**DEÜ MÜHENDİSLİK FAKÜLTESİ**
**FEN VE MÜHENDİSLİK DERGİSİ**
**Cilt: 5 Sayı: 1 sh. 1-7 Ocak 2003**

# TÜRKÇE KÜLLİYAT OLUŞTURULMASI VE TÜRKÇE METİNLERDE KULLANILAN KELİMELERİN UZUNLUK DAĞILIMLARININ BELİRLENMESİ

## *(CREATING A TURKISH CORPUS AND DETERMINING WORD LENGTH DISTRIBUTIONS THAT ARE USED IN TURKISH TEXT)*

**Gökhan DALKILIÇ*, Yalçın ÇEBİ***

**ÖZET/*ABSTRACT***

Bu çalışmada, Türkçe içeriğe sahip 10 ayrı web sitesinden yararlanılarak, 30MB büyüklüğünde bir külliyat oluşturulmuştur. Bu külliyatı oluşturan web sitelerinde kullanılan kelime sayıları hesaplanmış, tüm külliyat için kelime uzunluk dağılımları incelenmiş ve Türkçe'nin ortalama kelime uzunluğu belirlenmiştir.

*In this work, by making use of 10 different web sites having Turkish content, a corpus of 30MB is generated. The number of words used on the web sites that forms this corpus is calculated, word length distribution of all the corpus is analyzed, and average word length of Turkish is determined.*

**ANAHTAR KELİMELER/*KEYWORDS***

Türkçe kelime dağılımı, Türkçe külliyat
*Turkish word distribution, Turkish corpus*

*Dokuz Eylül Üniversitesi, Bilgisayar Mühendisliği, Bornova, İZMİR

## 1. GİRİŞ

Bir dildeki kelimelerin ardalanmalarının olasılıklarının tespiti, o dilin genel yapısının belirlenmesi, konuşma tanıma (speech recognition), el yazısı tanıma (hand-writing recognition), engellilerin iletişimleri ve dilbilgisi hatalarının giderilmesi gibi konular için öncelikle gerekli olan bir konudur.

Cümle içindeki kelimelerden yola çıkarak, önceki kelimeyi temel alarak bir sonraki kelimenin tespiti, özellikle engellilerin iletişimlerinde oldukça büyük kolaylıklar sağlamaktadır. Engellilerin iletişimlerinde, çok küçük el hareketleri ile ekrana gelen menüden kelimelerin seçilmesi ve bunların bir cümle halinde, ses araçları da kullanılarak karşı tarafa aktarılması oldukça karmaşık sistemleri gerektirmektedir. Bazen kullanıcının seçtiği kelimeye bağlı olarak karşısına gelen kelimeler, kullanıcının ekranına sığandan daha fazla sayıda olabilmekte, bu da kullanıcının kelime seçimini zorlaştırmaktadır. Bu nedenle, kullanıcının ekranına, seçilen kelimeye bağlı olarak, o dildeki tüm kelimeler değil, yalnızca seçilen kelimeden sonra gelme olasılığı en yüksek olan kelimelerin gelmeleri sağlanmaktadır (Jurafsky ve Martin, 2000).

Ayrıca, günümüzde iletişimin elektronik ortamla yapılmasının yaygınlaşması ve bilgi aktarımında elektronik ortam kullanımının gelecekte de artacak olması, bilgi aktarımı sırasında çeşitli bilgi koruma yöntemlerini de beraberinde getirmektedir. Bunlardan başlıcası olan bilginin şifrelenerek bir ortamdan başka bir ortama aktarılması, aynı zamanda şifreleme ve şifre çözümleme tekniklerinin kullanımını da yaygınlaştırmaktadır. Bir dildeki kelime dağılımının ve kelime ardalanmalarının olasılıklarının bilinmesi, o dilde gönderilen bir metnin şifrelenmesi veya şifrelenmiş olan bir metnin çözümlenmesi sırasında işlemleri kısaltmak, dolayısıyla da şifreleme/çözümleme sürelerini düşürerek daha karmaşık algoritmaların kullanılmasına imkan tanıyacaktır.

