

**YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**TÜRKÇENİN BİÇİMBİLİMSEL YAPISININ
SIKIŞTIRMAYA UYGUNLUĞUNUN ARAŞTIRILMASI**

Bilgisayar Mühendisi Hayriye Göksu

**FBE Bilgisayar Mühendisliği Anabilim Dalında
Hazırlanan**

YÜKSEK LİSANS TEZİ

Tez Danışmanı : Yrd. Doç. Dr. Banu DİRİ

İSTANBUL, 2009

İÇİNDEKİLER

	Sayfa
SİMGE LİSTESİ	v
KISALTMA LİSTESİ	vi
ŞEKİL LİSTESİ	vii
ÇİZELGE LİSTESİ	ix
ÖNSÖZ	xi
ÖZET	xii
ABSTRACT	xiii
1. GİRİŞ	1
2. VERİ SIKIŞTIRMA	3
2.1 Kayıpsız Sıkıştırma Algoritmaları	3
2.2 Kayıplı Sıkıştırma Algoritmaları	7
3. ENTROPY KODLAMA	9
3.1 Entropy	9
3.2 Huffman Kodlama	9
3.3 Statik Huffman Kodlama	12
3.4 Huffman Kodu Çözümleme	12
3.5 Huffman Kodlamanın Başlık Yapısı	14
4. TÜRKÇENİN BİÇİMBİLİMSEL YAPISI	16
4.1 Heceler	16
4.2 Kelimeler	17
4.3 Kökler	17
4.4 Ekler	17
4.5 İsimler	17
4.6 Sıfatlar	18
4.7 Fiiller	18
4.8 Biçimbilimsel Analizin Yapıldığı Araç	18
5. İNGİLİZCENİN BİÇİMBİLİMSEL YAPISI	21

5.1	Temel Terimler	21
5.2	İngilizce Bükümlü Biçimbilimi	22
5.3	Düzenli ve Düzensiz Biçimbilim	23
5.4	İngilizce Türetme Biçimbilimi	24
5.5	Biçimbilimsel Analizin Yapıldığı Araç.....	24
6.	N-GRAM.....	25
7.	TEST ORTAMININ OLUŞTURULMASI.....	28
7.1	Türkçe Külliyatlar	28
7.1.1	Erkek(Heterojen) Külliyatı (Külliyat-I)	28
7.1.2	Erkek(Homojen) Külliyatı (Külliyat-II).....	28
7.1.3	Heterojen Külliyatı (Külliyat-III)	28
7.1.4	Homojen(Siyaset) Külliyatı (Külliyat-IV)	28
7.1.5	Kadın(Heterojen) Külliyatı (Külliyat-V).....	28
7.1.6	Kadın(Homojen) Külliyatı (Külliyat-VI)	29
7.1.7	Roman Anayasa Metinleri Külliyatı (Külliyat-VII)	29
7.1.8	Tez Dokümanları Külliyatı (Külliyat-VIII).....	29
7.2	İngilizce Külliyat(Külliyat-IX)	29
8.	DENEYSEL SONUÇLAR.....	30
8.1	Türkçe Üzerine Yapılan Çalışmalar.....	30
8.1.1	Külliyata Göre Deneysel Sonuçlar.....	30
8.1.1.1	Külliyat-I için Deneysel Sonuçlar.....	31
8.1.1.2	Külliyat-II için Deneysel Sonuçlar	34
8.1.1.3	Külliyat-III için Deneysel Sonuçlar	37
8.1.1.4	Külliyat-IV için Deneysel Sonuçlar.....	40
8.1.1.5	Külliyat-V için Deneysel Sonuçlar	43
8.1.1.6	Külliyat-VI için Deneysel Sonuçlar.....	46
8.1.1.7	Külliyat-VII için Deneysel Sonuçlar.....	49
8.1.1.8	Külliyat-VIII için Deneysel Sonuçlar	52
8.2	İngilizce Üzerine Yapılan Çalışmalar	57
8.2.1	Külliyata Göre Deneysel Sonuçlar.....	57
8.2.1.1	Külliyat-IX için Deneysel Sonuçlar.....	57
8.3	Ayrıştırma Yöntemine Dayalı Sıkıştırma Sonuçlarının Karşılaştırılması.....	60
9.	TÜRKÇENİN BİÇİMBİLİMSEL YAPISINA DAYALI İSTATİSTİKSEL DEĞERLER	69
9.1	Külliyat-VIII için Ayrıştırma Şekline Dayalı Kullanım Frekans Değerleri.....	69
10.	BİÇİMBİLİMSEL YAPISINA BAĞLI OLARAK TÜRKÇE VE İNGİLİZCENİN SIKIŞTIRMA PERFORMANSLARININ KARŞILAŞTIRILMASI.....	81

11.	SONUÇ	85
	KAYNAKLAR	87
	İNTERNET KAYNAKLARI	88
	ÖZGEÇMİŞ	89

SİMGE LİSTESİ

$KU(x)$	x sembolünün kod uzunluğu
P_i	i sembolünün olasılık değeri
$\log_2 P_i$	Pi' nin 2 tabanındaki logaritması
a_i	Herhangi bir sembol ifadesi
Σ	Toplam
$\leq$	Küçük veya eşit
m	Herhangi bir sembolü ifade etmek için ortalamanın üzerinde ihtiyaç duyulan en az bit sayısı

KISALTMA LİSTESİ

KET	Ekler ek halinde olmak üzere kök – ek
H	Hece
K	Karakter
İET	Ekler tek parça halinde olmak üzere isim – ek
FET	Ekler tek parça halinde olmak üzere fiil – ek
SET	Ekler tek parça halinde olmak üzere sıfat – ek
İEH	Ekler hece halinde olmak üzere isim – ek
FEH	Ekler hece halinde olmak üzere fiil – ek
SEH	Ekler hece halinde olmak üzere sıfat – ek
İEK	Ekler karakter olmak üzere isim – ek
FEK	Ekler karakter olmak üzere fiil – ek
SEK	Ekler karakter olmak üzere sıfat – ek
İKET	İngilizce için ekler ek halinde olmak üzere kök – ek
İKEK	İngilizce için ekler karakter olmak üzere kök – ek
İK	İngilizce için karakter
A.Ş.	Ayrıştırma şekli
LNR	Left node right
HHI	Header huffman initials
NOB	Number of bits in code values
SPT	Symbol part type
SPST	Symbol part special char1
CP	Code part
SP	Symbol part
E	Entropy
R	Redundancy
NLP	Natural language processing
JPEG	Joint photographic experts group
MPEG	Moving pictures experts group
RLE	Run length encoding
TIFF	Tagged image file format
PCX	Personal computer exchange
XML	Extensible Markup Language
BCPL	Basic Combined Programming Language

ŞEKİL LİSTESİ

Şekil 3.1 Örnek bir Huffman ağacı.	10
Şekil 3.2 Ağaç oluşturma ilk aşama	11
Şekil 3.3 Ağaç oluşturma ikinci aşama	11
Şekil 3.4 Ağaç oluşturma üçüncü aşama	12
Şekil 4.1 Heceyi oluşturan fonemlerin heceyi oluşturma şekilleri(Ediskun)	16
Şekil 4.2 Bilgi elemanları ve dilbilgisi çıkarım yapısı(Ediskun)	19
Şekil 8.1 Külliyat-I için sıkıştırma oranları (%)	32
Şekil 8.2 Külliyat-I için toplam sıkıştırma süreleri (sn)	34
Şekil 8.3 Külliyat-II için sıkıştırma oranları (%)	36
Şekil 8.4 Külliyat-II için toplam sıkıştırma süreleri (sn)	37
Şekil 8.5 Külliyat-III için sıkıştırma oranları (%)	39
Şekil 8.6 Külliyat-III için toplam sıkıştırma süreleri (sn)	40
Şekil 8.7 Külliyat-IV için sıkıştırma oranları (%)	41
Şekil 8.8 Külliyat-IV için toplam sıkıştırma süreleri (sn)	43
Şekil 8.9 Külliyat-V için sıkıştırma oranları (%)	45
Şekil 8.10 Külliyat-V için toplam sıkıştırma süreleri (sn)	46
Şekil 8.11 Külliyat-VI için sıkıştırma oranları (%)	48
Şekil 8.12 Külliyat-VI için toplam sıkıştırma süreleri (sn)	49
Şekil 8.13 Külliyat-VII için sıkıştırma oranları (%)	51
Şekil 8.14 Külliyat-VII için toplam sıkıştırma süreleri(sn)	52
Şekil 8.15 Külliyat-VIII için sıkıştırma oranları (%)	54
Şekil 8.16 Külliyat-VIII için toplam sıkıştırma süreleri(sn)	55
Şekil 8.17 Külliyat-IX için sıkıştırma oranları (%)	58
Şekil 8.18 Külliyat- IX için toplam sıkıştırma süreleri (sn)	59
Şekil 9.1 Külliyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek 50 kök ve kullanım frekans değerleri	71
Şekil 9.2 Külliyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek 50 hece ve kullanım frekans değerleri	73
Şekil 9.3 Külliyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek 50 karakter ve kullanım frekans değerleri	74
Şekil 9.4 Külliyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek 50 isim ve kullanım frekans değerleri	76
Şekil 9.5 Külliyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek 50 hece ve kullanım frekans değerleri	78
Şekil 9.6 Külliyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek 50 fiil ve kullanım frekans değerleri	80

Şekil 10.1 KET sıkıştırma yönteminin Türkçe ve İngilizce en iyi deneysel sonuçlarına göre karşılaştırılması	82
Şekil 10.2 K sıkıştırma yönteminin Türkçe ve İngilizce en iyi deneysel sonuçlarına göre karşılaştırılması	83
Şekil 10.3 2-GRAM sıkıştırma yönteminin Türkçe ve İngilizce en iyi deneysel sonuçlarına göre karşılaştırılması	84

ÇİZELGE LİSTESİ

Çizelge 2.1 Veri sıkıştırma yöntemlerinin gruplandırılması	4
Çizelge 3.1 Sembol – Bit katarı haritası	13
Çizelge 3.2 Huffman başlık yapısı	14
Çizelge 5.1 Bükümlü kategoriler ve İngilizce ekler.....	23
Çizelge 5.2 Çekim eki kategorileri ve İngilizce ekler	23
Çizelge 5.3 Birkaç İngilizce yapım eki örnekleri.....	24
Çizelge 8.1 Külliyat-I için sıkıştırma oranları (%).....	32
Çizelge 8.2 Külliyat-I için toplam sıkıştırma süreleri (sn)	33
Çizelge 8.3 Külliyat-II için sıkıştırma oranları (%)	35
Çizelge 8.4 Külliyat-II için toplam sıkıştırma süreleri (sn).....	36
Çizelge 8.5 Külliyat-III için sıkıştırma oranları (%)	38
Çizelge 8.6 Külliyat-III için toplam sıkıştırma süreleri (sn).....	39
Çizelge 8.7 Külliyat-IV için sıkıştırma oranları (%).....	41
Çizelge 8.8 Külliyat-IV için toplam sıkıştırma süreleri (sn)	42
Çizelge 8.9 Külliyat-V için sıkıştırma oranları (%).....	44
Çizelge 8.10 Külliyat-V için toplam sıkıştırma süreleri (sn).....	45
Çizelge 8.11 Külliyat-VI için sıkıştırma oranları (%)	47
Çizelge 8.12 Külliyat-VI için toplam sıkıştırma süreleri (sn).....	48
Çizelge 8.13 Külliyat-VII için sıkıştırma oranları (%).....	50
Çizelge 8.14 Külliyat-VII için toplam sıkıştırma süreleri(sn)	51
Çizelge 8.15 Külliyat-VIII için sıkıştırma oranları (%)	53
Çizelge 8.16 Külliyat-VIII için toplam sıkıştırma süreleri(sn).....	54
Çizelge 8.17 Orijinal dosya ile sıkıştırılıp açılan dosyanın içeriklerinin karşılaştırılması.....	56
Çizelge 8.18 Külliyat-IX için sıkıştırma oranları (%).....	57
Çizelge 8.19 Külliyat-IX için toplam sıkıştırma süreleri (sn).....	58
Çizelge 8.20 Orijinal dosya ile sıkıştırılıp açılan dosyanın içeriklerinin karşılaştırılması.....	59
Çizelge 8.21 Külliyatlara bağlı olarak elde edilen KET ayrıştırma yöntemi sıkıştırma sonuçları.....	60
Çizelge 8.22 Külliyatlara bağlı olarak elde edilen H ayrıştırma yöntemi sıkıştırma sonuçları	61
Çizelge 8.23 Külliyatlara bağlı olarak elde edilen K ayrıştırma yöntemi sıkıştırma sonuçları	61
Çizelge 8.24 Külliyatlara bağlı olarak elde edilen İET ayrıştırma yöntemi sıkıştırma sonuçları	62
Çizelge 8.25 Külliyatlara bağlı olarak elde edilen FET ayrıştırma yöntemi sıkıştırma sonuçları.....	63
Çizelge 8.26 Külliyatlara bağlı olarak elde edilen SET ayrıştırma yöntemi sıkıştırma sonuçları	63
Çizelge 8.27 Külliyatlara bağlı olarak elde edilen İEH ayrıştırma yöntemi sıkıştırma sonuçları	64

Çizelge 8.28 Külliyyatlara bağlı olarak elde edilen FEH ayrıştırma yöntemi sıkıştırma sonuçları	65
Çizelge 8.29 Külliyyatlara bağlı olarak elde edilen SEH ayrıştırma yöntemi sıkıştırma sonuçları	65
Çizelge 8.30 Külliyyatlara bağlı olarak elde edilen İEK ayrıştırma yöntemi sıkıştırma sonuçları	66
Çizelge 8.31 Külliyyatlara bağlı olarak elde edilen FEK ayrıştırma yöntemi sıkıştırma sonuçları	67
Çizelge 8.32 Külliyyatlara bağlı olarak elde edilen SEK ayrıştırma yöntemi sıkıştırma sonuçları	68
Çizelge 8.33 Külliyyatlara bağlı olarak elde edilen 2-GRAM ayrıştırma yöntemi sıkıştırma sonuçları	68
Çizelge 9.1 Külliyyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek 50 kök ve kullanım frekans değerleri	70
Çizelge 9.2 Külliyyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek 50 hece ve kullanım frekans değerleri	72
Çizelge 9.3 Külliyyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek 50 karakter ve kullanım frekans değerleri	73
Çizelge 9.4 Külliyyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek 50 isim ve kullanım frekans değerleri	75
Çizelge 9.5 Külliyyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek 50 sıfat ve kullanım frekans değerleri	77
Çizelge 9.6 Külliyyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek 50 fiil ve kullanım frekans değerleri	79
Çizelge 10.1 KET sıkıştırma yönteminin Türkçe ve İngilizce en iyi deneysel sonuçlarına Göre karşılaştırılması	82
Çizelge 10.2 K sıkıştırma yönteminin Türkçe ve İngilizce en iyi deneysel sonuçlarına göre karşılaştırılması	83
Çizelge 10.3 2-GRAM sıkıştırma yönteminin Türkçe ve İngilizce en iyi deneysel sonuçlarına göre karşılaştırılması	84

ÖNSÖZ

Yapay Zekâ' nın bir dalı olan Doğal Dil İşleme alanında yapılmış olan bu çalışmanın amacı, Türkçenin biçimbilimsel yapısının sıkıştırmaya uygunluğunu tespit etmektir. Çalışma, farklı biçimbilimsel yapılara bağlı olarak farklı boyutlardaki Türkçe ve İngilizce dokümanların, Huffman sıkıştırma yöntemiyle sıkıştırma oranlarının bulunmasına dayanmaktadır. Bu çalışma, Türkçe ve İngilizce dokümanlar için uygulanan sıkıştırma işlemleri içerisinde, biçimbilimsel ayrıştırma yöntemini kullanan kapsamlı bir çalışma olması ve iki dilin benzer ayrıştırma yöntemleri ile sıkıştırılmasından elde edilen sonuçların karşılaştırılması ile dillerin biçimbilimsel yapılarına göre sıkıştırma işlemlerinden alınan verime ulaşılabilmesi yönünden de önem taşımaktadır.

Çalışmam boyunca hiç eksik etmediği isteklendirme ve zaman yönetimi, yönlendirmesi, esirgemediği yardımı için ve zamanını bana tüm titizliğiyle ayırdığı için Yrd. Doç. Dr. Banu Diri' ye teşekkür ederim.

Bu uzun çalışma boyunca bana karşı hep anlayışlı ve sevgi dolu olan aileme ve sonsuz desteğini daima yanımda hissettiğim Mennan Tekbir' e de çok teşekkür ederim.

ÖZET

İnternet’ in yaygınlaşmasıyla elektronik ortamdaki doküman sayısı oldukça artmıştır. Gittikçe artan bu bilgiye daha kolay ve hızlı erişmek amacıyla metin sıkıştırma önem kazanmaktadır. Son yıllarda, metin sıkıştırma alanında yapılan çalışmaların bir kısmı, dilin biçimbilimsel yapısını kullanmayı amaçlayan çalışmaları kapsamaktadır.

Bu çalışmada, Türkçe ve İngilizce dokümanların sıkıştırılma verimlerinin belirlenmesinde farklı ayrıştırma yöntemleri ve bu yöntemlerin sıkıştırma oranına etkileri araştırılmıştır. Dokümanlar Türkçe ve İngilizcenin biçimbilimsel yapısı kullanılarak ayrıştırılmıştır. Sonraki aşamada ayrıştırılan dokümanlardaki yapılara sıkıştırma işlemi uygulanmıştır. Sonuçta, 17 farklı ayrıştırma tekniği oluşturulmuş ve bunlar ile farklı külliyatlar üzerinde denemeler yapılmıştır.

Kullanılan Türkçe külliyatlar, siyaset, spor, güncel gibi farklı konularda yazan 18 erkek, 4 kadın olmak üzere toplam 22 yazara ait, 900 farklı dokümandan oluşturulmuştur. İngilizce için oluşturulan külliyat, güncel konularda yazan farklı kişilere ait 290 farklı dokümandan elde edilmiştir. Farklı doküman gruplarının, farklı konularda yazılan metinlerin, yazar sayısının ve yazar cinsiyetinin başarıya olan etkisini gözlemek amacıyla 9 farklı külliyat grubu oluşturulmuştur.

Sıkıştırma oranını belirlemede, doküman boyutunun sıkıştırmaya olan etkisini tespit edebilmek amacıyla her külliyatın altında 10 farklı boyuttaki dokümana Huffman sıkıştırma metodu uygulanarak elde edilen performanslar karşılaştırılmıştır.

Yapılan denemelere göre, en başarılı sonuçlar, 1024Kb 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.

Anahtar kelimeler: Türkçenin biçimbilimsel yapısı, Huffman Kodlama, N-Gram, Zemberek, Türkçe dilbilgisi, İngilizce dilbilgisi.

ABSTRACT

With the rapid growth of online information, the number of documents in electronic media is very common increased. Easy and quick access to this information gets more important for the purpose of text compression. In recent years, a portion of the work in the field of text compression, covers study aimed to the morphological structure of the language.

In this study, Turkish and English documents are compressed in the determination of the different decomposition methods and efficiency, this method has been to investigate the effects of compression. Turkish and English documents that are parsed by using morphological structure. The next stage in the parsed document structure is applied to the compression process. As a result, created 17 different parsing techniques with which attempts were made on a different corpus.

Corpus used in Turkish politics, sports, current issues such as by 18 men, 4 women for a total of 22 authors, from 900 different documents were created. Created for the English corpus, belonging to different people by the current issues of 290 different documents have been obtained. Of different documents, written texts in different subjects, the number of writers and authors to observe the effects of gender on the success of the group was created 9 different corpus.

In determining the compression ratio, the effect is to compress the size of the document in order to detect 10 different size at the bottom of each corpus document obtained by applying Huffman compression method performances are compared.

According to testing, the most successful results, 1024Kb in size from files and writing the thesis has been provided from the corpus. Different issues included articles by women writers of the document have been obtained if the failure rates of compression.

Keywords: The morphological structure of Turkish language, Huffman Encoding, N-Gram, Zemberek, Turkish grammar, English gramer.

1. GİRİŞ

20. yüzyılda, bilgi teknolojileri sektöründe, bilginin toplanması, işlenmesi, ayrımı ve dağıtımı alanlarında çok fazla gelişme meydana gelmiştir. Veri teknolojilerindeki büyük gelişimler, veri kapasitesinin, işleme gücünün ve iletişim hızının artmasına yol açmıştır.

Günümüz sistemlerinde veri iletimi artmış ve veri iletişim performansı çok önemli hale gelmiştir. Bu önemin farkında olan uzmanlar, daha yüksek kapasiteli bilgisayar sistemleri ile, daha fazla veri transferi yapabilen iletişim altyapısını oluşturabilmek ve bunun yanı sıra yer ve zamandan kazanç sağlayabilmek için, kaydedilen bilginin boyutunu azaltmaya çalışmaktadırlar.

Zamandan ve yerden kazanç sağlamak için, kayıplı ve kayıpsız yöntemlerle veri üzerinde işlem yapmak anlamına gelen veri sıkıştırma, veri sayısal ortamda daha fazla yer kaplayıp, maliyet artımına yol açtığı ve bunun yanı sıra, belli bir zamanda iletişim kanallarından transfer edilen verinin miktarı söz konusu olduğunda, çok büyük bir önem kazanmaktadır. Bu öneme istinaden veri teknolojilerinde geliştirme yapanlar, yazılım ve donanım elemanları ile verinin özellikleri açısından değişebilen çeşitli sıkıştırma yöntemleri kullanarak verinin boyutunu azaltma çalışmaları yapmaktadırlar. Yapılan çalışmalar, ihtiyaçlar doğrultusunda farklı alanlarda olabilmektedir.

Veri sıkıştırma çalışmalarının üzerinde yoğunlaştığı alanlardan biri olan ve “ana işlevi doğal bir dili çözümleme, yorumlama, anlama ve üretme olan bilgisayar sistemlerinin tasarımını ve gerçekleştirilmesini konu alan bilim ve mühendislik dalı” olarak tanımlanan Doğal Dil İşleme üzerinde, bilgilerin saklanması ve iletimi ile ilgili ortaya çıkan sorunları çözmek için yapılan çalışmalar son zamanlarda artış göstermiştir. Bu kapsamda, her dilde olduğu gibi Türkçe diline özgü bir takım çalışmalar da mevcuttur, fakat çok fazla değildir.

Diri ve Karşılıgil, (Diri ve Karşılıgil) Türkçe dilinin biçimbilimsel yapısından yararlanarak Türkçe bir metin içerisindeki her kelimeyi kök-ek, hece ve karakter gibi farklı imgelere ayırıp, sıkıştırma yapmış ve elde edilen sonuçları karşılaştırmışlardır.

Bu çalışmada Türkçe dilinin biçimbilimsel yapısının sıkıştırılmaya uygunluğu araştırılmaktadır. Bu araştırma esnasında karşılaştırma yapmak amacıyla İngilizce dili için de benzer bir çalışma yapılmıştır. Bu bağlamda doküman, biçimbilimsel yapısına göre internet ortamında geliştirmeye açık olan Zemberek isimli araç ve n-gram isminde bir algoritma ile

ayrıştırılmaktadır. Elde edilen giriş verisinin sıkıştırılması için kayıpsız sıkıştırma algoritmalarından biri olan Huffman Kodlama (Huffman Encoding) yöntemi kullanılmaktadır. Sıkıştırma işleminin sonucunda, giriş verisinin yapısına bağlı olarak sıkıştırma verimi ölçülmektedir.

Kelimeleri hece, kök ve eklerine ayırabilmek için Zemberek isimli program kullanılmıştır. Analiz yapılacak olan metin dosyası bloklar halinde hafızaya çekilmiştir. Zemberek, blok içerisinde bir kelimeyi yakaladığında gerekli algoritmaları çağırarak, bu kelimenin en uygun şekilde kök ve eklerine ayrılmasını sağlamıştır. Bunun yanı sıra kelimeler üzerinde n-gram isimli yöntem ile de ayrıştırma işlemi yapılmış ve sıkıştırma sonuçlarına etkisi araştırılmıştır. İmgelerin sıkıştırma işleminin kodlanması için Huffman sıkıştırma algoritması temel alınarak farklı algoritmalar geliştirilmiş ve bunların verimlerine göre karşılaştırılması yapılmıştır. Ortalama sıkıştırma verimi en yüksek % 61 olmak üzere, kullanılan algoritmaya bağlı olarak değişmektedir. Metin dosyalarının uzunluğu arttıkça daha iyi sıkıştırma verimi alınmıştır.

Tezin ikinci bölümünde, veri sıkıştırmanın anlamı ve özellikleri üzerinde durulmuştur. Üçüncü bölümde entropy kodlama başlığı altında Huffman kodlama ve başlık yapısı anlatılmış, dördüncü bölümde de Türkçe dilinin biçimbilimsel yapısı detaylandırılmıştır. Beşinci bölümde İngilizce dilinin biçimbilimsel yapısı verilmiş ve altıncı bölümde n-gram ayrıştırma yönteminin detayları anlatılmıştır. Yedinci bölümde test ortamının oluşturulması başlığı altında deneysel sonuçların elde edilmesinde kullanılan külliyatların yapısı üzerinde durulmuş, sekizinci bölümde de deneysel sonuçlar başlığı altında Türkçe ve İngilizce dilleri için elde edilen sonuçlar sunulmuş ve karşılaştırmalar yapılmıştır. Dokuzuncu bölümde Türkçenin biçimbilimsel yapısına dayalı istatistiksel değerler üzerinde durulmuş, onuncu bölümde ise Türkçe ve İngilizce dillerinin biçimbilimsel olarak başlığı altında, iki dilin sıkıştırılmasında kullanılan ayrıştırma yöntemlerine bağlı olarak elde edilen sıkıştırma sonuçları karşılaştırılmıştır.

2. VERİ SIKIŞTIRMA

Veri sıkıştırma, eldeki verinin orijinal boyutundan daha az yer tutacak şekilde bir biçimden başka bir biçime dönüştürülmesidir. Bir başka deyişle, verinin saklanacağı sayısal ortamda daha fazla yer tutmasının fiziksel ve maddi açıdan sorun olmasının yanı sıra iletişim kanallarından birim zamanda aktarılan veri miktarını etkilemesi ile de önemlidir.

Veri sıkıştırma, gerçek bağlantı hızının üstünde bilgi iletilmesini sağlar. Veri, özellikle metin ve grafikler, genellikle de birbiriyle aynı, yinelenen kısımlar içerir. Sıkıştırma işlemi ile, yinelenen bilgilerin çok sayıda karakterini birkaç karakterle değiştirme ve verinin yinelenen kısımlarının yalnızca bir kopyasını iletme işlemi yapılır. Bilişim teknolojilerindeki gelişmeler veri sıkıştırma yöntemlerinin hem yazılım hem de donanım elemanları ile gerçekleştirilmesini sağlamaktadır. Veri sıkıştırmanın temel özellikleri aşağıda belirtilmiştir:

- Var olan verinin daha az yer kaplayacak şekilde yeniden düzenlenmesidir.
- Zaman ve boyuttan kazanım sağlar. Bu da doğrudan maliyeti minimize eder.
- Sıkıştırma işlemi, verinin tipi ve yapısına en uygun yöntemin seçilmesi ile uygun değer verimli hale getirilebilir.

Veri sıkıştırma işlemleri temel olarak kayıpsız ve kayıplı sıkıştırma yöntemleri olarak gruplandırılır. Ses, görüntü ve video gibi verilerin sıkıştırma işlemleri kayıplı olurken, dokümanların sıkıştırılması kayıpsız olmaktadır. Bu gruplandırma ve altında yer alan yöntemler Çizelge 2.1’ de verilmiştir. Bu tabloda, yapılan genel grupta altı yer alan yöntemler, içerik tabanlı olarak da sınıflandırılıp, alt gruplara ayrılmışlardır. Bu alt gruplar altında yer alan yöntemler, kullanım sıklıklarına bağlı olarak alınmışlardır.

2.1 Kayıpsız Sıkıştırma Algoritmaları

Kayıpsız sıkıştırma algoritmalarında, sıkıştırılan veriden orijinal veri elde edilebilmektedir. Doküman sıkıştırma gibi işlemlerde, sıkıştırılan veri açıldığında tek bir karakterin bile kaybedilmemesi gerekmektedir. Kaybedilen veri, orijinal verinin bütünlüğünü tamamen bozabilmektedir. Mesela dokümanda yer alan 1.000.000 sayısı için, sıkıştırma işlemi sırasında bir 0 karakterinin silinmesi anlamı bozmaktadır.

Çizelge 2.1 Veri sıkıştırma yöntemlerinin gruplandırılması

Veri sıkıştırma yöntemleri		
Kayıpsız	Teori	Entropy, Complexity, Redundancy
	Entropi kodlama	Huffman, Adaptive Huffman, Arithmetic (Shannon-Fano, Range), Golomb, Exp-Golomb, Universal (Elias, Fibonacci)
	Sözlük kodlama	RLE, DEFLATE, LZ Family (LZ77/78, LZSS, LZW, LZWL, LZO, LZMA, LZX, LZRW, LZJB, LZT)
	Diğerleri	CTW, BWT, PPM, DMC
Ses	Teori	Convolution, Sampling, Nyquist-Shannon theorem
	Ses kodlayıcı	LPC (LAR, LSP), WLPC, CELP, ACELP, A-law, μ -law, MDCT, Fourier transform, Psychoacoustic model
	Diğerleri	Dynamic range compression, Speech compression, Sub-band coding
Görüntü	Terimler	Color space, Pixel, Chroma subsampling, Compression artifact
	Yöntemler	RLE, Fractal, Wavelet, EZW, SPIHT, DCT, KLT
	Diğerleri	Bit rate, Test images, PSNR quality measure, Quantization
Video	Terimler	Video Characteristics, Frame, Frame types, Video quality
	Video kodlayıcı	Motion compensation, DCT, Quantization
	Diğerleri	Video codecs, Rate distortion theory (CBR, ABR, VBR)

Kayıpsız sıkıştırma algoritmalarına, Static Huffman Encoding, Adaptive Huffman Coding, Arithmetic Coding, LZW vs. yöntemleri örnek olarak gösterilebilmektedir[1].

Huffman kodlama yöntemi, var olan sembollerin olasılıklarına göre küçükten büyüğe doğru sıralanması ile başlamaktadır. Daha sonra oluşturulan bir ağaca, bu semboller her biri bir yaprak olacak şekilde aşağıdan yukarıya doğru yerleştirilmektedirler. Bu işlem her adımda yapılmakta olup, her adımda en düşük olasılıklar toplanıp ağaca yerleştirilmekte ve bu sembollerin olasılık değerleri listeden silinmektedir. Silinen semboller yerine onları ifade edebilecek yardımcı semboller eklenmektedir. Bu yardımcı semboller, elde edilen olasılık toplamalarının negatifi şeklinde olmaktadır. Listede sadece bir tane yardımcı sembol

kaldığında ağaç tamamlanmaktadır. Daha sonra sembollerin kodlarını ifade etmek için elde edilen bu ağacın tersi alınır. Burada herhangi bir x sembolünün Huffman kod uzunluğu(KU) Eşitlik 1' deki formülün sonucundan daha küçük veya eşit olmaktadır(Diri).

$$KU(x) \leq \lceil -\log_2 P_i \rceil \quad (1)$$

Shannon-Fano kodlaması, Huffman kodlamasına benzemekte, sadece kodlama ağacının oluşturulması aşamasında farklılık göstermektedir. Bu yöntemde de, sembollerin kullanım sıklığı olasılıkları hesaplanır ve olasılığı en büyük olan yukarıda olacak şekilde, büyükten küçüğe doğru sıralama gerçekleştirilir. Kodlama ağacının oluşturulması için ilk önce kümedeki karakterler iki alt kümeye bölünür. Her iki kümede bulunan sembollerin olasılıklarının toplamının eşit olmasına dikkat edilir. Burada önemli olan olasılıklar toplamıdır. Yukarıda kalan alt küme içinde bulunan sembollere atanacak kodun ilk bitine 0, aşağıda kalan alt küme içinde bulunan sembollere atanacak kodun ilk bitine 1 atanır. Daha sonra bu iki alt küme de aynı şekilde alt kümelerle bölünerek sembollere atanacak kodların bit haritaları elde edilir. Bölme işlemi yapılmasına alt kümelerde 1 sembol kalana kadar devam edilir. Huffman kodlaması ile Shannon-Fano kodlaması, ağacın oluşturulması aşaması dışında birbirine çok benzeyen algoritmalarlardır. Eğer sembollerin kullanım sıklığı olasılıkları 2'nin eksi kuvvetine sahip ise Shannon-Fano kodlaması iyi sonuç vermektedir[7].

