TÜRKİYE SOSYAL ARAŞTIRMALAR DERGİSİ

Türkiye Sosyal Araştırmalar Dergisi * 2026 * Cilt: 30 * Sayı: 1 * (287-296)

Görece Küçük Ölçekli Bir Derlem Üzerinden Kelime Sıklık ve Dağılım Analizi

Word Frequency and Distribution Analysis on a Relatively Small-Scale Corpus

Rifat Ramazan BERK *

Makale Bilgisi

Geliş Tarihi:14.10.2025

Kabul Tarihi:13.02.2026

Doi:10.20296/tsadergisi.1803436

Anahtar Sözcükler:
Kelime sıklığı
Kelime dağılımı
Dinleme metinleri
Yabancılara Türkçe öğretimi

ÖZET

Bu çalışma, yabancı dil olarak Türkçe öğretim materyallerindeki sözcük seçiminde yalnızca mutlak sıklığa dayalı geleneksel yaklaşıma alternatif olarak dağılım (range) ve ortalama indirgenmiş frekans (average reduced frequency - ARF) gibi istatistiksel ölçütlerin birlikte kullanılmasının önemini araştırmaktadır. Çalışma, Yunus Emre Enstitüsü'nün Yedi İklim Türkçe Öğretim Seti (YİTÖS) B1 ve B2 seviyelerindeki dinleme metinlerini derlem olarak kullanmıştır. Metinlerdeki sözcükler, mutlak frekans, dağılım ve ARF değerlerine göre ayrı ayrı sıralanmış ve ilk 30'ar sözcüğün listeleri karşılaştırmalı olarak analiz edilmiştir. Bulgular, "bir", "ve", "ol-", "bu" gibi temel sözcüklerin tüm listelerde üst sıralarda yer aldığını göstermiştir. Ancak, yalnızca mutlak frekansa göre yapılan sıralamada, "gazete", "festival", "Selvihan" (B1) ve "su", "çalışma", "hayat" (B2) gibi belirli birkaç metinde yoğun kullanılan sözcüklerin yapay olarak üst sıralara çıktığı tespit edilmiştir. Buna karşılık, dağılım ölçütü "sonra", "yap-", "gibi" gibi metinler arasında yaygın kullanılan sözcükleri öne çıkarmıştır. ARF ölçütü ise hem mutlak sıklıktaki yapay yükselmeleri dengelemiş hem de dengeli dağılım gösteren sözcükleri koruyarak daha güvenilir bir liste sunmuştur. Öğretim materyali geliştiricilerinin ve araştırmacıların, sözcüklerin önem derecesini belirlerken yalnızca mutlak sıklığa değil, dağılım ve ARF gibi bileşik ölçütlere de başvurmaları daha dengeli söz varlığı seçimine imkân sağlayacaktır.

Article Information

Submission: 14.10.2025

Acceptance: 13.02.2026

Doi:10.20296/tsadergisi.1803436

Key Words:
Word frequency,
Word distribution,
Listening texts,
Teaching Turkish as a foreign Language,

ABSTRACT

Selecting appropriate vocabulary is a critical challenge in teaching Turkish as a foreign language, particularly when traditional methods rely primarily on absolute frequency counts. This study investigates the effectiveness of combining statistical criteria, distribution (range) and average reduced frequency (ARF), as an alternative approach to vocabulary selection. The research is based on a corpus of listening texts from B1 and B2 levels of the *Yedi İklim Türkçe* series published by the Yunus Emre Institute. Words in the corpus were ranked separately according to absolute frequency, distribution, and ARF values. The top 30 words from each list were then compared to evaluate differences across methods. The findings indicate that core functional words such as "bir," "ve," "ol-," and "bu" consistently rank highly across all criteria. However, relying solely on absolute frequency leads to distortions, as some words (e.g., "gazete," "festival," and proper nouns at B1; "su," "çalışma," and "hayat" at B2) appear artificially prominent due to their concentration in specific texts. In contrast, the distribution criterion highlights words used more evenly across the corpus, while ARF provides a balanced measure by combining frequency and dispersion. The study concludes that integrating distribution and ARF with absolute frequency offers a more reliable basis for vocabulary selection. It is recommended that material developers and researchers adopt multi-criteria approaches to better reflect authentic language use.

Atıf İçin

Berk, R. R. (2026). Görece küçük ölçekli bir derlem üzerinden kelime sıklık ve dağılım analizi. Türkiye Sosyal Araştırmalar Dergisi, 30(1), 287-296 doi:10.20296/tsadergisi.1803436

* Dr. Öğr. Üyesi, Bayburt Üniversitesi, Sosyal Bilimler MYO, Büro Hizmetleri ve Sekreterlik Bölümü, Bayburt, rifatberk@bayburt.edu.tr, ORCID: https://orcid.org/0000-0002-7505-7787

GİRİŞ

Kelime edinimi, yabancı dil öğrenim sürecinin merkezinde yer alan ve dört temel dil becerisinin gelişimini doğrudan etkileyen önemli bir unsurdur. Dolayısıyla öğretim materyallerinde sunulan sözvarlığının niteliği ve niceliği büyük önem taşımaktadır. Öğretim materyallerindeki kelime seçimini yönlendiren temel ölçütlerden biri olan kelime sıklığı (frekans), belirli bir uzunluktaki bir konuşma ya da yazıda aynı dilsel olgu ya da birimin gerçekleşme sayısı olarak tanımlanmaktadır (Vardar, 2002: s.174). Daha basit bir ifadeyle mutlak frekans ya da sıklık, bir kelimenin belirli bir metin ya da metin dizisi içinde tam olarak kaç kez geçtiğinin basit bir sayımıdır.