Kelimelerin ardalanma olasılıklarının tesbitinden önce, o dilde kullanılan kelimelerin metinlerden ayıklanması, kelime uzunluklarının ve dağılımlarının belirlenmesi gereklidir. Bu işlem için de o dile ait bir külliyatın oluşturulması ön şart olmaktadır.

Bu çalışmada, elektronik ortamda bulunan değişik dokümanlardan (Internet siteleri ve dokümanlar) yola çıkılarak bir Türkçe külliyat oluşturulmuş, oluşturulan külliyatta bulunan kelimeler uzunluklarının dağılımı açısından incelenmiştir.

## 2. N-GRAM VE TÜRKÇE KÜLLİYATIN OLUŞTURULMASI

Elde bulunan bir metnin hangi dile ait olduğu, ancak o metin içindeki harflerin ve kelimelerin birbirleri ile ilişkilerinin bilinmesi ile mümkün olabilmektedir. Bu ilişkilerin ortaya konması için, istatistiksel yöntemler (olasılık) kullanılmaktadır. Olasılıklar, nesnelerin sayımına dayanmaktadır. Olasılıklardan bahsetmeden önce neyin sayılacağı ve sayılacak olan nesneler için nerelere başvurulacağının bilinmesi gerekmektedir. Bir dilin istatistiksel olarak işlenebilmesi için o dilin *külliyatının* (yazılı ve sözlü anlatımda kullanılan kelimelerin) oluşturulması gerekmektedir. Ancak bu işlemden sonra külliyat üzerinde istatistiksel analizler yapılabilir.

Bir dilde, sözel anlatımda kullanılan kelime sayısı, yazılı anlatımda kullanılan kelime sayısından daha az olmakta, ayrıca, sözlü anlatımdaki kelime yapısı, gerek lehçe farklılıkları nedeniyle, gerekse de diğer nedenlerle yazılı anlatımlara göre değişiklikler gösterebilmektedir. Bu da dildeki kelime sayısının net olarak belirlenmesini zorlaştırmaktadır (Jurafsky ve Martin, 2000).

Bir sonraki kelimenin belirlenebilmesi için yapılan çalışmalar sırasında, kelime dizilerinin (cümlelerin) oluşum olasılıklarının da incelenmesi gerekmektedir. Dilbilgisi açısından son

derece düzgün görünen bir cümle, oluşma olasılığı çok az olan bir yapıda bulunabilir. Örneğin; "*Bir arabayı vurdum.*" cümlesi dilbilgisi açısından doğru olsa bile, Türkçe'de karşılaşılma olasılığı çok düşük olan bir durumu ifade etmektedir.

Bir metin içinde bulunan bir harfin o metinde bulunma olasılığı Eşitlik 1'de bulunmaktadır (Shannon, 1951; Garrett, 2001).

$$\text{Olasılık} = \frac{\text{Metin içindeki "seçilen harf" sayısı}}{\text{Metin içindeki toplam harf sayısı}} \tag{1}$$

Aynı şekilde bir harf dizisinin (kelimenin) bir metin içinde bulunma olasılığı da

$$\text{Olasılık} = \frac{\text{Metin içindeki "seçilen kelime" sayısı}}{\text{Metin içindeki toplam harf sayısı}} \tag{2}$$

şeklinde Eşitlik 2'de hesaplanmaktadır.

