Turkish Text Readability Degree Classification from EEG Signals

Chapter · November 2017

CITATIONS
0

READS
40

3 authors:

Server Göksel Eraldemir
iskenderun technical university
11 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Mustafa Turan Arslan
Mustafa Kemal University
18 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Esen Yildirim
Adana Science and Technology University
46 PUBLICATIONS **313** CITATIONS
SEE PROFILE

TÜRKÇE METİNLERİN OKUNABİLİRLİK DERECESİNİN EEG İLE SINIFLANDIRILMASI

TURKISH TEXT READABILITY DEGREE CLASSIFICATION FROM EEG SIGNALS

Server Göksel Eraldemir*
İskenderun Teknik Üniversitesi
sgoksel.eraldemir@iste.edu.tr

Mustafa Turan Arslan
Mustafa Kemal Üniversitesi
mtarslan@mku.edu.tr

Esen Yıldırım
Adana Bilim ve Teknoloji Üniversitesi
esenyildirim@gmail.com

ÖZET: Eğitim içerikli materyallerin geliştirilmesinde en önemli hususlardan bir tanesi okunabilirlik derecesidir. Okunabilirlik çalışmaları İngilizce için 1920'li yıllarda başlamıştır. İngilizcenin dil yapısı Türkçeden farklı olduğu için bu çalışmalar ile elde edilen formüllerin dilimize uygun olmadığı birçok çalışma ile gösterilmiştir. Türkçe için okunabilirlik çalışması 1997 yılında Akman tarafından geliştirilen bir yöntem ile başlamıştır. Bu yöntem, Türkçe metinlerin okunabilirlik düzeylerinin belirlenmesine yönelik birçok çalışmada kullanılmıştır. Bu çalışmada, okunabilirlik düzeyleri kolay ve zor olan metinlerin sessiz okunması esnasında toplanan EEG sinyallerinin ayırt edilmesi hedeflenmiştir. Çalışmada, yükseköğrenim gören 18 sağlıklı denek kullanılmıştır. Kayıt edilen EEG sinyalleri, Wavelet dalgacık yöntemi ile analiz edilerek öznitelikler elde edilmiştir. Elde edilen öznitelikler, k en yakın komşuluk algoritması ile sınıflandırılmıştır. Çalışmada elde edilen sonuçlara göre, Akman okunabilirlik endeksine göre zor ve kolay olarak derecelendirilen metinlerin sessiz okunması esnasında toplanan EEG işaretleri ortalama olarak %86.18 doğruluk oranıyla sınıflandırılmıştır.

Anahtar sözcükler: Türkçe Metin Okunabilirliği, EEG, Dalgacık Dönüşümü, Sınıflandırma

ABSTRACT: One of the most important aspects in the development of educational materials is the degree of readability of the text. Readability studies have started in 1920s for English language. Studies have shown that the formulas developed for English are not valid in Turkish texts because of the structural differences. Readability degree studies for Turkish started with a method developed by Akman in 1997. This method has been used in many studies to determine the readability degrees of Turkish texts. In this study, we aim to detect the readability degrees of the text, as easy or difficult, from EEG signals collected while texts are read silently. EEG signals are collected from 18 healthy higher education students. EEG signals are analyzed with Wavelet transform to extract the features. Extracted features are classified by k-nearest neighbor algorithm. Results show that, EEG signals collected during silent reading of Turkish text having readability degrees of easy and difficult, according to Akman's readability index, are classified with 86.18% accuracy on average.

