DEÜ MÜHENDİSLİK FAKÜLTESİ
FEN BİLİMLERİ DERGİSİ
Cilt: 12 Sayı: 3 sh. 77-85 Ekim 2010

BİÇİMBİLİME DAYALI DOKÜMAN SIKIŞTIRMA

(MORPHOLOGY BASED TEXT COMPRESSION)

Hayriye GÖKSU*, Banu DİRİ**

ÖZET/*ABSTRACT*

İnternet'in yaygınlaşmasıyla sayısal ortamdaki doküman sayısı gittikçe artmakta ve bu bilgiye daha kolay ve hızlı bir şekilde erişme isteği doküman sıkıştırmayı önemli hale getirmektedir. Doküman sıkıştırma alanında yapılan çalışmaların bir kısmı, dilin biçim bilimsel yapısını kullanmayı amaçlayan çalışmalardır. Bu çalışmada, Türkçe ve İngilizce dokümanların sıkıştırılma verimlerinin belirlenmesinde dilin biçim bilimsel yapısı kullanılarak 10 farklı ayrıştırma yöntemi uygulanmış ve bu yöntemlerin sıkıştırma başarısına olan etkileri karşılaştırmalı olarak verilmiştir.

With the rapid growth of online information, the number of documents in digital media is very common increased and access request to this information easier and quickly makes important the document compression. A part of studies on the document compression, the morphological structure of the language used is intended to work. In this study, Turkish and English language documents to determine the compression efficiency by using the morphological structure of 10 different decomposition methods applied and the effect on the compression success of this method are given in comparison.

ANAHTAR KELİMELER/*KEYWORDS*

Türkçe'nin biçim bilimsel yapısı, İngilizce'nin biçim bilimsel yapısı, Doküman sıkıştırma, Entropi kodlama, N-gram

The morphological structure of Turkish language, The morphological structure of English language, Document compression, Entropy coding, N-gram

* Groupama Emeklilik A.Ş., Eski Büyükdere Cad., No. 2, Maslak, İSTANBUL
** Yıldız Teknik Ün., Bilgisayar Müh. Bölümü, Barbaros Bulvarı, 34349 Yıldız, İSTANBUL

1. GİRİŞ

Zamandan ve yerden kazanç sağlamak için, kayıplı ve kayıpsız yöntemlerle veri üzerinde işlem yapmak anlamına gelen veri sıkıştırma, verinin sayısal ortamda daha fazla yer tutup, maliyet artımına yol açtığında ve belli bir zaman diliminde iletişim kanallarından transfer edilen verinin miktarı söz konusu olduğunda çok büyük bir önem kazanmaktadır. Bilişim teknolojilerindeki gelişmeler veri sıkıştırma yöntemlerinin hem yazılım hem de donanım elemanları ile gerçekleştirilmesini sağlamıştır. Veri sıkıştırmanın temel özellikleri, var olan verinin daha az yer kaplayacak şekilde yeniden düzenlenmesi, zaman ve boyuttan kazanım sağlayarak maliyetin minimize edilmesi olarak özetlenebilir.

Veri sıkıştırma işlemleri temel olarak kayıpsız ve kayıplı sıkıştırma yöntemleri olarak iki gruba ayrılır. Ses, görüntü ve video gibi verilerin sıkıştırma işlemleri kayıplı olurken, dokümanların sıkıştırılması kayıpsız olmaktadır. Kayıpsız sıkıştırma algoritmaları istatistiksel yöntemler ve sözlük tabanlı yöntemler olarak iki gruba ayrıabilir. LZ77-78 (Lempel Ziv), LZW (Lempel-Ziv-Welch), Huffman Kodlama, Aritmetik Kodlama, Run-Length Kodlama ve Dinamik Markov Zinciri bu algoritmalara örnek olarak verilebilir (Nelson, 1996).