Yukarıda verildiği gibi, bir cümleyi oluşturan kelimelerin o cümle içinde bulunma olasılıkları $P(w_1, w_2,..., w_{n-1}, w_n)$ olduğunda, zincir kuralının kullanılması ile bu olasılıkların biraraya getirilmeleri mümkündür ve Formül 3'de gösterilmiştir (Jurafsky ve Martin, 2000).

$$\begin{aligned} P(w_1^n) &= P(w_1)\, P(w_2 \mid w_1)\, P(w_3 \mid w_1^2)...P(w_n \mid w_1^{n-1}) \\ P(w_1^n) &= \prod_{K=1}^{N} P(w_N \mid w_1^{n-1}) \end{aligned} \tag{3}$$

Bu hesaplama sonucunda da, kelimelerin veya kelime içindeki harflerin ardalanmalarının olasılıkları hesaplanmaktadır.

Bir cümlenin içindeki bir harf dizisinin bulunabilme olasılığını belirleyen algoritmalar, aynı zamanda tamamlanmamış bir cümle içinde bulunan ve gelmesi muhtemel bir sonraki kelimenin belirlenmesinde de kullanılabilir. Kelime tahmininde kullanılan modeller '***n-gram'*** olarak adlandırılmaktadır. *N-gram* modeli, yukarıda da belirtildiği gibi, bir sonraki kelimenin o metin içinde bulunabilme olasılığını belirleyebilmek için önceki n-1 adet kelimeyi kullanmaktadır. Konuşma tanıma işleminde, bu türdeki kelime ardalanmalarının istatistiksel modellerini belirtmek için Dil Modeli-DM (*Language Model-LM*) terimi kullanılmaktadır (Jurafsky ve Martin, 2000; Shannon, 1951).

Bir dilin *n-gram* analizinin yapılabilmesi için, o dildeki yazılı ve sözlü metinlerin, mümkün olabildiğince geniş bir boyutta toplanması gerekmektedir. Ayrıca kelimelerin toplanması sırasında, dilin yapısı da göz önünde bulundurulmalı, noktalama işaretleri, boşluklar ve diğer işaretlerin kullanım yerleri de dikkate alınmalıdır. Doğal dillerin işlenmeleri sırasında, özellikle son işaretlerinin oldukça büyük önemi bulunmaktadır. Günümüzde kullanılan birçok yazılım, noktalama işaretlerini de kelime olarak kabul etmektedir (Jurafsky ve Martin, 2000).

Yazılı külliyattan farklı olarak, sözlü külliyat çoğunlukla noktalama işaretleri içermemekle beraber, kelime olarak işlenip işlenmeyeceği belirsiz olan kelimeler de içerebilmektedir. Kelimeler yarıda bölünebilmekte, yazılı külliyatta olmayan sözlü (hı, hım, eee, gibi) ve sözsüz (sessizlik) duraksama ifadeleri bulunabilmektedir. Ayrıca bu ifadelerin her birinin kendisine özgü bir anlamı da bulunmaktadır. Bu anlamların da araştırılması ve bu kelimelerin o dile özgü ve külliyatta yer alabilecek kelimeler olup olmadıkları belirlenmelidir.

Külliyatın oluşturulması sırasında bileşik kelimeler, çoğul kelimeler gibi aynı kökten türeyen, fakat farklı anlamlar da içerebilen kelimelerin de ne şekilde değerlendirileceği belirlenmelidir. Bileşik kelimelerin veya çoğul kelimelerin külliyat içinde ayrı kelimeler olarak değerlendirilebilirliği, külliyatı oluşturan kelime sayısını etkileyecek, analiz algoritmalarının çeşitli değişikliklere uğramasını veya tüm olasılıkları da dikkate alan algoritmaların geliştirilmesini de beraberinde getirecektir.