Keywords: Turkish Text Readability, EEG, Wavelet Transform, Classification

GİRİŞ

Metin okunabilirliği ilk olarak 1920'li yıllarda Amerika Bileşik Devletleri'nde (ABD) ortaya çıkan ve geliştirilen bir kavramdır. Metin okunabilirliği üzerine yapılan çalışmalar ABD'de 1920 yılından sonra hız kazanmış ve bu alanda 50'den fazla formül geliştirilmiştir (Crossley, Greenfield, ve Mcnamara, 2008). Bu formüller özellikle eğitim materyallerinin hazırlanmasında, askeri belgelerin hazırlanmasında

ve sağlık alanındaki belgelerde kullanılmıştır (Bezirci ve Yılmaz, 2010)

İngilizce için geliştirilen ve en çok kullanılan formüllerden bazıları FLESCH (Flesch, 1948), GUNNING FOG (Gunning, 1952), COLEMAN (Coleman, 1965), ARI (Smith ve Senter, 1967), ve Smog (Laughlin, 1969) okunabilirlik formülleridir. En çok kullanılan formüllerden biri olan Flesch okunabilirlik formülü denklem 1'de verilmiştir.

$$Flesch\ Okunabilirlik\ Formülü = 206.835 - 1.015x(KCO) - 84.6xHKO \quad (1)$$

Bu formülde belirtilen;

KCO= Kelime sayısı/Cümle Sayısı

HKO=Hece Sayısı/Kelime Sayısı

Türkçe için geliştirilen ilk ve en önemli formül olan Ateşman formülü en çok kullanılan okunabilirlik formülüdür (Ateşman, 1997). Bu formül İngilizce için geliştirilmiş olan Flesch okunabilirlik formülünün Türkçe metinlere uygun hale getirilmesi ile geliştirilmiştir. Denklem 2'de verilen Ateşman formülü Türkçe içerikli eğitim metinlerinin okunabilirlik açısından derecelendirilmesi için birçok çalışmada kullanılmıştır (Çeçen ve Aydemir, 2004; Okur ve Ari, 2013; Temur, 2002; Zorbaz, 2013). Bu formül Flesch okunabilirlik formülündeki sabit değişkenlerin Türkçe metinler için uygun hale getirilmesi ile geliştirilmiştir.

$$Ateşman\ Okunabilirlik\ Formülü = 8.825 - 40.175x(KHC) - 2.610x(KCO) \quad (2)$$

Ateşman formülü ile bulunan sonuçlar aşağıdaki Tablo 1'de gösterilen ölçeğe göre değerlendirilmektedir.

Tablo-1 Metin Okunabilirlik Ölçütleri

Düzey	Okunabilirlik Aralığı
Çok Kolay	90-100
Kolay	70-89
Orta güçlükte	50-69
Zor	30-49
Çok zor	1-29

Bu formüllerin hepsinde temel olarak kelime uzunluğu, cümle uzunluğu, hece sayısı ve bunların birbirlerine oranları kullanılmıştır (Bezirci ve Yılmaz, 2010; Çetinkaya, 2010). Bulunan sonuçlara göre metinlerin okunabilirlik dereceleri belirlenmektedir.

Elektroensefalografi (EEG), beynin her tülü aktivitesinde oluşan elektriksel sinyallerin kayıt edilmesine verilen isimdir. Kayıt edilen bu sinyaller sinyal işleme yöntemleri ile temizlenebilmekte ve özniteliklerine ayrılarak analizleri yapılabilmektedir. EEG sinyalleri 0.5-80 Hz arasında değişkenlik göstermektedir.

EEG sinyalleri kullanılarak sayısal ve sözel işlemler üzerine birçok çalışma yapılmıştır. Bu çalışmaların bazılarında sadece matematiksel işlemlerin ve sayısal ifadelerin tespiti üzerine olmuştur (Nath et al., 2015; Rao, Gawali, Mehrotra, Rokade, ve Deore, 2012). Bazıları ise sadece metinsel veya ifadeler işlemler üzerine olmuştur (Wahed, Rana, Hasan, ve Ahmad, 2016).

