İstatistik Araştırma Dergisi: Cilt 01 No 01 Sayfa 113-130 Nisan 2002

Basılı Türkçe'nin Önemli Bazı İstatistiksel Özellikleri

Mehmet E. DALKILIÇ* Gökhan DALKILIÇ**

ÖZET

Bu çalışmanın amacı, basılı Türkçe'nin bazı istatistiksel değerlerinin belirlenmesidir. Derlenen istatistikler tekli, ikili, ... , beşli harf gruplarının sıklık dağılımları, ilk/son harf çözümlemeleri, harf başına belirsizlik (entropi) ve fazlalık, rastgelelik endeksi, sözcük uzunluk dağılımı, sesli/sessiz harf oranı'nı içermektedir. Hürriyet gazetesinin internet arşivinden bir Türkçe külliyat (corpus) oluşturularak anılan değerler elde edilmiştir. Bununla yetinilmeyip, Türkçe'ye ilişkin öteki çalışmalar da kullanılarak, tüm bu çalışmaların ağırlıklı bileşkesi olan, bugüne kadar elde edilen en geniş Türkçe külliyat tabanı ve metin çeşitliliğine sahip, en kapsamlı sonuçlar elde edilmiştir. Farklı çalışmalarda elde edilen sonuçların birbiriyle uyumluluk derecesini belirlemek amacıyla bir benzerlik ölçütü geliştirilmiş ve mevcut çalışmaların sonuçlarına uygulanmıştır.

Anahtar Kelimeler: *Türkçe'nin istatistiksel özellikleri, n-gram sıklık dağılımları ,belirsizlik, ilk/son harf çözümlemesi, sözcük uzunlukları, sıralı liste benzerlik ölçütü*

1. GİRİŞ

Doğal Dil İşleme'nin (Natural Language Processing) önemi gün geçtikçe artmaktadır. Doğal Dil İşleme; ses tanıma, yazılan bir metnin yanlışlarının düzeltilmesi, optik karakter okuyucudan geçirilmiş bir metindeki yanlış ve eksik kısımların düzeltilmesi gibi alanlarda kullanılmaktadır. Otomatik metin sınıflandırma, yazarın kimliğini saptama (author identification) türü uygulamaların yanında geniş kullanım alanı olan veri sıkıştırma (data compression) ve veri güvenliği (data security) alanlarında da dillerin istatistiksel özellikleri yoğun olarak kullanılmaktadır. Bu tür çalışmaların yapılabilmesi için dilin bazı özelliklerinin ortaya çıkarılması gerekmektedir.

* Doç.Dr., Ege Üniversitesi Uluslararası Bilgisayar Enstitüsü, 35100 Bornova, İzmir, (Haberleşme adresi)

** Dokuz Eylül Üniversitesi Bilgisayar Mühendisliği Bölümü, 35100 Bornova, İzmir

Doğal Dil İşleme'den sıkıştırma ve veri güvenliğine kadar çok önemli uygulama alanlarında gereksinim duyulan doğal dillerin karakteristik özelliklerinin saptanması konusunda bazı diller için (örneğin İngilizce) yüzlerce araştırma yapılmışken, Türkçe için yapılan araştırmalar hem sayı, hem de kapsam olarak yetersiz kalmıştır. Bu çalışma ile bu konudaki açığın biraz olsun azaltılması hedeflenmiştir.

İnternet ve Türkçe içerikli WEB sitelerinin yaygınlaşması sonucu, elektronik ortamda bulunan Türkçe metinlerin miktarı her geçen gün artmaktadır. Bu da Türkçe ile ilgili yapılan çalışmalara kaynak oluşturması açısından yardımcı olmaktadır. Yapılan bu çalışmada Hürriyet gazetesinin arşivi kullanılarak 1.473.738 karakter içeren bir külliyat yaratılmış (Dalkılıç G., 2001) ve buna *Hürriyet Külliyatı* adı verilmiştir. Bu külliyat www.hürriyet.com.tr adresinden 1 Ocak 1998 – 26 Haziran 1998 tarihleri arasında anasayfa, 1 Ocak 1998 – 30 Haziran 1998 tarihleri arasında köşe yazarlarının sayfaları kullanılarak oluşturulmuştur. Gerek Hürriyet Külliyatı gerekse öteki külliyatlar 29 Türkçe harf ve boşluk dışında karakter içermemektedir.

Elimizde elektronik ortamda işlenebilir veri olarak mevcut diğer külliyatlar Yıldız Teknik Üniversitesi'nde Banu Diri tarafından (Diri, 2000) daha çok farklı yazarlara ait köşe yazıları ve hikayelerden derlenen 4.263.847 karakter uzunluğundaki *YTÜ Külliyatı* ile Ege Üniversitesi'nde Ahmet Koltuksuz tarafından 22 ayrı Türk yazarın toplam 24 eserinden rastgele biçimde derlenmiş 6.095.457 karakterden (Koltuksuz, 1995) oluşan *Koltuksuz Külliyatı*'dır.

Hacettepe Üniversitesi'nde Tuna Göksu ve Levent Ertaul tarafından çoğunluğu günlük gazetelerden ve makalelerden toplanan 547.004 harften (Göksu ve Ertaul, 1998) oluşan metne *Hacettepe Külliyatı* adı verilmiştir. Bu külliyat için yayınlanan sonuçlar boşluk karaktersiz tekli ve ikili dağılımlarından ibarettir. ODTÜ'de Ersin Töreci tarafından hazırlanan toplam 160.014 karakter uzunluğundaki Türkçe metinden oluşan külliyata (Töreci, 1975) *ODTÜ Külliyatı* adı verilmiştir. Bu külliyat için elimizdeki kaynaktaki (Gönenç, 1980) sonuçlar tekli dağılımı ile sınırlıdır.

Yukarıda sayılan bu beş külliyat toplam uzunluk olarak 12,5 milyon karakteri aşmakta, son beş yılda yoğunlaşmak üzere 1975-2000 yılları arasında yapılan çalışmaları ve hikaye, roman, köşe yazısı, haber, makale vb. zengin metin çeşitlerini içermektedir. Bu çalışmada, bu külliyatların ağırlıklı bileşkesini esas alarak Türkçe harf grupları (n-gram) dağılım çözümlemeleri yapılmış, tekli, ikili, üçlü, dörtlü, beşli dağılımları, harf başına belirsizlik ve fazlalık, rastgelelik endeksi, ilk/son ve sesli/sessiz harf oranları ve sözcük uzunluğu dağılımları saptanmıştır. Buna ek olarak, bu beş farklı külliyattan elde edilen sonuçların benzerliği (birbirleri ile uyumu) irdelenmiş, bu amaçla bir "benzerlik ölçütü" geliştirilmiştir. Bu çalışmanın amacı önceki çalışmalardan da yararlanarak basılı Türkçe'nin önemli bazı karakteristik değerlerini olanaklar ölçüsünde gerçeğe en yakın biçimde belirleyerek toplu olarak sunmak ve ilerisi için temel bir kaynak oluşturmaktır.

2. HARF GRUPLARI DAĞILIM ÇÖZÜMLEMELERİ

A adet simgeden oluşan bir dilde toplam A^n farklı n'li harf grubu (n-gram) bulunabilir. Örneğin, Türkçe için A=30 (29 harf ve boşluk karakteri) ve tekli harf grubu (1-gram) sayısı 30^*=30, ikili harf grubu (2-gram) sayısı 30^2=900, vb. şeklinde gider. Doğal dillerde yapı gereği, bu A^n adet n-gram'dan n arttıkça azalan küçük bir yüzdesi kullanılır. Örneğin, İngilizce'de A=27 ve olası 27^4=531.441 dörtlüden[1] yaklaşık %5'i kullanılmakta, beşliler için bu oran %1'in altına inmektedir (Dalkılıç G., 2001). Türkçe için kullanılan n-gram sayısını belirleme çalışması Hürriyet, YTÜ ve Koltuksuz külliyatları için gerçekleştirilmiş ve sonuçlar Tablo 1'de verilmiştir. Üç külliyatın ortalamasına göre Türkçe'de kullanılan n-gram oranları ikililer için %88,4 iken, üçlüler için %33,0'a, dörtlüler için %5,8 ve nihayet beşliler için %0,69'a düşmektedir. Aynı eğilim İngilizce için de saptanmıştır (Dalkılıç G., 2001).

