

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

**BİLGİ ERİŞİMİNDE
TAMLAMA TEMELLİ DİZİNLEME**

**YÜKSEK LİSANS TEZİ
Müh. Mehmet Kıvanç TÜRKEEŞ
(504041521)**

**Tezin Enstitüye Verildiği Tarih : 7 Mayıs 2007
Tezin Savunulduğu Tarih : 14 Haziran 2007**

**Tez Danışmanı : Prof.Dr. Eşref ADALI
Diğer Jüri Üyeleri Prof.Dr. A.Coşkun SÖNMEZ (Yıldız Teknik Ü.)
Yrd.Doç.Dr. Zehre ÇATALTEPE (İ.T.Ü.)**

HAZİRAN 2007

ÖNSÖZ

Tez çalışmalarım sırasında desteğini esirgemeyen değerli hocam ve tez danışmanım Prof. Dr. Eşref ADALI'ya teşekkürü bir borç bilirim.

Ayrıca, tüm eğitim hayatım boyunca desteklerini bir an olsun bile esirgemeyen, her türlü katkıyı sağlayan değerli aileme ve tez çalışmalarım sırasında destek ve yardımlarından dolayı değerli dostlarım Mesut TÜLEK ve Fatih KESGİN'e teşekkürlerimi sunarım.

Haziran, 2007

M. Kıvanç TÜRKEEŞ

İÇİNDEKİLER

KISALTMALAR	v
TABLO LİSTESİ	vi
ŞEKİL LİSTESİ	vii
SEMBOL LİSTESİ	viii
ÖZET	ix
SUMMARY	xi
1. GİRİŞ	1
2. TÜRKÇE	6
2.1 Türkçenin genel özellikleri	6
2.1.1 Kökeni bakımından Türkçe	6
2.1.2 Yapı bakımından Türkçe	7
2.1.2.1 Yalınlayan Diller	7
2.1.2.2 Bükümlü Diller	7
2.1.2.3 Eklemeli Diller	8
2.2 Biçim Bilgisi	8
2.2.1 Çekim Ekleri	9
2.2.2 Yapım Ekleri	9
2.3 Sözcük Bilgisi	9
2.3.1 İsim	10
2.3.2 Sıfat	10
2.3.3 Zamir	10
2.3.4 Eylem	11
2.3.5 Zarf (Belirteç)	11
2.3.6 Edat (İlgeç)	11
2.3.7 Bağlaç	11
2.3.8 Ünlem	11
2.4 Sözcük Grupları	11
2.4.1 İsim Tamlamaları	12
2.4.1.1 Belirtili İsim Tamlaması	13
2.4.1.2 Belirtisiz İsim Tamlaması	14
2.4.1.3 Takısız İsim Tamlaması	16
2.4.1.4 Zincirleme İsim Tamlaması	17
2.4.2 Sıfat Tamlaması	17
3. BİLGİ ERİŞİMİ VE DOĞAL DİL İŞLEME	19
3.1 Bilgi Erişimi	19
3.1.1 Bilgi Erişiminin Kısa Tarihçesi	19
3.1.2 Temel Bilgi Erişim Sistemi	20
3.1.3 Bilgi Erişiminde Temel Kavramlar	21
3.2 Doğal Dil İşleme	24
3.2.1 Sesbilim	25

3.2.2 Biçimbilim	26
3.2.3 Sözdizim Bilimi	28
3.2.4 Sözcüksel bilim	29
3.2.5 Anlambilim	30
3.3 Önceki Çalışmalar	30
3.3.1 Türkçe Dışındaki Dillerde Yapılmış Çalışmalar	30
3.3.2 Türkçe İçin Yapılmış Çalışmalar	32
3.3.3 Zemberek Doğal Dil İşleme Kütüphanesi	35
4. TAMLAMA ANALİZİ	43
4.1 İstatistiksel Tamlama Analizi	45
4.2 Dilbilimsel Tamlama Analizi	47
4.2.1 Belirtisiz İsim Tamlaması	49
4.2.2 Belirtili İsim Tamlaması	51
4.2.3 Zincirleme İsim Tamlaması	52
4.2.4 Sıfat Tamlaması	55
5. SİSTEM ANALİZİ	57
5.1 Sistem Tasarımı	57
5.1.1 Ön İşleme Birimi	60
5.1.2 Biçimbirimsel Analiz Birimi	60
5.1.3 Sözdizimsel Analiz Birimi	65
5.1.4 Dizinleme Birimi	65
5.1.5 Temsil Veritabanı	67
5.1.6 Bilgi Erişimi Çekirdeği	67
5.2 Yazılımın Kullanımı	68
5.2.1 Derlem Ekranı	68
5.2.2 Arama Ekranı	68
5.2.3 Ayarlar Ekranı	69
5.2.4 Sınama Ekranı	69
6. SONUÇLAR VE TARTIŞMA	70
KAYNAKLAR	76
EKLER	79
ÖZGEÇMİŞ	80

KISALTMALAR

BE	: Bilgi Eriřimi
DDİ	: Doęal Dil İřleme
NIST	: National Institute of Standards and Technology
TREC	: Text Retrieval Conference
BÇ	: Biçimbirimsel Çözümleme
SDM	: Sonlu Durum Makinesi
DBD	: Dilden Bağımsız Dizin
İS	: İsim
İS_İ	: İsim (-i tamlama eki almış)
İS_İN	: İsim (-in tamlama eki almış)
İS_İ_İN	: İsim (hem –i tamlama ekini hem de -in tamlama eki almış)
SI	: Sıfat
BA	: Bağlaç
Dİ	: Düzenli İfade

TABLO LİSTESİ

	<u>Sayfa No</u>
Tablo 3.1: Farklı diller için gövdeleme sonuçları	27
Tablo 3.2: Porter gövdeleme algoritması örnek kurallar ve örnekler	31
Tablo 3.3: Porter gövdeleme algoritmasının bulduğu yanlış sonuçlar.....	31
Tablo 3.4: Porter gövdeleyicisi ve Oflazer'in biçimbirimsel çözümleyicisinin sonuçlarının karşılaştırılması	33
Tablo 3.5: Zemberek kütüphanesi üretim kuralları örnekleri	37
Tablo 4.1: Bir istatistiksel tamlama analizi için örnek parametre değerleri.....	46
Tablo 4.2: Kullanılan düzenli ifadeler ve anlamları.....	48
Tablo 4.3: Tamlamaların sembolik ifadeleri	49
Tablo 5.1: Köke kadar gövdeleme yaklaşımına örnek	61
Tablo 5.2: Biçimbirimsel analiz birimi için örnek girdi ve çıktı.....	62
Tablo 6.1: Örnek belgeler için hesaplanan keskinlik ve anımsama değerleri.....	71
Tablo A.1: Durma sözcük listesi	79

ŞEKİL LİSTESİ

Sayfa No

Şekil 1.1: Doğal Dil İşleme destekli Bilgi Erişimi sisteminin genel tasarımı	5
Şekil 2.1: Zincirleme isim tamlaması örnekleri.....	17
Şekil 3.1: Temel bilgi sistemi tasarımı	21
Şekil 3.2: 3 Boyutlu örnek yöney uzayı	22
Şekil 3.3: Zemberek kütüphanesi dil yapısı elemanları.....	36
Şekil 3.4: Zemberek ek sözlüğünden bir örnek	37
Şekil 3.5: Zemberek kütüphanesi kök sözlüğü örneği.....	38
Şekil 3.6: Zemberek kütüphanesi kök ağacı örneği.....	39
Şekil 3.7: Zemberek kütüphanesi biçimbirimsel çözümleyici blok şeması.....	41
Şekil 3.8: Biçimbirimsel çözümleyici örnek çıktısı.....	42
Şekil 4.1: Belirtisiz isim tamlaması Adım 1 için örnek, düzenli ifade ve SDM.....	50
Şekil 4.2: Belirtisiz isim tamlaması Adım 2 için örnek, düzenli ifade ve SDM.....	50
Şekil 4.3: Belirtisiz isim tamlaması için düzenli ifade ve SDM.....	51
Şekil 4.4: Belirtili isim tamlaması için düzenli ifade ve SDM.....	52
Şekil 4.5: Zincirleme isim tamlaması Öbek 1, 2, 3 ve 4 için düzenli ifade ve SDM	54
Şekil 4.6: Zincirleme isim tamlaması Öbek 5 için düzenli ifade ve SDM	55
Şekil 4.7: Sıfat tamlaması Adım 1 için örnek, düzenli ifade ve SDM	55
Şekil 4.8: Sıfat tamlaması Adım 2 için örnek, düzenli ifade ve SDM	56
Şekil 4.9: Sıfat tamlaması için düzenli ifade ve SDM.....	56
Şekil 5.1: Doğal Dil İşleme destekli Bilgi Erişim sistemin yapısı ve birimleri.....	59
Şekil 5.2: Biçimbirimsel analiz biriminin akış şeması	64
Şekil 5.3: Veritabanı varlık ilişki çizelgesi.....	67
Şekil 6.1: Örnek belgeler için keskinlik ve anımsama grafiği (tam karşılaştırma) ...	72
Şekil 6.2: Örnek belgeler için keskinlik ve anımsama grafiği (hata paylı karşılaştırma).....	72

SEMBOL LİSTESİ

$\vec{b}_i$	: Derlemdeki i. belgeyi temsil eden yöney
$\vec{s}_k$	: k. sorguyu temsil eden yöney
$w_{i,k}$	: i. belge veya sorgudaki k. terimin ağırlığı
ts_t	: t teriminin terim sıklığı
tbs_i	: i belgesinin ters belge sıklığı
ϵ	: Sonlu durum makinelerinde boş geçiş

BİLGİ ERİŞİMİNDE TAMLAMA TEMELLİ DİZİNLEME

ÖZET

Bilgi, belki de günümüzde en hızlı artan, en değerli şeydir. Bilginin tarihinde yazının icadı, Gutenberg'in matbaası ve internet önemli yer tutmaktadır. Yazının icadı ile gündelik bilgi sonraki kuşaklara kısıtlı da olsa aktarılabilmiş, Gutenberg'in matbaası ile bilgi daha hızlı çoğaltılıp daha geniş kitlelere ulaşabilmiştir. 20. yüzyılın son yarısında internetin ortaya çıkması ile bilginin üretilmesi ve paylaşılması çok kolay hale gelmiş ve insanlık tarihi boyunca ulaştığı en büyük bilgi seviyesine erişmiştir. Bu bilgi bolluğunda, önemli boyutlarda bilgi bilinçli veya bilinçsiz olarak unutulmakta veya göz ardı edilmektedir. Çoğu durumda insanlık bilgi bolluğunda ilgi yoksulluğundan dolayı istenen bilgiye ulaşamamaktadır. Bu durumda bilgiye sahip olmak kadar bilgiye erişmek de önem kazanmaktadır.

Bilgi Erişimi, bilginin temsil edilmesi, saklanması, düzenlenmesi ve gerektiği zamanda erişilebilmesini mümkün hale getirmek için yöntemlerin geliştirildiği araştırma konusudur. Bilgi Erişimi, ses, görüntü gibi çoklu ortam bilgi yığınları üzerinde çalışsa da en çok yazılı belgeler üzerinde çalışmaktadır. Bilgi erişimi, belgeler üzerinde çalışırken belgeleri, sözcük torbası olarak değerlendirir ve genellikle sözcük sayısı veya sözcük sıklığı gibi genel yöntemler kullanır. Fakat sayısal ortamda saklanan yazılı belgelerin birçoğu doğal dille yazılmış olduğundan, doğal dillerin barındırdığı dilbilimsel çeşitliliği ve anlamsal karmaşıklığı da barındırırlar. Sadece sözcük sıklığına dayanan genel bilgi erişim yöntemleri bu farklılıklar karşısında yetersiz kalmaktadır. Bu durum metinlerin temsil edilmesi ve işlenmesinde yeni yöntemlerin geliştirilmesine yani Doğal Dil İşleme konusu üzerine çalışmalara yol açar.

Doğal Dil İşleme (DDİ), insanlar tarafından kullanılan sözlü ve yazılı doğal dilleri inceleyen, anlaşılması ve çözümlenmesi üzerine çalışan bir bilim dalıdır. Bilgi Erişimi alanında ele alınan metinlerdeki dilbilimsel farklılıkları gidermek ve Bilgi Erişimi sistemlerinin başarımını artırmak için Doğal Dil İşleme yöntemlerinden yararlanılması gerekmektedir. Türkçe, sondan eklemeli yapısı ile Doğal Dil İşleme çalışmaları için karmaşık ve zor, fakat kurallı yapısı ile de bir o kadar elverişlidir.

Bu tez kapsamında, Türkçe için doğal dil işleme destekli bir bilgi erişim sistemi tasarımı yapılmıştır. Bu bilgi erişim sistemi, Türkçenin özelliklerini ve yapısını göz önüne alarak, bilgi erişimindeki başarımı arttırmayı amaçlamaktadır. Tasarlanmış olan bu sistemde Türkçenin biçimsel ve sözdizimsel farklılıkları ele alınmış ve bu farklılıkların giderilmesinin bilgi erişimi üzerindeki etkileri irdelenmiştir. Birimsel farklılıklar biçimbirimsel çözümlemeye dayanan gövdeleme, sözdizimsel farklılıklar ise tamlama analizi ile giderilmeye çalışılmıştır. Tamlama analizinde hem

istatistiksel hem de dilbilimsel yaklaşımlar incelenmiş ve dilbilimsel tamlama analizi gerçekleştirilmiştir. İstatistiksel yaklaşımda, birbirini belirli sıklıklarla izleyen ve belge genelinde sıkça geçen sözcük öbekleri tamlama olarak kabul edilmiş, dilbilimsel yaklaşımda ise isim ve sıfat tamlamaları bulunmaya çalışılmıştır. İstatistiksel yaklaşımda sözcükler arasındaki yakınlık ve sıklık ilişkileri üzerinde durulurken dilbilimsel yaklaşımda biçimbirimsel çözümleme de kullanılarak sözcük türleri ve sözcüklere gelen tamlama ekleri incelenip Türkçedeki isim ve sıfat tamlamaları sonlu durum makineleri ve düzenli ifadeler olarak temsil edilmiştir. Tamlama analizi ile bulunan tamlamalar tek sözcüklerle beraber belgeleri temsil etmek için kullanılarak tamlama temelli dizinleme gerçekleştirilmiştir. Ayrıca sözcük temelli dizinleme ile tamlama temelli dizinleme karşılaştırılarak bilgi erişimi ve başarıımı üzerine etkileri araştırılmıştır.

PHRASE BASED INDEXING IN INFORMATION RETRIEVAL

SUMMARY

Today, information is probably the fastest growing and most valuable thing. The invention of writing, Gutenberg's movable-type printing press and internet are milestones in the history of information. Information could be carried over to next generations restrictedly by the invention of writing and was spread much faster and reached much more people by Gutenberg's movable-type printing press. In the second half of the 20th century, the production and sharing of information became much easier by internet and humanity reached its highest volume of information in its history. In this abundance of information, an important amount of information is discarded or not taken into consideration consciously or unconsciously. Generally, humanity is having poverty of information because of lack of attention and can not reach the desired information. In this situation, the retrieval of information becomes as important as having information.

Information Retrieval (IR) is the research subject that deals with the representation, storage, organization and retrieval of information. As it is possible to use multimedia information like sound, image or video as collection of information for IR, text documents are most common. When IR deals with text documents, it evaluates the documents with a bag of words approach and uses classical methods like word count or word frequency. But documents in digital media are mostly written in natural languages and they also contain linguistic variations and semantical complexity of natural languages. Methods which are only based on word frequency are not adequate to deal with these variations and complexity. This situation pioneers studies for Natural Language Processing to develop new methods for representing and processing of text documents.

Natural Language Processing (NLP) is a field of science which studies the analysis of the verbal and written natural languages used by human. NLP methods should be used in IR to consider this kind of linguistic variations in text documents and improve the performance. Turkish is an agglutinative language; this, on the one hand makes NLP studies hard and increases its complexity but, on the other hand, it's regular structure makes it very suitable.

In this study, an Information Retrieval system supported by Natural Language Processing for Turkish is designed. This IR system aims at increasing performance in IR by considering features and structure of Turkish language. Designed IR system deals with morphological and syntactical variations of Turkish and studies of the effects on overcoming these variations on IR. Morphological variations are handled with stemming based on morphological analysis and syntactical variations with

phrase analysis. For phrase analysis, both statistical and linguistic approaches are studied and linguistical phrase analysis is implemented. In statistical approach, words which follow each other and are frequently repeated in a document are considered as a phrase and terms of proximity and frequency for words are studied. In linguistic approach, noun and adjective phrases are analyzed. The type and suffix information for words are studied to find noun and adjective phrases in Turkish. Finite state machines and regular expressions are formed for the analysis of linguistic phrases. Phrases, which are determined by phrase analysis, are used for representing documents instead of single words. Phrase based indexing is implemented for this purpose. Comparison between phrase based indexing and term based indexing is done and their effect on IR are studied.

1. GİRİŞ

Bilgi, belki de günümüzde en hızlı artan, en değerli şeydir. Özellikle teknik ve bilimsel bilgi çağlar boyu önemli aşamalardan geçmiş, katlanarak artmıştır. Bilginin tarihinde yazının icadı, Gutenberg'in matbaası ve internet önemli yer tutmaktadır. Yazının icadı ile gündelik bilgi sonraki kuşaklara kısıtlı da olsa aktarılabilmiş, Gutenberg'in matbaası ile bilgi daha hızlı çoğaltılıp daha geniş kitlelere ulaşabilmiştir. 20. yüzyılın son yarısında internetin ortaya çıkması ile bilginin üretilmesi ve paylaşılması çok kolay hale gelmiş ve insanlık tarihi boyunca ulaştığı en büyük bilgi seviyesine erişmiştir.

Milattan önce 3. yüzyılda İskenderiye Kütüphanesi'nde yaklaşık 700 bin belge olduğu tahmin edilmektedir. Bu kütüphanenin yok olmasından sonra insanlığın aynı büyüklükte belgeye sahip kütüphaneler oluşturması 19. yüzyılın ortalarını bulacaktır. Şuan bile dünyanın en büyük kütüphanesi olan ABD Kongre Kütüphanesi'nde, 20. yüzyılın ikinci yarısında yüz milyon belge toplanacaktır. İnternet, 20. yüzyılın sonlarında ortaya çıkışıyla önemli bir bilgi kaynağını ortaya çıkarmış ve bilginin boyutu da oldukça artmıştır. İnternetin boyutu 1990'lı yıllarda birkaç yüz milyon sayfa iken 2001 yılında 550 milyon sayfa, 2005 yılında 11.5 milyar sayfaya ulaşmıştır.

İnsanlık sadece 2000 yılında bir ile iki exabyte büyüklüğünde yani dünya üzerindeki her insan için 250 megabyte büyüklüğünde tekil bilgi ortaya çıkmıştır. Bu rakam 2003 yılına gelindiğinde 5 exabyte büyüklüğüne yani her insan için 800 megabyte büyüklüğe ulaşmıştır [1]. Görüldüğü gibi sadece bilgi değil, bilginin artış hızı da artmaktadır.

Bilgi bolluğu her ne kadar gelişme için önemli olsa da önemli sorunlar da doğurmaktadır. Bilginin boyutunun ivmelenmeye başladığı 20. yüzyılın ikinci yarısında Donald Knuth şu ifadeyi kullanmıştır: “Bazen gerçekten kullanabileceğimizden daha çok bilgi ile yüzleşiyoruz ve en ustaca olanı bu bilgiyi

unutmak veya yok etmek olabilir.”. Gerçekten de, bilgi bolluğunda önemli boyutlarda bilgi bilinçli veya bilinçsiz olarak unutulmakta veya göz ardı edilmektedir. Çoğu durumda insanlık bilgi bolluğunda ilgi yoksulluğundan dolayı istenen bilgiye ulaşamamaktadır. Bu durumda bilgiye sahip olmak kadar bilgiye erişmek de önem kazanmaktadır.

Bilgi erişimi, bilginin temsil edilmesi, saklanması, düzenlenmesi ve gerektiği zamanda erişilebilmesini mümkün hale getirmek için yöntemlerin geliştirildiği araştırma konusudur [2]. Bilgi Erişim sistemleri üzerinde çalıştırıldıkları düzenli veya düzensiz bilgi yığınlarında bilgiyi temsil eder ve kullanıcının isteği veya sorgulaması ile bilgiyi yüksek doğrulukta ve kısa zamanda bulan ve getiren sistemlerdir.

Bilgi erişimi, ses, görüntü gibi çoklu ortam bilgi yığınları üzerinde çalışsa da en çok yazılı belgeler üzerinde çalışmaktadır. Bilgi erişimi, belgeler üzerinde çalışırken belgeleri, *sözcük torbası* yaklaşımıyla değerlendirir ve genellikle sözcük sayısı veya sözcük sıklığı gibi genel yöntemler kullanır. Fakat sayısal ortamda saklanan yazılı belgelerin birçoğu doğal dille yazılmış olduğundan, doğal dillerin barındırdığı dilbilimsel çeşitliliği ve anlamsal karmaşıklığı da içerirler. Bu çeşitliliğe örnek olarak sokak, sokaklar, sokağa ve sokağın gibi aynı sözcüğün diğer sözcükler ile ilişkilerinden dolayı oluşan biçimsel farklılıklar, öykü, hikaye gibi eşanlamlı sözcüklerin oluşturduğu sözcüksel farklılıklar, yüz, kara ve koyun gibi birden fazla temel anlamı olan eşsesli sözcüklerin oluşturduğu anlamsal farklılıklar ve benzer sözcüklerin farklı sıralanışıyla oluşmuş sözcük öbeklerinin farklı anlamlar yaratması ile oluşan sözdizimsel farklılıklar verilebilir. Sadece sözcük sıklığına dayanan genel bilgi erişim yöntemleri bu farklılıklar karşısında yetersiz kalmaktadır. Bu durum metinlerin temsil edilmesi ve işlenmesinde yeni yöntemlerin geliştirilmesine yani Doğal Dil İşleme konusu üzerine çalışmalara yol açar.

Doğal Dil İşleme (DDİ), insanlar tarafından kullanılan sözlü ve yazılı doğal dilleri inceleyen, anlaşılması ve çözümlenmesi üzerine çalışan bir bilim dalıdır. DDİ, yapay zeka ve dilbilimi alanlarının bir alt dalı olmakla birlikte felsefe ve psikoloji gibi diğer bilim dallarından da faydalanan bir bilimler arası disiplindir [3]. DDİ tanımlanırken kullanılan *doğal* sözü, insanlar tarafından konuşma ve yazıda kullanılan dili, matematiksel yazım ya da bilgisayar dillerinden ayırmak için kullanılmaktadır [4].

Bilgi erişimi alanında ele alınan metinlerdeki dilbilimsel farklılıkları gidermek ve bilgi erişimi sistemlerinin başarımını artırmak için Doğal Dil İşleme yöntemlerinden yararlanılması gerekmektedir.

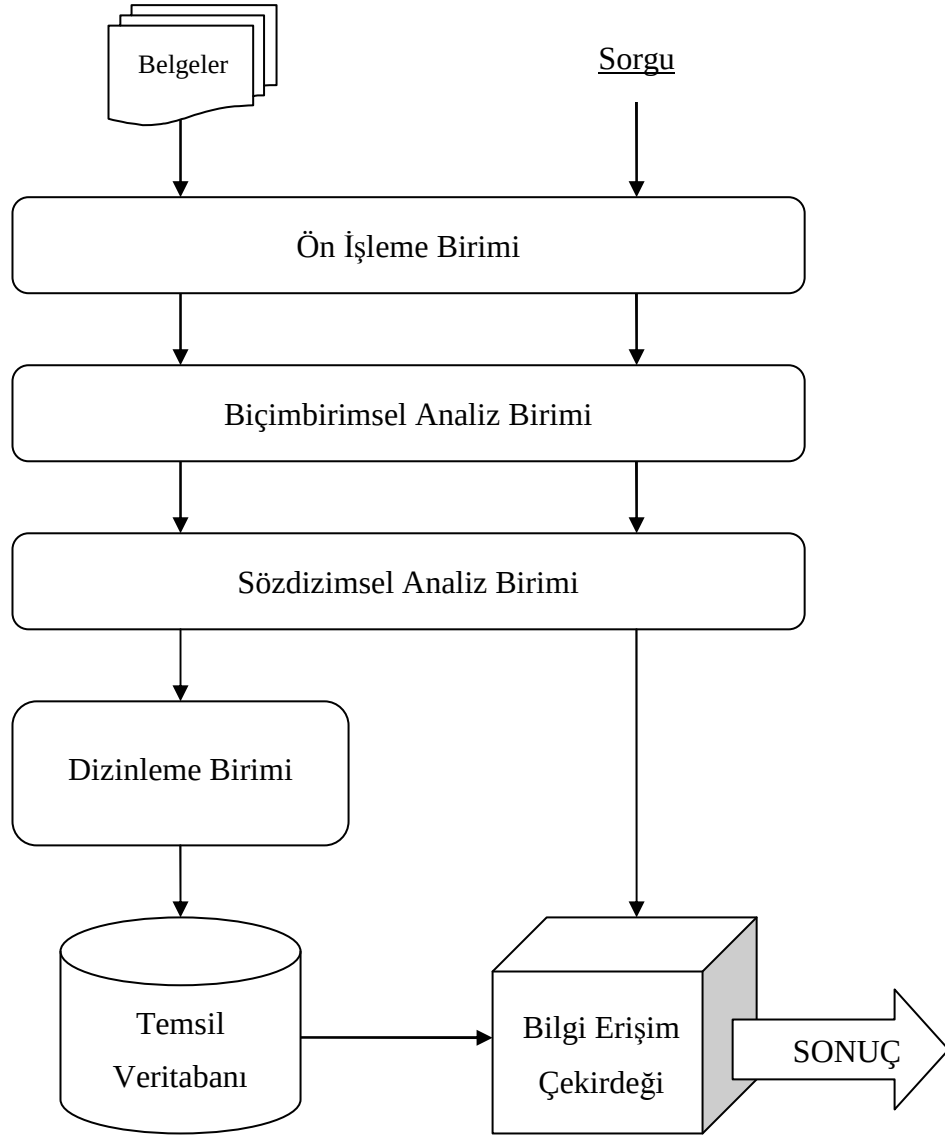
Türkçe, sondan eklemeli yapısı ile Doğal Dil İşleme çalışmaları için karmaşık ve zor, fakat kurallı yapısı ile de bir o kadar elverişlidir. Türkçenin cümle, sözcük ve hatta harf seviyesinde geçmişten günümüze kadar korunarak gelmiş olan kuralları vardır. Sözcük üretimi oldukça kurallıdır ve istisna durumlar azdır. Her ek bir anlam taşır ve eklendiği sözcüğün anlamını o yönde değiştirir.

