TÜRKÇE İÇİN BİR SIKLIK ANALİZİ PROGRAMI

OKTAY, Melek[*]-KURT, Atakan[**]-KARA, Mehmet[***]
TÜRKİYE/ТУРЦИЯ

ÖZET

İki kelime ile *metin analizi* olarak nitelendirebileceğimiz uygulamalar, birçok bilim dalında değişik bağlamlarda ortaya çıkmaktadır. İşletmecilikte içerik ve doküman yönetimi uygulamalarını, doğal dil işlemede metin özetleme ve makine çevirisini, veri madenciliğinde doküman sınıflama ve gruplamayı, dilbilgisinde okunabilirlik analizini buna örnek olarak verebiliriz. Benzer yüzlerce uygulama ve teknoloji mevcuttur. Metin analizi, temelde *sıklık analizine* dayanmaktadır. Sıklık analizi; metin içerisindeki değişik ses, ek, kelime vb. dil öğelerini saydırarak göreceli ve mutlak istatistiklerin elde edilmesidir. Başta İngilizce olmak üzere önde gelen Batı dilleri için sıklık analizi uygulamaları önceden geliştirilmiştir. Bildiğimiz kadarıyla Türkçe için şu ana kadar sağlıklı bir sıklık analizi programı ortaya konulup yaygın olarak kullanılır hâle gelmemiştir. Türkçe yapı olarak eklemeli bir dil olduğundan İngilizce için hazırlanmış uygulamalar ile Türkçe metinlerin sıklık analizi yapılamamaktadır. Türkçe hem alfabesi ve fonetiği hem de morfolojisi ve cümle yapısı açısından farklı bir dil olduğu için sıklık analizi bakımından bu dilin ayrıca ele alınması gerekir. Sunacağımız bildiride bir Türkçe sıklık analizi uygulamasının geliştirilme süreci işlenecektir. Bu bağlamda Türkiye Türkçesi ve bazı Türk lehçelerini de destekleyecek olan bu uygulamanın geliştirilme süreci içerisinde yer alacak *gereksinim analizi* ve *arayüz tasarımı* konuları dikkate sunulacaktır.

Anahtar Kelimeler: Doğal dil işleme, bilgisayarlı dilbilim, Türkçe, Türk lehçeleri, metin analizi, sıklık analizi.

ABSTRACT

Text analysis is an important tool used in many applications in a diverse spectrum of fields such as document management applications in business administration, text summarization and machine translation in natural language processing, document classification and clustering in text data mining, readability anaysis in linquistics. Many other applications using text analysis can be found in the literature. Text analysis is based on the frequency and the important statistical characteristics of various textual elements such as phonems, affixes, words in

[*] Fatih Üniversitesi Mühendislik Fakültesi, **e-posta:** moktay@fatih.edu.tr

[**] Fatih Üniversitesi Mühendislik Fakültesi, **e-posta:** akurt@fatih.edu.tr

[***] Fatih Üniversitesi Fen-Edebiyat Fakültesi, **e-posta:** mkara@fatih.edu.tr

texts. Many frequency analysis studies for English and other Western languages have been done and applications based on these studies have been developed in the West. To the best of our knowledge, there is not a commonly-used well-established appplication for the frequency analysis of Turkish texts. Beacuse Turkish is an inflectional language, the frequency analysis applications developed for English is not appropriate for Turkish. Since Turkish has its own phonetics, morhology and syntax, her frequency analysis has to be studied on its own. We will put forward the development process of an frequency analysis application cuurenly being developed for Turkish texts in this paper. In this context we will emphasize the requirement analysis and graphical user interface stages of the applcaiton which will also support some of the dialects of Turkish language.

Key Words: Natural language processing, computational linquistics, Turkish, Turkic languages, text analysis, frequncy anlysis.

GİRİŞ (MAHİYET, FAYDA)

Zaman içerisinde değişik sebeplerden ses, yapı, anlam değişikliklerine uğrayan Türkçede meydana gelen değişiklikleri, kalabalık metin kümelerini (corpora) inceleyerek analiz edebiliriz. Bu analizin en önemli dayanak noktası; belirlenen metin kümelerindeki ses, hece, kelime vs. sıklıklarını ortaya koymak olacaktır. Bu sıklıkları belirlemek, bir araştırmacının kısa zamanda tek başına altından kalkabileceği bir iş değildir. Günümüzde Türkçenin temel metinlerinin birçoğu bilgisayar ortamına aktarılmıştır. Yeni üretilen metinler ise ya doğrudan bilgisayar ortamında veya internette oluşmakta ya da kısa zamanda sayısal ortama geçirilmektedir. Dolayısıyla bilgisayarlı bir Türkçe sıklık çalışması, metin analizlerinde hem süreyi çok azaltacak, hem de hataları en aza indirecektir. Ayrıca sonuçlar sayısal ortamda oluşturulacağı için elde edilen veriler başka bilgisayar uygulamaları ve kişiler tarafından daha ileri söz dizimi ve anlam analizleri için doğrudan kullanıma hazır olacaktır.