Aritmetik kodlama yönteminde, Huffman ve Shannon-Fano kodlamalarında olduğu gibi bir kodlama ağacı oluşturulmasına ihtiyaç duyulmamaktadır. Ancak benzer olarak sembollerin kullanım sıklıklarının bilinmesi gerekmektedir. Aritmetik kodlamanın temelinde, kodlanacak veri parçasının 0 ile 1 aralığında gerçek bir sayı olarak temsil edilmesi bulunmaktadır. Veri parçasındaki her sembol bu gerçek sayı aralığını daraltır. Aralığın daralması ile doğru orantılı olarak, ilgili sembolü temsil etmek için gerekli olan bit sayısı da artar. Kullanım sıklığı olasılığı küçük olan semboller bu aralığı hızlı bir şekilde daraltırken, kullanım sıklığı olasılığı büyük olanlar daha az daraltırlar. Eğer veride kullanım sıklığı olasılığı büyük olan sembollerin sayısı çok ise sayı aralığı az daralacağı için, veri az sayıda bit ile kodlanabilir. Kodlama işlemine geçmeden önce, kümedeki semboller, olasılıkları, toplanmış olasılıkları(kendi ve kendinden önceki sembollerin olasılıkları toplamı şeklinde kümülatif olarak) ve her sembolün gerçek sayı temsiline yer aldığı bir çizelge oluşturulur. Kodlama işlemi bu çizelgedeki değerlerin aritmetik işlemlerde kullanılması ile gerçekleştirilir[8].

RLE(Run Length Encoding) algoritmasında, birbirini takip eden sembollerin hepsini tutmak yerine, sadece sembolü ve kaç defa tekrar ettiğini tutma mantığı kullanılmaktadır. RLE

algoritması yüksek sıkıştırma oranlarına ulaşamasa da, kolay uygulanabilir olması ve hızlı çalışması nedeniyle BMP, TIFF(Tagged Image File Format) ve PCX(Personal Computer Exchange) gibi görüntü dosya formatlarının çoğu tarafından desteklenmektedir[9].

LZ77 algoritmasında pencere olarak adlandırılan bellek alanı ve katar uyuşma yordamı kullanılır. Kodlanacak veriye ait karakterler bu pencere içine dâhil edilir. Herhangi bir t anında, kodlanacak veriye ait bir sembolün pencere içinde olup, olmadığına bakılır. Eğer bu karakter pencere içinde yer alıyorsa, bulunduğu adres(m), ve uyuşma gösteren sembolün uzunluğu(karakter uzunluğu)(x) kodlamada kullanılır. Bu iki değişken(adres, karakter uzunluğu), (m, x) biçiminde gösterilir. Bu gösterim şekli, uyuşma gösteren sembolü işaret ettiğinden, işaretçi olarak adlandırılır. Örneğin işaretçi değeri (32, 22) ise, 32. karakterden sonra 22 karakter uzunluğunda; (11,1) ise, 11. karakterden sonra 1 karakter uzunluğunda uyuşma olduğu durumu anlaşılır. LZ algoritmasında pencere boyu büyüdükçe sıkıştırma başarısı artar, ancak pencere boyunun büyümesine bağlı olarak, sembollerin ifadesinde kullanılan işaretçiler için gerekli bit sayısı artacağı durumu da göz önünde bulundurulduğunda belirli bir boyuttan sonra artırma devam etmenin verimli olmadığı görülür. Pencere boyunun büyümesi, uyuşma yordamının uyuşan sembolü bulması için oluşturacağı çevrim sayısını artıracağından, kodlama zamanını da artırır. Pencere içinde uyuşma gösteren sembolün etkin bir şekilde aranması ve pencere boyutunun ne olacağı birçok çalışmaya kaynak olmuştur. Öyle ki, LZ77 algoritmasının türevleri olan algoritmaların hemen hepsinin temelini, sıkıştırma başarısını azaltmadan, pencere ve uyuşma yordamının etkin bir şekilde kullanılması oluşturmuştur. Bu algoritmaların çoğunda pencere ve uyuşma yordamı arasında denge kurmak amacıyla pencere sabit tutulur ve hızlı arama yapabilmek için doğrusal dizi yerine ikili ağaç, sözlük ağacı(trie tree) veya çırpı fonksiyonu(hash function) kullanılır[10].

LZ78 algoritmasında, sözlük olarak adlandırılan bellek alanı, sözcükleri heceleme ve katar uyuşma yordamları kullanılır. Veri içerisinde yer alan semboller, kodlama işlemi anında, heceler ve sözcükler şeklinde sözlükte saklanır ve onlara birer kod verilir. Herhangi bir t anında, sembol düzeyinde gelen veri parçası(s) sözlükte aranır. Eğer bulunursa(uyuşma gösteriyorsa) bir sonraki sembol(A) o anki sözcüğe(s) eklenerek yeni sözcük(sA) oluşturulur ve arama işlemine devam edilir. Bu işlem en uzun sözcüğü bulana kadar sürdürülür. Yeni oluşturulan sözcük (sA) sözlükte yok ise, bir önceki sözcüğün kodu kullanılır ve yeni sözcük kendisine yeni kod eklenerek sözlüğe eklenir. Bu işlem sözlük dolana kadar sürer. Sözlük dolduktan sonra yeni eklemeler yapılmadan kodlamaya devam edilip, sözlük doldurulur(FLUSH yöntemi olarak anılır). Fakat LZ78'in türevleri olan algoritmalarda

sözlük doldurulmadığı zaman sıkıştırma başarısının arttığı görülmektedir. Sözlüğü doldurma yerine boşaltıp yeniden baştan başlama, iki sözlük oluşturulup doldurulduğunda diğeriyle kodlama işlemine geçme, sözlük dolduğunda az kullanılanı çıkarıp yerine yeni sözlük ekleme gibi yöntemler ile daha iyi sıkıştırma başarısı elde edilmektedir. LZ78’de sözlüğün kurulması için sözlük ağacı kullanılırsa zaman maliyeti iyi olur; ancak gerekli bellek alanı çok büyür[10].

LZW algoritması, hızlı uyarılama için tasarlanmış olmasına rağmen, verinin sınırlı analizini yapması nedeniyle, bu konuda başarısı en iyi olan algoritma değildir. Bu algoritma, sıkıştırılan metinden bir kelime çeviri tablosu oluşturur. Kelime çeviri tablosu sabit uzunluktaki kodlarla kelimeleri eşler. Kelime çeviri tablosu 8 biti temsil eden sembol değerleri ile başlangıç durumuna getirilir. LZW algoritması ile tüm metin seri bir şekilde taranır. Karşılaşılan veri dizeleri, kelime 20 çeviri tablosunda tutulur. Aynı veri dizesiyle karşılaşıldığında veri dizesi yerine, sadece dizinin tablodaki numarası kaydedilir[11].

2.2 Kayıplı Sıkıştırma Algoritmaları

Kayıplı sıkıştırmanın günlük hayattaki kullanımı insan algı mekanizmalarının algılayamayacağı bilgiyi verinin içinden temizlemeye dayanmakta ve sıkıştırılan veriden orijinal veri elde edilememektedir[1].

Görüntü sıkıştırma işlemlerinde kullanılan JPEG(Joint Photographic Experts Group) standardında olduğu gibi, bir resimdeki çok ince detaylar insan gözünün çözümleyemeyeceği düşünülerek atılabilmektedir. Yapılan bu eksiltme, ana hatlarda bozulmalara yol açmadığı sürece göz ardı edilebilmektedir. Burada önemli olan görüntüdeki kayıpların göz tarafından fark edilir derecede şikâyet oluşturmamasıdır[12].

Ses ve video sıkıştırma algoritması olarak MPEG(Moving Pictures Experts Group) standardı verilebilmektedir. Bir şarkıdaki çok tiz ve çok kalın seslerin insan kulağı tarafından algılanamayacağı düşünülerek vazgeçilebilmektedir. Bunlar gibi, insan algısının sınırlarını dikkate alan pek çok sıkıştırma yöntemi bulunmakta ve örneklerini her ‘jpg’, ‘gif’ uzantılı resim, ‘mpg’, ‘mp2’, ‘mp4’ uzantılı video, her Mp3 uzantılı ses dosyasına baktığımızda görebilmekteyiz[13].

Kayıplı sıkıştırma tekniklerinde, sıkıştırma oranı artırıldığında orijinal veride meydana gelen bozulmalar arttığından, kalitede düşme meydana gelmektedir. Örneğin sıkıştırılmış resim formatı olan JPEG dosyalarının kaliteli ya da az kaliteli olmasının nedeni sıkıştırma katsayısı

olmaktadır. Yani benzer iki resim dosyasından daha az disk alanı kaplayan daha kötü kalitededir denebilmektedir[12].

Kayıplı sıkıştırma algoritmalarına, JPEG standardında yer alan resimdeki piksellerin azaltılması(downsampling), nicemleme(quantization) vs. gibi yöntemler örnek olarak gösterilebilmektedir[12].

Kullanılan kayıpsız sıkıştırma yönteminin verimini ifade etmek için çok sayıda yöntem kullanılmaktadır. Bu yöntemlerden biri Eşitlik 2' deki formülde belirtilmiştir.

$$\text{Sıkıştırma Oranı} = \frac{\text{Çıkış dosyasının boyutu}}{\text{Giriş dosyasının boyutu}} \quad (2)$$

Bu tezde de bazı uygulamalarda olduğu gibi, yüzdesel sıkıştırma oranı $100 \cdot (1 - \text{Sıkıştırma Oranı})$ şeklinde ölçülür.

3. ENTROPY KODLAMA

Entropy kodlamada en çok bilinen iki yöntem Huffman ve Shannon Fano kodlamasıdır[14]. Bu bölümde tez kapsamında kullanılan Huffman kodlaması ve Entropy kavramından bahsedilecektir.

3.1 Entropy

Bilgi teorisinde *Entropy*(E) ve *Redundancy*(*Artıklık*(R)) önemli kavramlardır. a_i gibi tek bir sembolün entropy' si, $-P_i \log_2 P_i$ olarak ifade edilmektedir. P_i ise, veri içerisinde a_i sembolünün olma olasılığı olarak tanımlanmaktadır. a_i sembolünün entropy' si, bu sembolü ifade etmek için ortalamanın üzerinde ihtiyaç duyulan en az bit sayısı(m) Eşitlik 3' deki gibi ifade edilmektedir(Diri).

$$m = - \sum_{i=1}^n P_i \log_2 P_i \quad (3)$$

Eşitlik 3 ile ortalamanın üzerinde ihtiyaç duyulan en az bit sayısı ifade edilmiştir. a_i sembolünün entropy' sine bağlı olan P_i ifadesi, tüm n olasılıkları eşit olduğunda en küçük değerini almaktadır. Bu gerçek verideki R artıklığını tanımlamak için kullanılmaktadır. Bu ifade entropy ve en küçük entropy arasındaki farklılık Eşitlik 4' deki gibi tanımlanmaktadır.

$$R = \left(- \sum_{i=1}^n P_i \log_2 P_i \right) - \log_2 \left(\frac{1}{P} \right) = - \sum_{i=1}^n P_i \log_2 P_i + \log_2 n \quad (4)$$

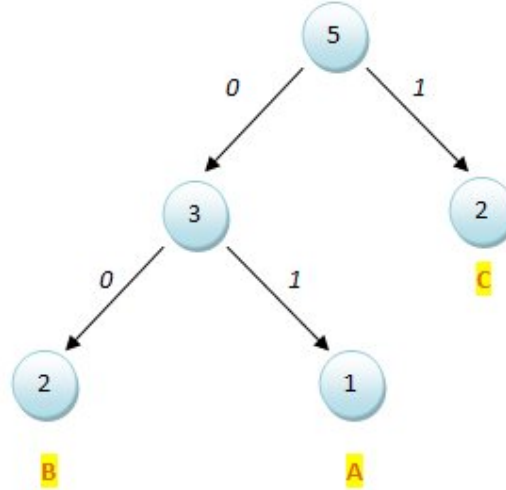
3.2 Huffman Kodlama

Huffman kodlama, kayıpsız veri sıkıştırma yapabilmek için yaygın olarak kullanılan bir entropy kodlama yöntemidir. Kişisel bilgisayarlarda bulunan çok sayıdaki program içerisinde bu sıkıştırma yöntemi kullanılmaktadır. Huffman kodlama, bu programların bazıları için temel yöntem olmakta, bazıları için ise kullanmış oldukları çok adımlı sıkıştırma işlemleri içerisinde, belirli bir adımda kullanılmaktadır[2].

Huffman kodlaması 1952 yılında D. Huffman tarafından geliştirilmiştir. Kodlamanın prensibi, kaynak mesaja bir kod kelimesi atamak, bu işlemi gerçekleştirirken de kaynak içerisindeki sembollerin kullanım sıklıklarını dikkate almaktır. Bu yöntem ile ortalama % 25 oranında

sıkıştırma yapılabilmekte, hatta bazı durumlarda bu oran % 50 - % 60' lara ulaşabilmektedir(Gibson).

Huffman kodlama yönteminde, veri içerisindeki karakter ya da sembollerin kullanım sıklıklarına(frekanslarına) göre bir ağaç oluşturulur. Ağacın en tepesinden aşağıya doğru ilerlerken sola ayrılan dal için 0, sağa ayrılan dal için 1 kodu verilir. Şekil 3.1 örnek bir huffman ağaç yapısını göstermektedir.



Şekil 3.1 Örnek bir Huffman ağacı.

Şekil 3.1' de yer alan daireler içindeki sayılar, karakterlerin kullanım sıklıklarını(frekanslarını) ifade etmektedir. Eğik rakamlar ise bit kodlarını ifade etmektedir. Bu ağaç, "ABC" karakterlerinden oluşan bir veri kümesi için üretilmiştir. Ağaca göre karakterler için bit haritaları şu şekildedir:

B: 00

A: 01

C: 1

Oluşturulan bit haritaları, karakterlerin veri içerisindeki konumlarına göre yerleştirilir. Ortaya çıkan bit haritası sıkıştırılmış veridir. Örneğin; "BAACC" verisi, elde edilen bit haritalarına göre yeniden düzenlenirse;

00 01 01 1 1 = 00010111 = 17H

elde edilir. Yani 1/5 oranında bir sıkıştırma elde edilmiş olur.

Ağacın oluşturulması aşamasında, ilk önce karakterlerin kullanım sıklıkları(frekansları) hesaplanır. Örneğin elimizdeki veri “ABBCC” olsun,

$A: 1$

$B: 2$

$C: 2$

$A: 1 \quad B: 2 \quad C: 2$

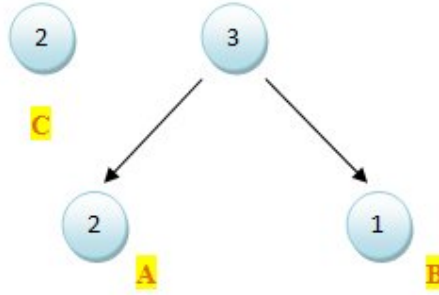


Şekil 3.2 Ağaç oluşturma ilk aşama

En küçük iki frekans toplanır ve frekans tablosu yeniden düzenlenir.

$A: 1 + B: 2 \quad C: 2$

$C: 2 \quad BA: 3$

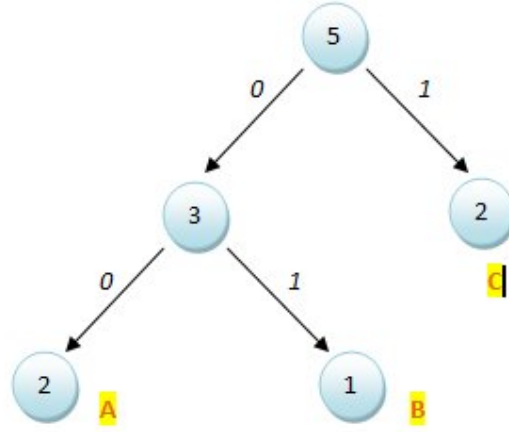


Şekil 3.3 Ağaç oluşturma ikinci aşama

Tek bir ağaç oluşturulana kadar sürekli en küçük frekanslar toplanır.

$C: 2 + BA: 3$

$CBA: 5$



Şekil 3.4 Ağaç oluşturma üçüncü aşama

Huffman algoritması, az sayıda karakter çeşidine sahip ve büyük boyutlardaki verilerin sıkıştırılmasında çok kullanışlı olmaktadır[2].

3.3 Statik Huffman Kodlama

Statik Huffman kodlama, Dinamik Huffman kodlama ile benzer mantığa sahiptir. Bu iki yöntem arasındaki fark sadece elde edilen Huffman sembol kodlarının gelen dokümana göre değişiklik göstermemesidir. Huffman sembol kodlarının elde edilmesi için oluşturulan geniş bir veri kümesinden yararlanılmaktadır. Bu veri kümesindeki sembollerin Huffman kodları çıkarılmakta ve kaydedilmektedir. Sıkıştırılmak istenen doküman ele alındığında Huffman kodunu doküman içeriğine bağlı olarak tekrar oluşturma bölümü yapılmamakta, sembolleri kodlamak için var olan Huffman kodları sembol eşleştirmesi ile alınıp kullanılmaktadır.

3.4 Huffman Kodu Çözümleme

Kod çözücü, semboller için oluşturulan Huffman ağacının en başına erişerek gelen bit katarının başlangıcının nerede olduğunu bilmek zorundadır. Huffman ağacının oluşturulma şekli bilindikten sonra geriye kalan bit katarının çözülmesi oldukça basit bir algoritmaya dayanmaktadır. Ağacın kökünden başlanır ve bit katarının ilk bit değeri okunur. Eğer bu bitin değeri sıfır ise, aşağıdaki daldan devam edilir. Eğer bu bitin değeri bir ise yukarıdaki daldan devam edilir. Bir sonraki bit okunur ve ağacın yapraklarına doğru başka bir dala yönelir. Kod çözümleyici bir yaprağa vardığında, yaprakta bulunan sembolün değeri bulunmuş olunur (LNR). Bu şekilde bir sonraki bitin kökünden başlayarak aynı işleyiş süregelir.

Örneğin; elimizdeki veri $17h$ olsun. $17h$ sayısının 16 tabanında olduğu bilinmektedir. 16 tabanındaki 17 sayısının 2 tabanındaki karşılığı $(00010111)_2$ bit dizisini oluşturur. Elimizdeki ağaç yapısının da Şekil 3.4’ deki gibi olduğu bilinmektedir. Bu ağaç yapısından yola çıkarak Tablo 3.1’ deki sembol – bit katarı haritası elde edilir.

Çizelge 3.1 Sembol – Bit katarı haritası

Sembol	Bit Katarı
A	00
B	01
C	1

Elde edilen bit katarının ilk karakterinden başlanarak Tablo 3.1’ de yer alan haritadan yararlanılıp bit katarı – sembol eşleştirmesi yapılır. Bit katarının ilk biti olan 0 değeri okunur. 0 değerine sahip bir sembolün varlığı aranır. Olmadığı için bir sonraki bit olan 0 değeri okunur. Elimizdeki bit katarı 00 elde edilir. 00 değerine sahip sembol aranır ve A sembolü ile eşleşme sağlandığı bulunur.

$00 \rightarrow A$

Çözümlenen sembol dizisi: A

Eşleşme sağlandığı için işleme sıradaki bitten devam edilir. Bit katarının sıradaki biti olan 0 değeri okunur. 0 değerine sahip bir sembolün varlığı aranır. Olmadığı için bir sonraki bit olan 1 değeri okunur. Elimizdeki bit katarı 01 elde edilir. 01 değerine sahip sembol aranır ve B sembolü ile eşleşme sağlandığı bulunur.

$01 \rightarrow B$

Çözümlenen sembol dizisi: AB

Eşleşme sağlandığı için işleme sıradaki bitten devam edilir. Bit katarının sıradaki biti olan 0 değeri okunur. 0 değerine sahip bir sembolün varlığı aranır. Olmadığı için bir sonraki bit olan 1 değeri okunur. Elimizdeki bit katarı 01 elde edilir. 01 değerine sahip sembol aranır ve B sembolü ile eşleşme sağlandığı bulunur.

$01 \rightarrow B$

Çözümlenen sembol dizisi: *ABB*

Eşleşme sağlandığı için işleme sıradaki bittten devam edilir. Bit katarının sıradaki biti olan *I* değeri okunur. *I* değerine sahip bir sembolün varlığı aranır ve *C* sembolü ile eşleşme sağlandığı bulunur.

$I \rightarrow C$

Çözümlenen sembol dizisi: *ABBC*

Eşleşme sağlandığı için işleme sıradaki bittten devam edilir. Bit katarının sıradaki biti olan *I* değeri okunur. *I* değerine sahip bir sembolün varlığı aranır ve *C* sembolü ile eşleşme sağlandığı bulunur.

$I \rightarrow C$

Çözümlenen sembol dizisi: *ABBCC*

Bit katarı dizisinin sonuna gelindiğinde çözümlenen sembol dizisi olarak “*ABBCC*” elde edilmiş olur.

3.5 Huffman Kodlamanın Başlık Yapısı

Sıkıştırılan verinin tekrar eski haline getirilebilmesi için bir başlık bilgisine ihtiyaç duyulmaktadır. Bu Huffman başlık yapısı özel bir formata sahiptir ve Çizelge 3.2’ de görülmektedir.

Çizelge 3.2 Huffman başlık yapısı

HHI	NOB	SPT	SPST	CP	SP
-----	-----	-----	------	----	----

HHI(Header Huffman Initials): Huffman başlık yapısının ilk karakteridir. Her zaman ‘H’ olmalıdır. Bu alanın kullanımı bir zorunluluk olmamakla birlikte, oluşturulan yeni dosyanın(sıkıştırılmış verinin gönderildiği dosya) başlangıcının ifade etmektedir.

NOB(Number Of Bits In Code Values): CP(Code Part) içinde bulunan bit sayısını belirtmektedir. Örneğin, CP “10010” olduğunda, boyut 5 olmaktadır. NOB/8 değeri CP

içindeki byte sayısını vermektedir. Verilen örnek için ofset değeri 1 olmaktadır. 1 byte başlık yapısından okunmakta ve bit değerine dönüştürülmektedir(Bit değeri “10010000” olacaktır). NOB değeri alınarak gerçek CP bulunmuş olmaktadır(“10010000” ın ilk 5 biti “10010” dır).

SPT(Symbol Part Type): Bu alanda, bütün sembollerin uzunluğu eşit olduğunda(fixed length symbols) “10”, bütün sembollerin uzunluğu eşit olmadığına ise “100” değeri yer almaktadır.

SPST(Symbol Part Special Char1): Bu alan SP “10” olduğunda, sembol uzunluğunu ifade etmektedir. SP “100” olduğunda ise Symbol Part Delimiter olarak işlevsellik kazanmaktadır.

CP(Code Part): Bu alan LNR(Left-Node-Right) metodu ile çıkarılmış Huffman ağacıdır.

SP(Symbol Part): Bu alan CP içindeki sembollerin değerlerini ifade etmektedir.

Sıkıştırılmış metin dosyasının uzantısı .huff olarak tanımlanmıştır.

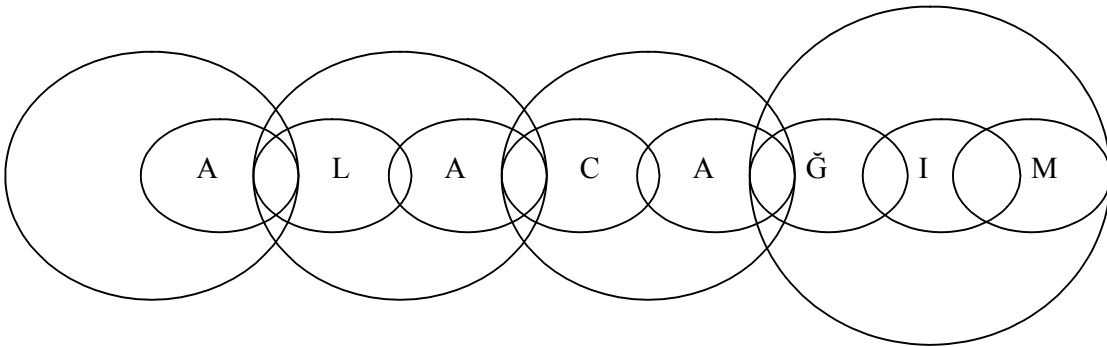
4. TÜRKÇENİN BİÇİMBİLİMSEL YAPISI

Bu çalışmada Türkçenin karakter, hece, kök, ek, isim, sıfat, fiil yapılarına göre giriş metninin ayrıştırması yapıldığından Türkçe biçimbilimsel yapısı denilince akla gelen tüm yapılar yerine, tez çalışmasında kullanılan bu yapılar üzerinde durulmuştur.

Biçimbilgisi, bir dilin kelimelerini, kelime yapılarını, türeme yollarını, çekim bilgilerini inceleyen dilbilgisi koludur. Her kelimenin bir ses yapısı, bir de biçim yapısı bulunmaktadır. Kelimenin ses yapısı denildiğinde akla öncelikli olarak kelimenin heceleri, hece yapıları gelmektedir. Seslerin öbekleşmesinden, bir ya da birkaç fonemin bir araya gelmesinden heceler ve kelimeler oluşmaktadır(Ediskun).

4.1 Heceler

Bir solukta çıkan bir tek ya da bileşik sese hece denmektedir. Örneğin, *bağımızdakilerden* kelimesinin hecelerine ayrılmış şekli *ba/ğı/mız/da/ki/ler/den* olarak gösterilebilmektedir. Hecenin niteliği anlamlı olup, olmamasında değil, bir solukta söylenip söylenmemesindedir. Bir hecenin boğumlanmasında üç basamak bulunmaktadır. Konuşmada, bir hecedeki fonemlerin tespih taneleri gibi değil, bir zincirin halkaları gibi dizildikleri görülmektedir. Konuşma organlarımız birinci fonemi çıkarma durumundan çözülüp gevşeme basamağına geçmeden ikinci fonemin gerilme basamağına sıçrarlar. Birinci fonemin gevşeme basamağı ile ikinci fonemin gerilme basamağı aynı anda olmaktadır. Böylece, fonem öbeğini yani heceyi oluşturan fonemler, bir tek solukta söyleniyormuş gibi, bir çırpıda çıkarılırlar(Ediskun). Bu yapının örneklendirilmesi Şekil 4.1' de görülmektedir.



Şekil 4.1 Heceyi oluşturan fonemlerin heceyi oluşturma şekilleri(Ediskun)

Türkçe’ de, her hecede bir tek ünlü fonem bulunmaktadır. Ünsüz fonemler, bir tek ünlü fonem çevresinde öbkleşerek heceleri oluşturmaktadırlar. Her ünlü fonem, kendinden önceki ünsüz fonemi kendi hecesi içine almaktadır. Örnek olarak, *el-im*, *be-şer*, *dört-er*, vs. verilebilmektedir.

4.2 Kelimeler

Kelimelerin biçim yapısı ise, kelimenin anlamlı ve anlamsız ögeleridir. Kelimelerde biçim yapısı bakımından anlamlı ögeler ve anlamsız ya da görevli ögeler olmak üzere iki öge bulunmaktadır.

4.3 Kökler

Kök, bir kelimenin anlamı ve yapısı bozulmadan, daha küçük parçaya ayrılmayan ögesidir. Örneğin, *dişçilik*, bir kelimedir; bunun sonundaki *-lik*’ i ve *-çi*’ yi çıkarabiliriz. Ama geriye kalan *diş* kelimesini, anlamı ve yapıyı bozmadan parçalayamayız. Bu nedenle, *diş* kelimesine bir heceli kök denir. *Gölgesizlik*, bir kelimedir; bunun sonundaki *-siz*’ i, *-lik*’ i çıkarabiliriz. Ama *gölge* kelimesini artık parçalayamayız. Bunun için, *gölge* kelimesine de iki heceli kök denir. Kökler, anlamları bakımından isim kökleri ve fiil kökleri olmak üzere ikiye ayrılırlar.

4.4 Ekler

Morfemler yani ekler, anlamları olmayan, ancak kelimelerden, kelimeler türetmekle ya da kelimelere durum, ilgi, nicelik, kip, zaman ve kişi ile ilgili birer anlam katmakla görevli dilbilgisi ögeleridir. Örneğin, *-i*, *-e*, *-de*, *-den*, *-ler*, *-in*, *-ci*, *-gin*, *-gen*, *-yor*, *-im*, *-m* gibi ögeler, ya kelime türetmeye ya da kelimeleri cümlelerin anlamına göre biçimlendirmeye yararlar. Dilbilgisi açıklamalarında ekleri göstermek için, bir ekin çeşitli biçimlerinden herhangi biri yalnızca e, i ünlü harflerini içeriyor ise ve bu ek ünsüz bir harfle başlıyorsa ve de bu harfin hem yumuşak, hem de sert ünsüzü var ise, o zaman yumuşak olanını söylemek ya da yazmak gerekir. Ekler görevleri bakımından, çekim ekleri ve yapım ekleri olmak üzere ikiye ayrılırlar

4.5 İsimler

Varlıkları ya da eylemleri bildiren, tanıtan kelimeye isim denmektedir. Örneğin, *sokak*, *bulut*, *düşünce*, *Osman*, *Ayla* birer varlık ismi, *uyumak*, *gölüş*, *anlatma* ise birer eylem isimleridirler. İsimler, varlıklara verilişlerine, oluşlarına ve sayılarına göre olmak üzere bazı bölümlere

ayrılmaktadırlar. Bu bölümler özel isim, cins isim, somut isim, soyut isim, tekil isim, çoğul isim ve topluluk isim şeklinde gruplandırılabilirler.

4.6 Sıfatlar

Varlıkları niteleyen ya da belirten kelimelere sıfat denmektedir. *Mavi gök, beyaz bulutlar, çalışan çocuk, bu kitap, birinci sınıf, bir milyon insan, altışar elma* kelime gruplarında yer alan *mavi, beyaz, çalışan, bu, birinci, bir, altışar* kelimeleri sıfatlara örnek olarak verilebilirler. Bu örneklerle göre, *mavi, beyaz, çalışan* kelimeleri *gök' ün, bulutlar' ın, çocuk' un* nasıl olduklarını, yani vasıflarını göstermektedirler. Bu, *birinci, bir milyon, altışar* kelimeleri ise *kitap, sınıf, insan* ve *elma* varlıklarının yerlerini ya da sayılarını belirtmektedirler. Sıfatlar anlamları bakımından, niteleme sıfatları ve belirtme sıfatları olmak üzere ikiye ayrılmaktadırlar.