Bu makalede, Türkçe'nin biçim bilimsel yapısının sıkıştırmaya uygunluğu araştırılmış ve elverdiği ölçüde İngilizce dili için de benzer bir çalışma yapılmıştır. Türkçe dilinin yapısını kullanan veya dilden bağımsız olarak n-gram'ları kullanarak yapılan literatürde bazı çalışmalar vardır (Çebi ve Dalkılıç, 2004). Diri, Türkçe'nin biçim bilimsel yapısından yararlanarak doküman içerisindeki her kelimeyi kök-ek, hece ve karakter gibi farklı sembollere ayırıp kayıpsız sıkıştırma gerçekleştirmiştir (Diri, 2000). Çelikel ise hem Türkçe hem de İngilizce için sabit ve esnek kelime tabanlı bir yöntem, Rueda ise İkili Arama Ağacı ve Fano kodlamasını kullanarak yeni bir sıkıştırma yöntemi geliştirmiştir (Çelikel vd., 2005; Rueda ve Oommen, 2005). Topaloğlu, İngilizce dili için n-gram'ları kullanarak dokümanların kayıpsız olarak sıkıştırılmasını incelemiştir (Topaloğlu ve Bayrak, 2005). Yapılan bu çalışma Diri'in yaptığı çalışmanın yeni ayrıştırma metotları eklenerek daha genişletilmiş bir halidir (Diri, 2000).

Çalışmanın ikinci bölümünde Entropi Kodlamadan, üçüncü bölümde deneysel sonuçlardan bahsedilmiş ve son bölümde de çalışmanın değerlendirilmesi yapılmıştır.

2. ENTROPİ KODLAMA

Bu çalışmada, entropi kodlama yöntemi olarak en çok tercih edilen Huffman Kodlama yöntemi seçilmiştir. Entropi (E-Entropy) ve Artıklık (R-Redundancy) kavramları bilgi teorisinde önemlidir. a_i gibi tek bir sembolün entropi'si, $-P_i \log_2 P_i$ olarak ifade edilmekte olup, P_i, doküman içerisindeki a_i sembolünün oluşma olasılığı olarak tanımlanmaktadır. Doküman içerisinde n tane farklı sembol var ise, a_i sembolü için ihtiyaç duyulan ortalama bit sayısı Eşitlik 1 deki gibi ifade edilmektedir (Salomon, 1997).

$$-\sum_{1}^{n} P_i \log_2 P_i \tag{1}$$

a_i sembolünün entropi'sine bağlı olan P_i ifadesi, tüm n olasılıkları eşit olduğunda en küçük değerini almaktadır. Bu da gerçek verideki R artıklığını tanımlamak için kullanılmaktadır. Bu ifade entropi ve en küçük entropi arasındaki fark alınarak Eşitlik 2'deki gibi tanımlanmaktadır.

$$R = \left(-\sum_{1}^{n} P_i \log_2 P_i \right) - \log_2 \left(1/P \right) = -\sum_{1}^{n} P_i \log_2 P_i + \log_2 n \qquad (2)$$

Huffman algoritmasında yarı dinamik (adaptive) kodlama tercih edilmiştir. İlk adımda sıkıştırılacak olan doküman içerisiden, seçilen ayrıştırma yöntemine göre karakter, kök, ek, hece, isim, sıfat, fiil, n-gram, vs. gibi sembollerin kullanım frekansları çıkarılır. Daha sonra kodlama ağacı oluşturularak sembollerin Huffman kodları elde edilir. Sıkıştırılmış dosyanın başına da kod çözme sırasında ihtiyaç duyulacak olan başlık (header) bilgisi eklenir. Başlık bilgisi içerisinde sembol ve sembollerin kullanım sıklıkları gibi bilgiler gönderilir. Ancak sıklık bilgisinin gönderilmesi çok yer tutacağından ve kod çözme işleminde ağacın tekrar oluşturulması zaman alacağından bu bilgilerin yerine başlıkta oluşturulan kodlama ağacının kendisi gönderilir. Ağaçtaki her sembol ve sembole ait kod değerleri sol-düğüm-sağ (left-node-right-LNR) notasyonu ile okunarak bir bit katarı oluşturulur. Başlığın oluşturulması hakkındaki detaylı bilgiye Diri'nin çalışmasından ulaşılabilir (Diri, 2000).