Türkçe, İngilizce, Almanca ve Fransızca gibi Hint – Avrupa dilleri ile karşılaştırıldığında iki önemli farklılık öne çıkar. Bunlardan biri Türkçenin daha az sayıda köke sahip olmasına rağmen sahip olduğu zengin ek hazinesine sahip olmasıdır. Bu özellik Türkçenin anlamsal olarak yetersiz bir dil olduğu gibi yorumlanabilecekse de gerçek bunun tam tersidir. Türkçe daha öz sözcüğe sahip olmasına rağmen az sözcük ile çok şey anlatabilen, anlam zenginliği yaratılabilen bir dildir. Bu diller ile Türkçe karşılaştırıldığında ortaya çıkan diğer bir önemli farklılık da yapılarıdır. Türkçe sondan eklemeli yapısı ile bir köke teorik olarak sonsuz sayıda ek eklenebilmesine olanak vermektedir. Fakat örneğin İngilizcede bir veya en fazla iki ek peşpeşe bir sözcüğe eklenebilmektedir. Bu özellik de ilk bakışta Türkçenin zor bir dil olduğu izlenimi uyandırabilir. Fakat tam tersine Türkçe yapı olarak kurallıdır ve bir o kadar da önemli olarak kuralları birbiri içerisinde tutarlıdır, istisnai durumlar azdır. Örneğin; İngilizcedeki çoğul ekleri göz önüne alınırsa, genel kural sözcüklerin sonuna (-s) eki gelmesidir. Fakat *foot* sözcüğü, *feet* olmaktadır. Benzer harf dizimine sahip olan *boot* sözcüğü *beet* olarak değil de *boots* olarak kullanılmaktadır. Bu iki özellik bilgisayar bilimleri perspektifinden incelendiğinde Türkçe üzerine DDİ çalışmalarının önemini de ortaya koymaktadır. Özellikle Bilgi Erişimi konusunda DDİ çalışmaları ile bu anlamsal zenginliğin iyi analiz edilmesi ve Türkçenin kurallarının düzgün ifade edilip yapısal farklılıkların ortadan kaldırılması ile başarımlar artırılabilir.

Bu tez kapsamında, Türkçe için doğal dil işleme destekli bir bilgi erişim sistemi tasarımı yapılmıştır. Bu bilgi erişim sistemi, Türkçenin özelliklerini ve yapısını göz önüne alarak, bilgi erişimindeki başarımları arttırmayı amaçlamaktadır. Tasarlanmış olan bu sistemde Türkçenin biçimsel ve sözdizimsel farklılıkları ele alınmış ve bu

farklılıkların giderilmesinin bilgi erişimi üzerindeki etkileri irdelenmiştir. Birimsel farklılıklar biçimbirimsel çözümlemeye dayanan gövdeleme, sözdizimsel farklılıklar ise tamlama analizi ile giderilmeye çalışılmıştır. Tamlama analizinde hem istatistiksel hem de dilbilimsel yaklaşımlar incelenmiş ve dilbilimsel tamlama analizi gerçekleştirilmiştir. İstatistiksel yaklaşımda, birbirini belirli sıklıklarla izleyen ve belge genelinde sıkça geçen sözcük öbekleri tamlama olarak kabul edilmiş, dilbilimsel yaklaşımda ise isim ve sıfat tamlamaları bulunmaya çalışılmıştır. İstatistiksel yaklaşımda sözcükler arasındaki yakınlık ve sıklık ilişkileri üzerinde durulurken dilbilimsel yaklaşımda biçimbirimsel çözümleme de kullanılarak sözcük türleri ve sözcüklere gelen tamlama ekleri incelenip Türkçedeki isim ve sıfat tamlamaları sonlu durum makineleri ve düzenli ifadeler olarak temsil edilmiştir. Tamlama analizi ile bulunan tamlamalar tek sözcüklerle beraber belgeleri temsil etmek için kullanılarak tamlama temelli dizinleme gerçekleştirilmiştir. Ayrıca sözcük temelli dizinleme ile tamlama temelli dizinleme karşılaştırılarak bilgi erişimi ve başarımı üzerine etkileri araştırılmıştır.

Bilgi Erişim sistemi, dört ana birime bölünmüş, bu sayede her birim kendi görevlerinden sorumlu olup her birimin başarımı ayrı ayrı ölçülebilmektedir. Ön işleme birimi, belgeleri birimsel biçimbirimsel analiz birimindeki işlemlere hazırlar. Belgeyi cümlelere ve sözcüklere böler ve anlamsal olarak belgeyi temsil edemeyecek sık geçen sözcükleri temizler. Biçimbirimsel analiz birimi, her sözcük için biçimbirimsel çözümleme yapar ve farklı yöntemlere göre çözümleme sonucundan uygun gövdeyi ve sözcük türünü bulur. Sözdizimsel analize birimi ise, tamlamaları hem istatistiksel hem de dilbilimsel yöntemler ile bulmaya çalışır. İstatistiksel tamlama bulma yöntemi sözcüklerin birbirlerine yakınlıklarına dayanırken dilbilimsel tamlama bulma yöntemi tez kapsamında hazırlanmış sonlu durum makineleri ve düzenli ifadeler çerçevesinde isim ve sıfat tamlamalarını bulmaya dayanır. Dizinleme birimi, belgeleri temsil edecek içerik tanımlayıcılarını seçer ve temsil veritabanına kaydeder. Belgeler ve sorgular bütün bu birimlerde işlendikten sonra sonuçlar ortaya çıkar. Tasarlanan Doğal Dil İşleme destekli Bilgi Erişimi sisteminin genel tasarımı Şekil 1.1'deki gibidir.



Şekil 1.1: Doğal Dil İşleme destekli Bilgi Erişimi sisteminin genel tasarımı

Tez, altı bölümden oluşmaktadır. İkinci bölümde, tezin kapsamına giren Türkçenin genel kuralları incelenmiştir. Üçüncü bölümde, Bilgi Erişimi ve Doğal Dil İşleme konusu ele alınıp bu konularda daha önce yapılmış çalışmalar incelenmiştir. Dördüncü bölümde tamlama analizi yöntemleri üzerinde durulmuştur. Beşinci bölümde tasarlanmış olan Doğal Dil İşleme destekli Bilgi Erişimi sisteminin detayları ve kullanımı açıklanmıştır. Altıncı bölümse ise çalışmanın sonuçlarına ve sonraki çalışmalara yer verilmiştir.

2. TÜRKÇE

2.1 Türkçenin genel özellikleri

Türkçe, Doğu Avrupa'dan Sibiryaya ve Çin'in batısına kadar yaygın bir şekilde kullanılan Türk dilleri ailesinin en çok konuşulan dilidir. Türkçe, Türkiye Cumhuriyeti, Kuzey Kıbrıs Türk Cumhuriyeti, Kıbrıs Rum Kesimi ve Makedonya'nın Kosova gibi belli bölgelerinde resmi dildir. Yapılan araştırmalar sonucunda dünya üzerinde 100 milyon civarında insan Türkçeyi ana dili olarak konuşmaktadır. UNESCO, 1980'li yıllarda yaptığı araştırma sonucu Türk lehçelerini ikinci ya da üçüncü dil olarak konuşanlar da dahil edildiğinde tüm Türk lehçelerini 200 milyon kişinin konuştuğunu ortaya çıkarmıştır [5].

Türkçede, diğer Türk dillerinde olduğu gibi büyük ve küçük ses uyumu vardır, yazımda sözcükler son ekler alarak uzarlar ve cümle yapısı özne-nesne-fiil sırasıyla oluşturulur.

2.1.1 Kökeni bakımından Türkçe

Yeryüzünde, bugüne kadar 2796 farklı dil konuşulduğu tahmin edilmektedir. Bunlardan bugün ancak 118 tanesi devlet dili olarak kullanılmıştır [6]. Kullanılan bu diller incelendiğinde diller arasında bir takım yakınlıklar ve akrabalıklar olduğu gözlenmiştir.

Dilleri sınıflandırırken kullanılan en yaygın yöntem kökendir. Kökenleri bakımından akraba olan dillerin bir ana kaynaktan, bir ana dilden çıktıkları düşünülmektedir. Diller kökenlerine göre sınıflandırıldığında 5 ana dil ailesi ortaya çıkar.

1. Ural – Altay Dilleri Ailesi
2. Hint – Avrupa Dilleri Ailesi

3. Çin – Tibet Dilleri Ailesi

4. Sami Dilleri Ailesi

5. Bantu Dilleri Ailesi

Türkçe, kökeni itibari ile Ural – Altay dil ailesinin Altay koluna ait bir dildir.

2.1.2 Yapı bakımından Türkçe

Dillerin sınıflandırılmasında kullanılan diğer bir yöntem de dillerin yapısıdır. Dillerin yapıları, sesler, kelime türetirken kullanılan teknik, sözcükleri anlamlı cümleler haline getirirken sözcük aralarındaki ilişkilerin nasıl belirtildiği, cümle içerisindeki kelimelerin sıralanışı gibi dilin kendisi ile ilgili özelliklerini içine alır [7].

Yapı bakımında sınıflandırmada dillerin ses, dilbilgisi ve kelime hazinelerindeki benzerlik esas alınır. Bu konudaki en yaygın ölçüt sözcük türetiminde kullanılan yöntemidir. August Wilhelm Schlegel (1767-1845) ve Wilhelm von Humboldt (1767-1835) tarafından 19. yüzyıl başında ortaya konan biçimbilgisi ölçütüne göre diller üç ana sınıfta toplanır:

1. Yalınlayan Diller

2. Bükümlü Diller

3. Eklemeli Diller

2.1.2.1 Yalınlayan Diller

Yalınlayan dillerde diğer adı ile tek heceli dillerde, hemen hemen bütün sözcükler tek hecelidir. Sözcükler hiçbir ek almazlar. Sözcükler tümce içinde vurguya bağlı olarak anlam kazanırlar. Çince, Vietnamca ve Tibet dilleri yalınlayan diller kümesinde yer alırlar.

2.1.2.2 Bükümlü Diller

Bükümlü dillerde dilbilgisi işlevleri sözcük kökündeki değişikliklerle ifade edilir. Bükümlü dillerde sözcük türetilirken kökün ünsüz harfleri değişmez, ünlü harfleri

değişir. Ekler, köklere önek, içek ve sonek olarak getirilebilir. Hint – Avrupa ve Sami dilleri bu sınıfta yer alır.

2.1.2.3 Eklemeli Diller

Eklemeli dillerde ise bir ya da birden çok heceli sözcük kökleri ile ekler bulunur. Her sözcükten eklerle yeni anlamda sözcükler türetilbildiği gibi, her sözcük, eklerle geçici durumlara da sokulabilir. Bu gruptaki dillerde kökler ve ekler birbirinden kolaylıkla ayırt edilebilir.

Eklemeli dillerde ekler, sözcüğün başına ya da sonuna getirilebilir. Bu özelliğe göre bu gruptaki diller, önden eklemeli ya da sondan eklemeli olarak adlandırılabilir. Türkçe, Japonca, Fince, Macarca ve Moğolca eklemeli diller grubunda yer alır.

Özellikle Türkçe, sondan eklemeli yapısı ile bilgisayar bilimleri disiplini olarak doğal dil işleme konusunda üzerinde çalışılması karmaşık fakat bir o kadar da verimli sonuçlar veren bir dildir.

2.2 Biçim Bilgisi

Biçim bilgisi, bir dilin sözcüklerini, sözcük yapılarını türeme yollarını, çekim biçimlerini inceleyen dilbilgisi koldur. Her sözcüğün bir ses yapısı bir de biçim yapısı vardır. Sözcüğün biçim yapısı anlamlı ve anlamsız öğelerden oluşur [6]. *Kök*, bir sözcüğün anlamı ve yapısı bozulmadan daha küçük parçalara ayrılmayan parçasıdır. *Ek* ise, tek başlarına anlamları olmayan ancak sözcük türetmekle ya da sözcüklere ilgi, nicelik, kip, zaman ve kişi ile ilgili anlam katmakla görevli dilbilgisi öğeleridir. Örneğin, *gölgesizlik* sözcüğü gölge köküne –*siz* ve –*lik* ekleri eklenerek oluşmuştur.

Kökler, anlamları bakımından *isim kökleri* ve *fiil kökleri* olarak ayrılabilir. Ekler ise görevleri bakımından *çekim ekleri* ve *yapım ekleri* olarak ikiye ayrılır.

Türkçe, sondan eklemeli bir dil olduğu için çekim ve yapım ekleri bakımından oldukça zengindir. Bu özellik, birçok dilin birer bağımsız sözcük ile ifade ettiği anlamın, eklerle ifade edilmesini sağlamıştır.

Türkçede, aynı görevde yapım eki veya aynı çekim eki üst üste gelemez; fakat farklı yapım ve çekim ekleri üst üste gelebilir. Çekim eki her zaman yapım ekinden sonra gelir; ama çekim ekinden sonra yapım eki gelemez. –ki yapım eki bu durum için oluşmuş istisnadır.

2.2.1 Çekim Ekleri

Çekim ekleri, kalıplaşanlar dışında eklendikleri kök veya gövdelerden yeni sözcükler türetmeyen, sözcükler arasında durum, iyelik, çokluk, kip, zaman ve şahıs gibi ilgiler kuran eklerdir. Çekim ekleri cümleden çıkarılırsa sözcükler arasındaki anlam ilgileri kopar ve yapı bozulur.

Çekim ekleri, iki gruba ayrılabilirler: isim çekim ekleri ve fiil çekim ekleri

2.2.2 Yapım Ekleri

Yapım ekleri, eklendikleri kök ve gövdelerin anlamını ve sınıfını değiştiren eklerdir. Türkçede iki yüze yakın yapım eki vardır. Bu eklerin bazıları çok işlektir ve yüzlerce sözcüğün yapısında yer almıştır; bazıları ise ancak bir iki sözcükte yaşamaktadır ve yeni sözcük türetiminde kullanılmaz. Dil, işlek olan eki kullanmayı tercih eder.

Yapım eklerinin bazıları, zamanla eklendikleri kökle kaynaşır, ayrılmaz hale gelir. Bu sözcüklerde kök neredeyse unutulur ve gövde kök niyetine kullanılır. Örneğin; *barış, bekçi, inan* vb.

Yapım ekleri dört gruba ayrılabilirler: isimden isim yapan ekler, isimden fiil yapan ekler, fiilden isim yapan ekler ve fiilden fiil yapan ekler.

2.3 Sözcük Bilgisi

Türkçede sözcükler isim soylu sözcükler, görevli sözcükler ve eylemler olmak üzere üç ana gruba ayrılırlar. Bu ayırmada sözcüklerin anlamları kadar sonlarına alabildikleri ekler de önemli bir ölçüttür. Örneğin Türkçede isim soylu sözcüklerle fiillere gelen ekler birbirinden tamamen farklıdır.

2.3.1 İsim

Varlıkları tanıtmaya yarayan sözcüklere denir. İsimler varlıklara verilişlerine göre özel isim ve cins ismi; varlıkların sayılarına göre tekil isim ve çoğul isim ; varlıkların oluşlarına göre somut isim ve soyut isim olarak türlere ayrılırlar.

Bütün özel isimler ve özel isimlerin ayrı yazılan her sözcüğü büyük harfle, cins isimleri cümle başında bulunmadıkça küçük harfle yazılırlar. İsimler kök ve gövdeler olarak kök sözlüğüne katılmışlardır. Ayrıca alabilecekleri ekler ve uyabilecekleri modeller de sistemde yer almaktadır.

2.3.2 Sıfat

Varlıkları niteleyen ya da sayılarını, yerlerini belirten sözcüklere denir. Bu görevleri bakımından sıfatlar ikiye ayrılırlar :

1. Niteleme sıfatları : Varlıkların durumlarını, biçimlerini, renklerini, kısaca nasıl olduklarını gösteren sözcüklere denir.
2. Belirtme sıfatları : Varlıkları türlü yönlerden belirten sözcüklerdir. Görevleri bakımından işaret, sayı, belgisiz ve soru sıfatları olarak dörde ayrılırlar.

Sıfatlar, belirteçlerle derecelendikleri gibi bazı öneklerle de kuvvetlenir. Bu öneklerle kuvvetlendirmeye pekiştirme denir. Sıfatlar yalın ve pekiştirilmiş olarak kök sözlüğünde yer alırlar.

2.3.3 Zamir

Cümlede isimlerin yerini tutan ve söyleyenle kendisine söylenen kimseleri anımsatan sözcüklere zamir denir. Zamirler, kişi zamirleri, işaret zamirleri, iyelik zamirleri, dönüşlü zamirler ve belirsiz zamirler olarak beşe ayrılır. Zamirler isimlere benzerler, çoğul olur, çekim eki alırlar.

2.3.4 Eylem

Varlıkların yaptıkları işleri, kılışları ya da onlarla ilgili oluşları zamana ve kişiye bağlayarak anlatan sözcüklerdir. Her tam eylemde edim, zaman, kip, kişi ve çatı kavramları kaynaşmıştır. Edimler, eylemlerin kök ya da gövde anlamıdır.

2.3.5 Zarf (Belirteç)

Sıfatların, eylemlerin ve görevce kendilerine benzeyen sözcüklerin anlamlarını berkiten ya da kısıp sınırlayan sözcüklere zarf ya da belirteç denir. Belirteçler görevleri bakımından beş çeşittir: zaman, yer, durum, azlık-çokluk ve soru belirteçleri.

2.3.6 Edat (İlgeç)

Kavramlar arasında türlü anlam ilgisi kurmaya yarayan ve anlamları ancak bu görevleriyle beliren sözcüklere edat veya ilgeç denir. Başlıcaları şunlardır: gibi, kadar, ile, için, sanki, dolayı, üzere, yalnız, ancak ve vb.

2.3.7 Bağlaç

Anlamca ilgili tümceleri, görevdeş öğeleri bağlamaya yarayan sözcüklere bağlaç denir. Başlıcaları şunlardır: ve, ki, çünkü, halbuki, madem, mademki, nitekim, ile, hele, zira, kah...kah, fakat, lakin, ya...ya, yahut, veyahut, veya, ancak, hem...hem.

2.3.8 Ünlem

Bir coşkunun etkisiyle içten kopup gelen;sevinç, korku, üzüntü, acı, şaşma gibi duyguları canlı canlı anlatmaya yarayan sözcüklerdir. Ünlemlerin başlıcaları şunlardır:a!, ha!, ey!, hey!, ay!, vay!, ah!, oh!, be!, eyvah!, aman!, yaşa!.

2.4 Sözcük Grupları

Birden fazla sözcüğün bir araya gelerek, anlam ve yapı bakımından tek sözcük bütünlüğü oluşturmalarına sözcük grubu denir. Sözcük grupları, cümle içinde tek bir

sözcük gibi görev alırlar, değerleri bağımsız, tek bir sözcük ile aynıdır. Bundan dolayı, ekler sözcük gruplarının sonuna getirilir [7].

Başlıca sözcük grupları isim tamlamalar, sıfat tamlamaları, sıfat fiil grupları, zarf fiil grupları, isim fiil gruplarıdır. Tezin konusu gereği ilerleyen bölümlerde sadece isim ve sıfat tamlamaları üzerinde durulacaktır.

2.4.1 İsim Tamlamaları

Birden fazla adın, ilgi ve iyelik eki aracılığıyla bir araya gelip tek kelime gibi anlam yüklenmesi sonucu oluşturduğu gruba isim tamlaması denir. Ev-in kapı-sı (burada ana öge, kapı'dır; yardımcı öge ise ev'dir. Kapı, eve aittir. Kapının sahibi ev'dir) - Ev kapı-sı (burada ana öge kapı'dır; yardımcı öge ise ev'dir. Ev ile ilgili kapı demektir).

İsim tamlamalarında yardımcı, tamamlayıcı ögeye, yani birinci isme belirten ya da tamlayan; ana ögeye, yani ikinci isme belirtilen ya da tamlanan denir:

Pencerenin camı pencerenin : yardımcı öge (yani tamlayan ya da belirten)

camı : ana öge (tamlanan ya da belirtilen)

Pencere camı pencere : yardımcı öge (tamlayan ya da belirten)

cam : ana öge(tamlanan ya da belirtilen)

İsim tamlamalarında tamlayan yardımcı, tamlanan asıl ögedir. Burada da, Türkçenin sözdizimindeki "*asıl unsurun sonda bulunacağı*" kuralı işlemektedir [7]. İsim tamlamaları, ilgi ekini alıp almamalarına göre belirtili isim tamlaması ve belirtisiz isim tamlaması olmak üzere ikiye ayrılır. Genel olarak bu sınıflandırmaya takısız isim tamlaması ve zincirleme isim tamlaması da dahil edilir.

İsim tamlaması belirli de olsa, belirsiz de olsa, tamlanan öge kesinlikle iyelik eki (-/, -si) alır. Bugün *Kadıköy*, *Mardinkapı*, *Galatasaray* şeklinde iyelik eki olmadan kullandığımız kelimelerin aslının belirsiz isim tamlaması olduğu ve iyelik eki taşıdığı unutulmamalıdır: *Kadıköyü*, *Mardinkapısı*, *Galatasarayı* [7].

2.4.1.1 Belirtili İsim Tamlaması

Belirtili isim tamlaması, tamlayanı söz söyleyen ve dinleyence bilinen ve tamlananı ile tamlananı arasında geçici bir ilgi bulunan tamlamadır. Bu tamlamada tamlayan eki -in; tamlanan eki ise -i'dir. Her iki ek de büyük ünlü uyumuna göre değişim gösterebilir.

Bu tamlamalarda, tamlayan öge belirlidir; bu kullanımla ona belirlilik kazandırılmıştır. Örneğin,

kızın eldiveni → kız – ın eldiven – i

Tamlayan olan “kız”, söz söyleyen ve dinleyence bilinen bir “kız”dır. Tamlanan olan “eldiven” ise belirli ve bilinen bir kızıdır. Kız ile eldiven arasındaki ilgi geçicidir; çünkü eldiveni, bu kıza ait olduğu sürece düşünebiliriz.

Belirtili isim tamlamasının özellikleri şu şekildedir:

1. Tamlayan ile tamlanan öge arasındaki anlam ilişkisi geçicidir. Bundan dolayı, belirtili isim tamlamaları kalıcı isim özelliği kazanmazlar. Eğer kalıcı isim olarak kullanılacak olursa, tamlama belirtisiz isim tamlaması olur: *öğrencinin çantası, öğrenci çantası*.
2. Tamlayan ile tamlanan ögeler yer değiştirebilir: *Uçtu kafesinden kuşu gönlümün*.
3. Tamlayan ile tamlanan ögenin arasına başka bir sözcük girebilir: *Çatının her rüzgârda uçan kiremitleri gittikçe azalıyordu*.
4. Tamlamanın başındaki sıfat, tamlamanın hepsini değil, tamlayan ögeyi nitelendirir: *eski yolun çamuru*
5. Tamlayan öge zamir veya özel isim ise, tamlama belirli olur: *onun masası, benim suyum, Ankara'nın havası*.
6. Tamlanan öge sayı veya belirsizlik zamiri ise, ilgi ekinin yerine ayrılma hâli eki -dan kullanılabilir: *gençlerin üçü → gençlerden üçü; gençlerin birçoğu → gençlerden birçoğu*.

2.4.1.2 Belirtisiz İsim Tamlaması

Belirtisiz isim tamlaması, tamlayanı söz söyleyen ve dinleyence belirli olmayan ve tamlayan ile tamlananı arasında sürekli bir bağ bulunan isim tamlamasıdır. Tamlayan, ek almaz; tamlanan ise -i ekini alır. Bu ek, büyük ünlü uyumuna göre değişim gösterebilir.

Bu tamlamalarda, tamlayan öge belirlidir; bu kullanımla ona belirlilik kazandırılmıştır. Örneğin,

ev duvarı → *ev duvar – ı*

Tamlayan olan ev, söz söyleyen ve dinleyence belirli olmayan bir evdir; bir ev cinsidir. Ev ile duvar arasındaki bağ sürekli; çünkü ev cins ismi olduğuna ve her evin de duvarı bulunduğuna göre, ev ile duvar arasındaki bağ sürekli olarak kalacaktır.

Belirtisiz isim tamlamasının özellikleri şu şekildedir:

1. Tamlayan ile tamlanan öge arasındaki anlam ilişkisi kalıcıdır. Bundan dolayı, belirtisiz isim tamlamaları kalıcı isim özelliği kazanır: *öğrenci çantası, tahta kalemi, aslan ağzı, kalkan balığı, devlet bakanı*. Belirtisiz isim tamlamaları bu yönüyle herhangi bir eşyanın, nesnenin, kavramın vb. adı hâline gelir ve anlam bakımından da uygun olursa, bitişik dahi yazılabilir.
2. Belirtisiz isim tamlamasında tamlayan ile tamlananın arasına, bu öğelerin anlam bakımından birbirlerine kalıcı olarak bağlanmalarından dolayı, başka bir kelime giremez: “*okulun kırık kapısı*” denilebilir; fakat, “*okul kırık kapısı*” denilemez. Bundan dolayı, bu tamlamalarda niteleyici kelime başta kullanılır ve tamlamanın hepsini nitelendirir: yüksek inşaat mühendisi.
3. Belirsiz isim tamlamasında öğeler yer değiştiremez. Öğrenci çantası tamlaması, çantası öğrenci şeklinde söylenemez; eğer söylenirse bu, “*Çantası öğrenci(dir).*” şeklinde algılanacak ve yapı bakımından doğru, anlam bakımından ise mantıksız bir ifade ortaya çıkacaktır.

Belirtili ve belirtisiz isim tamlamalarının yapı özellikleri şu şekilde özetlenebilir:

1. Belirtili ve belirtisiz isim tamlamasının bir ögesi olan her sözcük, dilbilgisindeki sözcük çeşidi ne olursa olsun, isimleşmiş sayılır: *Yeşilin rengi, suyun yeşili, yeşilin açığı, düşenin dostu, okuma zevki, okumanın zevki, köylünün okumuşu, okur mektupları gibi.*
2. İsimler gibi zamirler de belirtili isim tamlamalarında yer alabilirler.
 - a. “Kendi” zamiri dışındaki kişi zamirleri, yalnızca tamlayan olabilirler; “kendi” zamiri ise hem tamlayan hem de tamlanan olabilir: *Onun kitabı, bizim evimiz, kendisinin kitabı, adamın kendisi gibi.*
 - b. Kalıplaşmış sözcükler dışında işaret zamirleri yalnızca tamlayan olabilirler: *Şunun defteri, bunların işi, bunun yaramazlığı* (Onun şusu busu sözü, kalıplaşmış bir sözdür).
 - c. Belgisiz zamirlerle soru zamirleri, tamlayan ya da tamlanan olabilirler: *Birçoğunun giysisi, bazısının parası, evlerin her birisi, çocukların heps, hangisinin kitabı?, memleketin nereleri?... gibi.*
3. Zamirler, belirtisiz isim tamlamalarının hiçbir ögesi olamazlar.
4. Sıfatlar da belirtili isim tamlamalarında bulunabilirler.
 - a. Niteleme sıfatları, de belirtilen de ya da aynı tamlamada her iki öge de olabilirler: *Suyun yeşili, yeşilin rengi, yeşilin açığı...* gibi.
 - b. İşaret sıfatları hiçbir öge olamazlar.
 - c. Sayı sıfatları, tamlayan, tamlanan ya da aynı tamlananın her iki ögesi de olabilirler: *On üçün uğursuzluğu, birincinin kitabı, çeyreğin yarısı, elmaların ikisi.*
 - d. Sayı sıfatı belirtilen olursa, çoğul anlamı taşıyan tamlayan, -in eki yerine -den eki alabilir: *Elmaların (elmalardan) ikisi, çocukların (çocuklardan) üçü...* gibi.
 - e. Belgisiz sıfatlarla soru sıfatları, genellikle tamlanan olarak kullanılırlar. O zaman çoğul anlamı veren tamlayan, -in yerine -den

ekini alabilir: *Çocukların (çocuklardan) birkaçı, memurların (memurlardan) herhangi biri*

5. Sıfatlar belirtisiz isim tamlamalarında da yer alabilirler.
 - a. Niteleme sıfatları, tamlayan da tamlanan da olabilirler; ancak aynı tamlamanın iki ögesi de olamazlar: *tembel harcı, deniz mavisi*
 - b. İşaret sıfatları, hiçbir öge olamazlar.
 - c. Sayı sıfatlarından sıra sayı sıfatları, yalnızca, tamlana olabilirler: *sınıf birincisi*
 - d. Belgisiz sıfatlarla soru sıfatları hiçbir öge olamazlar.
6. -mek'li isim fiiller, belirtili isim tamlamalarında hiçbir öge olamazlar, -me'li ve -iş'li isim-fiiller, belirtili isim tamlamalarında hem tamlayan hem de tamlanan ya da aynı tamlamada her iki öge de olabilirler: *okumanın zevki, Ahmet'in okuyuşu, çalışmanın belirtilmesi, gidişin dönüşü, yükselmenin alçalması*
7. -mek'li isim-fiiller, belirtisiz isim tamlamalarında yalnızca belirten; -me'li ve -iş'li isim-fiillerse, belirtisiz isim tamlamalarında ya tamlayan, ya da tamlanan olabilirler; ancak aynı tamlamada her iki öge oluşları pek seyrek: *okumak zevki, kazanmak hırsı, uyuma gereksinimi, giriş parası*
8. Ortaçlar, belirtili isim tamlamalarında hem tamlayan, hem tamlana ya da aynı tamlamada her iki öge olabilirler: *ölenin kızı, köylünün okuryazarı, gidenlerin döneni*
9. Ortaçlar, belirtisiz isim tamlamalarında yer almazlar.