4.7 Fiiller

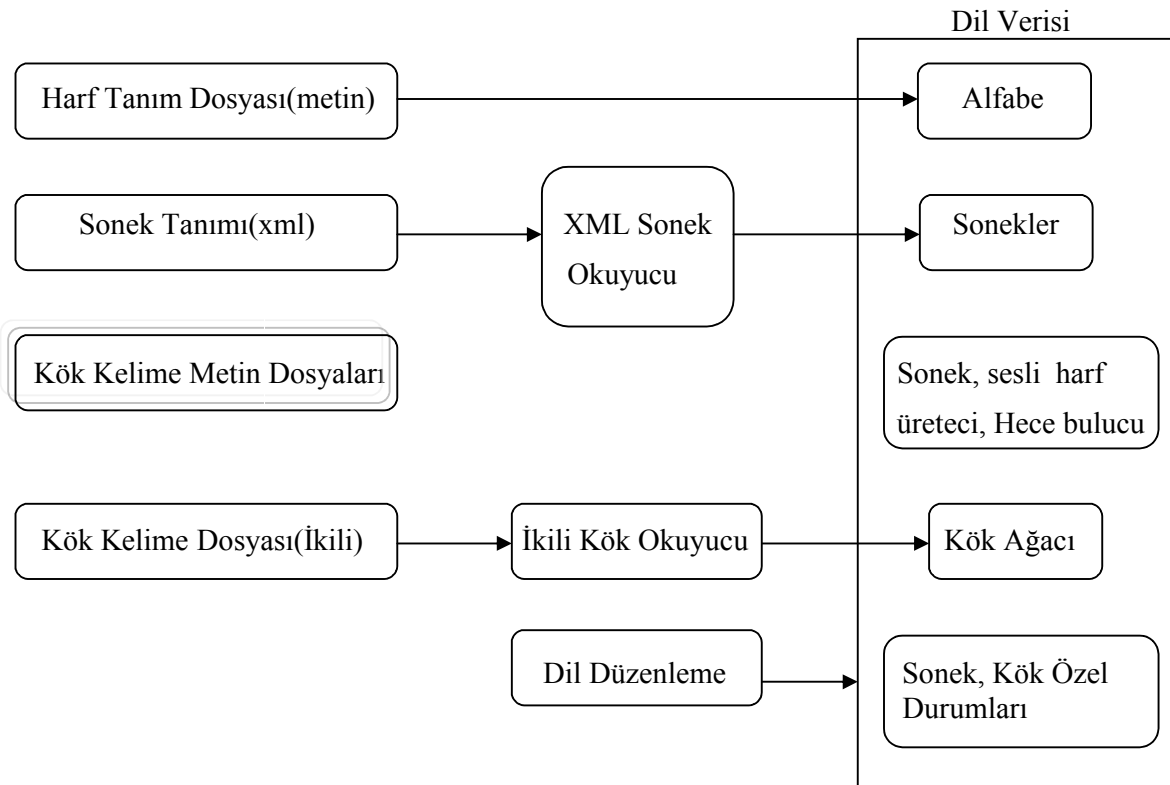
Fiil, varlıkların yaptıkları işleri zamana, biçime ve kişiye bağlayarak anlatan kelimelere verilen adlandırmadır. Fiiller, bulunduğu cümlelerin temellerini oluşturmaktadırlar. Her fiilde, eylem zaman ya da biçim, kişi kavramları kaynaşmış bir halde bulunurlar. Örneğin, *döndürdüm, bakmalısın* fiillerinde eylem kavramlarını gösteren kök ya da gövdeler *döndür-* ve *bak-* tır. Zaman ya da biçim kavramlarını gösteren ekler, *-dü*(geçmiş zaman), *-malı*(gerekliklik, tarz)' dır. Kişi kavramı gösteren ekler, *-m*(birinci tekil kişi), *-sın*(ikinci tekil kişi)' dır.

4.8 Biçimbilimsel Analizin Yapıldığı Araç

Bu çalışmada Türkçe metinleri biçimbilimsel yapılarına göre ayırma işleminde Zemberek isimli açık kaynak kodlu bir program kullanılmıştır.

Zemberek (Önceki adı ile Tspell) projesi, Türkçe diline ilişkin çeşitli bilgi işlem problemlerinin çözülmesi için açık kaynak kodlu, platform bağımsız bir kütüphane oluşturulması amacı ile başlatılmıştır. Proje şu anda Türkçe kelime denetleme, kelime çözümleme, kelime önerme, oluşturma, Türkçe karakter kullanılmadan yazılan yazıların dönüştürülmesi, heceleme gibi işlemleri gerçekleştirmektedir. Yine bu proje kapsamında Zemberek çekirdek kütüphanesi kullanılarak projeler de geliştirilmeye başlanmıştır. Zemberek OpenOffice eklentisi ve Zemberek Sunucusu yakın zamanda hayata geçirilmiştir[3].

Zemberek çalışması, dil yapısı bilgisi ve NLP işlemleri olmak üzere iki ayrı kütüphaneden meydana gelmektedir. Çekirdek kütüphane NLP' ye özgü algoritmaları içermektedir ve dil üzerinde yapılacak olan işlemler için gerekli araçları sağlamaktadır. Çekirdek kütüphane özellikle Türkçe dili için tasarlanmasına rağmen, dile bağımlı herhangi bir düzenleme içermemektedir. Bu esnekliği sağlamak için çok sayıda yardımcı mekanizmalar geliştirilmiştir. Her dil düzenlemesi, önceden tanımlanmış dilbilgisi gerekliliklerine uygun olmak ve gerekli dil verisini sağlamak ile sorumludur. Benzer şekilde, işlemlere bağımlı NLP yeni bir dil düzenlemesi için tekrar değiştirilmek zorunda değildir. Bir dil için düzenleme yapıldığında, çekirdek NLP fonksiyonları bu bilgiyi genel davranış işlemlerinde kullanırlar ve yazılım erişim mekanizmasını kullanabilmek için son kullanıcılara hizmet sağlarlar. Bir Türk dili için yapılacak olan düzenleme oldukça kolaydır. Zemberek uygulamasında, yazılım geliştiricilerin işlerini daha kolay yapabilmeleri için dil verileri, metin tabanlı düzenleme dosyalarından çıkarım yapılmaktadır. Fakat bu çıkarım işlemi genellikle, esnekliği ve performansı kötü yönde etkilemektedir[3]. Bu nedenle, özel durumlar ve son ek üretim mekanizması gibi bilgiler kodun içinde tutulurlar. Şekil 4.2' de dil bilgisi elemanları ve gerekli çıkarım dil bilgisi gösterilmektedir.



Şekil 4.2 Bilgi elemanları ve dilbilgisi çıkarım yapısı(Ediskun)

Elde edilen deneysel sonuçların alınmasında KET(ekler ek halinde olmak üzere kök - ek şekline) isimli ayrıştırma yöntemi ile sıkıştırma işlemi yapıp, tekrar sıkıştırılan metin eski haline getirildiğinde, orijinal metin ile sıkıştırma sonucu açılan metin arasında bazı farklılıklar gözlemlenmektedir. Bu sorunun oluşma nedeni, Zemberek isimli programın, yapmış olduğu çözümleme işlemlerinde, hece düşmelerini göz ardı etmesidir. Örneğin, ‘*bağırına*’ kelimesi çözümleme işlemine tabi tutulduğunda, “*bağır*” kökü ve “*ına*” eki elde edilmektedir[16].

Zemberek ile “*bağırına*” kelimesinin çözümlenmesi:

[Kök:*bağır*, Tip:İSİM | Ekler:İSİM_KÖK, İSİM_TAMLAMA_IN, İSİM_YÖNELME_E]

[Kök:*bağır*, Tip:İSİM | Ekler:İSİM_KÖK, İSİM_TAMLAMA_I, İSİM_YÖNELME_E]

[Kök:*bağır*, Tip:İSİM | Ekler:İSİM_KÖK, İSİM_SAHİPLİK_O_I, İSİM_YÖNELME_E]

[Kök:*bağır*, Tip:İSİM | Ekler:İSİM_KÖK, İSİM_SAHİPLİK_SEN_IN,İSİM_YÖNELME_E]

“*bağırına*” kelimesi için sıkıştırma işlemi sırasında Huffman algoritmasına gönderilen semboller “*bağır*”, “*ın*” ve “*a*” olmakta ve bu semboller ile sıkıştırma işlemi yapılmaktadır. Sıkıştırılan metnin tekrar eski haline dönüştürülmesi için semboller bir araya getirildiğinde, “*bağırına*” kelimesini elde etmek için “*bağır*”, “*ın*” ve “*a*” sembolleri bir araya getirileceğinden, orijinal kelimedenden farklı olarak “*bağırına*” kelimesi elde edilmektedir. Örnekten de anlaşılacağı üzere, hece düşmesi yaşanmış olan kelimelerin sıkıştırılıp, tekrar açılmasında, düşen karakterin metne yansması, eşleşme sorunu yaratmaktadır. Bu sorunu gidermek için sezgisel bir yaklaşım izlenmiş ve Zemberek isimli programın “*öner*” isimli fonksiyonu kullanılmıştır. “*Öner*” fonksiyonunun “*bağırına*” kelimesi için önerdiği sözcükler [*bağırın, bağırına, bağırinsa, bağırınla, bağırını, bağırında, bağırınca, bağırıc, bağırma, bağına, bağlına, bağcına, bağsına, ağırına, çağırına, barına, bağrışa, sağırına, yağırına, bayırına*] olmaktadır. Burada programsal olarak en iyi sonucu alabilmek için yapılan çok sayıda denemeden sonra, gelen dizideki ikinci önerinin alınmasına karar verilmiştir. Fakat bu karar sezgisel olduğundan, doğru öneri ikinci sırada olmadığında, sonuç yine olumsuz olmuştur.

5. İNGİLİZCENİN BİÇİMBİLİMSEL YAPISI

Bu çalışmada İngilizcenin karakter, kök, ek yapılarına göre giriş metninin ayrıştırılması gerçekleştirildiğinden İngilizcenin biçimbilimsel yapısı üzerine bazı açıklamalar yapılmıştır[4].

5.1 Temel Terimler

- Morfem: Dilin anlamlı en küçük birimidir. Örneğin *'items'* kelimesi *'item'* ve çoğul anlam veren *'s'* olan anlamlı iki parçaya bölünebilir.

- Bağımsız Morfem: Bağımsız olarak kendi başına anlam ifade eden morfemdir.

- Bağımlı Morfem: Bağımsız olarak kendi başına anlam ifade etmeyen, anlamlı hale gelebilmesi için başka bir morfeme eklenmesi gereken morfemdir. Örnek olarak *'kept'* kelimesinde yer alan *'kep-*' veya *'receive'* kelimesinde yer alan *'-ceive'* verilebilir.

- Taban: Morfemlerin eklendiği bir elementtir ve gövde olarak da adlandırılır. Bir taban, *'kindness'* kelimesindeki *'kind'* gibi temel bir kök morfemden ibaret olabilir. Fakat bir ya da daha fazla morfemden ibaret olan bir kelime de olabilir. Örneğin, *'kindnesses'* kelimesini elde etmek için taban morfem olan *'kindness'* kelimesine çoğul özellik veren *'es'* morfemi eklenmiştir.

- Kök: Eklerle beraber kullanılabilen bir morfemdir. Kök, yanına eklenen eklerle nazaran anlamlı ifadelerle sahiptirler. Örneğin *'kind'* kökü yanına ekler alarak *'kindly'*, *'kindness'*, *'kinder'*, *'kindest'* gibi kelimelerin oluşturulmasını sağlayabilir. Karmaşık bir kelimenin diğer tüm morfemlerini çıkardığınızda geriye kalan yapı köktür. Örneğin *'dehumanizing'* kelimesindeki *'-ing'*, *'de-*', *'-ize'* morfemleri çıkarıldığında geriye kalan *'human'* morfemi bir köktür ve bu yapı anlamlı daha küçük bir yapıya parçalanamaz.

- Takı: Bir kök ya da gövdeye eklenen bağımlı bir morfemdir. Önekler, bir kök ya da gövdenin önüne eklenirken, son ekler bir kök ya da gövdenin sonuna eklenirler. Örnek olarak *'rewrite'* kelimesinde yer alan *'re-*' eki verilebilirken, son ekler örnek olarak da *'critical'* kelimesinde yer alan *'-al'* eki verilebilir.

- Çekim Eki: Zaman ya da çoğulluk gibi kavramları ifade etmek için kökler ile birleşen eklerdir. Örnek olarak *'cats'* kelimesindeki çoğulluğu ifade eden *'-s'* takısı, *'talked'* kelimesinde yer alan ve eylemin geçmişte yapılmış olduğunu ifade eden *'-ed'* takısı

verilebilir. Çekim ekleri, yeni kelimeler oluşturmada, var olanlara ekler ekleyerek çok genel anlamlar elde edilmesini sağlarlar.

- Yapım Eki: Köklere eklenerek, yeni kelimeler üretilmesini sağlarlar. Örneğin *'modern-ize'*, *'read-er'* kelimelerindeki *'-ize'* ve *'-er'* ekleri yeni kelime üretmede kullanılan yapım ekleridir. Türeme ile var olan kelimelerden yeni kelimeler üretilir. Çekim ekleri ile yapım ekleri arasındaki fark, gri mutlak bir sınır değildir. Türeme daha az düzenlidir ve çekim ekleri biçimbilimine nazaran daha az tahmin edilebilir. Örneğin, İngilizce kelimeleri çoğul hale getirmek için *'-s'*, *'-es'* ve *'-ies'* ekleri eklenir. Tüm kelimeleri çoğul biçime getirmek için aynı işlem yapılır. Fakat fiillerden isim türetme işlemi daha az tahmin edilebilir özellikler içerir. Örneğin, *'payment'* kelimesini türetmek için kullanılan *'pay'* fiiline *'-ment'* yapım eki getirilirken, *'refusal'* kelimesini türetmek için neden *'-al'* ekinin kullanımının nedenini tahmin etmek oldukça güçtür.

- İçerik Morfemi: İçerik morfemleri isim, sıfat, fiil ve zarf sınıflarına ayrılmaktadır.

- İşlev Morfemi: İşlev morfemlerine örnek olarak, *'a'*, *'the'* artikelleri, *'of'*, *'at'* edatları, *'was'*, *'have'* gibi yardımcı filleri verilebilir.

- Basit Kelime: Tek bir morfemden meydana gelen kelimelerdir ve anlamlı daha küçük yapıları ayrılamazlar. Örnek olarak *'item'*, *'five'* kelimeleri verilebilir.

- Karmaşık Kelime: Bir ya da daha fazla kelimedenden oluşan yapılardan meydana gelirler. Örnek olarak *'items'*, *'walked'*, *'dirty'* kelimeleri verilebilir.

- Birleşik Kelime: İki veya daha fazla basit veya karmaşık kelimedenden meydana gelen kelimelerdir. Örnek olarak *'langlord'*, *'red-hot'*, *'window cleaner'* yapıları verilebilir.

5.2 İngilizce Bükümlü Biçimbilimi

İngilizce bükümlü olarak ifade edilen üç anlam kategorisine sahiptir. Bunlar, kelimelerdeki sayı, fiillerdeki zaman çekimi, sıfatlardaki karşılaştırmadır. İngilizce bu kategoriler ile, İspanyolca ve Rusça dilleri ile kıyaslandığında eklerden oluşan küçük bir envantere sahiptir. Fakat bu kategoriyi ifade etmek için de daima eklerden yardım almamaktadır[5].

Çizelge 5.1 Bükümlü kategoriler ve İngilizce ekler

Bükümün Uygulandığı İngilizce Sınıfları	Bükümlü Kategori	Kategorileri İfade Etmek için Kullanılan Ekler
İsimler	Sayı	-s, -es: book/books, bush/bushes
.	Aitlik	-'s, -': the cat's tail, Charles' toe
Fiiler	Geniş zaman üçüncü tekil kişi	-s, -es: it rains, Karen writes, the water sloshes
.	Geçmiş zaman	-ed: paint/painted
.	Tamamlanmamış geçmiş zaman	-ed: paint/painted ('has paint ed) (past participle)
.	Devam eden zaman	-ing: fall/falling, write/writing (present participle)
Sıfatlar	Karşılaştırma	-er: tall/taller
.	En üstünlük ifadesi	-est: tall/tallest

5.3 Düzenli ve Düzensiz Biçimbilim

Burada İngilizce kelimelerin düzensiz bükümlü yapısı gösterilmektedir.

Çizelge 5.2 Çekim eki kategorileri ve İngilizce ekler

Düzensizlik Tipleri	İsimlerin çoğul halleri	Filer: Geçmiş zaman	Fiiller: Geçmişte ortaç
Olağandışı soneki	oxen, syllabi, antennae	-	taken, seen, fallen, eaten
Köklerdeki ünlü değişimi	foot/feet, mouse/mice	run/ran, come/came, flee/fled, meet/met, fly/flew, stick/stuck, get/got, break/broke	swim/swum, sing/sung
Olağandışı soneki ile köklerdeki ünlü değişimi	brother/brethren/	feel/felt, kneel/knelt	write/written, do/done, break/broken, fly/flown
Kök/gövde değişimi	-	send/sent, bend/bent, think/thought, teach/taught, buy/bought	send/sent, bend/bent, think/thought, teach/taught, buy/bought
Değişimsiz	deer, sheep, moose, fish	hit, beat	hit, beat, come

5.4 İngilizce Türetme Biçimbilimi

Çizelge 5.3' te İngilizce yapım ekleri örnekleri bulunmaktadır. Fakat aşağıda gösterilmemiş olan çok daha fazla sayıda yapım eki bulunmaktadır.

Çizelge 5.3 Birkaç İngilizce yapım eki örnekleri

Ekler	Eklerin uygulandığı sınıflar	Anlamdaki değişim	Örnekler
Önek 'non-'	İsim, sıfat	Karşıt anlam verme	Noun: non -starter Adj.: non -partisan
Sonek '-ity'	Sıfat	İsme dönüşme	electric/electricity obese/obesity
Önek 'un-'	Fiilimsi	Eylemi tersine dönüştürme	tie/ untie , fasten/ unfasten clear/ unclear , safe/ unsafe
Sonek '-ous'	İsim	Sıfata dönüşme	fame/famous, glamor/glamorous
Önek 're-'	Fiil	Hareketi tekrar etme	tie/ retie , write/ rewrite
Sonek '-able'	Fiil	Sıfata dönüşme	print/printable, drink/drinkable

5.5 Biçimbilimsel Analizin Yapıldığı Araç

Bu çalışmada İngilizce metinleri biçimbilimsel yapılarına göre ayırma işleminde Porter Stemming isimli bir algoritma[6] kullanılmıştır.

Porter stemming algoritması, İngilizce kelimelerinde yer alan biçimbilimsel ve çekimsel ekleri ayırma işlemini gerçekleştirmektedir. Orijinal stemming algoritması 1979 yılında Cambridge' de Bilgisayar laboratuvarında IR projesinin bir parçası olarak yazılmıştır. Orijinal stemmer, bir zamanlar oldukça popüler olan fakat günümüzde geçerliliğini yitirmiş olan BCPL(Basic Combined Programming Language) dilinde yazılmıştır[15]. 1999 yılında diğer dillerde yazılmış olan biçimleri ortaya çıkarılmıştır. Fakat bu sürümler arasında işlevsellik açısından çok büyük farklar oluşmuştur.

İngilizce metinlerde, KET isimli ayrıştırma yöntemi ile sıkıştırma işlemi yapıp tekrar sıkıştırılan metin eski haline getirildiğinde, orijinal metin ile sıkıştırma sonucu açılan metin arasında hiçbir fark gözlemlenmemiştir. Zemberek ile yaşanan hece düşmesinin görmezden gelinmesi sorunu, Stemming isimli algoritmada yaşanmamıştır[6].

6. N-GRAM

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). Bir metin 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, 2 olan bigram, 3 olan trigram ve 4 ile üstü ise n-gram olarak adlandırılmaktadır. Çoğunlukla bigram ve trigram kullanılmaktadır(Dincer).

N-gram modellerin en genel kullanım amacı (n-1) birim kullanılarak n. birimin tahmin edilmesini sağlamaktır. Birimlerin olasılıkları kullanılarak meydana getirdikleri sözcük ya da cümlelerin olasılıkları hesaplanabilmektedir. Harflerin temel birim olarak kabul edildiği aşağıdaki örnekte çeşitli n-gram modellere göre sözcüğün nasıl ayrıştırıldığı görülmektedir(Dincer).

Örnek: pencere

1-gram : $p - e - n - c - e - r - e$

2-gram : $pe - en - nc - ce - er - re$

3-gram : $pen - enc - nce - cer - ere$

Benzer şekilde sözcükler de temel birim olarak kabul edilmektedir(Dincer).

Örnek : bugün süt içtim

1-gram : $bugün - süt - içtim$

2-gram : $bugün süt - süt içtim$

3-gram : $bugün süt içtim$

Olasılık hesabı yapılabilmesi için önce hangi birimlerin art arda gelme olasılıklarıyla ilgileniliyorsa o birimlerin her bir örneğinin kaç adet olduğu sayılmaktadır. Örneğin sözcüklerin art arda gelme olasılıklarıyla eğitim kümesindeki tüm sözcüklerin kaçar defa geçtiği sayılmaktadır. Daha sonra örneğin eğer bigram model kullanılıyorsa iki sözcüğün art arda gelme durumlarının sayısı bulunmaktadır. Bu sonuçlar kullanılarak iki sözcüğün art arda gelme olasılıkları hesaplanmaktadır(Dincer).

Basit n-gram modeline göre her birimin diğer bir birimin peşinden gelme olasılığı eşit olmaktadır(Jurafsky). Örneğin sözcük tabanındaki bir modelde, eğitim kümesi 100000 sözcük içeriyorsa bir sözcüğün herhangi başka bir sözcükten sonra gelme olasılığı $1/100000$ olacaktır. Daha gerçekçi bir yaklaşımla, bir sözcüğün başka bir sözcüğün ardından gelme olasılığı sözcüğün rastlanma sıklığına bağlı olmaktadır. Örneğin 1000000 sözcüklük örnek bir eğitim kümesi içerisinde “*bu*” sözcüğü 69971 defa, “*tavşan*” sözcüğü ise 11 defa geçmektedir. Buna göre bir sonraki sözcüğün “*bu*” olma olasılığı 0.07, “*tavşan*” olma olasılığı 0.00001 diyebiliriz. Ancak tek başına sözcüğün rastlanma sıklığına bakmak yeterli değildir. Örneğin, “*Hemen sonra bu*” cümlesinin sonuna “*tavşan*” gelme olasılığı “*bu*” gelme olasılığından daha yüksektir. Bu durumda bir sözcüğün gelme olasılığı kendisinden önce gelen sözcüklere göre koşullu olasılığı hesaplanabilmektedir. Örneğin “*beyaz*” dan sonra “*tavşan*” gelme olasılığı olan $P(tavşan|beyaz)$ diğer durumlarda “*tavşan*” gelme olasılığına göre daha yüksektir. Sözcük tabanlı n-gram bir modelde bir cümleinin olasılığı Eşitlik 6.1’ deki gibidir(Dincer).

$$P(w_1, w_2, \dots, w_{n-1}, w_n) \quad (6.1)$$

Zincir kuralı uygulanarak n. sözcüğün olasılığı Eşitlik 6.2’ deki gibi ifade edilmektedir(Dincer).

$$\begin{aligned} P(w^n) &= P(w_1)P(w_2 | w_1)P(w_3 | w^2_1) \dots P(w_n | w^{n-1}_1) \\ P(w^n) &= \prod P(w_k | w^{k-1}_1) \end{aligned} \quad (6.2)$$

Ancak $P(w_n | w^{n-1}_1)$ olasılığını hesaplamak için kendisinden n sözcük önce gelen sözcüğe göre olasılığını hesaplamak gerekmektedir. Bu nedenle $P(w_n | w^{n-1}_1)$ olasılığı için bir yaklaştırma yapılmaktadır; örneğin bigram bir modelde bir sözcüğün olasılığını sadece bir önceki sözcük belirler şeklinde bir kabul yapılmaktadır. Buna göre n-gram modeller için

genel olarak bir sonraki sözcüğün kendinden önceki N sözcüğe göre koşullu olasılığı Eşitlik 6.3' deki gibidir(Dincer).

$$P(w_n | w^{n-1}_1) \approx P(w_n | w^{n-1}_{n-N+1}) \quad (6.3)$$

Eşitlik 6.3' deki Wn sözcüğünün olasılığının sadece kendinden önce gelen N sözcüğün olasılığına yaklaştırılabileceğini göstermektedir(Dincer).

Sözcük tabanlı bigram bir modelde, bir cümleinin olasılığı, Eşitlik 6.3' ün 6.2' de yerine konması ile elde edilmekte ve Eşitlik 6.4' de yer alan sonuca ulaşılmaktadır(Dincer).

$$P(w^n_1) \approx \prod_{k=1}^n P(w_k | w_{k-1}) \quad (6.4)$$

Bu tez çalışmasında ise harf bigram' ları, külliyatlarda yer alan dosyaların ayrıştırılmasında kullanılmışlardır. Oluşturulan bigram' lar sıklık değerlerine göre Huffman ağacına yerleştirilmiş ve bu ağacın elemanları için sıkıştırma işlemi uygulanmıştır.

7. TEST ORTAMININ OLUŞTURULMASI

Bu çalışmada test ortamı hazırlanırken Türkçe ve İngilizce olmak üzere iki ayrı dilde külliyat oluşturulmuştur.

7.1 Türkçe Külliyatlar

Türkçe külliyatlar bölümü altında 8 farklı külliyat oluşturulmuştur. Her külliyat da kendi arasında 2Kb, 4Kb, 8Kb, 16Kb, 32Kb, 64Kb, 128Kb, 256Kb, 512Kb ve 1024Kb boyutlarında olmak üzere 10 farklı külliyat hazırlanmıştır.

7.1.1 Erkek(Heterojen) Külliyatı (Külliyat-I)

Popüler hayat, güncel, ekonomi, spor, dünya ve sağlık gibi farklı konularda yazan 18 erkek yazara ait 376 doküman alınarak 2Kb, 4Kb, 8Kb, 16Kb, 32Kb, 64Kb, 128Kb, 256Kb, 512Kb ve 1024Kb boyutlarında 10 farklı metin dosyası oluşturulmuştur.

7.1.2 Erkek(Homojen) Külliyatı (Külliyat-II)

Sadece güncel konularda yazan 7 erkek yazara ait 237 doküman alınarak 2Kb, 4Kb, 8Kb, 16Kb, 32Kb, 64Kb, 128Kb, 256Kb, 512Kb ve 1024Kb boyutlarında 10 farklı metin dosyası oluşturulmuştur.

7.1.3 Heterojen Külliyatı (Külliyat-III)

Popüler hayat, güncel, ekonomi, spor, dünya ve sağlık gibi farklı konularda yazan 18 erkek ve 4 kadın yazara ait 500 doküman alınarak 2Kb, 4Kb, 8Kb, 16Kb, 32Kb, 64Kb, 128Kb, 256Kb, 512Kb ve 1024Kb boyutlarında 10 farklı metin dosyası oluşturulmuştur.

7.1.4 Homojen(Siyaset) Külliyatı (Külliyat-IV)

Siyaset konusunda yazan 12 erkek ve 1 kadın yazara ait 271 doküman alınarak 2Kb, 4Kb, 8Kb, 16Kb, 32Kb, 64Kb, 128Kb, 256Kb, 512Kb ve 1024Kb boyutlarında 10 farklı metin dosyası oluşturulmuştur.

7.1.5 Kadın(Heterojen) Külliyatı (Külliyat-V)

Popüler hayat, güncel, ekonomi, spor, dünya ve sağlık gibi farklı konularda yazan 4 kadın yazara ait 140 doküman 2Kb, 4Kb, 8Kb, 16Kb, 32Kb, 64Kb, 128Kb, 256Kb, 512Kb ve 1024Kb boyutlarında 10 farklı metin dosyası oluşturulmuştur.

7.1.6 Kadın(Homojen) Külliyyatı (Külliyyat-VI)

Güncel konularda yazan 2 kadın yazara ait 98 doküman alınarak 2Kb, 4Kb, 8Kb, 16Kb, 32Kb, 64Kb, 128Kb, 256Kb, 512Kb ve 1024Kb boyutlarında 10 farklı metin dosyası oluşturulmuştur.

7.1.7 Roman Anayasa Metinleri Külliyyatı (Külliyyat-VII)

Türkiye Anayasa metni ile 6 öykü ve 1 romandan oluşan ait 46 doküman alınarak 2Kb, 4Kb, 8Kb, 16Kb, 32Kb, 64Kb, 128Kb, 256Kb, 512Kb ve 1024Kb boyutlarında 10 farklı metin dosyası oluşturulmuştur.

7.1.8 Tez Dokümanları Külliyyatı (Külliyyat-VIII)

Çeşitli kişilere ait 6 farklı tez dokümanı alınarak 2Kb, 4Kb, 8Kb, 16Kb, 32Kb, 64Kb, 128Kb, 256Kb, 512Kb ve 1024Kb boyutlarında 10 farklı metin dosyası oluşturulmuştur.

7.2 İngilizce Külliyyat(Külliyyat-IX)

Güncel konularda yazan farklı kişilere ait 290 farklı doküman alınarak 2Kb, 4Kb, 8Kb, 16Kb, 32Kb, 64Kb, 128Kb, 256Kb, 512Kb ve 1024Kb boyutlarında 10 farklı metin dosyası oluşturulmuştur.

8. DENEYSEL SONUÇLAR

Deneysel sonuçlar, Intel(R) Core(TM)2 Duo CPU T7500 @ 2.20GHz işlemcili, 2 GB bellekli ve sistem türü 32 bit işletim sistemi olan bir bilgisayardan elde edilmiştir.

Deneysel sonuçlar alınırken Türkçe ve İngilizce için farklı külliyatlar kullanılmıştır. Türkçe için farklı alanlarda yazan 22 yazarın 900 farklı yazısından 8 farklı alt külliyat ve her külliyat için 2Kb, 4Kb, 8Kb, 16Kb, 32Kb, 64Kb, 128Kb, 256Kb, 512Kb, 1024Kb boyutlarında dosyalar oluşturulmuştur. İngilizce için güncel konularda yazan farklı kişilere ait 290 farklı dokümandan 1 alt külliyat ve bu külliyat için 2Kb, 4Kb, 8Kb, 16Kb, 32Kb, 64Kb, 128Kb, 256Kb, 512Kb, 1024Kb boyutlarında dosyalar oluşturulmuştur.

Türkçe külliyatlar üzerinde sıkıştırma işlemi yapılırken, metinler biçimbilimsel yapılarına göre, ekler ek halinde olmak üzere kök – ek(KET), hece(H), karakter(K), ekler tek parça halinde olmak üzere isim – ek(İET), ekler tek parça halinde olmak üzere fiil – ek(FET), ekler tek parça halinde olmak üzere sıfat – ek(SET), ekler hece halinde olmak üzere isim – ek(İEH), ekler hece halinde olmak üzere fiil – ek(FEH), ekler hece halinde olmak üzere sıfat – ek(SEH), ekler karakter olmak üzere isim – ek(İEK), ekler karakter olmak üzere fiil – ek(FEK), ekler karakter olmak üzere sıfat – ek(SEK) ve bigram(2-GRAM) olarak 13 farklı şekilde ayrıştırılmışlardır.

İngilizce külliyat üzerinde sıkıştırma işlemi yapılırken, metinler biçimbilimsel yapılarına göre, ekler ek halinde olmak üzere kök – ek(İKET), ekler karakter olmak üzere kök – ek(İKEK), karakter(İK) ve bigram(2-GRAM) olarak 4 farklı şekilde ayrıştırılmışlardır. Türkçe külliyatlar üzerinde hecelerine göre sıkıştırma işlemi yapılabilirken, İngilizcede hecelerine ayırma işlemi yapılamadığından, İngilizce metinler hecelerine ayrılarak sıkıştırılamamışlardır.

8.1 Türkçe Üzerine Yapılan Çalışmalar

Bu bölümde Türkçe üzerine yapılan çalışmalar sırasıyla verilmiştir.

8.1.1 Külliyata Göre Deneysel Sonuçlar

Bu bölümde 8 farklı külliyat üzerinde yapılan deneysel sonuçlar sırasıyla verilmiştir.

8.1.1.1 Külliyyat-I için DeneySEL Sonuçlar

Popüler hayat, güncel, ekonomi, spor, dünya ve sağlık gibi farklı konularda yazan 18 erkek yazara ait 376 farklı doküman alınarak oluşturulan Külliyyat-I üzerinde, 13 farklı ayrıştırma yöntemi uygulanarak deneyler yapılmıştır. Çizelge 8.1’ de sıkıştırma başarı sonuçları verilmiştir.