Yabancı dil öğretiminde, özellikle kelime öğretimi söz konusu olduğunda, kelime sıklık listeleri önemli bir araç olarak görülmektedir. Nitekim kelimelerin öğrenirlere hangi sırayla tanıtılması gerektiğinin belirlenmesinde kelime sıklıklarının bilinmesi gerektiği (Aksan, 1998: s.20); öğretim materyallerinde kullanılacak kelimelerin seçiminde kullanım sıklıklarına başvurulması gerektiği (Göçen, 2016: s. 38) alan yazında belirtilmektedir. Bununla birlikte Türkçenin yabancı/ikinci dil öğretimine yönelik sıklık çalışmalarında (Göçen vd., 2020; Göçen & Aydın, 2021; Göçen & Bingöl, 2024; Göçen & Buluş, 2022; Göçen & Okur, 2016; Özdemirel & Dilidüzgün, 2018) kelimelerin sadece mutlak frekansına odaklanıldığı görülmektedir. Fakat kelime öğretiminde ya da öğretim materyallerinin hazırlanmasında kelimelerin sadece mutlak frekansına odaklanmak yeterli olmayabilir. Çünkü frekans değeri yalnızca hesaplamalar için kullanılan metne özgü bir durumdur ve metnin uzunluğuna, konusuna, yazarına ya da yazarlarına, yazarların üslûbuna sıkı sıkıya bağlıdır. Herhangi bir dille üretilmiş olan yazılı ve sözlü bütün ürünleri bir derlemde toplamak mümkün olamayacağına göre aynı durum derlemi oluşturan seçki için de geçerlidir. Bazı kelimeler belirli metinlerde (derleme dâhil edilen herhangi bir kitap ya da kitap bölümünde) yoğun bir şekilde kullanılabilir. Kilgarriff (1997), bu durumu "Deniz Salyangozu Problemi (The Whelk Problem)" şeklinde izah etmektedir. Araştırmacı, derleme dâhil edilen belirli bir konudaki tek bir kitabın, örneğin deniz salyangozlarıyla ilgili bir kitabın deniz yaşamına dair bir terimi yüzlerce kez tekrarlayarak kelimeyi sıklık listesinde yapay bir şekilde üst sıralara taşıyabileceğini ifade etmektedir. Bu sebeple, güvenilir dil verisi elde etmek için bir kelimenin sadece mutlak frekansına değil metinler arasındaki dağılımını ya da hem dağılımı hem frekansı ölçebilen istatistiksel yöntemlere başvurulmalıdır.

Dağılım, bir kelime ya da kelime grubunun bir metin ya da metin dizgesindeki yayılımını niceliksel olarak betimleyen ve onun yalnızca ne sıklıkta ortaya çıktığını değil aynı zamanda metni ya da derlemi oluşturan alt bölümler arasında ne derece dengeli dağıldığını ölçen istatistiksel bir kavramdır. Dağılımın tek ve mutlak bir ölçümü yoktur. Nitekim Gries (2008), derlem dilbiliminde kullanılan çok sayıda dağılım ölçümü ve indirgenmiş frekans yönteminden bahseder ve bunları üç temel kategoride sınıflandırır. İlk kategori, bir dil biriminin derlem içindeki yayılımını doğrudan istatistiksel bir değerle ölçmeyi amaçlayan yöntemleri içerir. İkinci kategori, bir dil biriminin mutlak frekansını, derlemin farklı bölümlerindeki dağılımına göre indirgemeye (düzeltmeye) yönelik yöntemlerdir. Üçüncü kategori ise, bir dil biriminin ardışık kullanımları arasındaki mesafeyi analiz eden yöntemlere odaklanır.

Bu çalışma, yabancı dil olarak Türkçe öğretimi materyallerindeki kelime seçimini değerlendirirken yalnızca mutlak frekansa odaklanmak yerine, dağılım (range) ve hem mutlak frekans hem de dağılımı birlikte ölçen ortalama indirgenmiş frekans (average reduced frequency - ARF) gibi istatistiksel ölçütleri de dikkate almanın önemini ortaya koymayı amaçlamaktadır. Çalışmanın odak noktası, bu farklı ölçütlerin bir ders materyalindeki kelimeleri sıralama biçimlerinin nasıl farklılaştığını karşılaştırmalı olarak incelemektir. Türkçenin yabancı dil olarak öğretimi alanında, öğretim materyallerindeki sözvarlığını değerlendirmek için mutlak frekans, dağılım ve indirgenmiş frekans ölçütlerini birlikte ele alan herhangi bir çalışmaya rastlanmamış olması, bu araştırmaya özgünlük kazandırmaktadır. Dolayısıyla çalışma, alanda yaygın olan yalnızca mutlak frekansa dayalı geleneksel yaklaşıma bir alternatif sunmayı hedeflemektedir. Bu bağlamda elde edilecek bulguların, gelecekteki öğretim setleri ve dil materyallerindeki sözvarlığının seçimi ve kademelendirilmesi için daha sağlam bir istatistiksel temel oluşturması ve alandaki kuramsal tartışmaya katkı sunması

beklenmektedir. Bu amaçla, alanda yaygın bir kullanıma sahip olan Yunus Emre Enstitüsü tarafından hazırlanan Yedi İklim Türkçe Öğretim Seti (YİTÖS) B1-B2 seviye dinleme metinleri analiz birimi olarak seçilmiştir. Çalışma, bu metinler üzerinden aşağıdaki sorulara yanıt aramaktadır:

1. YİTÖS B1 seviye dinleme metinlerinde, mutlak frekansa, dağılıma (range) ve ortalama indirgenmiş frekansa (ARF) göre belirlenen ilk 30'ar sözcük hangileridir ve bu ölçütlere göre sıralamaları nasıldır?

2. YİTÖS B2 seviye dinleme metinlerinde, mutlak frekansa, dağılıma (range) ve ortalama indirgenmiş frekansa (ARF) göre belirlenen ilk 30'ar sözcük hangileridir ve bu ölçütlere göre sıralamaları nasıldır?

