İSTANBUL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZİ

TÜRKÇE METİNLERDEKİ ANLAM BELİRSİZLİĞİ OLAN SÖZCÜKLERİN BİLGİSAYAR ALGORİTMALARI İLE ANLAM BELİRGİNLEŞTİRMESİ

Zeynep ORHAN
Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Programı

Danışman
Doç. Dr. Sabri ARIK
Kasım, 2006

İSTANBUL

ÖZET

TÜRKÇE METİNLERDEKİ ANLAM BELİRSİZLİĞİ OLAN SÖZCÜKLERİN BİLGİSAYAR ALGORİTMALARI İLE ANLAM BELİRGİNLEŞTİRMESİ

Sözcük anlamı belirginleştirme, anlam belirsizliği olan sözcüğün belirli bir kullanım alanında en uygun anlamının kullanıldığı tümcedeki diğer elemanları da göz önüne alarak belirlenmesi işlemidir.

Bu çalışmada, Türkçe metinlerde anlam belirsizliği olan sözcüklerin anlamlarının belirginleştirilmesini sağlayacak en uygun algoritmaların ve özelliklerin belirlenmesi ele alınmıştır. Türkçe için SAB çalışmalarında kullanılabilecek sözcükler ve anlam sınıfları oluşturulmuş, algoritmalarda kullanılacak metinlerin elle işaretlenmesi gerçekleştirilmiş ve kavramsal bir sözlük hazırlanarak bu alanda yapılacak araştırmalara önemli bir katkıda bulunulmuştur.

İlk bölümlerde öncelikle doğal dil işleme çalışmaları genel olarak ele alınmış ve buna paralel olarak Türkçe doğal dil işleme çalışmaları hakkında bilgi verilmiştir. Doğal dil işlemede SAB uygulama alanları, kullanılan bilgi tipleri ve kaynakları, uygulanan yöntemler ve karşılaşılan problemler incelenmiştir. Çalışma ile yakından ilişkili olan Senseval projesinin amacı, uygulamaları ve elde edilen sonuçları ayrıntılı olarak irdelenmiştir.

Yapılan çalışmanın aşamaları içerisinde Türkçe için derleme metin seçimi, sözcük anlamlarının oluşturulması, sözcük anlamlarına etki eden özelliklerin seçimi, kullanılan yöntemler, yazılımlar ve değerlendirme kriterleri hakkında bilgi verilmiştir.

Son üç yıl içerisinde çalışmanın bütün aşamalarında elde edilen sonuçlar grafikler ve tablolar şeklinde sunulmuştur. En son bölümde sonuçların ifade edildiği ve yorumlandığı, gelecekte yapılabilecek yeni çalışmalar üzerinde durulduğu genel bir değerlendirme bulunmaktadır.

SUMMARY

SENSE DISAMBIGUATION OF AMBIGUOUS WORDS IN TURKISH TEXTS BY MACHINE LEARNING ALGORITHMS

Word sense disambiguation is the process of selecting the most suitable sense of an ambiguous word in the given sentence by considering the other contextual features

In this study, determination of the most convenient algorithms and features that may lead to the successful disambiguation of the ambiguous words in Turkish texts have been discussed. Ambiguous words and their sense classifications that can be used for Turkish word sense disambiguation studies have been established, a limited ontology has been prepared and by providing manually sense tagged corpora, an important contribution has been achieved for the researches in this domain.

In the first chapters of the thesis, a brief introduction for natural language processing has been given and in parallel to this topic, important work on Turkish have been summarized. Then, the application areas of word sense disambiguation in natural language processing, knowledge types and sources, approaches in the literature and the problems of word sense disambiguation have been examined in the following sections. The objectives, applications and the results of the Senseval project, which is closely related to the thesis topic, have been exhaustively scrutinized

Corpora selection, sense classification, effective features determination, tested algorithms, software and evaluation criterion that have been utilized during the phases of the study have been explained.

The results obtained in the last three years from all phases of the study have been presented by graphics and tables. In the last section, a general evaluation and conclusion have been provided for commenting on the results and future work.

ÖNSÖZ

Yüksek lisans öğrenimim sırasında ve tez çalışmalarım boyunca gösterdiği her türlü destek ve yardımdan dolayı çok değerli hocam Doç. Dr. Sabri Arık'a, Yard. Doç. Dr. Zeynep Altan'a en içten dileklerimle teşekkür ederim. Tez izleme komitemde bulunmayı kabul edip çalışmamın ilerlemesinde katkıda bulunan Prof. Dr. Bülent Örencik'e, Prof. Dr. Osman Nuri Uçan'a da teşekkür ederim.

Bu çalışma boyunca yardımlarını esirgemeyen çalışma arkadaşlarıma ve İstanbul Üniversitesi'ne teşekkürü borç bilirim.

Her zaman yanımda olan ve beni her konuda yüreklendiren, destek olan aileme de teşekkür ederim.

Kasım, 2006

Zeynep ORHAN

İÇİNDEKİLER

ŞEKİL LİSTESİ

TABLO LİSTESİ

KISALTMALAR

A	**:**	All-words
AA	**:**	Anlamsal Ağlar
ACL	**:**	Association for Computational Linguistic
AODE	**:**	Aggregating One-Dependence Estimators
ARPA	**:**	Advanced Research Projects Agency
BÇ	**:**	Bilgisayarlı Çeviri
BÇK	**:**	Bilgi Çıkarımı
BİT	**:**	Bilgi Tabanlı
BNC	**:**	British National Corpus
BÖ	**:**	Bilgisayarla Öğrenme
BTÖ	**:**	Bellek Tabanlı Öğrenme
CART	**:**	Classification and Regression Trees
CB	**:**	Case-Based
CFG	**:**	Context Free Grammar-Bağlamdan Bağımsız Gramer
CM	**:**	Confussion matrix
CV	**:**	Cross Validation
DARPA	**:**	Defense Advanced Research Projects Agency
DMD	**:**	Derleme Metin Dilbilimi
DDA	**:**	Doğal Dil Anlama
DDİ	**:**	Doğal Dil İşleme
DDÜ	**:**	Doğal Dil Üretme
DM	**:**	Dil Mühendisliği
DT	**:**	Dil Teknolojisi
DMT	**:**	Derleme Metin Tabanlı
DTr	**:**	Decision trees
EB	**:**	Exemplar-based
EPSRC	**:**	The Engineering and Physical Sciences Research Council
ES	**:**	Elektronik Sözlükler
FN	**:**	False Negative(Yanlış negatif)
FP	**:**	False Positive(Yanlış pozitif)
HD	**:**	Hesaplamalı Dilbilim
IB	**:**	Instance Based
ID3	**:**	Induction Decision Tree
IG	**:**	Inflectional Groups
LBR	**:**	Lazy Bayesian Rules
LCSL	**:**	Laboratory for the Computational Studies of Language
LDOCE	**:**	Longman Dictionary of Contemporary English
LL	**:**	Lazy Learning
LOB	**:**	Lancaster-Oslo-Bergen
MBL	**:**	Memory Based Learning
MUC	**:**	Message Understanding Conferences

NB	**:**	Naive Bayes
O	**:**	Other-training.
OMWE	**:**	Open Mind Word Expert
OÖ	**:**	Otomatik Özetleme
OUP	**:**	Oxford University Press and Digital
P	**:**	Precision
R	**:**	Recall
S	**:**	Supervised-training
SAB	**:**	Sözcük Anlamı Belirginleştirme
SB	**:**	Similarity-based
SIGLEX	**:**	Special Interest Group on the Lexicon
TAN	**:**	Tree Augmented Naive Bayes
TDK	**:**	Türk Dil Kurumu
TEI	**:**	Text Encoding Initiative
TN	**:**	True Negative(Doğru Negatif)
TP	**:**	True Positive(Doğru Pozitif)
TREC	**:**	The Text Retrieval Conference
WEKA	**:**	Waikato Environment for Knowledge Analysis
WSD	**:**	Word Sense Disambiguation
YSA	**:**	Yapay Sinir Ağları
YZ	**:**	Yapay Zeka
YZT	**:**	Yapay Zeka Tabanlı

ÖZET

TÜRKÇE METİNLERDEKİ ANLAM BELİRSİZLİĞİ OLAN SÖZCÜKLERİN BİLGİSAYAR ALGORİTMALARI İLE ANLAM BELİRGİNLEŞTİRMESİ

Sözcük anlamı belirginleştirme, anlam belirsizliği olan sözcüğün belirli bir kullanım alanında en uygun anlamının kullanıldığı tümcedeki diğer elemanları da göz önüne alarak belirlenmesi işlemidir.

Bu çalışmada, Türkçe metinlerde anlam belirsizliği olan sözcüklerin anlamlarının belirginleştirilmesini sağlayacak en uygun algoritmaların ve özelliklerin belirlenmesi ele alınmıştır. Türkçe için SAB çalışmalarında kullanılabilecek sözcükler ve anlam sınıfları oluşturulmuş, algoritmalarda kullanılacak metinlerin elle işaretlenmesi gerçekleştirilmiş ve kavramsal bir sözlük hazırlanarak bu alanda yapılacak araştırmalara önemli bir katkıda bulunulmuştur.

İlk bölümlerde öncelikle doğal dil işleme çalışmaları genel olarak ele alınmış ve buna paralel olarak Türkçe doğal dil işleme çalışmaları hakkında bilgi verilmiştir. Doğal dil işlemede SAB uygulama alanları, kullanılan bilgi tipleri ve kaynakları, uygulanan yöntemler ve karşılaşılan problemler incelenmiştir. Çalışma ile yakından ilişkili olan Senseval projesinin amacı, uygulamaları ve elde edilen sonuçları ayrıntılı olarak irdelenmiştir.

Yapılan çalışmanın aşamaları içerisinde Türkçe için derleme metin seçimi, sözcük anlamlarının oluşturulması, sözcük anlamlarına etki eden özelliklerin seçimi, kullanılan yöntemler, yazılımlar ve değerlendirme kriterleri hakkında bilgi verilmiştir.

Son üç yıl içerisinde çalışmanın bütün aşamalarında elde edilen sonuçlar grafikler ve tablolar şeklinde sunulmuştur. En son bölümde sonuçların ifade edildiği ve yorumlandığı, gelecekte yapılabilecek yeni çalışmalar üzerinde durulduğu genel bir değerlendirme bulunmaktadır.

SUMMARY

SENSE DISAMBIGUATION OF AMBIGUOUS WORDS IN TURKISH TEXTS BY MACHINE LEARNING ALGORITHMS

Word sense disambiguation is the process of selecting the most suitable sense of an ambiguous word in the given sentence by considering the other contextual features

In this study, determination of the most convenient algorithms and features that may lead to the successful disambiguation of the ambiguous words in Turkish texts have been discussed. Ambiguous words and their sense classifications that can be used for Turkish word sense disambiguation studies have been established, a limited ontology has been prepared and by providing manually sense tagged corpora, an important contribution has been achieved for the researches in this domain.

In the first chapters of the thesis, a brief introduction for natural language processing has been given and in parallel to this topic, important work on Turkish have been summarized. Then, the application areas of word sense disambiguation in natural language processing, knowledge types and sources, approaches in the literature and the problems of word sense disambiguation have been examined in the following sections. The objectives, applications and the results of the Senseval project, which is closely related to the thesis topic, have been exhaustively scrutinized

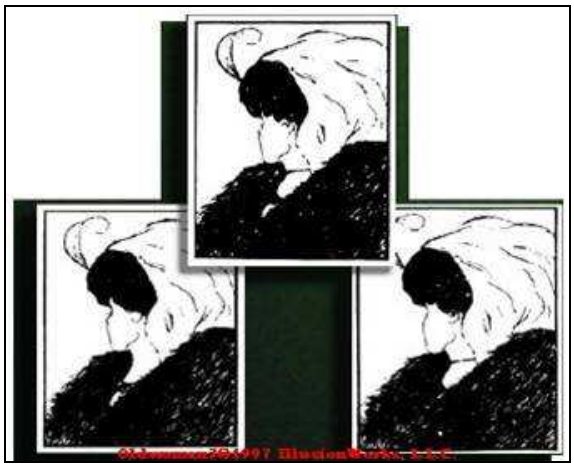
Corpora selection, sense classification, effective features determination, tested algorithms, software and evaluation criterion that have been utilized during the phases of the study have been explained.

The results obtained in the last three years from all phases of the study have been presented by graphics and tables. In the last section, a general evaluation and conclusion have been provided for commenting on the results and future work.

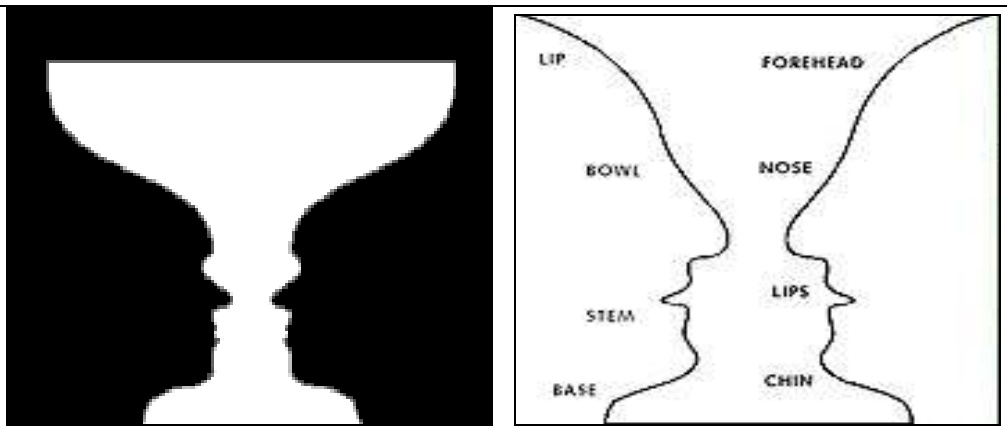
1. GİRİŞ

Doğal Dil İşleme (DDİ) otomatik araçlar kullanarak metin işlemek ve bu metnin sözdizimsel ve anlamsal yapısına ait çıkarımlar yapmak gibi, doğal dillere özgü çalışmalardır. Bu tip çalışmalar dilin yapısını anlamak, metinden bazı özel bilgiler elde etmek, bilgisayarla çevirileri etkinleştirmek, otomatik özet çıkarmak gibi amaçlarla yapılmaktadır.

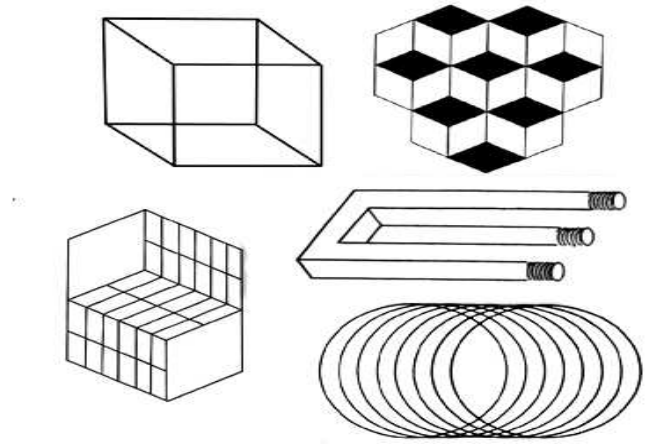
Belirsizlik (ambiguity) DDİ'de tümce öğelerinin çözümlenmesinden, birden fazla anlam içeren sözcüklerin anlamlarını belirlemeye kadar pek çok uygulamada çözülmesi gereken en önemli sorunlardan biridir. Belirsizlik kavramı sadece DDİ çalışmalarında değil, diğer bir çok alanda da karşılaşılan bir problemdir. Uygulandıkları yerlere göre de belirsizliğin giderilme yöntemleri farklı olmaktadır. Sözlük anlamına göre belirsizlik, birden fazla anlamı olma veya birden fazla şekilde yorumlanabilme kavramına karşılık gelir. Bu nedenle de, belirsizlik içeren herhangi bir durum bağlama göre birden fazla sonuç içerebilir; bu sonuçlardan ancak bir tanesi o bağlamda doğru olacaktır. Sonuçlardan bir diğeri ise, başka bir bağlamda doğru olabilir. Örneğin, Şekil 1.1'de görsel belirsizlikle ilgili bazı resimler görülmektedir. Resimlere bakan farklı iki kişi bu resimlerdeki yaşlı veya genç kadını, vazoyu veya birbirine bakan iki insanı ya da en son örnekte verildiği gibi şekillerin derinliğini farklı olarak görebilir. Daha da ilginç olarak; bu resimlere bakan aynı kişi olsa bile, farklı zaman dilimlerinde (hatta bir iki saniye gibi çok küçük aralıklarda) bu resimleri farklı algılayabilir; bu algıların hiçbirinin yanlış olduğu söylenemez.