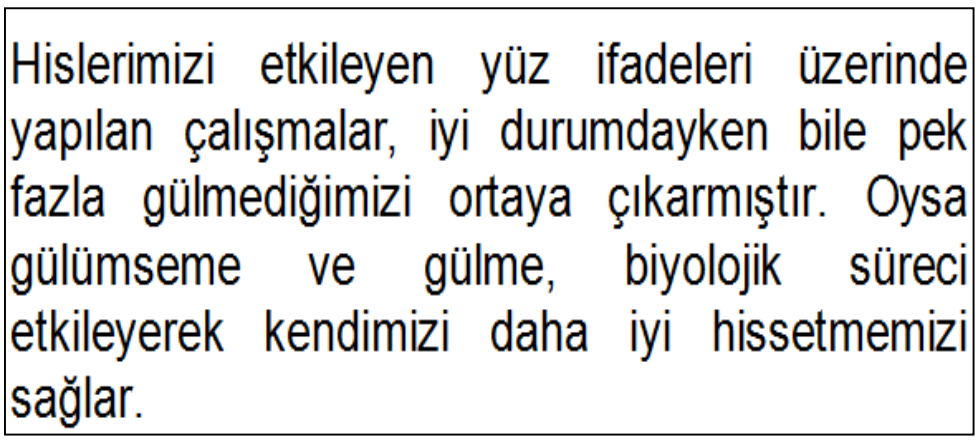
Rao ve arkadaşları 2012 yılında yaptıkları rakam tanıma çalışmasında yaşları 20-25 arasında değişen ve sağ elini kullanan erkek deneklere izlettirdikleri çalışmada EEG sinyallerinden rakam tanımaya beynin bir iş için çalıştığı sırada ürettiği sinyaller olan BETA dalga boyuna odaklanarak araştırmışlardır. Çalışma sonucunda rakamları %68.33 oranında tanımayı başarmışlardır (Rao, Gawali, Mehrotra, Rokade, ve Deore, 2012). Başka bir çalışmada, 2015 yılında Nath ve arkadaşları yaptıkları çalışmada deneklere slayt halinde çeşitli rakamlar gösterirken topladıkları EEG sinyalleri ile rakamların EEG sinyalleri ile dalga boylarının ilişkisini incelemiş ve rakamların tanınmasında beta dalgalarının diğer dalga boylarına göre daha iyi sonuçlar verdiğini göstermişlerdir (Nath et al., 2015). Wahed ve arkadaşları 2016 yılında yaptıkları çalışmada harflerin EEG sinyallerinden tanınması için yaptıkları çalışmada delta türü EEG sinyallerinin diğer sinyallere göre daha iyi başarı gösterdiğini göstermiştir. Çalışmada kayıt edilen EEG sinyalleri; bior2.4, symlet4, coiftlet6 ve db8 dalgacıkları ayrı ayrı özniteliklerine ayrıştırılmış ve bu öznitelikler birleştirilerek hepsine birlikte öznitelik seçimi uygulanmıştır. Sonuçta delta dalga boyunun diğer dalga boylarına göre daha iyi sonuç verdiği tespit etmiştir (Wahed, Rana, Hasan, ve Ahmad, 2016). Eraldemir ve Ark. 2015 yılında yaptıkları çalışma ile 18 sağlıklı denekten sayısal işlem ve sözel metinlerden oluşan toplam 60 slayttan oluşan sunumu izlerken kayıt ettikleri EEG sinyallerinden metinsel okuma ile matematiksel işlemleri birbirlerinden ayırt etmeye çalışmışlardır. Çalışma sonucunda bior2.4 dalgacığı ile özniteliklerine ayrılan veri J48 algoritması ile sınıflandırılarak %90.6 oranında birbirlerinden ayırt edilebilmiştir (Eraldemir ve Yildirim, 2015). Eraldemir ve arkadaşları 2014 yılında yaptıkları bir çalışmada matematiksel işlemlerden oluşan slaytları izlerken kayıt edilen EEG verilerinden çarpma-bölme ve toplama-çıkarma işlemelerini iki gurup halinde sınıflandırmaya çalışmışlardır. Bu çalışma sonucunda her iki gurup birbirinden k-NN algoritması ile %79.3 oranında ayırt edilebilmiştir (Eraldemir, Yildirim, ve Kutlu, 2014).

