Gazi Üniv. Müh. Mim. Fak. Der.
Cilt 23, No 2, 467-475, 2008

J. Fac. Eng. Arch. Gazi Univ.
Vol 23, No 2, 467-475, 2008

TÜRKÇE'DE KULLANILAN İŞLEV KELİMELERİN ZİPF 1. KANUNU ESASINDA DEĞERLENDİRİLMESİ

Senem KUMOVA METİN
Bilgisayar Bilimleri Fakültesi, İzmir Ekonomi Üniversitesi, 35330 Balçova-İzmir
senem.kumova@ieu.edu.tr

(Geliş/Received: 24.07.2007; Kabul/Accepted: 30.01.2008)

ÖZET

Bu çalışmada doğal dil içinde gramer yapısının oluşturulması amacıyla kullanılan, içinde bulunduğu metnin taşıdığı enformasyon miktarını değiştirmeyen kelimeler (işlev kelimeler) araştırılmıştır. Araştırmanın temelini Zipf'in 1. Kanunu'nun Türkçe metinler üzerinde sınanması ve yüksek frekanslı kelimelerin işlev kelime olacağı beklentisi oluşturmaktadır. Çalışmada önerilen yöntem ile ilgili testler Türkçe metinler içeren farklı derlemler üzerinde yapılmış, sonuçlar değerlendirilmiştir.

Anahtar Kelimeler: İşlev kelime, içerik kelime, test derlemleri.

EVALUATION OF FUNCTION WORDS IN TURKISH BASED ON THE ZIPF'S 1. LAW

ABSTRACT

In this study, function words that are used to construct the grammatical structure in Natural Language and that does not change the information content of the text have been investigated. Application of Zipf's first Law on Turkish texts and expectance of high frequency words to be function words constitute the fundamentals of the research. In the study the test for the proposed method has been performed on different corpus including Turkish texts ans results have been evaluated.

Keywords: Function word, content word, test corpus.

1. GİRİŞ (INTRODUCTION)

Dil içinde geliştirilen, hâlihazırda kullanılan kelimelerin sayısı ve görevlerini modellemek için pek çok çalışma yapılmıştır. Bu konuyla ilgili olarak geliştirilmiş yüzlerce sözlük bulunmasına rağmen sözlükler içinde bulunmayan kelimeler günlük hayatta karşımıza çıkmaktadır.

Bir dildeki farklı kelime sayısının değerlendirilmesinde iki farklı yaklaşım söz konusudur [1]. Bunlardan birincisi dillerin kapalı kelime dağarcığına sahip olduğu düşüncesidir. Bu varsayımda dil içindeki kök kelime ve ek sayısının sabit olduğu bilindiği için üretilebilecek yeni kelimelerin de kısıtlı olduğu kabul edilmektedir. Diğerinde ise dilin sürekli gelişen bir yapısı olduğu düşüncesi hâkimdir. Bu yaklaşımda dilin kelime dağarcığı yani dil içindeki farklı kelime sayısının sürekli arttığı ve geliştiği dolayısıyla kelime dağarcığının açık olduğu kabulü yapılmaktadır.

Dilin modellenmesi çalışmalarında karşılaşılan diğer bir problem ise anlam bütünlüğünün sağlanması için cümle içinde kullanılan bağlaç, zarf, edat, zamir vb. gibi harç görevi yapan kelimelerin saptanması ve kullanım miktarının belirlenmesidir. Bu kelimelere cümle içinde anlama katkıda bulunmaktan çok gramer yapısının oluşmasında görev aldıkları için *işlev (function) kelime* adı verilmektedir. Türkçe içinde yaygınlıkla kullanılan işlev kelimelere örnek olarak "ve, veya, gibi, ile, ki, ben, sen, o" gibi kelimeler verilebilir. İngilizce'de ise "and, a, be, but, about, above" kelimeleri sıklıkla rastlanan işlev kelimelerdir.

Dil içinde metin veya cümleye anlam kattığı kabul edilen kelimelere *içerik (content) kelime* denir. İçerik kelimeler genelde bir kavram veya olayı simgeleyen isim veya fiillerden oluşur.

Dilin kelime dağarcığı ve dilde kullanılan işlev kelimeler arasında bir ilişki olması gerektiği hâlihazırda yapılmış olan doğal dil işleme çalışmaları sonucunda kabul gören bir saptamadır. Makale içinde Zipf Birinci Kanunu ve İngilizce üzerinde Kornai'nin yaptığı çalışmalardan faydalanılarak Türkçe için işlev kelimelerin sayısı ve kelime dağarcığının miktarı incelenmektedir. Makale Zipf Kanunları, test derlemleri, yöntem, sonuçlar ve değerlendirme bölümlerinden oluşmaktadır.

2. ZİPF KANUNLARI (ZIPF LAWS)

Doğal dil işleme alanında kelimelerin kullanım sıklıkları, bir kelimenin taşıdığı anlam sayısı, kelimeler arası uzaklıklar gibi çok önemli unsurlar George K. Zipf (1902–1950) tarafından ortaya konulan kanunlar temel alınarak belirlenmektedir. Zipf'in bu konuda sıkça kullanılan 4 kanunu mevcuttur [2].