Tablo 1. Türkçe'de Toplam ve Kullanılan n-Gram Oranları

		Hürriyet Külliyatı		YTÜ Külliyatı		Koltuksuz Külliyatı	
	Toplam	Kullanılan	Oran	Kullanılan	Oran	Kullanılan	Oran
Tekli	30	30	%100	30	%100	30	%100
İkili	900	760	%84,44	803	%89,22	824	%91,56
Üçlü	27.000	7615	%28,20	9056	%33,54	10.027	%37,14
Dörtlü	810.000	36.181	%4,47	49.443	%6,10	55.892	%6,90
Beşli	24.300.000	113.270	%0,47	180.745	%0,74	208.745	%0,86

Tablo 1'de "Kullanılan" sütununda verilen değerler her bir örneklemde (külliyatta) görülen farklı n-gram adedini ifade etmekte ve örneklem boyu arttıkça kaydedilen farklı n-gram sayıları da artmaktadır. Bu sonuç doğaldır çünkü sıfıra yakın görülme sıklığına sahip n-gramlar ancak örneklem boyu arttıkça ortaya çıkma şansına kavuşmaktadırlar. Her sonlu külliyatta geçerli bazı n-gramların kayıp olması (görülmemesi) olgusu Doğal Dil İşleme'de "sıfır sıklık problemi" (zero frequency problem) olarak bilinir (Jurafsky ve Martin, 2000). Hürriyet'e kıyasla yaklaşık 2,9 ve 4,1 kat daha büyük olan YTÜ ve Koltuksuz külliyatlarında kullanılan n-gram oranları, beklenildiği gibi, ortalama %30 ve %44 daha yüksektir. İkili'ler üzerinde yapılan bir inceleme toplam 900 olası ikiliden 737'sinin (%81,9) her üç külliyatta, 57'sinin (%6,3) iki külliyatta (10 Hürriyet ve YTÜ, 10 Hürriyet ve Koltuksuz, 37 YTÜ ve Koltuksuz), 62'sinin (%6,9) tek külliyatta (3 Hürriyet, 19 YTÜ, 40 Koltuksuz) kaydedildiği, kalan 44 ikilinin (%4,9) ise külliyatların hiçbirinde görülmediği belirlenmiştir. Bu sonuçlar örneklemlerin kitlenin (Türkçe metinler) ortak özelliklerini büyük ölçüde yansıttıklarına, dolayısı ile kitleyi temsil etme özelliğine sahip olduklarına işaret etmektedir.

---

* Metinde ve tablolarda tekli harf grubu yerine tekli, ikili harf grubu yerine ikili, vb. ifadeleri değişmeli olarak kullanılmıştır.

Tablo 1'de dikkate değer bir diğer nokta ise toplam n-gram değerleri hızla artmakta ve $n \geq 2$ için n-gram sıklık sayımı, n-gram liste çapraz karşılaştırmaları vb. istatistiksel değer belirleme işlemleri ancak külliyat (ham verileri) bilgisayar ortamında mevcut ise mümkün olmaktadır. Dolayısı ile bu ve 4. Bölüm'deki çözümlemeler elektronik ortamda mevcut olan Hürriyet, YTÜ ve Koltuksuz külliyatları ile sınırlı tutulmuştur.

Türkçe metinler için geliştirilen bir bilgisayar programı aracılığı ile Hürriyet, YTÜ ve Koltuksuz külliyatlarında yer alan tekli, ikili, üçlü, dörtlü ve beşlilerin görülme sıklıkları ve yüzdeleri belirlenmiştir. Belirlenen bu n-gram dağılımları sıklıklarına göre sıralandıktan sonra Tablo 2'de görülen değerler elde edilmiştir. Bu tablo'da ilk 3, ilk 10, ilk 25, ilk 50, ilk 100, ilk 500 ve ilk 1000 değerin, toplamın yüzde kaçını oluşturduğu gösterilmektedir. Artan n değerleri için ilk-k n-gram toplamı hızla azalmaktadır. Örneğin, her üç külliyat için de, ilk-10 teklinin toplamı yaklaşık %69 iken bu oran ikililer için %16'ya, üçlüler için %5'e vb. düşmektedir. Yine de ilk-100 n-gram'ın ortalama görülme sıklığı tüm n-gramların eşit görülme sıklığında olması durumuna kıyasla, üç külliyatın ortalaması esas alındığında, ikililer için 0,681 / (100/900) = 6,1 kat, üçlüler için 0,254 / (100/27000) = 68,6 kat, dörtlüler için 939 kat ve beşliler için 14.779 kat daha çoktur.

Tablo 2. Türkçe İlk {3, 10, ..., 1000} n-Gram Sıklık(%) Toplamları

Hürriyet Külliyatı					
	Tekli	**İkili**	**Üçlü**	**Dörtlü**	**Beşli**
İlk 3	32,346	5,615	1,844	0,92	0,537
İlk 10	69,098	16,079	5,042	2,426	1,251
İlk 25	97,21	31,05	10,012	4,647	2,23
İlk 50	100	47,428	16,109	7,32	3,602
İlk 100	100	68,104	24,593	11,125	5,671
İlk 500	100	99,534	55,545	27,685	15,791
İlk 1000	100	100	75,416	39,659	23,276
YTÜ Külliyatı					
İlk 3	31,833	5,61	2,067	1,322	0,776
İlk 10	69,269	16,319	5,57	2,833	1,523
İlk 25	97,355	31,969	11,023	5,181	2,582
İlk 50	100	47,925	16,883	7,907	4,043
İlk 100	100	67,932	25,472	11,769	6,169
İlk 500	100	99,593	58,563	28,731	15,928
İlk 1000	100	100	76,18	40,342	22,953
Koltuksuz Külliyatı					
İlk 3	31,901	5,846	2,043	1,34	0,75
İlk 10	68,801	16,335	5,656	2,752	1,541
İlk 25	97,354	32,207	11,083	4,986	2,695
İlk 50	100	48,295	17,257	7,863	4,171
İlk 100	100	68,286	26,187	11,884	6,408
İlk 500	100	99,726	60,142	29,589	16,319
İlk 1000	100	100	77,762	41,573	23,375

Kıyaslama amacı ile "Penn State Üniversitesi Elektronik Klasikler Serisi Sitesi"nden (www2.hn.psu.edu/faculty/jmanis/jimspdf.htm) 1.757.413 karakterlik bir

İngilizce külliyat yaratılmış ve n-gram dağılımları belirlenerek Tablo 3 oluşturulmuştur. İngilizce ilk-k n-gram toplamı değerlerinin her zaman Türkçe eşdeğerlerinden yüksek olduğu görülmektedir. Bu da İngilizce dilinde sık kullanılan n-gram'ların ortalama kullanım yüzdesinin Türkçe eşdeğerlerinden yüksek olduğunu gösterir. İngilizce için de artan n değerleri için ilk-k n-gram toplamı hızla azalmaktadır. Örneğin, ilk-10 teklinin toplamı %74,8 iken bu oran ikililer için %23,6'ya, üçlüler için %9,9'a vb. düşmektedir. Yine de ilk-100 n-gram'ın ortalama görülme sıklığı tüm n-gramların eşit görülme sıklığında olması durumuna kıyasla, Tablo 3'deki değerler kullanılarak, ikililer için 0,773 / (100/729) = 5,6 kat, üçlüler için 0,373 / (100/19683) = 73,4 kat, dörtlüler için 1057 kat ve beşliler için 16.644 kat daha çoktur. Bu değerlere göre İngilizce'de sık kullanılan ilk-100 n-gramların ortalama kullanım yüzdeleri Türkçe eşdeğerlerinden, ortalama %4,8 (n=2,3,4,5 için sırasıyla -%8,2, %6,5, %10,6 ve %11,0) çoktur.