(a)



(b)



(c)

Şekil 1.1: Görsel belirsizlik örnekleri

Yukarıdaki resim ve şekillerdeki belirsizliklere benzer olarak yazılı veya sözlü metinlerde de bu tür durumlarla sık sık karşılaşılır. Örneğin,

Ali Ayşe'ye defterini vermesini söyledi

tümcesinde verilmesi istenen defterin Ali'nin mi, yoksa Ayşe'nin mi olduğu açık değildir. Eğer tümce

Ali Ali'ye defterini vermesini söyledi

olarak verilirse, belirsizlik daha açık bir biçimde görülebilir; burada ortamda kaç tane Ali olduğunu ve hangisinin defterini kimin istediğini anlamak gerekmektedir. Konuşma, yazı ve resimlerde ortaya çıkan bu belirsizlikler için, bağlama göre farklı belirginleştirmeler doğru kabul edilebilir. Ancak kesinlik gereken diğer bazı çalışma alanlarında herhangi bir durumdaki belirsizliği tamamen çözecek kuralların tanımlanmış olması gerekir. Örneğin, matematiksel bir ifade olan

*X=3*7+12/3*

eşitliği için X değeri olarak

(3*7)+(12/3)=25,
(((3*7)+12)/3)=11,
(3*(7+(12/3)))= 33,
(3*(7+12))/3=19

şeklinde farklı sonuçlar bulunabilir. Ancak; matematikteki işleçlerin öncelikleri kuralı kullanılarak bu ifadenin sonucunun 25 olduğu kesin olarak söylenir ve bu sonuç her yerde aynıdır.

Programlama dillerinde de belirsizliklerin çözülebilmesi için benzer kurallar tanımlanmıştır. Verilen bir program bazı belirsizlik durumları içerse de, bunlar programlama dillerinin işleme kuralları sayesinde çözümlenir ve sonuç hep aynı çıkar. Örneğin,

```
{
    int ali = 3;
    {
        int ali = 4;
        cout << ali;
    }
}
```

biçiminde yazılan bir C++ kodunda ekrana yazılacak değer *4* olacaktır. Diğer pek çok programlama dilinde olduğu gibi, C++ dilinin kurallarına göre değişkenin değeri en yakın tanımlanan değişkenden alınır. Matematik ve programlama gibi alanlarda belirsizlikleri gidermek ve kesin çözüme ulaşmak için kurallar tanımlanabilmiştir. Ancak konuşma, yazı, resim gibi pek çok alanda bu belirsizlikleri çözebilecek kesin kurallar henüz bulunamamış olmasına rağmen, insanlar bunları bağlama göre çok kolay ve başarılı bir biçimde çözümleyebilmektedir. Örneğin,

Geldi.
Göze **geldi.**
Göz göze **geldi.**

tümcelerinde geçen *geldi* sözcüğü, her bir tümcede farklı anlamlarda kullanılmıştır. Sadece *geldi* sözcüğünü duyan ya da okuyan bir kişi için bu tümce genellikle *birinin veya bir şeyin bir yere ulaştığı* anlamına gelir. Ancak *göze geldi* şeklinde kullanılan tümce, birinci tümceye çok benzemesine karşın anlamı tamamen farklı algılanacak ve *başa kötü bir şeyin gelmesi, nazar değmesi* olarak yorumlanacaktır. Üçüncü tümcenin sözcükleri ise birinci ve ikinci tümceyi içermesine karşın, buradaki anlam diğer iki tümceden farklı olarak *biriyle karşılaşma* biçiminde anlaşılacaktır. Bu örneklerde de görüldüğü gibi herhangi bir sözcük kullanıldığı yere göre oldukça farklı anlamlara sahip olabilmektedir.

Sözcük anlamı belirginleştirme (SAB) İngilizce word sense disambiguation (WSD) kavramına karşılık olarak kullanılmıştır. Aynı kavram için sözcük anlamı belirleme/kestirilmesi/indirgeme gibi karşılıklar da seçenek olarak kullanılabilir. SAB 1950'li yıllarda bilgisayarla DDİ çalışmalarının başlamasıyla birlikte ilgi çekmiş ve araştırılmaya başlanmış bir problemdir. SAB, hesaplamalı dilbilim (HD) alanında oldukça ilgi görmüş ve çok araştırma yapılmıştır. Ancak SAB alanındaki çalışmalar henüz tam olarak bir olgunluk kazanmadığı gibi, tanımında bile farklılıklar bulunmaktadır. SAB problemi genel olarak ele alındığında şu çıkarımlar yapılabilir:

- Pek çok sözcüğün hemen hemen bütün doğal dillerde birden fazla anlamı vardır. Bir sözcüğün nerede kullanıldığına bakılmaksızın, hangi anlamda olduğu konusunda genellikle bir belirsizlik mevcuttur.
- SAB araştırmalarının genel amacı sözlük, eşanlamlılar sözlüğü veya buna benzer bir kaynaktaki farklı anlamlar arasındaki belirsizliği çözümlemektir.
- Herhangi bir kişi, içinde anlam belirsizliği olan bir tümceyi anladığı zaman, belirsizliğe neden olan sözcüğün diğer anlamlarını elemiş ve sözcüğün sadece bir anlamını göz önüne almıştır. İnsanoğlu bu anlama işlemini gerçekleştiren bilişsel bir sisteme sahiptir. Belirsizlik içeren bir tümcenin anlaşılması demek "i*nsan dil anlama işleminde olası anlam kümesi içinden uygun anlamın seçilmesi"* demektir (Kilgarriff, 1999).

SAB, kısaca insan - dil anlama işleminin bir parçası olarak tanımlanabilir. Bu işlem bir SAB programı içerisinde modellenebilirse, pek çok DDİ uygulaması için gerekli olan önemli bir fonksiyon gerçekleştirilmiştir. Bu görüş Cottrell tarafından şu tümce ile çok net bir şekilde ifade edilmiştir:

"Sözcüksel belirsizlik Doğal Dil Anlama (DDA) sistemlerinin karşılaştığı en önemli problemlerdendir. Kullanılan sözcüklerin doğru anlamını belirleme DDA'nın temelidir. İnsanların belirsizliği çözme mekanizmasını anlamak çok önemlidi; çünkü bu işi yapma yöntemleri her ne ise gerçekten çok başarılı olmaktadır." (Cottrell, 1989).

Sonuç olarak SAB belirli bir kullanım alanında, anlam belirsizliği olan sözcüğün hangi anlamının seçildiğini, kullanıldığı tümcedeki diğer elemanları da göz önüne alarak bulmaya çalışma işlemidir. Bir sözcüğün bir kaynakta belirtilmiş sonlu sayıda farklı

anlamı vardır ve bir SAB programının işi kullanıldığı yere göre bu sözcüğün bir anlamını, daha doğrusu en uygun anlamını seçmektir.

Günümüz dünyasının oluşumunda çok önemli bir yapı taşı olan DDİ teknolojilerinin diğer teknik gelişmelerden ayrılan bir yönü vardır. Bu çalışmaların, her dil için o dili anadili olarak kullananlar tarafından uyarlanması önemlidir. Türkçe için Türk dilini konuşan araştırmacıların gerçekleştireceği sistemler çok daha verimli olacaktır. Dünyada yaklaşık üç yüz milyon kişinin Türkçe konuştuğu düşünüldüğünde ve Türk dilinin değişik lehçeleri de göz önüne alındığında, yapılacak işlerin yoğunluğu ve bunun karşılığındaki getirinin heyecan verici olacağı açıktır.

SAB çalışmalarının doğal dillerin bilgisayar ortamında incelendiği HD alanındaki önemi nedeni ile bu çalışmanın amacı, Türkçe metinlerde anlam belirsizliği olan sözcüklerin anlamlarının bilgisayar yardımıyla belirginleştirilmesini sağlayacak algoritmaları belirleme, bu algoritmaları test etme ve elde edilen sonuçları karşılaştırarak Türkçe için en uygun yöntemlerin hangileri olabileceğine karar vermektir. Ayrıca Türkçe için SAB çalışmalarında kullanılacak sözcükler ve anlam sınıfları oluşturularak, algoritmalarda kullanılacak metinlerin de elle işaretlenmesi gerçekleştirilmektedir. Ayrıca kullanılan metinlerde eylemlerle birlikte geçen sözcükler için kavramsal bir sözlük eklenmiştir.

SAB konusunda genel bilgilerin verildiği ilk bölümden sonra, İkinci Bölüm'de konunun anlaşılmasını kolaylaştırmak amacı ile, doğal dil işleme çalışmaları genel olarak ele alınmış ve buna paralel olarak Türkçe doğal dil işleme çalışmaları hakkında bilgi verilmiş; SAB çalışmasının ayrıntıları içinde SAB uygulama alanları, kullanılan bilgi tipleri ve kaynakları, uygulanan yöntemler ve karşılaşılan problemler incelenmiştir. Ayrıca İkinci Bölüm çalışma ile yakından ilişkili olan Senseval projesinin amacını, uygulamalarını ve elde edilen sonuçları Senseval 1, Senseval 2 ve Senseval 3 projeleri bağlamında ayrıntılı olarak irdelemektedir.

Yapılan çalışmanın aşamaları ile ilgili bilgiler Üçüncü Bölüm'de anlatılmaktadır. Derleme metin seçimi, sözcük anlamlarının oluşturulması, sözcük anlamlarına etki eden

özelliklerin seçimi, kullanılan yöntemler, yazılımlar ve değerlendirme kriterleri bu bölümde incelenmektedir.

Dördüncü Bölüm'de çalışmanın bütün aşamalarında elde edilen sonuçlar grafikler ve tablolar şeklinde sunulmuştur. Bu çalışmadan elde edilen sonuçların ifade edildiği ve yorumlandığı, gelecekte yapılabilecek yeni çalışmalar üzerinde durulduğu bölüm ise son bölüm olarak Beşinci Bölüm'dür.

2. GENEL KISIMLAR

2.1. DOĞAL DİL İŞLEME

Matematik alanındaki gelişmeler bilgisayar dünyasını da derinden etkilemiştir. Yüzyıllardır üzerinde çalışılan felsefe ve matematiğin mantık bilimi disiplini ile birlikteliği, günümüzde yapay zeka (YZ) çalışmaları ile daha da anlam kazanmaktadır. YZ kavramından yararlanarak çözülecek problemler güçlü mantıksal modellerle desteklendikleri ölçüde başarılı olacaktır. YZ alanındaki 20-25 yıl gibi oldukça kısa sayılabilecek bir zaman öncesinde başlayan araştırmaların, yakın bir gelecekte ulaşacağı sonuçların insan hayatını doğrudan etkilemesi kaçınılmazdır.

Alan Turing'in II. Dünya Savaşı sırasında Almanların şifreli haberleşmesini çözmek için yaptığı çalışmalar sonlu durum makineleri ve Turing makinelerinin temelini oluşturmuş, ortaya atılan programlanabilir makine fikri ile çağımızın en önemli gelişmesi olan bilgisayarlar ortaya çıkmıştır. Turing'in bu çalışmalarının YZ ve DDİ konularının da temelini oluşturduğu kabul edilir. Aslında, YZ kavramı ile ulaşılmak istenen hedef, insanda var olan insan yetisinin özgününden ayırt edilemeyecek bir şekilde benzetiminin gerçekleştirilmesidir. Yine Turing tarafından gerçekleştirilen Turing testi, bu hedefin ilk defa ortaya konulduğu uygulama olarak kabul edilmiştir. Turing testine göre; *"eğer görmeden konuşulan bir makine ile insan birbirinden ayırt edilemiyorsa, makine insanı mükemmel şekilde taklit etmektedir"* denilebilir. Turing testinden sonra bunu izleyen benzeri pek çok çalışma yapılmıştır. Örneğin, 1966 yılında Joseph Weizenbaum'un tasarladığı ELIZA adlı program çok basit kalıpları tanıyarak kullanıcısına psikolojik terapi uygulayabiliyordu (Weizenbaum, 1966). Bu program en çok tanınan ve en eski olarak kabul edilen YZ programıdır. Programın amacı, bir psikanalist ile hastası arasındaki konuşmaları makineyle basit bazı analizler yaparak sağlamaktır.

Başlangıçta YZ'nın küçük bir kolu olarak kabul edilen DDİ, kısa sürede uygulama alanlarını genişleterek tek başına bir disiplin olmuştur. Son yıllarda bilişim alanında gerçekleştirilen önemli gelişmeler, bilgiyi ön plana çıkararak bilgi toplumu kavramını doğurmuştur. Günümüzde özellikle İnternet üzerinden istenilen bilgiye ulaşma ve iletişim çok kolaylaşmıştır. Buna paralel olarak da bilgi ve teknolojinin farklı birey ve toplumlar arasında paylaşımını kolaylaştırma, diller arası çeviri veya belge indeksleme/bulma gibi geniş kapsamlı öte veri (meta data) içeren araştırmalar için daha gelişmiş dil işleme yazılımları gereksinimi ortaya çıkmıştır. Bilgisayarlar yardımıyla kolaylaşacak ve hızlanacak olan bu iletişim sayesinde iş dünyasında ve uluslararası ilişkilerde de önemli değişimler görülecektir. Ancak burada önemli bir sorun insan-makine iletişimidir.

Yukarıdaki örnekler İnternet dünyası için de arttırılabilir. 1996 yılındaki bir istatistik günümüz bilgi dağarcığını çok iyi açıklayacaktır. Her gün ortalama yirmi milyon sözcüklük teknik bilgi elektronik ortama aktarılmaktadır. Bir insanın dakikada ortalama bin sözcük okuyabileceği düşünülürse, bir günde eklenen bu bilginin okunabilmesi için gereken zaman günde sekiz saatten bir buçuk ay olarak hesaplanır. Ancak bu sürede eklenen yeni bilgiler de düşünüldüğünde, günlük bilgiyi okurken eklenen yeni bilgiler için de beş buçuk yıl gerektiği sonucu ortaya çıkar. Kısaca, normal standartlarda bir insanın elektronik ortamdaki yenilikleri takip edebilmesi için mutlaka başka yardımcılara gereksinimi vardır (Bird, 1996). İnsana güç kazandıran çok büyük bir bilgi okyanusuna sahip olmak değil, bu bilgi okyanusundan gereği gibi faydalanabilmektir. En üst düzeyde bu bilgi kaynağından faydalanmayı sağlayacak yardımcılardan biri DDİ'nin temel alt birimlerinden olan biçimbilimsel çözümleyicilerdir (morphological analyzers). Biçimbilimsel çözümleyiciler genellikle aşağıdaki alanlarda kullanılırlar:

- Sözcük işleme programlarında yazım düzeltme işlemlerinde,
- Sesten yazıya ve yazıdan sese çevrimlerde,
- Arama motorlarında indeksleme işlemi yapılırken (özellikle eklemeli dillerin sözcüklerinde),
- Sorgulamalardaki yazım yanlışlarını düzeltmede,
- *Türkiye Büyük Millet Meclisi* gibi sözcük öbeklerini algılamada, aranan sözcüklerin eşanlamlılarının sorguya dahil edilmesinde,
- Eklemeli dillerde kök bulunmasında.

Kısaca özetlenirse; DDİ, derleme metin dilbilimi (DMD), HD, dil mühendisliği (DM), dil teknolojisi (DT) gibi yeni bilim dallarının ortaya çıkmasına neden olan ve tüm bu alt dallarla bilginin tamlığının sağlandığı bir disiplindir.

2.1.1. Doğal Dil İşlemenin Amaçları

Doğal dil, Türkçe, İngilizce, Almanca, gibi insanların iletişim için kullandığı herhangi bir dil olup, yapay olarak insanlar tarafından geliştirilen programlama dillerinden farklıdır. DDİ ise doğal dili bilgisayarlarla işlemek için yapılan çalışmaların bütünü olarak tanımlanabilir.

Bilgisayarlar ve insanlar kendilerine verilen bir metni veya konuşmayı veri olarak aynı şekilde almalarına rağmen; bunlardan anladıkları birbirinden çok farklı olmaktadır. Bunun en önemli nedeni insanların yaşamları boyunca kazandıkları tecrübelerle iletişim için geliştirmiş oldukları ortak bilgi birikimi, çıkarım yapabilme yeteneği ve tecrübelerini kullanabilme yetisidir. Oysa bilgisayarlar sadece belirli donanım ve yazılım birimleri içerir; bunun dışında insanlarda bulunan potansiyel yetenek ve yetilere sahip değildir. Belirli işlemleri yerine getirebilmeleri ancak onlara bu işlemleri nasıl yapabileceklerinin öğretilmesi ile sınırlı olarak mümkündür; çünkü buradaki öğrenme kavramı insanlardaki öğrenmeden farklıdır ve daha çok bilgisayar dilinde verilen komutların belli bir sıra dahilinde işlenmesini sağlayarak sonuçlara ulaşmak anlamındadır.

Dil yeteneği, beynin çalışması konusunda önemli ipuçları verebilecek insan türüne özgü bir özelliktir. Bu nedenle de dilbilim, bilişsel bilimler alanında önemli bir yer tutar. İnsanoğlunun bu tür özelliklerinin bilgisayar ortamında modellenmesinin oldukça yararlı sonuçları olduğu gözlenmektedir.