Yazılı metinlerdeki kelime dağılımı ve çeşitliliği, dilin temsili konusunda önemli bir göstergedir. Bu sebeple kullanılan sayı, simge veya kelimelerin miktarı derlemin değerlendirilmesi çalışmalarında yer almaktadır. Bu konuda Zipf "Human Behavior and the Principle of Least Effort" kitabında en az gayret ilkesinin kelimelerin kullanımı konusunda da uygulanabileceğini vurgulamıştır, Zipf birinci kanunu şu şekildedir:

Tablo 1. Zipf birinci kanununun Tom Sawyer romanı üzerinde deneysel değerlendirmesi (Emprical evaluation of Zipf's 1. law on Tom Sawyer) [3]

Kelime	Sık(*f*)	Sıra(*r*)	*f x r*
the	3332	1	3332
and	2972	2	5944
a	1775	3	5325
he	877	10	8770
but	410	20	8200
be	294	30	8820
there	222	40	8880
one	172	50	8600
about	158	60	9480
more	138	70	9660
never	124	80	9920
Oh	116	90	10440
two	104	100	10400
turned	51	200	10200
you'll	30	300	9000
name	21	400	8400
comes	16	500	8000
group	13	600	7800
lead	11	700	7700
friends	10	800	8000
begin	9	900	8100
family	8	1000	8000
brushed	4	2000	8000
sins	2	3000	6000
Could	2	4000	8000
Applausive	1	8000	8000

Bir derlemdeki tüm kelimeler tek tek sayılıp, en yüksek sıklığa sahip kelimeden azalan sırada numaralandırıldığında her kelimenin gözlenme sıklığı (frekans, f) ve sıra numarasının çarpımı sabit bir değerdir.

$$f \cdot r \approx c(sabit) \qquad (2.1)$$

Bu kanun ilk olarak Estoup (1916) tarafından ortaya atıldıysa da Zipf tarafından yaygınlaştırıldığı için onun adı ile anılmaktadır.

En az gayret ilkesi gereği konuşmacılar birbirinden farklı az sayıda kelime kullanarak farklı kavramları ifade etme isteği duyarlar. Yani bir kelimenin birden fazla anlamı karşılamasını sağlayarak kelime dağarcıklarını sınırlı tutma istekleri vardır. Dinleyiciler ise her farklı kavram için farklı kelime duymak ve böylece anlama gayretlerini en aza çekmek isterler. Bu iki isteğin dengelenmesi sonucu kelimelerin gözlenme sıklığı ve sıra numarası çarpımlarının sabitlenmesi bu kanunun temelidir.

İngilizce üzerine yapılan araştırmalarda [3] Zipf birinci kanununda belirtilen sabit bulunmaya çalışılmıştır. Tablo 1'de görüldüğü üzere yaklaşık bir $f \cdot r$ değeri belirlenebilir.

Çalışma içinde temel alınan Zipf 1. kanunu dışında diğer kanunları kısaca şöyle listeleyebiliriz:

Zipf 2. Kanunu : *i* kere rastlanan farklı kelime şekli/biçimi sayısı (*V(i,N)*)ve bu *i* sıklığı arasında denklem 2.2'de belirtilen şekilde bir ilişki vardır

$$\log(i) = K_N - D_N \log(V(i,N)) \qquad (2.2)$$

Zipf 3. Kanunu: Bir kelimenin karşıladığı farklı anlam sayısı (w) ile kelimenin gözlenme sıklığının (f) karekökü arasında doğrusal bir ilişki vardır.

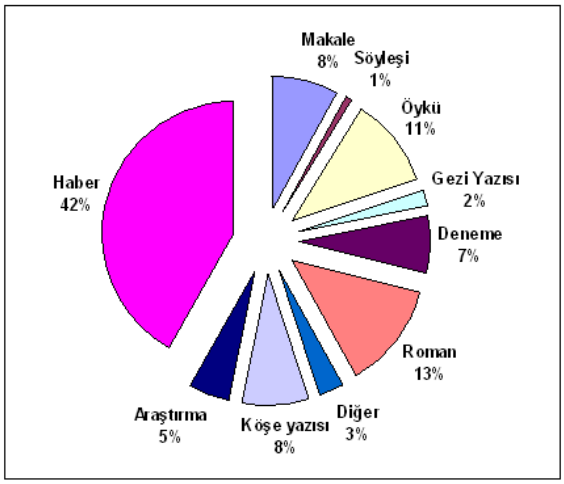
$$w \propto \sqrt{f} \qquad (2.3)$$

Zipf 4. Kanunu : İçeriği oluşturan kelimeler, metin içinde bazı yerlerde yığılım gösterir. Kelimenin metin içinde gözlendiği yerlerin araları (*I*) ile satır veya sayfa cinsinden sıklık (*F*) arasında denklem 2.4'de belirtildiği şekilde bir ilişki vardır.

$$F \propto I^{-\rho} \qquad (2.4)$$

3. TEST DERLEMLERİ (TEST CORPUS)

Yapılan çalışmada Türkçe derlemlerden faydalanılmıştır. Bunlar: Bilkent derlemi, ODTÜ derlemi ve çalışma amacıyla geliştirilen Makaleler derlemidir.