Külliyyat-I’ deki en iyi performans % 50 olup, metnin ekler ek halinde olmak üzere kök – ek(KET) şeklinde ayrıştırılması yönteminde 1024Kb boyutlu dosyanın sıkıştırılmasından elde edilmiştir. En kötü sonuç ise, metnin ekler ek halinde olmak üzere kök – ek(KET) şeklinde ayrıştırılması yönteminde 2Kb boyutlu dosyanın sıkıştırılması ile metnin hece(H) şeklinde ayrıştırılması yönteminde 2Kb boyutlu dosyanın sıkıştırılmasından elde edilmiştir. 2Kb ve 4Kb boyutlu dosyaların 2-GRAM ayrıştırma yöntemi ile sıkıştırılmasından da kötü sıkıştırma sonuçları ortaya çıkmıştır. Genel olarak başarının külliyyatın boyutuna ve ayrıştırma yöntemine bağlı olarak değişkenlik gösterdiği gözlemlenmiştir(Çizelge 8.1).

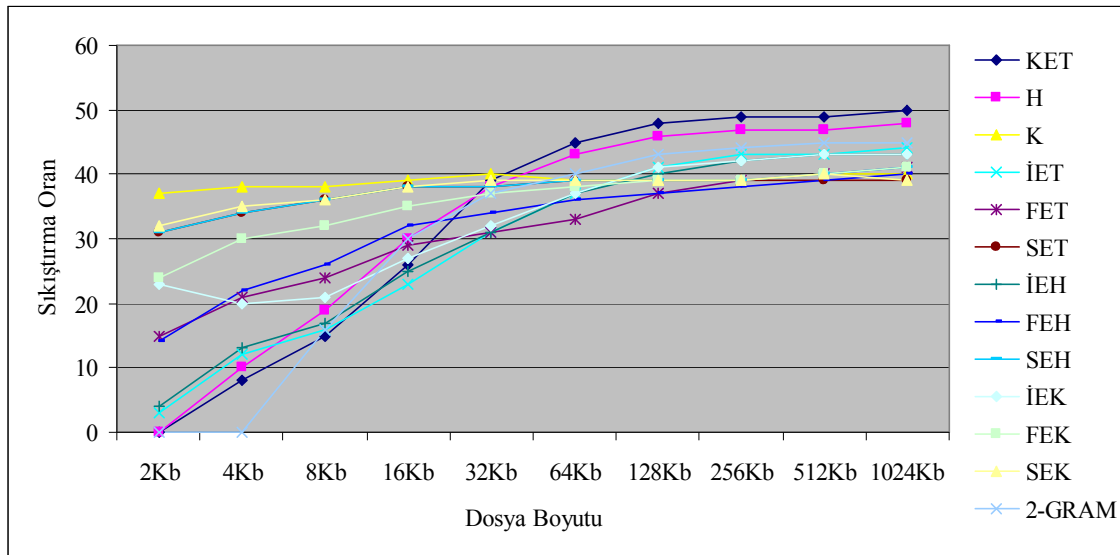
Tüm sıkıştırma sonuçlarının ayrıştırma yöntemine göre verdikleri sıkıştırma başarılarının ortalamaları alındığında karakter(K) ayrıştırma yönteminin % 38,9 ile en başarılı, ekler tek parça halinde olmak üzere isim – ek(İET) ayrıştırma yönteminin % 29,3 ile bu külliyyatta en başarısız yöntem olduğu gözlemlenmiştir(Çizelge 8.1). Başarı oranının düşük olması sebepleri arasında, dokümanlar içerisinde çok sayıda isim olması gösterilebilir. Bu durum kullanılan ağacın derinliğinin artması nedeniyle başlık bilgisinin boyutunun genişlemesinden kaynaklanmaktadır.

Tüm sıkıştırma sonuçlarının dosyanın boyutuna göre verdikleri sıkıştırma başarılarının ortalamaları alındığında 1024Kb boyutlu dosya ile yapılan sıkıştırma işleminin % 42,6 ile en başarılı, 2Kb boyutlu dosya ile yapılan sıkıştırma işleminin % 16,5 ile bu külliyyatta en başarısız sonuçların alındığı dosya boyutu olduğu gözlemlenmiştir(Çizelge 8.1). Başlık bilgisinin genişliğinden dolayı küçük dosyalarda her zaman başarısız sonuç alınması normal olmaktadır.

Çizelge 8.1 Külliyyat-I için sıkıştırma oranları (%)

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
KET	0	8	15	26	39	45	48	49	49	50	32,9
H	0	10	19	30	38	43	46	47	47	48	32,8
K	37	38	38	39	40	39	39	39	40	40	38,9
İET	3	12	16	23	31	37	41	43	43	44	29,3
FET	15	21	24	29	31	33	37	39	40	41	31
SET	31	34	36	38	38	39	39	39	39	39	37,2
İEH	4	13	17	25	31	37	40	42	43	43	29,5
FEH	14	22	26	32	34	36	37	38	39	40	31,8
SEH	31	34	36	38	38	39	39	39	40	41	37,5
İEK	23	20	21	27	32	37	41	42	43	43	32,9
FEK	24	30	32	35	37	38	39	39	40	41	35,5
SEK	32	35	36	38	39	39	39	39	40	39	37,6
2-GRAM	0	0	16	30	37	40	43	44	45	45	30
Ort Başarı %	16,5	21,3	25,5	31,5	35,8	38,6	40,6	41,5	42,2	42,6	

Popüler hayat, güncel, ekonomi, spor, dünya ve sağlık gibi farklı konularda yazan 18 erkek yazara ait 376 farklı doküman alınarak oluşturulan Külliyyat-I üzerinde, 13 farklı ayrıştırma yöntemi uygulanarak yapılan deneylerin grafiksel sonucu Şekil 8.1’ de verilmiştir.



Şekil 8.1 Külliyyat-I için sıkıştırma oranları (%)

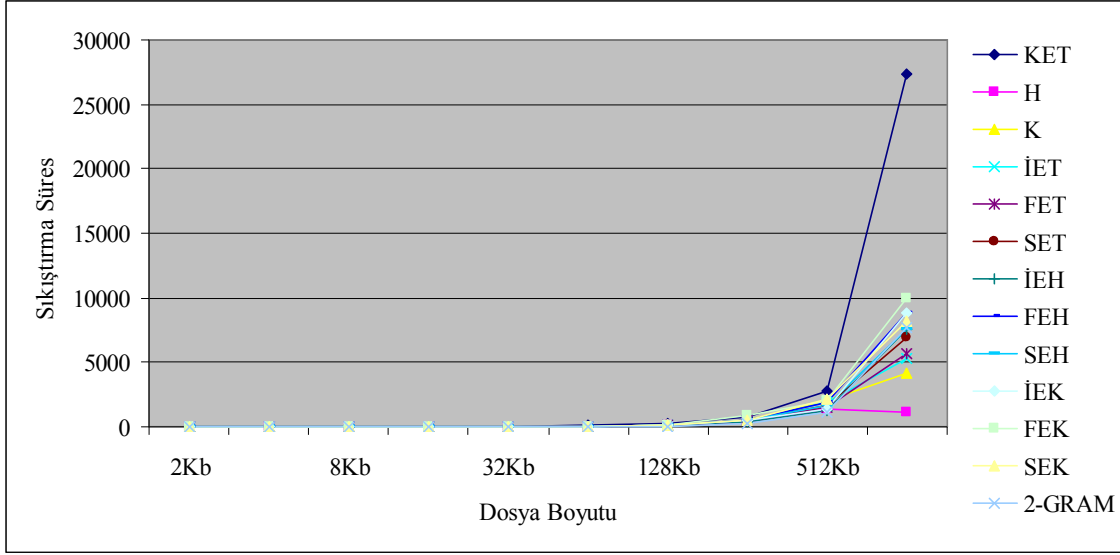
Çizelge 8.1’ de yer alan 0(sıfır) ifadesi, sıkıştırma işlemi uygulandığında sıkıştırma olmayıp, başlık yapısının boyutundan dolayı genişleme olmasından kaynaklanmaktadır.

Çizelge 8.2’ de Külliyyat-I’ de yer alan her dosyanın sıkıştırma ve geri açma süreleri saniye cinsinden verilmekte ve ön işleme, Huffman ağacının oluşturulması, sıkıştırma ve sıkıştırılan verinin açılması sürelerinin toplamını içermektedir.

Çizelge 8.2 Külliyyat-I için toplam sıkıştırma süreleri (sn)

	Dosya Boyutu									
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb
KET	1,6	2,08	6,46	15,812	34,83	75,488	238,186	775,773	2730,239	27343,275
H	0,14	0,21	0,85	2,333	4,994	10,471	64,148	403,699	1406,429	1188,812
K	0,06	0,09	0,26	1,009	3,867	16,054	120,013	553,504	2003,774	4173,983
İET	0,11	0,35	0,52	2,303	4,948	14,733	67,515	484,896	1863,872	5234,876
FET	0,09	0,17	0,57	1,317	4,482	19,889	81,433	551,071	1449,725	5686,781
SET	0,08	0,14	0,36	1,663	4,002	15,247	97,75	453,833	1629,343	6987,002
İEH	0,1	0,19	0,62	2,514	4,863	13,123	69,202	402,09	1309,231	7623,446
FEH	0,09	0,17	0,45	1,208	3,628	13,212	79,701	478,046	1953,76	8791,654
SEH	0,07	0,17	0,38	1,352	4,141	14,076	89,686	525,564	1669,776	7543,624
İEK	0,08	0,19	0,44	2,179	4,953	13,862	76,066	471,516	1409,494	8786,599
FEK	0,08	0,14	0,38	1,386	4,358	14,198	89,654	901,245	2014,084	9986,436
SEK	0,07	0,13	0,4	1,317	4,222	14,04	92,04	535,629	2134,932	8234,754
2-GRAM	0,41	0,56	1,13	2,401	5,274	14,18	60,388	280,588	1170,033	7635,215

18 farklı erkek yazarın yazılarının sıkıştırılması sırasında elde edilen sıkıştırma zamanı sonuçlarının saniye cinsinden grafiksel gösterimi Şekil 8.2’ de verilmiştir.



Şekil 8.2 Külliyyat-I için toplam sıkıştırma süreleri (sn)

8.1.1.2 Külliyyat-II için Deneysel Sonuçlar

Sadece güncel konularda yazan 7 erkek yazara ait 237 doküman alınarak oluşturulan Külliyyat-II üzerinde, 13 farklı ayrıştırma yöntemi kullanılarak deneyler yapılmıştır. Çizelge 8.3' de alınan sıkıştırma başarı sonuçları verilmiştir.

Külliyyat-II' deki en iyi performans % 50 olup, metnin ekler ek halinde olmak üzere kök – ek(KET) şeklinde ayrıştırılması yönteminde 1024Kb boyutlu dosyanın sıkıştırılmasından elde edilmiştir. En kötü sonuç ise, metnin 2-GRAM yöntemi ile ayrıştırılmasında 2Kb ve 4Kb boyutlu dosyaların sıkıştırılmasından elde edilmiştir. Genel olarak başarının külliyyatın boyutuna ve ayrıştırma yöntemine bağlı olarak değişkenlik gösterdiği gözlemlenmiştir(Çizelge 8.3).

Tüm sıkıştırma sonuçlarının ayrıştırma yöntemine göre verdikleri sıkıştırma başarılarının ortalamaları alındığında karakter(K) ayrıştırma yönteminin % 39,6 ile en başarılı, 2-GRAM ayrıştırma yönteminin ise % 29,6 ile bu külliyyatta en başarısız yöntem olduğu gözlemlenmiştir (Çizelge 8.3).

Tüm sıkıştırma sonuçlarının dosyanın boyutuna göre verdikleri sıkıştırma başarılarının ortalamaları alındığında 1024Kb boyutlu dosya ile yapılan sıkıştırma işleminin % 42,5 ile en başarılı, 2Kb boyutlu dosya ile yapılan sıkıştırma işleminin % 18,2 ile bu külliyyatta en başarısız sonuçların alındığı dosya boyutu olduğu gözlemlenmiştir(Çizelge 8.3).

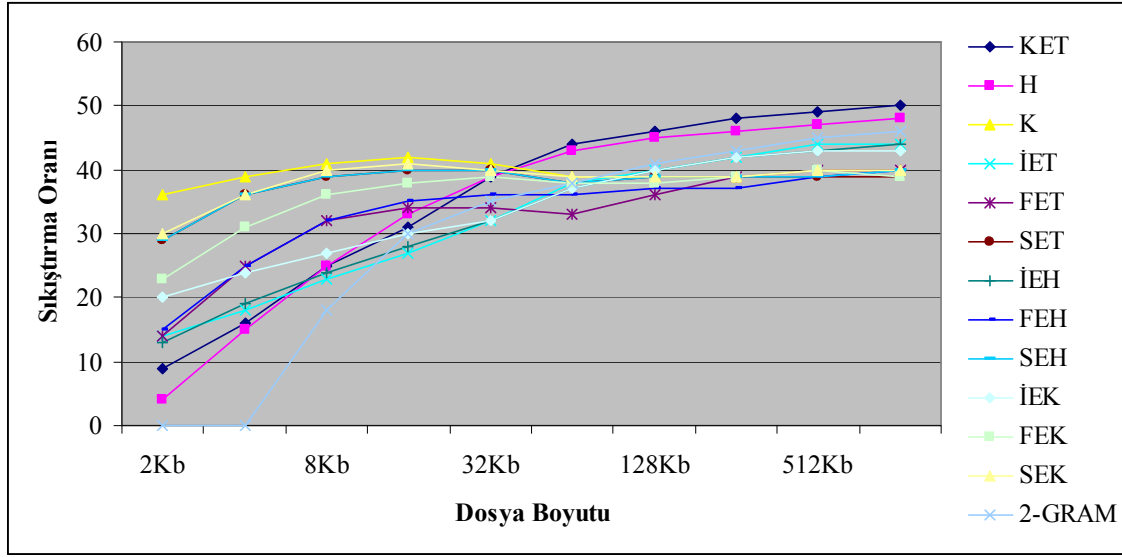
Çizelge 8.3 Külliyat-II için sıkıştırma oranları (%)

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
KET	9	16	25	31	39	44	46	48	49	50	35,7
H	4	15	25	33	39	43	45	46	47	48	34,5
K	36	39	41	42	41	39	39	39	40	40	39,6
İET	14	18	23	27	32	38	40	42	44	44	32,2
FET	14	25	32	34	34	33	36	39	40	40	32,7
SET	29	36	39	40	40	38	39	39	39	39	37,8
İEH	13	19	24	28	32	37	40	42	43	44	32,2
FEH	15	25	32	35	36	36	37	37	39	40	33,2
SEH	29	36	39	40	40	38	39	39	39	40	37,9
İEK	20	24	27	30	32	37	40	42	43	43	33,8
FEK	23	31	36	38	39	38	38	39	40	39	36,1
SEK	30	36	40	41	40	39	39	39	40	40	38,4
2-GRAM	0	0	18	30	35	38	41	43	45	46	29,6
Ort Başarı %	18,2	24,6	30,9	34,5	36,9	38,3	39,9	41,1	42,2	42,5	

Çizelge 8.3’ de yer alan 0(sıfır) ifadesi, sıkıştırma işlemi uygulandığında sıkıştırma olmayıp, başlık yapısının boyutundan dolayı genişleme olmasından kaynaklanmaktadır.

Sadece güncel konularda yazan 7 erkek yazara ait 237 doküman alınarak oluşturulan Külliyat-II üzerinde, 13 farklı ayrıştırma yöntemi uygulanarak yapılan deneylerin grafiksel sonucu Şekil 8.3’ de verilmiştir.

Külliyat-II ile yapılmış olan denemeler şunu göstermiştir ki eğer elimizdeki dokümanlar tek bir konu üzerine ise başarılar Külliyat-I’ e göre daha yüksek çıkmaktadır.



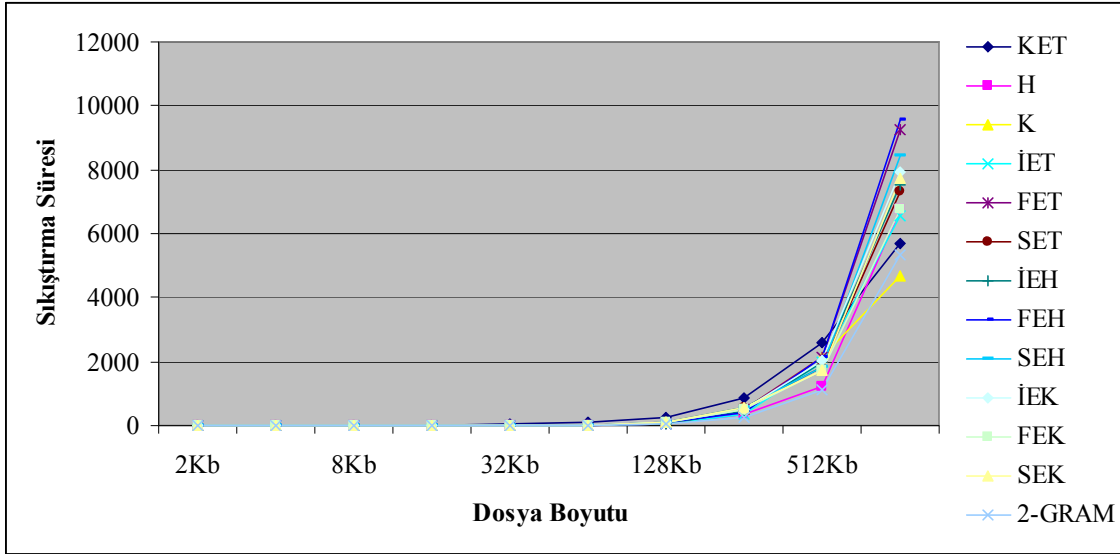
Şekil 8.3 Külliyyat-II için sıkıştırma oranları (%)

Çizelge 8.4' de Külliyyat-II' de yer alan her dosyanın sıkıştırma ve geri açma süreleri saniye cinsinden verilmekte ve ön işleme, Huffman ağacının oluşturulması, sıkıştırma ve sıkıştırılan verinin açılması sürelerinin toplamını içermektedir.

Çizelge 8.4 Külliyyat-II için toplam sıkıştırma süreleri (sn)

	Dosya Boyutu									
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb
KET	2,777	8,406	13,891	17,538	31,345	77,18	244,998	855,91	2611,416	5678,654
H	0,171	0,335	0,919	2,192	4,276	10,16	65,098	340,859	1221,093	6754,322
K	0,077	0,117	0,319	1,051	4,189	13,627	116,754	529,944	2130,122	4678,523
İET	0,128	0,312	0,526	2,045	5,456	14,25	68,161	373,967	1995,845	6545,687
FET	0,143	0,214	0,624	1,658	4,38	14,011	75,562	499,482	2137,731	9234,849
SET	0,122	0,189	0,438	1,167	4,071	14,766	89,294	558,278	1811,251	7345,012
İEH	0,163	0,269	0,531	1,679	5,173	13,834	38,446	476,566	1983,977	7549,981
FEH	0,067	0,285	0,459	1,269	4,483	14,264	77,861	399,921	2141,846	9566,391
SEH	0,065	0,374	0,424	1,18	4,353	14,396	87,684	562,414	1810,432	8440,967
İEK	0,088	0,191	0,551	1,406	5,22	13,368	79,119	496,011	2038,017	7934,652
FEK	0,06	0,322	0,474	1,268	4,278	13,759	85,244	531,825	1742,751	6772,342
SEK	0,145	0,171	0,514	1,143	4,185	14,567	91,695	566,17	1730,828	7738,544
2-GRAM	0,25	0,641	1,623	2,839	7,598	17,868	65,022	274,092	1111,091	5361,256

7 farklı erkek yazarın yazılarının sıkıştırılması sırasında elde edilen sıkıştırma zamanı sonuçlarının saniye cinsinden grafiksel gösterimi Şekil 8.4' de verilmiştir.



Şekil 8.4 Külliyyat-II için toplam sıkıştırma süreleri (sn)

8.1.1.3 Külliyyat-III için DeneySEL Sonuçlar

Popüler hayat, güncel, ekonomi, spor, dünya ve sağlık gibi farklı konularda yazan 18 erkek ve 4 kadın yazara ait 516 doküman alınarak oluşturulan Külliyyat-III üzerinde, 13 farklı ayrıştırma yöntemi kullanılarak deneyler yapılmıştır. Çizelge 8.5’ de sıkıştırma başarı sonuçları verilmiştir.

Külliyyat-III’ deki en iyi performans % 57 olup, metnin ekler karakter olmak üzere isim – ek(İEK) şeklinde ayrıştırılması yönteminde 512Kb boyutlu dosyanın sıkıştırılmasından elde edilmiştir. En kötü sonuç ise, metnin ekler ek halinde olmak üzere kök – ek(KET) şeklinde ayrıştırılması yöntemi ile metnin hece(H) şeklinde ayrıştırılması yönteminde 2Kb boyutlu dosyanın sıkıştırılmasından elde edilmiştir. 2Kb ve 4Kb boyutlu dosyaların 2-GRAM ayrıştırma yöntemi ile sıkıştırılmasından da kötü sıkıştırma sonuçları ortaya çıkmıştır. Genel olarak başarının külliyyatın boyutuna ve ayrıştırma yöntemine bağlı olarak değişkenlik gösterdiği gözlemlenmiştir(Çizelge 8.5).

Tüm sıkıştırma sonuçlarının ayrıştırma yöntemine göre verdikleri sıkıştırma başarılarının ortalamaları alındığında ekler karakter olmak üzere sıfat – ek(SEK) ayrıştırma yönteminin % 37,7 ile en başarılı, ekler tek parça halinde olmak üzere isim – ek(İET) ayrıştırma yönteminin % 29,2 ile bu külliyyatta en başarısız yöntemler oldukları gözlemlenmiştir (Çizelge 8.5).

Çizelge 8.5 Külliyat-III için sıkıştırma oranları (%)

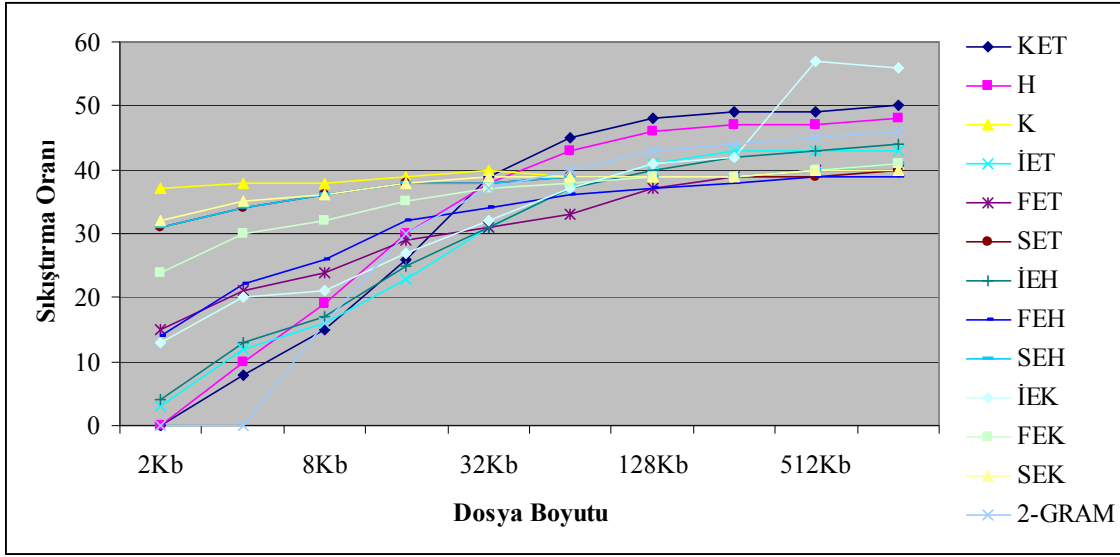
	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
KET	0	8	15	26	39	45	48	49	49	50	32,9
H	0	10	19	30	38	43	46	47	47	48	32,8
K	37	38	38	39	40	39	39	39	40	41	39
İET	3	12	16	23	31	37	41	43	43	43	29,2
FET	15	21	24	29	31	33	37	39	40	40	30,9
SET	31	34	36	38	38	39	39	39	39	40	37,3
İEH	4	13	17	25	31	37	40	42	43	44	29,6
FEH	14	22	26	32	34	36	37	38	39	39	31,7
SEH	31	34	36	38	38	39	39	39	40	41	37,5
İEK	13	20	21	27	32	37	41	42	57	56	34,6
FEK	24	30	32	35	37	38	39	39	40	41	35,5
SEK	32	35	36	38	39	39	39	39	40	40	37,7
2-GRAM	0	0	16	30	37	40	43	44	45	46	30,1
Ort Başarı %	15,69	21,31	25,54	31,54	35,77	38,62	40,62	41,46	43,23	43,77	

Tüm sıkıştırma sonuçlarının dosyanın boyutuna göre verdikleri sıkıştırma başarılarının ortalamaları alındığında 1024Kb boyutlu dosya ile yapılan sıkıştırma işleminin % 43,77 ile en başarılı, 2Kb boyutlu dosya ile yapılan sıkıştırma işleminin % 15,69 ile bu külliyatta en başarısız sonuçların alındığı dosya boyutu olduğu gözlemlenmiştir(Çizelge 8.5).

Çizelge 8.5’ de yer alan 0(sıfır) ifadesi, sıkıştırma işlemi uygulandığında sıkıştırma olmayıp, başlık yapısının boyutundan dolayı genişleme olmasından kaynaklanmaktadır.

Popüler hayat, güncel, ekonomi, spor, dünya ve sağlık gibi farklı konularda yazan 18 erkek ve 4 kadın yazara ait 516 doküman alınarak oluşturulan Külliyat-III üzerinde, 13 farklı ayrıştırma yöntemi uygulanılarak yapılan deneylerin grafiksel sonucu Şekil 8.5’ de verilmiştir.

Külliyat-III, Külliyat-II’ ye göre daha başarısız sonuçlar vermesine rağmen Külliyat-I’ e göre daha yüksek sonuçlar vermiştir. Külliyat-I heterojen bir yapıya sahip olan sadece erkek yazarlardan oluşurken, Külliyat-III içerisinde kadın yazarları da bulundurmaktadır.



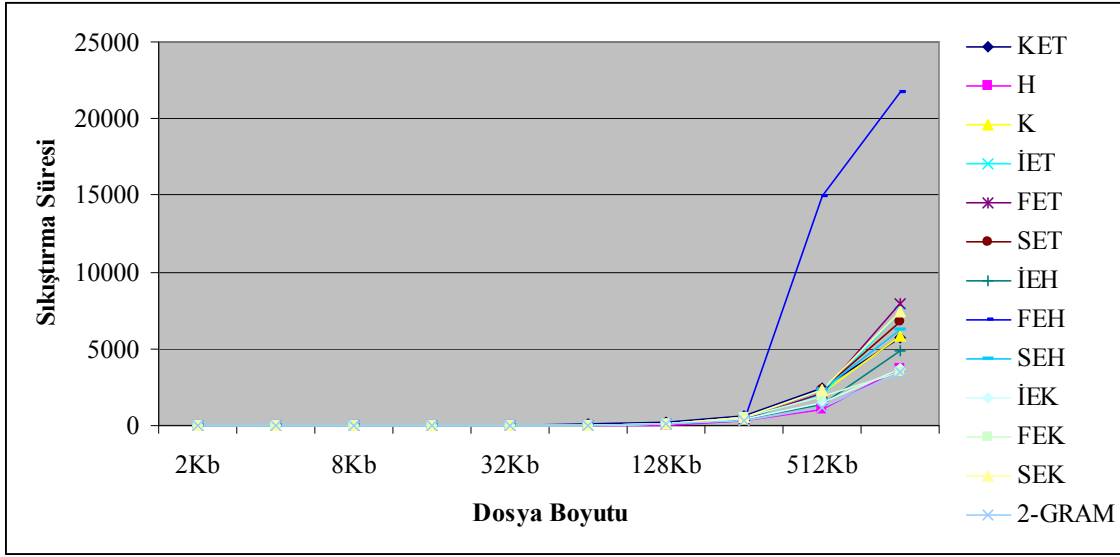
Şekil 8.5 Külliyyat-III için sıkıştırma oranları (%)

Çizelge 8.6’ da Külliyyat-III’ de yer alan her dosyanın sıkıştırma ve geri açma süreleri saniye cinsinden verilmiştir ve ön işleme, Huffman ağacının oluşturulması, sıkıştırma ve sıkıştırılan verinin açılması sürelerinin toplamını içermektedir.

Çizelge 8.6 Külliyyat-III için toplam sıkıştırma süreleri (sn)

	Dosya Boyutu									
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb
KET	1,53	1,397	6,099	15,162	30,358	64,195	206,046	675,419	2478,857	5765,899
H	0,112	0,188	0,782	2,419	4,37	9,626	46,973	361,891	1060,646	3654,78
K	0,077	0,109	0,265	0,922	3,417	1,897	86,097	474,272	2110,556	5812,734
İET	0,14	0,313	0,499	2,059	4,837	11,967	62,698	456,051	2014,04	7456,231
FET	0,109	0,172	0,423	1,358	3,993	13,354	70,184	432,683	2107,171	7900,752
SET	0,063	0,11	0,36	1,126	3,775	13,323	79,373	518,733	2313,201	6781,222
İEH	0,078	0,171	0,64	1,733	4,79	12,097	62,167	441,123	1384,767	4891,265
FEH	0,077	0,156	0,391	1,141	3,838	12,574	70,542	360,454	14895,582	21762,351
SEH	0,079	0,14	0,358	1,125	3,682	12,741	80,061	518,14	2307,428	6231,95
İEK	0,062	0,157	0,422	1,514	4,555	12,372	67,642	450,639	1467,274	3712,998
FEK	0,061	0,142	0,375	1,109	3,682	3,652	76,987	491,603	1919,364	3496,754
SEK	0,078	0,125	0,375	1,093	3,715	13,058	79,873	520,028	2293,981	7387,644
2-GRAM	0,389	0,546	1,063	2,358	5,165	14,109	57,737	270,895	1157,364	3457,394

18 farklı erkek ve 4 farklı kadın yazarın yazılarının sıkıştırılması sırasında elde edilen sıkıştırma zamanı sonuçlarının saniye cinsinden grafiksel gösterimi Şekil 8.6’ da verilmiştir.



Şekil 8.6 Külliyyat-III için toplam sıkıştırma süreleri (sn)

8.1.1.4 Külliyyat-IV için Deneysel Sonuçlar

Siyaset konusunda yazan 12 erkek ve 1 kadın yazara ait 271 doküman alınarak oluşturulan Külliyyat-IV üzerinde, 13 farklı ayrıştırma yöntemi kullanılarak deneyler yapılmıştır. Çizelge 8.7’ de sıkıştırma başarı sonuçları verilmiştir.

Külliyyat-IV’ deki en iyi performans % 61 olup, metnin ekler ek halinde olmak üzere kök – ek(KET) şeklinde ayrıştırılması yönteminde 1024Kb boyutlu dosyanın sıkıştırılmasından elde edilmiştir. En kötü sonuç ise 2Kb boyutlu dosyanın 2-GRAM ayrıştırma yöntemi ile sıkıştırılmasından elde edilmiştir. Genel olarak başarının külliyyatın boyutuna ve ayrıştırma yöntemine bağlı olarak değişkenlik gösterdiği gözlemlenmiştir(Çizelge 8.7).