YÖNTEM

Bu çalışma, Yunus Emre Enstitüsü tarafından hazırlanan Yedi İklim Türkçe Öğretim Seti B1-B2 seviyesi dinleme metinlerindeki kelimelerin sıklık ve dağılım özelliklerini incelemek amacıyla nicel bir araştırma deseni benimsemiştir.

Araştırmanın Modeli

Araştırma, nitel veri kaynaklarının nicel tekniklerle incelenmesine dayanan betimsel ve karşılaştırmalı bir doküman analizi çalışması olarak desenlenmiştir. Çalışma, mevcut öğretim materyallerinin içerdiği dilsel örüntüleri, önceden belirlenmiş istatistiksel göstergeler aracılığıyla tarif etmeyi ve farklı ölçütlerin (mutlak frekans, dağılım, ortalama indirgenmiş frekans) sonuçlarını karşılaştırmayı hedeflemektedir. Bu nedenle, araştırma süreci nicel paradigma içinde kurgulanmış olup, verilerin toplanması, işlenmesi ve yorumlanmasında sayısallaştırılmış dil verisi ve objektif analiz araçları esas alınmıştır. Bu analitik yaklaşım, öğretim materyallerinin sözvarlığına ilişkin somut, ölçülebilir ve genellenebilir bulgular elde etmeyi mümkün kılmaktadır. Temel analiz birimi olan metinlerin seçkisel bir derlem haline getirilmesi ve bu derlem üzerinde dilbilimsel istatistik yazılımları kullanılarak otomatik sorgulamalar yapılması, çalışmanın metodolojik çerçevesinin ana hatlarını belirlemiştir.

Evren ve Örneklem

Araştırmanın evrenini, yabancı dil olarak Türkçe öğretiminde kullanılan B1 ve B2 dil seviyelerini hedefleyen tüm dinleme materyalleri oluşturmaktadır. Çalışmanın örneklemi ise amaçlı örnekleme yöntemiyle seçilmiş, Yunus Emre Enstitüsü tarafından hazırlanan Yedi İklim Türkçe Öğretim Seti B1-B2 seviye dinleme metinleridir. Bu seçim, söz konusu setin yaygın kullanımı nedeniyle yapılmıştır. Dinleme metinleri, setin dinleme kitapçıklarından alınarak dijital ortama aktarılmış ve her biri ayrı bir metin birimi olarak değerlendirilmiştir. Örneklem kapsamına giren dinleme metinlerinin sayısal dağılımı Tablo 1'de sunulmuştur:

Tablo 1. *Çalışmanın örnekleminde yer alan dinleme metinleri*

Dil Seviyesi	Dinleme Metni Sayısı
B1	27
B2	29
Toplam	56

Belirlenen bu 56 adet dinleme metni, ilgili setin dinleme kitapçıklarından alınarak dijital ortama aktarılmış ve derlem analizi için hazırlanmıştır. Her bir dinleme kaydının yazılı transkripsiyonu, analiz sürecinde ayrı bir metin birimi (text file) olarak değerlendirilmiştir.

Veri Toplama Araçları

Araştırmanın birincil veri kaynağını, Yunus Emre Enstitüsü tarafından hazırlanan Yedi İklim Türkçe Öğretim Seti'nin (YİTÖS) B1 ve B2 seviyelerine ait dinleme kitapçıkları oluşturmaktadır. Bu

kitapçıklardaki tüm dinleme metinleri, bilgisayar ortamına metinsel olarak aktarılarak dijital dokümanlar haline getirilmiştir. Bu işlem sonucunda, B1 seviyesindeki 27 metin ve B2 seviyesindeki 29 metin olmak üzere toplam 56 metin bir araya getirilmiş ve analiz için temel veri setini oluşturmuştur. Daha sonra, karşılaştırmalı analizi mümkün kılmak amacıyla, bu metinler seviyelerine göre iki ayrı gruba ayrılarak iki farklı derlem oluşturulmuştur. Bir başka deyişle, veri toplama aracı olarak bizzat öğretim materyallerinin kendisi kullanılmış ve bu materyaller bilgisayar destekli dil analizine uygun hale getirilmiştir.