Tablo 3. İngilizce İlk {3, 10, ..., 1000} n-Gram Sıklık(%) Toplamları

	Tekli	İkili	Üçlü	Dörtlü	Beşli
İlk 3	37,086	9,071	4,119	2,905	2,001
İlk 10	74,745	23,563	9,907	6,244	3,435
İlk 25	99,918	42,116	17,387	9,957	5,481
İlk 50	100	58,561	25,978	13,96	7,968
İlk 100	100	77,27	37,256	19,901	11,629
İlk 500	100	99,996	72,162	42,04	26,303
İlk 1000	100	100	86,999	55,411	35,652

3. TÜRKÇE TEKLİ DAĞILIMLARI

Tekli dağılımı ham veriden kolaylıkla elde edilebilir ve dile ilişkin önemli istatistiksel değerlerin (örneğin, tekli belirsizliği ve fazlalığı, rastgelelik endeksi ve sesli/sessiz harf oranları) hesaplanmasına olanak sağlar. Bu bölümde tekli yüzdelerinin beş külliyata göre ağırlıklı ortalaması bulunmuş ve Tablo 4'de Türkçe simgeler ağırlıklı ortalamadaki değerlerine göre sıralı olarak verilmiştir.

Elimizdeki beş farklı külliyatın her birinin ayrı sütunda yer aldığı Tablo 4[1]'de ağırlıklı ortalamalar son sütunda verilmiştir. İlgili sütunlarda külliyatların ham veri büyüklükleri verilmiş ve ağırlıklı ortalamalar bu değerlere göre hesaplanmıştır. Elde edilen ağırlıklı ortalamalara göre, tahmin edilebileceği gibi, Türkçe metinlerde en çok görülen simge %13,3 ile boşluktur. **AEİNR** harfleri yüksek sıklık, **LIKD** harfleri üst-orta sıklık, **MYTUSBO** harfleri alt-orta sıklık, **ÜŞZGÇ** harfleri üst-düşük sıklık, **HĞVCPÖF** harfleri alt-düşük sıklık ve **J** harfi ise tek başına aşırı-düşük sıklık grubunu oluşturmaktadır. Şekil 3.1'de harfler ve görülme oranları grafiksel olarak gösterilmiştir.

3.1. Benzerlik Çözümlemesi

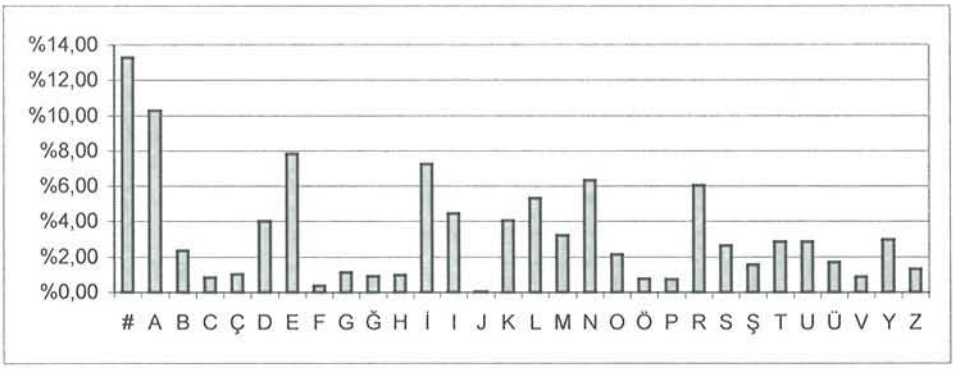
Tablo 4'deki tekli yüzdeleri yerine harflerin sıralamadaki yerleri esas alınarak Tablo 5 yaratılmıştır. Böylece Tablo 5'de her sütun farklı bir külliyata göre Türkçe

[1] Hacettepe külliyatı için referans değerler boşluk karakteri olmadan verildiğinden, Tablo 3.1'deki Hacettepe sütunu değerleri boşluk karakteri hesaba katılarak yeniden belirlenmiştir.

harflerin bir sıralamasını göstermektedir. Bu sıralamaların birbirleriyle oldukça benzer olduğu kısa bir inceleme ile anlaşılabilir. Dolayısı ile farklı külliyatlar ve farklı zaman dilimlerinde yapılan çalışmalarda elde edilen Türkçe tekli dağılım ölçümleri benzer sonuçlar vermiştir denilebilir. Fakat "ne kadar benzer?" sorusunun yanıtını verebilmek için sıralı listelerin benzerliğinin bir ölçüte bağlanması gerekir. Aynı küme öğelerinden oluşan ve farklı sayıda öğe içerebilen iki liste (L_1, L_2) için bir benzerlik katsayısı eşitliliği $B(L_1,L_2)$'yi geliştirdik. Bu eşitlikde benzerliklerini ölçmek istediğimiz listelerdeki sıralama farklarının toplamı ($\Sigma\Delta_\alpha$) 'nın aynı büyüklükte iki rastgele sıralı dizinin sıralama farklarının toplamına (μ) oranı ölçüt alınmıştır. Benzerlik ölçütünün örneklemin tüm öğeleri üzerinden yapılamadığı ya da yapılmak istenmediği durumlarda

Tablo 4. Külliyatlara Göre Türkçe Tekli Yüzdeleri

Tekli (Ağırlıklı Ortalamaya Göre)	**Koltuksuz** (6.095.457 karakter)	**YTÜ** (4.263.847 karakter)	**Hürriyet** (1.473.738 karakter)	**Hacettepe** (547.004 karakter)	**ODTÜ** (160.014 karakter)	**Ağırlıklı Ortalama**
# (Boşluk)	12,69	13,87	14,02	13,29	13,86	13,29
A	10,20	10,22	10,63	10,36	10,03	10,26
E	7,86	7,75	7,70	8,33	7,22	7,82
İ	7,22	7,29	7,12	7,43	6,71	7,23
N	6,31	6,34	6,28	6,81	5,88	6,33
R	6,07	6,07	5,95	5,75	6,38	6,04
L	5,02	5,53	5,57	6,05	5,06	5,30
I	4,54	4,35	4,28	4,21	4,89	4,44
K	4,11	4,06	3,90	4,00	4,17	4,07
D	4,25	3,80	3,65	3,69	4,12	4,00
M	3,27	3,18	3,09	2,89	3,34	3,20
Y	2,94	3,00	2,98	2,60	3,05	2,95
T	2,70	2,90	3,25	3,47	2,96	2,87
U	3,00	2,71	2,67	2,63	3,01	2,84
S	2,57	2,67	2,70	2,94	2,36	2,64
B	2,58	2,21	2,09	1,96	2,33	2,37
O	2,14	2,16	2,16	1,94	2,29	2,14
Ü	1,73	1,70	1,64	1,72	1,94	1,71
Ş	1,69	1,49	1,38	1,30	1,72	1,57
Z	1,31	1,32	1,28	1,02	1,20	1,30
G	1,17	1,12	1,09	1,08	1,17	1,14
Ç	1,10	0,97	0,90	0,82	1,15	1,02
H	0,99	0,93	0,95	0,90	0,79	0,96
Ğ	0,98	0,89	0,76	0,94	0,98	0,92
V	0,85	0,84	1,01	0,99	0,69	0,87
C	0,85	0,80	0,93	0,82	0,78	0,84
P	0,69	0,73	0,86	0,81	0,79	0,73
Ö	0,76	0,73	0,70	0,74	0,77	0,74
F	0,38	0,34	0,44	0,46	0,34	0,38
J	0,01	0,04	0,04	0,08	0,02	0,03