N-gram, bir dizide yer alan bir sonraki sembolü tahmin etmek için kullanılan olasılıklı bir modeldir. Klasik yaklaşımda dil modelleme önceki sözcüklere bakarak sonraki sözcüğün tahmin edilmesini ifade etmektedir (Manning ve Schütze, 1999). Bir metnin içinde bir sonraki harfi tahmin etme üzerine olan Shannon Oyunu'na atıfta bulunan bu yaklaşım, konuşma tanıma ve optik karakter tanıma uygulamalarının temelini oluşturmakta, aynı zamanda el yazısı tanıma, yazım hatası bulma ve istatistiksel çeviri gibi uygulamalarda kullanılmaktadır. Doğal dil işleme uygulamalarında n-gram modeller *n* adet birimin art arda sıralanma olasılıklarını kullanarak dilin modellenmesini sağlayan istatistiksel bir yöntemdir. Burada birim sözü ile ifade edilmek istenilen oluşturulacak modelin kullanım amacına yönelik olarak seçilen, ilgili dile ait sözcükler, harfler gibi yapı taşlarıdır. Sembol sayısı 1 olan n-gram unigram (tekli), 2 olan bigram (ikili), 3 olan trigram (üçlü) ve 4 ile üstü ise n-gram olarak adlandırılmaktadır. Çoğunlukla iki-gram ve üç-gram kullanılmaktadır (Eroğlu, 2005). Bu çalışmada, ayrıştırma işleminde, harf iki-gram'ların kullanım frekansları çıkarılarak sembollerin Huffman kodları elde edilmiş ve sıkıştırma işlemi uygulanmıştır.

3. TÜRKÇE' NİN BİÇİMBİLİMSEL YAPISI

Biçimbilgisi, bir dilin kelimelerini, kelime yapılarını, türeme yollarını, çekim bilgilerini inceleyen dilbilgisinin bir koludur. Biçim bilimsel yapısı incelenecek olan dil olarak Türkçe ve İngilizce seçilmiştir. Türkçe Ural-Altay dil ailesine bağlı olup, sondan eklemeli, İngilizce ise Hint-Avrupa dil ailesine bağlı olup, bükümlü bir dildir (Ediskun, 2005). Bu nedenle Türkçe ile İngilizce arasında sözdizim (sentaks), anlambilim (semantic), sesbilim (fonetic), sözcükbilim (lexicology), biçimbilim (morfolojic) yönlerinden önemli farklar vardır. Bu çalışmada Türkçe'deki her kelime karakter, hece, isim ve isme eklenen ekler, fiil ve fiile eklenen ekler, sıfat ve sıfata eklenen ekler, ekleri tek parça, ek, hece veya karakter halinde ayrı ayrı ele alarak, ayrıca dokümanın 2-gram'ları da çıkartılarak incelenmiştir.

Türkçe dokümanların biçim bilimsel yapılarına ayırma işleminde Zemberek isimli açık kaynak kodu kullanılmıştır (Akın vd., 2004). Kelimenin heceleri, kökü, eki/ekleri, isim, fiil veya sıfat olduğu bu araç yardımıyla bulunmaktadır. Ancak bazı çözümlemelerde hece düşmesinden kaynaklanan problemler yaşanmaktadır. Örneğin, 'bağrına' kelimesi çözümleme işlemine tabi tutulduğunda, kök "bağır", "ına" eki olarak analiz edilmektedir. Kök ve ek bu şekilde kodlanacak olursa kodçözme işlemi sırasında orijinal metnin aynısını elde edilemez. Yazılan program içerisinde bu gibi durumlar saptanarak Zemberek'ten gelen sonuçlar kontrol edilerek gerekli düzeltmeler yapılmıştır.

İngilizce dokümanların biçim bilimsel yapılarına göre ayırma işleminde Porter Stemming algoritması kullanılmıştır (Porter, 2006). Bu algoritma İngilizce kelimelerde yer alan biçim bilimsel ve çekimsel ekleri ayırma işlemini gerçekleştirmektedir. Bu çalışmada İngilizce'deki her kelime karakter, kök ve eki ek, kök ve eki karakter olmak üzere farklı şekillerde ayrıştırılarak kodlanmıştır.

Son aşamada da hem Türkçe hem de İngilizce dokümanların 2-gram'larını çıkaran bir kod yazılmış ve sadece 2-gram'lara göre kodlama gerçekleştirilmiştir.