Verilerin Analizi

YİTÖS B1-B2 seviye dinleme metinlerinde kullanılan kelimelerin sıklık ve dağılımına yönelik betimsel (nicel) analiz temel alınmıştır. Kelime sıklık ve dağılım istatistikleri yapılmadan önce oluşturulan derlemdeki kelimeler taban hâline getirilmiştir. Kurudayıoğlu ve Karadağ (2005), taban kavramını, çekim unsurları çıkarıldığında anlam ile biçimin kesiştiği ilk nokta olarak tanımlamaktadır. Aynı araştırmacılar, bu kesişme noktasının kelime kökünde, gövdesinde ya da birleşik kelime tabanında gerçekleşebileceğini belirtmektedir. Ayrıca Baş (2011) tarafından söz varlığı çalışmaları için belirtilen ölçütler dikkate alınmıştır. Çalışmada özel isimler, sayılar, tarih ve saatler, kısaltmalar, ikilemeler, deyimler muhafaza edilmiş ve derlemden silinmemiştir. Özel isimler, ayrı yazılan birleşik sözcükler istatistiksel hesaplamalar yapabilen derlem araçlarının tek bir birim olarak algılaması için birleştirilmiştir (örn: İdil Biret –> İdil-Biret). Aynı yöntemle, birleşik fiil oluşturan sözcükler arasına kısa çizgi konulmak suretiyle birleştirme yapılmıştır. Olumsuzluk eki, geçici isim yapan zarf-fiil ve sıfat fiil ekleri silinmiş bunlar fiil olarak ele alınmıştır. İsim-fiil eklerinin bulunduğu sözcükler madde başı olarak bırakılmıştır. Bütün bu işlemlerin ardından B1 seviye derlemi 5257toplam kelime ve 1888 farklı kelimeden oluşmuştur. B2 derlemi ise 6940 toplam kelime ve 2228 farklı kelimeden oluşmuştur. Ardından SketchEngine yazılımı kullanılarak frekans ve range değerleri hesaplanmıştır. Range, derlemde yer alan her bir kelimenin kaç farklı metinde geçtiğini göstermektedir. Bu yönüyle oldukça basit ama etkili bir dağılım göstergesidir. Fakat kelimenin farklı metinlerde kaç kez geçtiğini göstermemesi yönüyle zayıf bir dağılım istatistiği olarak değerlendirilebilir. Bu yüzden ARF (Average Reduced Frequency – ortalama indirgenmiş frekans kullanılmıştır. ARF, bir kelimenin veya ifadenin metinlerde ne sıklıkla kullanıldığını ölçen bir istatistiksel metrik olup özellikle psikolinguistik ve dilbilim araştırmalarında kullanılır. ARF, bir terimin mutlak frekansından ziyade, farklı metinlerdeki dağılımını ve "azaltılmış" (reduced) frekansını dikkate alarak daha güvenilir bir ölçüm sunmaktadır (Savický & Hlaváčová, 2002). ARF, bir kelimenin farklı metinlerdeki ortalama görülme sıklığını hesaplarken aşırı tekrarlanan kullanımların etkisini azaltmaktadır. Mevcut derlemdeki kelimelerin ARF'si ücretli bir yazılım olan SketchEngine kullanılarak hesaplanmıştır. ARF değeri hesaplanırken derlem, kelimenin toplam frekansı (k) kadar eşit uzunlukta (derlem toplam kelime sayısı/k) örtüşmeyen parçalara bölünmektedir. İndirgenmiş frekans (RF), kelimenin bu parçalardan kaçında geçtiğini göstermektedir. ARF, tüm kaydırmalı bölümler için hesaplanan RF değerlerinin ortalaması alınarak bulunmaktadır.

BULGULAR ve YORUMLAR

Birinci Araştırma Sorusuna Yönelik Bulgular

Çalışmanın ilk sorusu, YİTÖS B1 dinleme metinlerinde en sık geçen 30 kelimeyi, üç farklı ölçüte göre belirlemeyi amaçlamaktadır. Tablo 2, bu üç ölçütün ürettiği listeleri karşılaştırmalı olarak göstermektedir.

Tablo 2. *YİTÖS B1 seviye dinleme metinlerinde mutlak frekans, dağılım (Range) ve ortalama indirgenmiş frekansa (ARF) göre ilk 30 sözcüğün karşılaştırmalı sıralaması*

Kelime	**Mutlak Frekans**		**Kelime**	**Dağılım (Range)**		**Kelime**	**Ort. İnd. Frekans (ARF)**	
	Değer	**Sıra**		**Değer**	**Sıra**		**Değer**	**Sıra**
bir	152	1	bir	25	1	bir	93	1
ve	121	2	bu	24	2	ve	78	2

bu	84	3	ve	24	2	bu	53	3
ol-	72	4	ol-	21	4	ol-	41	4
için	56	5	için	21	4	için	33	5
çok	45	6	çok	20	6	çok	28	6
de	41	7	de	18	7	de	24	7
çocuk	36	8	da	18	7	da	22	8
da	36	8	sonra	17	9	o	17	9
ben	34	10	gel-	14	10	sonra	17	9
yıl	33	11	o	13	11	yıl	16	11
o	33	11	gün	13	11	gün	14	12
gazete	31	13	biri	12	13	ben	13	13
gün	27	14	var	12	13	çocuk	13	13
sonra	27	14	yıl	12	13	siz	12	15
siz	27	14	de-	12	13	var	12	15
film	26	17	git-	11	17	gel-	12	15
insan	24	18	zaman	11	17	de-	12	15
de-	23	19	mi	11	17	biz	11	19
yaş	22	20	ile	10	20	insan	10	20
var	22	20	iyi	10	20	dünya	10	20
biz	22	20	başla-	10	20	mi	10	20
festival	22	20	ama	10	20	kadar	9	23
en	21	24	kendi	10	20	ilk	9	23
mi	21	24	siz	10	20	git-	9	23
ilk	20	26	insan	10	20	en	9	23
dünya	20	26	biz	10	20	başla-	9	23
Selvihan	20	28	çocuk	10	20	biri	9	23
her	19	29	al-	10	20	zaman	9	23
gel-	18	29	daha	9	30	ile	8	30

Tablo 2 incelendiğinde, B1 seviyesi dinleme metinlerinde üç farklı ölçütün, ilk 30 kelime için ürettiği listelerin hem benzerlik hem de farklılıklar içerdiği görülmektedir. “bir” kelimesi her üç listede de birinci sırada yer almaktadır. “ve” ve “bu” kelimeleri ise tüm listelerde ilk üç içinde yer almakla birlikte, dağılım (range) listesinde aynı değere sahip oldukları için ikinci sırayı paylaşmaktadır. “ol-” ve “için” kelimeleri dördüncü ve beşinci sıralarda; “çok” kelimesi, “de” ve “da” bağlaçları da her üç listede ilk 10 kelime içinde yer almaktadır.