Şekil 1. Türkçe teklilerin beş külliyatın ağırlıklı ortalamasına göre görülme olasılıkları

da eşitliğin geçerli olması için $\Sigma\Delta_\alpha$ ölçüm yapılan öğe sayısı M ($M \leq N$:örneklem kümesindeki toplam öğe sayısı) ve μ ise N ile bölünerek normalize edilmiştir:

$$B(L_1,L_2) = 1 - \frac{\sum_{\alpha=1}^{M}\Delta_\alpha / M}{\mu/N} = 1 - \frac{N\sum_{\alpha=1}^{M}\Delta_\alpha}{M\mu} \qquad (1)$$

N : Örneklem kümesindeki toplam öğe (n-gram) sayısı
M : Karşılaştırmada kullanılmak üzere rastgele seçilen öğelerin (n-gram) sayısı
Δ_α : Öğe α'nın L_1 ve L_2 listelerindeki sıralama (yerleşim) farkı ($\Delta_\alpha \geq 0$)
μ : N uzunluğunda iki rastgele dizi için $\Sigma\Delta_\alpha$ değeri ($\sum_{i=0}^{N-1} i \cong N^2/2$)

Eşitlik (1) ile elde edilebilecek değerler sıfır ile bir aralığında değişir. İki dizi arasındaki benzerlik azaldıkça bulunan sonuç sıfıra benzerlik arttıkça bir'e yaklaşır. Birbirinin aynı olan iki dizi için benzerlik katsayısı bir'e eşittir.

Tablo 5'deki beş farklı külliyattan elde edilen Türkçe tekli sıralamalarının, ağırlıklı ortalama sıralaması ile benzerlik katsayılarının hesaplanarak Tablo 6'ün ilgili satırının oluşturulmasında Eşitlik (1)'de N=M=30, L_1=Tablo 5'deki Ağırlıklı Ortalama ve

Tablo 5. Beş Külliyata Göre Türkçe Teklilerin Sıralaması

Tekli (Ağırlıklı Ortalamaya Göre)	**Ağırlıklı Ortalama**	**Koltuksuz** (6.095.457 karakter)	**YTÜ** (4.263.847 karakter)	**Hürriyet** (1.473.738 karakter)	**Hacettepe** (547.004 karakter)	**ODTÜ** (160.014 karakter)
#	1	1	1	1	1	1
A	2	2	2	2	2	2
E	3	3	3	3	3	3
İ	4	4	4	4	4	4
N	5	5	5	5	5	6
R	6	6	6	6	7	5
L	7	7	7	7	6	7
I	8	8	8	8	8	8
K	9	10	9	9	9	9
D	10	9	10	10	10	10
M	11	11	11	12	13	11
Y	12	13	12	13	15	12
T	13	14	13	11	11	14
U	14	12	14	15	14	13
S	15	16	15	14	12	15
B	16	15	16	17	16	16
O	17	17	17	16	17	17
Ü	18	18	18	18	18	18
Ş	19	19	19	19	19	19
Z	20	20	20	20	21	20
G	21	21	21	21	20	21
Ç	22	22	22	25	26	22
H	23	23	23	23	24	24
Ğ	24	24	24	27	23	23
V	25	25	25	22	22	28
C	26	26	26	24	25	26
P	27	28	27	26	27	25
Ö	28	27	28	28	28	27
F	29	29	29	29	29	29
J	30	30	30	30	30	30

$$\mu = \sum_{i=0}^{29} i = \frac{30 \cdot 29}{2} = 435$$

değerleri kullanılmıştır. Örneğin, Tablo 6'teki Koltuksuz sütunundaki benzerlik katsayısı, L_2=Tablo 5'deki Koltuksuz sütunu ve $\Sigma\Delta_\alpha$=10 olmak üzere,

$$B(L_1, L_2) = 1 - \frac{30}{30} \cdot \frac{10}{435} = 0{,}977011$$

olarak hesaplanmıştır.

Bu bölümde karşılaştırılan listeler aynı uzunlukta ve örneklem kümesindeki tüm öğeleri içerdiğinden ($|L_1| = |L_2| = N$) *Spearman rank korelasyon katsayısı* (Siegel, 1956) olan,

$$r_S = 1 - \frac{6\sum_{i=1}^{N} (\Delta_\alpha)^2}{N^3 - N} \qquad (2)$$

değerini de bir tür benzerlik ölçütü olarak kullanabiliriz. Örneğin, Tablo 6'te Koltuksuz sütununda yeralan r_s değeri, $\Sigma(\Delta_\alpha)^2=12$ ve N=30 için 0,99733 olarak hesaplanmıştır.

Tablo 6'te görüldüğü gibi, Ağırlıklı Ortalama sıralaması ile en yüksek benzerlik katsayısına YTÜ Külliyatı'nda 1, en düşük benzerlik katsayısına ise 0,94 ile Hacettepe Külliyatı'nda rastlanmaktadır. Bu sonuçlar Spearman rank korelasyon sonuçları ile uyum içindedir.

Tablo 6. Külliyatların Tekli Ağırlıklı Ortalama ile Benzerlik ve Spearman r_s Katsayıları

	Koltuksuz	YTÜ	Hürriyet	Hacettepe	ODTÜ
Benzerlik Katsayısı	0,977011	1	0,954023	0,944828	0,972414
Spearman r_s	0,997330	1	0,991323	0,987096	0,995555

4. İKİLİ, ÜÇLÜ, DÖRTLÜ ve BEŞLİ DAĞILIMLARI

Bu bölümde n-gram ($2 \le n \le 5$) dağılımları Hürriyet, YTÜ ve Koltuksuz külliyatları üzerinde belirlenmiş ve Tablo 7'de kısmen (Hürriyet Külliyatı'nın ilk 50'si ve bu n-gramların YTÜ ve Koltuksuz külliyatlarındaki sonuçlarının sıraları) listelenmiştir.