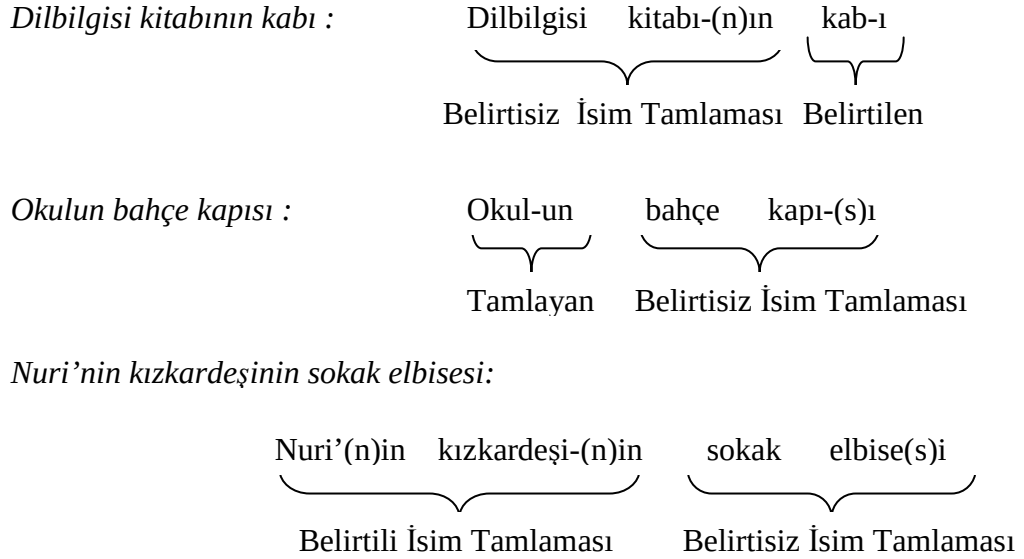
2.4.1.3 Takısız İsim Tamlaması

İsimler, ilgi ve iyelik eki almaksızın bir grup oluşturup daha çok "bir şeyden yapılma" veya "bir şeye benzeme" anlamı taşıyan bu kelime gruplara takısız isim tamlaması denir. Bu tamlama çeşidine örnek olarak "*demir kapı*", "*tahta masa*" ve "*ipek mendil*" verilebilir. Bu tamlamalar için takısız isim tamlaması teriminin

kullanılmasının temel sebebi, her iki ögenin de isim olması ve ek almamasıdır; ancak, Türkçede kelimeler, kullanıldıkları yere göre isim, sıfat veya zarf olabilmektedir, isimlerin arasındaki anlam ilişkisi farklı olsa da, takısız isim tamlamasında, birinci kelime, ikinci kelimeyi nasıl sorusu çerçevesinde nitelendirmiştir.

2.4.1.4 Zincirleme İsim Tamlaması

Birden fazla isim tamlaması birlikte, iç içe kullanıldığı durumlarda zincirleme isim tamlaması oluşur; ancak zincirleme isim tamlamaları da bir bütün olarak belirli veya belirsiz isim tamlamasıdır. Zincirleme isim tamlaması örnekleri ve çözümlemeleri Şekil 2.1'deki gibidir.



Şekil 2.1: Zincirleme isim tamlaması örnekleri

2.4.2 Sıfat Tamlaması

Bir sıfat ve bu sıfatın nitelendirdiği veya belirttiği bir isimden oluşan kelime gruplarına sıfat tamlaması denir. Bu tamlamada, sıfat niteleyen, isim ise nitelenen ögedir:

kırmızı elma, mavi kalem, bulutlu hava.

Sıfat tamlamasının özellikleri aşağıdaki gibidir:

1. Bir isim, birden fazla sıfatla nitelendirilebilir: *uzun, zayıf çocuk*.
2. Birden fazla isim tek sıfatla nitelendirebilir: *kırmızı kalemlik, çanta ve ayakkabı*.
3. Temel anlamıyla sıfat olmayan kelimeler, sıfat olarak kullanılabilir: *altın küpe*.
4. Sıfat tamlamasının çekim eki, nitelenen öge olan isme eklenir: *hızlı araba-da*.
5. Sıfat tamlamalarında, zaman zaman nitelenen isim düşer ve sıfat, ad gibi kullanılır: *Sınıfımızın başarılı öğrencileri geliyor. - Sınıfımızın başarılıları geliyor*.
6. Sıfat tamlaması, kelime grubu ve cümlede herhangi bir isim gibi kullanılabilir:

Demli çay geldi. (özne)

O, demli çayı seviyor. (nesne)

Demli çaya yöneldik. (tümleç)

7. Niteleyen ve nitelenen öğelerin kendileri birer sıfat tamlaması olabilir ve bu tür tamlamaları zincirleme sıfat tamlaması olarak adlandırabiliriz; fakat zincirleme sıfat tamlamaları da temelde bir sıfat tamlamasıdır: *kırmızı kollu yazlık gömlek*
8. Bir isim tamlaması ile sıfat tamlaması kelime grubu oluşturabilir. Bu tür tamlamalara karma tamlama adı verilmektedir; ancak, karma tamlamalar da temelde bir isim veya sıfat tamlamasıdır:

kahvaltılık peynirin fiyatı (karma tamlama / belirli isim tamlaması)

siyah ayakkabı boyası (karma tamlama / sıfat tamlaması)

3. BİLGİ ERİŞİMİ VE DOĞAL DİL İŞLEME

3.1 Bilgi Erişimi

Bilgi Erişimi, bilginin temsil edilmesi, saklanması, düzenlenmesi ve gerektiği zamanda erişilebilmesini mümkün hale getirmek için yöntemlerin geliştirildiği araştırma konusudur [2]. Bu konu, 50 yıllık geçmişe sahip olmasına rağmen son 10 yılda bilginin artması ve biçim değiştirmesi ile artan bir öneme sahip olmaya başlamıştır. Örneğin, sadece 2003 yılında yeryüzündeki her insan başına 800 megabyte bilgi üretilmiş ve bu bilginin ancak çok küçük bir kısmı basılı kaynaklardadır. Günümüzde sadece ulaşılan bilgi değil ayrıca bilginin artım hızı da insanlık tarihinde en yüksek seviyeye ulaşmıştır. Fakat aynı süre zarfında bilgi erişimi aynı hızda gelişmemiştir.

3.1.1 Bilgi Erişiminin Kısa Tarihçesi

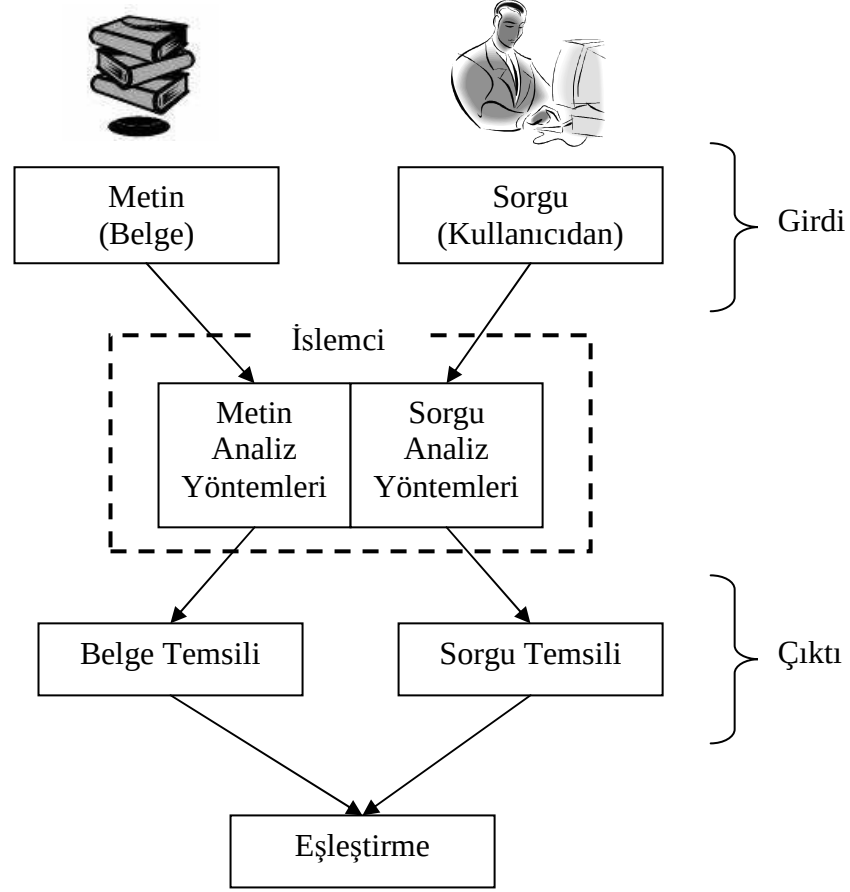
Bilgi Erişimi konusu, bilgisayar bilimleri altında değerlendirildiğinde yaklaşık 50 yıllık geçmişi olmasına rağmen, genel anlamda düşünüldüğünde yazı kadar geçmişi olduğu düşünülebilir. İnsanlık yazının bulunması ile beraber edindiği bilgileri saklayabilmiş ve daha sonra bunları erişip kullanabilmiştir.

Amerikalı bilim adamı Calvin Mooers, Bilgi Erişimi terimini ilk ortaya atan insandır. 1959 yılında ünlü Mooers Yasasını yayınlamıştır. Mooers yasası şu şekildedir: “Bilgi Erişim sistemleri, kullanıcı bilgiye sahip olması olmamasından daha zahmetli ve yorucu olana kadar kullanılmayacaktır.” Calvin Mooers, daha o yıllardan ortaya çıkan bilginin büyüklüğüne ve bilgi erişim sistemlerinin önemine dikkat çekmiştir. Modern bilgi erişiminin babası olarak da Gerard Salton kabul edilmektedir. Salton, Cornell Üniversitesindeki çalışmaları ile bilgi erişimi konusuna önemli bir ivme kazandırmış ve grubu ile birlikte SMART Bilgi Erişim Sistemini geliştirmiştir.

Bilgi Eriřiminin tarihsel süreci ve geliřimi bakımında önemli bir geliřme de 1992 yılında ABD Savunma Bakanlıęının Ulusal Standartlar ve Teknoloji Enstitüsü (National Institute of Standards and Technology (NIST)) ile beraber Metin Eriřim Konferanslarını (Text Retrieval Conference (TREC)) desteklesidir. TREC'in amacı bilgi eriřimi üzerine çalışan topluluęu bir araya getirmek, çalışabilecekleri uygun altyapıyı sağlamaktır. TREC, belirli aralıklarla konferanslar düzenler ve arařtırmacıların üzerinde çalışmaları için belge yığınları yayınlar. TREC belge yığınları TREC-3'de 2 gigabyte, TREC-6'da 5.8 gigabyte büyüklüęüne ulaşmıştır.

3.1.2 Temel Bilgi Eriřim Sistemi

Bilgi Eriřim sistemlerinin, üzerinde çalıştıkları bilgiden, uzmanlaşmış oldukları konulardan bağımsız olarak temel yapıları aynıdır. Bu sistemlerin amacı bilginin temsili ve kullanıcının sorgusunun bu temsil ile karşılaştırılmasıdır. Metin üzerine çalışan Bilgi Eriřim sistemlerinde derlemde yer alan belgeler, metin olarak değerlendirilir ve belirli analiz işleminden geçirilerek temsilleri oluşturulur. Benzer şekilde kullanıcı tarafından belirlenen sorgular da belirli analiz işlemlerinden geçirilerek temsili oluşturulur. Metin ve sorgu analiz yöntemleri çoęunlukla ortak olsa da farklı yöntemler de kullanılmaktadırlar. Belge ve sorguların temsilleri oluşturulduktan sonra sorgu temsili ile en iyi eşleşen belge temsili bulunmaya çalışılır. En iyi eşleşmeyi sağlayan belge, sorgu ile en ilgili belge olarak kabul edilir. Metin üzerinde çalışan temel bilgi eriřimi sistemi ve temel bileşenleri Şekil 3.1'deki gibidir.



Şekil 3.1: Temel bilgi sistemi tasarımı

3.1.3 Bilgi Erişiminde Temel Kavramlar

Başarım Ölçüm Kavramları

Bilgi Erişiminde başarımlar genellikle keskinlik ve anımsama ile değerlendirilir. *Keskinlik*, bulunan ilgili belgelerin tüm bulunan belgelere oranı olarak tanımlanabilir. *Anımsama* ise, bulunan ilgili belgelerin tüm ilgili belgelere oranıdır. Keskinlik ve anımsama kavramları Denklem (3.1) ve Denklem (3.2) ile temsil edilebilirler [8].

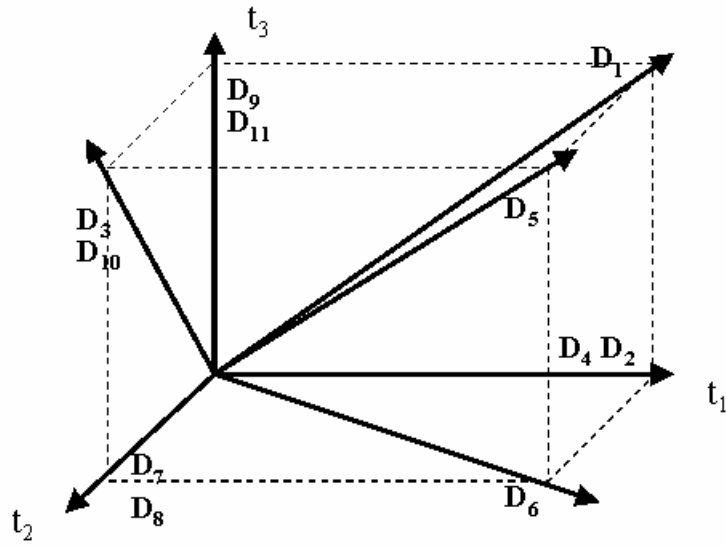
$$Keskinlik = \frac{\{ilgili\ belgeler\} \cap \{bulunan\ belgeler\}}{\{bulunan\ belgeler\}} \quad (3.1)$$

$$Anımsama = \frac{\{ilgili\ belgeler\} \cap \{bulunan\ belgeler\}}{\{ilgili\ belgeler\}} \quad (3.2)$$

Keskinlik ve anımsama başarımları değerli yüksek olan Bilgi Erişim sistemleri başarılı olarak kabul edilir.

Yöney Uzak Modeli

Yöney Uzak Modelinde, belgeler ve sorgular çok boyutlu uzayda birer yöney olarak temsil edilirler. Uzaydaki her boyut derlemdeki anahtar sözcük ya da tamlamaya denk düşer. Örneğin, Şekil 3.2'deki uzay, 3 boyutlu olup t_1 , t_2 , t_3 boyutlarında 11 adet belge temsil edilmiştir.



Şekil 3.2: 3 Boyutlu örnek yöney uzayı

Sorgu ve belge yöneylerini temsil eden ifadeler Denklem (3.3)'teki gibidir.

$$\begin{aligned} \vec{b}_j &= (w_{1,j}, w_{2,j}, w_{3,j}, \dots, w_{n,j}) \\ \vec{s}_k &= (w_{1,k}, w_{2,k}, w_{3,k}, \dots, w_{n,k}) \end{aligned} \quad (3.3)$$

Sorgu ile en ilgili belgeler, sorgu yöneyine en yakın yöneye sahip olan belgelerdir. Normalde bu benzerlik yöneyler arasında nokta çarpımı yapılarak hesaplanmalıdır. Fakat bu hesaplama boyutların büyüklüklerine karşı hassas olduğu için tercih edilmemektedir. Bunun yerine yöneyler arasındaki bu yakınlık yöneyler arasında kalan açının kosinüs değeri ile ölçülür [9]. İki yöney arasındaki benzerlik Denklem (3.4) ile hesaplanabilir.

$$ben(\vec{s}_k, \vec{b}_j) = \frac{\sum_{i=1}^n w_{i,k} \times w_{i,j}}{\sqrt{\sum_{i=1}^n w_{i,k}^2} \times \sqrt{\sum_{i=1}^n w_{i,j}^2}} \quad (3.4)$$

Terim Ağırlıklandırma

Bilgi Erişiminde, derlemde bulunan belgeler ile kullanıcı tarafından girilen sorgunun temsili oldukça önemlidir. Bu temsilde yöney uzay modelinde her anahtar sözcüğün yani uzayın ağırlığı öne çıkmaktadır. Yöneiy oluşturan uzayların ağırlıkları belgenin veya sorgunun temsilini de etkilemektedir.

Terim ağırlandırılmasında kullanılan iki parametre terim sıklığı ve belge sıklığıdır. Terim sıklığı, bir terimin bir belgede kaç kere geçtiği ile ilgilidir. Terim sıklığı en basit şekilde sözcük sayısı ile hesaplandığında ağırlık hesaplamalarında kısa ve uzun belgelerde çeşitli sorunlar çıkmakta, uzun belgelerde terimler daha çok yer aldığı için terimler uzun belgeler daha yüksek ağırlıklara sahip olacaklardır. Bu sebepten dolayı terim sıklığı düzenlenmelidir. Terim sıklığı, Denklem (3.5) ile hesaplanmaktadır.

$$ts_i = \frac{n_i}{\sum_k n_k} \quad (3.5)$$

Belge sıklığı ise bir terimin derlemde en az bir kere yer aldığı belge sayısı ile ilişkilidir. Ayrıca bir terimin önemini belirten ağırlığını hesaplarken belge sıklığı ters orantılı olarak kullanıldığı için belge sıklığı yerine ters belge sıklığı hesaplanır. Düzenlenmiş belge sıklığı Denklem 3.6 ile hesaplanmaktadır.

$$tbs_i = \log\left(\frac{N}{n_i}\right) \quad (3.6)$$

Terim ağırlığı, terim sıklığı ile doğru, belge sıklığı ile ters orantılıdır. Çünkü, bir terim bir belgede ne kadar çok ve bir derlemde ne kadar az belgede yer alıyorsa geçtiği belgeyi temsil etmekte o kadar başarılıdır. Terim ağırlığı, terim sıklığı ve ters belge sıklığının çarpımı olarak Denklem (3.7)'deki gibi hesaplanabilir.

$$w_i = bs_i \cdot tbs_i \quad (3.7)$$

Terim Seçimi

Terim seçimi, belgeyi temsil edecek terimlerin belirlenmesi ve edemeyecek geniş anlamlı sözcüklerin elenmesidir. Seçilecek terim, yer aldığı belgede farklı durumlarda bulunabilir. Bu yüzden terim seçiminden önce belgede yer alan sözcükler gövdeleme gibi doğal dil işleme işlemlerinden geçirilmelidir. Örneğin, *okul* ile ilgili bir belgede *okul* sözcüğü *okul*, *okula*, *okuldan* vb gibi yer alabilir.

Ayrıca, terim seçiminde dar anlamlı sözcüklerin tercih edilmesi de önemlidir. Türkçede hangi konuda yazılmış olursa olsun bütün belgelerde bir, ve, bu, eğer gibi sözcükler yoğun şekilde yer almaktadır. Bu tür sözcükler terim seçiminde dikkate alınma ve göz önüne alınmamalıdır.

3.2 Doğal Dil İşleme

Doğal Dil İşleme (DDİ), insanlar tarafından kullanılan sözlü ve yazılı doğal dilleri inceleyen, anlaşılması ve çözümlenmesi üzerine çalışan bir bilim dalıdır. DDİ, yapay zeka ve dilbilimi alanlarının bir alt dalı olmakla birlikte felsefe ve psikoloji gibi diğer bilim dallarından da faydalanan bir bilimler arası disiplindir [3]. DDİ tanımlanırken kullanılan *doğal* sözü, insanlar tarafından konuşma ve yazıda kullanılan dili, matematiksel yazım ya da bilgisayar dillerinden ayırmak için kullanılmaktadır [4].

Doğal Dil İşleme alanındaki çalışmaların önemi, doğal dilin insan hayatındaki önemi ile paralellik gösterir. İnsan, canlılar dünyasındaki en akıllı varlıktır. İnsan aklının karmaşıklığı genel anlamda kullanılan doğal dilin karmaşıklığına da yansımaktadır. Doğal Dil İşleme çalışmalarında karşılaşılan en önemli sorun doğal dilin karmaşıklığıdır.

Doğal Dil İşleme alanındaki çalışmaların genel amacı insan doğal dillerini daha iyi anlamak ve çözümlmek, doğal dilleri anlama, yorumlama ve birbiri arasında çevirme yetisine sahip bilgisayar sistemleri tasarlamaktır. Bu çalışmalar ilerledikçe insan ve bilgisayar sistemleri arasındaki iletişimin bugünkü sınırların çok dışına

çıkacağı, fare, tuş takımı gibi fiziksel, programlama dilleri gibi mantıksal arayüzlerin ortadan kalkacağı düşünülmektedir.

Doğal Dil İşleme'nin başlıca uygulama alanları şu şekilde özetlenebilir:

- Bilgi Çıkarımı
- Metin Madenciliği
- Soru – Cevap Sistemleri
- Yazıdan Konuşma ve Konuşmadan Yazı
- Makine Çevirisi
- Belge Anlama ve Üretme

Doğal Dil İşleme, doğal dili beş ana seviyede değerlendirir: sesbilim, biçimbilim, sözdizimbilim, anlambilim, kullanımbilim [10].

3.2.1 Sesbilim

Sesbilimi; bir dildeki harflerin seslerini, boğumlanma noktalarını, boğumlanma özellikleri ve bunların dilde kullanımlarını inceler. Günlük hayatta insanların anlamak için kullandıkları doğal diller, yazılı olduğu kadar, sesli olarak da kullanıldığından dolayı, doğal dil sesbilimi ile bütünleşiktir. Bu yüzden dil işlenirken kesinlikle ses özellikleri göz ardı edilmemelidir.

Türkçenin ses bilgisi kuralları seslere ilişkin organların hareketlerine uygun yapıdadır, herhangi bir zorlanma durumunda, sözcük içindeki sesler değişime uğrar. Bu da Türkçenin seslerinin kuralların izlemeyi kolaylaştırmaktadır. Ses değişimleri, ünlü – ünsüz uyumları, yabancı sözcüklerin uyarlanması kullanılan kurallar Türkçenin önemli ses bilgisi kurallarından birkaçıdır. Türkçenin ses kuralları, biçimbirim özelliklerini belirlemesi bakımından da son derece önemlidir.

Sesbilim özellikle yazıda konuşma, ve konuşmadan yazı çalışma alanlarında önem kazanmıştır.

3.2.2 Biçimbilim

Biçimbilim bir dilin sözcük yapılarını, türeme yollarını, çekim biçimlerini inceleyen dilbilgisi koludur. Biçimbilim sözcüklerin yapısı ile ilgilenir, sözcüklerin köklerinin belirlenmesi, eklerinin ayrıştırılması, incelenmesi ve görevlerinin belirlenmesi üzerine çalışır.

Biçimbilim, Çince ve Vietnamca gibi yapı bakımından yalınlayan dillerde neredeyse hiç, İngilizce, Almanca ve Fransızca gibi bükümlü dillerde ise oldukça sınırlıdır. Türkçe sonradan eklemeli bir dil olması sebebiyle biçimbilim çalışmaları için düzenli fakat bir o kadar da zor bir dildir. Türkçenin biçimbilimsel karmaşıklığı biçimbirimsel çeşitliliğe de arttırmaktadır. Örneğin Türkçede *göz* sözcüğünden türetilmiş yaklaşık 150 sözcük bulunmaktadır [11]. Aynı kökten türemiş bu sözcüklerin de çeşitli çekim ekleri olarak oluşturacakları sözcükler düşünüldüğünde Türkçenin biçimbirimsel çeşitliliği anlaşılabilir.

Türkçede biçimbirimsel türetme yapım ekleri ve çekim ekleri ile yapılır. Çekim ekleri kalıplaşmış kullanımlar dışında yeni bir sözcük türetmeyen, sözcükler arasında durum, iyelik, çokluk, kip, zaman, şahıs, gibi ilgiler kuran eklerdir. Yapım ekleri ise, eklendikleri sözcükten yeni bir sözcük türeten eklerdir [4]. Yapım ekleri eklenen kök veya gövdelerin anlamları değişir.

Doğal Dil İşleme'nin biçimbilim çerçevesinde ilgilendiği en önemli konu *gövdeleme* işlemidir. Gövdeleme işleminde amaç temel olarak sözcükten eklerin atılarak köke veya gövdeye ulaşmak ve atılan ek ve ek dizisini incelemektir [12]. Gövdeleme işleminin dayandığı mantık aynı kökten türemiş olan sözcüklerin aynı anlamda olduğu ya da en azından aralarında bir anlam ilişkisi olduğudur. Türkçe için gövdeleme işlemi yapılırken sadece çekim ekleri atılır. Çünkü aynı köke farklı yapım ekleri eklenerek oluşturulmuş sözcükler arasında bir anlam ilişkisi bulunsa da farklı anlamdadırlar.

Farklı diller için örnek kelimeler ve bu kelimeler için gövdeleme sonuçları Tablo 3.1'de verilmiştir.

Tablo 3.1: Farklı diller için gövdeleme sonuçları

	Türkçe	İngilizce	Almanca
Örnek Sözcükler	kitap kitabı kitaptan	connected connecting connection	lebensversicherungsgesellschaft
Kök	kitap	connect	leben, versicherung, gesellschaft

Gövdeleme yöntemleri iki ana grupta toplanabilir: biçimbirimsel analize dayalı yöntemler ve dayanmayan yöntemler. Biçimbirimsel analize dayalı yöntemlerde biçimbirimsel çözümleme yapıldıktan sonra en uygun kök çeşitli yöntemlerle seçilir. Biçimbirimsel analiz özellikle yapısı itibari ile eklemeli dillerde oldukça karmaşıktır. Ayrıca biçimbirimsel analiz için biçimbirim bilgisinin ve analiz sırasında kullanılacak olan sözlüğün bilgisayar ortamında hazırlanması da zaman ve emek açısından pahalı bir işlemdir [13]. Biçimbirimsel analize dayanmayan yöntemler ise şu şekildedir:

- *Tablo Arama Yöntemleri* genel olarak kökün sözlükte aranması ile gerçekleştirilen yöntemlerdir. Kolay bir yöntem olmasına karşın başarısı kullanılan sözlüğün başarısına bağlıdır. Herhangi bir dil için tüm kökleri içeren bir sözlük oluşturmak mümkün olmadığı için bu yöntem pratik değildir.
- *Halef Çeşitlilik Yöntemleri* kökün sınırlarının bulunmasına dayanır ve geniş bir yapısal dilbilgisi analizi gerektirir.
- *N-Grams Yöntemleri* sözcükte diagram ve triagramların bulunmasına dayanır.
- *Ek Atma Yöntemlerinde* oluşturulmuş kurallar ile eklerin sözcükten atılarak köke ulaşılmaya çalışılır. Bu yöntemler diğer gövdeleme yöntemlerine kıyasla daha sezgisel, basit ve kolay uygulanabilir [2].