## 3. TÜRKÇE KÜLLİYATIN İNCELENMESİ

Türkçe'de, kelimelerin metin içindeki dağılımlarının belirlenmesi için oluşturulacak külliyatın mümkün olabildiğince geniş olması, farklı konulardaki metinleri içermesi, hem konuşma, hem de yazı dilini mümkün olabildiğince temsil edebilmesi hedeflenmiştir. Bu amaca yönelik olarak, Türkçe metin miktarı fazla olan ve çeşitlilik içeren web sitelerinde bulunan sayfalar bir program yardımıyla kopyalanmıştır. Çizelge 1'de içeriklerinden yararlanılan web siteleri, bu sitelerden elde edilen sayfalarda kullanılan kelime sayıları, bu sitelerin hangi tarihte kopyalandığı, metin dosyası ve sadece kelimelerin bulunduğu (kelimeler birbirinden birer boşlukla ayrılmıştır) metin dosyası uzunlukları verilmektedir.

Doğal olarak, Çizelge 1'de verilen her bir site 'bir'den çok sayfadan oluşmaktadır ve bu sayfaların ayrı ayrı işlenmeleri oldukça zor olmaktadır. Bu nedenle, her site maksimum 30MB'lık dosyalar halinde birleştirilmiştir. Dosyaların tümü HTML formatında olduğundan, yazılı metinlerin HTML elemanlarından ayrılmaları ve düz metin dosyası (plain text) haline çevrilmeleri gerekmektedir. Bu nedenle, oluşturulan her dosya tek tek Internet Explorer 6.0 kullanılarak düz metin dosyası formatına çevrilmiştir. Bu çevrim işlemi sırasında farklı dosya

Çizelge 1.İçeriklerinden yararlanılan WEB siteleri

| Web Sitesi | Tarih | Kelime Sayısı | Metin Dosyası Büyüklüğü (karakter) | Sadece Kelimelerin Bulunduğu Metin Dosyası Uzunluğu (karakter) |
|---|---|---|---|---|
| www.tbmm.gov.tr | 02.11.2001 | 23.208.581 | 195.589.765 | 169.095.875 |
| www.stargazete.com.tr | 10.11.2001 | 9.749.058 | 109.539.680 | 69.102.032 |
| www.hurriyet.com.tr | 29.10.2001 | 9.503.049 | 114.212.205 | 69.785.084 |
| www.arabul.com | 29.10.2001 | 753.575 | 11.576.335 | 4.994.458 |
| www.pcmagazine.com.tr | 29.10.2001 | 527.746 | 5.109.376 | 3.720.579 |
| www.bilimteknoloji.com.tr | 29.10.2001 | 203.639 | 1.794.934 | 1.450.238 |
| www.abgs.gov.tr* | 19.03.2001 | 160.582 | 1.497.213 | 1.249.054 |
| www.lazland.com | 09.11.2001 | 135.530 | 1.226.056 | 954.395 |
| www.yeniasir.com.tr | 29.10.2001 | 96.858 | 949.910 | 706.598 |
| www.pankitap.com | 29.10.2001 | 59.024 | 532.998 | 425.318 |
| **Toplam:** | | 44.397.642 | 442.028.472 | 321.483.631 |
| | | | **Boşluk** | 44.397.641 |
| | | | **Sadece 29 harf** | 277.085.990 |

*Avrupa Birliği, Türkiye Ulusal Programı

büyüklükleri 512 MB bellek, Intel Pentium III/600 işlemci, Windows XP işletim sistemine sahip bir bilgisayarda denenmiştir. HTML'den metne çeviri işleminde Internet Explorer 6.0'ın daha büyük dosyalarda kilitlenmesi nedeniyle en uygun dosya büyüklüğü olarak maksimum 30 MB seçilmiştir.

Her bir site için oluşturulan dosyalar birleştirilmiş ve Çizelge 1'de belirtilen büyüklükteki metin dosyaları elde edilmiştir. Daha sonra bu metin dosyalarından noktalama ve özel işaretler dışında kalan ve yalnızca 29 Türkçe harf ve boşluktan oluşan karakter dizileri ayrılmıştır. Elde edilen kelimelerin başına ve sonuna birer boşluk gelecek şekilde kelimeler dizilmiş ve Çizelge 1'de belirtilen büyüklükte, yalnızca kelimelerin bulunduğu metin dosyaları elde edilmiştir.