Mutlak frekans listesi incelendiğinde çok sayıda tekrara dayalı olarak belirli metinlerde yoğunlaşmış kelimeleri üst sıralara taşımakta olduğu gözlenmektedir. Bu listede, “çocuk” (f:36-8), “gazete” (f:31-13), “film” (f:26-17), “festival” (f:22-20), “yaş” (f:22-20), “Selvihan” (f:20-28, özel isim) ve “her” (f:19-29) gibi kelimeler yer almaktadır. Buna karşılık, dağılımı esas alan listenin farklı bir öncelik sunduğu görülmektedir. Bu listede, “sonra” (9.), “gel-” (10.), “gün” (11.), “biri” (13.), “de-” (13.), “var” (13.), “git-” (17.), “al-” (20.), “ama” (20.), “başla-” (20.), “iyi” (20.) ve “kendi” (20.) gibi kelimeler, metinler arasında yaygın olarak kullanıldıkları için öne çıkmaktadır. Burada önemli olan bir diğer husus ise mutlak frekans listesindeki ‘gazete’, ‘film’, ‘festival’, ‘yaş’, ‘Selvihan’ ve ‘her’ gibi kelimelerin dağılım listesinin ilk 30’unda hiç yer almamasıdır. Bu durum, bu kelimelerin kullanımlarının birkaç metinle sınırlı kaldığını ve dolayısıyla derlem geneli düşünüldüğünde önemlerinin mutlak frekans listesinde olduğu kadar yüksek olmayabileceğine işaret etmektedir.

Ortalama indirgenmiş frekans (ARF) listesi incelendiğinde “dünya” (20.) kelimesinin listeye dâhil olduğu görülmektedir. Diğer taraftan birkaç metinde yoğun kullanımları nedeniyle ‘gazete’, ‘film’, ‘festival’, ‘yaş’, ‘Selvihan’ ve ‘her’ gibi kelimelerin listede yer almadığı gözlenmektedir. Hem mutlak frekansı hem de dengeli dağılımı dikkate aldığı için diğer listelerin ilk 30’unda yer almayan ‘kadar’ (23.), ‘ilk’ (23.) ve ‘zaman’ (23.) gibi kelimelerin bu listede yer alabildiği görülmektedir. Ayrıca sıralamalarda da düzeltici etkileri görülmektedir: dağılım listesinde alt sıralarda olan ‘çocuk’ (20.) ve mutlak frekansta 10. sırada olmasına rağmen dağılım listesinde görünmeyen ‘ben’ kelimesi, ortalama indirgenmiş frekans listesinde 13. sıraya yükselmiştir. Bu durum, “çocuk” kelimesinin geçtiği metinlerde nispeten yüksek, “ben” kelimesinin ise geçtiği az sayıda metinde yoğun bir kullanım sergilediğine işaret etmektedir.

İkinci Araştırma Sorusuna Yönelik Bulgular

Çalışmanın ikinci sorusu, YİTÖS B2 dinleme metinlerinde en sık geçen 30 kelimeyi, mutlak frekans, dağılım (range) ve ortalama indirgenmiş frekans (ARF) ölçütlerine göre sıralamayı amaçlamaktadır. B2 seviyesine ait derlem üzerinde yapılan analizlerin sonuçları Tablo 3'te sunulmaktadır.

Tablo 3. *YİTÖS B2 seviye dinleme metinlerinde mutlak frekans, dağılım (Range) ve ortalama indirgenmiş frekansa (ARF) göre ilk 30 sözcüğün karşılaştırmalı sıralaması*

Kelime	**Mutlak Frekans**		**Kelime**	**Dağılım (Range)**		**Kelime**	**Ort. İnd. Frekans (ARF)**	
	Değer	**Sıra**		**Değer**	**Sıra**		**Değer**	**Sıra**
ve	175	1	bu	28	1	ve	110	1
ol-	165	2	ol-	28	1	ol-	99	2
bir	150	3	ve	27	3	bir	85	3
bu	110	4	bir	26	4	bu	68	4
için	62	5	çok	23	5	için	36	5
de	59	6	için	22	6	de	35	6
çok	52	7	de	21	7	da	30	7
ne	49	8	yap-	20	8	çok	30	7
da	48	9	da	20	8	yap-	26	9
yap-	48	9	ne	18	10	ne	25	10
daha	44	11	gel-	18	10	daha	24	11
insan	38	12	daha	18	10	o	20	12
gel-	36	13	o	17	13	gel-	20	12
biz-	36	13	gibi	16	14	gün	18	14
çalış-	34	15	iste-	16	14	sonra	18	14
su	34	15	en	16	14	en	17	16
o	33	17	sonra	16	14	çalış-	16	17
gün	32	18	var	15	18	var	15	18
de-	32	18	de-	15	18	her	15	18
en	31	20	her	15	18	siz	15	18
sonra	30	21	gün	14	21	iş	15	18
iş	30	21	önce	14	21	iste-	15	18
kadar	29	23	kadar	13	23	kadar	15	18
her	29	23	siz	13	23	biz	14	24
çalışma	27	25	başla-	13	23	gibi	14	24
hayat	27	25	al-	12	26	insan	14	24
var	26	27	i-	12	26	al-	13	27
başla-	26	27	biz	12	26	hayat-	13	27
iste-	25	29	ver-	12	26	de-	13	27
siz	25	29	çalış-	12	26	başla-	12	30