Tablo 7. Hürriyet, YTÜ ve Koltuksuz Külliyatlarının Kısmi n-Gram Karşılaştırması

İkili	Hür.	YTÜ	Kol.	Üçlü	Hür.	YTÜ	Kol.	Dörtlü	Hür.	YTÜ	Kol.	Beşli	Hür	YTÜ	Kol.
N#	1	1	1	LAR	1	1	2	LARI	1	3	3	#BİR#	1	1	1
E#	2	2	3	AN#	2	6	8	#BİR	2	1	1	LARIN	2	2	2
AR	3	4	7	İN#	3	8	13	#VE#	3	7	13	LARI#	3	4	6
A#	4	5	5	LER	4	3	6	BİR#	4	2	2	INDA#	4	6	10
İ#	5	7	6	#YA	5	10	10	LERİ	5	4	4	LERİN	5	3	3
AN	6	9	10	#Bİ	6	2	1	INDA	6	9	11	LERİ#	6	5	9
#B	7	3	2	#KA	7	7	5	YOR#	7	10	32	BAKAN	7	153	6792
LA	8	8	8	ARI	8	11	12	LAR#	8	5	5	#TÜRK	8	8	746
R#	9	6	4	DA#	9	14	17	NDA#	9	16	25	İNDE#	9	10	13
ER	10	10	11	ERİ	10	12	11	ARIN	10	6	9	IYOR#	10	29	81
I#	11	15	14	#VE	11	26	38	#BU#	11	12	10	#İÇİN	11	7	7
LE	12	12	15	EN#	12	13	7	#BAŞ	12	28	30	NDAN#	12	11	11
İN	13	13	13	#BA	13	19	15	ARI#	13	15	20	İYOR#	13	14	65
#K	14	14	9	ARA	14	23	27	DAN#	14	13	7	BAŞKA	14	69	175
#D	15	11	12	#DE	15	9	9	ERİN	15	8	8	İÇİN#	15	19	26
#A	16	18	18	AR#	16	16	16	LER#	16	11	12	NLARI	16	20	56
DE	17	16	17	BİR	17	5	3	NLAR	17	23	38	#BAŞK	17	96	179
#Y	18	21	25	IN#	18	17	18	ERİ#	18	21	22	N#BİR	18	9	8
IN	19	17	19	İR#	19	4	4	NİN#	19	20	43	İNİN#	19	13	37
DA	20	19	24	#BU	20	20	20	IYOR	20	18	14	AŞKAN	20	424	15390
KA	21	29	27	DE#	21	15	19	DEN#	21	14	6	ARINI	21	15	29
YA	22	26	29	YOR	22	18	14	INDE	22	22	16	INDAN	22	35	35
#S	23	25	22	NDA	23	24	36	INI#	23	26	17	#KONU	23	44	142
K#	24	24	26	VE#	24	46	79	#YAP	24	43	96	KANI#	24	434	7679
EN	25	20	16	ER#	25	27	21	#VER	25	86	100	#OLDU	25	43	36
#İ	26	30	37	#OL	26	25	32	IYOR	26	29	29	#KARA	26	127	174
RA	27	34	35	#GE	27	22	22	İNE#	27	24	21	KİYE#	27	27	8095
İL	28	33	44	İLE	28	37	59	#KAR	28	41	36	RKİYE	28	31	12227
#G	29	27	23	N#B	29	28	23	NDE#	29	31	37	TÜRKİ	29	28	12038
MA	30	28	32	#SA	30	33	43	#DA#	30	19	40	ÜRKİY	30	30	12358
ND	31	31	33	#DA	31	21	33	#OLA	31	34	94	ALARI	31	36	82
AL	32	32	34	#HA	32	38	28	İYE#	32	30	74	ARIN#	32	21	48
İR	33	22	20	NDE	33	45	45	#GÜN	33	105	121	ARAK#	33	49	38
AK	34	35	30	OR#	34	62	137	İNİ#	34	25	15	#İLE#	34	138	200
#T	35	40	63	LE#	35	32	29	NIN#	35	27	27	ANLAR	35	26	67
Rİ	36	36	39	LAN	36	54	123	ESİ#	36	87	228	KLARI	36	23	24
Dİ	37	37	28	NI#	37	51	69	TÜRK	37	44	1490	ACAK#	37	60	133
Lİ	38	39	48	ANI	38	53	50	#TÜR	38	33	771	#BAKA	38	270	535
ET	39	64	92	ASI	39	35	52	ASI#	39	60	158	SONRA	39	55	23
Bİ	40	23	21	AK#	40	36	39	#YIL	40	114	769	#SONR	40	56	21
OR	41	43	42	#TA	41	58	83	ARA#	41	59	123	SINDA	41	46	54
U#	42	42	31	IND	42	49	68	#DE#	42	17	35	#MİLY	42	482	5903
#M	43	66	77	INI	43	31	37	ASIN	43	38	45	ERİNİ	43	22	33
BA	44	58	45	KAN	44	186	421	İĞI#	44	93	219	#NİN#	44	72	2158
ME	45	41	43	Dİ#	45	66	26	BAKA	45	281	1160	LMAZ#	45	601	874
#O	46	38	36	E#B	46	34	30	INA#	46	35	26	NLAR#	46	58	113
TA	47	53	59	İNİ	47	30	35	ALAR	47	40	66	ININ#	47	17	19
AY	48	46	57	ESİ	48	71	120	İLER	48	39	73	#YAPI	48	121	447
#V	49	68	78	NLA	49	48	64	AKAN	49	407	4073	#GÜND	49	1453	2052
EL	50	48	52	NE#	50	29	24	#MİL	50	248	2336	İLERİ	50	47	176

4.1. Benzerlik Çözümlemesi

Hürriyet, YTÜ ve Koltuksuz külliyatlarından elde ettiğimiz ve kısmi olarak Tablo 8'de verilen n-gram ($2 \leq n \leq 5$) listeleri için benzerlik katsayılarının hesaplanmasında aşağıdaki yöntem izlenmiştir. Üçüncü bölümde benzerlik çözümlemesini yaptığımız tekli listelerinden farklı olarak buradaki n-gram ($2 \leq n \leq 5$) listeleri için liste boyları ne birbirine ne de N'ye eşittir[2]. Örneğin, Hürriyet Külliyatı'ndan elde edilen ikili listesi olası toplam 900 ikiliden 760 öğe içerirken, YTÜ Külliyatı'ndan elde edilen liste 803 öğe içermektedir. 747 ikili her iki külliyat listesinde ve 69 ikili sadece bir listede yeralırken, 84 ikili her iki listede de yoktur. Dolayısıyla benzerlik katsayısı eşitlikteki (1) Δ_α sıralama farkı değeri için üç ayrı durum mevcuttur. Aranan öğe;

i) her iki listede de var,
ii) listelerin sadece birinde var ve
iii) her iki listede de yok.

İlk durum için Δ_α, sıralamalar arasındaki farkın mutlak değeri ve ikinci durum için $\Delta_\alpha=\mu/N$ (N uzunluğunda iki rastgele sıralı listede herhangi bir öğe için beklenen sıralama farkı) dir. Üçüncü durum için ise $\Delta_\alpha=\kappa(\mu/N)$; yani, μ/N'nin belli bir katsayı ile zayıflatılmış hali kullanılmıştır. Zayıflatma katsayısı (κ) öyle seçilmelidir ki,

i) listelerin birbirine tümüyle eşit ($L_1 = L_2$) fakat tüm öğeleri içermiyor ($|L_1| = |L_2| < N$) olması durumunda $B(L_1,L_2) = 1$ ve

ii) listelerin hiçbir ortak öğeleri olmaması ($L_1 \cap L_2 = \varnothing$) durumunda $B(L_1,L_2) = 0$ değerleri sağlansın:

$$\kappa = 1 - \frac{|L_1 \cap L_2|}{|L_1 \cup L_2|} \qquad (3)$$

Eşitlik (3) ile verilen (κ), listelerdeki ortak öğe sayısının listelerin birleşimindeki öğe sayısına oranınının bir'den farkını ölçü alır. Ayrıca,

i) $L_1 = L_2$, $|L_1| = |L_2| < N$ olması halinde $\kappa = 0$ ve her öğe için $\Delta_\alpha=0$ olduğundan $B(L_1,L_2) = 1$ ve
ii) $L_1 \cap L_2 = \varnothing$ durumunda $\kappa = 1$ ve her öğe için $\Delta_\alpha=\mu/N$ olduğundan $B(L_1,L_2) = 0$ koşulları sağlanır.

Tablo 8'deki değerlerin hesaplanmasında Eşitlik (1)'de M=N ve yukarıda açıklanan Δ_α kullanılmıştır. Külliyatlar arasındaki benzerlik katsayısının n (harf grubu boyu) büyüdükçe azaldığı gözlenmektedir. Bu olgu şöyle açıklanabilir: Benzerlik katsayısı $n>2$ için $1-\kappa$ (iki listedeki ortak öğe sayısının iki listenin birleşimindeki öğe

[2] Listeler aynı uzunlukta olmadığından ve örneklem kümesindeki tüm öğeleri içermediğinden Spearman rank korelasyon katsayısını, r_s, kullanamayız.

sayısına oranı) ile kuvvetle ilintilidir ve ortak n-gramlar ancak bir önceki listedeki ortak n-1 gramlardan belli bir kayıpla üretilebilmektedir. Örneğin, ortak üçlüler ancak ortak ikililerden, ortak dörtlüler ancak ortak üçlülerden (ve de kayıplı olarak) üremekte ve dolayısı ile n arttıkça ortak n-gram'ların sayısı ve dolayısı ile 1-κ katsayı azalmaktadır.