Günümüzde Avrupa, Amerika ve Japonya gibi dünyanın bilgi işleme ve iletişimi konularına oldukça fazla önem veren ülkelerinde DDİ alanına yönelik çok büyük yatırımlar yapılmakta ve bunun sonucunda da kullanıcılar için önemli yarar ve üstünlükler sağlayan yazılımlar ve bilgisayar sistemleri geliştirilmektedir. Bu konu ile

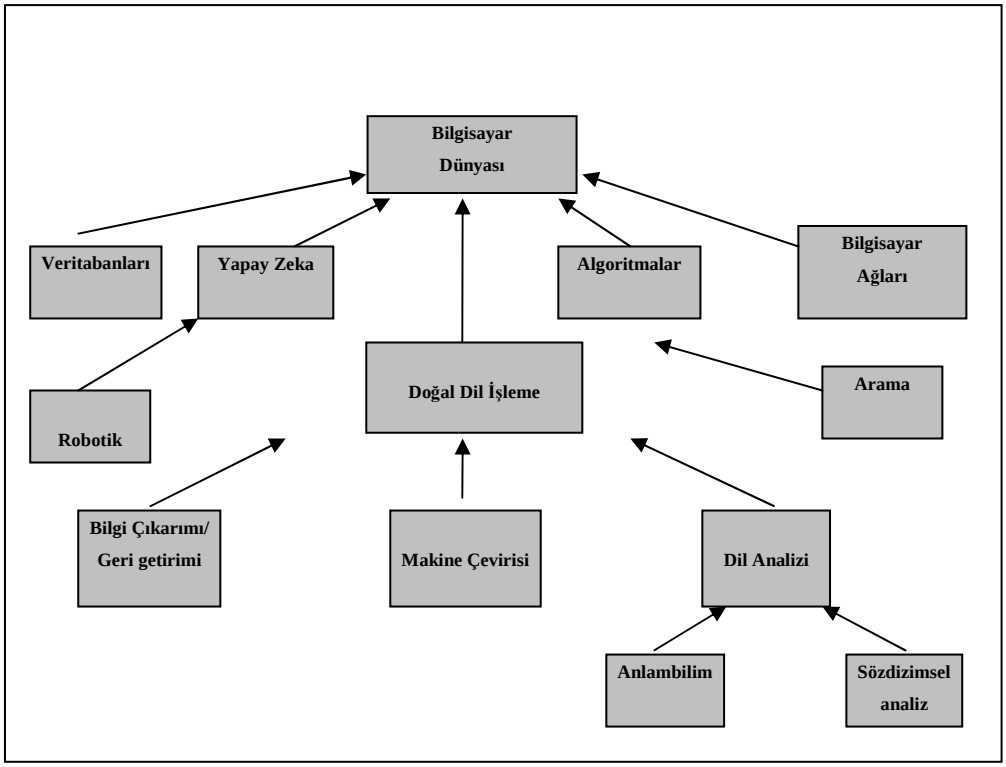
ilgili çalışmaların en çok yapıldığı ve uygulandığı ülkelerde geçerli dilin İngilizce olması nedeniyle, yapılan araştırmaların ağırlıklı olarak bu dilde olduğu gözlenmektedir. Türkçe dünyada üç yüz milyon kişinin konuştuğu yaygın bir dil olmasına rağmen, az incelenen diller kapsamındadır. Bu nedenle de Türkçe için uygulamalı dilbilim çalışmalarının henüz emekleme aşamasında olduğu ve bu alanda yeteri kadar çalışmanın tam olarak sonuçlanmadığı söylenebilir.

Diğer pek çok teknolojik gelişmenin aynen alınıp kopyalanması veya ufak değişikliklerle dünyanın farklı yerlerinde, farklı kültürlerinde kullanılması mümkün olmakla birlikte, DDİ alanındaki çalışmaların paylaşılması bu kadar kolay olmamaktadır. İngilizce veya başka herhangi bir dil için tanımlanan kurallar ve algoritmaların aynen Türkçe'ye veya başka bir dile uyarlanması dillerin yapısının farklılığından dolayı mümkün değildir. Mevcut sistemlerin kullanılacak dile uyarlanması mutlaka uzun ve yorucu çalışmalar sonunda gerçekleşmektedir Hatta çoğu zaman bunların uyarlaması bile mümkün olmamakta ve dile özgü pek çok sistemin baştan oluşturulması zorunluluğu ortaya çıkmaktadır. Ayrıca; bu alandaki çalışmalar pek çok konuda özellikle bilgisayar bilimi ve dilbilimi konusunda uzmanlık gerektirmektedir. Bu nedenle herhangi bir dil ile ilgili olarak yapılacak bilişimsel çalışmalar, hem o dilin dilbilimcileri ve bilgisayar bilimleri uzmanları, hem de incelenen dili çok iyi bilen diğer bilim insanları tarafından gerçekleştirilebilir.

DDİ çalışmalarının amaçları aşağıdaki gibi özetlenebilir:

- Dilin doğasını incelemek**:** Dilbilim.
- Bilişsel yetileri anlamaya yönelik bir kapı aralamak **:** Psikoloji
- Kullanıcı arayüzü teknolojisi olarak faydalanmak**:** İnsan-Bilgisayar Etkileşimi.
- Metin veya konuşmaların çevirisini yapmak**:** Bilgisayarlı Çeviri (BÇ).
- Bilgi yönetimi teknolojisi olarak kullanmak**:** Bilgi Geri Getirimi/Çıkarımı.

Bu amaçlar da göz önünde bulundurularak yapılacak bir değerlendirme ile, DDİ'nin disiplinlerarası bir çalışma alanı olarak pek çok farklı bilim dalıyla birlikte, bilgisayar biliminin alt dallarıyla da yakın ilişkisi bulunduğu sonucuna ulaşılır. Bu ilişkiler Şekil 2.1'de gösterilmiştir.



Şekil 2.1: DDİ'nin farklı bilim dallarıyla ve bilgisayar biliminin alt dallarıyla ilişkisi

2.1.2. Doğal Dil İşlemenin Gelişimi

BÇ, DDİ çalışmalarının ilklerindendir. İlk BÇ çalışmalarında basit bir yaklaşımla sözcükler çevirisi ve sıralaması üzerinde durulmuştur. Bu anlayışa dayanan sistemler; çeviri için sözcüklerin diğer dillerdeki karşılığını iki dilli bir sözlükten bulmuş ve çeviri yapılan dilin sözcük sıralama kurallarına göre bulunan sonuçları sıralamıştır. Doğal olarak bu yaklaşımdan elde edilen sonuçlar çok başarılı olamamıştır. Çeviri işlemini otomatik olarak yapma çalışmaları çok eskilere dayansa da gerçek anlamda uygulanabilirliği yirminci yüzyılda mümkün olmuştur. Elektronik hesap makinelerinin geliştirilmesiyle birlikte bilgisayarların çeviri için kullanılması çalışmaları da başlamıştır. II. Dünya Savaşı sırasında daha çok savaş teknolojilerinin geliştirilmesine harcanan fonlar savaşın bitmesiyle birlikte diğer araştırma alanlarına aktarılmaya başlanmıştır. Bu da diğer alanlarda yapılan çalışmaları hızlandırmıştır. II. Dünya Savaşı sırasında şifre çözme konusunda çalışan Weaver'a göre bütün uluslar, aslında farklı diller de konuşsalar birbirine benzemektedir (Weaver, 1949). Bu nedenle de bir dilde yazılmış bir metin, şifrelenmiş bir metin olarak düşünülebilir. Eğer bu şifre çözülebilirse

bu metnin başka bir dilde ifade edilmiş şekli elde edilebilir. Bu düşünceye göre Türkçe yazılmış bir metin, aslında İngilizce bir metnin şifrelenmiş halidir.

Bu ilginç önerilerin ardından Amerika, Rusya, Fransa ve İngiltere'de pek çok üniversitede BÇ çalışmaları başlatılmıştır. İlk çalışmalarda daha çok Almanca ve İngilizce çeviriler ele alınmıştır. Bu dillerin seçilmesindeki en önemli etken, savaştan kalan teknik belgelerin çevirilmesi gereksinimidir. Zamanla bu belgelere gereksinim kalmadığı için ve soğuk savaş nedeniyle çeviriler İngilizce, Rusça ve Fransızca dillerine kaymıştır. Her ne kadar iki dili, anadili gibi konuşan kişiler BÇ çalışmalarında yer alsa da, onların bilgi aktarımının bilgisayar programına dönüşmesi hiç de kolay olmamıştır.

Bu nedenle 1950'li yıllarda dilbilimciler de BÇ çalışmalarında yer almaya başladılar. Ancak dilbilim alanında da henüz güçlü teoriler bulunmuyordu. 1957 yılında Noam Chomsky, yazdığı *Syntactic Structures* isimli eseri ile teorik alanda günümüze kadar etkisini pek çok alanda gösterecek bir çalışma ortaya koydu (Chomsky, 1957). Biçimsel yapıların tanımlanması için ortaya koyduğu kurallar DDİ çalışmalarını derinden etkiledi.

DDİ konusunda yapılan çalışmalar bu dönemde konuşma işlemeyi de içine alacak şekilde çeşitli alanlara da yayıldı. Ancak, araştırmacılar arasındaki farklı görüşler bu alandaki çalışmaları ikiye böldü. Bir grup, tamamen dilbilimi üzerine yoğunlaşarak Chomsky'nin yaklaşımını benimseyip istatistiksel yöntemleri kullanmamayı tercih ederken, diğer bir grup da tamamen istatistiksel yöntemlere yöneldi.

1980'li yıllarda bilgisayarlar ve elektronik kaynaklardaki gelişmelerle DDİ çalışmalarında da önemli ilerlemeler olmuş, bu dönemde istatistiksel yöntemler de sembolik yöntemleri tamamlayıcı bir faktör olarak tekrar ele alınmıştır.

1990'lı yıllar ise elektronik metinlerin oldukça artması, bilgisayar kapasite ve hızının çok iyileşmesi ve İnternet sayesinde, DDİ de çok gelişmiş ve istatistiksel yöntemler farklı alanlarda kullanılır olmuştur. Yapılan çalışmalarda artık genel metinler incelenmekte ve çok farklı uygulamalar geliştirilmeye çalışılmaktadır. Bu dönemde IBM tamamen istatistksel yöntemler kullanan CANDIDE sisteminin sonuçlarını

açıklamıştır (Berger ve diğ, 1994). Japonya'daki çalışmalarda örnek tabanlı yaklaşım ortaya atılıp kullanılmaya başlanmıştır. İstatistiksel ve örnek tabanlı bu yaklaşımların ortak noktası biçimsel ve anlamsal kurallar yazma gereği bulunmayışı ve gerekli bilginin büyük derleme metinlerden çıkarılmasıdır.

Günümüzde yaygın olarak kullanılmakta olan bilgisayarlarda ses kontrolü ön plana çıkmaktadır. Burada ulaşılmak istenen hedef, insan-makine iletişimini daha özgür ve rahat bir konuma getirmektir. Çalışmalar esas olarak; makineye bir problemin iletilmesini ve çözümünün oluşturulup uygun bir biçimde kullanıcıya aktarılmasını gerçekleştirmeyi hedeflemektedir. Anlamsal analiz ise bu tür çalışmaların kalbidir ve bu konudaki incelemeler yoğun olarak devam etmektedir. Burada asıl amaç, mevcut bir bilgi tabanının bilgisayar tarafından yorumlanabilecek şekilde bir gösteriminin elde edilmesidir. Yapılan çalışmalarla sürekli olarak farklı alanlarda yeni kullanım olanakları ortaya çıkmaktadır.

2.1.3. Doğal Dil İşleme Kapsamında İncelenen Konular

DDİ bağlamındaki ana konular aşağıdaki ana başlıklarla özetlenebilir:

- *Konuşma sentezi:* İlk bakışta çok kolay gibi görünen bir işlem olmakla birlikte, doğal konuşma şekline yakın sentezleme işlemi teknik olarak oldukça karmaşıktır ve birleştirilecek konuşmanın doğru yapılabilmesi (örneğin, doğru vurgulama elde edilmesi) için bazı ayrıntıların belirginleştirilmesi gerekir.
- *Konuşma tanıma*: Temel olarak; sürekli ses dalgalarının sözcüklere dönüştürülmesidir.
- *DDA***:** Konuşma tanıma veya metinlerden elde edilen sözcüklerden, anlama dönüşümü gerçekleştiren basamaktır.
- *Doğal dil üretme (DDÜ):* Verilen girdi bilgilerine karşılık uygun doğal dil cevabının oluşturulması işlemidir.
- *BÇ***:** Verilen metnin bir doğal dilden diğerine bilgisayarlar yardımı ile çevrilmesidir.

Konuşma tanıma teknolojisinin kullanımı ile yapılabilecekler öncelikle bilim kurgu filmlerinde insanlığın ulaşmak istediği hedefler olarak ortaya konmaktadır.

Filmlerdekiler kadar kapsamlı olmasa da, artık günlük hayatta sıkça kullanılan bazı konuşma tanıma sistemleri modern toplumlarda yerini birer birer yerini almaktadır. Bunlara örnek olarak, operatör servislerindeki faturalama işlemlerinin otomatikleştirilmesi, aramaların sesle yönlendirilmesi, bazı standart formlar veya benzeri dökümanların otomatik olarak doldurulması verilebilir. Konuşma tanıma sistemleri daha çok konuşmacıyı doğrulama esasına göre çalışır ve insanlardan daha üstün bir başarı sağlar. Suçlu belirleme için hukuk alanında, güvenliği sağlamak için bankacılık sistemlerinde, özel güvenlik bölgelerine girişlerde, sese dayalı şifreleme sistemlerinde konuşma tanıma uygulamalarından sıkça yararlanılmaktadır.

DDA, pek çok DDİ uygulaması için önemlidir. Ancak herhangi bir yazı veya konuşmayı tam olarak anlama, her zaman gerekli olmasa da, kısmen anlama tam olarak anlama işlemine gerek olup olmadığı konusunda bilgi vermesi bakımından çoğu zaman faydalı bir ilk basamak oluşturur.

Yüzeysel veya kısmen yapılan metin analizleri, sınırlandırılmamış metinlerin etkin bir şekilde sınıflandırılmasında kullanılabilir. Bu ön işlemden sonra elde edilen bilgilerle anlamsal olarak metnin daha derinlemesine incelenmesi sadece gerekli bölümlerde yoğunlaşabilmektedir. Ayrıca, istatistiksel ve dilbilimsel bilgilerle birlikte kullanıldığında sistemlerin bilgi tabanlarına eklenebilecek bilinmeyen öğelerle ilgili dilbilimsel özelliklerin çıkarılmasında da faydalı olmaktadır.

Anlam, kavramlar ve bu kavramlar arasındaki ilişkilerden oluşan anlamsal modellerle gösterilmektedir. Bu anlamsal model kullanılarak herhangi bir şekilde istenen bir bilginin veya verilen bir sorgu ifadesinin bu ifade veya sorgunun dilinden veya anlatımından bağımsız olarak bir anlama eşlemesi yapılabilmektedir. Bu şekilde bilgiye farklı dillerle erişim olanağı dilden veya yapıdan bağımsız olarak sağlanmaktadır.

Dil üretme için temel olarak bir metnin veya konuşmanın anlamsal bir gösterimi kullanılabilir. Temel bir verinin yorumu, bir tümce veya tümceciğin anlamı yüzey metin dizgisine veya sese istenen dilde bazı kurallar çerçevesinde dönüştürülür.

Analiz ve üretimle birlikte kullanılan anlamsal bir modelle BÇ veya başka bir DDİ uygulaması gerçekleştirilebilir. Ancak günümüzdeki uygulamalarda bunun başarılabilmesi için genel olarak sınırlı sözcük ve kavram hazinesi kullanılmaktadır. Belge yapıları için çıkarılan bazı kalıplar, değişken parçaları olan sık kullanılan ifadeler yardımıyla yüksek kalitede metin üretimi gerçekleştirilmektedir.

Çeviri yapan sistemler daha karmaşık ve üst sistemlerdir ve diğer alanlarda yapılan ilerlemelere bağlı olarak gelişimlerini sürdürmektedirler. Bu konuda çalışma yapan labaratuvarlar ve sistemler dünyanın dört bir tarafında dev şirketler tarafından oldukça yüksek bütçelerle kurulmaktadır.

İnternet ve mobil teknolojinin dünyada hızla yaygınlaşmasıyla çeviri sistemlerine gereksinim de artmıştır. Bu nedenle Google gibi arama motorları da çeviri alanında önemli yazılımlar geliştirmiştir.

Genellikle bu işlemleri gerçekleştirebilmek için, sözcükten tümceye ve oradan da anlam çıkarma işlemleri için çeşitli analiz yöntemlerine gereksinim vardır.

2.2. HESAPLAMALI DİLBİLİM

HD, insanın dil yetisini ele alan bilimsel bir inceleme alanı olan dilbilim ile bilgisayar bilimi arasında bir disiplindir. HD, aynı zamanda bilişsel bilimlere de ait olarak insanın bilişsel yeteneğinin matematiksel modellerini incelerken, bilgisayar biliminin bir alt dalı olan YZ ile de örtüşen pek çok özelliğe sahiptir.