Gövdeleme yöntemleri, Bilgi Erişimi konusunda da verimliliği arttırmak için kullanılmaktadır. Sözcükler genellikle yalın hal yerine daha çok cümle içerisindeki görevlerine göre ek almış halde bulunurlar. Örneğin *çocuk* üzerine yazılmış bir metinde çok sayıda çocuk sözcüğü ile beraber *çocuklar*, *çocuğa*, *çocuktan* vb gibi

sözcükler de yer almaktadır. Eğer gövdeleme yöntemleri ile desteklenmeyen bir Bilgi Erişimi sistemi kullanılırsa *çocuklar*, *çocuğa*, *çocuktan* vb gibi çocuk sözcüğünden çekim ekleri ile türetilmiş sözcükler ile *çocuk* sözcüğü arasındaki ilişki bulunamaz. Gövdeleme yöntemleri Bilgi Erişimi'nde kullanılan verimlilik ölçütlerinden keskinlik üzerinde bir etki yaratmadan anımsamayı arttırmaktadır [12].

3.2.3 Sözdizim Bilimi

Sözdizim bilimi, sözcüklerin cümle içerisindeki yerleri, görevleri ve cümle içerisindeki diğer sözcüklerle anlam ilişkilerini inceler.

Sözdizim seviyesinde biçimbilimsel çözümleyicide ayrıştırılan sözcükler kullanılarak cümledeki öğelerin (isim, sıfat, zarf...) sözdizimsel formülleri oluşturulabilir. Türkçe sözdizimsel açıdan oldukça esnek bir dil olduğundan oluşturulabilecek formüllerin sayısı çok fazladır. Bu formüllerle sözdizimin anlamlı olup olmadığının ölçülebilir.

Türkçede cümleler en genel şekliyle özne, nesne ve yüklem bileşenlerinden oluşur. Cümleye eklenmek istenen anlamlar arttıkça cümleler, özne, yer tamlayıcısı, zarf tamlayıcısı, nesne ve yüklem gibi bileşenleri içerir. Ayrıca cümlenin anlamını kuvvetlendiren cümle dışı bileşenler de (bağlaç, edat, vb) cümlede bulunabilir. Bunlara örnek olarak *ile*, *için*, *ama*, *çünkü* sözcükleri de verilebilir. Türkçede özne ile yüklem cümlenin temel bileşenleridir ve genelde tüm cümlelerde yer alırlar. Yer tamlayıcısı, zarf tamlayıcısı, nesne gibi bileşenler bazı cümlelerde yer almayabilirler veya bazı cümlelerde sadece biri, bazılarında sadece ikisi bulunabilir. Bu bileşenlerin cümle içindeki sıralanışları da değişebilir.

Sözdizimsel seviyede de Türkçe için çok sayıda çeşitlilik oluşmaktadır. Örnek olarak “*bilgisayar mühendisliği*” ve “*inşaat mühendisliğinde kullanılan bilgisayar programları*” sözcük grupları verilebilir. Bu iki sözcük grubu da *bilgisayar* ve *mühendislik* sözcüklerini içermesine rağmen anlamları bakımından oldukça farklıdırlar.

Bilgi Erişimi sistemlerinde sözdizimsel çeşitlilikleri ele almak için iki ana yöntem vardır: sorgulara tamlamaların eklemesi ve dizinleme için sözdizimsel yapıların kullanılması. Bilgi Erişimi'nde kullanılan tamlama terimi birbiriyle ilişkili sözcük öbeği olarak tanımlanabilir. Tamlamalardaki sözcükler arasındaki ilişki istatistiksel

ya da sözdizimsel olabilir. *İstatistiksel tamlamalar* birbirleriyle aynı sıklıkta ve yüksek yakınlıkta olan sözcüklerden oluşur. *Sözdizimsel tamlamalar* ise dilin sözdizim seviyesinde oluşan sözcük öbekleridir. Türkçede sözdizimsel tamlamalara örnek olarak isim tamlamaları ve sıfat tamlamaları verilebilir. İstatistiksel tamlamalar sözcük sıklıkları ve uzaklıkları ile kolayca bulunabilirken sözdizimsel tamlamalar karmaşık doğal dil işleme yöntemleri ile bulunabilir. Tamlamaların sorgulara eklemesi ve dizinleme kullanılmaları ile keskinlikte bir artış sağlanırken, sıralamada daha doğru sonuçlar vermektedir [12].

3.2.4 Sözcüksel bilim

Sözcüksel bilim, sözcükler ve onların kullanımı üzerine çalışır. Türkçede diğer dilbilgisi seviyelerinde olduğu gibi bu seviyede de çeşitlilik oldukça yüksektir. Sözcüksel çeşitliliğe örnek olarak yazılışları ayrı, anlamları aynı olan eşanlamlı sözcükler verilebilir. Örneğin, *yüz – surat, hikaye – öykü* .

Sözcüksel çeşitliliğin Bilgi Erişimi sistemlerinde ele alınması oldukça önemlidir. Bilgi Erişimi sistemleri kullanıcının girmiş olduğu sorgu ile derlemde bulunan metinlerin karşılaştırması olduğuna göre eşanlamlı sözcüklerden bir tanesi sorguda diğeri ise derlemde bulunan bir belgede bulunması durumunda anlamsal olarak sorgu ve metin arasında bir ilişki bulunmasına rağmen bu sistem tarafından bulunamayacaktır. Örneğin *hikaye* sözcüğü içeren bir sorgu ile *öykü* üzerine yazılmış bir metnin eşleştirilmesi mümkün olmayacaktır.

Sözcüksel çeşitlilikler iki ana yöntem ile giderilebilirler: *sorgu genişletilmesi* ve *sorgu ve metindeki anahtar sözcüklerin kavramsal uzaklıklarının ölçümü*. Sorgu genişletilmesi yönteminde sorguya sorgudaki sözcüklerin eşanlamlıları eklenir. Diğer yöntemde ise sorgudaki sözcüklerle tür, çeşitlilik, karşıtlık vb gibi anlam ilişkilerinde olan sözcükler sorguya ilave edilir [12]. Her iki yöntem için de kavramsal sözcük veritabanları WordNet kullanılır [14].

Bilgi Erişimi sistemlerinde WordNet veritabanları kullanılarak yapılan sorgu genişletme yöntemleri sözcüksel çeşitlilikleri ele alarak anımsamada iyileştirmeler sağlamaktadır.

3.2.5 Anlambilim

Anlambilim, dilbilgisinin sözcük ve cümlelerin anlamları ile ilgilenen seviyesidir. Bilgisayar bilimleri açısından anlambilim üzerinde çalışılması en zor dilbilgisi daldır.

Anlambilim seviyesinde oluşan çeşitliliklere örnek olarak yazılışları aynı, anlamları farklı sözcükler verilebilir. Örneğin, yüz sözcüğü hem doksan dokuzdan sonra gelen tamsayı hem surat hem de yüzmek eyleminin emir kipi anlamındadır. Doğal Dil İşleme'nin anlambilimsel seviyede uğraştığı en önemli konu anlam belirsizliğinin giderilmesidir. Anlam belirsizliğinin giderilmesi aynı sözcüğün farklı anlamlardaki kullanımlarının birbirinden ayırt edilmesidir. Anlam belirsizliğinin giderilmesi Bilgi Erişimi sistemlerinde oldukça önemlidir. Bilgi Erişimi sistemlerinde oluşan sonuçlardaki yanlışlıkların önemli bir kısmı anlam belirsizliğinin giderilmemesinden kaynaklanmaktadır [12]. Doğal Dil İşleme alanında anlambilim seviyesindeki çalışmalar için de WordNet kavramsal sözcük veritabanları kullanılmaktadır.

Ayrıca anlambilim seviyesinde, doğal dillerde kullanılan cümleler diller arası bir biçime çevrilerek cümlelerin diğer dillere ya da makine diline çevrilebilmesi mümkün olabilir [15].

3.3 Önceki Çalışmalar

3.3.1 Türkçe Dışındaki Dillerde Yapılmış Çalışmalar

İngilizce için biçimbilim seviyesinde yapılmış en önemli ve en çok bilinen çalışma Porter gövdeleme algoritmasıdır. Bu algoritma Martin Porter tarafından geliştirilmiş ve 1980 yılında Program dergisinde yayınlanmıştır. Porter algoritması bir dizi ek atma kurallarından oluşmaktadır. Bu kurallar sözcüklere uygulanarak sözcüğün köküne ulaşmaya çalışılmaktadır [2,16]. Porter gövdeleme algoritmasında İngilizce için yaklaşık 1200 kural tanımlanmıştır. Porter gövdeleme algoritması ile ilgili örnek kurallar Tablo 3.2'deki gibidir.

Tablo 3.2: Porter gövdeleme algoritması örnek kurallar ve örnekler

Örnek Kural	Örnek
$s \rightarrow \varepsilon$	cats $\rightarrow$ cat
sses $\rightarrow$ ss	kisses $\rightarrow$ kiss
ational $\rightarrow$ ate	relational $\rightarrow$ relate
ing $\rightarrow \varepsilon$	motoring $\rightarrow$ motor

Porter gövdeleme algoritması yapı olarak basit ve kolay uygulanabilir olmasına karşın sonuçları oldukça iyidir. Krovetz yaptığı deneylerde Porter gövdeleme algoritmasının eksik ve hatalı çalıştığı noktaları ortaya çıkarmış ve bunları yöntem içerisinde gerçek bir sözlük kullanılmamasına bağlamıştır. Bu deneyler sonucunda bulunan hatalardan örnekler Tablo 3.3'deki gibidir [17].

Tablo 3.3: Porter gövdeleme algoritmasının bulduğu yanlış sonuçlar

Yanlış Bulunan Sonuçlar		Bulunamayan Sonuçlar	
organization	organ	European	Europe
doing	do	analysis	analyzes
policy	police	noise	noisy
university	universe	urgency	urgent

Sözcüksel ve anlamsal seviyede yapılmış en önemli çalışma WordNet kavramsal sözcük veritabanlarıdır. WordNet 1985 yılında Princeton Üniversitesi'nde geliştirilmeye başlanan, İngilizce için anlamsal bir sözlüktür. Proje hem çeşitli hükümetler hem de resmi kurumlar tarafından desteklenmektedir [14]. WordNet projesi başlangıcından itibaren ücretsiz olarak dağıtılmaktadır.

WordNet projesi anlamsal bir sözlük olarak İngilizce sözcükleri eş anlamlılar kümelerinde (synsets) sınıflandırır ve sözcüklerin kısa, genel tanımlamalarını yaparak bu eş anlamlılar kümeleri arasındaki çeşitli anlamsal ilişkileri oluşturur. Burada amaç iki farklı işlevi yerine getirmektir: İlki, sözcüklerin tanımlarının verildiği sözlüğü, sözcüklerden kavramları, kavramların özelliklerini ve kavramlar arasındaki ilişkileri oluşturmak iken, ikincisi yapay zeka uygulamalarını ve özellikle

otomatik metin analizini desteklemektir. WordNet projesinin 2.1 sürümünde isim, eylem, sıfat, belirteç olarak sınıflandırılmış toplam 155.327 farklı ögenin üst kavram (hyperonym), alt kavram (hyponym), eşanlamlılık (synonym), zıt anlamlılık (antonym), parçanın (bölümün-üyenin) bütünü (holonomy) ve bütünün parçası (bölümü –üyesi) gibi anlamsal ilişkiler yer almaktadır [18].

3.3.2 Türkçe İçin Yapılmış Çalışmalar

Doğal Dil İşleme'nin biçimbilim seviyesindeki önemli çalışma alanlarından gövdeleme için yapılmış olan yöntemler dört ana grupta incelenebilir: *tablo arama yöntemleri*, *halef çeşitlilik yöntemleri*, *n-gram yöntemleri*, *ek atma yöntemleri* [2]. Bu dört farklı yöntem de İngilizce gibi analitik diller ya da Almanca ve Fransızca gibi Hint – Avrupa dilleri için geliştirilmiştir. Bu yöntemler yüksek düzeyde biçimbirimsel zenginlik içeren sondan eklemeli Türk dilleri ailesi için yetersiz kalmaktadır [19]. Türkçe için yapılmış gövdeleme çalışmaları önceleri istatistik ve olasılığa dayalı yöntemler olsa da daha sonraları biçimbirimsel çözümlemeye önem verilmiştir.

Türkçe üzerindeki Bilgi Erişimi ve Doğal Dil İşleme araştırmaları 1980'li yıllara dayanmaktadır. Yayımlanan ilk çalışma 1981 tarihli A. Köksal'ın çalışmasıdır [20]. Köksal çalışmasında başlık, bölüm başlıkları, anahtar kelimeler ve özet içeren 570 belge ve bu belgeler üzerinde kullanılmak üzere hazırladığı 12 sorgu ile farklı dizinleme ve belge eşleştirme yaklaşımı kullanarak verimliliği ölçmüştür. Bu çalışmada dilbilimsel temellere dayanmayan bir gövdeleyici kullanılmıştır. Gövdeleme için sözcüğün sabit uzunluktaki bir kısmı alınıp kök olarak kabul edilmiştir. Köksal, yaptığı denemeler sonucunda sözcüklerin ilk 5 harfini alarak en iyi sonucun elde edildiğini gözlemlemiştir. Köksal'ın gövdeleme için kullandığı yöntem dilbilimsel yöntemlere dayanmasa da sonuçları itibari ile gövdelemenin bilgi erişiminin başarımını arttırdığını göstermektedir.

Köksal'ın çalışmasından sonra bir diğer önemli çalışma da Solak ve Can'ın 1994 yılında gerçekleştirmiş oldukları çalışmadır [21]. Bu çalışmada Solak ve Can 533 haber makalesi ve 71 sorgu kullanmışlardır. Gövdeleyici olarak biçimbirimsel çözümlemeye dayanan bir yaklaşım kullanmışlardır. Yöntem olarak sözcük, öncelikle sözlükte aranır, bulunamaması durumunda sözcüğün sağından harf atarak

sözlükte aranmaya devam edilir. Sözcüğün sözlükte bulunması durumunda biçimbirimsel çözümleme yapılır. Algoritmanın çalışması sonucunda bir sözcük için birden fazla kök bulunma olasılığı vardır. Solak ve Can yaptıkları çalışmalar sonucunda bir Türkçe sözcük karşılığında algoritmalarının ortalama 1.22 kök bulunduğunu belirtip kullandıkları biçimbirimsel gövdeleyicinin başarımı gövdeleme yapılmamasına göre arttırdığını ifade etmişlerdir.

Türkçe için gövdeleyici çalışmalarında Kut ve çalışma arkadaşları 1995 yılında yaygınca kullanılan sözlük tabanlı “En Uzun Eşleşme” algoritmasını kullanmışlardır. Bu sözlükte Türkçe kökler ve bu köklerin değişik biçimleri yer almaktadır. Bu algorithmada işlenen sözcük sağından harf atılarak sözlükte aranmakta ve algorithmaya da adını verdiği gibi en uzun eşleşme sağlanan adımda sözcüğün kökü bulunmuş kabul edilmektedir [22]. Bu yöntem uygulama açısından basit olmasına rağmen bazı sözcükler için yanlış kökler bulabilmektedir.

Her ne kadar doğrudan amacı gövdeleme olmasa da sonuçları itibari ile Oflazer’in Türkçenin biçimbirimsel analizini yaptığı çalışması da oldukça önemlidir. Bu çalışmanın sonuçlarından faydalanarak verilen sözcüğün biçimbirimsel analizi yapılarak olası tüm kökleri ve yapısal analizi yapılabilmektedir [23].

Eklemeli dillerde gövdeleme işleminin gerekliliğini görmek için Hint – Avrupa dil ailesinden İngilizce için gerçekleştirilen Porter Algoritması ile sondan eklemeli Türkçe için gerçekleştirilen Oflazer’in biçimbirimsel analiz sonuçları karşılaştırılabilir. Bu karşılaştırma için Porter’ın gövdeleyici 567.564 sözcük içeren derlem üzerinde, Oflazer’in biçimbirimsel çözümleyicisi 376.187 sözcük içeren bir derlem üzerinde çalıştırılmıştır [19]. Her iki çalışmanın sonuçları Tablo 3.4’de yer almaktadır.

Tablo 3.4: Porter gövdeleyicisi ve Oflazer’in biçimbirimsel çözümleyicisinin sonuçlarının karşılaştırılması

Derlem	Sözcük Sayısı	Farklı Sözcük Sayısı	Farklı Kök Sayısı	Sıkıştırma (%)
Türkçe	376.187	41.370	6.363	84,6
İngilizce	567.574	18.384	11.671	36,4

İki çalışmanın sonuçlarını gösteren Tablo 3.4’de görüldüğü gibi Türkçe derlem, İngilizce derleme göre daha az sözcük içermesine rağmen daha fazla sayıda farklı sözcük içermektedir. İki derlem üzerinde yapılan sonuçlar incelendiğinde farklı kök sayısının farklı sözcük sayısına oranının birden çıkarılması ile hesaplanan sıkıştırma oranı arasındaki büyük fark dikkat çekmektedir. Bu da Türkçe gibi sondan eklemeli dillerde daha az kök olmasına rağmen İngilizce gibi Hint – Avrupa dillerine göre daha fazla farklı kelime türetebilme yeteneğini sadece yapısal olarak değil istatistiksel olarak da ortaya koymuştur. Bu yetenek ayrıca Türkçe üzerinde yapılan Bilgi Erişimi ve Doğal Dil İşleme çalışmalarında gövdeleme yöntemlerinin önemini bir kez daha belirtmektedir.

Doğal Dil İşleme konusunda sözcüksel ve anlamsal seviyede yapılmış önemli bir çalışma da BalkaNet projesi içerisinde yer alan Türkçe Kavramsal Sözcük Projesi’dir. BalkaNet projesi Türkçe, Yunanca, Bulgarca, Çekçe, Romence ve Sırpça gibi Balkan dilleri için her bir dilin kendi kavramsal sözlüklerinin birleştirilmesi ile oluşturulacak çok dilli bir sözcüksel veritabanının tasarımı ve geliştirilmesini hedefleyen bir projedir [24]. BalkaNet projesinin Türkçe ayağını Sabancı Üniversitesi yürütmektedir. BalkaNet projesinin amaçları arasında ilerleyen aşamalarda çalışmaların Avrupa dillerini kapsayan EuroWordNet projesi ile birleştirilmesi de bulunmaktadır [25].

BalkaNet projesinde kavramsal sözlükteki tüm bilgiler elektronik bir veritabanında tutulur. Kavramsal sözlük, normal bir sözlükte olduğu gibi, dilin sözcüklerini, tamlamalarını ve bunların tanımlarını içerir. Normal bir sözlükten farklı olarak ise, eşanlamlı sözcükler, karşıt anlamlı sözcükler, alt-üst ilişkileri ve diğer anlam ilişkileri de yer alır. Temel yapıyı alt – üst ilişkileri oluşturmaktadır ve sözcükler değil kavramlar birbirine bağlıdır. Eşanlamlı sözcükler kümesi ve kavramsal sözlükleri birbirine bağlayan Dilden Bağımsız Dizin (DBD-Interlingual Index ILI) ile ilişkilendirilmesi ile proje kapsamında yer alan diller arasında makine çevirisi mümkün olabilecektir.

BalkaNet projesinin uygulama alanları arasında arama motorlarında daha iyi sonuçlar elde edilmesi, bir dilde arama yaparak tüm dillerdeki belgelere ulaşmak, otomatik olarak iki dilli veya çok dilli sözcükler oluşturmak ve kavramların tanımlarından yola çıkarak arama yapılabilmesidir. Proje şu anki durumu itibari ile 20.420 sözcük/isim

öbeği, 15.622 tekrarlanmayan sözcük/isim öbeği, 14.795 eşküme ve 23.540 ilişki içermektedir [25].

3.3.3 Zemberek Doğal Dil İşleme Kütüphanesi

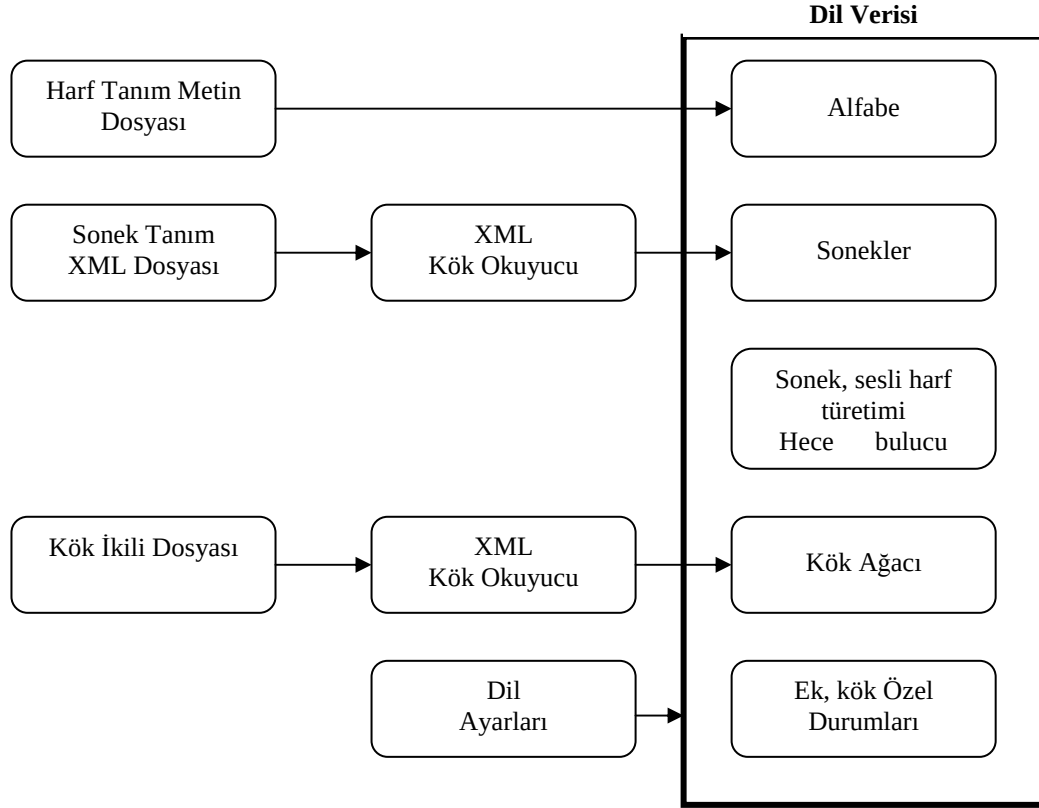
Zemberek Doğal Dil İşleme kütüphanesi, Ahmet Afşin Akın ve Mehmet Dünder Akın tarafından geliştirilmiş esnek, açık kaynak kodlu ve platform bağımsız bir doğal dil işleme kütüphanesidir. Doğal Dil İşleme konusunda yapılan çalışmalar genellikle Hint – Avrupa dilleri için yapılmaktadır. Zemberek projesi Türkçe üzerine yapılmış ilk açık kaynak kodlu çalışmadır. Zemberek projesi başlangıçta sadece Türkçe için tasarlanmış olsa da 2.0 sürümünde projenin altyapısı dilden bağımsız hale getirilip Türkiye Türkçesi dışında Türkmence, Azerice ve Tatarca'yı destekler hale getirilmiştir [26].

Zemberek kütüphanesi şu doğal dil işleme işlemlerini desteklemektedir:

1. Yazım denetleme
2. Biçimbirimsel analiz
3. Gövdeleme
4. Sözcük üretimi
5. Sözcük önerme
6. Sözcük heceleme
7. ASCII karakterlerden Türkçe harflere ve tersi

Zemberek kütüphanesi iki ana parçadan oluşmaktadır: dil yapısı bilgileri ve Doğal Dil İşleme işlemleri. Çekirdek kütüphane Doğal Dil İşleme algoritmaları ve dillerin gerçekleşmesi için gerekli araçları barındırır. Bu çekirdek kütüphane Türk dil ailesi için tasarlanmış olsa da içerisinde dillerin yapısı ile ilgili bir bilgi bulunmaz. Bu özelliği Zemberek kütüphanesine yeni dillerin eklenmesi konusunda büyük bir esneklik sağlar. Kütüphaneye yeni bir dil eklenmesi için gerekli olan bilgiler dilin yapısı ve dilbilgisidir. Zemberek kütüphanesi bünyesinde barındırdığı dillere ait genel ve dilbilgisi bilgilerini kütüphane dışında metin dosyalarında saklar. Zemberek

kütüphanesinde bir dilin çalışması için gerekli olan dosyalar Şekil 3.3’de gösterilmiştir.



Şekil 3.3: Zemberek kütüphanesi dil yapısı elemanları

Harf ve Alfabe

Her dil kendi harf ve alfabe bilgisini ayrı metin dosyalarında saklar. Bu dosyada ayrıca Türk dil ailesi dilleri için önemli olan ünlü – ünsüz ayrımı, ünlü ve ünsüz harfler için özellikler (sert – yumuşak, kalın – ince, düz – yuvarlak) de yer almaktadır.

Ekler

Zemberek kütüphanesi Türkçe için tasarlandığı için ek sözlüğü üzerinde özellikle durulmuştur. Şuan için sadece son ekleri desteklemektedir, önek desteği yoktur. Ek sözlüğü XML biçimde koddan bağımsız olarak dosyada tutulmaktadır. Bu dosyada ekler, eklerin hangi diğer eklerden önce ve sonra gelebileceği, eklere ait üretim

kuralları ve bu kurallara ait özel durumlar yer alır [26]. İsimden isim üreten –lık durum eki için ek dosyasındaki bilgi Şekil 3.4’deki gibidir:

```
<ek ad="ISIM_DURUM_LIK" uretim="Iık">
  <ozel-durum ad="SON_HARF_YUMUSAMA" uretim="Iığ" />
  <ardisil-ekler>
    <kume>ISIM_HAL</kume>
    <kume>ISIM_SAHİPLİK</kume>
    <kume>İMEK_ZAMAN</kume>
    <aek>ISIM_TANIMLAMA_DIR</aek>
    <aek>ISIM_COGUL_LER</aek>
    <aek>ISIM_BIRLIKTelik_LE</aek>
  </ardisil-ekler>
</ek>
```

Şekil 3.4: Zemberek ek sözlüğünden bir örnek

Ek sözlüğünde her ek için bir üretim kuralı yer almaktadır. Bu üretim kurallarında geçen harflerin anlamları şu şekildedir:

1. Küçük harfler üretim sırasında hiçbir değişikliğe uğramaz.
2. Büyük sesli harfler (A, I) ekin eklendiği sözcüğün son hecesindeki sesliye göre değişimi temsil eder.
3. Ek kuralının başındaki “+” işareti sözcüğün son harfi sesli ise kaynaştırma harfi ekleneceğini, “>” işareti ise sözcüğün son harfi sert sessiz ise ekin ilk sessizinin de sertleşeceğini gösterir.

Ekler ve üretim kuralları ile ilgili örnekler Tablo 3.5’ deki gibidir.