Şekil 1. ODTÜ derlem yapısı (ODTU corpus structure) [5]

Bilkent derlemi Bilkent Üniversitesinde hesaplamalı dilbilim çalışmalarının sonucu otomatik işaretlenmiş bir derlemdir [4]. Bilkent derlemi yalın haliyle yani sadece içindeki kelimelerin bulunduğu bir formatta ve Dinçer (2004) tarafından gerekli düzeltmelerin yapıldığı son haliyle tez içinde kullanılmıştır. Derlem bu haliyle kullanıldığında ~728172 adet toplam kelime ve ~24359 adet farklı kelimeden (kelime dağarcığı) oluşmaktadır. Derlemin içeriği hazırlandığı yıllarda çıkan gazete yazıları ve makalelerden oluşmaktadır [5]. Derlem içindeki kelimeler çekim eklerinden ayrılmış halde bulundukları için kullanılan diğer örnek derlemlerden daha farklı sonuçlar vermesi beklenen bir durumdur.

ODTÜ Derlemi yüzeysel formda (surface form) bulunan toplam 1987447 kelimeden oluşmaktadır. Bu derlem için kelime dağarcığı ~212852 adet kelimedir. Derlem içinde pek çok farklı konuda metin bulunmaktadır.

Tablo 2. Makaleler derleminin genel yapısı (Makaleler (Articles) corpus structure)

KELİME DAĞARCIĞI	TOPLAM KELİME SAYISI	YAZAR
893	1428	Y.Toruner
1042	1538	B.Coskun
1152	1776	M.Tamer
1494	2280	D. Sazak
1564	2351	D. Hızlan
1441	2461	S. Kohen
1596	2499	H. Pulur
1865	3150	M. Yılmaz
1950	3317	Y. Congar
2087	3810	E. Özkök
2289	3913	F. Altaylı
2056	4011	H. Güneş
2647	4204	M. Aşık
2608	4362	O. Ekşi
3142	7229	E. Kumcu
3888	6913	Ç. Altan
4915	11818	F. Bila
9177	18973	C. Dündar
7923	22408	H.Cemal
9702	24237	G. Civaoglu
20938	82593	M.A. Birand
22358	100603	E. Çölaşan

Genel istatistiklerin yanında konu dağılımı, derlemi oluşturan metinlerin yapısı ve metin yazarları derlemin enformasyon miktarını doğrudan etkilemektedir. Bu sebeple şekil 1'de de görüldüğü üzere farklı konuları içeren ODTÜ derlemine ait sonuçlar araştırma içinde önem kazanmaktadır. Bilkent derlemi ise sadece gazete yazılarından oluşmuştur dolayısıyla tek bir belge türüne sahiptir [5].

Makaleler derlemi özellikle güncel kelimeleri içermesi sebebiyle çalışma içinde kullanılmıştır. Derlem 22 farklı makale yazarı tarafından değişik zamanlarda gazetelerde yayınlanmış köşe yazılardan oluşmaktadır. Tablo 2'de derlemi oluşturan metinler, metin yazarları ve ilgili kelime adetleri bulunmaktadır.

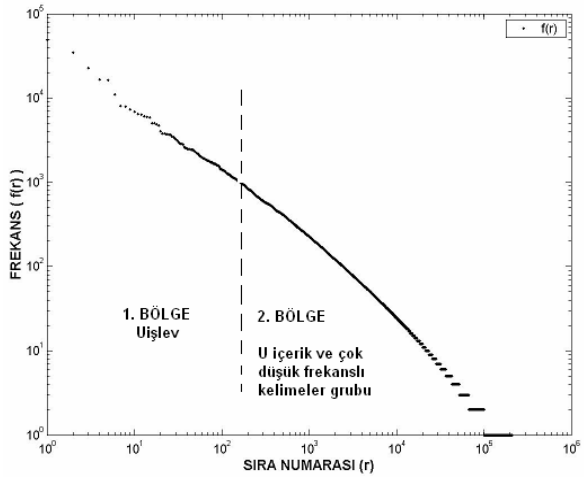
Makaleler derlemi ekonomi, siyaset ve güncel olayları işleyen makaleleri içermektedir. Makalelerde imlâ işaretleri metinlerden çıkartılarak, kelimeler metin içinde bulundukları halleriyle (çekim ve yapım ekleri varken) değerlendirilmiştir. Bu durumda derlem yaklaşık 310000 kelimeden oluşmaktadır. Derlemin kelime dağarcığı ise 57700 kelime civarındadır.

4. YÖNTEM (METHOD)

Kelimelerin yazılı bir metinde gözlenme sıklıkları ile taşıdıkları ve metne kattıkları anlam arasındaki ilişki Luhn (1958) tarafından tanımlanmıştır. Bu tanımlamada dilin bütünü düşünüldüğünde çok düşük ve çok yüksek frekansta rastlanan yani sıkça kullanılan kelimelerin anlama kattığı değer açısından önemsiz, orta frekansta kullanılan kelimelerin ise önemli olduğu belirtilmiştir. Önemli olduğu kabul edilen bu orta frekanstaki kelimeler içerik, yüksek frekanslı kelimeler ise işlev kelimeler olarak düşünülmektedir.

İlk olarak Herdan (1960) ile başlayarak içerik ve işlev kelimelerin farklı iki grup içinde değerlendirilmesi söz konusu olmuştur [1]. İşlev kelimeler $U_{işlev}$, içerik kelimeler $U_{içerik}$ grupları ile ifade edilebilir. Yapılan çalışmalarda bu iki grup arasındaki ayrım noktasının saptanması için frekans-sıra numarası grafikleri dikkate alınmaktadır. Bu grafiklerin elde edilmesi için pek çok farklı metnin bir araya getirilmesi ile oluşturulmuş dili modellediği varsayılan derlemler kullanılmaktadır.