HD'nin uygulamalı ve teorik olarak iki farklı çalışma alanı vardır (Uszkoreit, 2000). Teorik HD, bilişsel bilim ve dilbilim alanına giren teorik konuları inceler. Kısaca, insanın doğal dili anlaması ve üretmesi için gerekli olan dilbilgisinin biçimsel teorileri ile ilgilenir. Günümüzde bu teoriler bilgisayarların da yardımıyla incelenebilecek bir düzeye ulaşmıştır. HD alanında çalışan bir araştırmacı, dil yetisinin benzetimini yapacak biçimsel modeller geliştirir ve bunları bazen otomatik, bazen de yarı otomatik paket programlar şeklinde uygulamaya dönüştürür.

Uygulamalı HD, insanın dil yetisinin modellenmesi ve elde edilen pratik sonuçların değerlendirilmesi üzerinde çalışır. Günümüzde kullanılan HD uygulamaları insan mekanizmasını taklit edebilmekten çok uzak olsa bile, pek çok alanda çeşitli şekillerde kullanılmaktadır. Burada asıl amaç, insanın dil yetisi ile ilgili sınırlı da olsa bilginin içerildiği yazılım ürünleri geliştirmektir. Günümüzde insan-makine etkileşiminin önündeki en önemli engelin iletişim problemi olmasına rağmen, bu tip ürünler zaman içerisinde insan-makine etkileşimini önemli ölçüde değiştirecektir. Kullanılmakta olan bilgisayarlar doğal dilleri yeterli olarak çözümleyememektedir; ayrıca bilgisayar dillerini öğrenmek oldukça zordur ve yapıları insan düşünce sistemiyle de bire bir örtüşmemektedir.

HD çalışmalarının amacı üç temel nedenle ilişkilendirilebilir: (Uszkoreit, 2000)

- İnsanlar günlük yaşantıları içerisinde iletişim için dil işleme işlevini çok kolay yerine getirmektedir. HD terminolojisine göre bu işlev matematiksel olarak ifade edilecek olursa; insan verilen bir anlatım için bu anlatıma karşılık gelen anlamı ve de verilen bir anlama karşılık gelen anlatımı hesaplama yeteneğine sahiptir. Eğer böyle olmasaydı insanın düşüncelerini ifade etmesi mümkün olmayacaktı. HD, bu işlevin nasıl yerine getirildiğini çözerek doğal dil işleme konusundaki problemleri daha iyi algılamayı hedefler. Bu açıdan incelendiğinde ise HD, dilbilim ve bilişsel bilimin bir parçasıdır.
- HD çalışmaları bir başka amaca göre daha pratik nedenlerle gerçekleştirilir. Eğer anlam ve biçim arasındaki dönüşümün nasıl hesaplandığı biliniyorsa, bu işlevi gerçekleştirecek bilgisayar programlarının yazılması çok da zor olmayacaktır. Bu tip programların geliştirilmesi ile, sözlü bilgi sistemleri, bilgisayarlı çeviriler, birbirinden farklı doğal dil arayüzleri gibi pek çok doğal dil işleme uygulaması gerçekleştirmek mümkün olacaktır. Günümüzde bu tip uygulamalar hızla artmaktadır.
- HD çalışmalarının üçüncü nedeni ise teoriktir. Anlam ve biçim arasındaki dönüşüm hesaplamaları ilginç özelliklere sahiptir ve genel hesaplama teorisi ile bağlantılı olan önemli yönleri vardır. HD bu şekilde ele alınınca matematiksel dilbilim ve teorik bilgisayar bilimi ile yakından ilişkilidir.

2.3. DERLEME METİN DİLBİLİMİ

Birden fazla metinden oluşan metin topluluğuna *derleme metin (corpus)* denir. Tekil olan *corpus* sözcüğü Latince'den gelmektedir ve gövde anlamındadır; çoğulu ise *corpora* sözcüğüdür. Böylece derleme metinden herhangi bir gövde metni olarak bahsedilebilir. Ancak günümüzde bu sözcük modern dilbilim çalışmalarında verilen basit anlamın dışında farklı anlamlarda kullanılmaktadır. Buna göre derleme metin, dilbilimi analizi ve tanımlamalarında kullanılabilecek yazılı veya sözlü metinlerden oluşmaktadır (Arnold ve diğ., 1999). Bu anlamı ile kullanıldığında bir derleme metinde bulunması gereken birtakım özellikler vardır:

- Çalışılan alandaki olası diğer verileri örneklemeli ve simgeleyebilmelidir.
- Kapsamı yeterince büyük, boyutu sınırlı ve statik olmalıdır.
- Uygulama yapılan makine tarafından *okunabilir* olmalıdır. Burada okunabilirlik kavramı sadece yazılı metinleri değil, makine tarafından algılanabilecek herhangi bir sayısal formu da içine almaktadır.
- Standart bir referansa, kullanmak isteyen bütün araştırmacılara açık olmalıdır.
- Tüm uygulamacıların uyması gereken tasarım kriterlerine sahip olmalıdır.
- Orijinal olmalı ve yapay olarak üretilmemelidir.

Derleme metinler genellikle şu alanlarda kullanılmaktadır:

- Sözlükler,
- Dilbilim araştırmaları,
- HD ve DDİ,
    - Andaçlama ve çözümleme
    - BÇ
    - Otomatik bilgi çıkarımı
    - Otomatik metin özetleme
    - Konuşma tanıma ve sentezleme
- Öğretim yöntemleri ve araçları.

DMD, dili sosyolojik ve psikolojik açıdan inceleyen dilbiliminin diğer alt dallarından biraz farklıdır ve dilbilimin herhangi bir alanında kullanılabilecek bir metodolojidir. Aslında, DMD ve HD arasında da sıkı bir ilişki vardır. Örneğin, *Yağmur tenis maçının ertelenmesine neden oldu* tümcesi incelenmek istendiğinde, bir bilgisayarın bu tümceyi

anlaması, başka bir dile çevirmesi veya yeniden üretmesi için aşağıdakilere gereksinim vardır:

- Biçimbilimsel analiz,
- Sözdizimsel analiz,
- Anlamsal analiz,
- Kullanımla ilgili analiz.

Bilgisayarların bu analizleri yapabilmesi için gereken bilgileri elde etmesinin iki yolu vardır:

- Sınırlı bir veriden gerekli kuralların çıkarımı (hava durumu raporlarının kullanılması gibi).
- Bilgisayarda depolanmış çok büyük metinlerden analiz yoluyla kurallar elde edilmesi. Buradaki temel fikir yeterince büyük metinlerin analiz edilmesiyle, bu metinlerdeki kullanım sıklıklarından faydalanarak ve istatistiksel yöntemler kullanarak tahminler yapılmasıdır.

Günümüzde yukarıdaki özelliklere sahip ve kullanılmakta olan çok sayıda derleme metin vardır. Konu ile ilgili DDİ çalışmalarında kullanılan temel bazı derleme metinler bulunmaktadır. Örneğin, ilk modern elektronik ortamdaki derleme metin olarak *Brown Corpus of Standard American English* gösterilebilir. Derleme metinde bir milyon sözcüğü kapsayan ve 1961 yılında basılmış metinler bulunmaktadır. *Lancaster-Oslo-Bergen* (LOB) derleme metni İngiliz İngilizcesi, *Kolhapur Corpus* derleme metni ise Hindistan İngilizcesi için geliştirilmiş Brown derleme metnine uyumlu derleme metinlerdir. Her ikisi de bir milyon sözcük kapsayan Brown derleme metni gibi on beş farklı kategoriden toplanan yazılı metinlerden oluşmaktadır. 1995 yılında *British National Corpus* (BNC) geliştirilmiştir. Derleme metinde yüz milyon sözcük bulunmaktadır. Bu derleme metinlere daha pek çok modern derleme metin eklenebilir. Bunlar bilim dünyasında kabul görmüş ve birtakım standartları gerçekleştiren derleme metinlerdir (Mcenery ve Wilson, 1996). Türkçe için derleme metin çalışmaları da yakın zamanda başlamıştır. Bunlardan ODTÜ Türkçe Derleme metin Projesi "*Türkçe İçin Biçimbirimsel ve Sözdizimsel Olarak İşaretlenmiş Ağaç Yapılı Derlem Projesi"* başlığını taşımaktadır. Projenin amacı proje grubu tarafından *"Türkçe metinlerin biçimbirimsel ve sözdizimsel işaretlennıesine yönelik işaretleme ve ayrıştırma öğelerinin belirlenmesi ve bu işaretlemenin bilgisayar aracılığıyla yapılması için gerekli çekirdek işlevleri ve kullanıcı ara yüzlerini içeren programların geliştirilmesi ve*

bu programlar vasıtası ile yaklaşık 10,000 tümcenin öğelerinin bir ağaç yapısı çerçevesinde işaretlenerek, Türkiye ve dünyadaki diğer araştırmacıların hizmetine sunulması amaçlanmaktadır" olarak özetlenmiştir (Atalay ve diğ., 2003)

2.4. TÜRKÇE DOĞAL DİL İŞLEME VE HESAPLAMALI DİLBİLİM

Türkçe çok zengin bir dildir ve dilbilim açısından incelemeye değer bir kaynaktır. Bu nedenle özellikle 1980'lerden beri iki yılda bir Türkçe dilbilim konferansları düzenlenmektedir. Ancak Türkçe HD çalışmaları günümüze kadar sınırlı sayıda yapılmış olmasına rağmen, son zamanlarda daha aktif olarak ele alınmaktadır.

Bu konudaki ilk göze çarpan çalışma 1976 yılında yapılan biçimbilim uygulamasıdır (Köksal, 1976). 1981'de ODTÜ'de İngilizce'den Türkçe'ye BÇ'yi konu alan bir yüksek lisans tezi yapılmıştır (Sagay, 1981). Diğer önemli çalışmalara örnek olarak, ABD'de Jorge Hankamer'in (1986) Türkçe'nin biçimbilim çözümlemesi üzerine çalışması, Hollanda'da Albert Stoop'un Hollanda'ca ile Türkçe arasında BÇ konusundaki çalışması (Stoop, 1987), Kemal Oflazer'in Türkçe'nin biçimbilimsel çözümlemesi için iki düzeyli biçimbilim yaklaşımını kullanan çözümleyicisi (Oflazer, 1993) sayılabilir. Son yıllardaki çalışmalar arasında, yazım denetimi (Kibaroğlu, 1991; Akin ve diğ., 1993), veritabanı arayüzleri (Darcan, 1991), bulmaca (Berker ve Say, 1993) ve yarışma benzetimi yapan programlar (Say ve diğ., 1993), biçimbilimsel çözümleyicileri (Güngör ve Kuru, 1993; Cebiroğlu, 2002; Özgür, 2003), ELIZA programının Türkçe uyarlaması (Aytekin ve diğ, 1994), robotlarla iletişim çalışmaları (Keçeci, 1996), sözcük ve konuşmacı tanıma, anlama ve birleştirme sistemleri (Seven, 1997; Öğün, 2003; Sak, 2004), aritmetik problemleri çözen ve cevap üreten ALİ sistemi (Say, 2001), anlambilimsel ve sözcüksel araştırmalar (Kardeş, 2002; Bozşahin, 2002; Demir, 2003), bağlama bağımlı referans üretme sistemleri (Yüksel ve Bozşahin, 2002), BÇ çalışmaları (Şenkal, 2003), istatistiksel incelemeler (Güngör, 2003), bilgi çıkarımı ve belge sınıflandırma işlemleri (Pembe, 2004; Özgür, 2004) yer almaktadır.

1993 yılında, Bilkent Üniversitesi Bilgisayar ve Enformatik Mühendisliği Bölümü ve ODTÜ Bilgisayar Mühendisliği Bölümü'ndeki bir grup araştırmacı, Halıcı Bilgisayar

Şirketi'nin de katılımı ile, Türkçe DDİ konusunda bilgisayar ortamı üzerinde bir dizi temel dilbilimsel kaynak ve uygulama geliştirilmesi amacına yönelik bir proje önerisi hazırladılar. Bu proje önerisinin *NATO Science for Stability Programı*'na kabul edilmesi ile proje amaçlarına yönelik altyapı kurma ve araştırma çalışmaları başladı. Bilkent, ODTÜ, Boğaziçi, İTÜ'de günümüzde bu konu oldukça yaygın olarak çalışılmakta ve pek çok yüksek lisans, doktora tezi çalışmaları yapılmaktadır.

Bilkent Üniversitesi Türkçe Dil ve Konuşma İşleme Merkezi'nde DDİ ve konuşma işleme üzerine yapılan projeler bulunmaktadır. Bu projeler kapsamında Türkçe'nin yapısı incelenmiş, yazım düzeltme, biçimbilimsel ve sözdizimsel analiz, metinden konuşmaya çevirim, Türkçe sözlük, Türkçe-İngilizce iki dilli derleme metin, vs için yazılımlar geliştirilmiştir. ODTÜ'de Bilgisayar Mühendisliği ve Enformatik Enstitüsü'nde LCSL (Laboratory for the Computational Studies of Language) bu konularla ilgili çalışmalar devam etmektedir. Burada yapılan çalışmalar daha çok sözdizimsel analiz ve derleme metin dilbilimi üzerine yoğunlaşmıştır. Boğaziçi Üniversitesi'nde ise oldukça yeni sayılabilecek Bilişsel Bilim Yüksek Lisans Programı'nda bilişsel süreçlerin çeşitli boyutlarda irdelenme yöntemlerini tanıtmak amaçlı; bilgisayar mühendisliği, dilbilim, felsefe ve psikolojinin ilgi alanlarından oluşan disiplinler arası bir program başlatılmıştır. Sabancı Üniversitesi'nde de Türkçe WordNet hazırlanması ve bunu kullanan uygulamalar geliştirilmesi konusunda çalışmalar yapılmıştır (Stamou ve diğ., 2002; Bilgin ve diğ., 2004).

2.5. SÖZCÜK ANLAMI BELİRGİNLEŞTİRMEYE GENEL BAKIŞ

2.5.1. Sözcük Anlamı Belirginleştirme ve Uygulama Alanları

DDİ uygulamalarını ana ve ara uygulamalar şeklinde ikiye ayırmak mümkündür. Ana uygulamalar BÇ, otomatik özetleme (OÖ), bilgi çıkarımı (BÇK) gibi bütün olarak bir işlemi gerçekleştiren sistemlerdir. Ara uygulamalar ise, tümceyi öğelerine ayırma, çözümleme, biçimbilimsel analiz (sözcük ek ve köklerini bulma), SAB gibi ana uygulamalarda yararlı olan birtakım işlemleri gerçekleştirirler.

Bu ara uygulamalardan SAB, amacı anlama olan mesaj anlama, insan-makine iletişimi gibi uygulamalarda zorunlu olarak kullanılması gereken bir ara basamaktır. Amacı anlama olmayan bazı alanlarda ise en azından faydalı veya kullanımı gereklidir (Ide ve Veronis, 1998). SAB'nın kullanımının gerekli olduğu uygulamalardan bazıları aşağıda özetlenmektedir.

2.5.1.1. Bilgisayarlı Çeviri

BÇ araştırmacıları etkin SAB'ın kendi alanlarında çok önemli bir atılım sağlayacağına inanmaktadırlar. Çünkü, çevirilerde sözcüğün en uygun çevirisinin seçilmesi için SAB gereklidir. Bir BÇ sisteminde iki tür sözcük belirsizliğini çözmek gerekir (Hutchins, 1992). Sözcük kaynak dilde belirsiz olabilir veya hedef dilde birden fazla şekilde çeviri yapılabilir. Örneğin, Türkçe *yüz* sözcüğü kullanıldığı yere göre, İngilizce'ye *swim, float, skin, face, surface, cheek, hundred* vs. gibi farklı şekillerde çevirilebilir. Doğru olan sözcük ancak SAB ile seçilebilir.

2.5.1.2. Bilgi Çıkarımı ve Bağlantılı-Metin Taraması

Belirli anahtar sözcükler için tarama yapılırken, belgelerde bulunan o sözcüğün farklı anlamlarını elemek arama sonuçlarının kalitesini artırır. *Fare* sözcüğü bilgisayar terimi olarak aranırken, hayvan olarak kullanıldığı belgelerin elenmesi örnek olarak verilebilir.

2.5.1.3. İçerik ve Tema Analizinde Sözcük Anlamı Belirginleştirme

İçerik ve tema analizinde kullanılan yaygın bir yöntem, önceden belirlenmiş sözcük kategorilerinin ve fikir ya da tema hakkında belirleyici özelliğe sahip sözcüklerin metindeki dağılımının analizidir. Bu analizlerin gerçekleştirilmesinde SAB'ın rolü önemlidir. (Kelly ve Stone, 1975). Dağılımı incelenecek sözcüğün doğru anlamının kullanılması sonuçları etkileyecektir. Kelly ve Stone, (1975) incelen sözcükten önceki ve sonraki sözcükleri kullanarak karar ağaçları oluşturmuştur. Bu çalışmada anlamı belirginleştirilecek sözcüğün önündeki ve arkasındaki bazı sözcüklerin mevcut olup olmadıkları, olanların hedef sözcüğe olan uzaklıkları test edilerek oluşturulan karar ağacında ulaşılan son nokta çözüm için kullanılmıştır.