Çizelge 1'de görüldüğü gibi, elde edilen külliyat, toplam 44.397.642 adet kelimeden oluşmaktadır. Bu kelimelerin toplamda oluşturduğu büyüklük ise 44.397.641 karakter boşluk ve 277.085.990 karakter harf olmak üzere toplam 321.483.631 karakter olmaktadır. Boşluk büyüklüğünün tüm metin büyüklüğüne oranı %13,81, yalnızca harflerden oluşan metin büyüklüğüne oranı ise %16,02 olmaktadır. Buna göre, kullanılan külliyat dikkate alındığında, Türkçe metinlerde kelimeler arasında birer boşluk olduğu kabul edilirse boşluk bulunma oranının %13,81 olduğu görülmektedir.

Çizelge 2. Kelime uzunlukları dağılımı

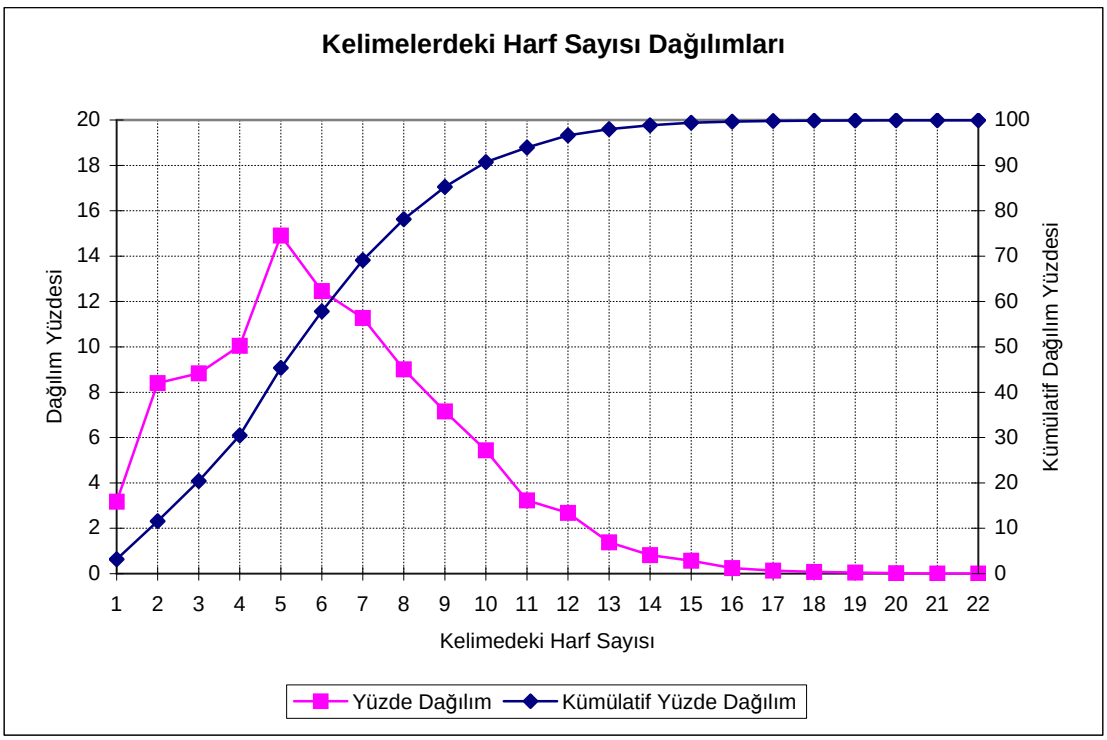
| Kelime Uzunluğu (harf) | Kelime Sayısı (adet) | Dağılım (%) | Kümülatif Dağılım (%) |
|---|---|---|---|
| **1** | 1.408.313 | 3,1720 | 3,1720 |
| **2** | 3.732.165 | 8,4062 | 11,5783 |
| **3** | 3.923.744 | 8,8377 | 20,4160 |
| **4** | 4.462.826 | 10,0519 | 30,4679 |
| **5** | 6.618.464 | 14,9072 | 45,3752 |
| **6** | 5.533.745 | 12,4641 | 57,8392 |
| **7** | 5.005.028 | 11,2732 | 69,1124 |
| **8** | 4.000.834 | 9,0114 | 78,1238 |
| **9** | 3.176.290 | 7, 1542 | 85,2780 |
| **10** | 2.414.330 | 5,4380 | 90,7160 |
| **11** | 1.434.338 | 3,2307 | 93,9466 |
| **12** | 1.188.762 | 2,6775 | 96,6242 |
| **13** | 612.643 | 1,3799 | 98,0041 |
| **14** | 362.578 | 0,8167 | 98,8207 |
| **15** | 253.310 | 0,5705 | 99,3913 |
| **16** | 110.865 | 0,2497 | 99,6410 |
| **17** | 60.065 | 0,1353 | 99,7763 |
| **18** | 34.119 | 0,0768 | 99,8531 |
| **19** | 20.401 | 0,0460 | 99,8991 |
| **20** | 9.881 | 0,0223 | 99,9213 |
| **21** | 4.173 | 0,0094 | 99,9307 |
| **22** | 4.259 | 0,0096 | 99,9403 |
| **23** | 2.828 | 0,0064 | 99,9467 |
| **Toplam** | **44.373.961** | **99,9467** | |