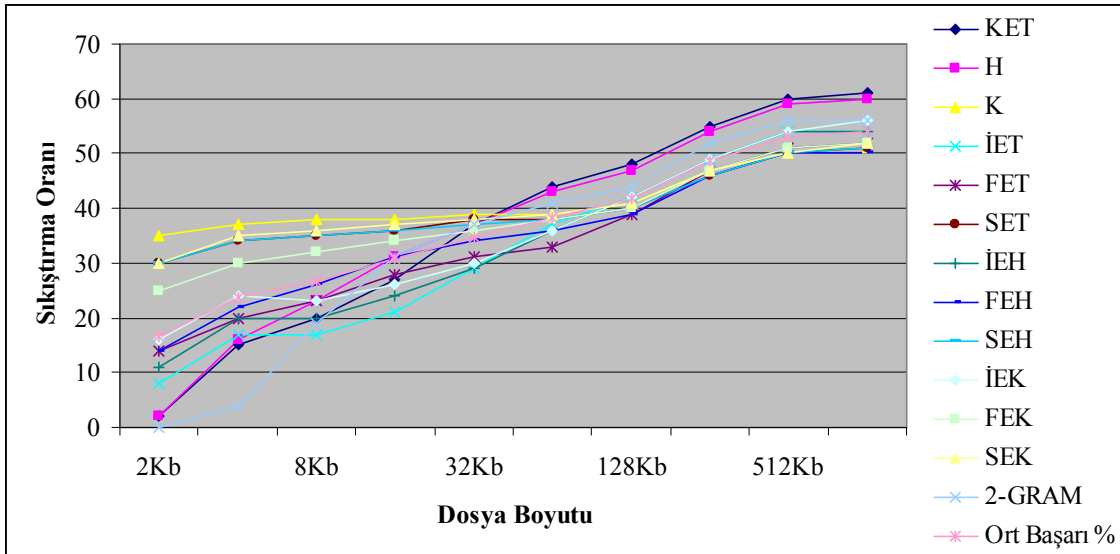
Tüm sıkıştırma sonuçlarının ayrıştırma yöntemine göre verdikleri sıkıştırma başarılarının ortalamaları alındığında karakter(K) ayrıştırma yönteminin % 41,5 ile en başarılı, ekler tek parça halinde olmak üzere isim – ek(İET) ayrıştırma yönteminin % 32,8 ile bu külliyyatta en başarısız yöntem olduğu gözlemlenmiştir(Çizelge 8.7).

Tüm sıkıştırma sonuçlarının dosyanın boyutuna göre verdikleri sıkıştırma başarılarının ortalamaları alındığında 1024Kb boyutlu dosya ile yapılan sıkıştırma işleminin % 53,85 ile en başarılı, 2Kb boyutlu dosya ile yapılan sıkıştırma işleminin % 16,69 ile bu külliyyatta en başarısız sonuçların alındığı dosya boyutu olduğu gözlemlenmiştir(Çizelge 8.7).

Çizelge 8.7 Külliyyat-IV için sıkıştırma oranları (%)

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
KET	2	15	20	27	37	44	48	55	60	61	36,9
H	2	16	23	31	37	43	47	54	59	60	37,2
K	35	37	38	38	39	39	41	47	50	51	41,5
İET	8	17	17	21	29	37	42	49	54	54	32,8
FET	14	20	23	28	31	33	39	47	51	52	33,8
SET	30	34	35	36	38	38	40	46	50	51	39,8
İEH	11	20	20	24	29	36	42	49	54	54	33,9
FEH	14	22	26	31	34	36	39	46	50	50	34,8
SEH	30	34	35	36	37	38	40	46	50	51	39,7
İEK	16	24	23	26	30	36	42	49	54	56	35,6
FEK	25	30	32	34	36	38	40	47	51	52	38,5
SEK	30	35	36	37	38	39	41	47	50	52	40,5
2-GRAM	0	4	19	31	37	41	44	52	56	56	34
Ort Başarı %	16,69	23,69	26,69	30,77	34,77	38,31	41,92	48,77	53	53,85	

Siyaset konusunda yazan 12 erkek ve 1 kadın yazara ait 271 doküman alınarak oluşturulan Külliyyat-IV üzerinde, 13 farklı ayrıştırma yöntemi uygulanılarak yapılan deneylerin grafiksel sonucu Şekil 8.7’ de verilmiştir.



Şekil 8.7 Külliyyat-IV için sıkıştırma oranları (%)

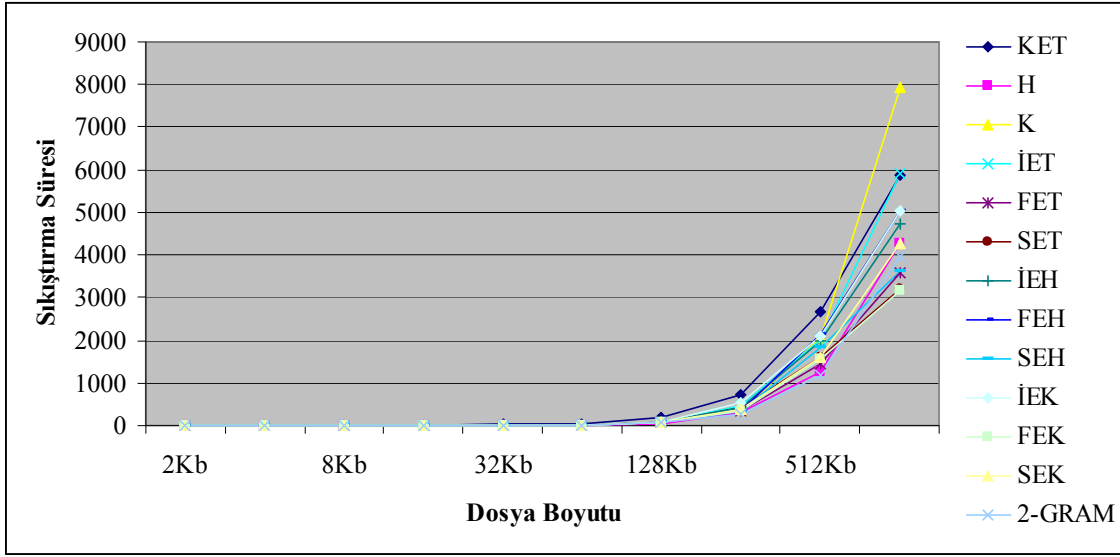
Külliyat-IV şimdiye kadarki en başarılı külliyat olmuştur. Tek bir konuda daha az sayıda üslup ile yazılmış dokümanlardan oluşmaktadır.

Çizelge 8.8’ de Külliyat-IV’ de yer alan her dosyanın sıkıştırma ve geri açma süreleri saniye cinsinden verilmiştir ve ön işleme, Huffman ağacının oluşturulması, sıkıştırma ve sıkıştırılan verinin açılması sürelerinin toplamını içermektedir.

Çizelge 8.8 Külliyat-IV için toplam sıkıştırma süreleri (sn)

	Dosya Boyutu									
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb
KET	1,26	2,3	5,11	10,884	22,85	52,192	204,501	732,517	2651,699	5871,776
H	0,17	0,21	0,8	1,936	4,478	10,79	55,24	292,033	1247,546	4276,124
K	0,06	0,13	0,33	0,968	3,449	14,432	82,994	353,201	2069,56	7922,391
İET	0,14	0,19	0,53	2,075	5,039	12,862	69,437	473,337	2012,28	5922,65
FET	0,13	0,2	0,64	1,358	4,166	13,043	73,508	336,025	1445,934	3601,941
SET	0,11	0,17	0,37	1,139	4,057	13,292	80,044	428,97	1588,736	3209,182
İEH	0,13	0,17	0,64	1,748	5,171	12,186	62,463	400,983	1964,462	4735,824
FEH	0,11	0,19	0,5	1,218	3,762	13,167	70,778	328,723	2123,099	5025,814
SEH	0,09	0,19	0,39	1,154	4,1	14,853	80,84	358,879	1829,99	3627,825
İEK	0,09	0,19	0,71	1,451	4,618	14,29	72,775	523,069	2081,693	5025,168
FEK	0,08	0,16	0,41	1,281	3,745	14,073	78,532	353,544	1557,786	3146,248
SEK	0,09	0,17	0,46	1,189	3,714	14,525	74,554	362,729	1596,677	4263,871
2-GRAM	0,36	0,58	0,91	2,185	5,444	13,728	60,903	280,256	1194,65	3954,291

12 farklı erkek ve 1 farklı kadın yazarın yazılarının sıkıştırılması sırasında elde edilen sıkıştırma zamanı sonuçlarının saniye cinsinden grafiksel gösterimi Şekil 8.8’ de verilmiştir.



Şekil 8.8 Külliyyat-IV için toplam sıkıştırma süreleri (sn)

8.1.1.5 Külliyyat-V için Deneysel Sonuçlar

Popüler hayat, güncel, ekonomi, spor, dünya ve sağlık gibi farklı konularda yazan 4 kadın yazara ait 140 doküman alınarak oluşturulan Külliyyat-V üzerinde, 13 farklı ayrıştırma yöntemi kullanılarak deneyler yapılmıştır. Çizelge 8.9’ da sıkıştırma başarı sonuçları verilmiştir.

Külliyyat-V’ deki en iyi performans % 51 olup, metnin ekler ek halinde olmak üzere kök – ek(KET) şeklinde ayrıştırılması yönteminde 1024Kb boyutlu dosyanın sıkıştırılmasından elde edilmiştir. En kötü sonuç ise, metnin ekler ek halinde olmak üzere kök – ek(KET) şeklinde ayrıştırılması yöntemi ile metnin hece(H) şeklinde ayrıştırılması yönteminde 2Kb boyutlu dosyanın sıkıştırılmasından elde edilmiştir. 2Kb boyutlu dosyanın 2-GRAM ayrıştırma yöntemi ile sıkıştırılmasından da kötü sıkıştırma sonuçları ortaya çıkmıştır. Genel olarak başarının külliyyatın boyutuna ve ayrıştırma yöntemine bağlı olarak değişkenlik gösterdiği gözlemlenmiştir(Çizelge 8.9).

Tüm sıkıştırma sonuçlarının ayrıştırma yöntemine göre verdikleri sıkıştırma başarılarının ortalamaları alındığında karakter(K) ayrıştırma yönteminin % 38,5 ile en başarılı, ekler tek parça halinde olmak üzere fiil – ek(FET) ayrıştırma yönteminin % 28,6 ile bu külliyyatta en başarısız yöntem olduğu gözlemlenmiştir(Çizelge 8.9).

Çizelge 8.9 Külliyat-V için sıkıştırma oranları (%)

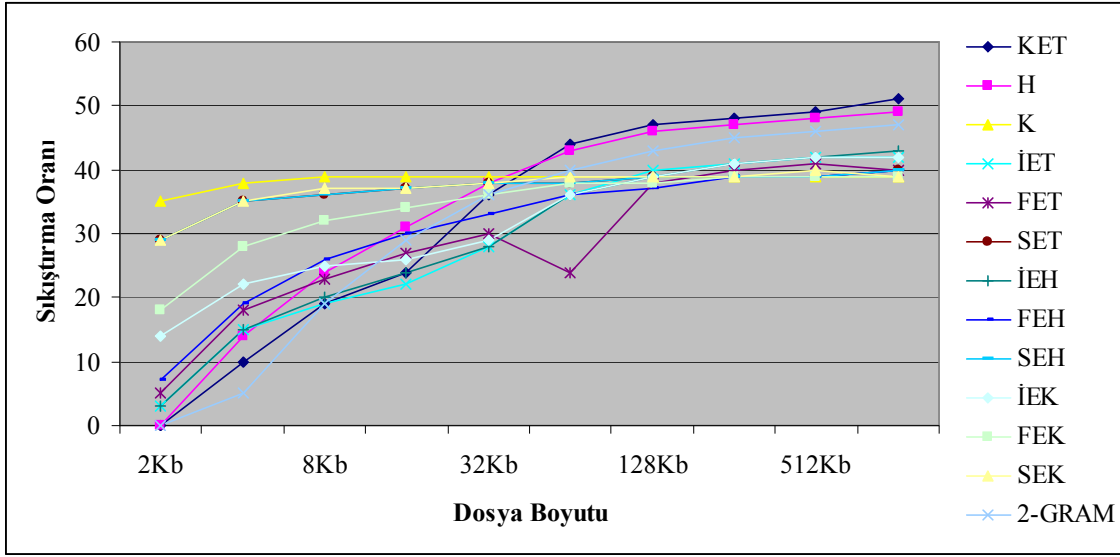
	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
KET	0	10	19	24	36	44	47	48	49	51	32,8
H	0	14	24	31	38	43	46	47	48	49	34
K	35	38	39	39	39	39	39	39	39	39	38,5
İET	3	15	19	22	28	36	40	41	42	42	28,8
FET	5	18	23	27	30	24	38	40	41	40	28,6
SET	29	35	36	37	38	38	39	39	39	40	37
İEH	3	15	20	24	28	36	39	41	42	43	29,1
FEH	7	19	26	30	33	36	37	39	39	40	30,6
SEH	29	35	36	37	38	38	39	39	39	40	37
İEK	14	22	25	26	29	36	39	41	42	42	31,6
FEK	18	28	32	34	36	38	38	39	39	39	34,1
SEK	29	35	37	37	38	39	39	39	40	39	37,2
2-GRAM	0	5	19	29	36	40	43	45	46	47	31
Ort Başarı %	13,23	22,23	27,31	30,54	34,39	37,46	40,23	41,31	41,92	42,39	

Tüm sıkıştırma sonuçlarının dosyanın boyutuna göre verdikleri sıkıştırma başarılarının ortalamaları alındığında 1024Kb boyutlu dosya ile yapılan sıkıştırma işleminin % 42,39 ile en başarılı, 2Kb boyutlu dosya ile yapılan sıkıştırma işleminin % 13,23 ile bu külliyatta en başarısız sonuçların alındığı dosya boyutu olduğu gözlemlenmiştir (Çizelge 8.9).

Çizelge 8.9’ da yer alan 0(sıfır) ifadesi, sıkıştırma işlemi uygulandığında sıkıştırma olmayıp, başlık yapısının boyutundan dolayı genişleme olmasından kaynaklanmaktadır.

Popüler hayat, güncel, ekonomi, spor, dünya ve sağlık gibi farklı konularda yazan 4 kadın yazara ait 140 doküman alınarak oluşturulan Külliyat-V üzerinde, 13 farklı ayrıştırma yöntemi uygulanarak yapılan deneylerin grafiksel sonucu Şekil 8.9’ da verilmiştir.

Külliyat-V heterojen yapıya sahip diğer külliyatlara göre daha yüksek performans vermiştir. Bu külliyat sadece kadın yazarlardan oluşmaktadır. 1024Kb dosya boyutunda daha yüksek başarı elde edilmiş olması kadın yazarların erkeklere göre daha az sayıda kök kelime kullandığını göstermektedir.



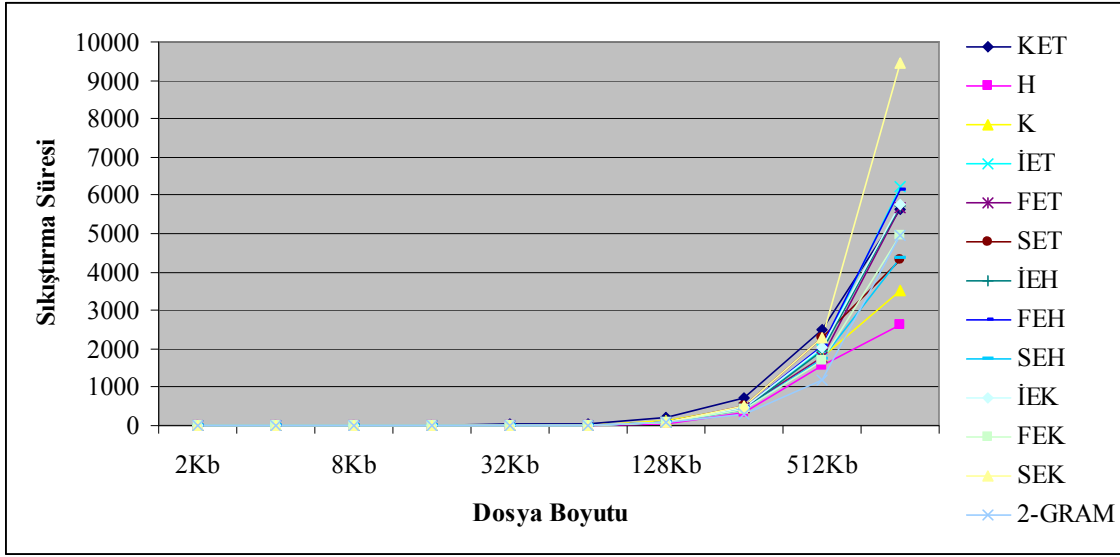
Şekil 8.9 Külliyyat-V için sıkıştırma oranları (%)

Çizelge 8.10' de Külliyyat-V' de yer alan her dosyanın sıkıştırma ve geri açma süreleri saniye cinsinden verilmiştir ve ön işleme, Huffman ağacının oluşturulması, sıkıştırma ve sıkıştırılan verinin açılması sürelerinin toplamını içermektedir.

Çizelge 8.10 Külliyyat-V için toplam sıkıştırma süreleri (sn)

	Dosya Boyutu									
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb
KET	1,72	2,481	5,15	13,495	26,069	53,792	209,12	709,161	2500,493	5621,458
H	0,13	0,22	0,73	2,216	4,338	9,596	57,019	360,095	1554,151	2643,589
K	0,09	0,11	0,27	0,955	3,401	12,497	106,361	459,015	1763,377	3526,437
İET	0,08	0,358	0,5	1,841	5,118	12,559	65,537	449,624	1991,451	6235,486
FET	0,08	0,204	0,48	1,563	4,274	13,026	69,327	465,411	1763,57	5693,248
SET	0,08	0,155	0,39	1,139	3,807	13,214	79,108	511,026	2292,716	4329,536
İEH	0,06	0,187	0,47	1,749	4,994	12,387	66,114	427,161	1967,349	5732,915
FEH	0,08	0,203	0,55	1,358	3,931	12,404	69,607	461,279	2069,388	6128,349
SEH	0,03	0,141	0,38	1,155	3,808	13,229	79,529	420,546	1738,355	4361,973
İEK	0,05	0,174	0,42	1,514	4,729	12,855	67,107	450,061	2029,529	5743,196
FEK	0,06	0,158	0,39	1,171	3,791	12,808	76,051	432,619	1675,551	4967,135
SEK	0,09	0,156	0,36	1,173	3,808	13,198	79,264	512,524	2295,448	9452,964
2-GRAM	0,27	0,441	0,92	2,605	5,929	14,65	66,062	276,136	1169,957	4972,545

4 farklı kadın yazarın yazılarının sıkıştırılması sırasında elde edilen sıkıştırma zamanı sonuçlarının saniye cinsinden grafiksel gösterimi Şekil 8.10' da verilmiştir.



Şekil 8.10 Külliyyat-V için toplam sıkıştırma süreleri (sn)

8.1.1.6 Külliyyat-VI için Deneysel Sonuçlar

Güncel konularda yazan 2 kadın yazara ait 98 doküman alınarak oluşturulan Külliyyat-VI üzerinde, 13 farklı ayrıştırma yöntemi kullanılarak deneyler yapılmıştır. Çizelge 8.11’ de sıkıştırma başarı sonuçları verilmiştir.

Külliyyat-VI’ daki en iyi performans % 49 olup, metnin ekler ek halinde olmak üzere kök – ek(KET) şeklinde ayrıştırılması yönteminde 1024Kb boyutlu dosyanın sıkıştırılmasından elde edilmiştir. En kötü sıkıştırma sonuçları ise, 2Kb boyutlu dosyanın 2-GRAM ayrıştırma yöntemi ile sıkıştırılmasından elde edilmiştir. Genel olarak başarının külliyyatın boyutuna ve ayrıştırma yöntemine bağlı olarak değişkenlik gösterdiği gözlemlenmiştir(Çizelge 8.11).

Tüm sıkıştırma sonuçlarının ayrıştırma yöntemine göre verdikleri sıkıştırma başarılarının ortalamaları alındığında karakter(K) ayrıştırma yönteminin % 39 ile en başarılı, ekler tek parça halinde olmak üzere isim – ek(İET) ayrıştırma yönteminin % 29,3 ile bu külliyyatta en başarısız yöntem olduğu gözlemlenmiştir(Çizelge 8.11).

Tüm sıkıştırma sonuçlarının dosyanın boyutuna göre verdikleri sıkıştırma başarılarının ortalamaları alındığında 1024Kb boyutlu dosya ile yapılan sıkıştırma işleminin % 42,54 ile en başarılı, 2Kb boyutlu dosya ile yapılan sıkıştırma işleminin % 18,77 ile bu külliyyatta en başarısız sonuçların alındığı dosya boyutu olduğu gözlemlenmiştir(Çizelge 8.11).

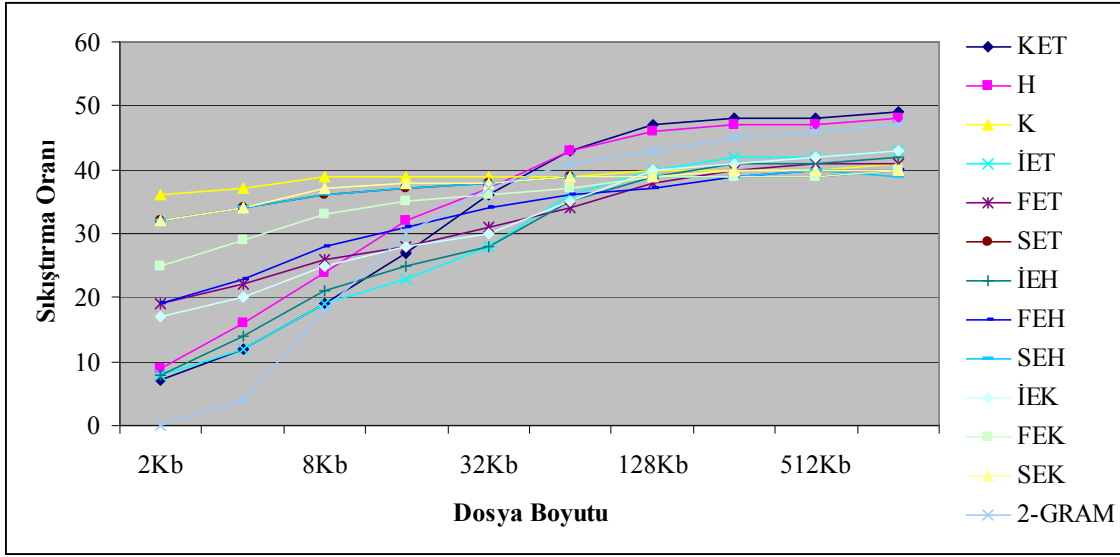
Çizelge 8.11 Külliyyat-VI için sıkıştırma oranları (%)

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
KET	7	12	19	27	36	43	47	48	48	49	33,6
H	9	16	24	32	37	43	46	47	47	48	34,9
K	36	37	39	39	39	39	40	40	40	41	39
İET	8	12	19	23	28	36	40	42	42	43	29,3
FET	19	22	26	28	31	34	38	40	41	41	32
SET	32	34	36	37	38	39	39	39	39	40	37,3
İEH	8	14	21	25	28	35	39	41	41	42	29,4
FEH	19	23	28	31	34	36	37	39	40	40	32,7
SEH	32	34	36	37	38	39	39	39	40	39	37,3
İEK	17	20	25	28	30	35	40	41	42	43	32,1
FEK	25	29	33	35	36	37	39	39	39	40	35,2
SEK	32	34	37	38	38	39	39	40	40	40	37,7
2-GRAM	0	4	18	30	37	41	43	45	46	47	31,1
Ort Başarı %	18,77	22,39	27,77	31,54	34,62	38,15	40,46	41,54	41,92	42,54	

Çizelge 8.11’ de yer alan 0(sıfır) ifadesi, sıkıştırma işlemi uygulandığında sıkıştırma olmayıp, başlık yapısının boyutundan dolayı genişleme olmasından kaynaklanmaktadır.

Güncel konularda yazan 2 kadın yazara ait 98 doküman alınarak oluşturulan Külliyyat-VI üzerinde, 13 farklı ayrıştırma yöntemi uygulanılarak yapılan deneylerin grafiksel sonucu Şekil 8.11’ de verilmiştir.

Külliyyat-VI tek bir konuda yazan kadın yazarlardan oluşurken Külliyyat-II erkek yazarlardan oluşmaktadır. Aynı dosya boyutu aynı yöntem ile her iki külliyyatta da en başarılı sonucu vermesine rağmen, Külliyyat-II daha başarılı sonuçlar vermiştir.



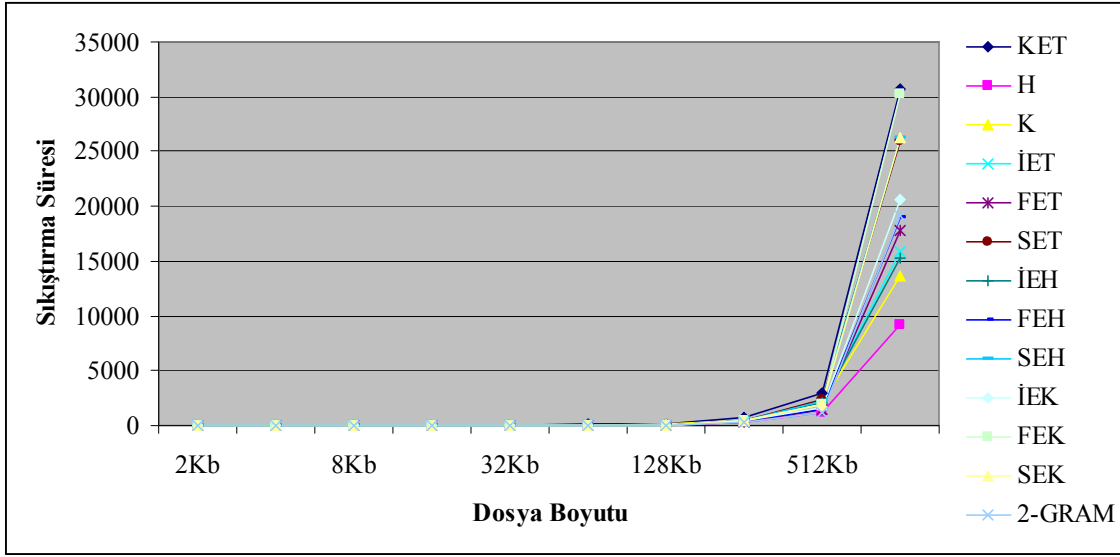
Şekil 8.11 Külliyyat-VI için sıkıştırma oranları (%)

Çizelge 8.12’ de Külliyyat-VI’ da yer alan her dosyanın sıkıştırma ve geri açma süreleri saniye cinsinden verilmiştir ve ön işleme, Huffman ağacının oluşturulması, sıkıştırma ve sıkıştırılan verinin açılması sürelerinin toplamını içermektedir.

Çizelge 8.12 Külliyyat-VI için toplam sıkıştırma süreleri (sn)

	Dosya Boyutu									
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb
KET	1,15	3,29	8,19	16,832	33,493	77,192	213,284	677,463	2991,598	30714,284
H	0,11	0,19	0,55	1,81	4,275	9,861	46,613	350,533	1228,032	9183,514
K	0,06	0,09	0,22	0,889	3,308	12,607	75,396	466,831	2125,984	13654,789
İET	0,13	0,22	0,45	1,763	5,134	12,402	57,986	447,612	2045,396	15854,749
FET	0,12	0,16	0,37	1,325	4,071	12,856	62,495	322,421	1500,331	17722,552
SET	0,09	0,16	0,34	1,217	3,699	13,184	71,497	505,032	2316,353	25923,954
İEH	0,08	0,19	0,42	1,561	4,962	12,419	58,143	440,28	2004,725	15207,116
FEH	0,06	0,16	0,39	1,266	3,761	12,262	63,666	462,384	1492,64	18920,929
SEH	0,05	0,16	0,34	1,123	3,666	13,167	70,997	509,917	2189,058	26210,069
İEK	0,11	0,14	0,37	1,577	4,557	12,98	60,749	391,545	1637,767	20601,536
FEK	0,08	0,31	0,47	1,093	3,604	12,872	68,922	481,464	1970,983	30322,453
SEK	0,05	0,16	0,34	1,094	3,729	13,199	71,652	511,431	1863,187	26186,136
2-GRAM	0,25	0,44	1	2,326	5,57	14,259	63,77	277,945	1166,941	19065,524

2 farklı kadın yazarın yazılarının sıkıştırılması sırasında elde edilen sıkıştırma zamanı sonuçlarının saniye cinsinden grafiksel gösterimi Şekil 8.12’ de verilmiştir.



Şekil 8.12 Külliyyat-VI için toplam sıkıştırma süreleri (sn)

8.1.1.7 Külliyyat-VII için Deneysel Sonuçlar

Türkiye Anayasa metni ile 6 öykü ve 1 romana ait 46 doküman alınarak oluşturulan Külliyyat-VII üzerinde, 13 farklı ayrıştırma yöntemi kullanılarak deneyler yapılmıştır. Çizelge 8.13’ de sıkıştırma başarı sonuçları verilmiştir.

Külliyyat-VII’ deki en iyi performans % 51 olup, metnin hece(H) şeklinde ayrıştırılması yönteminde 1024Kb boyutlu dosyanın sıkıştırılmasından elde edilmiştir. En kötü sonuç ise, metnin ekler ek halinde olmak üzere kök – ek(KET) şeklinde ayrıştırılması yöntemi, metnin hece(H) şeklinde ayrıştırılması yöntemi, metnin ekler tek parça halinde olmak üzere isim – ek(İET) şeklinde ayrıştırılması yöntemi ile metnin ekler hece şeklinde olmak üzere isim – ek(İEH) şeklinde ayrıştırılması yönteminde 2Kb boyutlu dosyanın sıkıştırılmasından elde edilmiştir. 2Kb ve 4Kb boyutlu dosyaların 2-GRAM ayrıştırma yöntemi ile sıkıştırılmasından da kötü sıkıştırma sonuçları ortaya çıkmıştır. Genel olarak başarının külliyyatın boyutuna ve ayrıştırma yöntemine bağlı olarak değişkenlik gösterdiği gözlemlenmiştir. Bununla birlikte bütünlük arz eden dokümanlarda isim, kök ve kullanılan hece sayısının çok fazla olması başarı sonucunu eksi yönde etkilemektedir(Çizelge 8.13).

Tüm sıkıştırma sonuçlarının ayrıştırma yöntemine göre verdikleri sıkıştırma başarılarının ortalamaları alındığında karakter(K) ayrıştırma yönteminin % 39,5 ile en başarılı, ekler tek parça halinde olmak üzere isim – ek(İET) ayrıştırma yönteminin % 28,8 ile bu külliyyatta en başarısız yöntem olduğu gözlemlenmiştir(Çizelge 8.13).

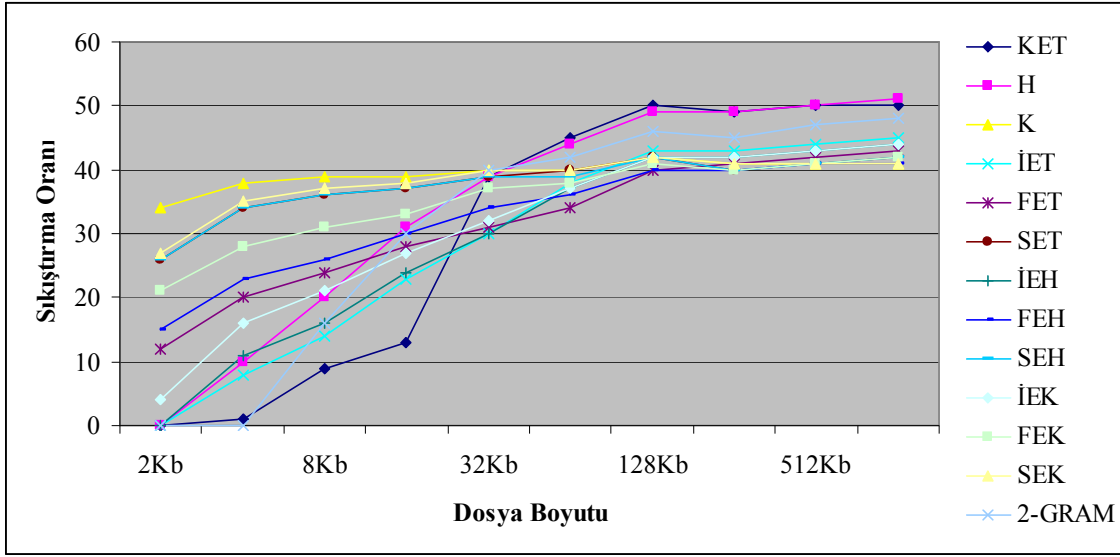
Çizelge 8.13 Külliyyat-VII için sıkıştırma oranları (%)

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
KET	0	1	9	13	39	45	50	49	50	50	30,6
H	0	10	20	31	39	44	49	49	50	51	34,3
K	34	38	39	39	40	40	42	41	41	41	39,5
İET	0	8	14	23	30	38	43	43	44	45	28,8
FET	12	20	24	28	31	34	40	41	42	43	31,5
SET	26	34	36	37	39	40	42	40	41	42	37,7
İEH	0	11	16	24	30	37	42	42	43	44	28,9
FEH	15	23	26	30	34	36	40	40	41	41	32,6
SEH	26	34	36	37	39	39	42	40	41	42	37,6
İEK	4	16	21	27	32	37	42	42	43	44	30,8
FEK	21	28	31	33	37	38	41	40	41	42	35,2
SEK	27	35	37	38	40	40	42	41	41	41	38,2
2-GRAM	0	0	16	30	40	42	46	45	47	48	31,4
Ort Başarı %	12,69	19,85	25	30	36,15	39,23	43,15	42,54	43,46	44,15	

Çizelge 8.13’ de yer alan 0(sıfır) ifadesi, sıkıştırma işlemi uygulandığında sıkıştırma olmayıp, başlık yapısının boyutundan dolayı genişleme olmasından kaynaklanmaktadır.