2.5.1.4. Dilbilgisi Analizinde Sözcük Anlamı Belirginleştirme

Tümcenin öğelerine ayrılması, biçimbilimsel analiz için sözcüklerin doğru anlamlarını bilmek önemlidir. Örneğin,

Belini büken **giderlerdi**

tümcesindeki *giderlerdi* sözcüğünün doğru işaretlenmesi için, aşağıdakilerden hangi anlamda olduğunun bilinmesi gereklidir.

Anlam 1: **git(eylem kök)+Geniş zaman+Çoğul+Geçmiş zaman**
Anlam 2: **gider(ad kök)+Çoğul+Geçmiş zaman**

2.5.1.5. Konuşma İşlemede Sözcük Anlamı Belirginleştirme

Konuşma sentezinde doğru seslendirme (hala ve hâlâ) ve konuşma analizinde sözcük bölümleri ve eşsesli ayrımı için sözcüklerin doğru anlamının bilinmesi gerekir.

2.5.1.6. Metin İşlemede Sözcük Anlamı Belirginleştirme

Yanlış yazılan sözcüklerin düzeltilmesinde, büyük küçük harf değiştirmede ve sesli harflerin kullanılmadığı Arapça, Maltaca, İbranice gibi semitik dillerdeki sözcüksel erişim için de faydalıdır.

2.5.2. Sözcük Anlamı Belirginleştirmede Gerekli Bilgi Tipleri

SAB çalışmalarında genel olarak kullanılan bazı bilgi tipleri vardır. Konu ile ilgili çalışmalarda en sık karşılaşılan ve SAB'da gerekli olduğu belirtilen bilgi kaynakları genel olarak aşağıdaki başlıklar altında özetlenebilir (Ide ve Veronis, 1998; Agirre ve diğ., 2001; Hirst, 1987):

- Öğeler:Sözcüklerin hangi öğe olarak kullanıldığının bilinmesi SAB için ayırt edicidir. Örneğin, *yüz lira* ve *denizde yüz* kullanımlarında *yüz* sözcüğü *ad* ve *eylem* olmasına göre anlamlandırılabilir.
- *Biçimbilim:* Kök ve türemiş sözcükler arasındaki ilişkiler de yardımcı olabilmektedir. *Git* sözcüğü *gittik* şeklinde kullanıldığında eylem olarak

kullanıldığı, *gideri* sözcüğünün ise ad olduğu biçimbilimsel analizle ayırt edilebilir.

- *Yardımcı Sözcükler:* Farklı anlamları olan *kara* sözcüğü *karaya çıkmamıza az kaldı* veya *kara kara düşünmek* olarak kullanıldığında hangi anlama geldiği çok nettir.
- *Anlamsal Sözcük Birliktelikleri:* Bu birliktelikler anlam-sözcük ilişkisi olarak tanımlanabilirse anlam için kuvvetli bir gösterge oluştururlar ve taksonomik (yüz-sayı), durum (yüz-deniz), konu (yüz-spor) gibi alt gruplarına ayrılır. Örneğin, *Gökyüzünün maviliği kara bulutlarla aniden değişti* tümcesinde *mavi* sözcüğü ile *karanın* ortak olduğu renk taksonomi sınıfından yararlanılarak renk anlamı seçilebilir.
- *Seçimsel Öncelikler:* Sözcükler argüman ilişkilerinden farklı şekilde, sözcük olarak değil, anlamsal sınıflar olarak ifade edilirler. Örneğin, *yemek* eylemi öznesine göre çok farklı anlamlara gelebilir. *Kafayı yemek, baba parası yemek* gibi.
- *Kullanım Alanı ve Şekli:* Müzik veya matematik konusunda yazılmış olan bir metinde veya bir söylemde geçen *üçgen* sözcüğü müzik konusunda bir alet, matematik konusunda ise bir şekil anlamında kullanılacağı için uygun anlamı konuya göre tercih edilebilir. Bazı durumlarda ise sözcüğün kullanılış şekli önemlidir. Örneğin, *genç çocuğa vuruldu* tümcesinde *vuruldu* sözcüğünün *sevmek* mi yoksa *dövmek* mi anlamında olduğunun belirginleştirilmesi tümce ile ilgili başka bilgi olmadan çok zordur.
- *Anlamların Frekansı:* Sözcüklerin anlamlarının kullanılış sıklığı da anlam belirsizliğini gidermede bir araç olabilir. *Gök* sözcüğünün *gökyüzü*, *renk* ve *ham* gibi anlamları olmasına rağmen genelde ilk anlamı kullanılmaktadır.

2.5.3. Sözcük Anlamı Belirginleştiren Sistemlerde Kullanılan Kaynaklar

Bölüm 2.5.2'de açıklanan bilgi tiplerini elde etmek için kullanılması gerekli olan bazı kaynaklar vardır. SAB algoritmalarında kullanılan bilgi kaynakları arasında elektronik sözlükler (ES), ontolojiler ve derleme metinler yer alır, ayrıca bunların birkaçının birlikte kullanıldığı uygulamalar da vardır.

- *Elektronik Sözlükler:* Sözlüklerdeki ilk anlam en sık kullanılan anlamın göstergesi olabilir. Bazı sistemler anlamsal sözcük birlikteliklerini modellemek için sözcüğün, sözlükte verilen anlam metnini işlemektedir. Ayrıca sözlüklerde bulunan konu kodlarını, alt sınıf bilgilerini ve temel seçimsel öncelikleri vs. kullanan sistemler de bulunmaktadır.
- *Ontolojiler:* Bazı az rastlanan sistemler hariç, pek çok sistem WordNet'i temel ontoloji olarak kullanmaktadır. Bunun en önemli nedeni ise, WordNet içerisinde önceki bölümde anlatılan bilgi kaynaklarının büyük bir kısmının bulunmasıdır.
- *Derleme metinler:* Derleme metin çeşitli dil modellerinin geliştirilmesine yardımcı olan örnek veriyi sağlar. Elle işlenen derleme metinler bilgisayarla öğrenme algoritmalarında öğrenme aşamasında kullanılmaktadır. Öğrenme verisi, sözcüğün anlamının bulunmasına yarayacak özelliklerin elde edilmesi için gereklidir. Uygulamalarda kolayca elde edilebilecek özellikler tercih edilmekte, dilin işlenmesini zorlaştıran algoritmalardan kaçınılmaktadır. Derleme metinler genel olarak yerel ve genel olmak üzere iki tip özellik kullanır. Yerel özellikler hedef anlam etrafındaki komşu sözcüklerin yakınlıklarını, küçük bir çerçeve içinde inceler. Bunlar yerel özellik değerleri sözcüklerin biçimleri, öğe tipleri vs. olabilir. Genel özellikler ise yerel özelliklere göre çok daha geniş bir çerçeveden elde edilirler. Hedef sözcük anlamları etrafında bulunan 50-100 sözcüklük bir alan taranabilir. Bir anlamla beraber sıkça kullanılan sözcükler ve o anlam arasında anlamsal bir ilişki kurulur. Bu ilişkiler daha çok durum ve konu ilişkileridir.

SAB çalışmalarında öğrenme aşamasında yukarıdaki özellikler çıkarılır. Test aşamasında ise, daha önceden elde edilen bu özelliklerle test edilen sözcüğün özellikleri karşılaştırılır ve en yakın özelliklere sahip olan anlam bu sözcüğün anlamı olarak seçilir.

- *Elektronik Sözlük ve Ontoloji Birleşimleri:* SAB kaynaklarından anlamsal sözcük birliktelikleri, WordNet gibi mevcut ontolojilerdeki anlamsal ilişki eksikliklerini ortadan kaldırmak için kullanılmaktadır. Bu sistemlerde, ES'lerde bulunan tanımlara ilave olarak ES'lerde bulunan veriler arasında yapılan birtakım sınıflandırmalar vs. kullanılmaktadır. Mihalcea ve Moldovan, (1998) çalışmalarında sınıflamaları ve WordNet tanımlarını birlikte kullanarak isim ve

eylem kavramlar arasında bir benzerlik ölçütü elde etmiştir. Bu benzerlik ölçütü olmadan WordNet'te kavramlar arasında bir bağlantı kurulamaz.

- *Elektronik Sözlük ve Derleme Metin Birleşimleri:* Yarowsky (1992), Roget eşanlamlılar sözlüğündeki sıradüzensel yapıyı kullanarak her bir anlam sınıfına ait sözcükleri çıkarmıştır. Yarowsky (1995) daha sonra, derleme metni elle işaretlemekten kaçınmak için bir ES'den alınan kök sözcükleri öğrenme verisini önyüklemek amacı ile kullanmıştır.
- *Ontoloji ve Derleme Metin Birleşimleri:* Ana amaç derleme metinlerin kullanılmasıyla otomatik olarak ontolojilerin elde edilmesidir. Bir kavramla konu bakımından ilişkili olan sözcük kümelerinden oluşan konu imzalarının elde edildiği çalışmalar mevcuttur (Agirre ve diğ., 2001).

2.5.4. Sözcük Anlamı Belirginleştirmede Karşılaşılan Problemler

SAB her ne kadar uzun yıllardır çalışılan ve pek çok DDİ uygulamasında kendisine kullanım alanı bulan bir konu da olsa karşılaştığı pek çok problem vardır. Aşağıda SAB çalışmalarında karşılaşılan bazı genel problemlerden bahsedilecektir.

2.5.4.1. İncelenen Çerçevenin Rolü

Bir sözcüğün anlamını belirginleştirmek için kullanılabilecek en önemli araç sözcüğün kullanıldığı yerdir. Bu nedenle bütün SAB yöntemleri sözcüğün kullanıldığı yerden elde edilen bilgiyi kullanırlar. Bazı çalışmalarda sözcüğün kullanıldığı yerdeki diğer sözcükler, hedef sözcüğe olan uzaklığı, dilbilgisi vs ilişkisine bakılmadan sadece birlikte kullanılmasına göre incelenirken, bazı çalışmalarda ise bu sözcüklerin hedef sözcükle olan ilişkileri önem kazanmaktadır. Ancak, genelde tüm bunlar çok sınırlı bir çerçevede incelenmiştir (yerel veya mikro çerçeve). SAB'nın mikro çerçevede incelenmesinde konu çerçevesi ve kullanım alanı önemlidir**;** ancak bunların birbirine göre rolleri, bunlardan gelen bilginin önem derecesi ve aralarındaki ilişkiler tam anlaşılamaz. Ayrıca hedef sözcüğün etrafında bulunan sözcüklerin hangi uzaklığa kadar inceleneceği de çok net değildir.

2.5.4.2. Anlamlar Kümesinin Saptanması ve Kullanımları

Bir sözcüğün kaç farklı anlamı olduğunu saptamak da oldukça zordur. Sözcüğün anlam sayısı uygulamadan uygulamaya veya kişiden kişiye değişebilmektedir. İnsanlar arasında bile anlamların belirlenmesinde tam bir uzlaşma yoktur. Sözlüklerde verilen anlamlar pratik DDİ uygulamaları için çok geniştir. Bazı anlamlar diğer anlamların özel durumu şeklindedir; hatta sözlüklerde bulunmayan anlamlara da gereksinim duyulabilir. Sözlüklerde bulunan anlamlar otomatik anlam ayrıştırması için zorluk çıkarmakta ve eğitim verisinin boyutunun çok büyük olmasını gerektirmektedir. Sözlük anlamlarını birleştirmek ise bir çözüm değildir**;** çünkü farklı problemlere farklı birleştirmeler gerekmektedir.

Ayrıca bir sözcüğün bütün anlamlarını belirlemenin çok kolay bir işlem olmadığı daha önce belirtilmişti. SAB için önceden belirlenen anlamların mı, yoksa metne göre oluşturulacak grupların mı kullanılması gerektiği tartışmalıdır.

2.5.4.3. Değerlendirme

SAB ile ilgili yapılan çalışmaların hem birbirleriyle karşılaştırılması, hem de ulaşılan sonuçların başka veriler kullanılarak elde edilecek sonuçlarla karşılaştırılması oldukça zordur. Bunlarla ilgili standart değerlendirme yöntemleri henüz tam olarak geliştirilmemiştir. Öncelikle kullanılan metinler çalışmadan çalışmaya farklılıklar göstermektedir. Bu metinlerden bazıları çok özel ve belirli konularla sınırlıdır ki; buralarda kullanılan sözcük anlamları da doğal olarak sınırlı olmaktadır. Bazılarında ise çok genel metinler kullanılmakta ve anlam sınırlaması da bulunmamaktadır. Test için seçilen sözcükler de çalışmaların başarısını etkileyebilmektedir. Sözcük anlamı kolay ayırt edilebilen bir sözcük ise, başarı oranını arttırmaktadır. Değerlendirmeler ile ilgili tüm bu güçlükler nedeniyle, SAB ile uğraşan bilim çevrelerinde standart bir değerlendirme yöntemi geliştirme konusunda çalışmalar devam etmektedir.

2.5.5. Sözcük Anlamı Belirginleştirmede Kullanılan Yöntemler

Günümüzde yaygın olarak kullanılan ve uygulaması yapılmış pek çok SAB çalışması, YZ tabanlı (YZT), bilgi tabanlı (BİT) ve derleme metin tabanlı (DMT) olarak sınıflandırılabilir. Bu bölümde farklı SAB yöntemleri ele alınmıştır.

2.5.5.1. Yapay Zeka Tabanlı Yöntemler

YZT yöntemler 1960'lı yıllarda kullanılmaya başlamış olup, DDA konusuyla yakından ilişkilidir. DDA için vazgeçilmez bir ara basamak olan SAB da bu çalışmalardan etkilenmiştir. Bu sistemler genelde insan dil anlama işlevini yerine getirebilmek için çok ayrıntılı sözdizimsel ve anlamsal bilgi kullanan sistemler olarak dikkat çekmektedir. YZT yöntemler, sembolik ve bağlantısal yöntemler olarak ikiye ayrılmaktadır**:**

- *Sembolik Yöntemler:* 1950'li yılların sonunda anlamsal ağlar (AA) geliştirildi ve sözcüklerin gösterimi problemine uygulandı (Masterman, 1961). Masterman'ın çalışmasında BÇ için tümcelerin gösteriminde dilden bağımsız, temel dil kavramlarını yansıtan bir AA kullanılmıştır. Anlam ayrımları, ağda birbirleri ile sıkı ilişkili düğümlerin seçilmesi ile kapalı olarak gerçekleştirilmekteydi. Bu AA THINGS, DO gibi 100 farklı temel kavram, 15000 kadar da sözlüksel kavram içermekteydi.

  Wilks'in (1968) çalışması anlamsal öncelikleri ve Masterman'ın temel kavramlarını kullanan ve ilk defa sadece SAB problemini çözmek üzere ortaya atılmış bir yaklaşımdı. Bu yaklaşımda bazı kısıtlamalar kullanılarak anlam seçimi yapılıyordu.

  Bu çalışmadan sonra sözcükler andaçlar ve kavramlar arası bağları kullanan ve bu bağlantılarla sözcükler arası ilişkilerin ifade edildiği çalışmalar yapılmıştır. (Quillian, 1961). Quillian çalışmasında sözlük tanımlarını esas alarak ağın ilk omurgasını oluştururken, ağın genişletilmesi için elle kodlama yapmıştır. Bu çalışmada iki sözcük verildiğinde sözcükler arasındaki en kısa yol bulunup, bu yolun geçtiği kavram düğümüne bakılarak anlam belirginleştirmesi yapılmaktadır. Bu yaklaşım daha sonraki ES kullanan sistemler için de bir başlangıç olmuştur.

  Bu çalışmaların çoğunda problem, sadece tek tümce bazında ele alınmaktaydı ve konu, kullanım alanı vs. gibi daha üst düzey bilgilerden yararlanılmıyordu.

Farklı bir yaklaşım ise, insanların dil anlama işleminde dil ile ilgili bilgilerinin kurallardan değil, sözcüklerle ilgili bilgilerden oluştuğu şeklindeydi. Buna göre, ayrıntılı bir çalışmayla dil uzmanlarının bilgilerini koda dönüştürmek ve böylece SAB gerçekleştirmek mümkün olabilir (Adriaens ve Small, 1988) Ancak böyle bir sistemin gerçekleştirilebilmesi için çok fazla sayıda dil uzmanına ve bunların uzun süreli çalışmasına ihtiyaç duyulması, bu sistemin pratikte uygulanabilirliğini ortadan kaldırıyordu.

Dahlgren (1988) pek çok yöntemin birlikte kullanılmasıyla daha başarılı sonuçlar elde edilebileceğini düşünüyordu. Geliştirdiği sistemde sözdizimsel ipuçlarından, seçimsel kısıtlamalardan ve mantıksal uslamlamadan yararlandı. SAB için önce sözdizimsel ipuçları kullanılıyordu**;** sözdizimsel ipuçları ile ve seçimsel önceliklerle SAB gerçekleştirilemediği durumlarda ise, ontolojilerden faydalanan mantıksal uslamlamaya başvuruluyordu.