4. DENEYSEL SONUÇLAR

Bu bölümde öncelikli olarak denemelerin yapıldığı 8 farklı külliyat oluşturulmuştur. Bu külliyatlardan K1-K7 arası Türkçe için, K8 ise İngilizce için hazırlanmıştır. K1 külliyatı; popüler hayat, güncel, ekonomi, spor, dünya ve sağlık gibi farklı konularda yazan 18 erkek yazara ait 376 farklı dokümandan, K2 külliyatı; sadece güncel konularda yazan 7 erkek yazara ait 237 dokümandan, K3 külliyatı; popüler hayat, güncel, ekonomi, spor, dünya ve sağlık gibi farklı konularda yazan 18 erkek ve 4 kadın yazara ait 516 dokümandan, K4 külliyatı; siyaset konusunda yazan 12 erkek ve 1 kadın yazara ait 271 dokümandan, K5 külliyatı; popüler hayat, güncel, ekonomi, spor, dünya ve sağlık gibi farklı konularda yazan 4 kadın yazara ait 140 dokümandan, K6 külliyatı; güncel konularda yazan 2 kadın yazara ait 98 dokümandan, K7 külliyatı; Türkiye Anayasa metni ile 6 öykü ve 1 romana ait 46 dokümandan, ve İngilizce için hazırlanan K8 külliyatı güncel konularda yazılmış 290 farklı dokümandan oluşturulmuştur. K1-K6 arasındaki külliyatlar günlük gazetelerin köşe yazılarından toplanmıştır. Farklı külliyatlar oluşturulmasının sebebi sadece kadın yazarların veya erkek yazarların ya da tek bir konuda veya farklı konularda yazılmış dokümanların biçim bilimsel analizi yapılarak sıkıştırıldığında performansın nasıl etkilendiğini görmektir. Herbir külliyat için 64 Kb, 256 Kb ve 1024 Kb'lık dokümanlar elde edilmiştir.

Türkçe için tüm külliyatlara 10 ayrıştırma yöntemi uygulanmıştır. Bunlar sırasıyla; KET-kelimenin kökü ve ekler ek bazında, H-kelimenin heceleri, K-kelimeyi oluşturan karakterler, İET-isim kökü ve ekler tek parça, FET-fiil kökü ve ekler tek parça, SET-sıfat kökü ve ekler tek parça, İEK-isim kökü ve ekler karakter bazında, FEK-fiil kökü ve ekler karakter bazında, SEK-sıfat kökü ve ekler karakter bazında, 2-Gr, kelimenin iki-gram olarak analiz edilmesidir. Bu ayrıştırma yöntemlerine örnek verilirse;

- KET: kitaplarda (kitap, -lar, -da),
- H: kitaplarda (ki, -tap, -lar, -da),
- K: kitaplarda (k, i, t, a, p, l, r, d),
- İET: kitaplarda (kitap, -larda),
- FET: yüzecek (yüz, -ecek),
- SET: güzellik (güzel, -lik),
- İEK: kitaplarda (kitap, l, a, r, d),
- FEK: yüzecek (yüz, e, c, k),
- SEK: güzellik (güzel, l, i, k),
- 2-Gr: soru (so, or, ru)

olarak göstelebilir. Ayrıştırma yöntemlerinden ilk 7 tanesi için alınan sıkıştırma başarılarına Çizelge 1'den Çizelge 7'ye kadar yer verilmiştir. İEH (isim kökü ve ekler hece bazında), FEH (fiil kökü ve ekler hece bazında) ve SEH (sıfat kökü ve ekler hece bazında) ayrıştırma yöntemleri diğerlerine yakın veya daha kötü sonuçlar verdiğinden araştırmanın dışında tutulmuştur. Sıkıştırma oranı Eşitlik 3'deki gibi hesaplanmaktadır.

So=100*(1-(cikisdosyaboyu/girisdosyaboyu)) (3)

Yedi külliyatta doğrudan görebilen şey, sıkıştırma yönteminden bağımsız olarak dosya boyutu arttıkça sıkıştırma oranının artmasıdır. Kısaca sıkıştırma oranı, dosya boyu ile doğru orantılıdır.

Çizelge 1. Türkçe için K1 külliyatına uygulanan 7 ayrıştırma yönteminin sıkıştırma oranları (%)

	KET	**H**	**K**	**İEK**	**FEK**	**SEK**	**2-Gr**
64Kb	45	43	39	37	38	39	40
256Kb	49	47	39	42	39	39	44
1024Kb	**50**	48	40	43	41	39	45
K1(Ort)	**48**	46	39.3	40.7	39.3	39	43

Çizelge 2. Türkçe için K2 külliyatına uygulanan 7 ayrıştırma yönteminin sıkıştırma oranları (%)

	KET	**H**	**K**	**İEK**	**FEK**	**SEK**	**2-Gr**
64Kb	44	43	39	37	38	39	38
256Kb	48	46	39	42	39	39	43
1024Kb	**50**	48	40	43	39	40	46
K2(Ort)	**47.3**	45.7	39.3	40.7	38.7	39.3	42.3

K1 ve K2 külliyatlarının her ikisi de erkek yazarlardan oluşmuştur. Aynı veya farklı konularda yazılmış olsalar da sıkıştırma oranları birbirine çok yakındır. En başarılı yöntem kök ve ekler ek olarak değerlendirildiğinde olmuştur. Ortalama sıkıştırma oranı başarısı K1 için % 48, K2 için de % 47.3'tür. Bunu izleyen diğer yöntem de her iki külliyat için hecelerin sıkıştırılmasıdır.