Tablo 3 incelendiğinde “ve”, “ol-” ve “bir” kelimeleri tüm listelerde ilk üç sırada yer almaktadır. Ancak “ve” kelimesi mutlak frekans ve ARF listelerinde birinci sıradayken, dağılım listesinde “bu” ve “ol-” kelimeleriyle aynı değere sahip olmamakla birlikte üçüncü sıradadır. “bu” kelimesi ise her üç listede de dördüncü sıradadır. “için”, “de”, “çok”, “da” ve “yap-” fiili her üç listede de ilk 10 kelime içinde yer almaktadır. Mutlak frekans listesi incelendiğinde, B1 seviyesinde olduğu gibi, belirli metinlerde yoğunlaşmış kelimelerin üst sıralarda olduğu gözlenmektedir. Bu listede, “ne” (f:49-8), “daha” (f:44-11), “insan” (f:38-12), “biz” (f:36-13), “çalış-” (f:34-15), “su” (f:34-15), “çalışma” (f:27-25) ve “hayat” (f:27-25) gibi kelimeler mutlak frekanslarından dolayı üst sıralarda yer almaktadır. Buna karşılık, dağılımı esas alan listenin farklı bir öncelik sunduğu görülmektedir. Bu listede, “yap-” (8.), “gel-” (10.), “gibi” (14.), “iste-” (14.), “sonra” (14.), “var” (18.), “de-” (18.), “her” (18.), “önce” (21.), “kadar” (23.), “başla-” (23.), “al-” (26.), “ise” (26.) ve “ver-” (26.) gibi kelimeler, metinler arasında yaygın olarak kullanıldıkları için öne çıkmaktadır. Burada önemli olan bir diğer husus ise mutlak frekans listesindeki ‘insan’, ‘biz’, ‘su’, ‘çalışma’ ve ‘hayat’ gibi kelimelerin dağılım listesinin ilk 10'unda hatta bazılarının ilk 20'sinde bile yer almamasıdır. Bu durum, bu kelimelerin kullanımlarının birkaç metinle sınırlı kaldığını ve dolayısıyla derlem genelindeki yaygınlıklarının mutlak frekansları kadar yüksek olmadığına işaret etmektedir.

Ortalama indirgenmiş frekans (ARF) listesi incelendiğinde diğer listelerdeki aşırı uçları dengeleyici bir rol oynadığı görülmektedir. Mutlak frekans listesinde üst sıralarda olan 'insan' (12.) ve 'hayat' (25.) gibi kelimeler ARF listesinde sırasıyla 24. ve 27. sıralara gerilemiş; dağılımı yüksek olan 'gün' (dağılımda 21., ARF'de 14.), 'sonra' (dağılımda 14., ARF'de 14.), 'iş' (mutlak frekansta 21., ARF'de 18.) ve 'her' (dağılımda 18., ARF'de 18.) gibi kelimeler ise nispeten daha üst sıralara yükselmiştir. Ayrıca diğer listelerin ilk 30'unda görülmeyen 'biz' (24.) kelimesinin bu listeye girebildiği görülmektedir.

TARTIŞMA / SONUÇ ve ÖNERİLER

Bu çalışma, YİTÖS B1 ve B2 seviye dinleme metinlerinde kullanılan kelimeleri mutlak frekans, dağılım ve ortalama indirgenmiş frekans gibi ölçütlerle değerlendirerek kelime sıralamasının, bu farklı istatistiksel yaklaşımlarla nasıl değiştiğini ortaya koymaya çalışmıştır. Analiz sonuçları, Kilgarriff'in (1997) "Deniz Salyangozu (The Whelk Problem) Problemi" olarak tanımladığı, bazı kelimelerin belirli bir veya birkaç metindeki yoğun kullanımının mutlak frekanslarını yapay bir şekilde arttırabilmesi sorununun Türkçe öğretim materyalleri bağlamında da geçerli olabileceğini göstermektedir. Bulgular, yalnızca mutlak frekansa dayalı bir yaklaşımın, kelimelerin dil kullanımındaki gerçek temsiliyet gücünü tespit etmede yetersiz ve yanıltıcı olabileceğine işaret etmektedir.

"bir", "ve", "ol-", "bu", "için" gibi kelimeler hem B1 hem de B2 seviye dinleme metinlerinde her listenin ilk sıralarında yer almaktadır. Bu durum, bu kelimelerin derlem genelinde hem sık hem de yaygın bir biçimde kullanılmış olmalarından kaynaklanmaktadır. Bununla birlikte farklı ölçütlerin kelime listelerinde önemli farklılıklar yarattığı gözlenmiştir. B1 seviyesinde "gazete", "festival", "Selvihan"; B2 seviyesinde "su", "çalışma", "hayat" gibi kelimeler, mutlak frekanslarının yüksek olmasına rağmen, dağılımı esas alan listenin ilk 30'unda ya hiç yer almamakta ya da çok alt sıralara kaymaktadır. Bu durum, bu kelimelerin kullanımlarının belirli birkaç metinle sınırlı kaldığını ve dolayısıyla derlem genelindeki gerçek önemlerinin mutlak frekanslarının gösterdiğinden düşük olduğunu göstermektedir. Bu yüzden materyal tasarımında sadece mutlak frekansı yüksek olan kelimelerin öncelenmesi, öğrenirleri, belirli bağlamlarla sınırlı, genel geçerliliği düşük bir kelime dağarcığıyla karşı karşıya bırakma riski taşıyabilir. Buna karşılık dağılım listeleri "sonra", "yap-", "gibi", "kadar", "başla-" gibi metinler arası yaygınlık gösteren kelimeleri öne çıkarmaktadır. Bu liste kelimelerin derlem içindeki dağılımını göstererek kelimelerin genel kullanım değeri hakkında daha ayrıntılı bir fikir vermektedir. Ortalama indirgenmiş frekans listeleri ise hem mutlak frekanstaki yapay artışları (gazete, su) düzeltmiş hem de dağılımı yüksek olan kelimeleri korumuştur. Daha da önemlisi her iki ölçüt için makul değerlere sahip olan kelimeleri (dünya, iş, her, insan) görünür kılmıştır. Ayrıca "çocuk" (B1) örneğinde olduğu gibi dağılımı nispeten düşük ama geçtiği metinlerde yoğun kullanımı olan ve bu nedenle önemini koruması gereken kelimelerin de listede yer almasını sağlamıştır. Bu nedenle ortalama indirgenmiş frekans, kelimenin derlem içindeki gerçek değerini daha isabetli bir şekilde tespit etmiştir.