Benzer programlar İngilizce ve diğer diller için geliştirilmiş olsa da bu programların Türkçe için kullanılmasında bazı önemli engeller bulunmaktadır. Türkçenin alfabesi, sesleri, heceleme kuralları, kelime (kökler, ekler) ve cümle yapısı İngilizce ve diğer dillerden farklıdır. Bu sebeplerden dolayı yabancı diller için geliştirilmiş uygulamalar Türkçe metinler için kullanılamamakta, kullanılsa da tam ve güvenilir sonuç almak mümkün olamamaktadır.

Bu programın geliştirilmesinin ana amacı, Türkçe metin örgüsü içerisindeki sayısız özelliği, bilgisayar yardımı ile tespit etmek; yalnızca Türkçe öğretimi ve araştırmalarına değil, aynı zamanda iş yönetiminden psikolojiye kadar birçok alandaki değişik çalışmalara yardımcı olmaktır. Klâsik tarzda bunları yapmak, oldukça güç ve zaman alıcı bir iştir. Bundan dolayı geliştirilecek bilgisayar uygulamasıyla anadili Türkçe olanlar için ses, hece, kelime öğretimi

daha kolaylaşacak; yabancılara Türkçe öğretiminde kolaylıklar sağlanacak; diğer alanlarda ise içerik analizleri daha rahat yapılır hâle gelecektir.

Öte yandan karmaşık metin örgüsü içerisinden kısa sürede elde edilebilecek isabetli analizler sayesinde dil öğretimi konusunda yazılacak kitaplarda yeni verileri ortaya koymak mümkün olabilecek, bu yolla Türkçenin değişik açılardan araştırılmasına katkı sağlanabilecek, üniversitelerin yanı sıra lise ve dengi okullarda Türkçe dil bilgisi öğretiminin verimliliği artacaktır.

Bilgisayarın yaygınlaşması ile her alanda olduğu gibi dil alanında da bilgisayarlı uygulamaların sayısının zaman içerisinde çoğalacağı muhakkaktır. Bilgisayarın hız, saklama kapasitesi ve hata yapmaması gibi özelliklerinden dolayı Türkçe öğretiminde ve Türkçe araştırmalarında da kullanım alanları bulması, geliştirilen bu tür programlar sayesinde mümkün olacaktır.

Bir metin içerisindeki harf, hece, ek, kelime gibi birimlerin sıklık analizleri; metin ve yazarı hakkında daha detaylı yorumların yapılabilmesine imkan sağlar. Yani metin analizinin daha sağlıklı yapılmasına yardımcı olur. Aynı analiz; bir metin değil de bir metin kümesi (corpus) üzerinde yapıldığında ise, Türkçenin belirli bir alanına (edebiyat, siyaset vb.) veya belirli bir zaman dilimine ya da belirli bir yaş grubuna ait özelliklerin incelenmesinde de faydalı olacaktır.

Bu uygulama, önde gelen bazı Türk lehçelerinde de kullanılabilecek tarzda geliştirildiğinde, ileride Türk lehçeleriyle ilgili daha kapsamlı çalışmalara örnek ve taban teşkil edebilir. Örnek olarak; bir ileri aşamada tarihî ve çağdaş Türk lehçelerini bilgisayarla işleyebilen; ses, yapı ve cümle analizleri yapabilen bilgisayar destekli çalışmalar ortaya konulabilir.

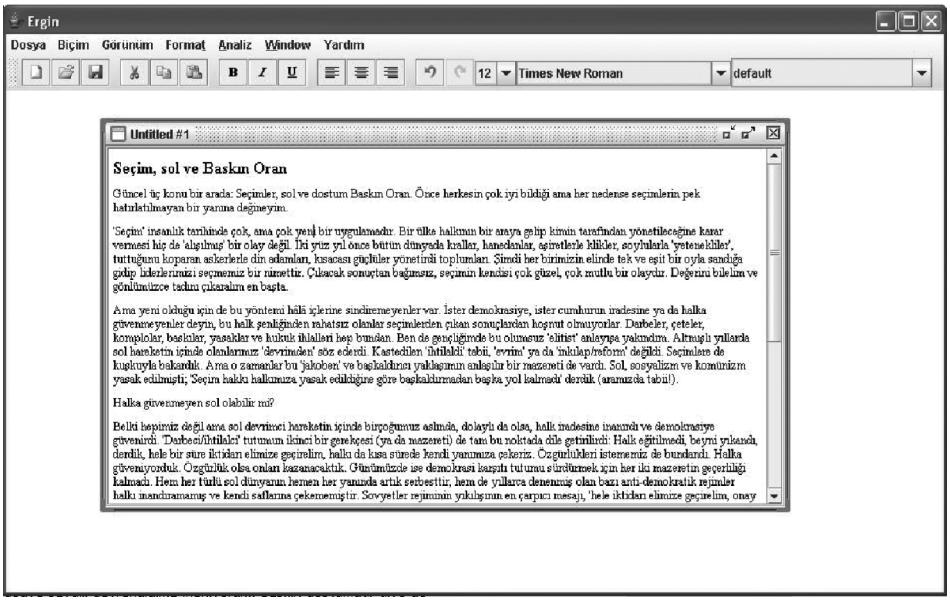
Geliştirdiğimiz bu program, Türk dil bilgisinden bahsedildiğinde adı ilk sıralarda anılan rahmetli Prof. Dr. Muharrem Ergin'in soyadıyla anılacaktır: **Ergin.** Aslında biz, söz konusu programın adını **Ercilasun** koymayı düşünmüştük. Yaşayan önemli Türk dil bilginlerinden biri olan Prof. Dr. Ahmet B. Ercilasun'a bunu açtığımızda kendisi büyük bir alçakgönüllülük göstererek programa **Ergin** adını vermemizi istemiştir.