Tüm sıkıştırma sonuçlarının dosyanın boyutuna göre verdikleri sıkıştırma başarılarının ortalamaları alındığında 1024Kb boyutlu dosya ile yapılan sıkıştırma işleminin % 44,15 ile en başarılı, 2Kb boyutlu dosya ile yapılan sıkıştırma işleminin % 12,69 ile bu külliyyatta en başarısız sonuçların alındığı dosya boyutu olduğu gözlemlenmiştir(Çizelge 8.13).

Türkiye Anayasa metni ile 6 öykü ve roman yazarına ait 46 doküman alınarak oluşturulan Külliyyat-VII üzerinde, 13 farklı ayrıştırma yöntemi uygulanarak yapılan deneylerin grafiksel sonucu Şekil 8.13’ de verilmiştir.



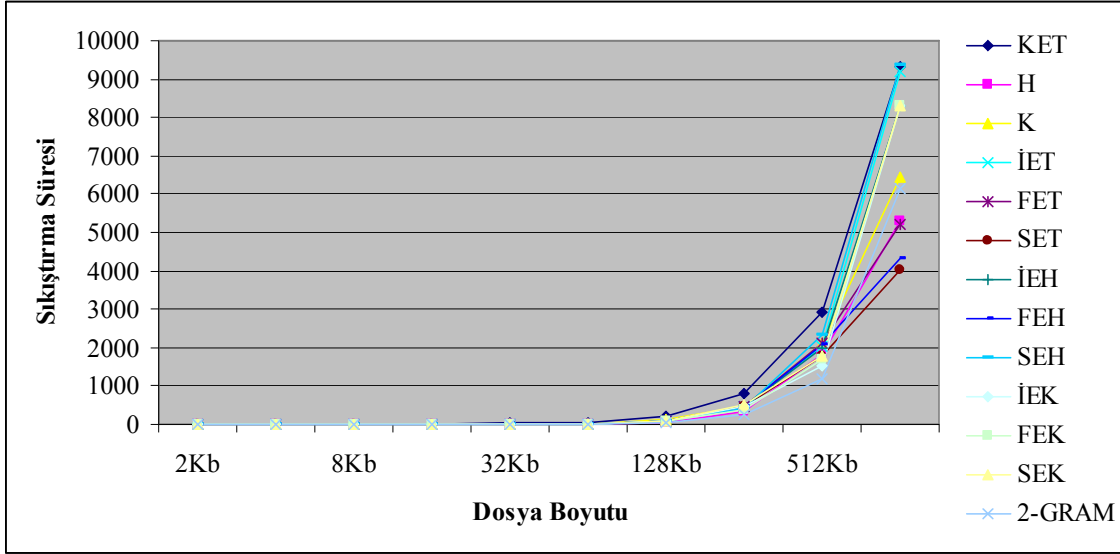
Şekil 8.13 Külliyyat-VII için sıkıştırma oranları (%)

Çizelge 8.14' de Külliyyat-VII' de yer alan her dosyanın sıkıştırma ve geri açma süreleri saniye cinsinden verilmiştir ve ön işleme, Huffman ağacının oluşturulması, sıkıştırma ve sıkıştırılan verinin açılması sürelerinin toplamını içermektedir.

Çizelge 8.14 Külliyyat-VII için toplam sıkıştırma süreleri(sn)

	Dosya Boyutu									
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb
KET	1,386	3,963	9,37	19,199	26,476	63,405	211,693	824,169	2930,303	9325,169
H	0,125	0,241	0,762	1,938	4,599	9,585	57,639	323,674	1803,408	5316,752
K	0,048	0,118	0,258	0,969	3,539	13,456	109,552	472,34	2117,841	6432,698
İET	0,105	0,259	0,691	2,256	5,241	13,54	63,772	451,918	2013,585	9201,364
FET	0,088	0,204	0,457	1,438	4,36	14,252	70,836	473,156	2105,658	5213,627
SET	0,067	0,167	0,368	1,168	3,893	13,893	81,003	478,411	1774,036	4023,816
İEH	0,083	0,231	0,619	1,807	4,807	12,437	64,092	446,792	1991,575	8301,691
FEH	0,067	0,19	0,41	1,354	3,981	12,605	71,071	468,894	2077,732	4309,745
SEH	0,059	0,164	0,487	1,171	4,551	13,547	81,128	411,007	2316,851	9374,619
İEK	0,075	0,204	0,439	1,582	4,515	13,165	70,686	458,29	1533,793	8306,419
FEK	0,069	0,172	0,381	1,175	3,817	12,82	78,988	391,444	1686,486	8307,612
SEK	0,063	0,164	0,36	1,203	3,926	13,522	81,973	518,184	1776,513	8304,628
2-GRAM	0,281	0,64	1,062	2,449	5,055	12,934	57,876	275,029	1203,15	6137,26

Türkiye Anayasa metni ile 6 öykü ve roman yazarın yazılarının sıkıştırılması sırasında elde edilen sıkıştırma zamanı sonuçlarının saniye cinsinden grafiksel gösterimi Şekil 8.14’ de verilmiştir.



Şekil 8.14 Külliyyat-VII için toplam sıkıştırma süreleri(sn)

8.1.1.8 Külliyyat-VIII için Deneysel Sonuçlar

Çeşitli kişilere ait 6 farklı tez dokümanı alınarak oluşturulan Külliyyat-VIII üzerinde, 13 farklı ayrıştırma yöntemi kullanılarak deneyler yapılmıştır. Çizelge 8.15’ de sıkıştırma başarı sonuçları verilmiştir.

Külliyyat-VIII’ deki en iyi performans % 54 olup, metnin ekler ek halinde olmak üzere kök – ek(KET) şeklinde ayrıştırılması yönteminde 1024Kb boyutlu dosyanın sıkıştırılmasından elde edilmiştir. En kötü sonuç ise, metnin hece(H) şeklinde ayrıştırılması yönteminde 2Kb boyutlu dosyanın sıkıştırılmasından elde edilmiştir. 2Kb boyutlu dosyanın 2-GRAM ayrıştırma yöntemi ile sıkıştırılmasından da kötü sıkıştırma sonuçları ortaya çıkmıştır. Genel olarak başarının külliyyatın boyutuna ve ayrıştırma yöntemine bağlı olarak değişkenlik gösterdiği gözlemlenmiştir(Çizelge 8.15).

Bu sonuçlar bize belli bir alanda yazılmış akademik çalışmalarda karşımıza çok fazla sayıda hece çıktığını göstermektedir. KET’ in başarısının yüksek olması da belli konuda yazıldığından çok farklı sayıda kökün olmadığını göstermektedir

Tüm sıkıştırma sonuçlarının ayrıştırma yöntemine göre verdikleri sıkıştırma başarılarının ortalamaları alındığında karakter(K) ayrıştırma yönteminin % 42,3 ile en başarılı, ekler tek parça halinde olmak üzere fiil – ek(FET) ayrıştırma yönteminin % 32,6 ile bu külliyatta en başarısız yöntem olduğu gözlemlenmiştir(Çizelge 8.15).

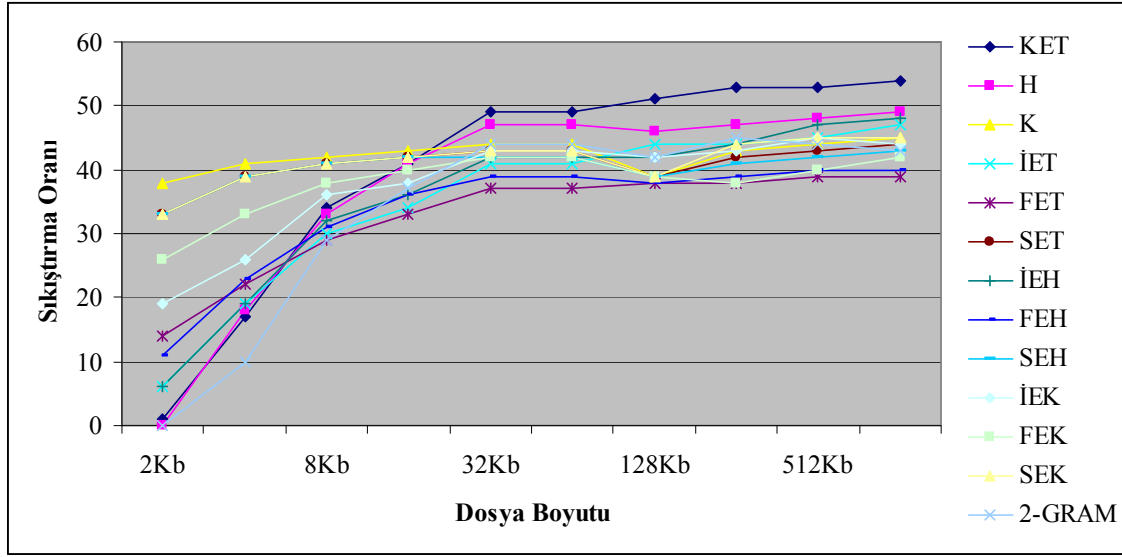
Tüm sıkıştırma sonuçlarının dosyanın boyutuna göre verdikleri sıkıştırma başarılarının ortalamaları alındığında 1024Kb boyutlu dosya ile yapılan sıkıştırma işleminin % 44,85 ile en başarılı, 2Kb boyutlu dosya ile yapılan sıkıştırma işleminin % 16,92 ile bu külliyatta en başarısız sonuçların alındığı dosya boyutu olduğu gözlemlenmiştir(Çizelge 8.15).

Çizelge 8.15’ de yer alan 0(sıfır) ifadesi, sıkıştırma işlemi uygulandığında sıkıştırma olmayıp, başlık yapısının boyutundan dolayı genişleme olmasından kaynaklanmaktadır.

Çizelge 8.15 Külliyat-VIII için sıkıştırma oranları (%)

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
KET	1	17	34	41	49	49	51	53	53	54	40,2
H	0	18	33	41	47	47	46	47	48	49	37,6
K	38	41	42	43	44	44	39	43	44	45	42,3
İET	6	19	30	34	41	41	44	44	45	47	35,1
FET	14	22	29	33	37	37	38	38	39	39	32,6
SET	33	39	41	42	43	43	39	42	43	44	40,9
İEH	6	19	32	36	42	42	42	44	47	48	35,8
FEH	11	23	31	36	39	39	38	39	40	40	33,6
SEH	33	39	41	42	42	42	39	41	42	43	40,4
İEK	19	26	36	38	43	43	42	43	45	44	37,9
FEK	26	33	38	40	42	42	39	38	40	42	38
SEK	33	39	41	42	43	43	39	44	45	45	41,4
2-GRAM	0	10	29	37	44	44	42	45	44	43	33,8
Ort Başarı %	16,92	26,54	35,15	38,85	42,77	42,77	41,39	43,15	44,23	44,85	

Çeşitli kişilere ait 6 farklı tez dokümanı alınarak oluşturulan Külliyat-VIII üzerinde, 13 farklı ayrıştırma yöntemi uygulanarak yapılan deneylerin grafiksel sonucu Şekil 8.15’ de verilmiştir.



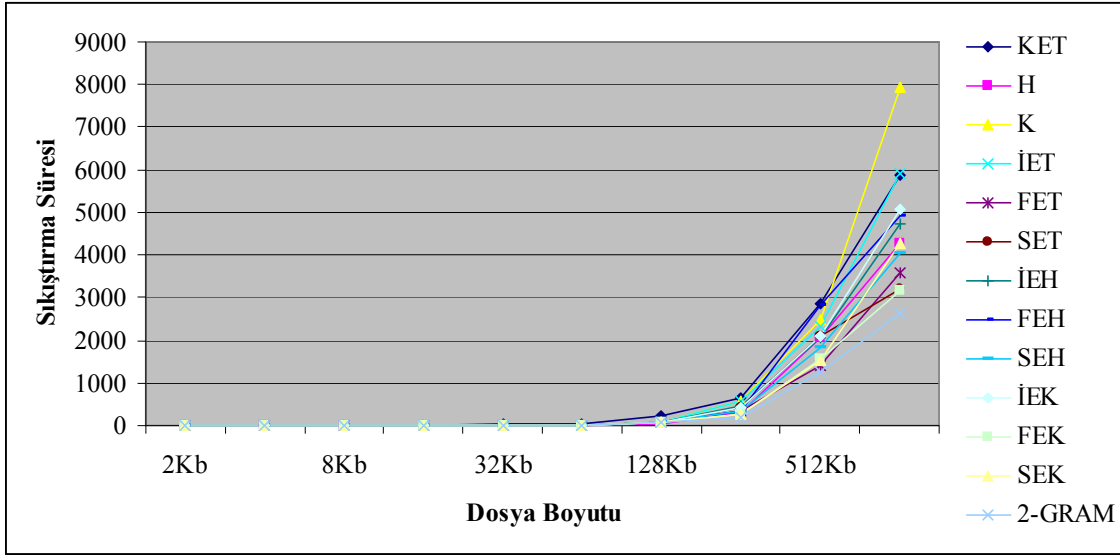
Şekil 8.15 Külliyyat-VIII için sıkıştırma oranları (%)

Çizelge 8.16’ de Külliyyat-VIII’ de yer alan her dosyanın sıkıştırma ve geri açma süreleri saniye cinsinden verilmiştir ve ön işleme, Huffman ağacının oluşturulması, sıkıştırma ve sıkıştırılan verinin açılması sürelerinin toplamını içermektedir.

Çizelge 8.16 Külliyyat-VIII için toplam sıkıştırma süreleri(sn)

	Dosya Boyutu									
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb
KET	1,047	2,028	5,165	11,981	22,386	20,327	232,847	632,917	2854,649	5871,676
H	0,095	0,186	0,361	0,579	3,044	2,951	48,548	289,083	2045,596	4276,424
K	0,032	0,126	0,281	1	3,324	3,308	87,767	559,281	2463,55	7921,391
İET	0,076	0,171	0,425	1,249	3,604	3,495	57,907	573,737	2311,27	5924,65
FET	0,08	0,154	0,436	1,14	3,719	3,651	71,245	356,035	1425,834	3601,941
SET	0,063	0,14	0,407	1,077	3,635	3,762	80,982	438,07	2078,726	3201,182
İEH	0,078	0,187	0,437	1,031	3,386	3,168	59,546	451,953	2064,462	4735,124
FEH	0,063	0,171	0,408	1,17	3,417	3,588	72,557	328,723	2823,099	4925,814
SEH	0,047	0,155	0,389	1,061	3,716	3,684	81,887	358,879	1829,99	4027,125
İEK	0,063	0,157	0,375	1,015	3,183	3,214	66,13	426,094	2081,643	5075,168
FEK	0,049	0,155	0,405	1,062	3,593	3,668	78,625	263,534	1555,786	3165,248
SEK	0,078	0,155	0,39	1,094	3,619	3,667	81,183	262,823	1516,677	4263,841
2-GRAM	0,204	0,297	0,781	1,484	3,73	3,761	66,442	194,438	1276,458	2637,49

Çeşitli kişilere ait 6 farklı tez yazılarının sıkıştırılması sırasında elde edilen sıkıştırma zamanı sonuçlarının saniye cinsinden grafiksel gösterimi Şekil 8.16’ da verilmiştir.



Şekil 8.16 Külliyat-VIII için toplam sıkıştırma süreleri(sn)

Tüm külliyatlar için sıkıştırma işlemi yapıp, sıkıştırılan dosya açıldığında, oluşan dosya ile orijinal dosyanın eşleşme yapısı Çizelge 8.17’ de gönderilmiştir. 0 sayısı % 100 eşleşmenin olmadığı, 1 ise % 100 eşleşmenin sağlandığı anlamına gelmektedir. KET ayrıştırma yönteminde % 100 eşleşmenin olmaması Zemberek isimli araçtan kaynaklanmaktadır. Zemberek ile kök – ek ayrımı işlemi yapıldığında, hece düşmelerinden kaynaklanan sorunlar oluşmaktadır. Hece düşmesi olan kelimelerde Zemberek’ ten birkaç farklı öneri gelmektedir. Orijinal kelime bazen önerilen birinci sonuç olurken bazen ikinci sonuç olmaktadır. Programsal olarak daha başarılı sonuçlar elde edildiğinden ikinci öneri kullanılmasına rağmen doğru sonucun birinci öneriden alındığı kelimelerde bu seçim soruna yol açmıştır.

Sonuç olarak, 8 adet külliyat ile yapılan çalışmalar şunu göstermiştir ki dosya boyutu ile alınan sıkıştırma başarısı doğru orantılıdır. Külliyatların çoğunda – Külliyat-III – hariç ortalama olarak en yüksek başarı K ile yapılan ayrıştırma şeklinde alınmıştır. Külliyat-III ve Külliyat-VII hariç yöntem bazında da KET ile daha başarılı sonuçlar elde edilmiştir. Külliyat-III’ de ise ortalamada SEH daha başarılı olmuştur.

Çizelge 8.17 Orijinal dosya ile sıkıştırılıp açılan dosyanın içeriklerinin karşılaştırılması

[illegible]

8.2 İngilizce Üzerine Yapılan Çalışmalar

Bu bölümde İngilizce üzerine yapılan çalışmalar sırasıyla verilmiştir.

8.2.1 Külliyata Göre Deneysel Sonuçlar

Bu bölümde Külliyat-IX olarak hazırladığımız küme ile deneysel sonuçlar alınmıştır.

8.2.1.1 Külliyat-IX için Deneysel Sonuçlar

İngilizce külliyat üzerinde sıkıştırma işlemi yapılırken, metinler biçimbilimsel yapılarına göre, ekler ek halinde olmak üzere kök – ek(İKET), ekler karakter olmak üzere kök – ek(İKEK), karakter(İK) ve bigram(2-GRAM) olarak 4 farklı şekilde ayrıştırılmışlardır. Güncel konularda yazan farklı kişilere ait 290 farklı dokümandan alınan metinler 2Kb, 4Kb, 8Kb, 16Kb, 32Kb, 64Kb, 128Kb, 256Kb, 512Kb, 1024Kb boyutlarındaki dosyalara ayrıştırılmış ve her bir dosya 4 yöntem ile sıkıştırılıp, Çizelge 8.18’ deki deneysel sıkıştırma sonuçları elde edilmiştir.

Külliyat-IX’ daki en iyi performans % 54 olup, metnin ekler karakter halinde olmak üzere kök – ek(İKEK) şeklinde ayrıştırılması yönteminde 64Kb boyutlu dosyanın sıkıştırılmasından elde edilmiştir. En kötü sonuç ise, 2Kb boyutlu dosyanın 2-GRAM ayrıştırma yöntemi ile sıkıştırılmasından elde edilmiştir. Genel olarak başarının külliyatın boyutuna ve ayrıştırma yöntemine bağlı olarak değişkenlik gösterdiği gözlemlenmiştir(Çizelge 8.18).

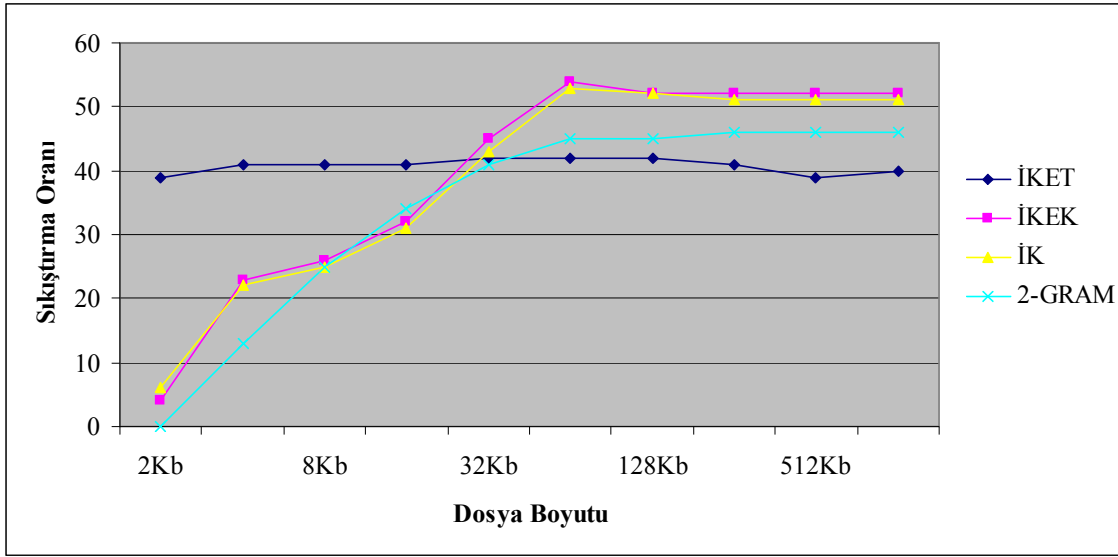
Tüm sıkıştırma sonuçlarının ayrıştırma yöntemine göre verdikleri sıkıştırma başarılarının ortalamaları alındığında ekler ek halinde olmak üzere kök - ek(İKET) ayrıştırma yönteminin % 40,8 ile en başarılı, 2-GRAM ayrıştırma yönteminin % 34,1 ile bu külliyatta en başarısız yöntem olduğu gözlemlenmiştir (Çizelge 8.18).

Çizelge 8.18 Külliyat-IX için sıkıştırma oranları (%)

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
İKET	39	41	41	41	42	42	42	41	39	40	40,8
İKEK	4	23	26	32	45	54	52	52	52	52	39,2
İK	6	22	25	31	43	53	52	51	51	51	38,5
2-GRAM	0	13	25	34	41	45	45	46	46	46	34,1
Ort Başarı %	12,3	24,75	29,25	34,5	42,75	48,5	47,75	47,5	47	47,25	

Tüm sıkıştırma sonuçlarının dosyanın boyutuna göre verdikleri sıkıştırma başarılarının ortalamaları alındığında 64Kb boyutlu dosya ile yapılan sıkıştırma işleminin % 48,5 ile en başarılı, 2Kb boyutlu dosya ile yapılan sıkıştırma işleminin % 12,3 ile bu külliyatta en başarısız sonuçların alındığı dosya boyutu olduğu gözlemlenmiştir (Çizelge 8.18).

Güncel konularda yazılan 290 farklı doküman alınarak oluşturulan Külliyat-IX üzerinde, 3 farklı ayrıştırma yöntemi uygulanarak yapılan deneylerin grafiksel sonucu Şekil 8.17’ de verilmiştir. İKET, her dosya boyutunda aynı sonuçları vermiş olmasına rağmen, İKEK, İK ve 2-GRAM yöntemleri 32Kb’ den sonra gözle görülür bir şekilde artış göstermişlerdir.



Şekil 8.17 Külliyat-IX için sıkıştırma oranları (%)

Çizelge 8.19’ da farklı kişilere ait güncel konulardaki yazıların sıkıştırılması sırasında elde edilen sıkıştırma zamanı sonuçları verilmiştir. Çizelge 8.19’ da yer alan zaman değerleri saniye cinsinden olmakta ve ön işleme, Huffman ağacının oluşturulması, sıkıştırma ve sıkıştırılan verinin açılması sürelerinin toplamını içermektedir.

Çizelge 8.19 Külliyat-IX için toplam sıkıştırma süreleri (sn)

	Dosya Boyutu									
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb
İKET	0,12	0,154	0,279	0,918	3,386	12,99	84,326	493,934	2251,736	4264,32
İKEK	0,328	0,218	0,624	2,169	3,853	8,658	83,664	397,161	1605,246	10101,94
İK	0,282	0,232	0,618	2,056	4,242	9,52	84,56	415,028	1826,855	1417,011
2-GRAM	0,235	0,547	0,702	1,935	4,228	10,998	57,565	268,805	1163,782	1324,624

8.3 Ayırıştırma Yöntemine Dayalı Sıkıştırma Sonuçlarının Karşılaştırılması

Bu bölümde KET, H, K, İET, FET, SET, İEH, FEH, SEH, İEK, FEK, SEK, 2-GRAM ayırıştırma yöntemleri ile yapılan sıkıştırma işlemlerinde Külliylatlara bağlı olarak elde edilen sonuçlar karşılaştırılmıştır.

KET ayırıştırma yöntemi ile yapılan sıkıştırma işlemlerinde külliylatlara ve dosya boyutlarına bağlı olarak elde edilen değerler Çizelge 8.21’ de gösterilmektedir. Çizelge 8.21’ deki değerler incelendiğinde en iyi ortalama sıkıştırma başarısının Külliyyat-VIII ile elde edildiği görülmektedir. Ortalama bazında yapılan sıkıştırma sonuçları karşılaştırmasında, elde edilen bu sonuca göre, tez dokümanlarından elde edilen kök ve ek çeşidinin diğerlerine göre daha az olduğu ortaya çıkmıştır. Kök ve ek çeşidinin az olmasından ötürü sıkıştırma işleminde kullanılan başlık yapısının boyutu küçük olmuş ve bu nedenle sıkıştırma verimi yüksek çıkmıştır.

Çizelge 8.21 Külliylatlara bağlı olarak elde edilen KET ayırıştırma yöntemi sıkıştırma sonuçları

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
Külliyyat-I	0	8	15	26	39	45	48	49	49	50	32,9
Külliyyat-II	9	16	25	31	39	44	46	48	49	50	35,7
Külliyyat-III	0	8	15	26	39	45	48	49	49	50	32,9
Külliyyat-IV	2	15	20	27	37	44	48	55	60	61	36,9
Külliyyat-V	0	10	19	24	36	44	47	48	49	51	32,8
Külliyyat-VI	7	12	19	27	36	43	47	48	48	49	33,6
Külliyyat-VII	0	1	9	13	39	45	50	49	50	50	30,6
Külliyyat-VIII	1	17	34	41	49	49	51	53	53	54	40,2

H ayırıştırma yöntemi ile yapılan sıkıştırma işlemlerinde külliylatlara ve dosya boyutlarına bağlı olarak elde edilen değerler Çizelge 8.22’ de gösterilmektedir. Çizelge 8.22’ deki değerler incelendiğinde en iyi sıkıştırma başarılarının Külliyyat-VIII ile elde edildiği görülmektedir. Ortalama bazında yapılan sıkıştırma sonuçları karşılaştırmasında, elde edilen bu sonuca göre, tez dokümanlarından elde edilen hece çeşidinin diğerlerine göre daha az olduğu ortaya çıkmıştır. Hece çeşidinin az olmasından ötürü sıkıştırma işleminde kullanılan başlık yapısının boyutu küçük olmuş ve bu nedenle sıkıştırma verimi yüksek çıkmıştır.

Çizelge 8.22 Külliyatlara bağlı olarak elde edilen H ayrıştırma yöntemi sıkıştırma sonuçları

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
Külliyat-I	0	10	19	30	38	43	46	47	47	48	32,8
Külliyat-II	4	15	25	33	39	43	45	46	47	48	34,5
Külliyat-III	0	10	19	30	38	43	46	47	47	48	32,8
Külliyat-IV	2	16	23	31	37	43	47	54	59	60	37,2
Külliyat-V	0	14	24	31	38	43	46	47	48	49	34
Külliyat-VI	9	16	24	32	37	43	46	47	47	48	34,9
Külliyat-VII	0	10	20	31	39	44	49	49	50	51	34,3
Külliyat-VIII	0	18	33	41	47	47	46	47	48	49	37,6

K ayrıştırma yöntemi ile yapılan sıkıştırma işlemlerinde külliyatlara ve dosya boyutlarına bağlı olarak elde edilen değerler Çizelge 8.23’ de gösterilmektedir. Çizelge 8.23’ deki değerler incelendiğinde en iyi sıkıştırma başarılarının Külliyat-VIII külliyatından elde edildiği görülmektedir. Ortalama bazında yapılan sıkıştırma sonuçları karşılaştırmasında, elde edilen bu sonuca göre, tez dokümanlarından elde edilen karakter çeşidinin diğerlerine göre daha az olduğu ortaya çıkmıştır. Karakter çeşidinin az olmasından ötürü sıkıştırma işleminde kullanılan başlık yapısının boyutu küçük olmuş ve bu nedenle sıkıştırma verimi yüksek çıkmıştır.

Çizelge 8.23 Külliyatlara bağlı olarak elde edilen K ayrıştırma yöntemi sıkıştırma sonuçları

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
Külliyat-I	37	38	38	39	40	39	39	39	40	40	38,9
Külliyat-II	36	39	41	42	41	39	39	39	40	40	39,6
Külliyat-III	37	38	38	39	40	39	39	39	40	41	39
Külliyat-IV	35	37	38	38	39	39	41	47	50	51	41,5
Külliyat-V	35	38	39	39	39	39	39	39	39	39	38,5
Külliyat-VI	36	37	39	39	39	39	40	40	40	41	39
Külliyat-VII	34	38	39	39	40	40	42	41	41	41	39,5
Külliyat-VIII	38	41	42	43	44	44	39	43	44	45	42,3

İET ayrıştırma yöntemi ile yapılan sıkıştırma işlemlerinde külliyatlara ve dosya boyutlarına bağlı olarak elde edilen değerler Çizelge 8.24’ de gösterilmektedir. Çizelge 8.24’ deki değerler incelendiğinde en iyi sıkıştırma başarılarının Külliyat-VIII ile elde edildiği görülmektedir. Ortalama bazında yapılan sıkıştırma sonuçları karşılaştırmasında, elde edilen bu sonuca göre, tez dokümanlarından elde edilen isim ve ek çeşidinin diğerlerine göre daha az olduğu ortaya çıkmıştır. İsim ve ek çeşidinin az olmasından ötürü sıkıştırma işleminde kullanılan başlık yapısının boyutu küçük olmuş ve bu nedenle sıkıştırma verimi yüksek çıkmıştır.

Çizelge 8.24 Külliyatlara bağlı olarak elde edilen İET ayrıştırma yöntemi sıkıştırma sonuçları

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
Külliyat-I	3	12	16	23	31	37	41	43	43	44	29,3
Külliyat-II	14	18	23	27	32	38	40	42	44	44	32,2
Külliyat-III	3	12	16	23	31	37	41	43	43	43	29,2
Külliyat-IV	8	17	17	21	29	37	42	49	54	54	32,8
Külliyat-V	3	15	19	22	28	36	40	41	42	42	28,8
Külliyat-VI	8	12	19	23	28	36	40	42	42	43	29,3
Külliyat-VII	0	8	14	23	30	38	43	43	44	45	28,8
Külliyat-VIII	6	19	30	34	41	41	44	44	45	47	35,1

FET ayrıştırma yöntemi ile yapılan sıkıştırma işlemlerinde külliyatlara ve dosya boyutlarına bağlı olarak elde edilen değerler Çizelge 8.25’ de gösterilmektedir. Çizelge 8.25’ deki değerler incelendiğinde en iyi sıkıştırma başarılarının Külliyat-IV ile elde edildiği görülmektedir. Ortalama bazında yapılan sıkıştırma sonuçları karşılaştırmasında, elde edilen bu sonuca göre, siyaset konusunda yazan kadın ve erkek yazarların dokümanlarından elde edilen fiil ve ek çeşidinin diğerlerine göre daha az olduğu ortaya çıkmıştır. Fiil ve ek çeşidinin az olmasından ötürü sıkıştırma işleminde kullanılan başlık yapısının boyutu küçük olmuş ve bu nedenle sıkıştırma verimi yüksek çıkmıştır.