Tablo 8. Külliyatların n-Gram Benzerlik Katsayıları

Benzerlik Katsayısı	Tekli	İkili	Üçlü	Dörtlü	Beşli
YTÜ-Koltuksuz	0,98	0,84	0,70	0,60	0,52
Hürriyet-YTÜ	0,95	0,86	0,74	0,61	0,49
Hürriyet-Koltuksuz	0,94	0,80	0,63	0,5	0,39
PennState-Hürriyet	0,49	0,41	0,31	0,16	0,06

Karşılaştırılan listeler makalenin 2. Bölümü'nde açıklandığı gibi n>2 için seyrek ($|L_1| << N$ ve $|L_2| << N$) olduğundan "her iki listede de olmama" durumu Eşitlik (1)'de baskın konuma geçmekte ve bunun sonucu olarak benzerlik katsayısı (1-κ) değeri tarafından belirlenir hale gelmektedir. Somut bir örnek olarak Hürriyet-YTÜ karşılaştırmasını ele alalım. İkili karşılaştırması için iki listedeki ortak öğe sayısı 747, listelerin bileşimindeki öğe sayısı 816, $\kappa = 0,08$ ve $1-\kappa = 0,92$ (benzerlik katsayısı 0,86) iken üçlülere gelindiğinde bu 747 ortak ikiliden üretilen üçlülerin hepsi ortak olamayacağından (kayıp olgusu), ortak üçlülerin üretilen toplam üçlülere oranı $1-\kappa = 7147 / 9534 = 0,75$'e (benzerlik katsayısı 0,74) düşmektedir. Bir sonraki Hürriyet ve YTÜ 4-gram listelerinde ise bu 7147 ortak üçlülerden üretilen dörtlülerin sadece $1-\kappa = 32453 / 53178 = 0,61$'i (benzerlik katsayısı 0,61) ortaktır. Beşlilere gelindiğinde bu oran $1-\kappa = 0,49$'a (benzerlik katsayısı 0,49) kadar düşmektedir.

Tablo 8'de YTÜ-Koltuksuz külliyatlarının en yüksek n-gram benzerlik katsayılarına sahip olduğu, Hürriyet-Koltuksuz külliyat karşılaştırmasının ise Türkçe-Türkçe n-gram liste karşılaştırmaları içinde en düşük benzerlik değerlerine sahip olduğu görülmektedir. Elde edilen bu Türkçe-Türkçe külliyat benzerlik katsayılarının değerlendirilmesinde bir İngilizce-Türkçe külliyat karşılaştırmasının sonuçlarının referans olarak yararlı olacağı düşüncesi ile makalenin 2. bölümünde bahsedilen PennState Külliyatı kendisine en yakın büyüklükteki Türkçe külliyat olan Hürrriyet Külliyatı ile karşılaştırılmış ve elde edilen benzerlik katsayıları Tablo 8'nin son satırında verilmiştir. Bu sonuçlara bakıldığında Türkçe Hürriyet Külliyatı ve İngilizce PennState Külliyatı için tekli (0,49), ikili (0,41) ve hatta üçlü (0,31) listeleri arasında kayda değer benzerlik olduğu, fakat dörtlü (0,16) ve beşli (0,06) listeleri için benzerlik katsayılarının hızla düştüğü görülmektedir. Hürriyet Külliyatı'nın diğer iki Türkçe külliyat ile olan benzerlik katsayılarının ortalamasını PennState Külliyatı ile olan benzerlik katsayıları ile karşılaştırırsak; tekliler için 0,945/0,49 =1,93, ikililer için 0,83/0,41=2,02, üçlüler için 0,685/0,31=2,21, dörtlüler için 0,595/0,16=3,71 ve beşliler için 0,44/0,06=7,33 kat Türkçe-Türkçe lehine bir fark görürüz. Bu eğilim n>5 için olan listelerde bu oranın daha da büyüyeceğine işaret etmektedir.

5. BELİRSİZLİK, FAZLALIK ve RASTGELELİK ENDEKSİ

Belirsizlik (H), ortalama olarak bir metindeki her bir karakterin taşıdığı bilgi miktarıdır. Bir kesikli (n farklı değer alabilen) rastgele değişken (*discrete random variable*) **X** için, **X**'in belirsizliği,

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i = \sum_{i=1}^{n} p_i \log_2 (\frac{1}{p_i}) \qquad (4)$$

biçiminde tanımlanır. Burada p_i, i=1,...,n, X'in *i* değerini alması olasılığıdır (Shannon, 1951). Tüm olasılıkların eşit olduğu bir dağılımda belirsizlik en yüksek değeri olan $\log_2 n$ değerine ulaşır. Öte yandan *fazlalık* (R) dilin yapısı tarafından metne yüklenen kısıtlamaların ölçüsüdür:

$$R(X) = \text{En Büyük Belirsizlik} - H(X) = \log_2 30 - H(X) \qquad (5)$$

Dil ne kadar çok kısıt altında ise o kadar düzenli ve dolayısı ile o kadar düşük belirsizlik değerine sahip olur. Rastgele değişken ***X*** için belirsizlik ile fazlalığın toplamı sabit ve en büyük belirsizlik değerine eşittir:

$$H(X) + R(X) = \log_2 n \qquad (6)$$

Bunun anlamı belirsizlik arttıkça fazlalığın azalacağı, belirsizlik azaldıkça fazlalığın artacağıdır. Bir metinde fazlalık ne kadar çok (belirsizlik ne kadar az) ise, o metnin sıkıştırılabilme potansiyeli o kadar yüksek olur.

Tablo 9'de Hürriyet, YTÜ ve Koltuksuz külliyatlarından elde edilen n-gram ($1 \le n \le 5$) belirsizlikleri ve bunların ağırlıklı bileşkesinden elde edilen n-gram belirsizlik değerleri verilmiştir. Bu tablodan, her üç külliyattaki değerlerin birbiriyle oldukça uyumlu olduğu ve beklendiği gibi ***n*** arttıkça simge başına düşen bit cinsinden (bit/simge) belirsizliğin azaldığı gözlenmektedir. Çok büyük ***n*** değerleri için n-gram belirsizliğinin dilin belirsizliğine yakınsayacağı açıktır (Stinson, 1995). Pratikte 5'ten büyük ***n*** değerleri için n-gram dağılımlarını elde etmek çok güç olduğundan, doğal dillerin belirsizlikleri farklı yöntemlerle belirlenmektedir (Shannon, 1951; Cover ve King, 1978). Bu tür bir yöntem yakın geçmişte Türkçe için uygulanmış ve belirsizlik 1,47 bit/simge olarak bulunmuştur (Dalkılıç M.E. ve Dalkılıç G, 2000). Dolayısı ile Türkçe'nin fazlalığı 3,43 bit/simge olup, bu Türkçe metinlerin yaklaşık %70 oranında sıkıştırılabileceğini ifade etmektedir. Türkçe n-gram ($1 \le n \le 5$) fazlalık değerleri Tablo 10'de verilmiştir.

Tablo 9. Türkçe Külliyatların Belirsizlik Değerleri

	Belirsizlik Değeri (bit/simge)			
	Hürriyet	YTÜ	Koltuksuz	Ağırlıklı Ortalama
Tekli	4,35	4,35	4,38	4,37
İkili	3,95	3,93	3,96	3,95
Üçlü	3,6	3,59	3,62	3,61
Dörtlü	3,27	3,27	3,33	3,30
Beşli	2,97	3	2,97	2,98

Tablo 10. Türkçe Külliyatların Fazlalık Değerleri

	Fazlalık (bit/simge)			
	Hürriyet	YTÜ	Koltuksuz	Ağırlıklı Ortalama
Tekli	0,56	0,56	0,52	0,54
İkili	0,96	0,98	0,95	0,96
Üçlü	1,31	1,32	1,28	1,30
Dörtlü	1,64	1,64	1,58	1,61
Beşli	1,94	1,91	1,93	1,92

X'in rastgelelik endeksi ($I_c(X)$), X'den seçilen rastgele iki öğenin aynı olma olasılığıdır ve klasik şifreleyicilerin gizlilik çözümlemesinde yaygın kullanıma sahiptir. Rastgelelik endeksi aşağıda verilmiştir (Stinson, 1995):

$$I_c(x) \approx \sum_{i=1}^{n} p_i^2 \qquad (7)$$

Tablo 4'deki tekli dağılımları kullanılarak tüm külliyatlar için rastgelelik endeksleri hesaplanmış ve ağırlıklı ortalamalar ile birlikte Tablo 11'te gösterilmiştir.