Bu çalışmada, üniversite öğrencilerinden Ateşman formülü kullanılarak derecelendirilmiş metinleri sessiz bir şekilde okunmaları istenmiş ve bu esnada EEG sinyalleri toplanmıştır. Toplanan EEG sinyalleri ayrık dalgacık yöntemleri ile analiz edilerek öznitelik vektörleri oluşturulmuş ve sınıflandırma algoritması ile analiz edilmiştir. Bu çalışmanın geri kalan organizasyon şekli şu şekildedir: 2.bölümde kullanılan veri seti ve yöntemler hakkında bilgi verilmiş, 3. Bölümde ise yapılan deneysel analizlerden sonuçlar elde edilmiştir. Son bölümde ise yapılan çalışma özetlenmiştir.

YÖNTEM

Verilerin Toplanması

Bu çalışmada yükseköğrenim gören, gönüllülük esasına göre 18 sağlıklı denekten metin okunması sırasında çekilen EEG verileri kullanılmıştır. Kayıtlar esnasında, deneklere toplam 30 adet slayt gösterilerek sessiz okuma yapmaları istenmiştir. Bu slaytlardaki metinler Ateşman formülüne göre derecelendirildiğinde 4 adet zor, 3 adet kolay, 1 adet çok kolay ve 22 adet orta güçlükte metnin olduğu görülmüştür. Çalışmada zor ve kolay/çok kolay metinlerin okunması esnasında toplanan EEG verileri sınıflandırılmıştır. Kolay metin örneği Şekil-1'de, zor metin örneği ise Şekil-2'de gösterilmiştir.



Şekil 1. Kolay Olarak Derecelendirilmiş Slayt Örneği

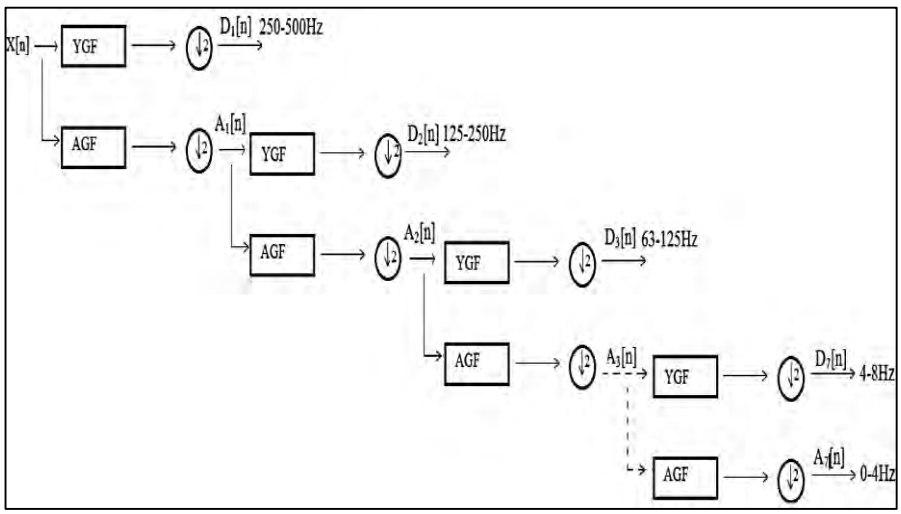
Çağımız bilgi ve teknoloji çağıdır. Uluslararası rekabette bilgi ve teknolojiden en iyi şekilde yararlananların ilk sıralarda yarıştığı bilinmektedir. Bu nedenle, gelişmiş teknolojinin en son yeniliklerini siz değerli öğretmenlerimizin hizmetine sunmaktan, iş ve işlemlerin yürütülmesindeki zaman ve kaynak israfının önlenmesine yönelik tedbirleri sizinle paylaşmaktan mutluluk duyuyoruz.

Şekil 2. Zor Olarak Derecelendirilmiş Slayt Örneği

Çalışmada kullanılan slaytların uzunluğu 13.25 saniye, EEG için örnekleme frekansı 1 kHz'dir.