K3 ve K4 külliyatları hem erkek hem de kadın yazarlardan, genel ve sadece tek bir konudan oluşmaktadır.

Çizelge 3. Türkçe için K3 külliyatına uygulanan 7 ayrıştırma yönteminin sıkıştırma oranları (%)

	KET	**H**	**K**	**İEK**	**FEK**	**SEK**	**2-Gr**
64Kb	45	43	39	37	38	39	40
256Kb	49	47	39	42	39	39	44
1024Kb	50	48	41	**56**	41	40	46
K3(Ort)	**48**	46	39.7	45	39.3	39.3	43.3

K3 külliyatında ortalama en başarılı yöntem KET olmasına karşılık, isim ve ekler karakter bazında kodlanmış olan İEK yönteminde % 56'lık bir başarı alınmıştır. Bu durum bize kadın yazarların erkeklere göre yazılarında daha fazla isim kullandığını ve bunların tekrarladığını göstermektedir.

K4 külliyatında en yüksek başarı hem ortalamada hem de dosya bazında KET yönteminden alınmıştır. Aynı konularda yazılmış olan K2 ve K4 külliyatları arasında 1024Kb'lık dosyadaki farkın yine kadın yazarlardan kaynaklandığı düşünülmektedir. Başarı

daha yüksek olduğu için tekrar eden aynı kök sayısı fazladır. Bu da kadın yazarların daha az kelime dağarcığı kullandığını göstermektedir.

Çizelge 4. Türkçe için K4 külliyatına uygulanan 7 ayrıştırma yönteminin sıkıştırma oranları (%)

	KET	**H**	**K**	**İEK**	**FEK**	**SEK**	**2-Gr**
64Kb	44	43	39	36	38	39	41
256Kb	55	54	47	49	47	47	52
1024Kb	**61**	60	51	56	52	52	56
K4(Ort)	**53.3**	52.3	45.7	47	45.7	46	49.7

K3 ve K4 külliyatlarında ikinci başarılı yöntem H'dir. K4 külliyatının başarılı olmasının sebebi tek bir konu hakkında yazıldığında elde edilen iki-gram'ların sayısının daha az ve kullanım frekans değerlerinin yüksek olmasıdır.

Çizelge 5. Türkçe için K5 külliyatına uygulanan 7 ayrıştırma yönteminin sıkıştırma oranları (%)

	KET	**H**	**K**	**İEK**	**FEK**	**SEK**	**2-Gr**
64Kb	44	43	39	36	38	39	40
256Kb	48	47	39	41	39	39	45
1024Kb	**51**	49	39	42	39	39	47
K5(Ort)	**47.7**	46.3	39	39.7	38.7	39	44

Çizelge 6. Türkçe için K6 külliyatına uygulanan 7 ayrıştırma yönteminin sıkıştırma oranları (%)

	KET	**H**	**K**	**İEK**	**FEK**	**SEK**	**2-Gr**
64Kb	43	43	39	35	37	39	41
256Kb	48	47	40	41	39	40	45
1024Kb	**49**	48	41	43	40	40	47
K6(Ort)	**46.7**	46	40	39.7	38.7	39.7	44.3

K5 ve K6 külliyatları ise sadece kadın yazarlara aittir ve her ikisinde en başarılı yöntem KET olmuştur. Diğer yöntemlerin başarıları birbirlerine oldukça yakındır. Külliyatlar ister karışık, isterse tek bir konuda yazılmış olsa bile kadın yazarların yazım stilleri birbirlerine oldukça yakındır.

Çizelge 7. Türkçe için K7 külliyatına uygulanan 7 ayrıştırma yönteminin sıkıştırma oranları(%)

	KET	**H**	**K**	**İEK**	**FEK**	**SEK**	**2-Gr**
64Kb	45	44	40	37	38	40	42
256Kb	49	49	41	42	40	41	45
1024Kb	50	**51**	41	44	42	41	48
K7(Ort)	**48**	**48**	40.7	41	40	40.7	45

K7 külliyatı ise anayasa, öykü ve romandan oluşmakta olup, içerik olarak gazete köşe yazılarından oldukça farklıdır. Bu külliyat için dosya bazında en başarılı ayrıştırma yöntem, H olmuştur. Ortalama başarıda ise KET ve H ayrıştırma yöntemleri daha başarılıdır. Genelde tüm ayrıştırma yöntemlerinde 64 Kb gibi kısa dosya boyutunda başarı sonuçları düşüktür. Ancak, külliyat K7'de en yüksek sıkıştırma oranı alınmıştır. Bunun nedeninin K7 içerisinde yer alan dokümanların daha edebi yapıtlar olduğu ve dilin güzel kullanılmış olduğu düşünülmektedir.

