. It has been demonstrated that, rather than relying on a single criterion, different linguistic parameters can be examined using a holistic approach. Indeed, proposed analyses conducted on the M2a and M2b texts confirm the measurable comparison capabilities offered by the tool. In this respect, the study, rather than developing a direct scale for clinical assessment purposes, offers a complementary approach that provides experts and researchers with information about the quality of texts used or potentially usable in existing scales. Therefore, it can be stated that the developed tool has the potential to guide evaluators both in the selection of texts to be used in educational interventions and in clinical applications.

In the study, a computational text analysis tool developed for the first time in Turkish using the R language. This R-based text analysis tool was developed and proposed to evaluate the quantitative linguistic characteristics of texts used in the clinical assessment of reading skills. In addition, six texts written by the researchers were tested using this developed tool and similarity analyses of the texts developed by the researchers were also conducted. The study is original and innovative in these respects. Furthermore, it is anticipated that the proposed tool will be applied to students and undergo validity and reliability studies, progressing towards becoming a standard tool. In conclusion, the study proposes the development of a text analysis tool in the R language, using quantitative criteria and enabling holistic evaluation, for texts to be used in the clinical assessment of reading skills. Thus, the most suitable texts for assessing the reading skills of students in grades 2 and 3 and for advanced diagnosis and intervention tools were sought. On the other hand, studying has some limitations. The first is that no automatic process was used to determine the number of morphemes; instead, the number of phonemes was determined using the system (v.01.16, Mutlu et al., 2021). The second is that the coherence and consistency of the texts were not examined. The last limitation is that the text analysis tool has not yet been applied to students.

Thanks to this proposed R-based linguistic text analyses tool, researchers studying reading skills and specialists such as special education teachers, classroom teachers, guidance and psychological counseling teachers, and speech therapists might assess the reading skills of students in grades 2 and 3. In this context, it is recommended that the tool and texts be used in schools, counseling centers, and Guidance Research Centers. On the other hand, the proposed tool can be effectively used by researchers working in various fields such as special education and speech therapy in the processes of assessing reading skills, making diagnoses, or implementing intervention programs. The tool can also be used in studies conducted in fields such as linguistics and Turkish language education in the context of parameters such as word frequency and letter-bigram frequency. Furthermore, researchers using advanced technologies such as eye-tracking can conduct comprehensive studies with the proposed tool.

Acknowledgments

This study was funded by Scientific and Technological Research Council of Türkiye (TUBITAK). Project no 122K308. TUBITAK1001 Project Title: Neurolinguistic Approach to the Diagnosis of School Age Dyslexia: Investigation and Comparison of Phoneme Discrimination, Word and Text Reading Processes with Psychometric, Electrophysiological and Eye Tracking Techniques.

Ethics Committee Approval

This research was conducted with the permission obtained by the İzmir Bakırçay University Scientific Research and Publication Ethics Social and Human Sciences Board's decision dated 28/09/2022 and numbered 717.

Conflict of Interest

The authors declare that they have no conflict of interest.

Author Contribution

Esmehan Özer: *Writing texts, writing the introduction and discussion sections of the study, and reviewing the manuscript for reporting purposes.* ***Sema Acar-Ünalgan:*** *Obtaining research permission, writing texts, developing the introduction, methods, and discussion sections.* ***Hazal Artuvan-Korkmaz:*** *Writing texts and reviewing the manuscript for reporting purposes.* ***Rahime Duygu Temeltürk:*** *Writing texts and reviewing the manuscript for reporting purposes.* ***Özgür Aydın:*** *Writing the methods section, analyzing the data, and writing the findings.*

Orcid

Esmehan Özer https://orcid.org/0000-0001-5919-8072

Sema Acar-Ünalgan https://orcid.org/0000-0001-9129-2229

Hazal Artuvan-Korkmaz https://orcid.org/0000-0001-7739-8676

Rahime Duygu Temeltürk https://orcid.org/0000-0002-9303-5944

Özgür Aydın https://orcid.org/0000-0003-2925-4146

REFERENCES

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review, 4*(4), 439-461. https://doi.org/10.3758/BF03214334

Arnold, T., Ballier, N., Lissón, P., & Tilton, L. (2019). Beyond lexical frequencies: Using R for text analysis in the digital humanities. *Language Resources and Evaluation, 53*(4), 707-733. https://doi.org/10.1007/s10579-019-09456-6

Ateşman, E. (1997). Türkçede okunabilirliğin ölçülmesi. *Ankara Üniversitesi Türkçe ve Yabancı Dil Uygulama ve Araştırma Merkezi Dil Dergisi, 58*, 171-174.