1. Geliştirilen Program: Girdi, Arayüz ve Çıktı (Input, GUI, Output)

Öncelikle geliştirilen bu program temel metin özelliklerini (dosya açma, dosya kapama, dosya kaydetme vs.) ve editör özelliklerini (kes, kopyala, yapıştır) desteklemektedir. Geliştirilen bu metin editörünün temel fonksiyonları Mila projesinden alındı [MILA] ve üzerine yapılan eklentiler ve iç mimarisinin iyileştirilmesi ile daha kullanışlı ve düzgün bir hale getirildi. Bu program "txt" ve "rtf" uzantılı metin dosya tipleri desteklemektedir. Bunun yanında herhangi bir kaynaktan kopyalanan metinlerin editörün açılan penceresine yapıştırılması ile de analiz yaptırılabilir. Geliştirilen programın arayüzü **Tablo 1**'de verilmiştir.

Söz konusu program, aynı anda birden fazla metin üzerinde çalışılabilir; istenilirse metinler birleştirilerek de tek bir metin hâlinde analiz edilip sonuçlar bir pencerede görülebilir. Bunun yanında, metinler birleştirilmeden her biri farklı bir metin şeklinde analiz edilip, sonuçlar farklı pencerelerde kullanıcıya gösterilebilir. Bu gibi seçenekler tamamen kullanıcının isteği doğrultusunda belirlenir ve buna göre işleme tabi tutulur.

Programı üç ana başlık altında ele almak istiyoruz: **Karakter**, **Hece** ve **Kelime**.

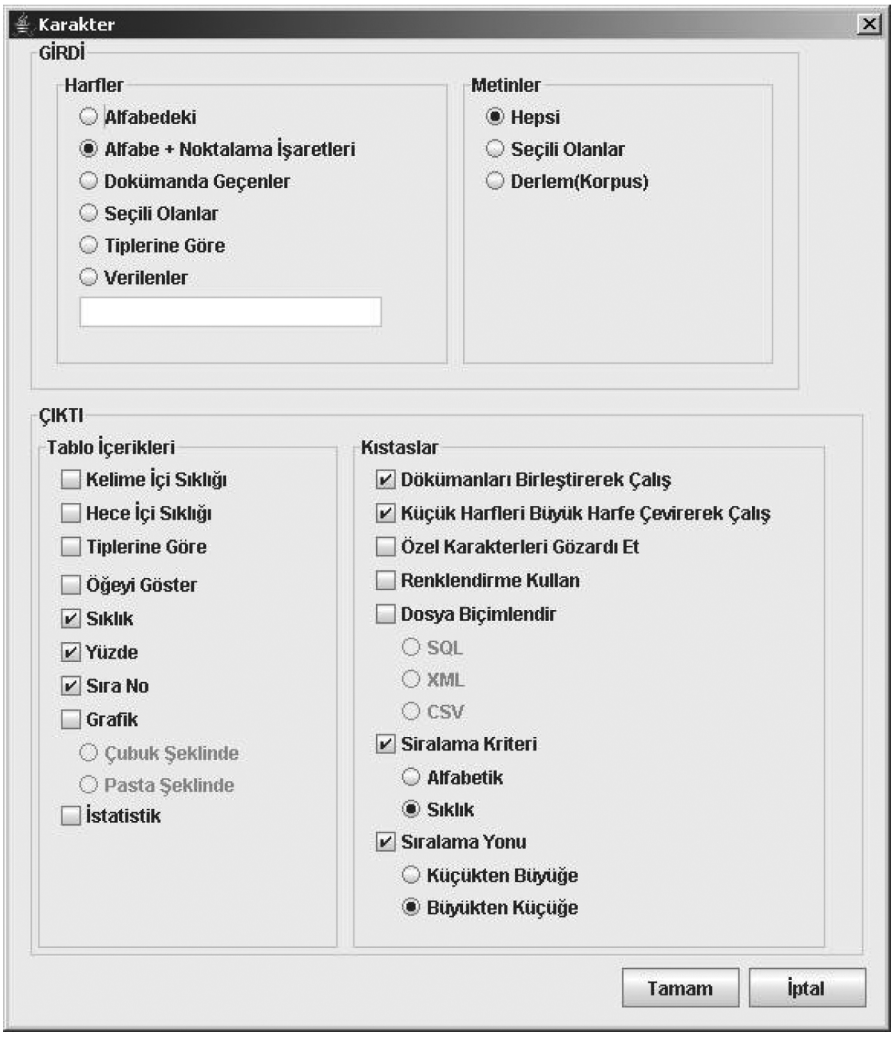


Tablo 1: Metin İşleme/Sıklık Analiz Programı Arayüzü ve Örnek Metin-I

1.1. Karakter

Karakter modülünde "Girdi" bölümü, işleme tabi tutulacak karakter ve metin tiplerini içermektedir. Kullanıcı, "Karakter Penceresi"nin "Girdi" bölümünde, Harfler kısmında işleme tabi tutacağı kesiti belirler. Bunlar, Türkiye Türkçesi alfabesinde bulunan karakterler veya alfabedeki karakterler ve noktalama işaretleri ya da sadece sayılması istenilen karakterler olabilir.