Tablo 3.5: Zemberek kütüphanesi üretim kuralları örnekleri

Ek	Üretim Kuralı	Örnek
İsim Çoğul Eki (-lar)	lAr	okull ar , ev ler
İsim Andırma Eki (-sı)	+msI	odun su , acı msı
İsim Küçültme Eki (-cık)	>cIk	kedi cık , kurt çuk

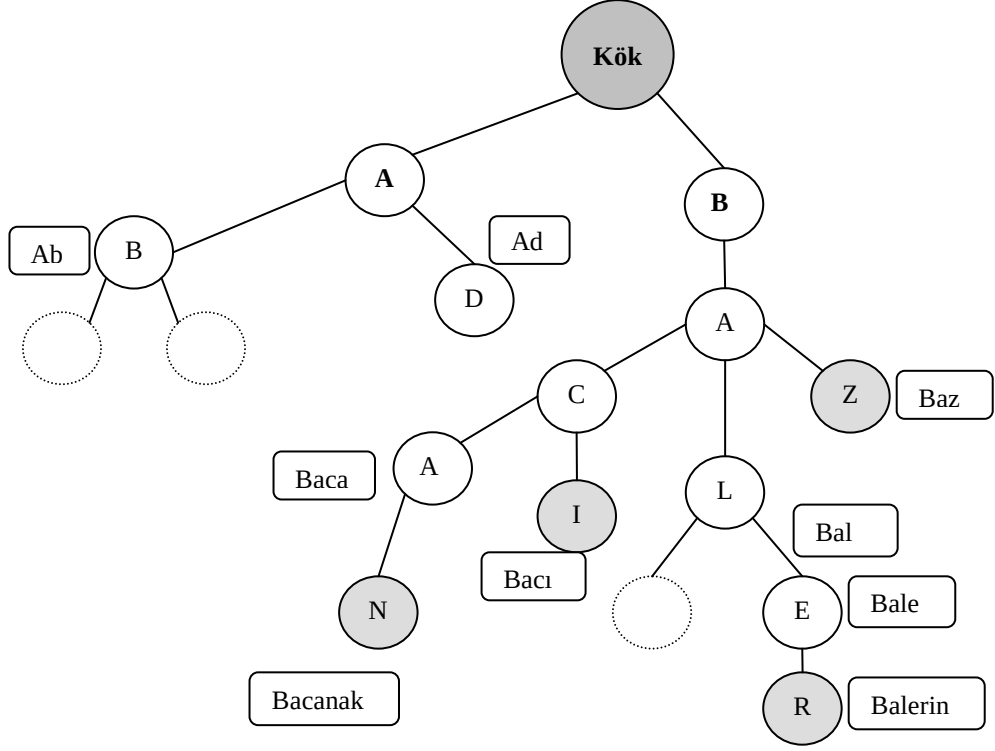
Kök Sözlüğü

Kök sözlüğü, dildeki yer alan anlamlı ek içermeyen yani kök halinde bulunan kelimeleri barındırır. Bir kök sözlüğünde dilde yer alan tüm kökleri toplamak oldukça güçtür. Fakat zengin kök sözlüğü kullanılan biçimbirimsel çözümlemelerin daha başarılı olacağı kesindir. Zemberek kütüphanesi, barındırdığı her dil için ayrı birer kök sözlüğünü metin biçimde bir dosyada tutar. Sözlükte ayrıca kökün kendisi dışında kökün sözcük türü ve varsa barındırdığı özel durumlar yer alır. Zemberek kütüphanesinin Azerice için kullandığı kök sözlüğünden örnek kayıtlar Şekil 3.5’de verilmiştir.

su	AD		balıq	AD
sağlıq	AD	YUM	bir	RA
al	EY		iki	RA
gel	EY		dörd	RA
tuz	AD		dünen	ZAMAN
istiot	AD		sabah	ZAMAN

Şekil 3.5: Zemberek kütüphanesi kök sözlüğü örneği

Zemberek kütüphanesi, işlem hızını arttırabilmek için kök sözlüğünü ikili dosyalarda saklar. Kütüphane başlatıldığı zaman bu dosyayı okur ve Doğrudan Düz Sözcük Çizge’si (DAWG) oluşturur [26]. Bu ağaç yapısı sayesinde yapılan çözümlemelerde hızlı arama ve sözlüğü kolay genişletilebilme yeteneği sağlar. Şekil 3.6’de kök sözlüğünden oluşturulmuş kök ağacının örnek temsili yer almaktadır.



Şekil 3.6: Zemberek kütüphanesi kök ağacı örneği

Türkçede ünsüz sertleşmesi veya yumuşaması gibi kurallardan dolayı bazı durumlarda aranan kök sözlüğünden bulunamayabilir. Bu durumları karşılamak için Zemberek kütüphanesi her yeni kök ağaca eklendikten sonra kökün dilbilgisi kurallarına göre değişik durumlarını da sözlüğe ekler. Bu şekilde kökün değişikliğe uğradığı durumlarda da kök sözlükte bulunabilir. Örneğin *ağac* sözcüğü kök sözlüğü kaynak dosyasında yer alırken sözcüğün ünlü bir ek aldığına değişikliğe uğramış hali *agac* yer almamaktadır ama oluşturulan ağaç yapısına kendiliğinden eklenmektedir.

Zemberek kütüphanesinde kök sözlüğü içeren ağaç yapısı ayrıca sözlük üzerindeki aramalarda da kullanılır ve bu yapı sayesinde çeşitli kök seçiciler kolaylıkla geliştirilebilir. Kütüphanenin şu anki yapısında üç tane kök seçici bulunmaktadır. İlk kök seçici; verilen bir sözcük için normal tam bir kök seçimi yapar. Örneğin; “elmaslar” sözcüğü için “el”, “elma”, “elmas” köklerini bulur. İkinci kök seçici bir harf dizisi benzerlik algoritması kullanıp belirli özür değeri içinde kalan kök adaylarını seçer. Doğal olarak ilk yöntemle göre daha fazla kök bulur. Üçüncü seçici ise ASCII kodlamasında yer almayan harfler için belirli özür değerleri arasında kalan kökleri seçer.

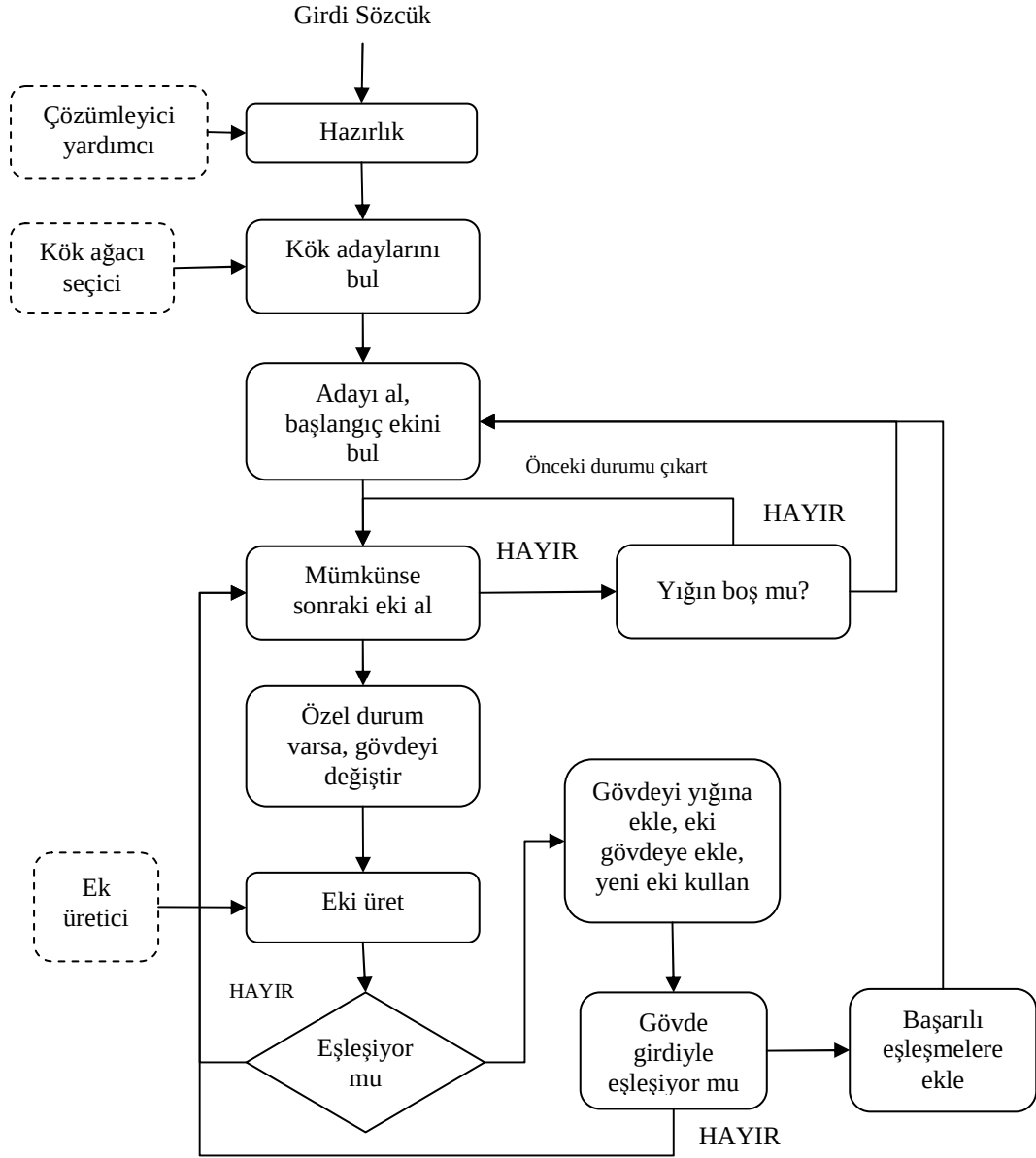
Biçimbirimsel Çözümleyici

Zemberek kütüphanesinde yer alan biçimbirimsel çözümleyici temel olarak verilen sözcük için olası kök ve ekleri bulur. Bu çözümleyici yapısı itibari ile basit sözlük tabanlı yukarıdan aşağıya çözümleyicidir. Biçimbirimsel çözümleyicinin verilen girdi sözcük üzerinde uyguladığı adımlar şu şekildedir:

1. Sözcük önce ön işlemde geçirilir.
2. Uygun kök seçici ile sözcük için kök adayları seçilir.
3. Her kök adayı için, olası ekler eklenerek sözcüğe ulaşılmaya çalışılır. Bu adım sözcüğe ulaşılan veya eklenecek ek kalmayana kadar devam eder.
4. Son olarak çözümleme sonuçları son işlemde geçirilir.

Sözcük, ön işlemde geçirilerek çözümleme işlemine hazırlanır. Ön işlemde vurgu, tire işaretleri kaldırılır, sözcük küçük harf dizisi haline getirilir. Eğer sözcük daha önceden tanımlanmış alfabede yer almayan bir harf içermesi durumunda çözümleme işlemine geçilmez. Ön işleme adımından sonra sözcüğün başından başlanarak kök adayları bulunmaya çalışılır. Daha önce bahsedildiği gibi Zemberek kütüphanesi sadece son ekleri desteklemektedir, ön ek desteği yoktur. Kök adayları bulunduğundan sonra sistem kökün tipine göre ilk eki seçer. Zemberek kütüphanesi içerisinde her ek için o ekten sonra gelebilecek eklerin listesi yer almaktadır. Bu listeler kullanılarak uygun ek dizisi elde edilmeye çalışılır. Eklerin eklenmesi aşamasında dikkat edilmesi gereken durum, özel durumlardır. Kök adayına bir ek eklendiğinde kökün yapısı bozulabilir. Örneğin, “*burun*” kökü sesli harf ile başlayan bir ek aldığı anda kökteki son sesli düşer ve kök “*burn*” olur. Zemberek kütüphanesinde ek üretim kurallarını düzenleyen yapıya ek üretici denir. Her ek üretici, dil bağımlıdır ve bağımlı olduğu dilin kurallarına göre somut ekler üretmekle sorumludur. Ek üretildikten sonra, doğruluğu girdi sözcük ile karşılaştırılarak bulunur. Sözcük karşılaştırmaları için farklı karşılaştırmacılar kullanılabilir: tam, ASCII toleranslı, hata toleranslı. Eğer ek doğru olarak yaratılmışsa, sözcüğün yeni şekli ve ekin durumu bir yığında saklanır ve işlemlere devam edilir. Girdi sözcük ve başarılı ekler eklenmiş kök düzgün şekilde eşleştirildiğinde sistem sonucu döner veya eşleşen ek kalmayana kadar diğer olası çözümleri denemekle devam eder.

Genel olarak biçimbirimsel çözümleme işlemi dilden bağımsızdır. Fakat doğru olası çözümlemeler bulunduğu anda uygulanan son işlem adımı dile bağımlıdır ve her dil için ayrı ayrı tanımlanmıştır. Bu adımda önışlem adımıında çıkarılan işaretler ve düzeltilen küçük harfler girdi sözcüğüne göre tekrar düzeltilir. Biçimbirimsel çözümleyicinin açıklanan adımlarını özetleyen blok şeması Şekil 3.7’deki gibidir.



Şekil 3.7: Zemberek kütüphanesi biçimbirimsel çözümleyici blok şeması

Biçimbirimsel çözümleyicinin ürettiği sonuçlar dilbilgisi kurallarına göre doğru olsa da anlamı ve cümle içerisindeki görevi itibarı ile doğru olmayabilir. Örneğin “*koyun*” sözcüğü incelendiğinde biçimbirimsel çözümleyicinin sonuçları Şekil 3.8’deki gibi

olacaktır. Fakat bu sözcüğün bir hayvan cinsi anlamında mı, yoksa koy eyleminin emir kip olarak kullanıldığı yoksa küçük körfez anlamına gelen koy olarak mı kullanıldığı içinde bulunduğu cümle ve anlamsal ilişkide bulunduğu diğer sözcükler incelenerek bulunabilir.

```
{Icerik: koyun Kok: koyun tip:ISIM} Ekler:ISIM_KOK
{Icerik: koyun Kok: koyun tip:ISIM} Ekler:ISIM_KOK
{Icerik:koyun Kok: koyu tip:SIFAT} Ekler:ISIM_KOK +
ISIM_SAHİPLİK_SEN_IN
{Icerik: koyun Kok: koy tip:FIIL} Ekler:FIIL_KOK + FIIL_EMİR_SİZ_IN
{Icerik: koyun Kok: koy tip:ISIM} Ekler:ISIM_KOK + ISIM_TAMLAMA_IN
{Icerik: koyun Kok: koy tip:ISIM} Ekler:ISIM_KOK +
ISIM_SAHİPLİK_SEN_IN
```

Şekil 3.8: Biçimbirimsel çözümleyici örnek çıktısı

Sonuç

Zemberek kütüphanesi, diğer diller için yapılmış olan çalışmaların Türkçe için uyarlaması olan bir proje olmaktan ziyade, açık kaynak kodlu, esnek, kullanım kolaylığı sağlayan, yüksek performanslı ve sadece Türkçe için değil diğer Türk dil ailesinde yer alan diller için geliştirilmiş olan bir doğal dil işleme kütüphanesidir. Bu özellikleri ile Zemberek kütüphanesi bu tez kapsamında kullanmak için tercih edilmiştir. Zemberek kütüphanesinin desteklediği doğal dil işleme işlemleri çok sayıda olmasına rağmen tez kapsamında sadece biçimbirimsel çözümleyici gövdeleyici olarak kullanılmıştır. Biçimbirimsel çözümleyicinin ürettiği olası çözümlemelerden biri farklı yöntemlerle seçilerek sözcüklerin gövdelerine ulaşılması amaçlanmaktadır.

4. TAMLAMA ANALİZİ

Tüm insan müdahalesi olmayan bilgisayar destekli Bilgi Erişimi sistemlerinde metinler ve doğal dil sorgularını temsil etmek için içerik tanımlayıcıları kullanır. Bu tanımlayıcılar tek sözcükten oluşan basit tanımlayıcılar olabileceği gibi birden fazla sözcükten veya tamlamadan oluşan karmaşık tanımlayıcılar da olabilir.

Bilgi Erişiminde içerik tanımlayıcıları temsil ettikleri metni içerik bakımından tam olarak karşılayamaması önemli sorunlardan bir tanesidir. Bunun sebebi genellikle sözcüklerin anlamlarının özgüllüğünün oldukça değişmesidir. Bazı sözcüklerin anlamları özeldir ve tanımladıkları kavramlar dar bir alandadır. Diğer bir yandan bazı kelimelerin anlamları geneldir ve bu sözcüklerle tanımlanan kavramlar geniş kapsamlıdır. Bilgi Erişim sistemlerinde metinlerin temsili için çok geniş veya çok dar kapsamlı kavramlar kullanılması durumunda erişim sonucunda çok az ya da çok fazla sonuç bulunacaktır [27].

Bilgi Erişimi teorik olarak ilk ortaya atıldığı zamanlardan itibaren metinler için içerik tanımlayıcısı olarak anahtar sözcükler kullanılmış ve Anahtar Sözcük Erişimi Hipotezi kabul görmüştür. Anahtar Sözcük Erişimi Hipotezi'ne göre bir belge ve bir sorguda ortak anahtar sözcük bulunduğunda belge sorgu ile ilgilidir [28]. Tabi ki, ne kadar çok ortak anahtar sözcük varsa, belge de sorgu ile o kadar ilgili olacaktır. Anahtar Sözcük Erişimi Hipotezi, metinleri anlamlarında uzak *sözcük torbası* olarak değerlendirir. Bu bakış açısı anlamsal ve sözdizimsel çeşitliliklerden kaynaklanan sorunlar ile ilgilenememekte ve çözememektedir. Bu sorunların çözümü için artık Anahtar Sözcük Erişimi Hipotezi'nin yerini Tamlama Erişimi Hipotezi almaktadır. Tamlama Erişimi Hipotezi'ne göre bir belge ve bir sorguda ortak tamlama bulunduğunda belge sorgu ile ilgilidir [28]. İçerik tanımlayıcısı olarak tamlamaları kabul eden bu hipotezin dayandığı temel aynı tamlama içerisinde yer alan sözcükler aynı anlamsal ilişkiyi paylaşmasıdır. Örneğin, kullanıcı sorgu olarak “Türk Tarih Kurumu”nu girdiğinde, bilgi erişim sistemi eğer metinleri tek sözcüklerden oluşmuş bir dizi halinde temsil ediyorsa arama sonucu olarak Türk, tarih ve kurum

sözcüklerinin tek, ikili ve üçlü geçtiği metinleri bulacaktır. Bu sözcüklerin tek ve ikili geçtiği sonuçlar yanlış olduğundan ilk aşamada kolayca elenebilir. Fakat bu üç sözcüğün de yer aldığı metinler kullanıcının tam olarak aradığı sonuçlar da olmayabilir. Örneğin bu sonuçlar arasında Türk bürokrasinin kurumsal tarihi üzerine yazılmış kaynaklar da gelecektir. Fakat aynı sözcüklerin yer almasına rağmen “Türk Tarih Kurumu” sorgusu ile Türk bürokrasinin kurumsal tarihinden bahseden bir belge arasındaki anlamsal benzerlik oldukça düşüktür.

Bilgi Erişimi sistemlerinin başarımı *keskinlik* ve *anımsama* gibi kavramlarla ölçülür. *Keskinlik*, bulunan ilgili belgelerin tüm bulunan belgelere oranı, *anımsama* ise bulunan ilgili belgelerin tüm ilgili belgelere oranı olarak tanımlanabilir. Keskinlik ve anımsamanın yüksek olduğu Bilgi Erişim sistemleri başarılı olarak kabul edilmektedir. Keskinlik ve anımsama kavramları da dolaylı olarak üzerinde çalışılan metnin en kadar iyi temsil edildiği ile ilişkilidir. Metinlerin temsilinde tek anahtar sözcükler yerine tamlamaların kullanılmasının sonuçları şu şekilde olacaktır:

- Birden fazla sözcükten oluşan sözlüksel birimlerin (acele posta, balık bilimi, Osmanlı Türkçesi gibi) tek başına sözcüklerin yerine kullanılması ile hem keskinlik hem de anımsama artar.
- İsim ve sıfat tamlamaları gibi sözdizimsel tamlamaların, tamlamayı oluşturan sözcüklerin tek tek kullanılmasına göre anımsamayı etkilemeden keskinlik artar.

Özellikle tamlamaları, tek anahtar sözcükler yerine içerik tanımlayan erişim birimi olarak ele almak önemlidir. Bunun en önemli nedeni tamlamaların, tüm doğal dillerde sözdizimsel olarak cümlelerde özne ve nesne görevinde bulunması ve cümlenin tanımlayıcı, açıklayıcı elemanlarının bu görevlerde yer almasıdır.

Metin içindeki tamlamaların belirlenmesi için bir çok yöntem vardır. Bu yöntemler sözcüklerin sıklıkları ve yakınlıklarına dayanan *istatistiksel yöntemler* ve dilbilgisinin sözdizimsel ve anlamsal seviyelerinde yapılan *dilbilimsel yöntemler* olarak ikiye ayrılabilir.

İstatistiksel tamlamalar; herhangi bir uzunlukta (ikili, üçlü veya daha uzun) bir birini izleyen sözcüklerden oluşan tamlamalardır. Örneğin, *Amerika Birleşik Devletleri*

hakkında yazılmış bir metinde *amerika*, *birleşik* ve *devletleri* sözcükleri yüksek sıklıkta peş peşe yer alacaktır.

Dilbilimsel tamlamalar ise aralarında dilbilgisi kurallarına uygun anlam ilişkileri bulunan sözcük öbekleridir. Örneğin *Büyük Ortadoğu Projesi* sözcük dizisinde *büyük* sıfatı *ortadoğu* ismini nitelendirmekte ve *ortadoğu* sözcüğü de *proje* sözcüğünü tamlamaktadır. Bu üç sözcük iç içe geçmiş bir sıfat ve belirtisiz isim tamlaması oluşturmaktadır. Dilbilimsel tamlamalarda, sözcükler arasında sadece bitişiklik ilişkisinden ziyade dilbilimsel ilişki vardır. Bu dilbilimsel ilişki de anlamsal ilişkiye yol açar.

4.1 İstatistiksel Tamlama Analizi

İstatistiksel tamlamalar; herhangi bir uzunlukta (ikili, üçlü veya daha uzun) bir birini izleyen sözcüklerden oluşan tamlamalardır. İstatistiksel tamlama analizinde kullanılan istatistiksel yöntem beş ana parametrenin kullanımına dayanmaktadır.

- *Uzunluk*: Tamlamada yer alabilecek en fazla eleman sayısıdır. Yapılacak deneylerde tamlamalar iki sözcükten oluştuğu varsayılarak uzunluk iki olarak belirlenir. Ancak bu parametreye ikiden daha büyük bir değer verilecek daha uzun sözcük grupları aranabilir.
- *Etki Alanı*: Tamlamanın elemanlarının tümünün bulunması gereken alandır. Bu değer cümle, paragraf ve belge olabilir.
- *Yakınlık*: Tamlama elemanlarının her biri, bir önceki ve bir sonraki eleman ile belirli bir yakınlıkta olmalıdır. Yakınlık değerinin bir olduğu durumlarda tamlama elemanları ardı ardına gelirken birden büyük olduğu durumlarda tamlama elemanlarının arasına tamlamada yer almayan sözcükler girebilir.
- *Belge Sıklık Eşik Değeri (BS_s)*: Tamlamada yer alan her sözcüğün derlemde aynı belge sıklığında geçmesi beklenemez. Tamlama elemanlarından en azından birinin önceden belirlenmiş olan belge sıklık eşik değerini geçmiş olması gerekir.

- *Tamlama için Belge Sıklık Eşik Değeri (BS_T):* Tamlamanın sadece elemanları değil bütün olarak derlemde belirli bir belge sıklık değerinde olmalıdır. Belirlenecek eşik değerinin hem alt hem de üst sınırı olabilir. Alt eşik sınırından küçük olan tamlamalar derlemde çok az geçtikleri için, üst eşik sınırından büyük olan tamlamalar ise çok fazla geçtikleri için derlemdeki değerlerini yitirirler.

Yukarıda belirtilen parametreler Tablo 4.1'deki değerler verilerek tanımlanan bir istatistiksel tamlama analizinden şu sonuçlar çıkarılabilir:

- Tamlama en fazla 3 sözcükten oluşabilir.
- Tamlamayı oluşturan sözcükler aynı cümlede yer alır.
- Durma sözcükleri çıkarıldığında tamlama elemanları birbirine komşu olurlar.
- Her tamlama derlemde en az 25 belgede geçen en az bir eleman içerir.
- Tamlama derlemde en az 1 belgede geçer.

Tablo 4.1: Bir istatistiksel tamlama analizi için örnek parametre değerleri

Parametre	Değer
Uzunluk	3
Etki Alanı	Cümle
Yakınlık	1
Belge Sıklık Eşik Değeri (BS _S)	25
Tamlama için Belge Sıklık Eşik Değeri (BS _T)	1

Tamlama analizi, en çok dizinleme sistemlerinde kullanılır. İstatistiksel tamlama analizi içeren bir dizinleme sisteminde öncelikle sırasıyla paragraf, cümle ve sözcükler ayrıştırılır. Daha sonra belgede yüksek sıklıkta bulunan fakat belgenin içeriği temsil etmeyen *durma sözcükleri* ayıklanır. Bu işlemten sonra genellikle gövdeleme işlemi uygulanarak farklı eklerin aynı kök ve gövdelere gelerek oluşturdukları biçimbirimsel çeşitlilikler ortadan kaldırılır. Daha sonra her sözcük için terim sıklığı, belge sıklığı ve tanımlanan parametrelere göre tamlama elemanı

olup olamayacağı hesaplanır. tf_{tv} , t teriminin, v vektöründeki ağırlığıdır. Bu değer Denklem 4.1 ile düzeltilir. Denklem 4.2'deki n , derlemdeki belge sayısı, df_t , t belgesinin belge sıklığını belirtir. Terim ve belge sıklıkları Denklem 4.1 ve 4.2'ye göre düzeltilebilir.

$$norm_tf_{tv} = \frac{tf_{tv}}{\max_tf_v} \quad (4.1)$$

$$tf_idf = norm_tf_{tv} \cdot \ln \frac{n}{df_t} \quad (4.2)$$

Son olarak da terimler için Denklem 4.3 ve olası tamlamalar için Denklem 4.4'e göre ağırlık değerleri hesaplanır.

$$w_{tv} = \frac{tf_idf_{tv}}{\sqrt{\sum_{i=1}^k tf_idf_{iv}^2}} \quad (4.3)$$

$$w_{pv} = \frac{\sum_{i=1}^m w_{mv}}{m} \quad (4.4)$$

Seçilen tamlamalar için Denklem 4.4'ün kullanılmasının nedeni; tamlamanın tamamının ağırlığı tüm tamlama elemanlarının ağırlıkları ile ilişkili olmalıdır, tamlamanın ağırlığının büyüklüğü, tamlama elemanlarının ağırlıklarının büyüklüğünde çok farklı olmamalıdır.

4.2 Dilbilimsel Tamlama Analizi

Dilbilimsel tamlamalar ise aralarında dilbilgisi kurallarına uygun anlam ilişkileri bulunan sözcük öbekleridir. Dilbilimsel tamlamalarda, sözcükler arasında sadece bitişiklik ilişkisinden ziyade dilbilimsel ilişki vardır. Bu dilbilimsel ilişki de anlamsal ilişkiye yol açar.