Bu çalışmanın benzeri İngilizce için elimizde bulunan dil çözümleyicisinin imkanları dahilinde yapılmıştır (Porter, 2006). Metinler biçim bilimsel yapılarına göre, ekler ek bazında olmak üzere kök-ek (İKET), ekler karakter olmak üzere kök-ek (İKEK), karakter (İK) ve 2-Gr olarak 4 farklı şekilde ayrıştırılmışlardır. Güncel konularda yazan farklı kişilere ait 290 farklı dokümandan oluşan K8 külliyatı için başarı sonuçları Çizelge 8'de verilmektedir. İngilizce'de en başarılı ayrıştırma yöntemi kök ve ekler karakter halinde kodlanan İKEK yöntemi'dir. Ortalama başarısı % 52.7'dir. İkinci ayrıştırma yöntemi de karakter bazında yapılan İK kodlama yöntemidir.

Çizelge 8. İngilizce için hazırlanan K8 külliyatına uygulanan 4 ayrıştırma yönteminin sıkıştırma oranları (%)

	İKET	**İKEK**	**İK**	**2-İGr**
64Kb	42	**54**	53	45
256Kb	39	52	51	46
1024Kb	40	52	51	46
K8(Ort)	40.3	**52.7**	51.7	45.7

Bu iki dili karşılaştıracak olursak, KET ve İKET ayrıştırma yöntemlerinde Türkçe İngilizce'ye göre daha yüksek sıkıştırma başarısı vermiştir. Türkçe'nin sondan eklemeli bir dil olması sebebiyle toplam kök sayısı, İngilizce diline göre daha azdır. Türkçe'de kelime köküne eklenen yapım ekleri ile farklı kelimeler elde ediliyorken, İngilizce için böyle bir durum söz konusu değildir. Karakter bazında ayrıştırma yapan K ve İK yöntemlerini karşılaştırdığımızda ise İngilizce dili için daha başarılı sıkıştırma performansı elde edilmiştir. İki-gram ayrıştırma yöntemi K4 külliyatı haricinde, İngilizce dilinde daha iyi sıkıştırma performansı vermiştir.

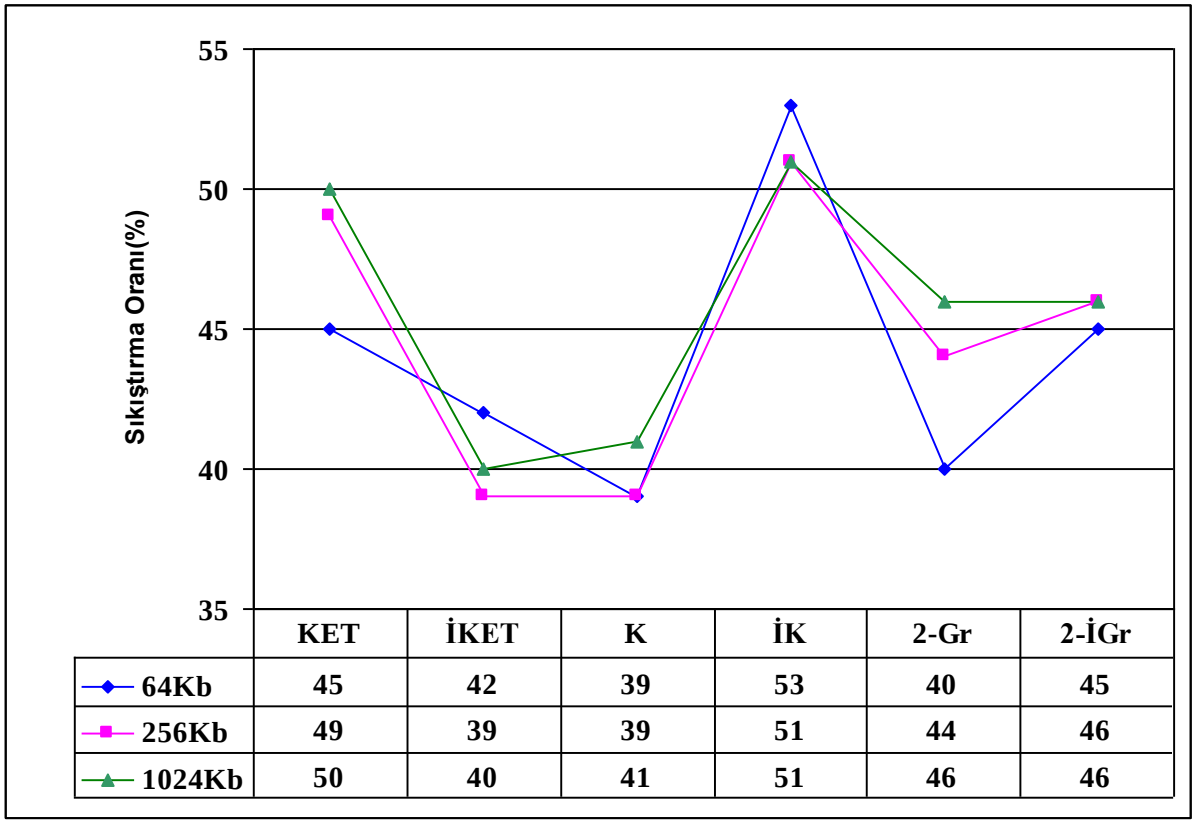
Türkçe'de başarılı sonuç vermediğinden bu çalışmada yer verilmeyen kök ve ekler karakter olarak alınan İKEK ayrıştırma yöntemi İngilizce için % 52.7 gibi ortalama başarı vermiştir.