"Girdi" bölümünde, hangi metin üzerinde çalışma yapılacağının belirlenmesi gerekir. Bu, o esnada seçili olan metin ya da metinler olabilir. Karakter Penceresi, **Tablo 2**'de verilmiştir:



Tablo 2: Karakter Penceresi

Çıktı olarak hesaplanacak fonksiyonlar, "Karakter Penceresi"nin alt kısmında yer almaktadır. Bu fonksiyonlar, sırasıyla şöyledir:

• **Harf Sıklığı**

Bir karakterin verilen metinde ne kadar sıklıkla kullanıldığını tespit etmeye yarar. **Tablo 1**'deki örnek metnin harf sıklığını gösteren kesit, **Tablo 3**'te verilmiştir:

HARF SIKLIĞI

Sıra No	Karakter	Sıklık	Oran
1	E	117	11,48
2	İ	103	10,11
3	A	87	8,54
4	L	63	6,18
5	T	60	5,89
6	R	58	5,69
7	K	56	5,50
8	S	53	5,20
9	N	51	5,00
10	M	44	4,32
11	Ü	40	3,93
12	D	31	3,04
13	Y	29	2,85
14	U	26	2,55
15	I	25	2,45
16	B	19	1,86

Tablo 3: Harf Sıklığı

• **Kelime İçi Sıklığı**

Bir karakterin kelime içinde kaçıncı sırada/sıralarda yer aldığını tespit etmeye yarar. **Tablo 1**'deki örnek metinde bulunan karakterlerin "kelime içi sıklığı", **Tablo 4**'te gösterilmiştir:

KELİME İÇİ SIKLIK

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Toplam	Yüzde
E	9	23	5	14	14	4	19	2	8	9	5	3	1	0	1	0	0	0	0	0	117	%11,48
İ	8	9	8	16	17	5	9	6	9	2	3	5	2	3	1	0	0	0	0	0	103	%10,11
A	4	30	7	9	5	9	9	2	6	2	2	0	1	0	0	1	0	0	0	0	87	%8,54
L	1	2	11	10	9	9	3	8	4	5	0	1	0	0	0	0	0	0	0	0	63	%6,18
T	9	0	11	9	5	11	1	6	4	1	1	2	0	0	0	0	0	0	0	0	60	%5,89
R	0	2	15	4	5	8	3	7	0	6	3	0	1	2	1	1	0	0	0	0	58	%5,69
K	11	3	13	1	3	10	0	8	1	3	3	0	0	0	0	0	0	0	0	0	56	%5,50
S	14	2	10	5	5	5	3	7	0	0	2	0	0	0	0	0	0	0	0	0	53	%5,20
N	0	8	8	0	9	2	9	5	3	2	0	2	2	0	1	0	0	0	0	0	51	%5,00
M	7	0	7	0	9	9	2	4	3	0	1	0	0	2	0	0	0	0	0	0	44	%4,32
Ü	4	15	0	9	5	0	6	0	1	0	0	0	0	0	0	0	0	0	0	0	40	%3,93
D	3	4	3	3	0	3	2	4	1	3	1	1	1	1	1	0	0	0	0	0	31	%3,04
Y	10	1	4	4	1	6	0	2	1	0	0	0	0	0	0	0	0	0	0	0	29	%2,85
U	0	7	0	5	1	2	1	1	5	0	4	0	0	0	0	0	0	0	0	0	26	%2,55
I	0	4	1	4	4	6	0	1	1	0	1	1	1	0	1	0	0	0	0	0	25	%2,45
B	13	0	1	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	19	%1,86
F	13	0	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	%1,86

Tablo 4: Kelime İçi Sıklığı

• **Hece İçi Sıklığı**

Bir karakterin kelime içinde hecelerde kaçıncı karakter olarak yer aldığını tespit etmeye yarar. Tablo-1'deki örnek metinde bulunan karakterlerin "hece içi sıklığı", Tablo-5'te gösterilmiştir:

HECE İÇİ SIKLIĞI

	1	2	3	4	5	6	Toplam
E	9	108	0	0	0	0	117
İ	8	95	0	0	0	0	103
A	7	80	0	0	0	0	87
L	48	0	15	0	0	0	63
T	54	0	5	1	0	0	60
R	29	1	28	0	0	0	58
K	30	2	24	0	0	0	56
S	39	2	12	0	0	0	53
N	16	5	29	1	0	0	51
M	27	0	17	0	0	0	44
Ü	4	36	0	0	0	0	40
D	31	0	0	0	0	0	31
Y	27	0	2	0	0	0	29
U	0	26	0	0	0	0	26
I	0	25	0	0	0	0	25
B	19	0	0	0	0	0	19
F	19	0	0	0	0	0	19