Elde edilen kelime dosyaları toplam büyüklüğünün, toplam kelime sayısına bölünmesi sonucunda, incelenen külliyatta kullanılan kelimelerin ortalama uzunluğu 6,241 harf olarak ortaya çıkmaktadır. Daha önce kelime boyu dağılımının belirlenmesi için yapılan bir çalışmada 11,5 MB büyüklüğünde bir külliyat kullanılmış, ortalama kelime uzunluğu olarak da 6,13 harf elde edilmiştir (Dalkılıç ve Dalkılıç, 2001). Külliyat büyüklüğü yaklaşık 26,5 kat artmış olmasına karşın, ortalama kelime büyüklüğü yaklaşık olarak aynı kalmıştır.

Hazırlanan yazılımlarla, kelime uzunlukları da belirlenmiş olup, Çizelge 2 ve Şekil 1'de verilmektedir. Benzer şekilde dağılımı %0,01'in altında olan kelimeler 1'de yer almamaktadır.

Kelime uzunluklarının dağılımına bakıldığında, "bir" harften oluşan kelime sayısının 1.408.313 adet ile %3,1720 gibi bir oranda olduğu görülmektedir. Bu oran, kelimelerde kullanılan harf sayısı arttıkça artmakta, en yüksek değerine 5 harften oluşan kelimelerde 6.618.464 kelime ve %14,9072 ile ulaşmakta, daha sonra kelimeyi oluşturan harf sayısının artması ile azalmakta ve 13 harften oluşan kelimelerden itibaren de %1'in altına inmektedir.

Kümülatif dağılımlara bakıldığında ise, en fazla 4 harften oluşan kelimeler %30,4679 gibi bir oran oluştururken, 5 harften oluşan kelimelerin eklenmesi ile bu oran %45,3752'ye çıkmakta, 6 harften oluşan kelimelerin eklenmesi ile ise %57,8392 değerine erişmektedir. Bu oran, 7 harften oluşan kelimelerin de eklenmesi ile %69,1124'e erişmekte, artış hızı bundan sonra azalarak 10 harften oluşan kelimelere kadar %90,7160, 15 harften oluşan kelimelere kadar ise %99,3913 değerine ulaşmaktadır. Toplam uzunluğu 16 ile 23 harf arasında olan kelimeler ise bu orana ancak %0,5554 kadar etki edebilmektedir. 23 karakterden daha uzun kelimeler dağılımları %0,01'in altında olduğu için Çizelge 2'de yer almamaktadır. 23 karakterden uzun kelimeler toplam kelimelerin %0,0533'ünü oluşturmaktadır.