Aziz, A., Fook, C. Y., & Alsree, Z. (2010). Computational text analysis: A more comprehensive approach to determine readability of reading materials. *Advances in Language and Literary Studies, 1*(2), 200-219. https://doi.org/10.7575/aiac.alls.v.1n.2p.200

Başaran, M., & Ayol, H. (2009). Okuduğunu anlama ve metne karşı geliştirilen tutum üzerinde metnin bilgi verici veya hikâye edici olmasının etkisi. *Uşak Üniversitesi Sosyal Bilimler Dergisi, 2*(1), 11-23. https://doi.org/10.12780/UUSBD41

Baştuğ, M. (2012). İlköğretim I. kademe öğrencilerinin akıcı okuma becerilerinin çeşitli değişkenler açısından incelenmesi [Investigation of primary school first stage students' reading fluency skills in terms of certain variables] (Tez Numarası: 311007) [Doktora tezi, Gazi Üniversitesi]. Yükseköğretim Kurulu Ulusal Tez Merkezi.

Bezirci, B., & Yılmaz, A. E. (2010). Metinlerin okunabilirliğinin ölçülmesi üzerine bir yazilim kütüphanesi ve Türkçe için yeni bir okunabilirlik ölçütü. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 12*(3), 49-62.

Borleffs, E., Maassen, B. A. M., Lyytinen, H., & Zwarts, F. (2019). Cracking the code: The impact of orthographic transparency and morphological-syllabic complexity on reading and developmental dyslexia. *Frontiers in Psychology, 9*, 1-19. https://doi.org/10.3389/fpsyg.2018.02534

Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing, 12*(3), 169-190. https://doi.org/10.1023/A:1008131926604

Çetinkaya, G. (2010). Türkçe metinlerin okunabilirlik düzeylerinin tanımlanması ve sınıflandırılması [Identifying and classifying the readability levels of the Turkish texts] (Tez Numarası: 265580) [Doktora tezi, Ankara Üniversitesi]. Yükseköğretim Kurulu Ulusal Tez Merkezi.

Çetinkaya, G., & Uzun, L. (2010). Türkçe ders kitaplarındaki metinlerin okunabilirlik özellikleri. H. Ülper (Ed.), *Türkçe ders kitabı çözümlemeleri* (4, 141-153) içinde. Pegem Akademi.

Duran, E., & Kargın, T. (2022). Okuma yazma eğitiminde güncel konular: Ne çalışmalı, nasıl çalışmalı?. *Milli Eğitim Dergisi, 51*(235), 1859-1876. https://doi.org/10.37669/milliegitim.886879

Erten, B., Bozşahin, C., & Zeyrek, D. (2014). Turkish resources for visual word recognition. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 2106-2110). European Language Resources Association (ELRA).

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221-233. https://doi.org/10.1037/h0057532

Güneş, F. (2013). Türkçede metin öğretimi yerine metinle öğrenme. *Adıyaman Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *11*, 603-637. https://doi.org/10.14520/adyusbd.454

Hamzadayı, E., & Batmaz, Ö. (2022). Okuduğunu anlamayı etkileyen etmenlere yönelik bir inceleme. *Ege Eğitim Dergisi, 23*(2), 190-209. https://doi.org/10.12984/egeefd.1105439

Hebert, M., Bohaty, J. J., Nelson, J. R., & Brown, J. (2016). The effects of text structure instruction on expository reading comprehension: A meta-analysis. *Journal of Educational Psychology, 108*(5), 609-629. https://doi.org/10.1037/edu0000082

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2*(2), 127-160. https://doi.org/10.1007/BF00401799

Kassambara, A. (2023). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.2. https://rpkgs.datanovia.com/rstatix/.

Keuleers, E. (2013). _vwr: Useful functions for visual word recognition research_. R package version 0.3.0. https://CRAN.R-project.org/package=vwr

Kirkgöz, Y., & Ünaldi, I. (2012). Coh-metrix: Introduction and validation of an online tool for text analysis. *Çukurova University Faculty of Education Journal, 41*(2), 1-17.

Kucer, S. B. (2014). *Dimensions of literacy: A conceptual base for teaching reading and writing in school settings*. Routledge.

Kumar, A., & Paul, A. (2016). Mastering text mining with R. Packt Publishing Ltd.

Kwartler, T. (2017). Text mining in practice with R. John Wiley & Sons.