Tablo 5: Hece İçi Sıklığı

• **Tiplerine Göre**

Ünlü ve ünsüzlerin türlerine göre sıklığını tespit etmeye yarar. **Tablo 1**'deki örnek metinde bulunan karakterlerin türlerine göre sıklığı, **Tablo 6**'da gösterilmiştir:

TİPLERİNE GÖRE HARF SIKLIĞI		
Sesli Karakter Sayısı	:	419
Sessiz Karakter Sayısı	:	548
Sesli --> Kalın	:	152
Sesli --> İnce	:	267
Sesli --> Düz	:	332
Sesli --> Yuvarlak	:	87
Sesli --> Genis	:	225
Sesli --> Dar	:	194
Sessiz --> Sedalı	:	333
Sessiz --> Sedasız	:	215
Sessiz --> Sürekli	:	367
Sessiz --> Süreksiz	:	181
Sessiz --> Akıcı	:	245
Sessiz --> Sızıcı	:	122
Sessiz --> Patlayıcı	:	181
Sessiz --> Dudak	:	69
Sessiz --> Diş Dudak	:	35
Sessiz --> Diş	:	199
Sessiz --> Diş Dudak	:	15
Sessiz --> Ön Damak	:	206
Sessiz --> Art Damak	:	13
Sessiz --> Gırtlak	:	11

Tablo 6: Tiplerine Göre Ünlü-Ünsüz Sıklığı

Bunların yanında karakterlerin sıra numarası, sıklığı ve oranı da kullanıcının isteği doğrultusunda hesaplanabilir. İstatistik seçeneğinde ise metin içerisinde toplamda kaç karakter olduğu, bunlardan kaç tanesinin işleme tabi tutulduğu, toplamda kaç ünlü ve ünsüz bulunduğu hesaplanabilir.

"Kıstaslar" bölümünde, sonuç olarak ekranda gösterilecek öğelerin hangi kıstasa göre sıralanacağı belirlenir. Kullanıcı isterse sıklığa göre, isterse alfabetik sıraya göre bunları sıralayabilir. Sonuçlar ayrıca artan ya da azalan sırada listelenebilir. Kullanıcı isterse dokümanları birleştirerek ve küçük harfleri büyük harflere çevirerek de çalışabilir. Bunlara ek olarak sonuçların daha belirgin olarak görülebilmesi için "Renklendirme Kullan" seçeneği mevcuttur. Sonuçlar, satır satır farklı renkte gösterilerek daha anlaşılır bir hale getirilebilir. "Özel Karakterleri Göz Ardı Et" seçeneği de alfabe dışı bazı özel karakterlerin, sıklık analizi yapılmadan metin içerisinden ayıklanmasına yarar.

2.2. Hece

Geliştirilen bu programın içerisindeki modül seçenekleri birbirlerine benzemektedir. Karakter modülünde olduğu gibi Hece modülü de "Girdi" ve "Çıktı" şeklinde iki bölümden oluşmaktadır (bkz.: **Tablo 7**). "Girdi" bölümünde bulunan Heceler kısmı içerisinden kullanıcı işlem yapacağı hece veya heceleri belirler. "Dokümandakiler" seçeneğini işaretlerse, metin içerisindeki bütün heceler üzerinde işlem yapmak istiyor demektir. "Verilenler" seçeneği işaretlenirse, kullanıcının belirlemiş olduğu hece veya heceler üzerinde işlem yapılır. Girdi bölümünde Metinler seçeneği, üzerinde çalışılacak olan metinlerin belirlenmesi için kullanılmaktadır. Kullanıcı seçeneğine bağlı olarak, editörde o esnada açık olan bütün metinler üzerinde veya o esnada seçili olan metin üzerinde işlem yapılabilir.

Tablo 7: Hece Penceresi

“Türkiye Sağlık ve Tedavi Vakfı tarafından kurulan Fatih Üniversitesi, 18.11.1996 tarihinde Dokuzuncu Cumhurbaşkanımız Sayın Süleyman Demirel tarafından eğitim - öğretime açılmıştır. On yedi üyesi bulunan Mütevelli Heyeti ile yönetilmektedir. Üniversitemiz, Büyükçekmece Kampüsü’nde Fen - Edebiyat, İktisadi ve İdari Bilimler, Mühendislik Fakülteleri, Fen ve Sosyal Bilimler Enstitüleri ve İstanbul Meslek Yüksekokulu; Ostim Kampüsü’nde Tıp Fakültesi, Sağlık Bilimleri Enstitüsü, Hemşirelik Yüksekokulu, Sağlık Bilimleri Meslek Yüksekokulu ve Ankara Meslek Yüksekokulu ile eğitim – öğretim faaliyetlerini sürdürmektedir.