- *Bağlantısal (Connectionist) Yöntemler:* 1960 ve 70'li yıllarda yapılan çalışmalar anlamsal ateşleme kavramını gündeme getirdi. Buna göre belirli bir kavramın öğrenilmesiyle önceden öğrenilen ve mantıksal olarak ilişkili olan kavramların işlenmesi etkilenmekteydi. İnsanların SAB mekanizmasında da bu etkileşim söz konusuydu. Bu fikir yayılan aktivasyon modelleri olarak kullanıldı (Collins ve Loftus, 1975). Bu modellerde bir AA'da bir kavram kullanıldığında aktif olmakta ve bu aktiflik bağlı olan diğer düğümlere de yayılmaktaydı. Aktiflik yayılma arttıkça azalmaktaydı, ancak birden fazla yayılmadan etkilenen düğümlerde etki artıyordu. Bu modele yasaklama kavramı eklendi (McClelland ve diğ., 1981). Buna göre THROW *ball* sözcüğünün *fiziksel nesne* anlamının aktifliğini arttırırken *sosyal olay* gibi diğer anlamlarının aktif olmasını engelliyordu.

2.5.5.2. Bilgi Tabanlı Yöntemler

YZT yöntemlerde kullanılması gereken bilginin otomatik olarak değil de insanların uzun çalışmaları ile elde edilmesi ihtiyacı bu sistemlerin gelişmesini sınırladı. Böyle çok büyük hacimli bilginin elde edilme maliyetinin yüksek olması bilgi edinimi darboğazı olarak biliniyordu. 1980'lerde büyük hacimli elektronik dil kaynakları

kullanılmaya başlandı. Bu gelişme, elektronik kaynaklardan bilginin otomatik elde edilmesi çalışmalarına da hız kazandırdı ve ES'lerden daha çok faydalanılmaya başlandı.

Sözcüklerin sözlükteki anlamını veren tümceden o anlam için imza denen bir sözcük listesi çıkaran (Lesk, 1986) veya o anlamla beraber geçen sözcüklerin sıklığını kullanan çalışmalar yapıldı (Wilks ve Fass, 1990).

2.5.5.3. Derleme Metin Tabanlı Yöntemler

ES'lerdeki tutarsızlıklar ve sözcüksel ayrıntılı bilgiye rağmen kullanımla ilgili bilginin olmayışı BİT sistemlerin dezavantajıydı. Ayrıca dil analizi için istatistiksel çalışmaların yapılması gereksinimi ve deneysel çalışmaların artmasıyla birlikte büyük hacimli derleme metinlere ihtiyaç duyuldu. 1960'lardan itibaren Brown Corpus, LOB Corpus gibi pek çok derleme metin ortaya çıktı. Çalışmasında 5 sözcük ve her sözcük için 20 öğretim tümcesi kullanmış olmasına rağmen, Weiss (1973) elle anlam işaretlemesi yapılmış bir derleme metinden SAB kuralları öğrenilebileceğini göstermiş, bu da DMT yöntemler için önemli bir adım olmuştur. Ayrıca Harvard Belirginleştirme Projesi olarak bilinen ve daha önce içerik analizinde bahsedilen Stone'un çalışması da yarım milyon sözcüklük bir derleme metinden 1800 sözcük için SAB yapan bir sistem olarak ortaya çıkmıştır. Daha sonra karar ağaçlarına dayanan bir modelde 22 milyon andaca sahip bir derleme metin kullanıldı (Black, 1988). Beş test sözcük için 2000 satırda elle anlam işaretlemesi yapıldı. Bu çalışma öğreticiyle öğrenme çalışmalarının başlangıcı oldu.

Ancak öğretim verisinin anlamlarının elle işaretlenmesinin zor olması ve veri seyrekliği nedeniyle otomatik işaretleme yöntemleri arandı. Bu yöntemler de genellikle şu başlıklar altında toplanıyordu:

- *Önyüklemeli (bootstrapping) yöntemler:* Elle işaretlenmiş sözcük anlamlarından istatistiksel bilgi toplayıp, yeni sözcük anlamlarını ayırmada başarılı oldukça istatistiksel bilgilere ekleme yapan yöntemlerdir (Hearst, 1994).
- *Kümeleme/öbekleme (cluster) yöntemleri:* Önce metindeki sözcükleri gruplandıran sonra hedef sözcükleri bir vektörle gösterip sözcükler yerine

gruplara elle anlam işaretlemesi yapan yöntemlerdir (Schütze ve Pedersen, 1995).

- *İki dilli derleme metin yöntemleri:* Kanada Parlementosu'nun Hansard Derleme metni gibi birbirinin çevirisi olan metinleri kullanarak otomatik anlam işaretlemesi yapan yöntemlerdir (Gale ve diğ., 1992).

SAB çalışmalarında oldukça farklı yöntemler kullanılmakta ve bu yöntemleri uygulamak için de pek çok bilgi kaynağına başvurulmaktadır. En çok kullanılan yöntemler arasında daha çok öğreticiyle öğrenme yöntemleri bulunur. Hafıza tabanlı yöntemler, karar ağaçları, arttırma (boosting) yöntemleri, oylama yöntemleri, Bayes yöntemleri, anlamsal sınıflandırma ağaçları bunlardan bazılarıdır. Yapılan çalışmaların değerlendirilmesi oldukça zordur; çünkü tam olarak standartlaşmış kriterler henüz tanımlanmamıştır. Fakat çok farklı dillerdeki SAB çalışmalarını değerlendiren Senseval çalıştayları halen de periodik olarak devam etmektedir (Resnik ve Yarowsky, 1997).

2.6. SÖZCÜK ANLAMI BELİRGİNLEŞTİRMEDE ÖNEMLİ ÇALIŞMALAR

SAB konusu ile ilgili olarak yapılan toplantılardan en önemlileri Senseval projesi resmi sayfasında da belirtildiği gibi şu şekilde sıralanabilir:

- SIGLEX-97 (Special Interest Group on the LEXicon): "Tagging Text with Lexical Semantics: Why, What, and How?"(Metinleri Sözlüksel Anlamları ile Andaçlama: Niçin, Ne ve Nasıl?) ANLP-97 ile birlikte Nisan 1997'de Washington'da gerçekleştirilmiştir.
- Senseval 1: 1998 yılında 2-4 Eylül tarihlerinde İngiltere'de Herstmonceux Castle'da İngilizce, Fransızca ve İtalyanca için bir çalıştay olarak düzenlenmiştir.
- SIGLEX-99: Standardizing Lexical Resources (Sözlüksel Kaynakları Standartlaştırma) ACL (Association for Computational Linguistics) 1999'la birlikte Maryland Üniversitesi'nde Haziran 1999 tarihinde gerçekleştirilmiştir.
- Senseval 2: 2001 yazında düzenlenen bu toplantıyı Temmuz 2001'de Toulouse'da ACL-2001 ile birlikte gerçekleştirilen bir çalıştay izlemiştir. Senseval 2 Baskça, Çince, Çekçe, Danimarkaca, Hollandaca, İngilizce, Estçe, Japonca, Korece, İspanyolca ve İsveççe dillerini kapsamıştır. Metinde geçen

bütün sözcüklerin belirginleştirilmesinin yapıldığı *bütün sözcükler* sınıfında Çekce, Hollandaca, İngilizce ve Estçe dillerinde çalışılmıştır. İngilizce bütün sözcükler sınıfında isim, eylem, sıfat ve belirteçlerden oluşan açık sınıf sözcüklerinden 1331 tanesi bulunmaktadır. Çeviride Japonca'nın diğer dillerdeki sözcük anlamları ile ilgili çalışmalar yapılmıştır.

- ACL-02: Word Sense Disambiguation: Recent Successes and Future Directions (SAB: Son Başarılar ve Gelecek Planları) çalıştayı ACL 2002 kapsamında Philadelphia'da Haziran ayında düzenlenmiştir.
- Senseval 3: Mart-Nisan 2004'te Barcelona'da gerçekleştirilmiş ve ardından da 2004 Haziran ayında yine de Barcelona'da ACL 2004 ile birlikte düzenlenen bir çalıştay yapılmıştr. Senseval 3 kapsamında SAB için temel 14 işin yanında anlamsal rollerin tanınması, çok dilli açıklamalar, mantıksal biçimler, alt sınıflandırma edinimi gibi konular çalışılmıştır. Daha çok dil kapsanmıştır. İngilizce bütün sözcükler sınıfında 2073 tane sözcük bulunmaktadır.
- Senseval 4 çalıştayının düzenlenmesi çalışmaları tamamlanmıştır. 2007'de gerçekleştirilecek olan çalıştayın kapsamına ilk defa Türkçe'nin de alınması yapılan bu çalışmayla gerçekleştirilmiştir.

2.6.1. Senseval 1 ile İlgili Genel Bilgiler

SAB sistemlerinin başarısı bu sistemlerin değerlendirilmesiyle yakından ilişkilidir. 1997 yılında toplanan ilk çalıştayın amacı, SAB sistemlerinin farklı sözcükler, kullanılan dilin farklı özellikleri, hatta farklı diller açısından güçlü ve zayıf yönlerini çıkarmak üzere gerekli değerlendirmelerin yapılmasına ve uygun etkinliklerin düzenlenmesine öncülük etmekti. Böylece, sözcüksel anlambilim ve çok anlamlılık konularındaki çalışmalara katkı sağlanması mümkün olmuştur. Bu proje ACL denetimindeki SIGLEX isimli daha küçük bir grup tarafından gerçekleştirilmekte ve TREC (The Text REtrieval Conference) ve MUC (Message Understanding Conferences) gibi dil alanında diğer değerlendirme gruplarının araştırmalarından bağımsız olarak çalışmaktadır (Edmonds, 2002).

Senseval herkese açık, çok katılımcılı ilk SAB değerlendirme çalışmasıdır. Senseval'daki değerlendirme yaklaşımı MUC ve diğer ARPA (Advanced Research

Projects Agency) değerlendirmelerine benzemektedir (Hirschman, 1998). İlk Senseval çalışmasının İngilizce, Fransızca ve İtalyanca dilleri üzerinde yapılması hedeflenmiş olmakla birlikte, katılımcılar ve ayrılan fonlar nedeni ile daha çok İngilizce için çalışılmıştır.

SAB için yapılabilecek işlemlerden birisi bütün sözcüklerin anlam belirginleştirmesidir; burada katılımcılar bütün sözcükleri veya bütün açık sınıf sözcükleri ele almak durumundadır. Sözcüksel örneklerde ise seçilen bir grup sözcüğün her biri için derleme metinden seçilen örnek tümcelerdeki anlam belirginleştirilmesinin yapılması beklenir. Sözcüksel örnekler ilk Senseval çalışmasında seçilen işlemler olmuştur.

Senseval projesi kapsamında kullanılan sözlük ve derleme metin 1990'lı yılların başlarında *OUP (Oxford University Press and Digital)* projesi olarak geliştirilen *HECTOR* çalışmasına dayandırılarak oluşturulmuştur. Hector, bir sözlükle derleme metnin bağlantılı olduğu veritabanıdır. BNC projesine de başlangıç olan bu proje 20 milyon sözcük içermektedir. Derleme metindeki sözcüğün her geçtiği yere bir anlam andacı konulmuştur. İşaretlemeler iki kişi tarafından yapılmış, birinci işaretleyicinin işaretledikleri ikincisi tarafından kontrol edilmiş, ancak bundan başka tutarlılık kontrolü yapılmamıştır. Bu veritabanının seçilme sebebi tamamen maliyete dayalıdır; çünkü Senseval için gerekli özellikleri taşımakta ve elle işaretleme için gereken zaman ve parayı belirgin bir şekilde azaltmaktadır (Kilgarriff ve Rosenweig, 2000).

Daha sonra Senseval kapsamında işaretleme için fon bulunmuş ve değerlendirme verilerinin anlam işaretlemelerinin hepsi iki kez daha kontrollü olarak işaretlenmiştir. Eğitim verileri Hector'un işaretleme yapılmamış verileri üzerinde çalışılarak elde edilmiştir. Kullanılan sözlük diğer basılı sözlüklerden veya WordNet'ten çok daha zengin olup, SAB için daha avantajlıdır. Sözlüğün derleme metin güdümlü olması nedeni ile, daha sonraki DDİ çalışmalarında kullanılabilecek sözlükler için de bir örnek olabileceği düşünülmüştür. Ayrıca bu derleme metinde kullanılan metinler bir alana özgü olmayıp, genel İngilizce metinler olduğu için değerlendirmeye daha uygundur.

Senseval öncesi yapılan değerlendirme çalışmalarında en çok eleştirilen yönlerden birisi sözcük örneklerinin seçiminin kişiye/sisteme bağımlı olmasıdır. Senseval'de katmanlı

rastgele örnekler seçimi yapılmış, basit bir rastgele seçim yapılmamıştır. Çünkü Zipf dağılımı nedeniyle seçilen sözcüklerin bir kısmının veya tamamının az kullanılan sözcükler olma olasılığı vardır. Sık kullanılan sözcükler kullanmak hem daha anlamlıdır; hem de kullanım sıklığı ile anlamsal karmaşıklık arasında kuvvetli bir ilişki olduğu için SAB sistemlerinde çözülmesi daha zor olan bir problemdir.

Senseval için sözcük seçimi yapılırken sözcükler BNC'deki geçiş sıklıklarına ve WordNet'teki çok anlamlılık düzeylerine göre sınıflanmıştır. İncelenen bütün sözcük türleri için bu iki kritere göre yapılan sınıflandırma dörde bölünmüş ve 4x4'lük bir matris oluşturulmuştur. Değerlendirmeler ve kullanımlar için 40 sözcüklük bir örnek oluşturulmuştur. Bu örnekler ise matris hücrelerine, hücredeki sözcük sayısı ve derleme metinde bu sözcüklerin karşılık geldiği andaçların oranına göre dağıtılmıştır. Hector sözcükleri ile sınırlı olan çalışmada, daha sonra gerekli sözcük sayısı kadar eleman rastgele hücrelerden seçilmiştir. Bu sözcüklere karşılık altın standardında olacak örnekler için de anlam sayısı ve geçiş sıklığı gözönüne alınarak, sık kullanılan veya anlam sayısı fazla olan sözcükler için daha fazla örnek kullanılmıştır. Burada en az kullanılan veya anlam sayısı az olanlar için en az 160 ve en çok kullanılan veya anlam sayısı fazla olanlar için ise en fazla 400 örnek alınmıştır (Kilgarriff ve Rosenweig, 2000).

2.6.1.1. Senseval Projesi Verileri

Senseval kullanıcılarına üç çeşit veri sunulmuştur (Kilgarriff ve Rosenweig, 2000):

- *Kuru yürütüm (dry run) verisi:* Eğitim ve test verilerine benzer şekilde oluşturulan bu veri, katılımcıların kullanılacak biçim ve biçeme uyum sağlaması için dağıtılmıştır**.**
- *Eğitim verisi:* Bu veride sözcüksel bilgiyle, değerlendirmede kullanılacak sözcükler için örnekler bulunmaktadır. Sözcüksel bilgi, sistemlerin kendilerini bu bilgiye uyarlamaları ve gerekiyorsa ekleme yapmaları için oluşturulur. Örnekler ise beş sözcük dışında eğitimli sistemlerin diğer sözcükleri kullanarak eğitilmelerini sağlar. Eğitimli sözcükler için 26-2008 arasında değişen örnekler vardır.
- *Değerlendirme/test verisi:* Bu grupta sadece her iş için bir küme anlam işaretlemesinin yapıldığı örnekler bulunur. Her bir sözcük için en az üç

işaretleme yapılmış ve bu işaretlemeler araştırmaya katılanlara verilmemiştir. Değerlendirme verisinde toplam 8448 örnek bulunmaktadır.

Değerlendirmedeki temel yaklaşım şöyledir: Bir grup sözcük seçilmekte ve bu sözcüklerin kullanılan derleme metinlerin bazı tümcelerindeki anlamları işaretleyiciler tarafından belirlenmektedir. Daha sonra katılımcı, SAB sistemleri oluşturulmuş derleme metinlerdeki aynı sözcüklerin hangi anlamda kullanıldıklarını tahmin eder. Farklı sistemlerin atadıkları anlamlar, eğer işaretleyicilerin belirlediği anlamlarla tam örtüşüyorsa doğru, kısmen örtüşüyorsa kısmen doğru veya tamamen farklı ise yanlış olarak kabul edilmiş ve toplamda aldıkları sonuçlara göre değerlendirilmişlerdir.

2.6.1.2. Senseval 1 Projesi Değerlendirme Yöntemi ve Ölçütleri

Senseval projesinde farklı değerlendirme kriterleri kullanılmıştır. Bu bölümde bu kriterler anlatılacaktır (Kilgarriff ve Rosenweig, 2000).