Frekans-sıra numarası grafiklerinde işlev ($U_{işlev}$) ve içerik kelimeler ($U_{içerik}$) şekil 2'de temsili olarak gösterilmiştir. $U_{işlev}$ yüksek frekanslı kelimeler, $U_{içerik}$ grubu ise orta ve düşük frekanslı kelimelerden oluşmaktadır.



Şekil 2. ODTÜ derleminde logaritmik frekans ve sıra numarası eğrisi, işlev ve içerik kelimeler için muhtemel bölgeler belirtilmiştir. (Logaritmic frequency-rank curve for ODTU corpus, area for function and content words has been defined)

Grafikte sıra numaraları [1-*V*] aralığında artan değerlere sahiptir (*V*: Toplam kelime dağarcığı). Eğriyi ayrık değerler yerine sürekli doğrusal bir fonksiyon olarak kabul edersek eğrinin altında kalan alan şöyle ifade edilebilir:

$$N = \int_0^V f(r) \cdot dr \qquad (4.1)$$

Denklem 4.1'de *N* toplam kelime sayısını dolayısıyla derlem büyüklüğünü, *V* ise en yüksek sıra numarasını dolayısıyla derlemin kelime dağarcığını temsil etmektedir. Grafikteki frekans $f(r)$ ve sıra numarası r değerleri toplam kelime sayısına (*N*) bölünerek grafik birim kareye dönüştürülüp değerler normalize edilebilir. Bu durumda göreceli sıra numarası değeri r_N, göreceli frekans eğrisi $f_N(r_N)$ olarak ifade edilirse eğri altında kalan alan şöyle tanımlanabilir ($f_N = f/N$ ve $r_N = r/N$ için):

$$\int_0^{V/N} f_N(r_N) \cdot dr_N = \frac{1}{N} \qquad (4.2)$$

Kornai (2002), $f_N(r_N)$ eğrisi üzerinde yaptığı incelemede bu eğrinin $\exp(-c \cdot r_N)$ fonksiyonu ile ifade edilebileceğini ortaya koymuştur. Grafik ile ilgili olarak önermeleri şu şekildedir:

(1) $f_N(r_N) = \exp(-D_N \cdot r_N)$

(2) sol limit dikkate alınırsa
$f_N(1/N) = sabit = \exp(-c)$

(3) doğrusal alan kuralından

$$\int_{1/2N}^{(V+1/2)/N} f_N(r_N) \cdot dr_N = \frac{1}{N}$$

Bu önermeler göz önüne alınarak eğri altında kalan alan hesaplandığında 1 ve 2 önermeleri ile $D_N = c \cdot N$ ve integral sonucu ise şu şekilde olur:

$$1/N = \int_{1/2N}^{(V+1/2)/N} \exp(-c \cdot N \cdot r_N) dr_N$$
$$= \frac{1}{cN}(\exp(-c/2) - \exp(-c(V(N)+1/2)) \qquad (4.3)$$

Derlemin yeterince büyük olduğu kabul edilirse $N \to \infty$ ve $V \to \infty$ alınabilir bu durumda *c*=0.7035 gibi bir değer elde edilir. Bu değer en yüksek frekanslı kelimenin %49.4866 gibi bir yoğunlukla derlem içinde bulunduğunu göstermektedir [1].

Kornai, yaptığı çalışmada denklem 4.3 sonucunda derlemin yarısının işlev kelimelerden oluştuğunu belirtmiştir.

İşlev kelimelerin yüksek frekanslı kelimelerden oluşması genel bir görüştür. Ancak yapılan çalışmalarda işlev kelime olmasına rağmen düşük frekanslara sahip olan kelimelere ve yüksek frekanslı içerik kelimelere de rastlanmaktadır. Ayrıca bir cümlede içerik kelime olarak kullanılan bir kelime diğer bir cümlede işlev kelime olarak da kullanılabilir. Bu sebeple işlev ve içerik kelimeler arasında kesin bir sınır belirlemek belli bir hata payı ile kabul edilebilir.

Kornai (2002), çalışmasında frekans-sıra numarası değerlerinden faydalanarak $U_{işlev}$, $U_{içerik}$ gruplarının belirlenmesi ve içerik-işlev kelimeler arasında muhtemel sınırın çizilmesi üzerinde durmuştur.

$U_{işlev}$, $U_{içerik}$ gruplarının belirlenmesi çalışmasında dikkate alınabilecek derlem özellikleri şunlardır:

- kelime sayısı - kelime dağarcığı ilişkisi
- frekans-sıra numarası ilişkisi

Dil içindeki bütünlüğün sağlanması gerekliliği bu iki veri grubunun ilişkili olması gerekliliğini de ortaya çıkarmıştır. Dolayısıyla frekans-sıra numarası ve kelime sayısı-kelime dağarcığı eğrileri bir arada değerlendirilmelidir.

Kelime sayısı-kelime dağarcığı ilişkisinin tanımlanması ve ilgili denklemin belirlenmesi ile ilgili olarak pek çok rasyonel ve deneysel

yöntemler bulunmaktadır [6]. Genel anlamda bu ilişki artan yönde kabul edilmektedir. *V-N* eğrilerinin genel yapısı dikkate alınarak şu şekilde ifadesi mümkündür [1]:

$$V(N) = N^{\rho} \qquad (4.4)$$

Denklem 4.4'de $V(N)$ değerlerinin *N*'in bir kuvveti ile arttığı belirtilmektedir. ρ katsayısı, [0 1] aralığında olup yazım tarzı, yazar vb. gibi bir takım özelliklere bağlı ancak *N* değerinden bağımsız bir parametredir.