Tablo 11.Türkçe Külliyatların Rastgelelik Endeksleri

	Koltuksuz	Hürriyet	YTÜ	Hacettepe	ODTÜ	Ağırlıklı Ort.
Rastgelelik Endeksi	0,0604	0,0637	0,0632	0,0634	0,0618	0,0624

6. SÖZCÜK BOYU, İLK/SON VE SESLİ/SESSİZ HARF ÇÖZÜMLEMELERİ

Ortalama sözcük uzunluğu ve sözcük uzunluğu dağılımları doğal diller arasında önemli farklılıklar taşıyan özelliklerdir. Elimizde elektronik ortamda bulunan Hürriyet (181.172 sözcük) ve YTÜ (591.357 sözcük) külliyatları için sözcük uzunluğu olasılıkları hesaplanmış, Koltuksuz Külliyatı (876.688 sözcük) için (Koltuksuz, 1995)'de verilen değerler ve ağırlıklı ortalamalarla birlikte Tablo 6.1'de gösterilmiştir. Sonuçlar büyük ölçüde birbirleriyle uyumlu çıkmıştır. Ortalama sözcük uzunluğu ve standart sapma 6,134; 5,187 (Hürriyet), 6,210; 4,822 (YTÜ), 6,070; 5,302 (Koltuksuz) ve 6,128; 5,057 (ağırlıklı ortalama) olarak bulunmuştur. Ağırlıklı

ortalamaya göre Türkçe metinlerde en çok rastlanan sözcük uzunlukları 5 (%16,71), 6 (%12,03), 7 (%11,80), 4 (%10,81) ve 3 (%10,67)'dir. Bu sıralama çok küçük farklarla her üç külliyatta da geçerlidir. Tablo 6.1'deki değerler Türkçe'de 1-10 harf arası uzunluğa sahip sözcüklerin, toplam kullanımın büyük çoğunluğunu oluşturduğunu (%92,2) göstermektedir.

Tablo 12. Türkçe Külliyatlarda Sözcük Boyu Olasılıkları

	Ağr. Ort.	Hürriyet	YTÜ	Koltuks		Ağr. Ort.	Hürriyet	YTÜ	Koltuksuz
1	0,0088	0,0141	0,0103	0,0066	11	0,0321	0,0302	0,0350	0,0305
2	0,0809	0,0872	0,0887	0,0744	12	0,0215	0,0216	0,0243	0,0196
3	0,1067	0,0950	0,1065	0,1093	13	0,0113	0,0106	0,0136	0,0099
4	0,1081	0,1009	0,1017	0,1139	14	0,0064	0,0057	0,0082	0,0054
5	0,1671	0,1538	0,1553	0,1778	15	0,0032	0,0028	0,0044	0,0025
6	0,1203	0,1346	0,1172	0,1194	16	0,0016	0,0015	0,0023	0,0012
7	0,1180	0,1223	0,1163	0,1183	17	0,0007	0,0006	0,0012	0,0005
8	0,0904	0,0924	0,0881	0,0915	18	0,0004	0,0003	0,0006	0,0002
9	0,0715	0,0732	0,0730	0,0702	19	0,0002	0,0002	0,0003	0,0001
10	0,0506	0,0528	0,0527	0,0487	20+	0,0002	0,0002	0,0003	0,0001

İlk/son harf çözümlemeleri, ikili dağılımlar kullanılarak #X (#:boşluk karakteri, X:29 Türkçe harften biri) ve X# örüntüleri üzerinden gerçekleştirilmiştir. Hacettepe Külliyatı'nda boşluk karakteri yer almadığından, ODTÜ Külliyatı'ndaki sonuçlarda ise ikililer hiç olmadığından ilk/son harf çözümlemeleri geri kalan üç külliyat üzerinden (Tablo 13) yapılmıştır.

Tablo 13'nin ilk harf olma olasılığı sütununda ikililerde boşluğu izleyen harflerin (ilk harflerin) Hürriyet, YTÜ ve Koltuksuz külliyatlarındaki oranları verilmektedir. Tabloda da görüldüğü gibi, en çok kullanılan ilk harf sıralaması **B, D, K, A** şeklindedir. Tablo 13'deki ağırlıklı ortalamalar, verilen bir Türkçe metinde bir sözcüğün ilk harfinin en yüksek olasılıkla **B** (%13,22), sesli harflerin içinde ilk harf olma olasılığı en yüksek olan harfin %7,09 ile **A** ve ilk beş sıradaki harflerin yüzdeleri toplamının %44,78 ve ortalamasının %8,96 olduğunu göstermektedir.

Tablo 13'nin son harf olma olasılığı sütununda ikililerde kendisinden sonra boşluk gelen harflerin (son harflerin) Hürriyet, YTÜ ve Koltuksuz külliyatlarındaki oranları verilmektedir. En çok kullanılan son harf sıralaması **N, E, A, R** şeklinde devam etmektedir. Ağırlıklı ortalamalar, verilen bir Türkçe metinde bir sözcüğün son harfinin en yüksek olasılıkla **N** (%15,17), sesli harflerin içinde son harf olma olasılığı en yüksek olan harfin %12,44 ile **E**, ilk 4 sıradaki harflerin ikisinin sesli harfler olduğunu ve ilk beş sıradaki harflerin yüzdeleri toplamı %61,92 ve ortalamasının %12,38 olduğunu göstermektedir. Bu sonuçlar ilk harf sonuçları ile karşılaştırıldığında son harflerin büyük olasılıkla (%83,3) çok daha küçük bir gruptan (NEARİIKUM) çıktığı görülmektedir.

Tablo 13'deki ilk/son harf olma olasılıkları ağırlıklı ortalama değerleri esas alınarak karşılaştırıldığında ortaya şu ilginç sonuç çıkmaktadır. Yazılı Türkçe'de sesli harflerin sözcüklerde ilk harf olma olasılığı %24,69, buna karşın son harf olma olasılıkları ise bunun iki katından çok, %50,27'dir. Ayrıca Tablo 13 incelendiğinde, ilk harf olma olasılık dağılımının belirsizliğinin (4,27 bit/simge) son harf olma olasılık

dağılımının belirsizliğinden (3,72 bit/simge) daha büyük olduğu görülmektedir. Bunun anlamı

Tablo 13. Türkçe Külliyatlarda İlk/Son Harf Olma Olasılıkları (%)