Öznitelik Çıkarma

Elde edilen sinyaller, ayrık dalgacık yöntemi ile özniteliklerine ayrılmıştır. Toplanan EEG sinyalleri Ayrık Dalgacık Dönüşümü (ADD) sayesinde detay ve yaklaşım katsayılarına ayrılmaktadır. Çalışmadaki verilere, Şekil 3'te görüldüğü gibi D4, D5, D6, D7 ve A7 katsayıları elde edilecek şekilde, ADD uygulanmıştır. Öznitelik olarak, katsayıların mutlak değerlerinin ortalaması, mutlak değerlerinin maksimumu, mutlak değerlerinin minimumu ve standart sapmasından oluşan istatistiksel özellikler kullanılmıştır. Kanal sayısı 26 olduğundan toplamda her slayt için (26 Kanal * (4 Detay Katsayısı + 1 Yaklaşım Katsayısı)* 4 İstatistiksel Veri) 520 adet öznitelik elde edilerek öznitelik matrisi oluşturulmuştur.



Şekil 3. Dalgacık Analizi

k-En Yakın Komşu Algoritması

k-En yakın komşu yöntemi (k-NN), veri madenciliğinin en temel, basit, etkili ve güçlü örüntü sınıflandırma yöntemlerinden birisidir. Bu yöntem, makine öğrenme algoritmaları arasında popüler olarak kullanılan danışmanlı öğrenme algoritmalarından birisidir.

Eğitim işlemi çok kısa sürmesine rağmen k-NN algoritması yüksek bir hesaplama maliyetine sahiptir (Shmueli, Patel, ve Bruce, 2010). Ayrıca, bu algoritma, yüksek bellek gereksinimlerine sahiptir, komşu sayısı ve uzaklık ölçütü gibi parametrelere duyarlıdır (Duda, Peter, ve David G. Stork, 2012).

k-NN sınıflandırıcısı örnek tabanlı bir yöntem olduğundan sınıf etiketi bilenen eğitim örnekleriyle kıyaslama yapar. Bu algoritmanın çalışma mantığına göre, eğitim esnasında bütün örnekler algoritmanın belleğinde tutulur. Test örneğinin sınıfı belirlenirken bu örneğe en yakın k adet komşu eğitim örneği seçilir. Daha sonra k adet komşu eğitim örneğinin sınıflarına bakılarak test örneği baskın olan sınıfa dâhil edilir.

k-NN yönteminde eğitim ve test örnekleri arasındaki mesafeyi ölçmek için çeşitli yaklaşımlar kullanılmaktadır. Bunlar Öklid (Euclidean), Minkowski, Manhattan, Dilca, Chebyschev gibi mesafe ölçütleridir. Öklid (Euclidean) yöntemi, sınıflandırma ve kümeleme algoritmalarında en sık kullanılan uzaklık ölçütü olduğundan dolayı bu çalışmada da Öklid yöntemi kullanılmıştır.

$$\left(\sqrt{\sum_{i=1}^{n}\left(x_i - y_i\right)^2}\right) \tag{3}$$

Öklid ölçütü, uzayda iki nokta arasındaki $C=\left(x_1,x_2,x_3,...,x_n\right)$ ve $D=\left(y_1,y_2,y_3,...,y_n\right)$ doğrusal uzaklığı ölçmek için denklem x e göre hesaplanır (Kresse ve Danko, 2012).

k değerinin veriye en uygun şekilde seçilmesi önemlidir. k değeri azaldıkça veriye en uygun örneklerin seçimi kolaylaşırken k değerinin artışı veri ile olan benzeşme oranını düşürür (Mitchell, 1997). k değeri bulunurken sınıflandırma işlemi farklı k değerleri ile yapılarak sonuçlar karşılaştırılır ve en iyi sonucu veren k değeri sınıflandırma için kullanılır. Eraldemir ve arkadaşlarının yaptığı çalışmada matematiksel dört işlemin sınıflandırmasında k-değerlerini sırası ile 1,3,5 ve 10 olarak alarak k değerlerinden en iyi sonucu veren k değerini 1 olarak bulmuşlardır (Eraldemir, Yildirim, ve Kutlu, 2014). Bu çalışmada da k değeri 1 olarak alınmıştır.