İçerik ve işlev kelimelerin ayrımı konusunda dikkate alınacak bir diğer özellik olan frekans ile sıra numarası arasındaki ilişkinin doğru denklemini ise Zipf birinci kanunu şu şekilde betimlemektedir:

$$\log(f_N) = H_N - B_N \log(r) \qquad (4.5)$$

Denklem 4.5'de f_N kelimelerin gözlenme sıklığını (göreceli frekans), *r* sıra numarasını ve H_N derlem büyüklüğüne bağlı sayısal bir sabiti simgelemektedir. B_N ise frekans-sıra numarası eğrisinin eğimini simgeleyen derlem büyüklüğüne bağlı bir sabittir.

İçerik ve işlev kelime gruplarının belirlenmesinde belli bir sıra numarası örneğin *k* sıra numarasına sahip kelimenin ssınır kelime olduğu kabul edilirse işlev kelimelere ait toplam olasılık şu şekilde ifade edilebilir [1]:

$$P_K = \sum_{r=1}^{k} p_r \qquad (4.6)$$

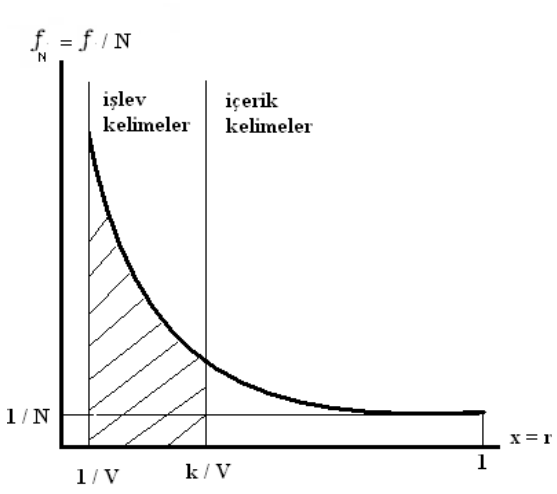
Denklem 4.6'da verilen olasılık değeri aslında frekans-sıra numarası eğrisinin altında kalan alanın hesaplanması ile elde edilebilen bir değerdir. Şöyle ki, örnek derlemin tüm dili temsil ettiği düşünülürse her kelimenin dil içinde rastlanma olasılığı, frekans değerinin toplam kelime sayısına (*N*) bölümüdür.

Denklem 4.5 gereği frekans–sıra numarası eğrisinin frekans değerleri *N* ile normalize edilerek sıra numarası değerleri ise *V* yani kelime dağarcığı ile normalize edilerek grafik birim kare haline dönüştürülebilir. Şekil 3'de temsili bir frekans-sıra numarası eğrisi verilmiştir.

Şekil 3'de gösterilen taralı alan işlev kelimelerin toplamını simgelemektedir. Grafik incelendiğinde eğride görülen özellikler şunlardır:

D1. sağ limit $f_N(1) = 1/N$

D2. sol limit $f_N(k/V(N)) = sabit$



Şekil 3. Temsili frekans-sıra numarası eğrisi, frekans değerleri *N* ile sıra numarası değerleri *V* ile normalize edilmiştir. (Schematic frequency-rank curve, frequency values are normalized by N and rank values are normalized by V)

D3. *k/V* ile sağ limit arasındaki alan,

$$\int_{k/V(N)}^{1} f_N(x)dx = (1 - P_k)/V(N)$$

Denklem 4.5'de *r* yerine *x.V* koyularak (*r=x.V*) ilgili grafik birim kareye dönüştürülür ve şekil 4.4'de belirtilen eğriye ait denklem olarak kullanılabilir. Bu durumda şu eşitlik elde edilir :

$$f_N(x \cdot V) = \exp(H_N - B_N \cdot \log(x \cdot V)) \qquad (4.7)$$

D1, D2, D3 özellikleri denklem 4.7'de değerlendirilirse $f_N(1) = \exp(H_N) = 1/N$ durumundan $H_N = -\log(N)$ elde edilir. Sonuçta denklem 4.7 şu hale dönüşür:

$$f_N(x \cdot V) = \frac{1}{N \cdot (x \cdot V)^{B_N}}$$

yani

$$f_N(x) = \frac{1}{N \cdot x^{B_N}} \qquad (4.8)$$

Tüm derlemin hem Zipf birinci kanununa hem de denklem 4.4'de verilen ilişkiye uygunluğu kabul edilirse *N* büyümesine rağmen $f_N(k/V(N)) = 1/(N \cdot (k/V(N))^{B_N})$ eşitliğinin yani $(N \cdot (k/V(N))^{B_N})$ değerinin sabit kalması beklenir.

Tablo 3. Merc derleminde sınır kelime olması muhtemel kelimeler ve ilgili değerleri (Possible threshold words and their B values)

Kelime	Sıra numarası	Frekans	B
be	30	0.0035	1.66
had	75	0.0019	1.45
other	140	0.0012	1.36
me	220	0.00051	1.41

Dolayısıyla $\log(N)+B_N\cdot\log(k)-B_N\cdot\log(V(N))$ bir sabittir. $B_N\approx B$ kabulüyle $B\cdot\log(k)$ değeri sabit bir değer olarak alınır. Bu durumda $\log(N)-B\cdot\log(V(N))$ değerinin sabit kalması için $\log(N)\approx B\cdot\log(V(N))$ olarak kabul edilebilir.