Tamlamalar, Türkçede önemli yer tutar. İsim ve sıfat tamlamaları hem acele posta, Osmanlı Türkçesi ve acemi ocağı gibi terim anlamları ile sözlüklerde yer alırken diğer bir yandan günlük hayatta hem yazılı hem de sözlü dilde anlatım ögesi olarak yoğun bir biçimde kullanılırlar. Ayrıca Türkçedeki isim ve sıfat tamlamaları oldukça düzenlidir. Tamlama elemanlarının sırası ve görevleri kurallara uyar ve istisnalar yok denecek kadar azdır [7]. Örneğin İngilizce’de *information retrieval* ve *retrieval of information* tamlamaları aynı anlamdadır. Fakat iki tamlamada tamlama elemanlarının sırası değişik olup ikinci tamlamada araya *of* edatı gelmiştir. Bu gibi durumlar için İngilizce’de dilbilimsel tamlama analizi yapılırken sözdizimsel düzeltmeler yapılması gereklidir [28]. Fakat Türkçede isim ve sıfat tamlamalarında bu gibi durumlar olmadığı için sözdizimsel düzeltmelere de gerek yoktur.

Dilbilimsel tamlamaların bulunması için geliştirilmiş olan yöntem sözcükler arasında tamlama ilişkisi olup olmadığını arar ve ilişkide olan sözcükleri ayırtmaya çalışır. Bu işlem için öncelikle Türkçedeki tamlama çeşitlerinin sembolik olarak ifade edilmesi gerekmektedir. Yazılacak sembolik ifadelerde ifade kolaylığı için düzenli ifadelerden faydalanılacaktır [29]. Kullanılan düzenli ifadeler ve anlamları Tablo 4.2’deki gibidir.

Tablo 4.2: Kullanılan düzenli ifadeler ve anlamları

Düzenli İfade Sembolü	Anlamı
(X)?	X ifadesi ya yer almaz ya da bir kere yer alır
(X)*	X ifadesi 0, 1 veya herhangi bir sayıda tekrarlanır
(X)+	X ifadesi en az bir kez tekrarlanır
[X Y]	X veya Y ifadesinde biri yer alır

Geliştirilen yöntem belirtili, belirtisiz ve zincirleme isim tamlamalarını ve sıfat tamlamalarını ayırtmaktadır. Bu tamlama çeşitleri için oluşturulan ilkel sembol ifadeler Tablo 4.3’deki gibidir. Tablo 4.3’de görüldüğü gibi zincirleme isim tamlamaları için tek bir ifade oluşturulmak yerine kapsanan durumlar beş ana öbekte toplanmıştır.

Tablo 4.3: Tamlamaların sembolik ifadeleri

Tamlama Çeşidi	Sembolik İfade	
	Tamlayan	Tamlanan
Belirtisiz İsim Tamlaması	İS	İS_IN
Belirtili İsim Tamlaması	İS_I	İS_IN
Zincirleme İsim Tamlaması (1)	İS İS_I_IN	İS_I
Zincirleme İsim Tamlaması (2)	İS_IN İS_I_IN	İS_I
Zincirleme İsim Tamlaması (3)	İS İS_I_NİN	İS İS_I
Zincirleme İsim Tamlaması (4)	İS_IN İS_I_IN	İS İS_I
Zincirleme İsim Tamlaması (5)	İS_IN	İS İS_I
Sıfat Tamlaması	SI	İS

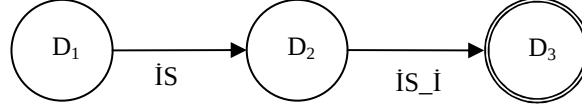
Sözdizimsel ayrıştırma işlemi için sonlu durum makineleri kullanılmaktadır. Sonlu durum makineleri, Doğal Dil İşleme çalışmalarında tam ayrıştırma gerektirmeyen durumlarda yeterlidir. Belirtili, belirtisiz ve zincirleme isim tamlamaları ve sıfat tamlamaları adım adım analiz edilmiş ve düzenli ifadelerle sonlu durum makineleri oluşturulmuştur.

4.2.1 Belirtisiz İsim Tamlaması

Belirtisiz isim tamlaması, tamlayanı söz söyleyen ve dinleyence belirli olmayan ve tamlayan ile tamlananı arasında sürekli bir bağ bulunan isim tamlamasıdır. Tamlayan, ek almaz; tamlanan ise -i ekini alır.

Adım 1: Belirtisiz isim tamlaması, bir isim ve onu takip eden –i tamlama eki almış isimden oluşur. Örneğin, *araba camı*. Belirtisiz isim tamlamalarının Adım 1 için düzenli ifade ve SDM Şekil 4.1’deki gibidir.

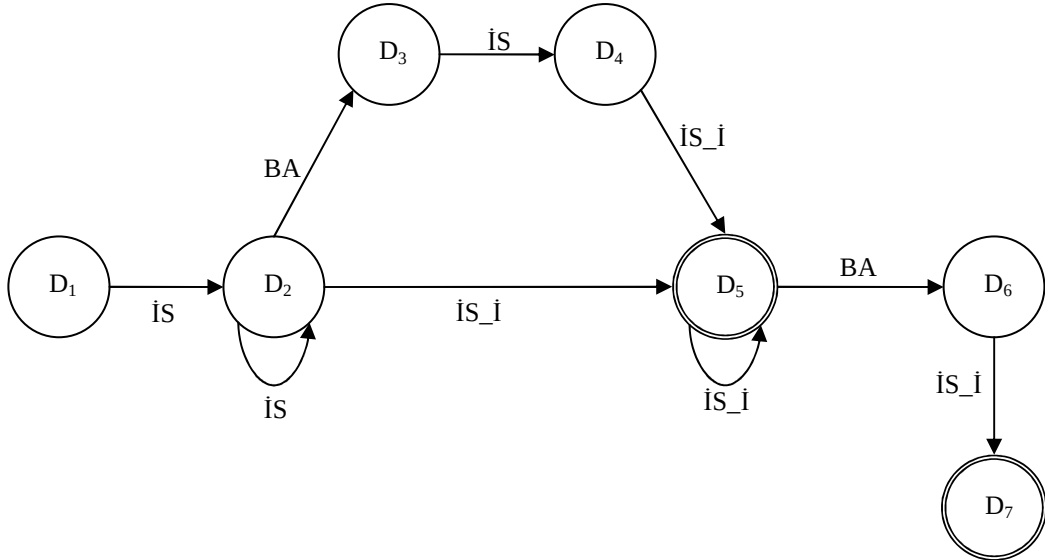
<u>araba</u> <u>camı</u> İS İS_İ	Düzenli İfade : İS İS_İ
-------------------------------------	--------------------------------



Şekil 4.1: Belirtisiz isim tamlaması Adım 1 için örnek, düzenli ifade ve SDM

Adım 2: Belirtisiz isim tamlamasında tamlayan ve/veya tamlanan birden fazla olabilir. Tamlayanın birden fazla olduğu durumda son tamlayan bağlaç ile ayrılabilir. Aynı kural tamlanan için de geçerlidir. Örneğin, *araba ve ev camı*, *araba satışı ve tamiri*. Belirtisiz isim tamlamalarının Adım 2 için düzenli ifade ve SDM Şekil 4.2’deki gibidir.

<u>araba</u> <u>ve</u> <u>ev</u> <u>camı</u> İS BA İS İS_İ
Düzenli İfade : İS+ (BA İS)? İS_İ+ (BA İS_İ)?



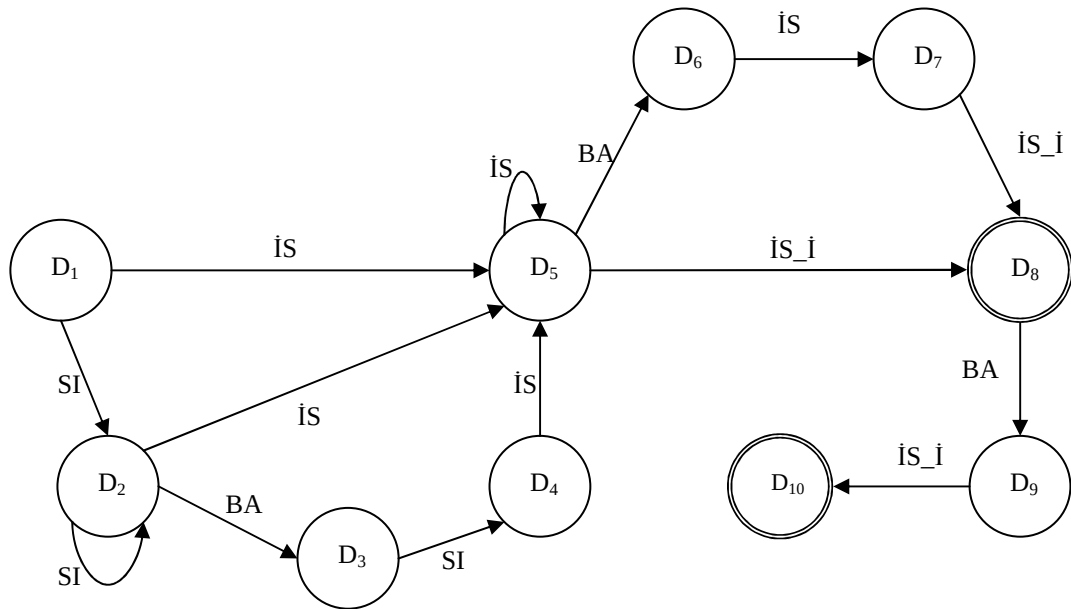
Şekil 4.2: Belirtisiz isim tamlaması Adım 2 için örnek, düzenli ifade ve SDM

Adım 3: Belirtisiz isim tamlamasında, sıfat ancak tamlayandan önce gelir. Bu sıfat da tamlayanı tek başına değil tamlamanın tamamını niteler. Birden fazla sıfat

tamlamayı nitelendirdiğinde son sıfat, sıfat öbeğine bağlaçla bağlanabilir. Belirtisiz isim tamlamasında öğeler anlam bakımından birbirlerine kalıcı olarak bağlanmalarından dolayı tamlayan ile tamlananın arasına başka bir sözcük giremez. Örneğin, *kırık araba camı* denebilir fakat *araba kırık camı* denemez. Belirtisiz isim tamlaması için tüm bu adımları kapsayan düzenli ifade ve SDM Şekil 4.3'deki gibidir.

<u>eski</u> ,	<u>kırık</u>	<u>ve</u>	<u>bakımsız</u>	<u>ev</u>	<u>duvarı</u>	<u>ve</u>	<u>kapısı</u>
SI	SI	BA	SI	İS	İS_İ	BA	İS_İ

Düzenli İfade : (SI+ (BA SI)?)? İS+ (BA İS)? İS_İ+ (BA İS_İ)?
--



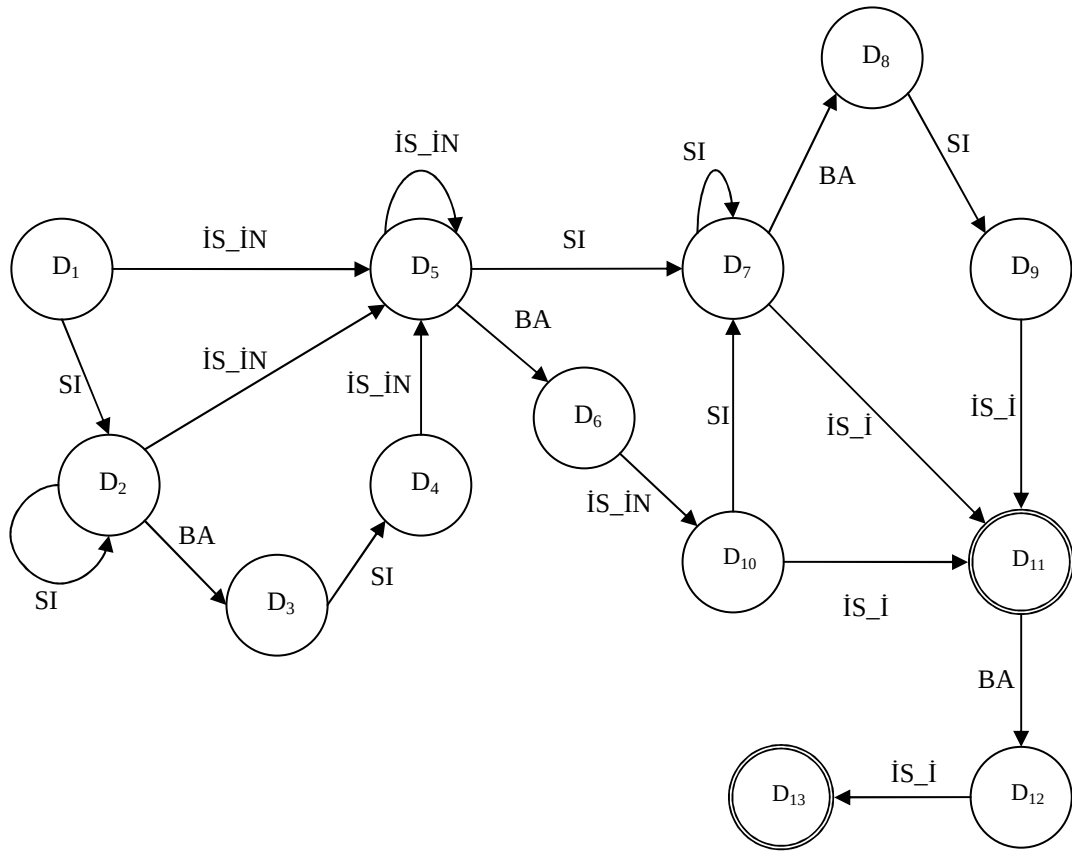
Şekil 4.3: Belirtisiz isim tamlaması için düzenli ifade ve SDM

4.2.2 Belirtili İsim Tamlaması

Belirtili isim tamlaması, tamlayanı söz söyleyen ve dinleyence bilinen ve tamlananı ile tamlananı arasında geçici bir ilgi bulunan tamlamadır. Bu tamlamada tamlayan eki -in; tamlanan eki ise -i'dir. Belirtili isim tamlamasının yapısı ve özellikleri belirtisiz isim tamlamasına oldukça benzemektedir. Belirtili isim tamlamasında da belirtisiz isim tamlamasında olduğu gibi tamlayan ve tamlanan birde fazla isimden oluşabilir ve isim öbeğindeki son isim bağlaç ile ayrılabilir. Belirtili isim

taamlamasında anlam ilişkisi geçici olduđu için hem taamlayan hem de taamlanan sıfat veya sıfat öbeđi ile nitelendirilebilir. Belirtili isim taamlaması, bu durum ile belirtisiz isim taamlamasından ayrılmaktadır. Belirtili isim taamlaması için belirtisiz isim taamlaması için bulunmuş olan düzenli ifade ve SDM kullanılarak Şekil 4.4'deki düzenli ifade ve SDM elde edilir.

D1 : (SI+ (BA SI)?)? IS_IN+ (BA IS_IN)? (SI+ (BA SI)?)? IS_I+ (BA IS_I)?



Şekil 4.4: Belirtili isim tamlaması için düzenli ifade ve SDM

4.2.3 Zincirleme İsim Tamlaması

Birden fazla isim tamlaması birlikte, iç içe kullanıldığı durumlarda zincirleme isim tamlaması oluşur; ancak zincirleme isim tamlamaları da bir bütün olarak belirli veya belirsiz isim tamlamasıdır. Tamlama analizi yapılırken zincirleme isim tamlamaları beş öbek halinde değerlendirilir.

Öbek 1: Bu öbekte, tamlayanı belirtisiz isim tamlaması, tamlananı ise –i tamlama eki almış isimden oluşan zincirleme isim tamlamaları yer alır. Örneğin, *okul duvarının boyası*. Bu öbekteki örneklerde ikinci sözcük hem de –i tamlama hem de –in tamlama ekini alarak tamlayan olan belirtisiz isim tamlaması ile tamlanan olan isim arasındaki ilişkiyi sağlar.

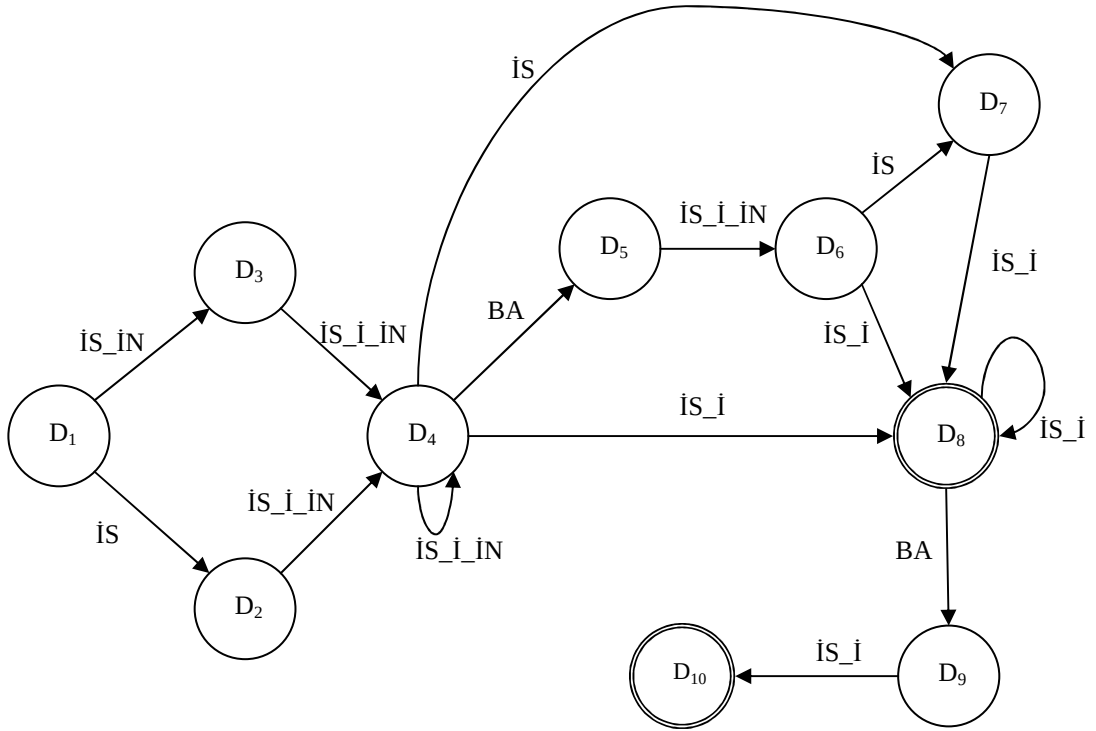
Öbek 2: Bu öbekte, tamlayanı belirtili isim tamlaması, tamlananı ise –i tamlama eki almış isimden oluşan zincirleme isim tamlamaları yer alır. Örneğin, *okulun bahçesinin kapısı*.

Öbek 3: Bu öbekte, hem tamlayanı hem de tamlananı belirtisiz isim tamlaması olan zincirleme isim tamlamaları yer alır. Örneğin, *okul bahçesinin sokak lambası*. Bu öbekte yer alan örneklerde her iki tamlama ilk tamlamanın tamlananına gelen –in tamlama eki ile birbirine bağlanır.

Öbek 4: Bu öbekte, tamlayanı belirtili isim tamlaması, tamlananı ise belirtisiz isim tamlaması olan zincirleme isim tamlamaları yer alır. Örneğin, *Kıvanç'ın kardeşinin sokak elbisesi*.

Zincirleme isim tamlamalarının Öbek 1, Öbek 2, Öbek 3 ve Öbek 4 için oluşturulan düzenli ifade ve sonlu durum makinesi Şekil 4.5'deki gibidir.

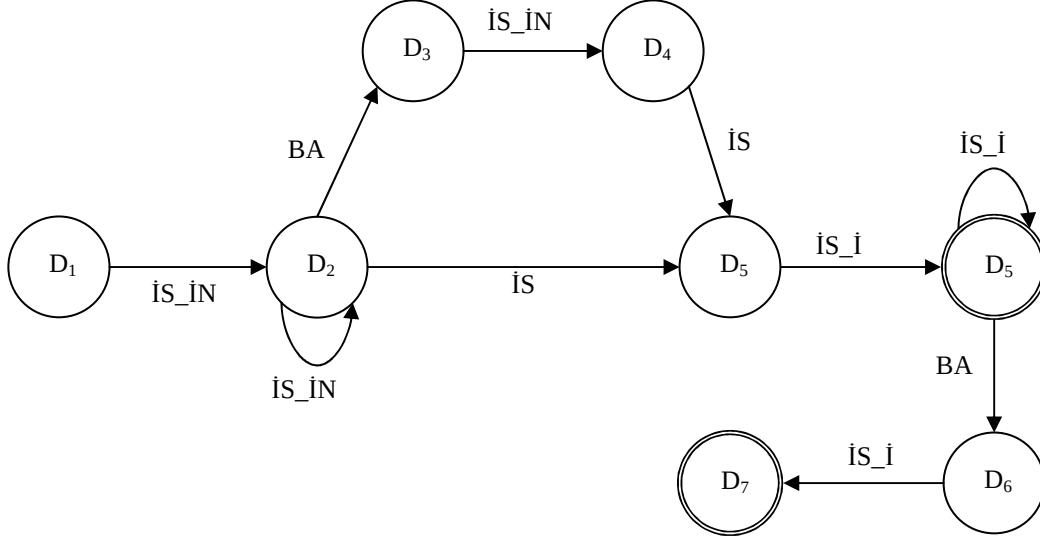
Dİ : [İS İS_IN] İS_I_IN+ (BA İS_I_IN)? İS? İS_I+ (BA İS_I)?



Şekil 4.5: Zincirleme isim tamlaması Öbek 1, 2, 3 ve 4 için düzenli ifade ve SDM

Öbek 5: Bu öbekte, tamlayanı isim, tamlananı ise belirtisiz isim tamlamasından oluşan zincirleme isim tamlamaları yer alır. Örneğin, *okulun bahçe kapısı*. Öbek 5 için düzenli ifade ve sonlu durum makinesi Şekil 4.6'deki gibidir.

Düzenli İfade : $\dot{I}S_IN (BA \dot{I}S_I)? \dot{I}S \dot{I}S_I+ (BA \dot{I}S_I)?$



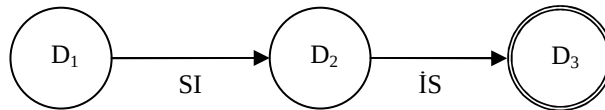
Şekil 4.6: Zincirleme isim tamlaması Öbek 5 için düzenli ifade ve SDM

4.2.4 Sıfat Tamlaması

Sıfat tamlamaları, bir sıfat ve bu sıfatın nitelendirdiği veya belirttiği bir isimden oluşan sözcük öbekleridir. Sıfat niteleyen, isim ise nitelenen ögedir.

Adım 1: Sıfat tamlaması, bir sıfat ve onu izleyen bir isimden oluşur. Örneğin, *güzel kız*. Sıfat tamlaması Adım 1 için SDM ve düzenli ifadeler Şekil 4.7’deki gibidir.

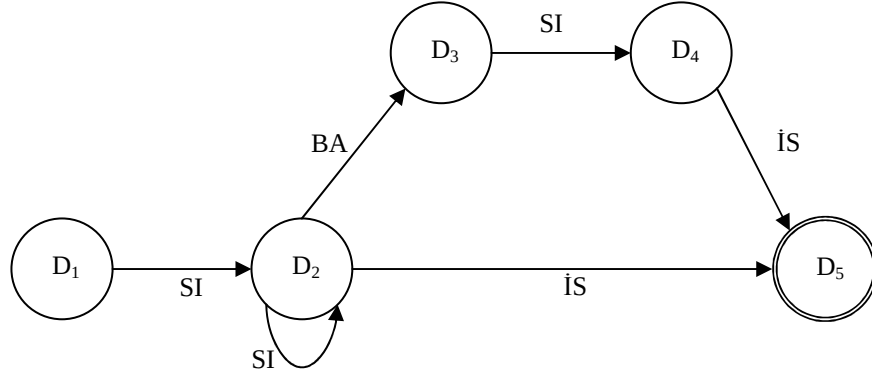
$\frac{\text{güzel}}{SI} \quad \frac{\text{kız}}{IS}$	Düzenli İfade : $SI \dot{I}S$
---	--------------------------------------



Şekil 4.7: Sıfat tamlaması Adım 1 için örnek, düzenli ifade ve SDM

Adım 2: Sıfat tamlamasında, sıfat birden fazla olabilir ve sıfat öbeğinde son sıfattan önce bağlaç gelebilir. Örneğin, *büyük, eski ve güzel ev*. Sıfat tamlamalarının Adım 2 için SDM ve düzenli ifadeler Şekil 4.8’deki gibidir.

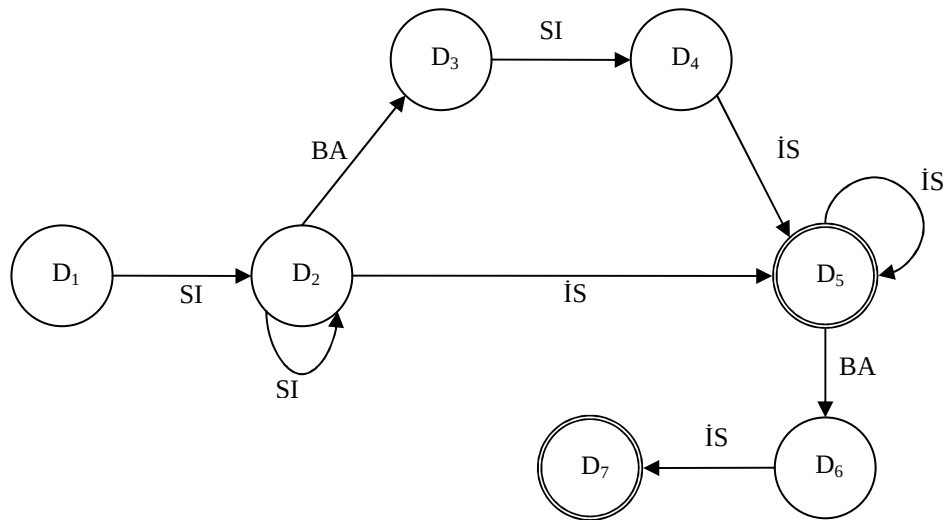
<u>büyük</u> , <u>eski</u> <u>ve</u> <u>güzel</u> <u>ev</u> SI SI BA SI İS	Düzenli İfade : SI+ (BA SI)? İS
---	--



Şekil 4.8: Sıfat tamlaması Adım 2 için örnek, düzenli ifade ve SDM

Adım 3: Sıfat tamlamasında, isim birden fazla olabilir ve isim öbeğinde son isimden önce bağlaç gelebilir. Örneğin, *kırmızı elbise ve ayakkabı*. Sıfat tamlamalarının 3. adımı için SDM ve düzenli ifadeler Şekil 4.9'deki gibidir.

<u>yeni</u> , <u>kırmızı</u> <u>elbise</u> <u>ve</u> <u>ayakkabı</u> SI SI İS BA İS
Düzenli İfade : SI+ (BA SI)? İS+ (BA İS)?