BULGULAR

Çalışmada kullanılan metin içerikli slaytların Ateşman Formülü uygulanmış sonuçları Tablo 2' de verilmiştir.

Tablo 2. Kullanılan Metinleri Ateşman Formülü Sonuçları

Slayt Numarası	Toplam Kelime Sayısı	Toplam Hece Sayısı	Cümle Sayısı	Ateşman Sonuçları
Kolay1	37	97	5	74.19
Kolay2	36	72	3	87.155
Kolay3	32	73	6	93.26
Kolay4	31	74	3	75.95339
Zor1	29	102	4	38.60
Zor2	34	98	2	38.66
Zor3	46	143	3	33.91315
Zor4	28	86	2	38.89036

Tablo 1'de görüldüğü gibi sonuçları 70-89 arası olan kolay slaytlardan üç adet ve sonucu 90-100 arasın

olan çok kolay slaytlardan bir adet bulunmaktadır. Bu iki gurup kolay gurup olarak isimlendirilmiştir. Ayrıca sonuçları 30-49 arası olan zor slayttan 4 adet bulunmakta ve bunlarda kendi arasında zor metin gurubu olarak isimlendirilmiştir.

Çalışmada kullanılan bu sekiz slayttan BayesNet algoritması ile sınıflandırma yapıldığında elde edilen sonuçlar Tablo 3'de gösterilmiştir.

Tablo 3. Kullanılan Metinlerin k-NN Sınıflandırma Sonuçları

Denek No	Doğruluk(%)	Kesinlik (%)	F-Ölçeği
Denek1	89.00	89.50	89.00
Denek2	82.50	82.60	82.50
Denek3	77.50	77.70	77.50
Denek4	80.00	80.00	80.00
Denek5	87.50	87.50	87.50
Denek6	88.00	88.00	88.00
Denek7	88.50	88.60	88.50
Denek8	88.50	88.50	88.50
Denek9	90.00	90.00	90.00
Denek10	88.80	88.80	88.70
Denek11	90.00	90.00	90.00
Denek12	91.80	91.80	91.70
Denek13	88.80	88.80	88.70
Denek14	81.80	81.80	81.70
Denek15	85.50	85.50	85.50
Denek16	91.30	91.30	91.20
Denek17	74.50	74.50	74.50
Denek18	87.30	87.30	87.20
Ortalama	86.18	86.23	86.15

Tablo 3'de görüldüğü gibi çalışma sonucunda kolay olarak gruplandırılan metinler ile zor olarak sınıflandırılmış metinler %86.18 doğruluk oranıyla sınıflandırılmıştır. En yüksek sınıflandırma sonucunun denek20'de %91.80 doğruluk oranıyla gerçekleştiği görülmüştür. Ayrıca en düşük sınıflandırma sonucunun %74.50 doğruluk oranıyla denek 17' e ait olduğu görülmüştür.

Toplam 4 deneğin %90 ve üstü başarı ile sınıflandırıldığı görülmüştür. Çalışmaya katılan 12 deneğin ise %80 - %90 arasında doğru pozitif oranı ile kolay ve zor metinler arası sınıflandırma derecesi elde edilmiştir.

SONUÇ

Bu çalışmada, Ateşman formülüne göre zor ve kolay/çok kolay olarak derecelendirilmiş metinlerin EEG sinyalleri kullanılarak k-NN algoritması ile analizi gerçekleştirilmiştir. Elde edilen sonuçlara göre, Türkçe metinlerin formüller uygulanmadan EEG sinyalleri kullanılarak derecelendirilebileceği görülmüştür.

KAYNAKLAR

Ateşman, E. (1997). Türkçede Okunabilirliğin Ölçülmesi. *Dil Dergisi*, *58*, 71–74.