Denklem 4.4 dikkate alındığında $\log(N)\approx B\cdot\log(V(N))$ sonucu frekans-sıra numarası ve kelime dağarcığı- kelime sayısı eğrilerinin birbirine bağımlı olduklarını vermektedir. İki eğri arasındaki ilişkiyi $B=1/\rho$ eşitliği vermektedir.

Frekans-sıra numarası eğrisi dikkate alındığında $B=1$ olduğu durum düşük frekanslı kelimelerin bulunduğu kısmı ifade eder. $B>1$ kabul ederek, yüksek frekanslı işlev kelimeler grubunu incelersek D3 özelliği ve $B=1/\rho$ sonucu dikkate alınarak şu eşitlik elde edilir:

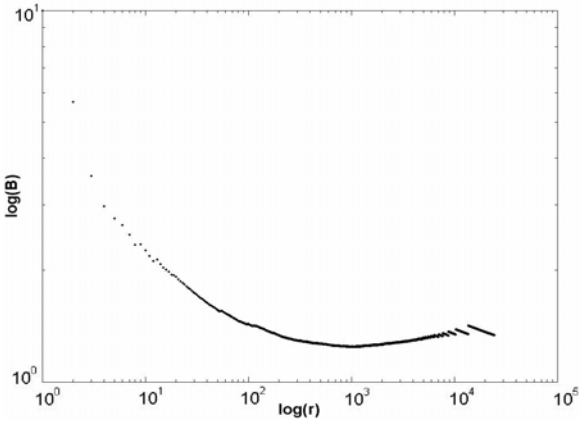
$$\int_{k/N^{\upsilon}}^{1}\frac{1}{N\cdot x^{B}}dx=(1-(k/N^{\rho})^{1-B})/N(1-B) \quad (4.9)$$

Denklem 4.9'da $k=x\cdot N^{\rho}$ alınarak türevi alınırsa $\partial P_k/\partial k=k^{-B}$ elde edilir. İçerik ve işlev kelimelerin ayrımının olduğu sınırda $p_k\approx 1/k^{B}$ alınabilir. k değerinin belirlenmesi esnasında sonuç olarak şu denklem kullanılabilir:

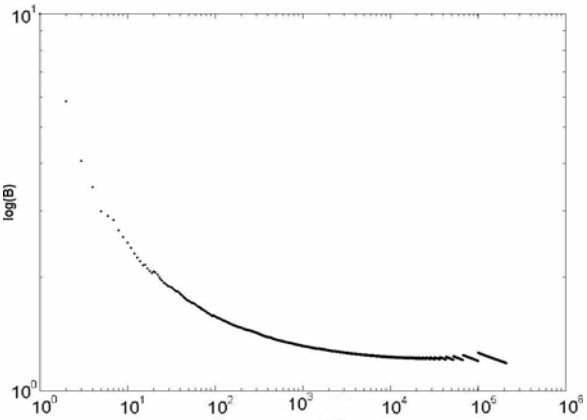
$$B=-\log(p_k)/\log(k) \quad (4.10)$$

Denklem 4.10'da k işlev kelimeler ile içerik kelimeler arasındaki sınır kelime olduğu kabul edilen kelimenin sıra numarasını, p_k ise bu kelimenin *frekans/N* oranını belirtmektedir.

Kornai (2002) çalışmasında Merc derlemi üzerinde k değerinin sınamalarını yapmıştır. Muhtemel sınır kelimeler arasında kesin bir geçiş gözlenmemesine rağmen muhtemel sınır kelimelerin B değerlerinin denklem 4.10'u doğrular özellikte olduğu belirtilmiştir. Tablo 3'de ilgili sonuçlar verilmektedir.



Şekil 4. Bilkent derleminde B ve sıra numarası değerlendirmesi (B and rank values in Bilkent corpus)



Şekil 5. ODTÜ derleminde B ve sıra numarası değerlendirmesi (B and rank values in ODTU corpus)

5. SONUÇLAR (RESULTS)

Çalışma içerisinde işlev ve içerik kelime ayrımını belirlemek için bölüm 3'de önerilen yöntem elimizde bulunan ODTÜ, Makaleler ve Bilkent derlemleri üzerinde değerlendirilmiştir.

Denklem 4.10'a dayanarak ODTÜ ve Bilkent derlemleri için hesaplanan B değerleri şekil 4 ve 5'te görülmektedir.

Şekil 4 ve 5'te r değeri kelimelere ait sıra numaralarını belirtmektedir. Kelime dağarcığının yüksek olması sebebiyle grafiğin daha anlamlı olması için değerlerin logaritmaları alınarak grafik hazırlanmıştır. Şekiller incelendiğinde B değerlerinin yüksek frekanslı kelimelerden düşük frekanslı kelimelere doğru azalarak değiştiği görülmektedir. Tüm değerler listelendiğinde Bilkent, ODTÜ ve makaleler derlemleri için en yüksek frekanslı kelimelerdeki benzerlik beklendiği şekilde görülmektedir. Tablo 4, Tablo 5 ve Tablo 6'da üç derlem için ilk 30 kelimenin göreceli frekans değerleri ve hesaplanan B değerleri gösterilmektedir.