Şekil 1'de her iki dil için, kök ve ekler ek halinde, karakter ve iki-gram olarak ayrıştırılmış, 3 dosya boyutundaki sıkıştırma performansları karşılaştırmalı olarak verilmektedir. Şekil 1'de görüldüğü gibi kök ve ekler ek halinde olan ayrıştırma yöntemi Türkçe dili için daha başarılı olmuşken karakter tabanlı ayrıştırma İngilizce dili için oldukça yüksek bir performans sergilemiştir. İki-gram ayrıştırma yöntemi ise dosya boyutu arttıkça her iki dil için de yaklaşık aynı sıkıştırma başarısını vermektedir.

5. SONUÇLAR

Bu çalışmada, Türkçe ve İngilizce dokümanlarda dilin biçim bilimsel yapısı kullanılarak farklı ayrıştırma yöntemleri ile entropi kodlama ile kayıpsız sıkıştırma uygulanmıştır. Türkçe için yedi farklı külliyat, genel veya tek bir konuda, erkek, kadın veya karışık yazarlardan oluşturulmuştur. Biçim bilimsel olarak kelime kökü ve köke eklenen ekler ek, karakter, hece,

isim ve isme eklenen ekler tek parça, karakter, hece, fiil ve fiile eklenmiş ekler tek parça, karakter, hece, sıfat ve sıfata eklenmiş ekler tek parça, karakter, hece bazında ve iki-gramlar olmak üzere yapılmıştır. İngilizce için ise genel bir külliyat oluşturulmuş, sadece kelime kökü ve köke eklenen ekler ek, kök ve köke eklenen ekler karakter bazında, karakter ve iki-gram'lar olmak üzere dört farklı ayrıştırma yöntemi tercih edilmiştir. Herbir ayrıştırma yöntemiyle elde edilen sıkıştırma sonuçlarının karşılaştırılması yapılmıştır. Hangi ayrıştırma yönteminin hangi dilde daha başarılı olduğu tartışılmıştır.



	KET	İKET	K	İK	2-Gr	2-İGr
64Kb	45	42	39	53	40	45
256Kb	49	39	39	51	44	46
1024Kb	50	40	41	51	46	46

Şekil 1. Türkçe ve İngilizce dilleri için 3 farklı ayrıştırma yönteminin karşılaştırılması

Yapılan denemelere göre, en başarılı sonuçlar, 1024 Kb boyutundaki dosyalarda ve tez yazılarından elde edilen külliyattan sağlanmıştır. Farklı konularda yazan kadın yazarların yazılarının yer aldığı dokümanlardan ise en başarısız sıkıştırma oranları elde edilmiştir.