1997-1998 akademik yılında Büyükçekmece Kampüsü’nde eğitim-öğretime başlayan Fatih Üniversitesi; Fen - Edebiyat Fakültesi, İktisadi ve İdari Bilimler Fakültesi, Mühendislik Fakültesi sosyal tesisleri ve öğrenci yurtlarıyla modern bir eğitim ortamına sahiptir. Sosyal tesis binasında kütüphane, sinema salonu, kafeterya, yemekhane, kitabevi, kırtasiye, terzi, kuaför ve internet kafe bulunmaktadır. Fakültelerin bünyesinde kurulan laboratuvarlarda eğitim öğretim faaliyetlerinin yanı sıra araştırma çalışmaları da sürdürülmektedir.”

Tablo 8: Örnek Metin II

Hece için “Çıktı” bölümünde, Tablo-8’deki örnek metin kullanılmıştır. Heceler için hesaplanabilecek tablolar ve bunların örnek çıktıları ise aşağıda verilmiştir.

• Hece Sıklığı: Bu kısımda, analiz edilen metin öğeleri hecelerine ayrılıp kullanıcının isteğine bağlı olarak sıklığına göre veya alfabetik olarak sıralanabilir. Biz, sıklığına göre yapılmış bir sıralamayı **Tablo 9**’da gösterdik:

HECE SIKLIĞI

Sıra No	Hece	Sıklık	Oran
1	te	16	3,82
2	si	12	2,86
3	fa	10	2,39
4	ri	10	2,39
5	le	10	2,39
6	de	9	2,15
7	ku	8	1,91
8	ve	8	1,91
9	ti	8	1,91
10	bi	8	1,91
11	tim	7	1,67
12	ta	7	1,67
13	da	7	1,67
14	e	7	1,67
15	ye	7	1,67
16	yük	6	1,43

Tablo 9: Hece Sıklığı

• **Kelime İçi Sıklığı**

Bu seçenekte isminden de anlaşılacağı üzere, hecelerin kelime içindeki sırasına göre sıklığı hesaplanır. Örnek olarak "tedavi" kelimesindeki "te", kelimenin ilk hecesi; "da" ikinci hecesi ve "vi" de üçüncü hecesi olarak belirlenir. Hesaplanan bu bilgilerin tablo şeklinde sıralanmış biçimi Tablo-10'da verilmiştir. Tablonun en son sütunu, hecenin toplam sıklığını göstermektedir:

KELİME İÇİ SIKLIK

Hece	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Toplam
te	3	1	6	1	5	0	0	0	0	0	0	0	0	0	0	16
si	1	0	2	7	0	2	0	0	0	0	0	0	0	0	0	12
fa	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10
ri	0	1	2	3	2	2	0	0	0	0	0	0	0	0	0	10
le	0	2	3	3	2	0	0	0	0	0	0	0	0	0	0	10
de	1	2	1	5	0	0	0	0	0	0	0	0	0	0	0	9
ku	3	1	0	4	0	0	0	0	0	0	0	0	0	0	0	8
ve	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
ti	0	4	3	0	0	1	0	0	0	0	0	0	0	0	0	8
bi	6	0	2	0	0	0	0	0	0	0	0	0	0	0	0	8
tim	0	1	6	0	0	0	0	0	0	0	0	0	0	0	0	7

Tablo 10: Kelime İçi Hece Sıklığı

• **Hece Tipleri Kelime İçi Sıklığı**

Türkçede altı çeşit hece türü bulunmaktadır. Bu hece türleri ve örnekleri **Tablo 11**'de verilmiştir. Bu tablodaki **V** – Ünlüyü (Vowel) , **C** – Ünsüzü (Consonant) temsil etmektedir. Türkiye Türkçesi'ndeki alıntı kelimelerin hece tipleri, çok azı hariç (tren vb.), Türkçe hece tiplerine benzemektedir:

Hece Tipleri	Örnek
V	a, e, ı, i, o, ö, u, ü
VC	at, aç,iş…
CV	ba, be, bı…
CVC	bel, gel, köy, tır…
VCC	alt, üst, ırk…
CVCC	kurt, yurt, renk, Türk…

Tablo 11: Türkçe Hece Tipleri

Örnek olarak verilen metnin "hece tipleri kelime içi sıklığı" **Tablo 12**'de gösterilmiştir.

HECE TİPLERİ KELİME İÇİ SIKLIĞI

Hece Türü	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Toplam
V	18	3	0	0	0	0	0	0	0	0	0	0	0	0	0	21
VC	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12
CV	62	63	43	44	18	5	3	0	0	0	0	0	0	0	0	238
CVC	31	44	45	14	4	5	1	0	0	0	0	0	0	0	0	144
VCC	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
CVCC	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2

Tablo 12: Hece Tipleri Kelime İçi Sıklığı

• Hece Uzunluğu Kelime İçi Sıklığı: Hece uzunluğu, hecenin sahip olduğu karakter sayısını göstermektedir. Örnek metnin "hece uzunluğu kelime içi sıklığı", **Tablo 13**'te verilmiştir:

HECE UZUNLUĞU KELİME İÇİ SIKLIĞI

Harf Uzunluğu	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Toplam
1	18	3	0	0	0	0	0	0	0	0	0	0	0	0	0	21
2	74	63	43	44	18	5	3	0	0	0	0	0	0	0	0	250
3	33	44	45	14	4	5	1	0	0	0	0	0	0	0	0	146
4	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tablo 13: Hece Uzunluğu Kelime İçi Sıklığı

• **Hece Uzunluğu Sıklığı:** Hece uzunlukları (harf sayısına göre) ve bu hece uzunluklarına ait sıklık, **Tablo 14**'te gösterilmiştir:

HECE UZUNLUĞU SIKLIĞI

Hece Uzunluğu	Sıklık
4	2
3	146
2	250
1	21

Tablo 14: Hece Uzunluğu Sıklığı

• Hece Tipi Sıklığı:

Tablo 11'de belirtilen hece tiplerinin örnek metin için hesaplanan sıklıkları **Tablo 15**'te gösterilmiştir.

HECE TİPİ SIKLIĞI

Hece Tipi	Sıklık
V	21
CVC	144
CV	238
CVCC	2
VC	12
VCC	2

Tablo 15: Hece Tipleri Sıklığı

3.1. Kelime

Bu uygulamadaki en kapsamlı kesit, kelimeler bölümüdür. Kelimelerle ilgili pencere, **Tablo 16**'da gösterilmiştir. Bu pencere, daha önceki hece ve harf penceresiyle bazı ortak özellikler içermektedir.

Tablo 16: Kelime Penceresi

Bu kısımda, öncelikle hangi metinde kelimelerin hangi özelliklerine göre bir çalışma yapılacağına karar verilmesi gerekir. Bu aşamadaki önemli fonksiyonlar ve bunların **Tablo 8**'deki örnek metne göre çıktılarının bir bölümü aşağıda verilmiştir.

• **Kelime Sıklığı**

Metin içerisindeki kelimelerin sıklık ve oranlarını tespit etmeye yarar. Örnek metne ait kelime sıklığı, **Tablo 17**'de gösterilmiştir:

KELİME SIKLIĞI

Sıra no	Kelime	Sıklık	Oran
1	VE	8	5,97
2	YÜKSEKOKULU	4	2,99
3	EĞİTİM	4	2,99
4	FAKÜLTESİ	4	2,99
5	BİLİMLER	3	2,24
6	MESLEK	3	2,24
7	FEN	3	2,24
8	SOSYAL	3	2,24

Tablo 17: Kelime Sıklığı

• **Harf Sayısı Sıklığı**

Kelimelerin içerdikleri harf sayısına göre sıklıklarının belirlenmesini sağlar. **Tablo 8**'de verilen metindeki kelimelerin harf sayısına göre sıklığı, **Tablo 18**'de verilmiştir. Örnek olarak; 16 ve 15 harfli 3'er adet kelime bulunmaktadır:

HARF SAYISI

Harf Sayısı	Sıklık
16	3
15	3
14	2
13	2
12	5
11	13
10	5
9	17
8	18
7	7
6	20
5	11
4	6
3	7
2	12
1	3

Tablo 18: Kelimelerin Harf Sayısına Göre Sıklığı

• Hece Sayısı Sıklığı: Kelimelerin içerdikleri hece sayılarına göre sıklıklarını tespit etmeye yarar. Bununla ilgili çıktı, **Tablo 19**'da verilmiştir:

HECE SAYISI SIKLIĞI

Hece Sayısı	Sıklık
7	4
6	6
5	12
4	36
3	30
2	23
1	15

Tablo 19: Kelimelerin Hece Sayılarına Göre Sıklığı

• Kelime Kökü Sıklığı

Yapım ve çekim eklerini ayıklayarak kelime kökü sıklığının bulunmasını sağlar. Bu çıktı, stilistik çalışmalarında son derece önemlidir. Örnek metnin kelime kökü sıklığı, **Tablo 20**'de gösterilmiştir:

KELİME KÖKÜ SIKLIĞI

Kelime Kökü	Sıklık
hemşire	1
bünye	1
fen	3
baş	1
faaliyet	2
idari	2
edebiyat	2
heyet	1
bina	1
il	2
bir	1

Tablo 20: Kelimelerin Köklerine Göre Sıklıkları

• **Ekler Sıklığı**

Kelimelerin almış olduğu eklerin sıklığını belirler. **Tablo 21**'de örnek metinde geçen eklerin sıklığı verilmiştir. Bu tabloda yer alan eklerdeki büyük harfler, bir ekin farklı ünlü veya ünsüz (kalın/ince vb.) almış biçimlerini tek simgeyle göstermede kullanılır. Bu özel karakterlerin neyi ifade ettiği, dilci ve dilbilimciler tarafından bilinmektedir. Örnek vermek gerekirse, **Tablo**'daki "lAr", metin içerisindeki "-lar" veya "-ler" eki yerine geçmektedir:

EKLER SIKLIĞI

Ekler	Sıklık
yA	2
nHn	1
Hn	2
ndA	2
lA	1
yH	1
lHk	1
sH	3
yAn	1
lAr	2

Tablo 21: Ekler Sıklığı

• **Kelime Gövdesi Sıklığı**

Kelime gövdesi, bir kelime kökünün yapım eki almış biçimidir. Örnek metne göre kelimelerin gövde sıklıkları, **Tablo 22**'de verilmiştir:

KELİME GÖVDESİ SIKLIĞI

Gövde	Sıklık
başlayan	1
bünye	1
fen	3
faaliyet	2
hemşirelik	1
idari	2
edebiyat	2
heyet	1
bina	1
il	2
bir	1

Tablo 22: Kelime Gövdesi Sıklığı

SONUÇ VE GELECEKTE YAPILACAKLAR

Türkiye Türkçesi'ne ait metinlerdeki ses, hece, ek, kelime sıklıklarını analiz edecek bir uygulamanın geliştirilmesini amaç edinen bu çalışma, birçok dil örgüsünün incelenebilmesine imkan sağlayacak biçimde tasarlanmıştır.

Burada öncelikle Türkiye Türkçesi'nin sıklık analizi hedeflenmiştir. Diğer Türk lehçelerine ait sıklık analiz uygulamalarında ise, birikimlerimizin Türkiye Türkçesine göre daha kısıtlı olmasından dolayı bazı problemlerle karşılaşabileceğimizi; bunların bir kısmını başlangıçta, diğerlerini ise ilerleyen zamanlarda çözebileceğimizi ümit ediyoruz.

Sözü edilen çalışmanın başarısı, geliştirilen uygulamanın ne kadar iyi ortaya konduğu ve ne kadar iyi test edildiği kadar, kullanılan kaynakların (kök-ekler, morfolojik çözümleyici, sözlük vb.) ne kadar doğru bilgi içerdiğine de bağlı olacaktır.

Türkçe metinleri analiz eden bir uygulamanın Türkçenin değişik lehçelerinde yazılmış metinleri -uygulama bu işleme uygun hale getirildiğinde- işleyebilmesi mümkün olabilir. Çünkü Azerice, Türkmence gibi bazı lehçeler Türkiye Türkçesi ile önemli oranlarda benzeşmektedirler. Bu çalışmanın ana amaçlarından biri de Türkçe metin işleme uygulamasına hiç olmazsa Türkiye Türkçesi dışında bir başka lehçede daha metin işleme özelliğini kazandırmaktır. Bu iş için Türkmence düşünülmektedir. Çünkü Türkmence, hem Türkiye Türkçesine yakın bir lehçedir hem de bu lehçe üzerine yaptığımız/yaptırdığımız tezler ve bilimsel çalışmalarla gerekli altyapı bir dereceye kadar hazırlanmıştır. Lehçeler üzerine yapılmış çalışmaların sınırlı olmasından dolayı, bunlarla ilgili sıklık analizlerinin Türkiye Türkçesi için yapılan sıkılık analizlerine göre bazı yönlerden eksikliklerinin bulunması doğaldır. Fakat uygulama genişletilebilir olarak geliştirileceği için diğer lehçelerin zaman içerisinde programa eklenmesi gerekli bilgi birikimi ortaya çıktığında kısa zaman içerisinde yapılabilecektir.

Şu ana kadar, sözü edilen uygulamanın harf, hece ve kelime kısmı gerçekleştirildi. Bu uygulamaya ileride cümle ve paragraf kısmı da eklenecektir. Program yeni geliştirildiği için bazı eksiklikler bulunabilir. Bu eksiklikler, denemelerden sonra düzeltilecektir.

KAYNAKÇA

Adalı, O., (2004), **Türkiye Türkçesinde Biçimbirimler**, Papatya Yayıncılık, Ankara.

Banguoglu, T., (2000), **Türkçenin Grameri,** Türk Dil Kurumu Yayınları, Ankara.

Ergin, M., (1998), **Türk Dil Bilgisi**, Boğaziçi Yayınları, İstanbul.

Eryiğit, G.-Oflazer, K., (2006), "Statistical Dependency Parsing of Turkish", **Proceedings of EACL 2006 11th Conference of the European Chapter of the Association for Computational Linguistics**, Trento, Italy, April.

Göz, İ., (2003), **Yazılı Türkçenin Kelime Sıklığı Sözlüğü**, Türk Dil Kurumu Yayınları, Ankara.

Jukka, K. K., (2006), **Unicode Explained**, O'Reilly, New York.

Karaman, L., (1997), **Türkçede Söz Dizimi**, Akçağ Yayınları, Ankara.

Oflazer, K., (1994), "Two-level Description of Turkish Morphology", **Literary Linguistic Computing**, 9, 137-148.

Tantuğ A. C.-Adalı, E., Oflazer, K., (2006) "A Prototype Machine Translation System Between Turkmen and Turkish", **Proceedings of the Turkish Artificial Intelligence and Neural Networks**, TAINN 2006, Muğla,Turkey.

Tekcan, A.-Göz, İ., (2005), **Türkçe Kelime Normları**, Boğaziçi Üniversitesi Yayınevi, İstanbul.

The Official Unicode Web Site: **http://unicode.org.**

The Resource Bundle Class: http://java.sun.com/j2se/1.4.2/docs/api/java /util/ ResourceBundle.html