Bezirci, B., & Yılmaz, E. A. (2010). Metinlerin Okunabilirliğinin Ölçülmesi Üzerine Bir Yazılım Kütüphanesi ve Türkçe için Yeni Bir Okunabilirlik Ölçütü. *DEÜ Mühendislik Fakültesi Fen Bilimleri Dergisi*, *12*(3), 49–62.

Coleman, E. B. (1965). On Understanding Prose: Some Determiners of its Complexity. *NSF Final Report GB-2604*.

Crossley, S. A., Greenfield, J., & Mcnamara, D. S. (2008). Assessing Text Readability Using Cognitively Based

Indices. *TESOL Quarterly*, *42*(3), 475–493.

Çeçen, M. A., & Aydemir, F. (2004). Okul Öncesi Hikâye Kitaplarının Okunabilirlik Açısından İncelenmesi. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *8*(16), 185–194.

Çetinkaya, G. (2010). *Türkçe Metinlerin Okunabilirlik Düzeylerinin Tanımlanması Ve Sınıflandırılması*. Ankara Üniversitesi.

Duda, R. O., Peter, E. H., & David G. Stork. (2012). *Pattern Classification*. John Wiley & Sons.

Eraldemir, S. G., & Yildirim, E. (2015). Comparison of wavelets for classification of cognitive EEG signals. In *2015 23nd Signal Processing and Communications Applications Conference (SIU)* (pp. 1381–1384). IEEE.

Eraldemir, S. G., Yildirim, E., & Kutlu, Y. (2014). Classification of Mathematical Tasks from EEG Signals Using k-NN Algorithm. In *Electrical - Electronics - Computer and Biomedical Engineering Symposium (Eleco 2014)* (pp. 551–554). Bursa.

Flesch, R. (1948). A New Readability Yardstick. *Journal of Applied Psychology*, *32*(3), 221–233.

Gunning, R. (1952). *The Technique of Clear Writing*. New York: McGraw-Hill International Book Co.

Kresse, W., & Danko, D. M. (2012). *Springer Handbook of Geographic Information*. Springer Science & Business Media.

Laughlin, G. H. M. (1969). SMOG Grading-a New Readability Formula. *Journal of Reading*, *12*(8), 639–646.

Mitchell, T. M. (1997). Machine learning. *Burr Ridge*, *45*(37), 870–877.

Nath, D., Uddin, M. B., Rana, M. M., Biswas, P. C., Wahed, S., & Ahmad, M. (2015). Number recognition using salient features of beta rhythmic EEG signal. In *2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)* (pp. 1–6). Dhaka: IEEE.

Okur, A., & Ari, G. (2013). Readability of Texts Turkish Textbooks in Grades 6, 7, 8. *Elementary Education Online İlköğretim Online*, *12*(121), 202–226.

Rao, S., Gawali, B., Mehrotra, S. ., Rokade, P., & Deore, R. (2012). Number Recognition System Using Electroencephalogram (Eeg) Signals. *Advances in Computational Research*, *4*(1), 975–3273.

Shmueli, G., Patel, N. R., & Bruce, P. C. (2010). *Data Mining for Business Intelligence*. New Jersey: John Wiley & Sons.

Smith, E. A., & Senter, R. J. (1967). Automated Readability Index. *Aerospace Medical Research Laboratories*, 1–14.

Temur, T. (2002). *İlköğretim 5.Sınıf Türkçe Ders Kitaplarında Bulunan Metinler ile Öğrenci Kompozisyonlarının Okunabilirlik Düzeyleri Açısından Karşılaştırılması*. Gazi Üniversitesi.

Wahed, S., Rana, M., Hasan, S. M. K., & Ahmad, M. (2016). Toward letter recognition system: Determination of best wavelet and best rhythm using EEG. In *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)* (pp. 1–4). IEEE.

Zorbaz, K. Z. (2013). Türkçe Ders Kitaplarındaki Masalların Kelime – Cümle Uzunlukları ve Okunabilirlik Düzeyleri Üzerine Bir Değerlendirme. *Journal of Theory and Practice in Education*, *3*(1), 87–101.