Çizelge 8.25 Külliyatlara bağlı olarak elde Edilen FET ayrıştırma yöntemi sıkıştırma sonuçları

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
Külliyat-I	15	21	24	29	31	33	37	39	40	41	31
Külliyat-II	14	25	32	34	34	33	36	39	40	40	32,7
Külliyat-III	15	21	24	29	31	33	37	39	40	40	30,9
Külliyat-IV	14	20	23	28	31	33	39	47	51	52	33,8
Külliyat-V	5	18	23	27	30	24	38	40	41	40	28,6
Külliyat-VI	19	22	26	28	31	34	38	40	41	41	32
Külliyat-VII	12	20	24	28	31	34	40	41	42	43	31,5
Külliyat-VIII	14	22	29	33	37	37	38	38	39	39	32,6

SET ayrıştırma yöntemi ile yapılan sıkıştırma işlemlerinde külliyatlara ve dosya boyutlarına bağlı olarak elde edilen değerler Çizelge 8.26’ de gösterilmektedir. Çizelge 8.26’ deki değerler incelendiğinde en iyi sıkıştırma başarılarının Külliyat-VIII ile elde edildiği görülmektedir. Ortalama bazında yapılan sıkıştırma sonuçları karşılaştırmasında, elde edilen bu sonuca göre, tez dokümanlarından elde edilen sıfat ve ek çeşidinin diğerlerine göre daha az olduğu ortaya çıkmıştır. Sıfat ve ek çeşidinin az olmasından ötürü sıkıştırma işleminde kullanılan başlık yapısının boyutu küçük olmuş ve bu nedenle sıkıştırma verimi yüksek çıkmıştır.

Çizelge 8.26 Külliyatlara bağlı olarak elde edilen SET ayrıştırma yöntemi sıkıştırma sonuçları

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
Külliyat-I	31	34	36	38	38	39	39	39	39	39	37,2
Külliyat-II	29	36	39	40	40	38	39	39	39	39	37,8
Külliyat-III	31	34	36	38	38	39	39	39	39	40	37,3
Külliyat-IV	30	34	35	36	38	38	40	46	50	51	39,8
Külliyat-V	29	35	36	37	38	38	39	39	39	40	37
Külliyat-VI	32	34	36	37	38	39	39	39	39	40	37,3
Külliyat-VII	26	34	36	37	39	40	42	40	41	42	37,7
Külliyat-VIII	33	39	41	42	43	43	39	42	43	44	40,9

İEH ayrıştırma yöntemi ile yapılan sıkıştırma işlemlerinde külliyatlara ve dosya boyutlarına bağlı olarak elde edilen değerler Çizelge 8.27’ de gösterilmektedir. Çizelge 8.27’ deki değerler incelendiğinde en iyi sıkıştırma başarılarının Külliyat-VIII ile elde edildiği görülmektedir. Ortalama bazında yapılan sıkıştırma sonuçları karşılaştırmasında, elde edilen bu sonuca göre, tez dokümanlarından elde edilen isim ve hece çeşidinin diğerlerine göre daha az olduğu ortaya çıkmıştır. İsim ve hece çeşidinin az olmasından ötürü sıkıştırma işleminde kullanılan başlık yapısının boyutu küçük olmuş ve bu nedenle sıkıştırma verimi yüksek çıkmıştır.

FEH ayrıştırma yöntemi ile yapılan sıkıştırma işlemlerinde külliyatlara ve dosya boyutlarına bağlı olarak elde edilen değerler Çizelge 8.28’ de gösterilmektedir. Çizelge 8.28’ deki değerler incelendiğinde en iyi sıkıştırma başarılarının Külliyat-VIII ile elde edildiği görülmektedir. Ortalama bazında yapılan sıkıştırma sonuçları karşılaştırmasında, elde edilen bu sonuca göre, tez dokümanlarından elde edilen fiil ve hece çeşidinin diğerlerine göre daha az olduğu ortaya çıkmıştır. Fiil ve hece çeşidinin az olmasından ötürü sıkıştırma işleminde kullanılan başlık yapısının boyutu küçük olmuş ve bu nedenle sıkıştırma verimi yüksek çıkmıştır.

Çizelge 8.27 Külliyatlara bağlı olarak elde edilen İEH ayrıştırma yöntemi sıkıştırma sonuçları

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
Külliyat-I	4	13	17	25	31	37	40	42	43	43	29,5
Külliyat-II	13	19	24	28	32	37	40	42	43	44	32,2
Külliyat-III	4	13	17	25	31	37	40	42	43	44	29,6
Külliyat-IV	11	20	20	24	29	36	42	49	54	54	33,9
Külliyat-V	3	15	20	24	28	36	39	41	42	43	29,1
Külliyat-VI	8	14	21	25	28	35	39	41	41	42	29,4
Külliyat-VII	0	11	16	24	30	37	42	42	43	44	28,9

Çizelge 8.28 Külliyatlara bağlı olarak elde edilen FEH ayrıştırma yöntemi sıkıştırma sonuçları

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
Külliyat-I	14	22	26	32	34	36	37	38	39	40	31,8
Külliyat-II	15	25	32	35	36	36	37	37	39	40	33,2
Külliyat-III	14	22	26	32	34	36	37	38	39	39	31,7
Külliyat-IV	14	22	26	31	34	36	39	46	50	50	34,8
Külliyat-V	7	19	26	30	33	36	37	39	39	40	30,6
Külliyat-VI	19	23	28	31	34	36	37	39	40	40	32,7
Külliyat-VII	15	23	26	30	34	36	40	40	41	41	32,6
Külliyat-VIII	11	23	31	36	39	39	38	39	40	40	33,6

SEH ayrıştırma yöntemi ile yapılan sıkıştırma işlemlerinde külliyatlara ve dosya boyutlarına bağlı olarak elde edilen değerler Çizelge 8.29’ de gösterilmektedir. Çizelge 8.29’ deki değerler incelendiğinde en iyi sıkıştırma başarılarının Külliyat-VIII ile elde edildiği görülmektedir. Ortalama bazında yapılan sıkıştırma sonuçları karşılaştırmasında, elde edilen bu sonuca göre, tez dokümanlarından elde edilen sıfat ve hece çeşidinin diğerlerine göre daha az olduğu ortaya çıkmıştır. Sıfat ve hece çeşidinin az olmasından ötürü sıkıştırma işleminde kullanılan başlık yapısının boyutu küçük olmuş ve bu nedenle sıkıştırma verimi yüksek çıkmıştır.

Çizelge 8.29 Külliyatlara bağlı olarak elde edilen SEH ayrıştırma yöntemi sıkıştırma sonuçları

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
Külliyat-I	31	34	36	38	38	39	39	39	40	41	37,5
Külliyat-II	29	36	39	40	40	38	39	39	39	40	37,9
Külliyat-III	31	34	36	38	38	39	39	39	40	41	37,5
Külliyat-IV	30	34	35	36	37	38	40	46	50	51	39,7
Külliyat-V	29	35	36	37	38	38	39	39	39	40	37
Külliyat-VI	32	34	36	37	38	39	39	39	40	39	37,3
Külliyat-VII	26	34	36	37	39	39	42	40	41	42	37,6
Külliyat-VIII	33	39	41	42	42	42	39	41	42	43	40,4

İEK ayrıştırma yöntemi ile yapılan sıkıştırma işlemlerinde külliyatlara ve dosya boyutlarına bağlı olarak elde edilen değerler Çizelge 8.30’ de gösterilmektedir. Çizelge 8.30’ deki

değerler incelendiğinde en iyi sıkıştırma başarılarının Külliyyat-VIII külliyyatından elde edildiği görülmektedir. Ortalama bazında yapılan sıkıştırma sonuçları karşılaştırmasında, elde edilen bu sonuca göre, tez dokümanlarından elde edilen isim ve karakter çeşidinin diğerlerine göre daha az olduğu ortaya çıkmıştır. İsim ve karakter çeşidinin az olmasından ötürü sıkıştırma işleminde kullanılan başlık yapısının boyutu küçük olmuş ve bu nedenle sıkıştırma verimi yüksek çıkmıştır.

Çizelge 8.30 Külliyyatlara bağlı olarak elde edilen İEK ayrıştırma yöntemi sıkıştırma sonuçları

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
Külliyyat-I	23	20	21	27	32	37	41	42	43	43	32,9
Külliyyat-II	20	24	27	30	32	37	40	42	43	43	33,8
Külliyyat-III	13	20	21	27	32	37	41	42	57	56	34,6
Külliyyat-IV	16	24	23	26	30	36	42	49	54	56	35,6
Külliyyat-V	14	22	25	26	29	36	39	41	42	42	31,6
Külliyyat-VI	17	20	25	28	30	35	40	41	42	43	32,1
Külliyyat-VII	4	16	21	27	32	37	42	42	43	44	30,8
Külliyyat-VIII	19	26	36	38	43	43	42	43	45	44	37,9

FEK ayrıştırma yöntemi ile yapılan sıkıştırma işlemlerinde külliyyatlara ve dosya boyutlarına bağlı olarak elde edilen değerler Çizelge 8.31’ de gösterilmektedir. Çizelge 8.31’ deki değerler incelendiğinde en iyi sıkıştırma başarılarının Külliyyat-IV ile elde edildiği görülmektedir. Ortalama bazında yapılan sıkıştırma sonuçları karşılaştırmasında, elde edilen bu sonuca göre, siyaset konusunda yazan kadın ve erkek yazarların dokümanlarından elde edilen fiil ve karakter çeşidinin diğerlerine göre daha az olduğu ortaya çıkmıştır. Fiil ve karakter çeşidinin az olmasından ötürü sıkıştırma işleminde kullanılan başlık yapısının boyutu küçük olmuş ve bu nedenle sıkıştırma verimi yüksek çıkmıştır.

SEK ayrıştırma yöntemi ile yapılan sıkıştırma işlemlerinde külliyyatlara ve dosya boyutlarına bağlı olarak elde edilen değerler Çizelge 8.32’ de gösterilmektedir. Çizelge 8.32’ deki değerler incelendiğinde en iyi sıkıştırma başarılarının Külliyyat-VIII ile elde edildiği görülmektedir. Ortalama bazında yapılan sıkıştırma sonuçları karşılaştırmasında, elde edilen bu sonuca göre, tez dokümanlarından elde edilen sıfat ve karakter çeşidinin diğerlerine göre daha az olduğu ortaya çıkmıştır. Sıfat ve karakter çeşidinin az olmasından ötürü sıkıştırma

işleminde kullanılan başlık yapısının boyutu küçük olmuş ve bu nedenle sıkıştırma verimi yüksek çıkmıştır.

Çizelge 8.31 Külliyatlara bağlı olarak elde edilen FEK ayrıştırma yöntemi sıkıştırma sonuçları

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
Külliyat-I	24	30	32	35	37	38	39	39	40	41	35,5
Külliyat-II	23	31	36	38	39	38	38	39	40	39	36,1
Külliyat-III	24	30	32	35	37	38	39	39	40	41	35,5
Külliyat-IV	25	30	32	34	36	38	40	47	51	52	38,5
Külliyat-V	18	28	32	34	36	38	38	39	39	39	34,1
Külliyat-VI	25	29	33	35	36	37	39	39	39	40	35,2
Külliyat-VII	21	28	31	33	37	38	41	40	41	42	35,2
Külliyat-VIII	26	33	38	40	42	42	39	38	40	42	38

2-GRAM ayrıştırma yöntemi ile yapılan sıkıştırma işlemlerinde külliyatlara ve dosya boyutlarına bağlı olarak elde edilen değerler Çizelge 8.33’ de gösterilmektedir. Çizelge 8.33’ deki değerler incelendiğinde en iyi sıkıştırma başarılarının Külliyat-VIII ile elde edildiği görülmektedir. Ortalama bazında yapılan sıkıştırma sonuçları karşılaştırmasında, elde edilen bu sonuca göre, tez dokümanlarından elde edilen 2-GRAM çeşidinin diğerlerine göre daha az olduğu ortaya çıkmıştır. 2-GRAM çeşidinin az olmasından ötürü sıkıştırma işleminde kullanılan başlık yapısının boyutu küçük olmuş ve bu nedenle sıkıştırma verimi yüksek çıkmıştır.

FET ve FEK ayrıştırma yöntemlerinde ortalama ve tek başına elde edilen başarılar Külliyat-IV ile alınmıştır. Bu külliyat sadece siyaset alanında yazılmış dokümanlardan oluşmaktadır. Bu durum sınırlı sayıda fiil kullanılmasından kaynaklanmaktadır.

Çizelge 8.32 Külliylatlara bağlı olarak elde edilen SEK ayrıştırma yöntemi sıkıştırma sonuçları

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
Külliyyat-I	32	35	36	38	39	39	39	39	40	39	37,6
Külliyyat-II	30	36	40	41	40	39	39	39	40	40	38,4
Külliyyat-III	32	35	36	38	39	39	39	39	40	40	37,7
Külliyyat-IV	30	35	36	37	38	39	41	47	50	52	40,5
Külliyyat-V	29	35	37	37	38	39	39	39	40	39	37,2
Külliyyat-VI	32	34	37	38	38	39	39	40	40	40	37,7
Külliyyat-VII	27	35	37	38	40	40	42	41	41	41	38,2
Külliyyat-VIII	33	39	41	42	43	43	39	44	45	45	41,4

Çizelge 8.33 Külliylatlara bağlı olarak elde edilen 2-GRAM ayrıştırma yöntemi sıkıştırma sonuçları

	Dosya Boyutu										Ort Başarı %
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb	
Külliyyat-I	0	0	16	30	37	40	43	44	45	45	30
Külliyyat-II	0	0	18	30	35	38	41	43	45	46	29,6
Külliyyat-III	0	0	16	30	37	40	43	44	45	46	30,1
Külliyyat-IV	0	4	19	31	37	41	44	52	56	56	34
Külliyyat-V	0	5	19	29	36	40	43	45	46	47	31
Külliyyat-VI	0	4	18	30	37	41	43	45	46	47	31,1
Külliyyat-VII	0	0	16	30	40	42	46	45	47	48	31,4
Külliyyat-VIII	0	10	29	37	44	44	42	45	44	43	33,8

Sonuç olarak, 13 ayrıştırma yöntemi 8 farklı külliyyata uygulandığında 11 ayrıştırma yönteminin Külliyyat-VIII’ de ortalama olarak daha başarılı olduğu görülmüştür. Bu 11 ayrıştırma yönteminde Külliyyat-IV sadece 1024Kb boyutundaki dosyada başarılı olmuştur.

9. TÜRKÇENİN BİÇİMBİLİMSEL YAPISINA DAYALI İSTATİSTİKSEL DEĞERLER

Bu bölümde sıkıştırma işleminde kullanılan sembol kullanım frekans değerlerine ilişkin istatistiksel değerler dosya boyutu tabanlı olarak grafiksel anlatımlarıyla birlikte yer almaktadır. Deneysel çalışmalarımızı Türkçe için 8 farklı külliyat ile gerçekleştirmiştik. 13 farklı ayrıştırma yönteminde ortalama olarak en başarılı külliyat olan 8 numaralı külliyat ile bundan sonraki çalışmalarımıza devam ediyoruz.

9.1 Külliyat-VIII için Ayrıştırma Şekline Dayalı Kullanım Frekans Değerleri

Çeşitli kişilere ait 6 farklı tez dokümanı alınarak oluşturulan Külliyat-VIII üzerinde ayrıştırma işlemi yapıldığında en yüksek sıklık değerlerine sahip 50 kök ve kullanım frekans değerleri Çizelge 9.1’ de gösterilmektedir.

Külliyat-VIII üzerinde ayrıştırma işlemi yapıldığında en yüksek kullanım frekans değerlerine sahip 50 hece ve kullanım frekans değerleri Çizelge 9.2’ de gösterilmektedir.

Külliyat-VIII üzerinde ayrıştırma işlemi yapıldığında en yüksek kullanım frekans değerlerine sahip 50 karakter ve kullanım frekans değerleri Çizelge 9.3’ de gösterilmektedir.

Külliyat-VIII üzerinde ayrıştırma işlemi yapıldığında en yüksek kullanım frekans değerlerine sahip 50 isim ve kullanım frekans değerleri Çizelge 9.4’ de gösterilmektedir.

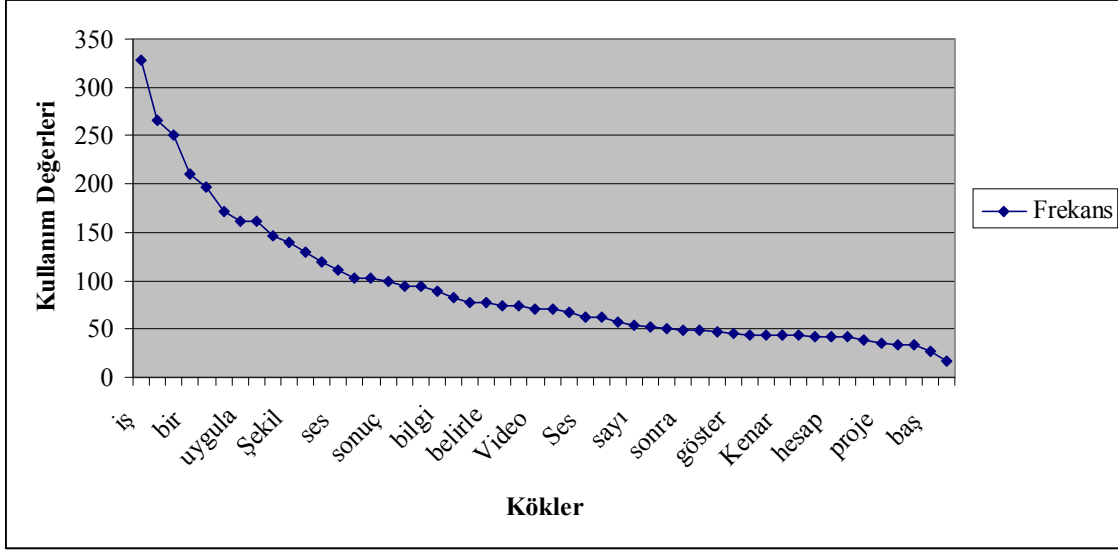
Külliyat-VIII üzerinde ayrıştırma işlemi yapıldığında en yüksek kullanım frekans değerlerine sahip 50 sıfat ve kullanım frekans değerleri Çizelge 9.5’ de gösterilmektedir.

Çizelge 9.1 Külliyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek

50 kök ve kullanım frekans değerleri

Sembol	Frekans	Olasılık	Ağaç İndeks	Sembol	Frekans	Olasılık	Ağaç İndeks
iş	328	0.00465	562	sistem	70	0.00099	2422
değer	266	0.00377	613	Gabor	68	0.00096	2445
video	250	0.00354	622	Ses	62	0.00088	2486
bir	211	0.00299	1097	işlem	62	0.00088	2527
resim	197	0.00279	1023	haber	57	0.00080	2850
gör	172	0.00244	1106	sayı	54	0.00076	2936
uygula	162	0.00229	1137	çerçeve	52	0.00073	2999
renk	161	0.00228	1138	Renk	51	0.00072	3004
Bu	146	0.00207	1185	sonra	49	0.00069	4099
Şekil	139	0.00197	1215	veri	48	0.00068	2076
et	129	0.00183	1233	bul	47	0.00066	2198
kenar	120	0.00170	1271	göster	46	0.00065	4354
ses	111	0.00157	1453	avi	44	0.00062	4388
çal	103	0.00146	1500	gerçek	43	0.00061	4465
su	102	0.00144	1503	Kenar	43	0.00061	4466
sonuç	100	0.00141	1506	eş	43	0.00061	4468
özet	95	0.00134	2089	art	42	0.00059	1514
özel	95	0.00134	2197	hesap	42	0.00059	4469
bilgi	90	0.00127	2178	sahne	42	0.00059	4499
fark	82	0.00116	2250	komut	39	0.00055	4654
hareket	78	0.00110	2335	proje	35	0.00049	4838
belirle	78	0.00110	2336	seviye	33	0.00046	4120
filtre	74	0.00105	2360	karşı	33	0.00046	4175
yöntem	74	0.00105	2367	baş	27	0.00038	4186
Video	71	0.00100	2419	üzeri	17	0.00024	4424

Çizelge 9.1’ te yer alan değerlerin grafiksel gösterimi Şekil 9.1’ de gösterilmiştir.



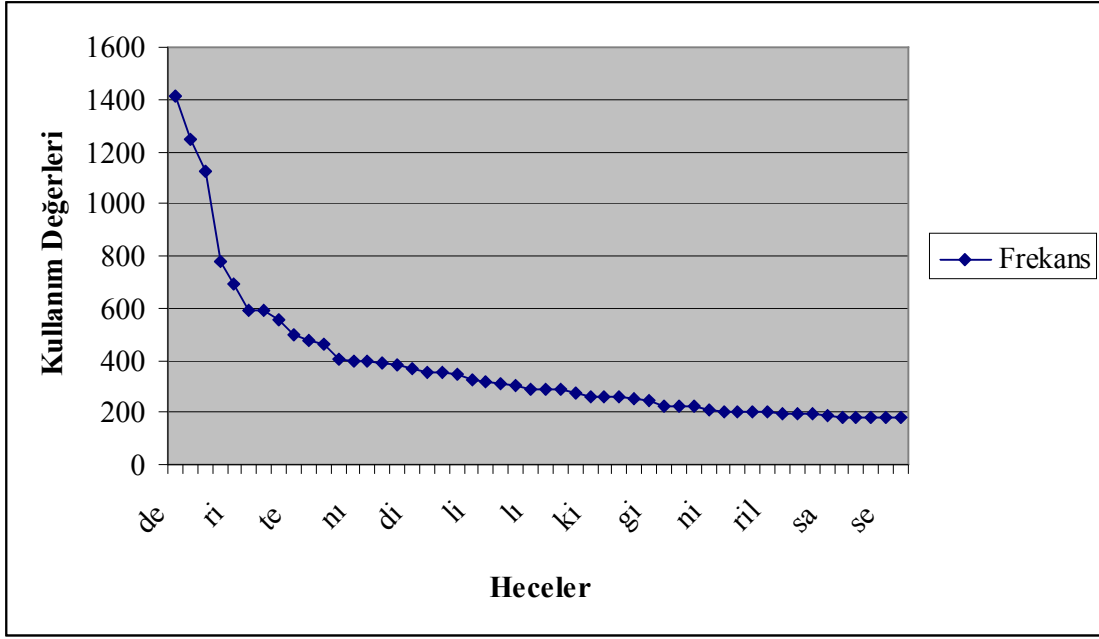
Şekil 9.1 Külliyyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek 50 kök ve kullanım frekans değerleri

Çizelge 9.2 Külliyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek

50 hece ve kullanım frekans değerleri

Sembol	Frekans	Olasılık	Ağaç İndeks	Sembol	Frekans	Olasılık	Ağaç İndeks
de	1413	0.01864	141	ler	288	0.00380	620
la	1245	0.01643	149	ne	285	0.00376	626
le	1122	0.01480	158	dir	272	0.00358	707
ma	775	0.01022	263	ki	262	0.00345	714
ri	692	0.00913	286	na	261	0.00344	715
ve	588	0.00776	305	rak	257	0.00339	720
re	588	0.00776	304	ğer	254	0.00335	725
da	558	0.00736	316	gi	242	0.00319	742
te	494	0.00651	368	mek	226	0.00298	756
me	473	0.00624	372	ge	225	0.00296	757
si	460	0.00607	375	ha	223	0.00294	762
sı	404	0.00533	519	ni	207	0.00273	1035
nı	397	0.00523	522	sim	204	0.00269	1037
len	394	0.00520	523	lir	202	0.00266	1041
ra	388	0.00512	529	bi	202	0.00266	1072
ta	385	0.00508	532	ril	201	0.00265	1042
di	370	0.00488	541	lar	198	0.00261	1049
rı	356	0.00469	562	ka	196	0.00258	1050
vi	355	0.00468	564	ke	193	0.00254	1062
ği	345	0.00455	575	sa	184	0.00242	1086
li	325	0.00428	585	rin	183	0.00241	1119
lan	314	0.00414	595	kul	182	0.00240	1071
ti	311	0.00410	597	mak	181	0.00238	1120
ya	301	0.00397	607	se	180	0.00237	1124
lı	291	0.00384	616	gö	177	0.00233	1128

Çizelge 9.2' te yer alan değerlerin grafiksel gösterimi Şekil 9.2' de gösterilmiştir.

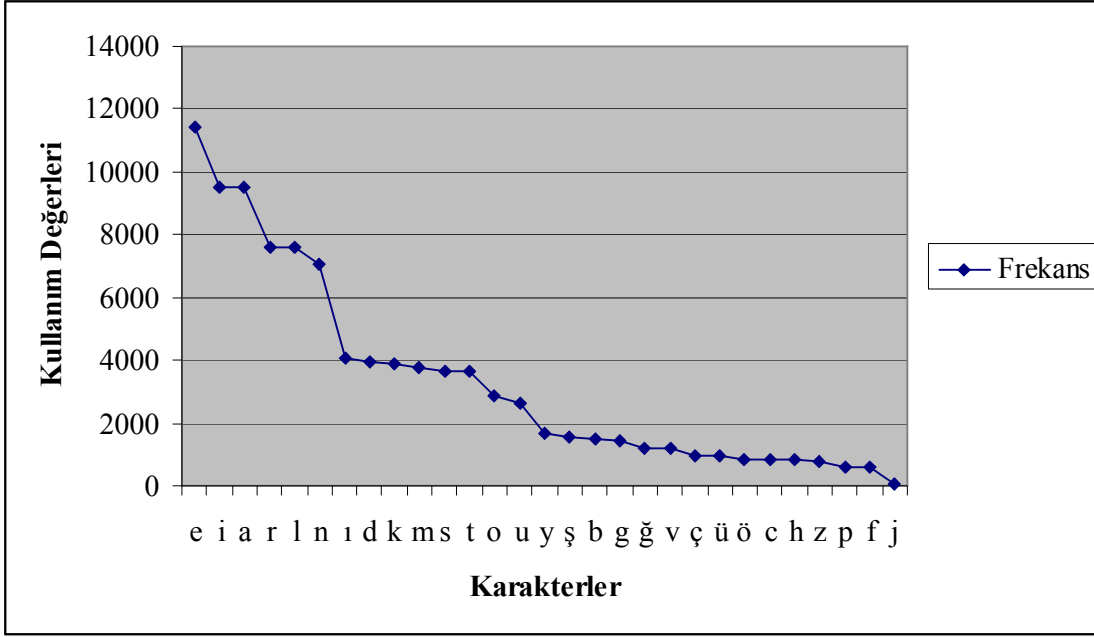


Şekil 9.2 Külliyyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek 50 hece ve kullanım frekans değerleri

Çizelge 9.3 Külliyyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek 50 karakter ve kullanım frekans değerleri

Sembol	Frekans	Olasılık	Ağaç İndeks	Sembol	Frekans	Olasılık	Ağaç İndeks
e	11446	0.08813	31	ş	1529	0.01177	263
i	9535	0.07342	33	b	1489	0.01146	265
a	9504	0.07318	34	g	1427	0.01098	184
r	7612	0.05861	40	ğ	1177	0.00906	266
l	7608	0.05858	41	v	1177	0.00906	314
n	7069	0.05443	43	ç	984	0.00757	360
ı	4069	0.03133	76	ü	963	0.00741	306
d	3922	0.03020	78	ö	856	0.00659	359
k	3903	0.03005	79	c	813	0.00626	365
m	3766	0.02899	80	h	809	0.00622	313
s	3635	0.02799	85	z	792	0.00609	368
t	3626	0.02792	86	p	619	0.00476	538
o	2861	0.02203	92	f	572	0.00440	1075
u	2652	0.02042	134	j	84	0.00064	1076
y	1679	0.01292	181				

Çizelge 9.3’ de yer alan değerlerin grafiksel gösterimi Şekil 9.3’ de gösterilmiştir.



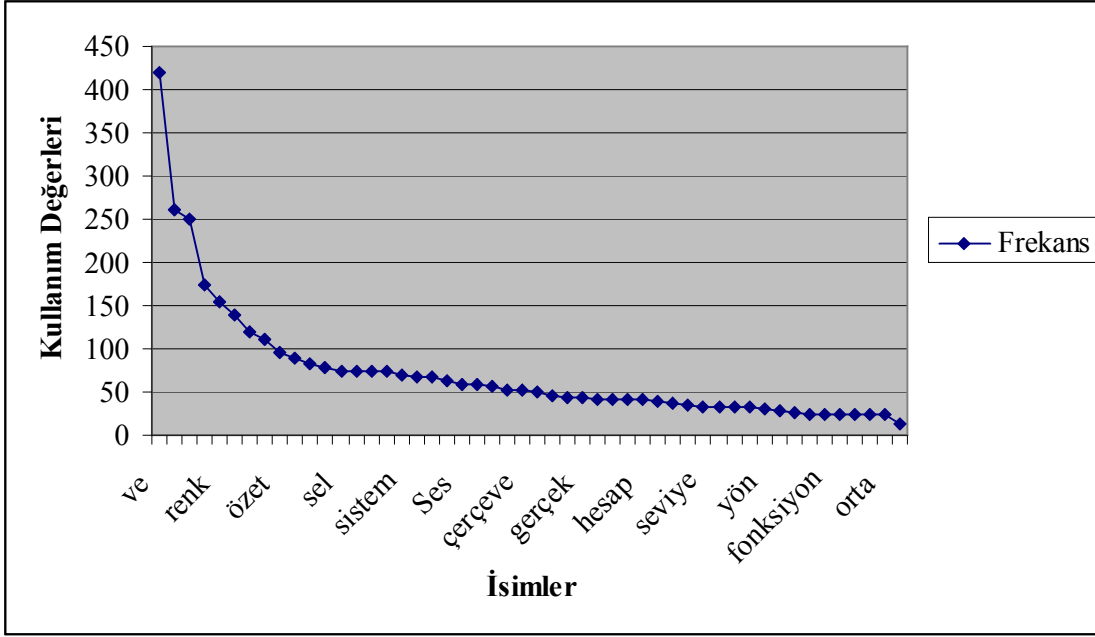
Şekil 9.3 Külliyyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek 50 karakter ve kullanım frekans değerleri

Çizelge 9.4 Külliyyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek

50 isim ve kullanım frekans değerleri

Sembol	Frekans	Olasılık	Ağaç İndeks	Sembol	Frekans	Olasılık	Ağaç İndeks
ve	420	0.00460	566	Renk	49	0.00053	4702
değer	261	0.00286	732	veri	46	0.00050	5024
video	250	0.00274	741	daha	44	0.00048	5049
resim	173	0.00189	1262	gerçek	43	0.00047	5062
renk	154	0.00168	1357	sonuc	42	0.00046	5074
Şekil	139	0.00152	1458	sahne	42	0.00046	5075
kenar	120	0.00131	1530	Kenar	41	0.00044	5077
ses	111	0.00121	2258	hesap	41	0.00044	5078
özet	95	0.00104	2507	komut	39	0.00042	5441
bilgi	90	0.00098	2514	yaka	37	0.00040	5468
fark	82	0.00089	2543	proje	35	0.00038	5832
hareket	78	0.00085	2558	seviye	33	0.00036	9401
sel	75	0.00082	2722	karşı	33	0.00036	5861
su	74	0.00081	2725	her	32	0.00035	5862
yöntem	74	0.00081	2728	temel	32	0.00035	5921
filtre	74	0.00081	2727	yön	30	0.00032	6120
sistem	70	0.00076	2916	algoritma	28	0.00030	9040
el	67	0.00073	2927	parça	27	0.00029	9069
Video	67	0.00073	2928	sağ	25	0.00027	9389
işlem	62	0.00067	2966	fonksiyon	25	0.00027	9382
Ses	59	0.00064	3065	resmi	25	0.00027	9390
sonuç	58	0.00063	4391	sınıf	24	0.00026	10009
haber	57	0.00062	4409	dalga	24	0.00026	9716
sayı	53	0.00053	4680	orta	24	0.00026	10010
çerçeve	52	0.00057	4687	kolay	12	0.00013	2505

Çizelge 9.4’ de yer alan değerlerin grafiksel gösterimi Şekil 9.4’ de gösterilmiştir.