Türkçe ve İngilizce dilleri için KET ile yapılan ayrıştırma şeklinde sıkıştırma işlemi yapılıp, sonuçlar karşılaştırıldığında en düşük ve en yüksek başarının Türkçe dokümanlardan elde edildiği gözlemlenmiştir. K ile yapılan ayrıştırma şeklinde sıkıştırma işlemi yapılıp, sonuçlar karşılaştırıldığında ise en düşük ve en yüksek başarı İngilizce dokümanlardan elde edilmiştir. 2-Gr ile yapılan ayrıştırma şeklinde sıkıştırma işlemi yapılıp, sonuçlar karşılaştırıldığında ise en düşük başarı İngilizce dokümanlardan, en yüksek başarı ise Türkçe dokümanlardan elde edilmiştir. İngilizce dokümanların sıkıştırma verimleri göz önünde bulundurulduğunda, dosya boyutuna bağlı olarak çok büyük artışlar görülmemekle birlikte, Türkçe dokümanların sıkıştırma verimlerinin dosya boyutuna bağlı olarak artış gösterdiği görülmüştür.

Bu çalışmanın amacı yüksek sıkıştırma oranına sahip bir yöntem bulmak değil, doğal bir dilin biçim bilimsel yapısı kullanılarak bir sıkıştırma yapılırsa nasıl bir sonuç alınacağı ve dillerin hangi özelliklerinin avantaj sağlarken, hangi özelliklerinin dezavantaj sağlayacağını

görmektir. Elde edilen deney sonuçlarının daha iyi olabilmesi için dile bağlı olarak kullanılan biçim bilimsel analiz yapan araçların işlem yapma başarımlarında iyileştirmeler yapılması gereklidir. Yine çalışma içerisinde sıkıştırma sürelerinden bahsedilmemiştir. Çünkü biçim bilimsel analizin yapılması ve kayıpsız sıkıştırmanın gerçekleştirilebilmesi için analiz sonucundaki çıktıların dilin özelliklerine bağlı olarak tekrardan işleme sokulması gerekmektedir. Bu süreç de işlem süresini artırmakta ve kıyaslanabilir durumdan çıkarmaktadır. Sıkıştırılmış dosyanın geri açılması ise daha hızlı olarak gerçekleşmektedir.

Bu çalışma, bugüne kadar Türkçe dokümanların biçim bilimsel yapılarını kullanarak sıkıştırma işleminin yapılmasını sağlayan en kapsamlı çalışmadır.

KAYNAKLAR

Akın M. D., Kaba S., Ahmet A. A. (2004): "Zemberek Projesi", https://zemberek.dev.java.net/

Çebi Y., Dalkılıç G. (2004): "Turkish Word N-gram Analysing Algorithms for a large Scale Turkish Corpus-TurCo", IEEE International Conference on Information Technology, Cilt 2, s. 226-240.

Çelikel E., Dalkılıç M. E., Dalkılıç G. (2005): "Word-Based Fixed and Flexible List Compression", Computer and Information Sciences-ESCIS, LNCS 3733, Springer Verlag, sp. 780-790.

Diri B. (2000): "A Text Compression System Based on the Morphology of Turkish Language", International Symposium on Computer and Information Sciences (ISCIS) XV, 11-13 October, Istanbul.

Ediskun H. (2005): "Türk Dilbilgisi Sesbilgisi-Biçimbilgisi-Cümlebilgisi", Remzi Kitabevi A.Ş., Selvili Mescit Sok. 3, Cağaloğlu 34440, İstanbul, Türkiye.

Eroğlu Ö. S. (2005): "Hece Tabanlı İstatistiksel Yöntemler ile Yazım Hatası Bulma ve Düzeltme", Yüksek Lisans Tezi, İ.T.Ü. Fen Bilimleri Enstitüsü.

Manning C., Schütze H. (1999): "Foundations of Statistical Natural Language Processing", MIT Press, ISBN 0-262-13360-1, Cambridge, USA.

Nelson M. (1996): "The Data Compression Book", NewYork, USA.

Porter M. (2006): "Stemming Algoritması", http://tartarus.org/~martin/PorterStemmer/

Rueda L., Oommen B. J. (2005) "Efficient Adaptive Data Compression Using Fano Binary Search Trees", Computer and Information Sciences-ISCIS, Cilt 3733, s. 768-779.

Topaloğlu U., Bayrak C. (2005): "Polymorphic Compression", Computer and Information Sciences-ISCIS, Cilt 3733, s. 759-767.

Salomon D. (1997): "Data Compression", Springer-Verlag, NewYork.