Levesque, K. C., Breadmore, H. L., & Deacon, S. H. (2021). How morphology impacts reading and spelling: Advancing the role of morphology in models of literacy development. *Journal of Research in Reading, 44*(1), 10-26. https://doi.org/10.1111/1467-9817.12313

Liu, Y., Groen, M. A., & Cain, K. (2024). The association between morphological awareness and reading comprehension in children: A systematic review and meta-analysis. *Educational Research Review, 42*, 100571. https://doi.org/10.1016/j.edurev.2023.100571

Milli Eğitim Bakanlığı (2019). *Türkçe Dersi Öğretim Programı (ilkokul ve ortaokul 1-8. sınıflar).*

Milli Eğitim Bakanlığı (2021). *Ders kitaplarında okunabilirlik.* Ders Kitapları ve Öğretim Materyalleri Daire Başkanlığı.

Milli Eğitim Bakanlığı (2023). *Taslak ders kitabı ve eğitim araçları ile bunlara ait elektronik içeriklerin incelenmesinde değerlendirmeye esas olacak kriterler ve açıklamaları.* Talim ve Terbiye Kurulu Başkanlığı.

Mouselimis, L. (2021). textTinyR: Text Processing for Small or Big Data Files}, R package version 1.1.8. https://CRAN.R-project.org/package=textTinyR.

Muncer, S. J., Knight, D., & Adams, J. W. (2014). Bigram frequency, number of syllables and morphemes and their effects on lexical decision and word naming. *Journal of Psycholinguist Research, 43*, 241-254. https://doi.org/10.1007/s10936-013-9252-8

Mutlu, M. U., Yetimaslan, N., & Atagün, İ. (2021). Dizge: The grammar analyzer for Turkish. https://github.com/dizge

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review, 13*(1), 45-52. https://doi.org/10.3758/BF03193811

Perfetti, C. A. (1999). Comprehending written language: A blueprint of the reader. In C. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 167-208). Oxford University Press.

Perfetti, C. (2012). Thru but not wisht: Language, writing, and universal reading theory. *Behavioral and Brain Sciences, 35*(5), 299-300. https://doi.org/10.1017/S0140525X12000234

Pilav, S., & Oğuz, M. (2013). Türkçe ders kitaplarında yer alan metin türleri üzerine bir araştırma. *Kırıkkale Üniversitesi Sosyal Bilimler Dergisi, 3*(2), 16-30.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372-422. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology, 62*(8), 1457-1506. https://doi.org/10.1080/17470210902816461

Rayner, K., & Reichle, E. D. (2010). Models of the reading process. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(6), 787-799. https://doi.org/10.1002/wcs.68

Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance, 37*(2), 514-528. https://doi.org/10.1037/a0020990

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Santangelo, D. M. (2023). *The influence of bigram frequency on naming speed: A meta-analysis.* [Master's thesis, The Florida State University].

Schmalz, X., & Mulatti, C. (2017). Busting a myth with the Bayes factor: Effects of letter bigram frequency in visual lexical decision do not reflect reading processes. *The Mental Lexicon, 12*(2), 263-282. https://doi.org/10.1075/ml.17009.sch

Silge, J., & David, R. (2016). tidytext: Text mining and analysis using tidy data principles in r. JOSS, 1 (3). https://doi.org/10.21105/joss.00037

van Heuven, W. (2024). _strngrams: ngram functions_. R package version 0.3.6. https://github.com/waltervanheuven/strngrams

Verhoeven, L., & Perfetti, C. A. (2011). Morphological processing in reading acquisition: A cross-linguistic perspective. *Applied Psycholinguistics, 32*(3), 457-466. https://doi.org/10.1017/S014271641 1000154

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). dplyr: A Grammar of Data Manipulation. R package version 1.1.4. https://dplyr.tidyverse.org.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software 59*(1), 1-23. https://doi.org/10.18637/jss.v059.i10.

Yap, M. J., & Balota, D. A. (2015). Visual word recognition. In A. Pollatsek & R. Treiman (Eds.), *The Oxford handbook of reading* (pp. 26-43). Oxford University Press.

Yıldız, M. (2013). Okuma motivasyonu, akıcı okuma ve okuduğunu anlamanın beşinci sınıf öğrencilerinin akademik başarılarındaki rolü. *International Periodical For The Languages, Literature and History of Turkish or Turkic, 8*(4), 1461-1478.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*(5), 971-979. https://doi.org/10.3758/PBR.15.5.971

Zevin, J. (2009). Word recognition. In L. R. Squire (Ed.), *Encyclopedia of neuroscience* (pp. 517-522). Academic Press.