Tablo 4. Bilkent derleminde ilk 30 kelime için frekans ve B değerleri (Frequency and B values of first 30 words in Bilkent corpus)

Kelime	Sıra numarası	B	Göreceli frekans f_N	Kelime	Sıra numarası	B	Göreceli frekans f_N
bir	1	∞	0,0074	**o**	16	2,01	0,0014
ol	2	5,66	0,0072	**çok**	17	1,98	0,0013
ve	3	3,58	0,0071	**ara**	18	1,94	0,0013
bu	4	2,97	0,0059	**var**	19	1,94	0,0012
de	5	2,75	0,0044	**kendi**	20	1,92	0,0012
et	6	2,64	0,0032	**daha**	21	1,89	0,0012
yap	7	2,49	0,0029	**ile**	22	1,87	0,0011
da	8	2,34	0,0028	**konu**	23	1,85	0,0011
için	9	2,34	0,0021	**gibi**	24	1,83	0,0011
ver	10	2,26	0,0020	**söyle**	25	1,81	0,0011
al	11	2,18	0,0020	**sonra**	26	1,79	0,0011
Türkiye	12	2,11	0,0019	**en**	27	1,78	0,0010
çık	13	2,14	0,0015	**iste**	28	1,76	0,0010
yıl	14	2,08	0,0015	**ne**	29	1,75	0,0010
gel	15	2,03	0,0015	**yer**	30	1,74	0,0010

Tablo 5. ODTÜ derleminde ilk 30 kelime için frekans ve B değerleri (Frequency and B values of first 30 words in ODTÜ corpus)

Kelime	Sıra numarası	B	Göreceli frekans f_N	Kelime	Sıra numarası	B	Göreceli frekans f_N
bir	1	∞	0,1536	**değil**	16	2,12	0,0159
ve	2	5,68	0,1088	**sonra**	17	2,07	0,0158
bu	3	3,70	0,0717	**olarak**	18	2,04	0,0154
da	4	3,47	0,0517	**her**	19	2,01	0,0149
de	5	3,00	0,0513	**Türkiye**	20	1,98	0,0127
için	6	2,92	0,0344	**var**	21	1,95	0,0119
çok	7	2,79	0,0252	**ki**	22	1,92	0,0118
ne	8	2,63	0,0248	**şimdi**	23	1,97	0,0118
daha	9	2,51	0,0229	**olan**	24	1,95	0,0118
o	10	2,44	0,0215	**büyük**	25	1,93	0,0116
gibi	11	2,36	0,0204	**son**	26	1,91	0,0115
ile	12	2,28	0,0201	**Ancak**	27	1,89	0,0110
kadar	13	2,27	0,0192	**böyle**	28	1,89	0,0108
ama	14	2,21	0,0187	**Türkiyenin**	29	1,87	0,0106
en	15	2,15	0,0185	**yok**	30	1,86	0,0101

Çizelgelerde görüldüğü üzere ilk 30 kelime içinde işlev kelime olması düşünülemeyecek "Türkiye" gibi özel isimler bulunmaktadır. Bu kelimeler hatalı olarak yakalanan işlev kelimelerdir.

En düşük frekans yani $f_k = 1/N$ değerine sahip kelimeler ve en yüksek frekanslı kelime dışındaki kelimeler için *B* değerlerinin ortalaması alındığında üç derlem için şu sonuçlar şu şekildedir:

- Bilkent derlemi için $B = 1.3254$
- ODTU derlemi için $B = 1.2232$
- Makaleler derlemi için $B = 1.2291$

Bu sonuçlar değerlendirilerek ilgili *B* değerleri *k* sayısının belirlenmesinde kullanılabilir. Bu değerler hesaplandığında

- Bilkent derlemi için ortalama *B* değerini veren k=211
- ODTÜ derlemi için ortalama *B* değerini veren k=10425
- Makaleler derlemi için ortalama *B* değerini veren k=6379

olarak belirlenmiştir.

Tablo 6. Makaleler derleminde ilk 30 kelime için frekans ve B değerleri (Frequency and B values of first 30 words in Makaleler (Articles) corpus)

Kelime	Sıra numarası	B	Göreceli frekans f_N	Kelime	Sıra numarası	B	Göreceli frekans f_N
bir	1	∞	0,1536	**değil**	16	2,12	0,0159
ve	2	5,68	0,1088	**sonra**	17	2,07	0,0158
bu	3	3,70	0,0717	**olarak**	18	2,04	0,0154
da	4	3,47	0,0517	**her**	19	2,01	0,0149
de	5	3,00	0,0513	**Türkiye**	20	1,98	0,0127
için	6	2,92	0,0344	**var**	21	1,95	0,0119
çok	7	2,79	0,0252	**ki**	22	1,92	0,0118
ne	8	2,63	0,0248	**şimdi**	23	1,97	0,0118
daha	9	2,51	0,0229	**olan**	24	1,95	0,0118
o	10	2,44	0,0215	**büyük**	25	1,93	0,0116
gibi	11	2,36	0,0204	**son**	26	1,91	0,0115
ile	12	2,28	0,0201	**Ancak**	27	1,89	0,0110
kadar	13	2,27	0,0192	**böyle**	28	1,89	0,0108
ama	14	2,21	0,0187	**Türkiyenin**	29	1,87	0,0106
en	15	2,15	0,0185	**yok**	30	1,86	0,0101

6. DEĞERLENDİRME (EVALUATION)

ODTÜ ve Makaleler derlemleri kelimelerin kök ve eklerinin üzerinde herhangi bir çalışma yapılmadığı, metinlerin yalın hallerini içeren derlemlerdir. Örneğin "o, onun, ona, onda" gibi aynı gövdeye sahip ancak çekim ekleri ile dilbilgisi kurallarına uygun hale getirilen işlev kelimelerin her biri farklı bir terim olarak değerlendirilmektedir. Bu değerlendirme işleminde hata payının yükselmesine neden olmaktadır. Bu sebeple *k* sayısı için Bilkent derleminden elde edilen sonuç kullanılmış ve değerlendirme yapılmıştır.