İlk Harf Olma Olasılıkları					Son Harf Olma Olasılıkları				
	Ağırlıklı Ortalama	Hürriyet (1.473.738 Karakter)	YTÜ (4.263.847 Karakter)	Koltuksuz (6.095.457 Karakter)		Ağr. Ort.	Hürriyet (1.473.738 Karakter)	YTÜ (4.263.847 Karakter)	Koltuksuz (6.095.457 Karakter)
B	13,22	11,156	12,421	14,275	N	15,166	15,721	15,113	15,069
D	9,022	7,703	9,39	9,083	E	12,435	12,267	12,917	12,138
K	8,92	7,808	8,713	9,332	A	11,715	11,539	12,018	11,546
A	7,09	7,291	7,019	7,09	R	11,296	9,684	11,377	11,628
S	6,533	6,18	6,391	6,717	İ	11,292	11,282	11,204	11,356
Y	6,494	6,823	6,775	6,218	I	8,264	9,044	8,244	8,089
G	6,255	5,507	5,894	6,687	K	6,285	6,159	6,399	6,235
İ	5,045	5,935	5,199	4,723	U	4,65	4,21	4,225	5,052
O	4,624	4,119	4,45	4,866	M	4,011	3,494	3,259	4,66
H	3,861	3,476	3,631	4,115	Z	2,729	2,525	2,825	2,71
T	3,695	4,883	4,258	3,015	Ş	2,06	1,417	1,894	2,331
E	3,443	3,698	3,308	3,476	T	2,008	2,981	2,18	1,653
M	3,036	4,141	3,092	2,731	L	1,963	2,678	2,235	1,601
V	3,02	3,849	3,039	2,806	P	1,263	1,351	1,161	1,314
Ç	2,925	2,264	2,587	3,321	Ü	1,257	1,245	1,225	1,282
N	2,277	2,407	2,622	2,005	Ç	0,779	0,652	0,75	0,83
Ö	1,678	1,855	1,852	1,514	Y	0,657	0,868	0,661	0,603
P	1,422	1,808	1,568	1,227	O	0,657	0,53	0,678	0,672
Ş	1,332	1,408	1,266	1,36	S	0,517	0,896	0,602	0,367
U	1,094	0,907	1,161	1,092	H	0,403	0,538	0,378	0,387
Ü	0,884	0,946	0,991	0,794	F	0,221	0,227	0,22	0,221
F	0,822	1,132	0,754	0,795	D	0,106	0,207	0,128	0,067
C	0,799	1,231	0,797	0,695	V	0,106	0,148	0,125	0,083
Z	0,778	0,605	0,734	0,851	Ğ	0,058	0,05	0,053	0,064
R	0,763	1,16	0,871	0,592	B	0,036	0,046	0,049	0,023
L	0,514	0,913	0,62	0,344	G	0,026	0,098	0,035	0,003
I	0,414	0,697	0,553	0,25	C	0,023	0,106	0,022	0,003
J	0,04	0,087	0,043	0,027	J	0,016	0,033	0,02	0,01
Ğ	0,002	0,012	0	0	Ö	0,001	0,003	0,002	0
Belirsizlik	4,27	4,37	4,29	4,21		3,72	3,76	3,71	3,71

sözcüklerde ilk harflerin taşıdığı ortalama bilgi miktarının, son harflerin taşıdığı ortalama bilgi miktarından daha çok olduğudur. Bir başka ifade ile Türkçe sözcüklerde ilk harfler son harflerden daha çok belirsizliğe sahiptir.

Tablo 14 .Türkçe Külliyatlarda Sesli/Sessiz Harf Oranları

	Hürriyet	YTÜ	Hacettepe	ODTÜ	Koltuksuz	Ağırlıklı Ortalama
Sesli (%)	42,89	42,84	43,07	42,79	42,90	42,89
Sessiz (%)	57,11	57,16	56,93	57,21	57,10	57,11

Tablo 14'de Hürriyet, YTÜ, Hacettepe, ODTÜ ve Koltuksuz külliyatlarındaki tekli dağılımlarından elde edilen sesli ve sessiz harf oranları verilmiştir. Görüldüğü üzere sonuçlar arasında büyük benzerlik bulunmaktadır. Elde edilen ağırlıklı ortalama da külliyatların değerleriyle neredeyse eşittir. Bu tablo Türkçe bir metinde ortalama olarak %42,89 oranında sesli, %57,11 oranında sessiz harf bulunduğunu göstermektedir. Alfabede bulunan sesli harflerin sayısı (8 tane), sessiz harflere (21 tane) göre çok az olmasına karşın, kullanım oranları yüksektir. İngilizce için 5 sesli harfin yüzdesi yaklaşık %38 ve kalan 21 sessiz harfin ise %62'dir (Dalkılıç G., 2001).

7. SONUÇ

Bu çalışmada, yazılı Türkçe'nin Doğal Dil İşleme'den veri sıkıştırma ve veri güvenliğine kadar uzanan geniş bir yelpazede kullanım alanı bulan n-gram dağılımları, belirsizlik, fazlalık, rastgelelik endeksi, ortalama sözcük uzunluğu ve sesli/sessiz, ilk/son harf oranları gibi önemli bazı istatistiksel özellikleri saptanmış ve çözümlenmiştir.

Türk diliyle ilgili üzerinde çalışma yapılmış, sonuçlar yayınlanmış, standart olarak kabul edilen bir külliyat olmaması (örneğin İngilizce Calgary Corpus gibi) Türkçe üzerinde yapılan çalışmaların zorluğunu arttırmakta, elde edilen sonuçların güvenilirliği konusunda fikir sahibi olmayı güçleştirmektedir. Bu güçlüğü aşmak ve ileride Türk dili üzerinde yapılacak çalışmalar için sağlam bir kaynak oluşturabilmek için yapılan çalışma bilinen beş ayrı Türkçe külliyat'ın 12,5 milyon karakteri aşan bileşkesi üzerinden gerçekleştirilmiş ve sonuçların tutarlılığı bu amaçla geliştirilen bir benzerlik ölçütü ile incelenmiştir.
Gelecek çalışma olarak tekli dağılımları için oluşturulan geniş tabanlı ağırlıklı ortalama değerler ikili dağılımları için de gerçekleştirilebilir. Benzer çalışmalar üçlü, dörtlü ve daha büyük harf gruplarına da genişletilebilir.

KAYNAKLAR

COVER, T. and KING, R. (1978), *A Convergent Gambling Estimate of the Entropy of English*, IEEE Transactions on Information Theory, IT-24, n.4, 413-421.

DALKILIÇ, G. (2001), *Günümüz Türkçesi'nin İstatistiksel Özellikleri ve Bir Metin Sıkıştırma Uygulaması*, Yüksek Lisans Tezi, Uluslararası Bilgisayar Enst., Ege Üniversitesi.

DALKILIÇ, and M.E. DALKILIÇ, G. (2000), *On the Entropy*, Redundancy and Compression of Contemporary Printed Turkish, Proc. of the XV International. Symposium on Computer and Information Sciences, 60-67.

DİRİ, B. (2000), *A Text Compression System Based on the Morphology of Turkish Language*, Proc. of the XV Int'l. Symp. on Computer & Information Sciences, 12-23.

GÖKSU, T. and ERTAUL, L. (1998), *Yer Değiştirmeli ve Dizi Şifreleyiciler için Türkçe'nin Yapısal Özelliklerini Kullanan Bir Kriptoanaliz*, BAS'98, 184-194.

GÖNENÇ, G. (1980), *Türkçe abece İçin 'En İyi' Kodlar*, 3. Ulusal Bilişim Kurultayı, Bilişim'80 Bildiriler Kitabı, 73-75.

JURAFSKY, D. and MARTIN, J.H. (2000), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Prentice Hall.

KOLTUKSUZ, A. (1995), *Simetrik Kriptosistemler için Türkiye Türkçesinin Kriptanalitik Ölçütleri*, Doktora Tezi, Bilgisayar Mühendisliği, Ege Üniversitesi.

SHANNON, C.E. (1951), *Prediction and Entropy of Printed English*, Bell System Technical Journal, 30(1), 50 - 64.

SIEGEL, S. (1956), *Nonparametric Statistics for the Behavioral Sciences*, McGrawHill

STINSON, D.R. (1995), *Cryptography Theory and Practice*, New York: CRC Press.

TÖRECİ, E. (1975), *Statistical Investigations on the Turkish Language Using Digital Computers*, Yüksek Lisans Tezi, ODTÜ (Gönenç, 1980 de referans edildiği şekilde).

Some Important Statistical Properties of Printed Turkish

ABSTRACT

The goal of this study is to determine some statistical properties of printed Turkish. Compiled statistics include the letter frequency (monogram, digram, ... , pentagram) distributions of Turkish, first/last letter analyses, per letter entropy and redundancy, index of coincidence, word length distribution, vowel/consonant proportion. These values are obtained by compiling a corpus from the Internet archive of daily Hurriyet newspaper. Furthermore, using existing studies on Turkish and combining them together, the largest Turkish corpus base to date with the widest text variety and the most comprehensive results are obtained. To determine the degree of agreement for the results of the different studies, a similarity rate measure has been developed and applied to the existing studies' results.

Key Words: *Statistical properties of Turkish, n-gram frequency distributions, entropy, first/last letter analysis, word lengths, similarity assessment of sorted lists.*