Şekil 4.9: Sıfat tamlaması için düzenli ifade ve SDM

5. SİSTEM ANALİZİ

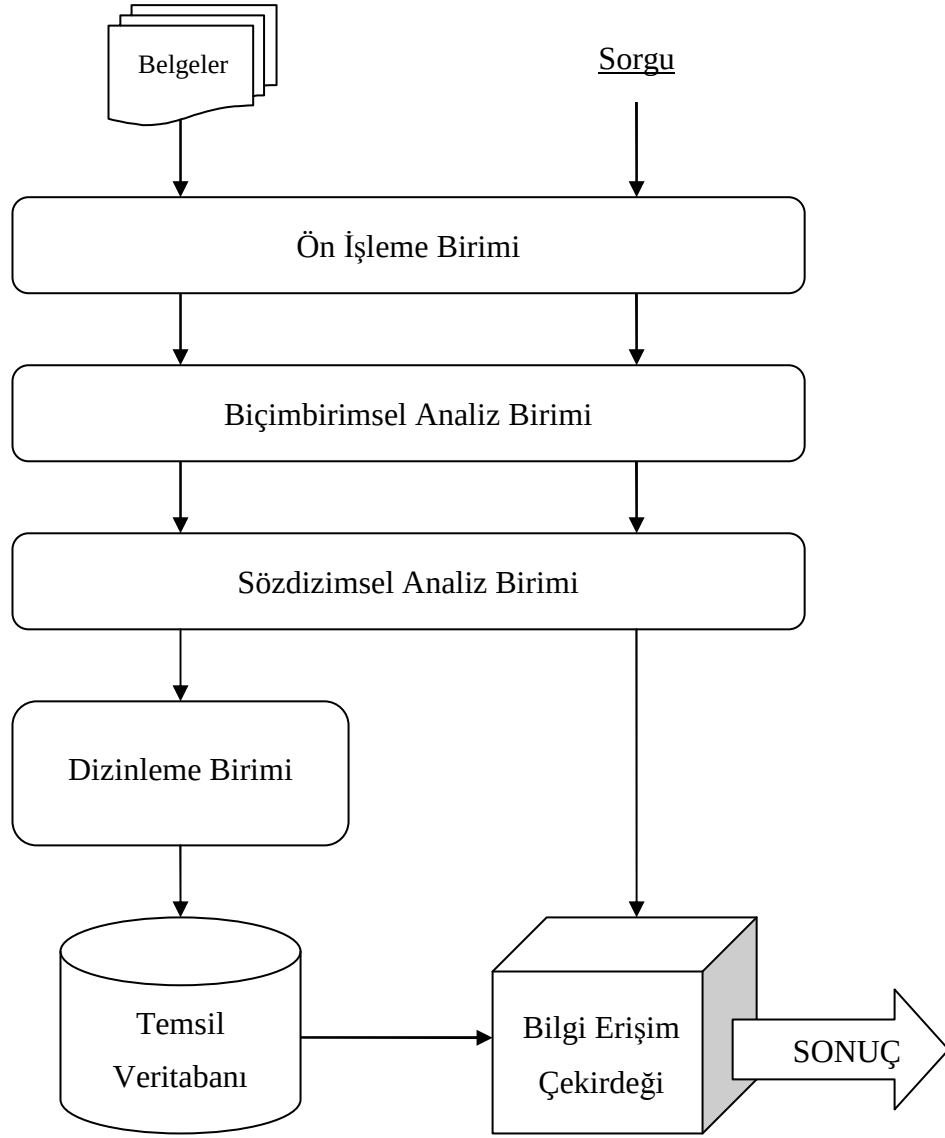
5.1 Sistem Tasarımı

Bu tez kapsamında Doğal Dil İşleme destekli Bilgi Erişimi sistemi tasarımı yapılmıştır. Bu sistem Türkçenin özelliklerini göz önünde alarak biçimsel ve sözdizimsel seviyede DDİ işlemleri yaparak bilgi erişiminin başarımını arttırmayı hedeflemektedir. Biçimsel seviyedeki farklılıklar için Bölüm 3.3.3’de bahsedildiği gibi Zemberek DDİ kütüphanesi kullanılarak biçimbirimsel çözümleme yapılır ve bu çözümleme sonuçlarından faydalananarak gövdeleme işleme gerçekleştirilir. Bu tezin, Bilgi Erişimi konusunda asıl yoğunlaştığı ve sonuçlarını ölçmeyi amaçladığı konu sözcükler yerine tamlamaların yani sözcük öbeklerinin belgelerin temsilde kullanılması ve bunun bilgi erişiminin başarımı üzerindeki etkisidir. Bunun için “Tamlama Temelli Dizinleme” konusu üzerinde çalışılmış ve tek sözcüklerin kullanıldığı genel terim temelli dizinleme ile karşılaştırılmaya çalışılmıştır.

Bilgi Erişimi sisteminde tamlama temelli dizinleme yapmak için öncelikle belgelerdeki tamlamaların bulunması gerekmektedir. Bu sebepten dolayı tamlama analizi üzerine farklı yaklaşımlar araştırılmış ve iki yaklaşım bu tez kapsamında incelenmiştir : istatistiksel tamlama analizi ve dilbilimsel tamlama analizi. İstatistiksel yaklaşımda, birbirini belirli sıklıklarla izleyen ve belge genelinde sıkça geçen sözcük öbekleri tamlama olarak kabul edilmiş, dilbilimsel yaklaşımda ise isim ve sıfat tamlamaları bulunmaya çalışılmıştır. İstatistiksel yaklaşımda sözcükler arasındaki yakınlık ve sıklık ilişkileri üzerinde durulurken dilbilimsel yaklaşımda biçimbirimsel çözümleme de kullanılarak sözcük türleri ve sözcüklere gelen tamlama ekleri incelenip Türkçedeki isim ve sıfat tamlamaları sonlu durum makineleri ve düzenli ifadeler olarak temsil edilmiştir. Tasarlanan Bilgi Erişimi sisteminde sadece dilbilimsel tamlama analizine yer verilmiş ve gerçekleştirilmiştir..

Sistem tasarımı yapılırken sadece tamlama tabanlı dizinleme konusunu gerçeklemek yerine genel bir bilgi erişimi sistemi tasarımına gidilmiştir. Tasarım sırasında DDİ işlemlerinde farklı yaklaşımların ortaya çıktığı durumlarda farklı seçenekler kullanıcı tarafından daha sonra ayar ekranından değiştirilebilir şekilde tasarlanmıştır. Bu sayede kullanıcı DDİ işlemlerinde farklı yaklaşımları sınavabilmekte ve sonuçlara olan etkisini gözlemleyebilmektedir.

Tasarım aşamasında dikkat edilen önemli bir husus da birimselliktir. Tasarım sırasında mümkün olduğu kadar birimsellik ön planda tutulmuş ve yapılan işlemlere göre sistem birimlere ayrılıp birimler birbirlerinden bağımsız olarak ele alınabilmektedir. Bir birimin çıktısı diğer bir birimin girdisi olarak kullanılmıştır. Fakat birimlerin iç yapıları birbirlerinden bağımsızdır. Doğal Dil İşleme destekli Bilgi Erişimi sisteminin genel yapısı ve birimleri Şekil 5.1'deki gibidir.



Şekil 5.1: Doğal Dil İşleme destekli Bilgi Erişim sistemin yapısı ve birimleri

Sistemin yazılım kısmı için Java programlama dili ve NetBeans uygulama geliştirme ortamı kullanılmıştır. Sistemin gerçekleştirilmesi sırasında Java 1.6.0 ve NetBeans 5.5 en yeni sürümler oldukları için tercih edilmiştir [30,31]. Programlama dili olarak Java'nın seçilmesinin sebebi nesneye dayalı, tutarlı, platform bağımsız yapısı ve yüksek performans sağlamasıdır. Ayrıca biçimbirimsel çözümleme için kullanılan Zemberek DDİ kütüphanesinin kaynak kodlarının Java programlama dilinde yazılmış olması da bu seçimde etkili olmuştur. NetBeans ise Java programlama dilindeki en gözde uygulama geliştirme ortamı olması ve Java'nın dağıtıcısı olan Sun Microsystem tarafından desteklenmesi sebebiyle tercih edilmiştir. Sistemin

veritabanı kısmında ise Microsoft Access veritabanı basit ve kurulum gerektirmeyen yapısı ve bilgilerin kolay taşınabilirliği sebebiyle tercih edilmiş ve kullanılmıştır.

5.1.1 Ön İşleme Birimi

Ön işleme birimi, belgeler üzerinde hazırlık işlemleri yapar ve biçimbirimsel analiz için hazırlar. Bu hazırlık için önce metinleri cümlelere ve daha sonra cümleleri sözcüklere ayırır. Bu ayrıştırma işlemi sırasında noktalama işaretleri ve boşluk karakterlerinden faydalanılır. Metinler, cümle dizisi, cümleler ise sözcük dizisi olarak saklanır.

Ön işleme biriminin yaptığı diğer bir işlem de durma sözcüklerinin metinden temizlenmesidir. Durma sözcükleri, metinlerde sıkça geçen fakat yer aldıkları belgeleri içerik olarak temsil edemeyen geniş anlamlı sözcüklerdir. Durma sözcüklerine örnek olarak *bir, bu, eğer, için* gibi sözcükler verilebilir. Durma sözcükleri listesi, bu daha önce yapılmış çalışmalar incelenerek oluşturulmuştur. Durma sözcüklerinin tam listesi Tablo A.1'deki gibidir.

5.1.2 Biçimbirimsel Analiz Birimi

Biçimbirimsel analiz biriminin görevi belgelerdeki ve sorgulardaki sözcükler üzerinde biçimbirimsel çözümleme yapıp, birden fazla olası çözümleme bulunması durumunda çeşitli yöntemlere göre en uygun çözümlemeyi seçmektir. Bu biçimbirimsel çözümleme sonucu, bu birim içerisinde gövdeleme işleminde ve bir sonraki adım olan sözdizimsel analizde tamlama eklerinin belirlenmesi için kullanılacaktır.

Bu birimde öncelikle Zemberek DDİ kütüphanesi kullanılarak her sözcük için biçimbirimsel analiz yapılır. Eğer analiz sonucu elde edilemezse sözcük üzerinde hiçbir işlem yapılmaz ve sözcük türü Hatalı olarak işaretlenir. Yapılan biçimbirimsel çözümlemeden sonra genellikle birden fazla çözümleme sonucu çıkmaktadır. Bu durumda her çözümleme adayı için gövde bulunur. Gövde bulunurken iki farklı yaklaşım kullanılmaktadır::

- *Köke kadar gövdeleme:* Sözcükten hem çekim hem de yapım eklerinin atılmasıyla köke ulaşılır.

- *En uzun gövdeyi bulma*: Sözcükten sadece çekim ekleri atılır, yapım ekleri atılmaz. Bu şekilde sözcükten en uzun gövdeye ulaşılmış olur.

Köke kadar gövdeleme yaklaşımı, bir bilgi erişimi sisteminde keskinliği kesin olarak arttıracak gibi anımsamayı aynı kökten türemiş, anlam olarak birbirlerinden uzak sözcüklerin kullanıldığı durumlarda düşürebilir. Örneğin, Tablo 5.1’de görüldüğü gibi Belge 1’de *sözel*, Belge 2’de ise *sözcü* sözcükleri sıkça yer almaktadır. Her iki sözcük de köke kadar gövdeleme yaklaşımıyla gövdelendiğinde *söz* sözcüğü bulunmaktadır. Bu durumda her iki belge de *söz* sözcüğü ile temsil edileceğinden dolayı benzer bulunacaktır. Fakat her iki belge de *en uzun gövdeyi bulma* yaklaşımı ile gövdelendiğinde farklı bulunacaktır. Yani, *en uzun gövdeyi bulma* yaklaşımında bilgi erişim sistemlerinde hem keskinlik hem de anımsama artacaktır. Bu özelliklerinden dolayı *en uzun gövdeyi bulma* yaklaşımı tercih edilmiştir.

Tablo 5.1: Köke kadar gövdeleme yaklaşımına örnek

Belge 1	Belge 2
ÖSS sınavında sözel bölümünde 90 soru bulunur...Sözel bölümü türkçe, tarih...	Grev sözcüleri saat yedide yerlerini aldılar... Sözcüler yaptıkları açıklamadan sonra...
sözel → söz	sözcü → söz

Her çözümleme adayı için gövde bulunduktan sonra çözümleme adaylarında en sık geçen gövde bulunur. Bu yaklaşımdaki amaç, aynı sözcük için bulunan köklerin genellikle aynı gövdeyi vermesidir. Örneğin, detayları Tablo 5.2’de verilen örnek biçimbirimsel çözümleme çıktısında *değişimin* sözcüğü için yedi farklı çözümleme bulunmuştur. Bu çözümleme adayları arasında kök olarak *değişim* sözcüğü iki, *değiş* sözcüğü iki ve *değ* sözcüğü üç kere yer almaktadır. Fakat her çözümleme için en uzun gövdeyi bulma yaklaşımı ile gövdeleme yapıldığında altı çözümleme *değişim* gövdesi, bir çözümleme ise *değiş* gövdesi bulunmaktadır.

Bu örnekte de açık olarak görüldüğü gibi bir sözcük için çözümleme adayları arasında en sık geçen gövdeyi seçmek yüksek oranda doğrudur. Diğer örnekler üzerinde yapılan çalışmalar birkaç istisna dışında bu yaklaşımın doğru olduğunu göstermiştir.

Tablo 5.2: Biçimbirimsel analiz birimi için örnek girdi ve çıktı

Sözcük	değişimin
Çözümleme adayları	Kok: değişim tip:ISIM Ekler:ISIM_KOK + ISIM_TAMLAMA_IN Kok: değişim tip:ISIM Ekler:ISIM_KOK + ISIM_SAHİPLİK_SEN_IN Kok: değış tip:FIIL Ekler:FIIL_KOK + FIIL_DONUSUM_IM + ISIM_TAMLAMA_IN Kok: değış tip:FIIL Ekler:FIIL_KOK + FIIL_DONUSUM_IM + ISIM_SAHİPLİK_SEN_IN Kok: değ tip:FIIL Ekler:FIIL_KOK + FIIL_BERABERLİK_IS + FIIL_DONUSUM_IM + ISIM_TAMLAMA_IN Kok: değ tip:FIIL Ekler:FIIL_KOK + FIIL_BERABERLİK_IS + FIIL_DONUSUM_IM + ISIM_SAHİPLİK_SEN_IN Kok: değ tip:FIIL Ekler:FIIL_KOK + FIIL_DONUSUM_IS + ISIM_SAHİPLİK_BEN_IM + ISIM_TAMLAMA_IN
Seçilen çözümleme	Kok: değişim tip:ISIM Ekler:ISIM_KOK + ISIM_TAMLAMA_IN
Gövdeleme sonucu	değişim

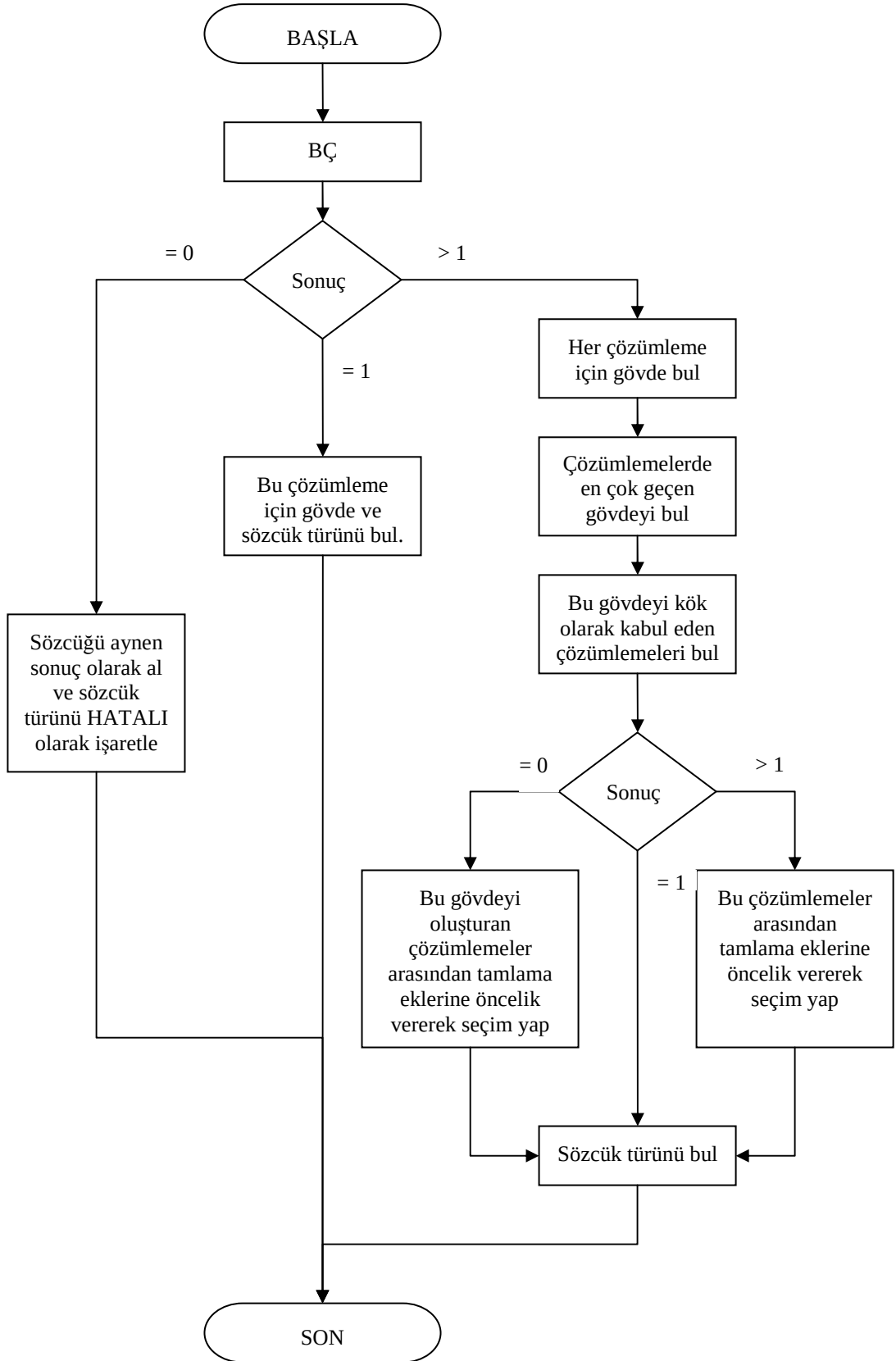
Sözcük için gövde belirlendikten sonra bu gövdeyi kök olarak veren çözümlemeler seçilir. Bundaki amaç sözcüğün türünü belirleme işlemlerini basitleştirmektir. Bu gövdeyi kök olarak bulan çözümleme adaylarının birden fazla olması durumunda bu çözümlemeler arasından tamlama eklerine öncelik vererek seçim yapılır. Eğer buluna gövdeyi kök olarak bulan çözümleme adayları yoksa bulunan gövdeye ulaşılabilen çözümlemeler arasından yine tamlama eklerine öncelik verilerek seçim yapılır. Çözümlemelerin seçiminde en yüksek öncelikli olan hem –i tamlama hem de –in tamlama ekini içeren çözümlemelerdir. Bundan sonra bu iki ekten birini içeren çözümlemelere öncelik verilir. En son olarak bulunan çözümleme için sözcük türü bulunur. Sözcük türünün bulunması özellikle tamlama analizi için oldukça önemlidir.

Biçimbirimsel çözümleme için kullanılan Zemberek DDİ kütüphanesi özellikle gövdenin ve sözcük türünün bulunması konusunda yetersiz kalmaktadır. Gövdenin

ve sözcük türünün bulunması için Zemberek DDİ kütüphanesinde çeşitli değişiklikler yapılması gerekmiştir. Kütüphane açık kaynak kodlu olması nedeniyle bu değişikliklere oldukça elverişlidir. En uzun gövdeyi bulma yaklaşımı ile yapılan gövdeleme için özellikle eklerin türleri çok önemlidir. Zemberek DDİ kütüphanesi bünyesinde ek kural ve üretimlerini içeren bir ek sözlüğü barındırmaktadır. Fakat bu ek sözlüğünde eklerin türleri yani çekim veya yapım eki oluşları belirtilmemiştir. Ek sözlüğü her ekin türü belirlenerek genişletilmiş, kütüphanede bu değişiklikleri okuyup değerlendirilebilecek şekilde güncellenmiştir. Yapılan bu değişiklik ile sadece çekim eklerinin atılması ile yapılan en uzun gövdeyi bulma yaklaşımı ile gövdeleme gerçekleştirilmiştir.

Zemberek DDİ kütüphanesinde yapılan bir diğer değişiklik ise sözcük türleri üzerinedir. Dilbilimsel tamlama analizi yapılırken cümle içerisindeki sözcüklerin türlerinin doğru olarak belirlenmesi analizin başarımını doğrudan etkileyen bir unsurdur. Zemberek DDİ kütüphanesinde sadece köklerin sözcük türleri bulunmaktadır. Çözümleme sonucunda bulunan gövdenin sözcük türü genellikle kök ve ona eklenen yapım eklerinden yararlanılarak bulunması gerekmektedir. Bu yüzden sözcük türünü değiştiren yapım ekleri için yine ek sözlüğünde değişiklik yapılmış ve bu eklerin eklendikleri sözcüklerin türünü ne yönde değiştirdiği konusunda bilgi eklenmiştir. Yapılan bu değişiklik sayesinde kök üzerine yapım ekleri getirilerek gövdenin sözcük türü yüksek başarı ile belirlenebilmektedir.

Ayrıca yapılan denemelerde ek sözlüğünden çeşitli yanlışlar bulunmuş ve ek kurallarının bir kısmının silinmesi ve yeni kuralların eklenmesi ile biçimbirimsel analiz başarımı arttırılmıştır.



Şekil 5.2: Biçimbirimsel analiz biriminin akış şeması

5.1.3 Sözdizimsel Analiz Birimi

Sözdizimsel Analiz Birimi, belgelerde yer alan ve daha sonra dizinleme işlemi sırasında belgeyi temsil etmeye aday olacak tamlamaları bulur. Belgeler, sözdizimsel analiz birimine geldiğinde biçimbirimsel çözümlemeleri yapılmış, bu çözümleme sonuçlarından uygun çözümleme seçilmiş sözcük dizileri halindedir.

Bu birim tamlama analizi yaparken dilbilimsel yaklaşımı kullanır. Dilbilimsel yaklaşımda ise, dilbilimsel anlamda isim ve sıfat tamlamaları bulunmaya çalışılır. Yapılan dilbilimsel tamlama analizinde hem basit hem de bileşik tamlamalar üzerinde çalışılır. Basit tamlamalar, isim tamlamaları için bir tamlayan bir de tamlanan isimden oluşurlar. Bileşik tamlamalar ise isim ve sıfat tamlamalarında tamlayan ve tamlananın birden fazla sözcükten oluştuğu ve kimi durumlarda bağlaçların da kullanıldığı, sıfat ve sıfat öbeklerinin isim tamlamalarını nitelediği durumları kapsamaktadır. Dilbilimsel tamlama analizinde Bölüm 4.2’de anlatılmış olan sonlu durum makineleri ve düzenli ifadeler kullanılmıştır.

5.1.4 Dizinleme Birimi

Dizinleme birimi, belgedeki sözcükleri ve sözdizimsel analiz sonucu bulunan tamlamaları kullanarak dizinleme işlemini gerçekleştirir. Bu birim hem sözcükler hem de tamlamalar için ağırlık hesaplaması yapar. Bu ağırlık hesaplaması terim sıklığı ve ters belge sıklıklarını içerir. Tek sözcükten oluşan terimler için ağırlık Denklem 5.1 ile hesaplanır.

$$w_{tv} = \frac{tf_{tv} \cdot idf_{tv}}{\sqrt{\sum_{i=1}^k tf_{iv} \cdot idf_{iv}^2}} \quad (5.1)$$

Tamlamalar için hesaplanan ağırlık hesaplamasında ise tamlamayı oluşturan sözcüklerin ağırlıklarında yararlanır. Bunun sebepleri, tamlamanın önemi onu oluşturan elemanların öneminin bir fonksiyonu olması ve tamlama ağırlığının büyüklüğünün, onu oluşturan elemanların ağırlıklarının büyüklüklerinden ayrı olamayacağıdır. Tamlamalar için ağırlık Denklem 5.2 ile hesaplanabilir.

$$w_{pv} = \frac{\sum_{i=1}^m w_{mv}}{m} \quad (5.2)$$

Bu denklem Türkçenin özellikleri daha iyi kavraması açısından değiştirilebilir. Türkçede tamlamalarda önce tamlayan yani belirten daha sonra da tamlanan yani belirtilen gelmektedir. Tamlama kurmaktaki amaç tamlananı anlam yönünden genişletmek olduğu için tamlananın anlam bakımından daha önemli olduğu varsayılabilir. Bu durumda, tamlama ağırlıkları hesaplanırken tamlanan elemanların ağırlıkları tamlayanlarınkine göre daha önemli olmalıdır. Türkçede her zaman tamlayan tamlanandan önce geldiği düşünüldüğünde tamlamadaki sözcüklerin sırası kullanılarak bu önem denkleme yansıtılabilir. Tamlamaların ağırlıklarının hesaplanmasında kullanılan denklemin son hali Denklem 5.3'deki gibidir.

$$w_{pv} = \frac{\sum_{i=1}^m \ln(i+1) \cdot w_{mv}}{m} \quad (5.3)$$

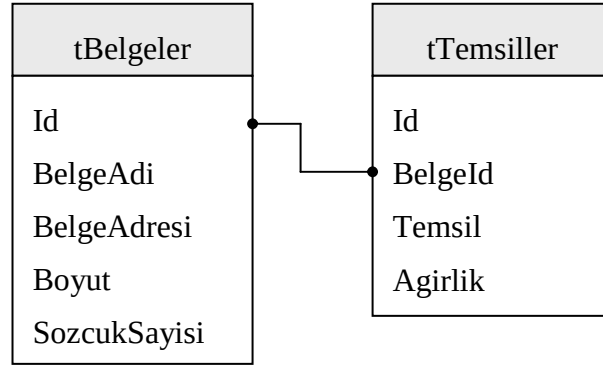
Dizinleme birimi, hem sözcükler hem de tamlamalar için ağırlık hesaplaması yaptıktan belgeleri temsil edecek terimleri seçer. Bunu amacı daha sonra kullanılacak olan yöney uzayında boyut sayısını düşürmek ve karmaşıklığı azaltmaktır. Terim seçerken terim ağırlıkları göz önünde bulundurulur. Terim seçimi iki farklı yaklaşım ile yapılabilir:

- *İlk n terimin seçilmesi:* Belgedeki terimler (sözcükler ve tamlamalar) ağırlık sırasında göre dizilip en yüksek ağırlığa sahip n tanesi alınır.
- *Eşik değerini geçen terimlerin seçilmesi:* Belgedeki terimler arasından ağırlığı belirli bir eşik değerini geçenler seçilir. Bu eşik değeri her belge için ayrı ayrı hesaplanabileceği gibi ağırlık hesaplaması normalize edilmiş olduğu için derlem için bir kere hesaplanabilir. Eşik değerinin her belge için hesaplandığı durumlarda eşik değeri olarak en yüksek ağırlık değerinin %75'i alınır.

5.1.5 Temsil Veritabanı

Temsil veritabanı, belgelerin temsilleri de dahil olmak üzere özellikleri daha sonra sorgulanabilir şekilde saklar. Bu birim, belgelerin isimlerini, bulundukları dizini, boyutunu, içerdiği toplam sözcük sayısını ve temsil edildiği sözcük ve tamlamaları saklar.

Veritabanında iki adet tablo bulunmaktadır: tBelgeler, tTemsiller. tBelgeler tablosunda belgelerin özellikleri saklanır ve derlemde bulunan her belge için bir kayıt bulunur. tTemsiller tablosunda ise derlemde yer alan belgeleri temsil etmekte kullanılan sözcükler ve tamlamalar kullanılır. Bu tabloda yer alan Agirlik alanında sadece sözcüğün veya tamlamanın terim sıklığı saklanmaktadır, ters belge sıklığı dinamik olarak sorgu sırasında hesaplanmaktadır. Veritabanının varlık ilişki çizelgesi Şekil 5.3'deki gibidir.