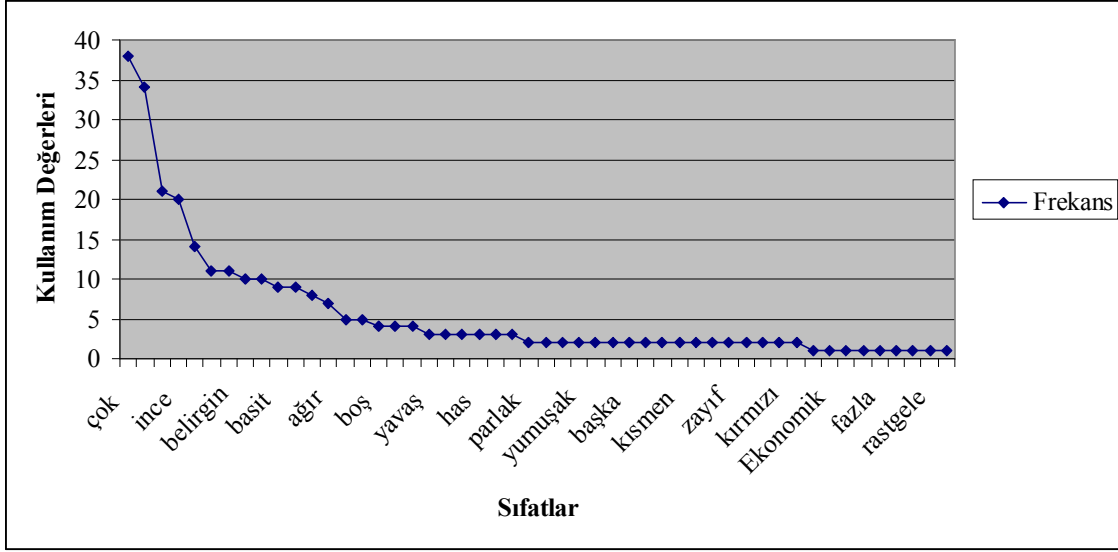
Şekil 9.4 Külliyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek 50 isim ve kullanım frekans değerleri

Çizelge 9.5 Külliyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek

50 sıfat ve kullanım frekans değerleri

Sembol	Frekans	Olasılık	Ağaç İndeks	Sembol	Frekans	Olasılık	Ağaç İndeks
çok	38	0.00029	626	Geçer	2	0.00002	157201
büyük	34	0.00029	9766	modern	2	0.00002	157207
DOĞRU	21	0.00016	16977	yumuşak	2	0.00002	157210
ince	20	0.00015	19479	yeşil	2	0.00002	271667
zor	14	0.00010	1483	Büyük	2	0.00001	44442
geniş	11	0.00009	33921	başka	2	0.00002	135834
belirgin	11	0.00009	33922	Zayıf	2	0.00002	157200
tüm	10	0.00008	38962	zayıf	2	0.00002	157205
mümkün	10	0.00008	39186	kısmen	2	0.00002	157208
basit	9	0.00007	39184	uzak	2	0.00002	157211
yeter	9	0.00007	22219	kırmızı	2	0.00002	271374
tamam	8	0.00006	39295	zayıf	2	0.00002	157205
ağır	7	0.00005	44522	kısmen	2	0.00002	157208
belirli	5	0.00004	19594	uzak	2	0.00002	157211
Çok	5	0.00004	88884	kırmızı	2	0.00002	271374
boş	4	0.00003	88257	nasıl	2	0.00002	313575
temiz	4	0.00003	88255	kötü	1	0.00008	9738
izole	4	0.00003	88267	Ekonomik	1	0.00001	313540
yavaş	3	0.00002	135679	ufak	1	0.00001	313548
mavi	3	0.00002	135682	uzun	1	0.00001	313551
yoğun	3	0.00002	176521	fazla	1	0.00001	313557
has	3	0.00002	39059	pembe	1	0.00001	313566
DOĞRU	3	0.00002	89053	nitelik	1	0.00001	313574
Atak	3	0.00002	135681	rastgele	1	0.00001	313585
parlak	2	0.00002	156785	Doygun	1	0.00001	313591

Çizelge 9.5’ de yer alan değerlerin grafiksel gösterimi Şekil 9.5’ de gösterilmiştir.



Şekil 9.5 Külliyyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek 50 hece ve kullanım frekans değerleri

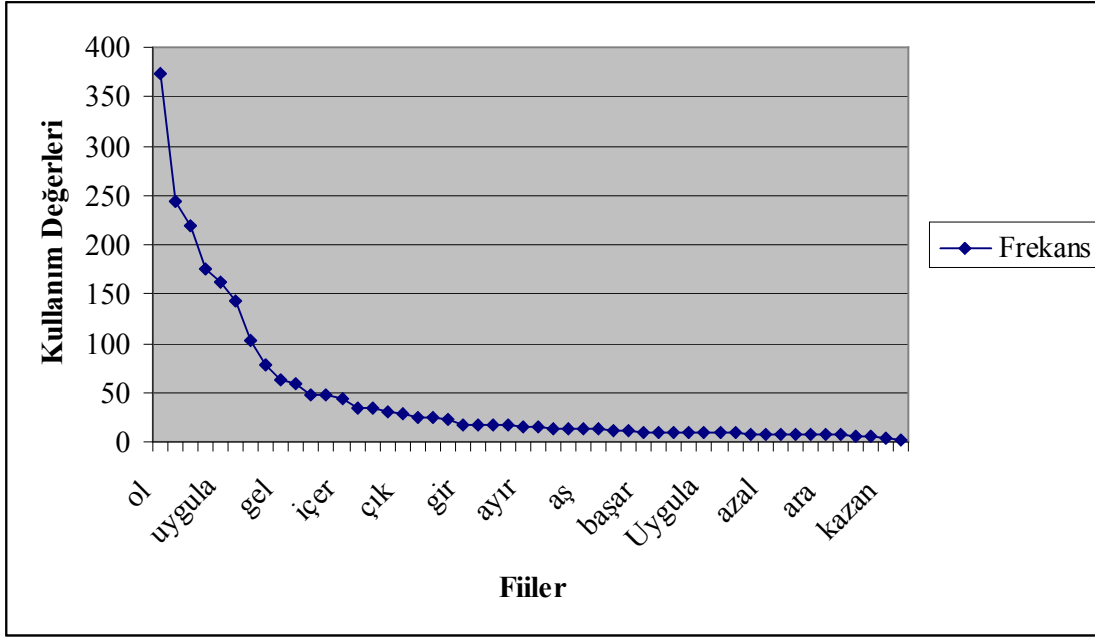
Külliyyat-VIII üzerinde ayrıştırma işlemi yapıldığında en yüksek kullanım değerlerine sahip 50 fiil ve kullanım frekans değerleri Çizelge 9.6' da gösterilmektedir.

Çizelge 9.6 Külliyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek

50 fiil ve kullanım frekans değerleri

Sembol	Frekans	Olasılık	Ağaç İndeks	Sembol	Frekans	Olasılık	Ağaç İndeks
ol	373	0.00335	727	bak	15	0.00013	19082
değ	243	0.00218	1181	dağıl	14	0.00012	19202
in	219	0.00197	1201	Çal	14	0.00012	19201
gör	175	0.00157	1463	aş	13	0.00011	20637
uygula	162	0.00145	1468	eğit	13	0.00011	20631
arak	143	0.00128	2321	Yap	12	0.00011	21126
çal	103	0.00092	2623	yaklaş	11	0.00001	23372
belirle	78	0.00070	4118	başar	10	0.00001	32946
gel	63	0.00056	4719	duy	10	0.00009	9370
yap	59	0.00053	4771	var	10	0.00009	32943
göster	48	0.00043	5281	düş	10	0.00009	32948
bul	47	0.00042	1283	Uygula	9	0.00008	33266
içer	43	0.00038	5861	kıs	9	0.00008	33268
art	35	0.00031	9367	benze	9	0.00008	33267
seç	35	0.00031	9344	artır	8	0.00007	38393
geç	31	0.00027	9442	azal	8	0.00007	37643
çık	28	0.00025	9588	Değ	8	0.00007	38429
böl	25	0.00023	10505	ölç	8	0.00007	42021
kal	25	0.00023	10507	çek	7	0.00006	581
Belirle	23	0.00020	10590	ara	7	0.00006	41225
gir	18	0.00016	18687	bit	7	0.00006	38383
çıkart	17	0.00015	18695	öğren	6	0.00005	42168
gerek	17	0.00015	18702	Yay	6	0.00005	42337
sür	17	0.00015	18698	kazan	3	0.00003	38169
ayır	15	0.00013	19085	taş	1	0.00001	2573

Çizelge 9.6' da yer alan değerlerin grafiksel gösterimi Şekil 9.6' da gösterilmiştir.



Şekil 9.6 Külliyyat- VIII dosyası için ayrıştırma işlemi uygulandığında elde edilen en yüksek 50 fiil ve kullanım frekans değerleri

Doğal diller, kullanım sıklıkları diğerlerinden daha fazla olan değişmez bazı kelimelere sahiptir. Bu varsayım tüm diller için geçerli olmaktadır. Bu düşünceyi ifade etmenin en genel yollarından biri Zipf' in kanunları olarak bilinmektedir. “The n^{th} most common word in a human language text occurs with a frequency inversely proportional to n .” (George, 1949). Bu kanun şu anlama gelmektedir, bir dilde bazı kelimeler diğerlerinden daha fazla kullanılır. Belirli bir konuya özel olan kelimeler için bu doğru bir yaklaşımdır. Bu yüzden önemli frekans değerlerine sahip sembolleri diğerlerinden ayırmak için bir eşik değeri (threshold) kullanılmaktadır. Bu bölümde en fazla sıklıkta kullanılan sembollerin dokümanların sıkıştırma performanslarında önemli etkisi olduğu düşünülerek, frekans değeri en yüksek olan 50 sembol alınarak istatistiksel değerler elde edilmiştir.

10. BİÇİMBİLİMSEL YAPISINA BAĞLI OLARAK TÜRKÇE VE İNGİLİZCENİN SIKIŞTIRMA PERFORMANSLARININ KARŞILAŞTIRILMASI

Bu bölümde Türkçe ve İngilizce için yapılan deneysel işlemlerin sonuçlarının karşılaştırılmasına dayalı analiz ve sonuçlar yer almaktadır.

Türkçe dili üzerinde yapılan çalışmalarda, 8 farklı külliyat üzerinde 13 farklı ayrıştırma yöntemine dayalı sıkıştırma işlemi uygulanmıştır. KET ayrıştırma yöntemi ile elde edilen sıkıştırma sonuçlarına bakıldığında en iyi sıkıştırma başarısının % 61 ile 1024Kb boyutlu Külliyat-IV ile sıkıştırılmasından elde edildiği görülmüştür. En kötü sıkıştırma başarısının ise 2Kb boyutlu bütün külliyatların sıkıştırılmasından elde edildiği görülmüştür. K ayrıştırma yöntemi ile elde edilen sıkıştırma sonuçlarına bakıldığında en iyi sıkıştırma başarısının % 51 ile 1024Kb boyutlu Külliyat-IV' ün sıkıştırılmasından elde edildiği görülmüştür. En kötü sıkıştırma başarısının ise % 34 ile 2Kb boyutlu Külliyat-VII' nin sıkıştırılmasından elde edildiği görülmüştür.

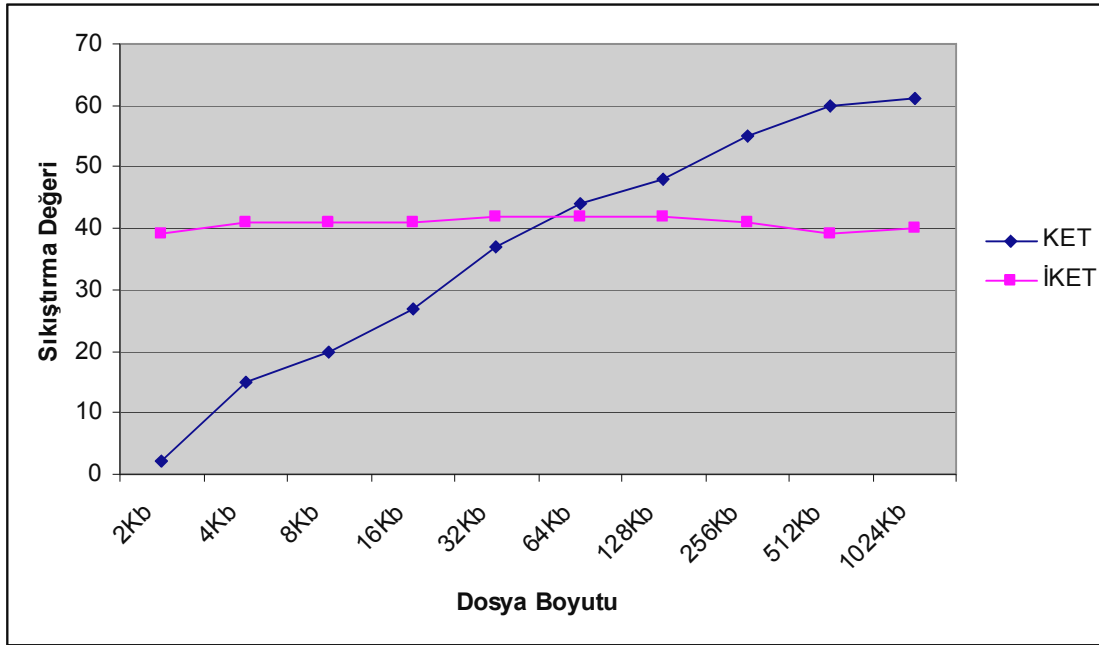
İngilizce dili üzerinde yapılan çalışmalarda, 1 külliyat üzerinde 4 farklı ayrıştırma yöntemine dayalı sıkıştırma işlemi uygulanmıştır. İKET ayrıştırma yöntemi ile elde edilen sıkıştırma sonuçlarına bakıldığında en iyi sıkıştırma başarısının % 42 ile 32Kb ve 64Kb boyutlu dosyanın sıkıştırılmasından elde edildiği görülmüştür. En kötü sıkıştırma başarısının ise % 39 ile 2Kb boyutlu dosyadan elde edildiği görülmüştür. K ayrıştırma yöntemi ile elde edilen sıkıştırma sonuçlarına bakıldığında en iyi sıkıştırma başarısının % 53 ile 64Kb boyutlu dosyanın sıkıştırılmasından elde edildiği, en kötü sıkıştırma başarısının ise % 6 ile 2Kb boyutlu dosyanın sıkıştırılmasından elde edilmiştir. 2-GRAM ayrıştırma yöntemi ile elde edilen sıkıştırma sonuçlarına bakıldığında en iyi sıkıştırma başarısının % 46 ile 256Kb, 512Kb ve 1024Kb boyutlu dosyalardan elde edildiği, en kötü sıkıştırma başarısının ise 2Kb boyutlu dosyanın sıkıştırılmasından elde edildiği görülmüştür.

KET ayrıştırma yöntemine göre Türkçe ve İngilizce sıkıştırma işlemleri karşılaştırıldığında en iyi sıkıştırma başarısı açısından yapılan değerlendirmede, 1024Kb boyutlu Türkçe dokümanın sıkıştırılmasından elde edilen % 61 sıkıştırma başarısının, 32Kb veya 64Kb boyutlu İngilizce dokümanın sıkıştırılmasından elde edilen % 42 sıkıştırma başarısından çok daha yüksek olduğu görülmektedir (Çizelge 10.1).

Çizelge 10.1 KET sıkıştırma yönteminin Türkçe ve İngilizce en iyi deneysel sonuçlarına Göre karşılaştırılması

	Dosya Boyutu									
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb
KET	9	15	20	27	37	44	48	55	60	61
İKET	39	41	41	41	42	42	42	41	39	40

Çizelge 10.1’de yer alan deneysel sonuçlar Şekil 10.1 üzerinde grafiksel olarak gösterilmiştir.



Şekil 10.1 KET sıkıştırma yönteminin Türkçe ve İngilizce en iyi deneysel sonuçlarına göre karşılaştırılması

Şekil 10.1’ den de görüldüğü gibi İngilizcede farklı dosya boyutlarında yaklaşık bir sıkıştırma oranı söz konusu iken Türkçede sıkıştırma başarısı dosyanın boyutu ile doğru orantılı olarak artmaktadır. Ayrıca 64Kb dosya kırılım noktasını oluştururken, üzeri dosya boyutlarında İngilizceden çok daha yüksek sıkıştırma başarısı alınmıştır.

K ayrıştırma yöntemine göre Türkçe ve İngilizce sıkıştırma işlemleri karşılaştırıldığında en iyi sıkıştırma başarısı açısından yapılan değerlendirmede, 64Kb boyutlu İngilizce dokümanın sıkıştırılmasından elde edilen % 53 sıkıştırma başarısının, 1024Kb boyutlu Türkçe dokümanın

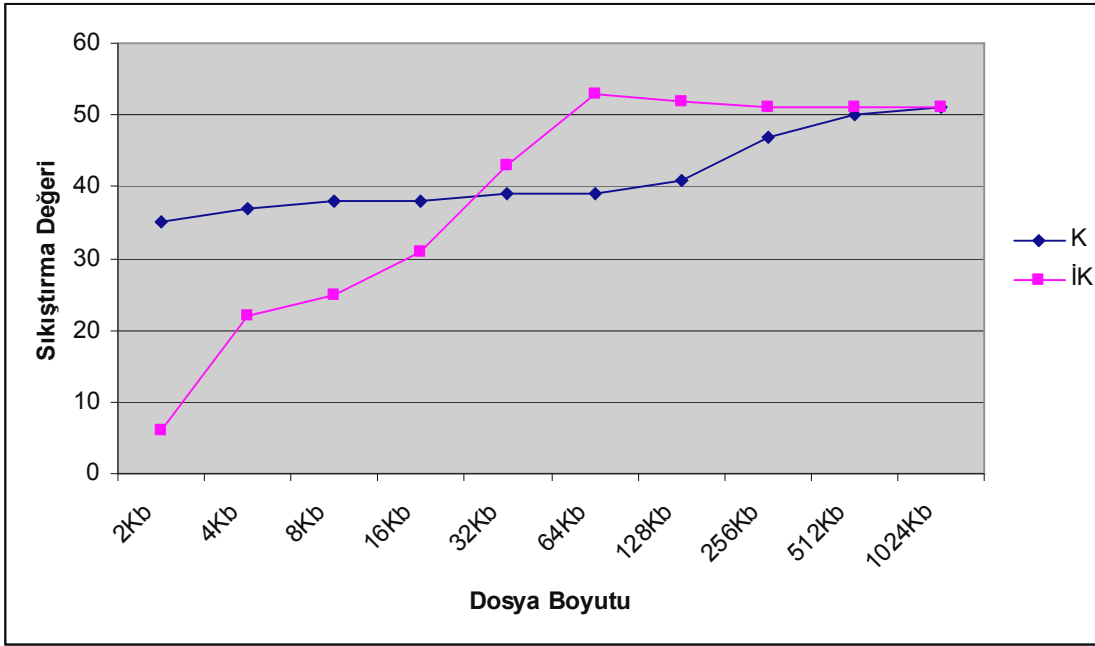
sıkıştırılmasından elde edilen % 51 sıkıştırma başarısından çok daha yüksek olduğu görülmektedir (Çizelge 10.2).

Çizelge 10.2 K sıkıştırma yönteminin Türkçe ve İngilizce en iyi deneysel sonuçlarına göre

karşılaştırılması

	Dosya Boyutu									
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb
K	35	37	38	38	39	39	41	47	50	51
İK	6	22	25	31	43	53	52	51	51	51

Çizelge 10.2’de yer alan deneysel sonuçlar Şekil 10.2 üzerinde grafiksel olarak gösterilmiştir.



Şekil 10.2 K sıkıştırma yönteminin Türkçe ve İngilizce en iyi deneysel sonuçlarına göre

karşılaştırılması

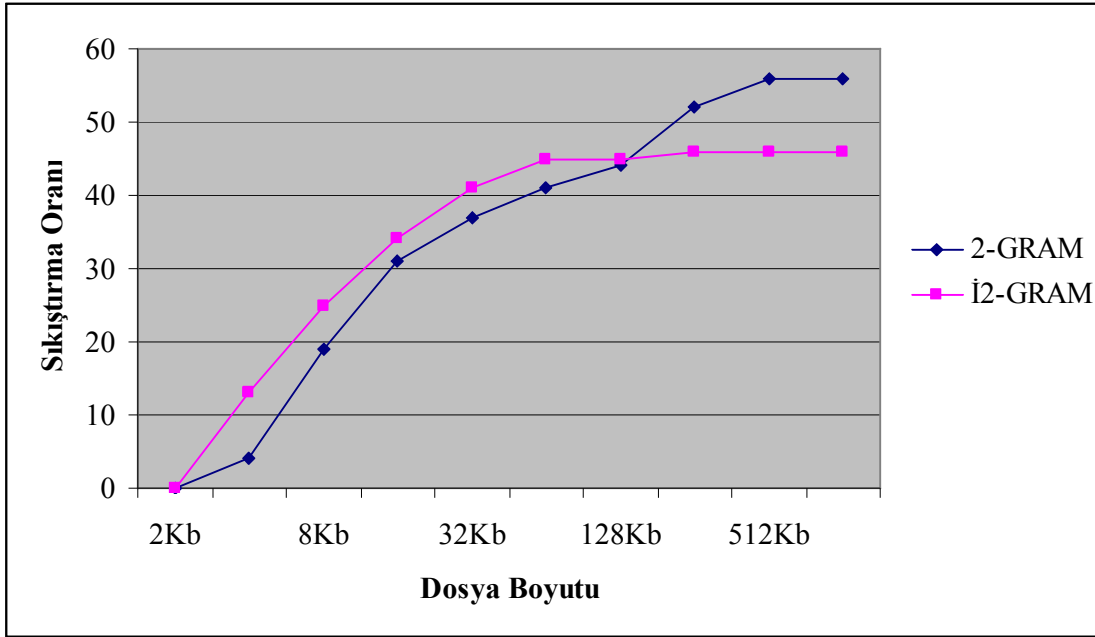
Şekil 10.2’ den de görüldüğü gibi dosya boyutu arttığında Türkçe ve İngilizce dilinde karakter tabanlı sıkıştırma kullanıldığında başarının arttığı ve birbirlerine yakın sonuçlar verdiği görülmektedir. İngilizcede 64Kb alınmış olan bu değer yanıltıcı olmamalıdır.

2-GRAM ayrıştırma yöntemine göre Türkçe ve İngilizce sıkıştırma işlemleri karşılaştırıldığında en iyi sıkıştırma başarısı açısından yapılan değerlendirmede, 512Kb ve 1024Kb boyutlu Türkçe dokümanın sıkıştırılmasından elde edilen % 56 sıkıştırma başarısının, 256Kb, 512Kb veya 1024Kb boyutlu İngilizce dokümanın sıkıştırılmasından elde edilen % 46 sıkıştırma başarısından çok daha yüksek olduğu görülmektedir (Çizelge 10.1).

Çizelge 10.3 2-GRAM sıkıştırma yönteminin Türkçe ve İngilizce en iyi deneysel sonuçlarına göre karşılaştırılması

	Dosya Boyutu									
	2Kb	4Kb	8Kb	16Kb	32Kb	64Kb	128Kb	256Kb	512Kb	1024Kb
2-GRAM	0	4	19	31	37	41	44	52	56	56
İ2-GRAM	0	13	25	34	41	45	45	46	46	46

Çizelge 10.3’ de yer alan deneysel sonuçlar Şekil 10.3 üzerinde grafiksel olarak gösterilmiştir.



Şekil 10.3 2-GRAM sıkıştırma yönteminin Türkçe ve İngilizce en iyi deneysel sonuçlarına göre karşılaştırılması

Şekil 10.3’ den de görüldüğü gibi dosya boyutu arttığında Türkçe ve İngilizce dilinde 2-GRAM ayrıştırma yöntemi ile sıkıştırma yapıldığında başarının arttığı ve birbirlerine yakın sonuçlar verdiği görülmektedir.

11. SONUÇ

Bu çalışmada, Türkçe ve İngilizce dokümanların sıkıştırılma verimlerinin belirlenmesinde farklı ayırıştırma yöntemleri ve bu yöntemlerin sıkıştırma oranına etkileri araştırılmıştır. Dokümanlar Türkçe ve İngilizcenin biçimbilimsel yapısı kullanılarak ayırıştırılmıştır. Sonraki aşamada ayırıştırılan dokümanlardaki yapılara sıkıştırma işlemi uygulanmıştır. Sonuçta, 17 farklı ayırıştırma tekniği oluşturulmuş ve bunlar ile farklı külliyatlar üzerinde denemeler yapılmıştır.

Kullanılan Türkçe külliyatlar, siyaset, spor, güncel gibi farklı konularda yazan 18 erkek, 4 kadın olmak üzere toplam 22 yazara ait, 900 farklı dokümandan oluşturulmuştur. İngilizce için oluşturulan külliyat, güncel konularda yazan farklı kişilere ait 290 farklı dokümandan elde edilmiştir. Farklı doküman gruplarının, farklı konularda yazılan metinlerin, yazar sayısının ve yazar cinsiyetinin başarıya olan etkisini gözlemek amacıyla 9 farklı külliyat grubu oluşturulmuştur.

Sıkıştırma oranını belirlemede, doküman boyutunun sıkıştırmaya olan etkisini tespit edebilmek amacıyla her külliyatın altında 10 farklı boyuttaki dokümana Huffman sıkıştırma metodu uygulanarak elde edilen performanslar karşılaştırılmıştır.

Yapılan denemelere göre, en başarılı sonuçlar, 1024Kb 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 için 8 adet külliyat ile yapılan çalışmalar şunu göstermiştir ki dosya boyutu ile alınan sıkıştırma başarısı doğru orantılıdır. Külliyatların çoğunda – Külliyat-III – hariç ortalama olarak en yüksek başarı K ile yapılan ayırıştırma şeklinde alınmıştır. Külliyat-III ve Külliyat-VII hariç yöntem bazında da KET ile daha başarılı sonuçlar elde edilmiştir. Külliyat-III’ de ise ortalama SEH daha başarılı olmuştur.

Ayırıştırma şekline göre deneyler yapıldığında ise 13 ayırıştırma yöntemi 8 farklı külliyata uygulanmış ve 11 ayırıştırma yönteminin Külliyat-VIII’ de ortalama olarak daha başarılı olduğu görülmüştür. Bu 11 ayırıştırma yönteminde Külliyat-IV sadece 1024Kb boyutundaki dosyada başarılı olmuştur.

Türkçe ve İngilizce dilleri için KET ile yapılan ayırıştırma şeklinde sıkıştırma işlemi yapıp, sonuçlar karşılaştırıldığında en düşük ve en yüksek başarının Türkçe dokümanlardan elde

edildiđi gözlemlenmiştir. K ile yapılan ayrıştırma şeklinde sıkıştırma işlemi yapıp, sonuçlar karşılaştırıldığında ise en düşük ve en yüksek başarı İngilizce dokümanlardan elde edilmiştir. 2-GRAM ile yapılan ayrıştırma şeklinde sıkıştırma işlemi yapıp, 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 bağılı olarak çok büyük artışlar görülmemekle birlikte, Türkçe dokümanların sıkıştırma verimlerinin dosya boyutuna bağılı olarak artış gösterdiği görülmüştür.

Bu çalışmada yer verilmeyen ayrıştırma şekilleri (zarf, zamir, 3-gram vs.) ileriki çalışmalarda ele alınıp, daha da kapsamlı ve genişletilmiş külliyatlar ile daha başarılı sıkıştırma sonuçları alınabilir.

Kısaca bu çalışma, bugüne kadar Türkçe dokümanların biçimbilimsel yapılarını kullanarak sıkıştırma işleminin yapılmasını sağlayan en kapsamlı çalışmadır.

KAYNAKLAR

- Çebi Y., and Dalkılıç, G., 2004, “Turkish Word N-gram Analyzing Algorithms for a Large Scale Turkish Corpus - TurCo”, ITCC 2004, IEEE International Conference on Information Technology, Vol:2, pp. 236-240.
- Çelikel, E., Dalkılıç, M.E. ve Dalkılıç, G., “Word-Based Fixed and Flexible List Compression”, Computer and Information Sciences-ISCIS, LNCS 3733, Springer Verlag, pp. 780-790, (2005)
- Dincer O., 2005, Hece Tabanlı İstatistiksel Yöntemler ile Yazım Hatası Bulma ve Düzeltme, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul, Türkiye.
- Diri B., 2002, Content Based Compression of Turkish Documents, Yıldız Technical University, Department of Computer Engineering, Yıldız 80750, İstanbul, Türkiye.
- Ediskun H., 2005, Türk Dilbilgisi Sesbilgisi-Biçimbilgisi-Cümlebilgisi, Remzi Kitabevi A.Ş., Selvili Mescit Sok. 3, Cağaloğlu 34440, İstanbul, Türkiye.
- Gibson, J. D, Berger, T.et al, 1998. Digital Compression for Multimedia, Morgan Kauffmann.
- Jurafsky D., Martin J.H., 2000, Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall, New Jersey.
- Manning C.D., Schütze H., 1999, Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge, Massachusetts, London, England.
- Nelson, M., The Data Compression Book, NewYork, USA, (1996)
- Rueda, L., Oommen, B.J., “Efficient Adaptive Data Compression Using Fano Binary Search Trees”, Computer and Information Sciences-ISCIS, Vol. 3733, p. 768-779, (2005)
- Salomon, D., 1998. Data Compression, Springer, NY, USA.
- Topaloğlu, U. ve Bayrak, C., “Polymorphic Compression”, Computer and Information Sciences-ISCIS, Vol. 3733, p. 759-767, (2005)
- Zipf, George K., “Human Behavior and the Principle of Least Effort”, an Introduction to Human Ecology, Addison-Wesley, Reading, Mass., 1949

İNTERNET KAYNAKLARI

- [1] http://en.wikipedia.org/wiki/Data_compression
- [2] http://en.wikipedia.org/wiki/Huffman_coding
- [3] <https://zemberek.dev.java.net/>
- [4] <http://cla.calpoly.edu/~jrubba/morph/morph.over.html>
- [5] [http://en.wikipedia.org/wiki/Morphology_\(linguistics\)](http://en.wikipedia.org/wiki/Morphology_(linguistics))
- [6] <http://tartarus.org/~martin/PorterStemmer/>
- [7] http://en.wikipedia.org/wiki/Shannon%E2%80%93Fano_coding
- [8] http://en.wikipedia.org/wiki/Arithmetic_coding
- [9] http://en.wikipedia.org/wiki/Run-length_encoding
- [10] http://en.wikipedia.org/wiki/LZ77_and_LZ78
- [11] <http://en.wikipedia.org/wiki/Lempel-Ziv-Welch>
- [12] <http://en.wikipedia.org/wiki/Jpeg>
- [13] <http://en.wikipedia.org/wiki/MPEG>
- [14] http://en.wikipedia.org/wiki/Entropy_coding
- [15] <http://en.wikipedia.org/wiki/BCPL>
- [16] <http://zemberek-web.appspot.com/>
- [17] www.hurriyet.com.tr
- [18] <http://www.milliyet.com.tr>
- [19] www.sabah.com.tr
- [20] <http://www.aksam.com.tr>

ÖZGEÇMİŞ

Doğum tarihi	05.05.1984	
Doğum yeri	Adıyaman	
İlköğretim	1995–1998	Yedikule İlköğretim Okulu
Lise	1998–2001	Yedikule Lisesi
Lisans	2001–2006	Yıldız Teknik Üniversitesi Elektrik-Elektronik Fak. Bilgisayar Mühendisliği Bölümü
Yüksek Lisans	2006–2009	Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği

Çalıştığı kurumlar

2006–2008	Garanti Teknoloji A.Ş., Yazılım Mühendisi
2008-Devam ediyor	Groupama A.Ş., Yazılım Mühendisi