- *Değerlendirme Öğe Boyu:* Her iş için katılımcı sistemler değerlendirme öğe boyuna (granularity of scoring) göre aşağıdaki üç grupta değerlendirilmiştir:
  - *Kaba taneli (Coarse-grained)* değerlendirmede alt sınıflandırmalar ihmal edilerek sadece ana anlam sınıfı kullanılmıştır. Cevap ve anahtar dosyalarındaki 1.1 veya 2.1 gibi alt anlam andaçları ana anlam sınıflarına indirgenerek kullanılmıştır. Bu durumda 1.1 alt anlam sınıflaması 1; 2.1 ise 2 olarak kabul edilmiştir. Bu nedenle cevap dosyasındaki 1.1 sonucu anahtar dosyasındaki 1, 1.1, 1.2 olarak gösterilen bütün anlam andaçları ile bire bir uyumlu olarak doğru kabul edilmiştir.
  - *Karışık taneli (Mixed-grained)* değerlendirmede yapılan bir tahmin anahtar dosyasında bulunan bir cevabı içeriyorsa tam puan verilir, eğer bu tahmin herhangi bir cevabı içeriyorsa yarım puan verilecektir. Bir andacın ana anlam kısmı (örneğin, 2), diğer bir andacın anlam andacındaki üst anlamla aynı ise (örneğin, 2.1) içerme ilişkisine sahiptir.
  - *İnce taneli (Fine-grained):* Değerlendirmede sadece özdeş olan andaçlar doğru kabul edilmektedir. Bu durumda andaçlar arasında içerme ilişkisi olsa bile, bunlar tamamen özdeş değilse, tam veya kısmi puan verilmemektedir.

- *En Az Puanlama:* Bir sistemin aldığı en az puan, anahtar dosyadaki andaçla aynı şekilde andaçlanan öğeler kümesi kadardır. Bu nedenle işaretleyiciler arasındaki farklılıktan veya daha farklı sebeplerden çift anlam andacına sahip olan sözcükler bu puanlamada hesaba katılmamıştır. En az puanlama tekniği bütün taneli değerlendirme yöntemlerinde kullanılmıştır. Ancak kaba taneli değerlendirmede en az puanlamaya dahil edilen sözcüklerin ince taneli değerlendirmeye dahil edilmesi gerekmez.
- *Sistem sınıfları:* Proje kapsamında sistemler üç şekilde sınıflandırılmıştır: *A* (Bütün sözcükler-All-words), *S* (Denetimli eğitim-Supervised-training) *O* (Diğer eğitim-Other-training). Karşılaştırmalar aynı sistem sınıfları arasında aşağıdaki gibi yapılmaktadır:
    - *A (Bütün Sözcükler)*: Bu sınıftaki sistemlerin bir metnin içeriğinde bulunan bütün sözcükleri, ya da en azından verilen bir türdeki bütün sözcükleri belirginleştirmesi gerekmektedir.
    - *S (Denetimli Eğitim)*: Bu sistemlerde belirginleştirilecek her bir sözcük için belli büyüklükte (30 veya üzeri) anlam işaretlemesi yapılmış örnek gerekmektedir.
    - *O (Diğer Eğitim)*: Bu grupta ise 30 ve daha fazla sayıda işaretlenmiş örnek öngörülmemekle birlikte, belirginleştirilmesi yapılacak her bir sözcük için bir öğrenme sürecine gereksinim vardır. Bu tür sistemler sadece sözcüksel örneklere uygulanırlar ve bunların 35 sözcük belirginleştirmesi yapan bir sistemden 20000 sözcük belirginleştirmesi yapan bir sisteme dönüştürülmesi henüz gerçekleştirilmemiştir. Ayrıca bu büyüklükte bir ölçeklendirme yapmak kolay gözükmemektedir.
- *Taban Puanları:* Karşılaştırma için puanlama özetlerinde ve ayrıntılı sonuçlarda, eğitim derleme metnini ve sözlük tanımlamalarını kullanan iki çeşit taban puan bulunmaktadır. Bunlardan derleme metin kullanan taban puan, denetimli yöntemlerle belirginleştirme yapan sistemlerin değerlendirilmesinde, sözlük kullanan taban puan ise, sadece sözlüklerden faydalanarak belirginleştirme yapan sistemlerde kullanılmıştır. Bu taban puanlar belirlenirken dosyaların adında bulunan sözcük türü, sözcük kökü gibi katılımcı sistemlerin de kullanımına açık olan bazı bilgiler dışında dilbilimsel bir özellik kullanılmamıştır.

Sözcük öbeği filtresi deyim, kalıplaşmış ifade vs. gibi çok sözcüklü ifadelerin işlenmesine yardımcı olarak, taban puanlama işlemini etkinleştirmek amacıyla kullanılan bir ön işlemcidir. Bu filtre test tümcesinde çok sözcüklü bir ifade bulursa diğer sözcüklerdeki anlam andaçalarını de elemektedir. Bu filtrenin yeterince örneği olmayan ifadeleri de eleyebilmesi diğer bir yararlı işlevidir. Bundan sonra, yeterince örneği olan çok sözcüklü ifadelerle çok sözcüklü olmadığı kesin olan sözcükler taban puanlama algoritmasına gönderilirler. Bu şekilde taban puanlama yönteminde gereksiz sözcüklerle uğraşılmamaktadır.

2.6.2. Senseval 2

Senseval ilk olarak 1998 yılında 2-4 Eylül tarihlerinde İngiltere'de Herstmonceux Castle'da bir çalıştay olarak düzenlendi. Bu ilk çalıştayın başarılı gelişmesi üzerine EURALEX, ELSNET, EPSRC ve ELRA tarafından desteklenen Senseval 2 2001'de düzenlendi. İkinci Uluslararası SAB Sistemlerini Değerlendirme Çalıştayı (Second International Workshop on Evaluating Word Sense Disambiguation Systems) 5-6 Temmuz 2001 tarihinde Toulouse'da düzenlenen ACL-2001 ile birlikte gerçekleştirildi.

2.6.2.1. İşler ve Katılımcılar

Senseval 2 daha çok bütün sözcüklerin anlam belirginleştirmesinin değerlendirilmesine yoğunlaşmıştır. Senseval 2'de 12 farklı dilde üç ayrı işte SAB sistemlerini değerlendirmiştir (Edmonds, 2002):

- *Bütün sözcükler (all-words/AW):* Çekçe, Hollandaca, İngilizce, Estçe dillerinde yapılmıştır. Metinlerde geçen hemen hemen bütün sözcükler belirginleştirilmeye çalışılmıştır.
- *Sözcüksel örnekler (Lexical sample/LS):* Baskça, İngilizce, İtalyanca, Japonca, Korece, İspanyolca ve İsveççe olarak tasarlanmış ve seçilen bir grup sözcüğün anlam belirginleştirilmesi verilen metinlerde gerçekleştirilmeye çalışılmıştır.
- *Çeviri (translation/TL):* Japonca'da ele alınmıştır. Sözcük anlamlarının farklılığını ayırdetmek için çevirilerden faydalanılmıştır.

2.6.2.2. Senseval 2 Verileri

Bu toplantıda üç çeşit veri bulunmaktadır:

- Anlamları açıklamak, tanımlamak veya ayırdetmek için WordNet gibi gerekli ek bilgilerle desteklenmiş sözcük-anlam eşlemesi yapan bir anlam stoku,
- Altın standart olarak kullanılacak; seçimli olarak test ve eğitim olmak üzere parçalara ayırılmış, elle işaretlenmiş derleme metin(ler),
- Kaba taneli ve ince taneli anlam belirginleştirmesi için seçimli olarak sunulan anlam sıradüzeni (Edmonds, 2002).

Anlamlar İngilizce için WordNet 1.7 sürümünden, İspanyolca, İtalyanca ve Estçe için de EuroWordNet sürümlerinden ilk defa alınmıştır. WordNet temel yaklaşım olarak *çokanlamlılık* yerine *eşanlamlılık* ilkesini benimsemiş ve eşanlamlılar kümeleri üzerine kurulmuştur. WordNet 1.7'de Senseval 2'de işaretleme yapan kişilerce önerilmiş bazı değişiklikler bulunmaktadır Altın standardı elde etmede en önemli amaç işaretleyiciler arasındaki uzlaşmanın yaklaşık %90 olmasıdır ancak bu oran uygulamada daha düşük gerçekleşmiştir. Genel olarak bir sözcüğün anlam işaretlemesi için en az iki işaretleyici görev almış, anlaşmazlık durumlarında daha fazla sayıda kişinin görüşüne başvurulmuştur.

2.6.2.3. Senseval 2 Değerlendirme Yöntemi ve Sonuçlar

İşin tipinden bağımsız olarak, sistemlerin değerlendirme verisindeki belirli sözcükleri bazı olasılık veya güvenilirlik katsayıları vererek anlam deposundan seçilecek bir veya daha fazla anlam andacı ile işaretlemesi istenmiştir. Eğitim verisi kullanan denetimli sistemlerle sadece test verisi kullanan denetimsiz sistemler katılmıştır. Sadece test verisi kullanıp sözlük, derleme metin gibi başka hiçbir kaynağı kullanmayan sistemlere yalın denetimsiz sistemler denmektedir ve bu tür sitemlerden sadece bir iki tanesi yarışmaya katılmıştır. Senseval 1'dekine benzer bir değerlendirme yöntemi izlenmiştir. Her bir iş için deneme verisi, varsa eğitim verisi ve değerlendirme verisi sunulmuştur. Sonuçlar gönderildikten sonra otomatik olarak değerlendirilmiştir. Senseval 1'de ortaya konan değerlendirme işlemi küçük bazı değişikliklerle burada da kullanılmıştır. Bütün sistemler için ince taneli değerlendirme yapılmış, ancak işte anlam sıradüzeni veya gruplaması varsa kaba taneli değerlendirme de yapılmıştır.

Sistemlerin bütün sözcükleri veya bir sözcüğün bütün geçişlerini işaretlemesi beklenmemiştir. Doğru cevapların değerlendirme verisindeki tüm elemanlara oranı olan *R*, sistemlerin toplamda ne kadar doğru belirginleştirme yapabildiklerinin bir göstergesi olduğu için bu çalışmada doğruluk oranı olarak öne çıkarılmıştır. Verilen cevaplar içerisindeki doğruluk oranını gösteren duyarlılık, küçük bir kümede anlam işaretlemesi yaptığı zaman çok net cevaplar verebilen sistemlerin lehine sonuçlar vermektedir. Ayrıca bir sistemin anlam belirginleştirmesi yapması gereken kullanımlardan kaç tanesini cevapladığını gösteren kapsam da kullanılmıştır. Taban puan başarımı farklı şekillerde elde edilmekle birlikte genellikle en sık kullanılan anlamın seçilmesi ile oluşturulmuştur. Senseval 2'de İngilizce sözcüksel örnekler için elde edilen sonuçlar, değerlendirme yöntemlerinin aynı ve katılımcıların da bir önceki katılımcılar olup geliştirilmiş yöntemlerle yarışmaya giriyor olmasına rağmen Senseval 1'deki sonuçlardan %14 daha kötü olmuştur.

Senseval 2'de en başarılı sistemler denetimli sistemler olmuştur. Bir sonraki aşamada makine öğrenme algoritmalarının farklı çok anlamlılık durumlarındaki başarımlarında etkili olan özelliklerin seçimi üzerine araştırmalar yapılması düşünülmüştür. Farklı yöntem ve özellikler kullanılarak çokanlamlılık üzerine daha ayrıntılı bilgi edinilebileceği görüşüne varılmıştır. Senseval 2'den çıkan diğer önemli bir sonuç ise iyi bir anlam sınıflandırmasının ve listesinin elde edilmesinin öneminin kavranmasıdır. İnsanların bile ayırdetmekte zorlandığı anlamlarda makinaların başarılı olmasını beklemenin ya da bu işlemi yapan yöntemleri değerlendirmenin olası olmadığı vurgulanmıştır. Bu nedenle uygun anlam sınıflandırmaları oluşturmak ve de bu anlamlarla derleme metinlerde işaretlemeler yapmak için yeni yöntemlerin geliştirilmesi için sözlükbilimciler ve anlambilimcilerle işbirliği içinde çalışmalar yapılması gerekmektedir. Senseval'deki çalışmaların önemli bir dezavantajı ise gerçek uygulamalardan bağımsız olarak SAB sistemleri arasında değerlendirme yapılmasıdır. Bu nedenle belirli bir alanda veya uygulamada bu sistemlerin ne kadar başarılı olacağı, kendi başına bir SAB sisteminin etkili olabileceği bir alanın bulunup bulunamayacağı ya da daha önemlisi ne kadar ayrıntılı anlam sınıflaması yapmanın gerekeceği cevap bekleyen sorular arasındadır. Senseval-3 çalışmaları hazırlık aşamasındayken bu nedenlerden dolayı SAB sistemlerini değerlendirebilecek uygulamaya bağımlı veya uygulamadan bağımsız farklı işlerin önerilmesi için çağrı yapılmıştır. Bu bağlamda

farklı dillerin çalışmaya dahil edilmesi, çok dilli uygulamalara ağırlık verilmesi, bilgisayarlı çeviri veya bilgi çıkarımı gibi özel bir DDİ alanında kullanılması gibi fikirler üzerinde yoğunlaşılmıştır. SAB ile ilgili olan anlamsal andaçlama ve alan sınıflandırması gibi konulardaki işlemler de yan ürünler olarak ortaya çıkabilecek çalışmalardır.

2.6.3. Senseval 3

Senseval 3 çalışması daha önceki Senseval çalışmalarından edinilen birikim ve olgunlukla biraz daha geliştirilmiştir.

Bu bölümdeki bilgiler için Senseval projesi resmi sayfası kullanılmıştır:

- *İngilizce bütün sözcükler:* 64 takım katılmıştır. Senseval 2'dekine benzer şekilde Penn Treebank metinlerinden alınan 500 sözcük WordNet 1.7 anlam andaçları ile andaçlanmaktadır. Olabildiğince çok sıfat ve belirteçle birlikte yüklemler ve özneler işaretlenmeye çalışılmıştır.
- *İtalyanca bütün sözcükler:* 7 takım katılmıştır. İtalyan ağaç bankasından seçilen 5000 sözcük içeren küçük bir kümede bütün sözcüklerle belirginleştirme işlemi yapılmaktadır. Tümcelerdeki sözcük türleri ve biçimsel bağımlılık işaretlemesi de verilebilmektedir. İçerikteki sözcüklerden isimler, eylemler ve sıfatlar ItalWordNet'teki anlamları ile işaretlenmişitir.
- *Baskça sözcüksel örnekle:* 8 takım katılmıştır. Baskça için denetimli ve yarı denetimli SAB öğrenme sistemleri için örnekler sunulmuştur. 40 sözcük için anlam işaretlilerin az ve işaretsizlerin buna oranla (yaklaşık on katı kadar) daha fazla olduğu örnekler kullanılmıştır. Denetimli sistemler işaretli veri üzerinde işlem yapmış, yarı denetimliler veya denetimsizler ise işaretsiz veriden yararlanmıştır. (75 + 15 x anlam sayısı + 7 x çok sözcüklü ifadeler) kadar elemanı olan örneklerin üçte ikisi eğitim geriye kalan kısmı da test için ayrılmıştır. Anlamların WordNet 1.6 ile ilişkisi elle kurulmuştur. 10 kadar sözcüğün diğer dillerdeki (Katalanca, İngilizce, İtalyanca, Romence, İspanyolca) örneklerle eşgüdümlü olabilmesi için çalışma yapılmıştır.
- *Katalanca sözcüksel örnekler:* 8 takım katılmıştır. Baskça sözcüksel örnekler özelliklerini taşımaktadır. Eleman sayısı 45 sözcük için (75 + 15 x anlam sayısı) kadardır.