Şekil 5.3: Veritabanı varlık ilişki çizelgesi

5.1.6 Bilgi Erişimi Çekirdeği

Bilgi erişim çekirdeği, kullanıcıdan gelen ve işlemlerden geçirilmiş olan sorgular ile derlemde bulunan belgeleri eşleştirir, ilgili belgeleri ilgi sırasına göre sıralı şekilde getirir ve sonuçları oluşturur. Sorgu bu birime geldiğinde terimler ve olası tamlama adaylarından oluşmaktadır. Bunlar, temsil veritabanında bulunan belgeleri temsilleri yani terimler ve tamlamalar ile eşleştirilirler.

Eşleştirme işlemi sırasında yöney uzay modeli kullanılır. Sorgu ve belgeler n-boyutlu uzayda birer yöney olarak temsil edilirler ve sorgu – belge benzerliğinin

bulunmasında yöney benzerliği esas alınır. Belge yöneyi ve sorgu yöneyi arasındaki benzerlik Denklem (5.4) ile hesaplanır.

$$ben(\vec{s}_k, \vec{b}_j) = \frac{\sum_{i=1}^n w_{i,k} \times w_{i,j}}{\sqrt{\sum_{i=1}^n w_{i,k}^2} \times \sqrt{\sum_{i=1}^n w_{i,j}^2}} \quad (5.4)$$

Sorgu ile her belge için benzerlik hesaplandıktan sonra en yüksek benzerliğin bulunduğu belge derlemde bulunan en ilgili belge olarak kabul edilir ve sonuçlar ilgiye göre azalan sırada sıralanarak kullanıcıya sunulur.

5.2 Yazılımın Kullanımı

5.2.1 Derlem Ekranı

Derlem ekranında, derlemde yer alan belgeler ve onlarla ilgili bilgiler vardır. Ekranın sol üst köşesinde derleme eklenmiş belgeler, sağda seçilen belgenin özellikleri, altta ise seçilen belgeyi temsil etmek için seçilen terimler ve tamlamalar ayrı ayrı görüntülenebilmektedir.

Derlem ekranı ilk açıldığında ekranda sadece derlemde yer alan belgeler gözükmemektedir. Kullanıcı bu belgelerden birini seçerek belgenin özelliklerini ve temsillerini görebilir. Kullanıcı bu bilgilerin görüntülenmesinin dışında ayrıca derleme yeni belgeler ekleyebilir. Kullanıcının tek yapması gereken “Yeni belge ekle” alanından “Gözet” düğmesini seçerek belgenin bilgisayarındaki yerini belirtmek ve sonrasında “Ekle” düğmesine basmaktır. Geri kalan işlemleri sistem kendiliğinden yapmaktadır.

5.2.2 Arama Ekranı

Arama Ekranı, kullanıcının istediği ölçütleri girerek derlemdeki belgeler arasında arama yapabileceği ekrandır. Kullanıcı, aramak istediği ölçütleri metin alanına girer ve “Ara” düğmesine basarak arama işlemini başlatır. Arama işlemi sonuçlandığında

sonular altta yer alan tabloda kullanıcının sorgusuna gre benzerlik sırasıyla yer alır. Sorguya en benzer belgeler en stte yer alır.

5.2.3 Ayarlar Ekranı

Kullanıcı, bilgi erişimi sisteminde kullanılan doğal dil işleme işlemlerinde kullanılan yöntemleri deęiştirebilir. Bu amaçla ayarlar ekranı kullanılmaktadır. Bu ekranda biçimbirimsel çözümleme sonuçlarından seçim yöntemleri, gövdeleme yaklaşımları ve terim – tamlama öncelik ayarları bulunmaktadır. Bu ayarlarla ilgili seçenekler şu şekildedir:

- Biçimbirimsel çözümleme sonuçlarından seçim yaklaşımları : *En kısa kök yöntemi, En uzun kök yöntemi, Kök sıklık yöntemi, Melez yöntem*
- Gövdeleme yaklaşımları: *Köke kadar gövdeleme, En uzun gövdeyi bulma*
- Terim – Tamlama ağırlıklandırmaları: Birbirinin bire tamamlayan iki ondalıklı sayı

5.2.4 Sınama Ekranı

Kullanıcı, sınama ekranını kullanarak bilgi erişimi sisteminden bağımsız olarak metinler üzerinde biçimbirimsel ve tamlama analizi yapabilir. Kullanıcı metin dosyasından yükledięi veya yazı alanına yerleştirdięi bir metin üzerinde bilgi erişim sisteminin doğal dil işleme yöntemlerini sınayabilir.

6. SONUÇLAR VE TARTIŞMA

Günümüzde, hem bilgi hem de bilginin artış hızı inanılmaz hızlarda artmaktadır. Bilgiyi saklanmak ve ilgili bilgiye ulaşmak bilgiye sahip olmak kadar önemli olmuştur. Bu durum Bilgi Erişimi konusunun ortaya çıkmasına ve internetin yaygınlaşması ile de daha da önem kazanmasına yol açmıştır. Bilgi Erişimi, bilginin temsil edilmesi, saklanması, düzenlenmesi ve gerektiği zamanda erişilebilmesini mümkün hale getirmek için yöntemlerin geliştirildiği araştırma konusudur. Bilgi Erişiminde, metin belgeleri üzerinde çalışırken dilbilimsel çeşitlilik ve anlamsal karmaşıklık Doğal Dil İşleme çalışmaları ile giderilebilmekte ve başarımlar üzerine iyileştirmeler elde edilebilmektedir.

Bu tez çalışması kapsamında Doğal Dil İşleme destekli bir Bilgi Erişim sistemi tasarlanmış ve gerçekleştirilmiştir. Gerçeklenen bu sistem Türkçenin biçimsel ve sözdizimsel farklılıklarını göz önünde bulundurarak Bilgi Erişimi başarımlarını arttırmayı hedeflemektedir. Biçimsel farklılıklar için biçimbirimsel çözümleme yardımıyla yapılan gövdeleme, sözdizimsel farklılıklar için tamlama analizinden faydalanılmıştır. Biçimsel farklılıkların giderilmesi ile anımsamanın arttığı, keskinliğin ise kimi durumlarda arttığı kimi durumlarda da azaldığı gözlenmiştir. Keskinliğin azalmasının temel sebebi aynı kök veya gövdeden türemiş farklı anlamlarda kullanılan sözcüklerden gövdeleme işlemi ile aynı köke ulaşılması ve birbiriyle ilgili olmayan belge veya sorguların ilgiliymişler gibi yorumlanmasıdır. Bu durum, gövdeleme yaklaşımlarının üzerinde çeşitli değişiklikler yapılarak azaltılmıştır.

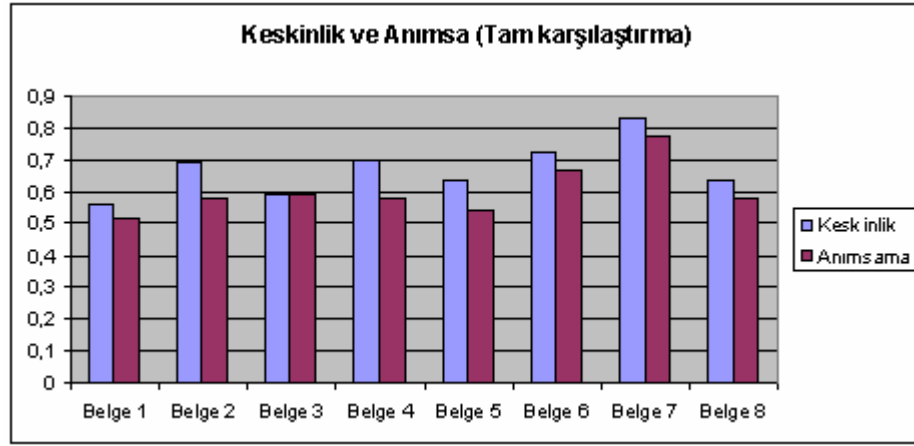
Sözdizimsel farklılıklar için tamlama analizi yapılmış ve bulunan tamlamalar tek sözcüklerle beraber dizinleme işleminde kullanılarak tamlama temelli dizinleme gerçekleştirilmiştir. Tamlama analizinde, belirtili, belirtisiz ve zincirleme isim tamlamaları ile sıfat tamlamaları incelenmiş ve sözcük türleri ve tamlama eklerine dayanarak bu tamlama çeşitleri için düzenli ifadeler ve sonlu durum makineleri oluşturulmuştur.

Tamlama analizinin başarımını ölçmek için derlemde rasgele belgeler seçilerek belgelerdeki tamlamalar farklı kişiler tarafından belirlenmiş ve tamlama analizi sonucunda bulunan tamlamalar ile karşılaştırılmıştır. Bu karşılaştırma yapılırken iki farklı karşılaştırma yöntemi kullanılmıştır : tam karşılaştırma, hata paylı karşılaştırma. Tam karşılaştırma yönteminde elle bulunan tamlamalar ile tamlama analizi sonucunda bulunan tamlamalar hiç hata payı olmadan, tamlamaların tüm elemanları göz önüne alınarak karşılaştırılır. Hata paylı karşılaştırmada ise üç ve daha fazla elemandan oluşan tamlamalarda bir elemanda hata olması durumunda da karşılaştırılan tamlamalar eşit sayılır. Bu karşılaştırmalar sonucunda her belge için keskinlik ve anımsama değerleri hesaplanmıştır.

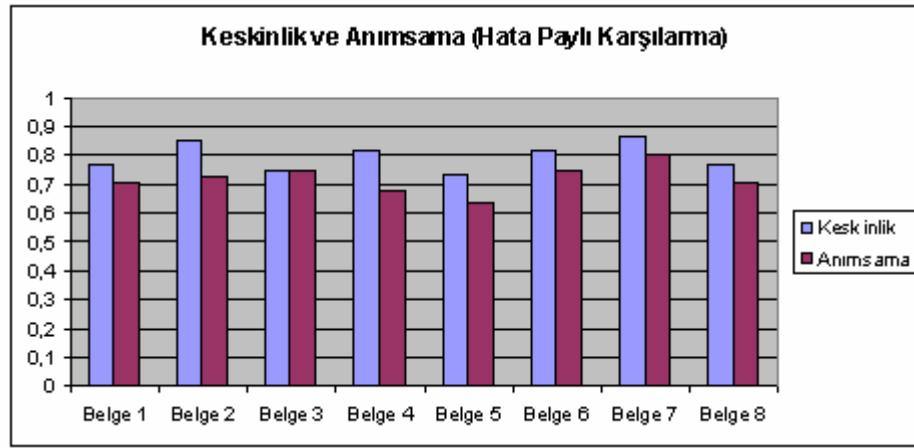
Derlemde seçilen 8 adet belgeden hem elle hem de geliştirilen tamlama analizi yöntemi ile tamlamalar çıkarılmış ve her iki yaklaşım ile sonuçlar karşılaştırılmıştır. Yapılan bu çalışmada tam karşılaştırma yöntemi ile ortalama 0.67 keskinlik ve 0.60 anımsama, hata paylı karşılaştırma ile ortalama 0.79 keskinlik ve 0.72 anımsama bulunmuştur. Bu 8 adet belge için farklı karşılaştırma yöntemleri ile hesaplanan keskinlik ve anımsama değerleri Tablo 6.1, keskinlik ve anımsama grafikleri Şekil 6.1 ve Şekil 6.2'deki gibidir.

Tablo 6.1: Örnek belgeler için hesaplanan keskinlik ve anımsama değerleri

	Tam Karşılaştırma		Hata Paylı Karşılaştırma	
	Keskinlik	Anımsama	Keskinlik	Anımsama
Belge 1	0,564	0,518	0,770	0,706
Belge 2	0,687	0,581	0,855	0,723
Belge 3	0,591	0,591	0,750	0,750
Belge 4	0,697	0,575	0,818	0,675
Belge 5	0,631	0,545	0,737	0,636
Belge 6	0,727	0,667	0,818	0,750
Belge 7	0,828	0,774	0,862	0,806
Belge 8	0,636	0,583	0,773	0,708



Şekil 6.1: Örnek belgeler için keskinlik ve anımsama grafiği (tam karşılaştırma)



Şekil 6.2: Örnek belgeler için keskinlik ve anımsama grafiği (hata paylı karşılaştırma)

Yukarıda belirtilen keskinlik ve anımsama değerleri değerlendirildiğinde ortalama 0.79 keskinlik, 0.72 anımsama değerleri yakalanmış olup bu veriler ışığında tasarımı yapılan dilbilimsel tamlama analizinin başarılı olduğu söylenebilir. Seçilen belgeler üzerinde yapılan çalışma sırasında keskinlik ve anımsama değerlerinin hesaplanması dışında yapılan analiz sonucundaki hataların nedenleri üzerinde de durulmuştur. Bu hatalar, Zemberek DDİ kütüphanesindeki hatalar, biçimbirimsel çözümleme ve sonrasındaki hatalar, üzerinde çalışılan belgenin özellikleri ve kapsanmayan durumlar gibi dört ana grupta toplanabilir. Zemberek DDİ kütüphanesindeki hataların bazıları geliştirme ve değiştirme uygun yapısı sayede düzeltilebilmiştir ve başarımlar artırılabilmiştir. Fakat bazı hataların çözümleri bulunamamıştır. Tamlama analizinin sonucunu en çok etkileyen etmen ise biçimbirimsel analiz sonucunda

sözcük için doğru çözümlemenin seçilememesidir. Örneğin; isim sözcüğü hem kökü olarak hem de is köküne –m sahiplik eki olarak çözümlenebilir. İlk çözümlemenin seçilmesi durumunda “isim tamlaması” tamlaması bulunabilirken ikincisinin seçilmesi durumunda bulunamamaktadır. Bu, hataların %60’ından fazlasının nedenidir. Diğer bir önemli etmen de üzerinde çalışılan belgenin özellikleridir. Belgede geçen sözcüklerin kök sözlüğünde bulunmaması durumunda çözümleme yapılamamakta ve bu sözcükler tamlama analizi dışında kalmaktadır. Özellikle teknik terimlerin ve yer adlarının sıkça geçtiği belgelerde tamlama analizinin başarısı ortalamadan düşük olmaktadır.

Örnek belgeler üzerinde yapılan çalışmada tamlama analizinde kapsanmayan durumlar da hata olarak kabul edilmiştir. Tamlama analizi, çeşitli durumları ve takısız isim tamlamalarını kapsamamaktadır. Takısız isim tamlamaları, tamlayan sözcüğün türünün belirsizliği nedeniyle analizin dışında bırakılmıştır. Bu sözcük türü belirsizliğinin giderilmesi için daha ileri çalışmalar yapılmalıdır.

Yapılan analiz, tamlayan ile tamlanan öğelerinin yer değiştirdiği durumları kullanımları çok yaygın olmadığı ve gerçekte tamlama olmayan sözcük öbeklerinin de tamlama olarak kabul edilmesine yol açabileceği için kapsamamaktadır. Örneğin, *Uçtu kafesinden kuşu gönlümün*. Bu sebepten dolayı bu durumların sıkça kullanıldığı şiir tipi yazılarda tamlama analizinin başarımı oldukça düşük olacaktır. Tamlayan ile tamlanan arasına giren ve tamlamanın ögesi olmayan sözcüklerin kullanımı sınırlı olarak tanınmaktadır. Eğer bu öğeler sıfat veya sıfat dizisi ise tanınabilmekte ve tamlamanın içerisine dahil edilebilmektedir.

Türkçede, belirtili isim tamlamalarında tamlananın sayı veya belirsizlik zamiri ise bazı hallerde tamlayan ekinin yerine -den eki kullanılır. Örneğin gençlerin üçü yerine gençlerden üçü gibi. Bu hallerde sözcüklerin arasında tamlama ilişkisinin bozulduğu varsayılarak bu sözcük öbekleri tamlama olarak değerlendirilmemiştir.

Türkçede kural olarak sıfat tamlamalarında sıfatlar ek alamazlar; ancak sıfat-fiillerle yapılmış sıfat tamlamalarında sıfat, yapım ve çekim eki alabilir. Yapılan analizde sıfat tamlamasının tamlayanı olan sıfat çekim eki almadığı varsayılmış ve bunun dışında kalan istisna durumlar tasarım dışında bırakılmıştır. Örneğin; yapılan tasarım ile *tanıdığım insanlar* tamlaması bulunamamaktadır. Bunun dışında sıfatların yapım

eki aldığı durumlar özellikle fiillerden yapım eki alıp sıfat oluşturdıkları durumlar ele alınmıştır.

Türkçede bazı durumlarda isim tamlamalarının tamlananı düşer ve tamlayana –ki eki gelir. Bu durum da tez kapsamında yapılan tamlama analizi dışında bırakılmıştır. Örneğin; “Kıvanç’ın tezi Ahmet’in tezinden daha zor.” cümlesindeki *Kıvanç’ın tezi* ve *Ahmet’in tezi* tamlamaları bulunabilirken “Kıvanç’ın tezi Ahmet’inkinden daha zor.” cümlesinde sadece *Kıvanç’ın tezi* tamlaması bulunabilir.

Türkçede birden fazla tamlanan olduğu ve son tamlananın bir bağlaç ile diğerlerinden ayrıldığı durumlarda bazen sadece tamlananlardan sonuncusu tamlama eklerini taşır. Örneğin; *TBMM’nin görev ve yetkileri*. Bu durum tamlama analizinin dışında bırakılmıştır.

Tamlama analizi sırasındaki hataların bir kısmının kaynağı olmasına rağmen Zemberek DDİ kütüphanesi yüksek performansı ve kapsadığı geniş konular ile tez çalışmasına önemli katkılarda bulunmuştur. Ayrıca kütüphanede yapılan değişiklikler ile tamlama analizinin başarımı % 25 civarında arttırılmıştır.

Bilgi Erişimi sistemi çerçevesinde yer alan tamlama temelli dizinleme için tek başına sözcükler ile birlikte tamlama analizi ile bulunan tamlamalar da belgelerin temsili için kullanılmıştır. Bu yaklaşımla aynı sözcüklerin farklı sözcük dizimleriyle oluşturdıkları anlam farklılıkları ele alınabilmiştir alınmaya çalışılmıştır. Dizinleme işleminde kullanılan elemanların ağırlıklandırılmasında tek sözcükler için terim sıklığı ve ters belge sıklığı değerleri, tamlamalar içinse tamlama elemanlarının ağırlıklarının ortalaması kullanılmıştır. Bu hesaplama ile tamlamanın ağırlığı ile tamlama elemanlarının ağırlıkları arasında bir ilişki kurulmuştur. Yapılan çalışmalar öncesinde teoride tamlama temelli dizinlemenin tek sözcük dizinlemesine oranla daha iyi sonuç vereceği tahmin edilmekteydi. Yapılan çalışmalar sonrasında elde edilen sonuçlar değerlendirildiğinde aksini gösterir sonuçlar elde edilmemesine rağmen iki dizinleme yönteminin karşılaştırılması eldeki derlemin yeterli olmaması nedeniyle yapılamamış ve niceliksel sonuçlara ulaşılmıştır. İki yöntemin karşılaştırılması için oluşturulan dar kapsamlı derlemin üzerinde yapılan çalışmalarda tamlama tabanlı dizinlemenin tek sözcük tabanlı dizinlemeye oranla az da olsa daha iyi sonuç verdiği gözlenmiştir.

Tez çalışmaları sırasında karşılaşılan en büyük sorun, Türkçe için yapılmış olan bilgi erişimi çalışmalarının belirli bir altyapıya oturtulmuş olmaması ve birbirlerinden habersiz oluşudur. Bunun sebebi Türkçe için TREC benzeri bir yapılanmanın olmamasıdır. Bu durum, Türkçe için geniş kapsamlı bir derlemin hala yaratılmamasındaki nedenlerden biridir.

Bu tezde tasarımı yapılan bilgi erişim sistemi, sözcüksel ve anlamsal farklılıkları da kapsayacak şekilde geliştirilebilir. Bu amaç doğrultusunda tasarıma ilgili birimler eklenip bir kavramsal sözcük veritabanı yardımıyla sorgu genişletme yapılabilir. Bu amaç doğrultusunda BalkaNet projesi kullanılabilir. Ayrıca tamlama analizinin sonucunu doğrudan etkileyen biçimbirimsel çözümleme sonuçlarının değerlendirilmesi ve doğru çözümlemenin seçilmesi konularına farklı yaklaşımlar getirilerek başarımın artması sağlanabilir.

KAYNAKLAR

- [1] **Lyman, P. and Varian, H.R.**, 2000. How much Information?, The Journal of Electronic Publishing, December 2000, Volume 6, Issue 2
- [2] **Baeza-Yates, R. and Ribeiro-Neto, B.**, 1999. Modern Information Retrieval, Addison-Wesley, England.
- [3] Dođal Dil İşleme, Vikipedi, Özgür Ansiklopedi. 22 Mart 2007 <http://tr.wikipedia.org/wiki/Do%C4%9Fal_dil_i%C5%9Fleme>
- [4] **Kesgin, F.**, 2007. Türkçe Metinler için Konu Belirleme Sistemi, *Yüksek Lisans Tezi*, İ.T.Ü. Fen Bilimleri Enstitüsü, İstanbul.
- [5] Türkiye Türkçesi, Wikipedia : Özgür Ansiklopedi, 22 Mart 2007 <http://tr.wikipedia.org/wiki/T%C3%BCrkiye_T%C3%BCrk%C3%A7esi>
- [6] **Ediskun, H.**, 1992. Türk Dilbilgisi. Remzi Kitapevi, İstanbul.
- [7] **Güzel, A.**, 2005. Üniversiteler için Türk Dili Ders Kitabı. Başkent Üniversitesi, Ankara.
- [8] Information Retrieval, Wikipedia: The Free Encyclopedia. , 3 Mart 2007 <http://en.wikipedia.org/wiki/Information_retrieval>
- [9] **Jurafsky, D. and Martin, J.**, 2000. Speech and Language Processing. Prentice-Hall, New Jersey.
- [10] **Maden, I., Demir, S., Özcan, E.**, 2003. Türkçe'den SQL Sorgularına Çeviri Yapan Bir Doğal Dil İşleme Uygulaması (NALAN-TS), *TBD 20. Ulusal Bilişim Kurultayı*, Aralık 2003.
- [11] **Sever, H. and Bitirim, Y.**, 2003. FindStem The Analysis and Evaluation of Stemming algorithms for Turkish, *10th International Symposium on String Processing and Information Retrieval (SPIRE'03)*, Manaus, Brazil, October 8-10, 2003, 238-251.
- [12] **Arampatzis, A. T., van der Weide, Th. P., Koster, C. H. A. and van Bommel, P.**, 2000. Linguistically Motivated Information Retrieval, *In Encyclopedia of Library and Information Science*. Marcel Dekker, Inc., New York, Basel.
- [13] **Jackson, P. and Moulinier, I.**, 2002. Natural language processing for online applications: text retrieval, extraction, and categorization, Amsterdam.

- [14] **Miller, G. A.**, 1995. WordNET : A Lexical Database for English. *Communications of the ACM*, **38**(11): 39-41.
- [15] **Kardeş, O., Güngör, T. and Kipman, E.**, 2003. Bir Uygulama Alanında Türkçe Metnin Anlambilimsel Gösterimi, *Akademik Bilişim Konferansı*, Çukurova Üniversitesi, Adana, February 2003.
- [16] **Porter, M.F.**, 1980, An algorithm for suffix stripping, *Program*, **14** (3) pp 130-137
- [17] **Krovetz, R.** 1993, Viewing morphology as an inference process, *SIGIR-93*, pp. 191-202, ACM.
- [18] **Orhan, Z., Altan, Z.**, 2006. Anlam Belirsizliği İçeren Türkçe Sözcüklerin Hesaplamalı Dilbilim Uygulamalarıyla Belirginleştirilmesi, XX. Ulusal Dilbilim Kurultayı, İstanbul, Mayıs 2006.
- [19] **Dinçer, T., Karaoğlu B.**, 2003. Stemming in Agglutinative Languages: A Probabilistic Stemmer for Turkish, *Proceedings of the Ninth Int. Symp. on Computer and Information Sciences*, Antalya, Turkey, November 1994, 244-251.
- [20] **Köksal A.**, 1981. Tümüyle Özdevimli Deneyisel Bir Belge Dizinleme ve Erişim Dizgesi, *TBD 3. Ulusal Bilişim Kurultayı*, Ankara, 6-8 Nisan, s. 37-44.
- [21] **Solak A., Can F.**, 1994. Effects of Stemming on Turkish Text Retrieval, *Proceedings of the Ninth Int. Symp. on Computer and Information Sciences*, Antalya, Turkey, November 1994, 49-56.
- [22] **Kut A., Alpkoçak A., Özkarahan E.**, 1995. Bilgi Bulma Sistemleri İçin Otomatik Türkçe Dizinleme Yöntemi, *Bilişim Bildirileri*, Dokuz Eylül Üniversitesi, İzmir, 247-253.
- [23] **Oflazer, K.**, 1994. Two-level Description of Turkish Morphology, *Literary and Linguistic Computing*, Vol. 9, No:2.
- [24] **Bilgin, O., Çetinoğlu, Ö., Oflazer, K.**, 2004. Building a Wordnet for Turkish, *Romanian Journal of Information Science and Technology*, Vol. 7, Num. 1-2.
- [25] BalkaNet Project Homepage, BalkaNet - Design and Development of a Multilingual Balkan WordNet. 30 Mart 2007 www.ceid.upatras.gr/Balkanet/
- [26] **Akın, A. A. and Akın, M. D.**, 2007. Zemberek, an open source NLP framework for Turkic languages.
- [27] **Fagan, J.L.**, 1987. Automatic Phrase Indexing for Document Retrieval: An Examination of Syntactic and Non-Syntactic Methods, *Proceedings of the Tenth Annual International ACM SIGIR Conference on Research*

and Development in Information Retrieval, New Orleans, Louisiana, USA, June 3-5, 1987, 91-101.

- [28] **Arampatzis, A. and Tsois, T. and Koster, C. and van der Weide, T.**, 1998. Phrase-based Information Retrieval, Technical Report CSIR9809, Computing Science Institute, University of Nijmegen, Nijmegen, The Netherlands.
- [29] Regular Expressions, Wikipedia: The Free Encyclopedia. , 5 Nisan 2007 <http://en.wikipedia.org/wiki/Regular_expression>
- [30] Java Technology, Sun Microsystems Inc., 7 Nisan 2007 <<http://java.sun.com>>
- [31] Welcome to NetBeans, Sun Microsystems Inc., 7 Nisan 2007 <<http://www.netbeans.org>>

EKLER

EK A

Tablo A.1: Durma sözcük listesi

bir	zaman	var	geçen	bin
ve	karşı	çünkü	tek	yok
bu	milyar	her	böyle	milyon
da	ya	o	biz	bazı
eğer	gün	bütün	ben	tüm
de	nasıl	ilk	yine	pek
için	oldu	son	kez	iki
ile	tarafından	ancak	elli	içinde
olarak	yer	değil	sadece	devam
daha	kendi	dedi	şimdi	ise
gibi	başka	diye	bunun	yapılan
artık	önemli	olduğunu	özel	diğer
sonra	hem	ki	şu	ortaya
olan	bile	mı	hiç	
ama	mi	önce	şöyle	
kadar	eden	yüzde	üzerine	
tam	aynı	olduğu	yaptığı	
ne	dün	göre	bugün	

ÖZGEÇMİŞ

1982 yılında Ankara’da doğan M. Kıvanç TÜRKEEŞ, 2000 yılında Beşiktaş Atatürk Anadolu Lisesi’nden, 2004 yılında da İstanbul Teknik Üniversitesi Elektrik – Elektronik Fakültesi Bilgisayar Mühendisliği bölümünden mezun olmuştur. M. Kıvanç TÜRKEEŞ, 2004 yılından beri özel sektörde yazılım uzmanı olarak çalışmaktadır.