Bu çalışmanın bulguları Türkçenin gerek ana dili gerekse yabancı dil olarak öğretiminde kelime seçimi ve materyal geliştirmeye ilişkin önemli sonuçlar ortaya koymaktadır. Öncelikle öğretim materyallerindeki söz varlığını değerlendirirken ve sıralarken yalnızca mutlak frekans listelerine odaklanmak bağlama aşırı bağımlı sonuçlar üretebilmektedir. Bu durum, alanyazında daha önce yapılan ve yalnızca mutlak frekansa odaklanan birçok çalışmanın bir sınırlılığına işaret etmektedir. Kelime listeleri oluşturmak için kelimelerin hem ne sıklıkta hem de ne kadar yaygın kullanıldığını birlikte ölçen istatistiksel yöntemlere başvurulmalıdır. Dağılım bu açıdan önemli bir ilk adımdır, ancak ortalama indirgenmiş frekans gibi her iki boyutu tek bir metrikte birleştiren bileşik ölçütler daha kapsamlı ve dengeli bir resim sunmaktadır. Ortalama indirgenmiş frekansın kullanılması, öğrencilere yalnızca "sık" değil, aynı zamanda "yaygın" kullanımı olan kelimeleri kazandırma potansiyeli taşımaktadır.

Sonuç olarak bu çalışma, Yabancı dil olarak Türkçe öğretiminde kelime seçimiyle ilgili geleneksel yaklaşıma istatistiksel bir alternatif önermekte ve alandaki kuramsal tartışmaya bir katkıda bulunmaktadır. Öğretim materyali geliştiricileri, sınav hazırlayıcıları ve dil araştırmacıları için

kelimelerin değerini belirlerken tek boyutlu bir frekans okumasının ötesine geçmek ve dağılım ile indirgenmiş frekans gibi kavramları analize dâhil etmek daha sağlam ve öğrenci ihtiyaçlarına daha uygun dil kaynaklarının ortaya çıkmasını sağlayabilir.

Öneriler

Bu çalışma B1-B2 seviye dinleme metinleri ile sınırlı görece küçük ölçekli bir derlem üzerinde gerçekleştirilmiştir. Benzer yöntemin okuma metinlerine, yazma ve konuşma materyallerine ve özellikle daha ileri veya daha temel seviyelere uygulanması, Türkçenin yabancı dil olarak öğretimine yönelik kapsamlı bir "önemli kelimeler" listesinin oluşturulmasına katkı sağlayacaktır. Bu çalışmada, temel bir dağılım ölçütü olan "range" ve birleşik bir ölçüt olan ortalama indirgenmiş frekans kullanılmıştır. Gries (2008) tarafından sınıflandırılan diğer dağılım ölçütleri veya başka indirgenmiş frekans formülleri ile karşılaştırmalı analizler yapmak ve bu ölçütlerin Türkçe gibi sondan eklemeli bir dildeki performansını değerlendirmek faydalı olacaktır. Kalıp ifadelerin ve dil bilgisel yapıların dağılımını ve ortalama indirgenmiş frekansın değerlerini inceleyen araştırmalar, öğretilmesi gereken dil birimlerinin seçimine daha bütüncül bir yaklaşım getirebilir. Son olarak, bu istatistiksel yöntemlerle belirlenen sözcük listelerinin etkililiği deneysel olarak test edilebilir.

KAYNAKLAR

Aksan, D. (1998). *Her Yönüyle Dil-Ana Çizgileriyle Dilbilim*: TDK yayınları.

Baş, B. (2011). Söz Varlığı ile İlgili Çalışmalarda Kullanılacak Ölçütler. *TÜBAR(XXIX),* 27-61.

Göçen, G., & Aydın, E. (2021). Yabancılar İçin Hazırlanmış Türkçe Okuma Kitaplarındaki Söz Varlığı: Çocuk Hikâyeleri Dizisi A1-A2 Örneği. *Uluslararası Yabancı Dil Olarak Türkçe Öğretimi Dergisi, 4(1), 93-126.*

Göçen, G., & Bingöl, S. (2024). Yabancı Dil Olarak Türkçe Öğretimi Ders Kitaplarında Yer Alan Söz Varlığı Unsurları: "Anahtar" B2 Ders Kitabı Örneği. *Journal of Sustainable Education Studies, Özel Sayı (Ö3), 74-101.*

Göçen, G., & Buluş, F. (2022). Yabancılar için "Hayat Boyu Türkçe" ders kitaplarında yer alan söz varlığı unsurları. *Journal of Sustainable Education Studies, Özel Sayı (Ö1), 70-87.*

Göçen, G., & Okur, A. (2016). Yabancılar için Türkçe ders kitaplarındaki sözcüklerin kullanım sıklığı ve yaygınlığı. *Milli Eğitim Dergisi, 45*(210), 447-476.

Göçen, G., Gümüş, B., Kışla, C., & Güdek, G. (2020). Türkçeyi Yabancı Dil Olarak Öğrenenler İçin Hazırlanan Okuma Kitaplarında Yer Alan Sözcüklerin İncelenmesi. *Journal of Sustainable Education Studies, 1*(1), 33-63.

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International journal of corpus linguistics, 13*(4), 403-437.

Kilgarriff, A. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography, 10*(2), 135–55.

Kurudayıoğlu, M., & Karadağ, Ö. (2005). Kelime Hazinesi Çalışmaları Açısından Kelime Kavramı Üzerine Bir Değerlendirme. *Gazi Eğitim Fakültesi Dergisi, 25*(2), 293-307.