Şekil 1. Kelimelerdeki harf sayısı dağılımları

## 4. SONUÇ VE ÖNERİLER

Bu çalışmada 44.397.641 karakter boşluk ve 277.085.990 karakter harfler olmak üzere toplam 321.483.631 karakterden (306MB) oluşan bir külliyat oluşturulmuş ve bu külliyatın hazırlanması sırasında çeşitlilik göz önünde bulundurulmaya çalışılmıştır. Oluşturulan bu külliyatın boyutu bugüne kadar oluşturulan ve sadece 29 Türkçe karakter ve boşluk içeren diğer bazı külliyatlara göre *~6MB, ~4MB, ~1,2MB, ~530KB, ~156KB* çok büyüktür (Diri, 2000; Dalkılıç, 2001; Göksu ve Ertaul, 1998; Töreci, 1975). Bunun nedeni daha önce yapılan çalışmaların harf tabanlı n-gramlar üzerine, bu çalışmanın ise kelime tabanlı n-gramlar üzerine yapılmış olmasıdır.

Yapılan çalışmalar sonucunda, külliyat içindeki boşluk oranının %13,81, ortalama kelime boyunun 6,241 harf olduğu belirlenmiş, külliyat içinde 7 harfe kadar kelimelerin %69,11 gibi bir oranda olduğu, 10 harfe kadar olan kelimelerin ise, oluşturulan külliyatın %91 gibi bir bölümünü oluşturduğu görülmüştür.

Bu çalışma sırasında bileşik kelimelerin ayrılması, köklerin ayrıştırılması, soru eklerinin ve diğer eklerin ayrıştırılması gibi çalışmalara girilmemiş, külliyatın oluşturulması sırasında, incelenen metinlerdeki kelimeler oldukları gibi alınmışlardır.

Bir dilin külliyatının oluşturulması sayesinde, o dile ait çeşitli kalıplar bulunabileceği gibi, farklı kişiler tarafından yazılan metinlerdeki saklı kalıplar (hidden patterns) da araştırılabilir. Buradaki kalıplardan yararlanılarak, şifreleme ve çözümleme işlemlerinde de ilave çalışmalar yapılabilir. Kullanılan külliyat büyüklüğü ne denli artarsa elde edilen sonuçlar da o dile ait karakteristiklerin belirlenmesinde o denli duyarlı olacaktır.

## 5. KAYNAKLAR

Dalkılıç G. (2001): "Günümüz Türkçesi'nin İstatistiksel Özellikleri ve Bir Metin Sıkıştırma Uygulaması", Yüksek Lisans Tezi, Uluslararası Bilgisayar Enstitüsü, Ege Üniversitesi.

Dalkılıç M.E., Dalkılıç G. (2001): "Some Measurable Language Characteristics of Printed Turkish", Proc. of the XVI. International Symposium on Computer and Information Sciences, 217-224 pp.

Diri 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.

Garrett P. (2001): "Making-Breaking Codes", ISBN 0-13-030369-0, Prentice Hall.

Göksu T., Ertaul L. (1998): "Yer Değiştirmeli ve Dizi Şifreleyiciler için Türkçe'nin Yapısal Özelliklerini Kullanan Bir Kriptoanaliz", BAS'98, 184-194.

Jurafsky D., Martin J.H. (2000): "Speech and Language Processing", Prentice Hall.

Shannon C.E. (1951): "Prediction and Entropy of Printed English", The Bell System Technical Journal, 30(1), 50-64pp.

Töreci E. (1975): "Statistical Investigations on the Turkish Language using Digital Computers", Yüksek Lisans Tezi, ODTÜ.