Kornai (2002) İngilizce üzerine yaptığı çalışmasında işlev kelimelerin dilden bağımsız olarak bir derlemin ~%49,5'lik kısmını işgal edeceğini göstermiştir. Bu önerme *k*=211 sonucu için değerlendirilirse Bilkent derleminde %44.03, ODTÜ derleminde %26.52, Makaleler derleminde %29.55 yoğunlukla işlev kelimeye rastlandığı gözlenir.

Tüm derlemlerde ilk 211 kelime değerlendirildiğinde sıfat, zamir, edat vb. gibi işlev olması muhtemel kelimelerin dışında özel isimler, güncel olaylara ait bir takım kelimeler, yer adları işlev kelimeler listesine girmektedir. Bunlara örnek olarak "Türkiye, devlet, başkan" verilebilir.

İşlev kelimeler içinde gözlenen içerik kelimelerin tümü incelendiğinde farklı derlemlerde aynı kelimelere rastlamak mümkündür. Bu kelimelerin pek çoğu derlemlerin oluşturulduğu sırada gelişen güncel olaylarla ilgili yer ve kişi isimleri, Türkçe'de çok kullanılan fiillerden oluşmuştur.

İlk 211 kelimenin işlev kelime olduğu kabul edilerek genel anlamda bir değerlendirme yapılırsa, ODTÜ derlemi için ~ %30, Bilkent derlemi için ~ %54, Makaleler derlemi için ~ %32,7 oranla hatalı işlev kelime saptaması yapılır. Bu sonuç şu şekilde yorumlanabilir, örneğin ODTÜ derlemi için işlev kelime olarak kabul edilen ilk 211 kelimeden yaklaşık 63 tanesi aslında içerik kelimedir. (63 /211*100 =%30). Ancak konu, yazar vb. gibi pek çok sebeplerden ötürü derlem içinde yüksek miktarda kullanılmıştır.

Çalışma sonucunda elde edilen değerlerin kesin bir ayrımı simgeleyemeyeceği ancak olası ayrım noktalarındaki *B* değerlerinin İngilizce'de (Kornai, 2002) elde edilen sonuçlarla benzer olduğu görülmüştür. Üç farklı test derlemi için elde edilen sonuçların birbirinden farklı olması derlemlerin dili modellemekte yetersiz olduğu görüşünü desteklemektedir. İleriki çalışmalarda kelime sayısı yüksek ve konu dağılımı dengeli bir derlemde araştırma yinelenecektir.

KAYNAKLAR (REFERENCES)

1. Kornai, A., How many words are there?, **Glottometrics 2002/4**, 61-86p., 2002.
2. Zipf, G. K., **Human Behaviour and the Principles of Least Effort**, Cambridge, MA, Addison-Wesley, 1949.
3. Manning, C.D., Schütze, H., **Foundations of Statistical Natural Language** Processing. The MIT Press, Cambridge, Massachusetts, London, England, 2003.
4. Hakkani-Tür, D.Z., Oflazer, K., and Tür, G., Statistical morphological disambiguation for agglutinative languages, **International Conference On Computational Linguistics**, Proceedings of the 18th conference on Computational Linguistics - Volume 1, Saarbrücken, Germany, 285 - 291 , 2000.

5. Dinçer, T., **Türkçe için İstatistiksel Bir Bilgi Geri-getirim Sistemi**, Doktora Tezi, U.B.E.,Ege Üniversitesi, 2004.
6. Tuldava, J., A Mathematical Model Of The Vocabulary-Text Relation. **COLING 1980**, 600-604, 1980.
7. Alpkoçak, A., Kut, A., ve Özkarahan, E., 1995, Bilgi Bulma Sistemleri için Otomatik Dizinleme Yöntemi, **Bilişim'95**, Dokuz Eylül Üniversitesi, İzmir, 6s., 1995.
8. Holmes-Higgin, P., Abidi S. R., Ahmad, K., A Description of Texts in a Corpus: 'Virtual' and 'Real' Corpora, **EURALEX'94**, 1994.
9. Van Rijsbergen, C.J., **Information Retrieval (2nd ed.)**, Butterworths, London, 1979.
10. Argamon, S., Levitan, S., Measuring the Usefulness of Function Words for Authorship Attribution, **Proceedings of ACH/ALLC Conference 2005 in Victoria**, BC, Canada, 2005.
11. Cleveland, D. B. , Cleveland, A. D., **Introduction to indexing and abstracting**, Libraries Unlimited, Inc., Littleton, Colorado, 1983.
12. Baayen, R. H., **Word Frequency Distributions,** Dordrecht: Kluwer Academic Publishers, 2001.
13. Kumova, S., **Derlem Hazırlama Kriterlerinin Oluşturulması**, Yüksek Lisans Tezi, U.B.E., Ege Üniversitesi, 2005.
14. Powers, D. M. W., Applications and Explanations of Zipf's law, **NEMLAP3/CONLL98, New methods in language processing and Computational natural language learning**, 151-160, 1998.