Özdemirel, A. Y., & Dilidüzgün, Ş. (2018). Yabancı Dil Olarak Türkçe ve İngilizce Ders Kitaplarındaki Sözcüklerin Kullanım Sıklığı Bağlamında Değerlendirilmesi. *OPUS International Journal of Society Researches, 9*(16), 1464-1505. https://doi.org/10.26466/opus.467804

Savický, P., & Hlavácová, J. (2002). Measures of Word Commonness. *Journal of Quantitative Linguistics, 9*(3), 215–231. https://doi.org/10.1076/jqul.9.3.215.14124

Vardar, B. (2002). *Açıklamalı Dilbilim Terimleri Sözlüğü*: Multilungual Yayınları.

Extended Abstract

This study aims to reevaluate the statistical approaches used in vocabulary selection for teaching Turkish as a foreign language, offering a perspective that moves beyond the traditional focus on the absolute frequency criterion. Vocabulary acquisition is a critical component underpinning language learning, directly influencing the development of all four core skills. Therefore, the methodology for selecting which words to present in instructional materials carries significant pedagogical weight. While word frequency lists are a fundamental tool in this selection process, relying solely on a word's raw count of occurrences within a specific corpus—its absolute frequency—entails considerable methodological risks. As highlighted by Kilgarriff's (1997) "Whelk Problem," a word can attain a high frequency rank artificially if it is repeated extensively within one or a few specialized texts (e.g., the word "whelk" in a book about marine life), without being widely useful across general language contexts. This scenario underscores that absolute frequency is inherently tied to the specific composition, topic, and length of the corpus analyzed. To obtain more reliable and pedagogically sound data, it is necessary to complement frequency with measures of a word's distribution—how widely it is spread across different texts—and to employ composite statistical metrics that balance both frequency and distribution. This research argues for the integration of such advanced criteria, specifically distribution (range) and Average Reduced Frequency (ARF), into the evaluation and design of Turkish language teaching materials. The study's originality stems from addressing a notable gap in the field, as no prior research on teaching Turkish as a foreign language has concurrently applied absolute frequency, distribution, and reduced frequency measures to assess instructional vocabulary. Consequently, it proposes a substantive alternative to the prevalent, one-dimensional frequency-based approach. The investigation adopted a quantitative, descriptive, and comparative document analysis design. The research sample consisted of the B1 and B2 level listening texts from the widely used Yedi İklim Türkçe Öğretim Seti (YİTÖS), comprising 27 and 29 texts respectively, for a total of 56 texts. These transcribed texts were converted into digital format and organized into two separate corpora (B1 and B2). Following lemmatization processes to consolidate word forms (e.g., reducing inflected verbs to their base form "ol-"), the corpora were analyzed using SketchEngine software. Three key statistical measures were calculated for the vocabulary in each corpus: the Absolute Frequency (raw count), the Distribution or Range (number of different texts a word appears in), and the Average Reduced Frequency or ARF (a metric that calculates frequency while mitigating the effect of clustered repetitions by considering distribution across text segments). For each level, the top 30 words based on these three distinct criteria were identified, and their comparative rankings were meticulously analyzed. The findings revealed both expected consistencies and divergences across the different lists. As anticipated, core function words such as "bir" (a/one), "ve" (and), "ol-" (to be), and "bu" (this) consistently ranked at the top across all three lists for both proficiency levels, confirming their fundamental role as high-frequency items with broad distribution. The critical insight, however, emerged from the discrepancies. The absolute frequency lists included words like "gazete" (newspaper), "festival," and the proper noun "Selvihan" in B1, and "su" (water), "çalışma" (study/work), and "hayat" (life) in B2. Most of these contextually bound words were absent from the top 30 of the corresponding distribution lists, clearly indicating their usage was concentrated in only a handful of theme-specific texts rather than being widespread across the material. Conversely, the distribution lists highlighted words like "sonra" (after), "yap-" (to do), "gibi" (like), and "başla-" (to start), which demonstrated utility across a wider array of texts and communicative situations. The ARF lists demonstrated a synthesizing and balancing effect. They effectively demoted words whose high rank was solely due to concentrated repetition in a few texts (e.g., "gazete," "su") while preserving and often promoting words with good distribution. Furthermore, ARF brought forward words like "dünya" (world) and "kadar" (until) in B1, and "iş" (job) and "her" (every) in B2, which possessed reasonably good scores in both frequency and distribution, thus offering a more nuanced picture of a word's overall importance within the corpus. The conclusions drawn from this comparative analysis are significant for material development and pedagogical practice. They empirically validate that an over-reliance on absolute frequency alone can distort vocabulary selection, potentially leading to the inclusion of words with low generalizability at the expense of more widely useful items. This poses a tangible risk of providing learners with a lexicon that is inadequately balanced for diverse communicative needs. While distribution serves as a vital corrective lens to identify widely-used vocabulary, the Average Reduced Frequency metric proves to be a particularly robust tool for material designers and researchers. By integrating both dimensions of occurrence—sheer count and spread—into a single, refined value, ARF offers a more reliable and pedagogically insightful ranking. It filters out statistical "noise" from topic-bound repetitions and highlights words that are both reasonably common and broadly applicable, or crucially important within their justified, specific contexts. Based on these findings, several avenues for future research are recommended. The methodological framework should be applied to other skills (reading, writing, speaking) and across all proficiency levels (A1 through C2) to contribute to a comprehensive, multi-criteria core vocabulary list for Turkish as a foreign language. Further comparative studies could explore the efficacy of other distribution

measures or reduced frequency formulas, particularly for agglutinative languages. The analysis could also be productively extended to lexical chunks, collocations, and grammatical structures. Ultimately, experimental studies are necessary to evaluate the pedagogical effectiveness of word lists generated using these combined statistical criteria in actual classroom settings. In summary, this study moves the theoretical discussion forward by demonstrating that a multi-faceted statistical approach, which strategically incorporates distribution and reduced frequency alongside traditional counts, can establish a firmer, more reliable foundation for selecting and grading vocabulary in teaching materials, ultimately aiming to create more effective and learner-centred resources for teaching Turkish as a foreign language.