- *Çince sözcüksel örnekler:* 16 takım katılmıştır. Bu örneklerde sözlük, eğitim ve test verisi olmak üzere üç çeşit veri bulunmaktadır. Sözlükte 20 sözcük için veri girişi vardır. Her sözcük için pek çok anlam HowNet bilgi tabanından hareketle tanımlanmıştır. Her bir anlamın sözlükteki veri girişi için bir numarası, türü, tanımı ve İngilizce çevirisi ile birlikte o anlamın farklılığı ile ilgili ek bilgi bulunmaktadır. Her bir sözcük için eğitim verisinde sözcüğün anlam sayısıyla orantılı olarak 20-100 arası örnek yer almaktadır. Eğitim verisinde sözcük türleri bilgisi bulunan ve bulunmayan olmak üzere iki grup vardır. Sözcük türü andaçlayıcı bir sistem de sağlanmıştır. Test verisinde eğitim verisinin yarısı kadar örnek bulunmaktadır.
- *İngilizce sözcüksel örnekler:* 65 takım katılmıştır. Bu grup için veriler Open Mind Word Expert (OMWE) arayüzü ile toplanmıştır. Her bir eleman için iki andaç toplanmış ve işaretleyiciler arası uzlaşma sağlama için testler yapılmıştır. Daha önce yapılan değerlendirmeler OMWE verilerinin yüksek kalitesini ve faydasını göstermiştir. Senseval 3 için bu veride 150 anlam belirsizliği olan isim, eylem, sıfat ve belirteç için veri toplanması öngörülmüştür. Test verisinin bir kısmı UNT Dilbilimi Bölümü sözlük bilimcileri tarafından oluşturulmuştur. Diğer kısım ise internetten toparlanan anlam işaretli derleme metinlerden elde edilmiştir. İnce taneli ve kaba taneli değerlendirmeler için anlam eşlemeleri de sağlanmıştır.
- *İtalyanca sözcüksel örnekler:* 11 takım katılmıştır. Katalanca sözcüksel örneklerle aynı özelliktedir.
- *Romence sözcüksel örnekler:* 8 takım katılmıştır. Katalanca ve Baskça sözcüksel örneklerle hemen hemen aynı özelliktedir. Bütün açık sınıf sözcük türlerini kapsayan ve farklı düzeyde anlam belirsizliği olan 50 sözcük alınmıştır. Her bir sözcük için örnek sayısı belirlenirken (75 + 15 x anlam sayısı + 10 x çok sözcüklü ifadeler) ifadesi kullanılmıştır. Anlamlar ve çok sözcüklü ifadeler Romence WordNet ve Romence için en çok bilinen ve kullanılan bir sözlük olan DEX'ten alınmıştır. Veriler OMWE'nin Romence için olan arayüzü ile toplanmıştır.
- *İspanyolca sözcüksel örnekler:* 18 takım katılmıştır. Katalanca sözcüksel örneklerle aynı özelliktedir.

- *İsveççe sözcüksel örnekler:* 4 takım katılmıştır. Senseval 2'deki İsveççe örneklere benzemektedir.
- *Otomatik alt sınıflama edinimi:* 35 takım katılmıştır. 30 eylemle sınırlıdır. Bunlar sık kullanılan ancak fazla sayıda anlamı olan *zor* eylemlerdir. Katılımcılara bu eylemlerin listesi verilerek bir eğitim aşamasına izin verilmiş ancak eğitim verisi sağlanmamış test verisi sunulmuştur. Her eylem için 1000 örnek verilerek katılımcıların bunları WordNet 1.7.1 anlamları ile işaretlemesi istenmiştir. Sistemlerin cevapları alınıp geniş Levin biçemi eylem sınıflarına dayanan anlamlarla karşılaştırılarak eşlenmiştir. Her sistemden gelen anlam işaretli veri Anna Korhonen'in alt sınıflama edinimi yazılımına girdi olmuştur. Edinilen çerçevelerle elle oluşturulan altın standardındaki çerçevelerle karşılaştırılmış ve bunun sonucuna göre sistemler sıralanmıştır.
- *Çok dilli sözcüksel örnekler:* 23 takım katılmıştır Burada çeviriler SAB için kaynak olarak kullanılması planlanmıştır. Çeviri metinlerdeki çok anlamlı sözcükler üzerinde yoğunlaşılmıştır
- *WordNet açıklamaları SAB:* 36 takım katılmıştır. WordNet'teki bütün içerik sözcüklerin elle işaretlemesi çok zaman alıcı bir işlem olduğu için bu zorluğun üstesinden gelme amacıyla örnek veri kullanılarak işaretlemelerin otomatik yapılıp yapılamayacağı konusu araştırılmıştır.
- *Anlamsal rollerin otomatik andaçlanması:* 36 takım katılmıştır. Otomatik değerlendirme yöntemlerinin başarımını incelemek üzere oluşturulmuştur. İsveççe anlamsal rollerin tanınması çalışmasına başlanmış ve 2 takım katılmıştır.
- *İngilizce'deki mantıksal biçimlerin tanınması:* 26 takım katılmıştır. İngilizce tümcelerin birinci dereceden mantıksal notasyonlara dönüştürülmesi hedeflenmiştir.

OMWE içinde tekrarlı elemanlar bulunmaktadır. Bu kaynak internet kullanıcılarına açıktır ve istenildiği zaman ekleme yapılabilmektedir, bu şekilde dinamik olarak büyümektedir. Image bir şekil koleksiyonundan elde edilen şekil başlıklarından oluşmaktadır.

3. MALZEME VE YÖNTEM

3.1. TÜRKÇE SÖZCÜK ANLAMI BELİRGİNLEŞTİRME ÇALIŞMASINA KISA BAKIŞ

Türkçe SAB uygulamaları 90'lı yıllarda yok denebilecek kadar az sayıda yapılmıştır. Bu konudaki çalışma sayısının günümüzde de az olmasının en etkili nedenlerinin başında Türkçe DDİ kaynaklarının sınırlı olması gelir. DDİ'nin açık bir araştırma alanı olan SAB çalışmalarında Türkçe için önceden elde edilmiş dilbilgisine dayanan bilgi tabanlı yöntemleri kullanmak pek mümkün olmamıştır. Fakat, yeni kaynakların elde edilmesine paralel olarak zaman içinde Türkçe için yapılan SAB çalışmasının içerik ve uygulamasında da çeşitli farklılıklar gözlenmiştir. Derleme metine bağlı olarak incelenen sözcükler, SAB için kullanılan yöntemler ve bunların çeşitli özellikleri üzerindeki çalışmalar ilerledikçe, araştırmalarda birtakım değişiklikler yapılması gereksinimi ortaya çıkmıştır. Bu yöntemlerin en belirgin özellikleri derleme tabanlı yöntemler ve denetimli öğrenme algoritmaları olmalarıdır. Kısaca, SAB çalışmalarında yeterince büyük metin örneklerinden dilin kullandığı modelleri çıkarmak üzere makine ile öğrenme ve istatistiksel teknikler kullanılmaya başlanmıştır. Bu çalışmada yer verilmeyen eğitim örneğindeki verinin sınıflandırmasının bilinmediği derleme metin tabanlı yöntemler grubundaki denetimsiz öğrenme, bazı araştırmacılar tarafından sınıf andaçlarının ihmal edildiği işlenmemiş örneklerin senaryoları olarak bilgi tabanlı öğrenme grubunda incelenirler (Marquez, 2004).

Yapılan belirginleştirme çalışmalarının farklı aşamalarında göz önüne alınmış olan konulardan bazıları şöyle özetlenebilir:

- kullanılacak derleme metin ve diğer bilgi kaynakları,
- sözcüklerin seçimi ve bu sözcüklere ait anlamların elde edilmesi,
- anlam sınıflamasında etkili olan özelliklerin belirlenmesi,

- uygulanacak yöntemlerle ilgili olarak seçilecek algoritmalar ve değerlendirmelerin nasıl yapılacağına karar verilmesi.

Belirginleştirme çalışmalarında yukarıdaki birbirini tamamlayan farklı kavramların betimlenmesinde kişilerin değerlendirmelerindeki öznel farklılıkların farklı sonuçlar doğurması kaçınılmazdır. Örneğin, herhangi bir hedef sözcüğün anlamını belirginleştiriken, incelenecek metnin büyüklüğü farklı araştırmacılar tarafından farklı seçilebilir. Bazıları sadece birkaç sözcüğü incelerken, diğerleri tümcenin çok daha fazla kısmını değerlendirebilir. Bu durum ise, aynı metin için farklı sonuçlar elde edilmesine yol açacaktır.

Yapılan çalışmada derleme metnin seçiminden kullanılan belirginleştirme algoritmalarına kadar farklılıklar gösteren iki farklı yaklaşım izlenmiştir. Bu çalışmalar Bölüm 3.2 ile Bölüm 3.5 arasında ayrıntıları ile anlatılmakadır.

3.2. DERLEME METNİN SEÇİLMESİ

Derleme metin tabanlı yöntemlerin diğer dillerde de SAB yöntemleri içinde daha başarılı sonuçlar vermesi nedeni ile, çalışmada kullanılan belirginleştirme yöntemleri de bu sınıf içerisinden seçilmiştir. Çalışmada metinlerle ilgili olarak, iki tip derleme metin üzerinde çalışılmıştır. Bunlardan birincisi dünya klasiklerinden seçilen yedi farklı hikaye olan Gulliver, Candide, Ivan Nikiforovic, Tours Papazı, Mozart Prag Yolunda, Mektuplar ve Kır Atlı'dan oluşmaktadır. İkincisi ise ODTÜ ve Sabancı Üniversitesi işbirliği ile geliştirilen derleme metindir.

Birinci derleme metin üzerinde herhangi bir dilbilimsel işleme yapılmamış, tarayıcıdan geçirilen hikayeler ham veri olarak, düz metin şeklinde elektronik ortama aktarılmıştır. Bu nedenle, metinlerin kullanılabilmesi için öncelikle bir ön işlemden geçirilmeleri gerekmiştir. Bu ön işlem oldukça uzun bir elle işaretleme süresi gerektirdiği için bazı zorluklara neden olmuştur. Çalışma sırasında genel kullanıma açık sözdizimsel ve biçimbirimsel çözümleme yapan programlara erişilemediği ve WordNet benzeri Türkçe bir bilgi kaynağına ulaşılamadığı için, ODTÜ tarafından geliştirilmiş olan derleme metnin kullanıma açılması ile çalışmaya bu derlem ile devam edilmiştir. Ancak buradaki bazı problemler de henüz çözümlenememiştir. Derlemdeki metinlerin büyük

çoğunluğu biçimbilimsel çözümlemeden geçirilmiş; fakat yapısal belirsizlikler tamamen çözülememiştir. Seçilen bazı tümceler için belirsizliklerin sonradan düzeltilmiş olmasına rağmen, bu tümceler metinden rastgele seçildikleri için kullanılan yöntemlerde bazı özellikler kaybolmuştur. Ayrıca, ODTÜ derleme metninde de bazı aşamalarda elle işaretlemelere devam edilmesi gerekmiştir. Yararlanılan her iki tip derlem ve bu derlemlerle birlikte kullanılan belirginleştirme algoritmaları aşağıda ayrıntılı olarak anlatılmaktadır.

3.2.1. Birinci Tip Derleme Metin İle Belirginleştirme Çalışması

Birinci tip derleme metnin hazırlanması sırasında ilk aşama olarak elektronik ortama dünya klasiklerinden yedi farklı hikaye aktarılmıştır. Bu metinlerde bulunan bazı hatalar zamanla ayıklanmış ve düzeltilmiştir. Metin incelemede kullanılması gereken araçların Türkçe için sınırlı olması ya da hiç olmaması nedeniyle çalışmada anlam incelemesi yapılacak sözcüklerin de sınırlandırılması gerekmiştir.

SAB için kullanılabilecek sözcüklerin seçiminde genellikle eylemler tercih edilmiştir. Örneğin, *git* sözcüğünün anlamları incelendiğinde, TDK (Türk Dil Kurumu) sözlüğünde bu sözcük için 21 farklı anlam bulunmuştur. Bu anlamlar birleştirilerek 10 farklı anlam içeren bir küme çalışma kapsamına alınmıştır (Tablo 3.1). Birleştirilen bazı anlamlar dışında, bazıları da diğer anlamların özel hali olduğu için elenmiştir.

Bir başka örnek olarak incelenen *çık* sözcüğünün TDK sözlüğünde 58 anlamı bulunmaktadır. Ayrıca *çıkar* sözcüğü ile de bazı anlamları birlikte değerlendirilebilmektedir ve TDK sözlüğünde *çıkar* sözcüğünün isim olarak 1 ve eylem olarak 25 anlamı bulunmaktadır. Bu nedenle işaretleme sırasında *çıkar* sözcüğünün anlamlarının da değerlendirilmesi gerekmiştir. Sonuçta sözlükteki bütün anlamlar ve incelenen metinler göz önüne alınarak, 17 elemanı olan bir anlam kümesi oluşturulmuştur (Tablo 3.2).

Tablo 3.1: *git* sözcüğünün çıkarılan anlamları

Anlam
Bir yere varmak, ulaşmak, bir yerden ayrılmak, yönelmek, terketmek
Devam etmek, sürekli olmak
Hislenmek, beğenmek, yardımcı eylemle kullanılan diğer anlamlar
Yol alma
Kaybolmak, yok olmak, bitmek, tükenmek,sonlanmak
Tekrarlamak, volta atmak
Duymak, işitmek (kulağına gitme)
Anmak,hatırlamak
Harcama
Kademeli olarak, zaman ilerledikçe, durum değiştikçe

Tablo 3.2: *çık/çıkar/çıkarmak* sözcüklerinin çıkarılan anlamları

Anlam
Anlamak
Atmak, eksiltmek
Ayak basmak,varmak
Bir şeyin çıkmasına neden olmak
Elde etmek, sağlamak, olmak
Gitmek, terketmek, varmak
Göstermek
Hoşlanmak
İşe başlamak
Rastlamak
Reddetmek, tavır koymak
Soyunmak
Üzerine veya yukarı doğru tırmanmak
Yetmek, yeterli olmak
Burun (çıkıntı)
Menfaat, yarar
Unutmak

Anlam sayılarının çıkarılması işleminden sonra, metinlerde bu sözcüklerin kullanıldığı tümceler bulunmuş ve tümcelerdeki sözcük anlamları numaralandırılmıştır. Daha sonra bu sözcüklerin yer aldığı tümcelerdeki ilgili bütün sözcüklerin öğeleri bulunmuş ve işaretlenmiştir. Bu işaretlemeler ile ilgili örnekler Tablo 3.3'te görülebilir. Burada anlam belirginleştirmesi yapılacak sözcük **$** işaretleri arasına alınmıştır. Açık metin üzerinde hedef sözcük araştırılırken, sözcüğün sadece incelenecek kök ile başlaması şartı aranmıştır. Bu nedenle işaretleme sırasında bu sözcükletam olarak aynı sınıfta

bulunmayan ifadeler de işaretlenmiştir. İncelenen sözcüğün hemen sonunda **{Numara}** ile belirtilen kısım sözcüğe elle atanan anlamı göstermektedir. **[ ]** arasında yazılan ifadeler ise, incelenen sözcükle ilgili olan sözcük veya sözcük gruplarını belirtmekte ve bunu izleyen **( )** arasında yazılan ifadeler de, bu sözcük veya sözcük grubunun türünü belirten özellikler ile ilgili elle yapılan işaretlemeleri göstermektedir.
Örneğin,

Umut Burnuna kadar [rüzgâr] (ÖZ) [çok iyi] (DuZ) $gitti$ {3}

tümcesinde *rüzgar* ilişkili sözcük olup *özne* durumundadır. ç*ok iyi* ikinci ilişkili sözcük grubu olup *durum zarfıdır*. *Gitti* ise incelenen sözcüktür ve anlamı *3* olarak işaretlenmiştir.

Tablo 3.3: Örnek işaretleme

No	Tümce
1	ülkeyi keşfetmek için yazar da [birlikte] (DuZ) $gidiyor$ {1} ve karada kalıyor
2	gemi, [Suratya] (YeZ) $gidiyordu$ {1}
3	Umut Burnuna kadar [rüzgâr] (ÖZ) [çok iyi] (DuZ) $gitti$ {3}
4	Biraz daha kuzeye dönerek Tataristanın kuzeybatısına ve [Buzdenizine] (YeZ) $gitmek$ {1} [olasılığı karşısında] (ZaZ), bulunduğumuz rotayı izlemenin daha iyi olacağını düşündük
5	[merakımı] (KEyl) $giderecek$ {5} bir şey de göremediğimden
6	[olanca hızımla] (DuZ) [o önce] (ZaZ) $gittiğim$ {1} [yana] (YeZ) koşmuştum
7	[sesim ve işaretlerim] (ÖZ) [hoşuna] (KEyl) $gitmiş$ {4} [gibiydi] (Eyl)
8	Fakat nasıl davrandığımı, kocasının işaretlerine göre ne kadar iyi davrandığımı görünce bana alıştı ve $gitgide$ {12} [artan bir sevgi] (DoT) beslemeye başladı
9	Bu yaptığım [pek] (MiZ) [hoşlarına] (KEyl) $gitmişti$ {4}
10	Ben de $gittim$ {1}, elini öptüm

Bu derlemi kullanarak seçilen sözcüklerin anlamlarını belirginleştirmek için farklı türde algoritmalar kullanılmıştır. Bunlardan biri istatistiksel bir yöntem olan Naive Bayes (NB); diğeri ise Exemplar-based (EB) örnek tabanlı yöntemidir. Bir başka olası açıklaştırma yöntemi ise sözcüklerin kullanıldığı yere göre belge, bölüm, paragraf vs. gibi metnin çeşitli bölümlerindeki komşu anlamlarına bakılmasıdır. Bu bilgiler ancak derleme metnin uygun olması durumunda kullanılabilir ve metinde en sık kullanılan

anlaın gerçek anlam olduğuna karar verilir. Birinci derleme metin bu uygulamayı test edecek özelliklere sahip değildir.

3.2.2. İkinci Tip Derleme Metin ile Belirginleştirme Çalışması

Yukarıda anlatılan birinci tip derleme metinde tümce öğelerini bulma işi elle yapılmaktaydı. Ayrıca sözcüklerin biçimbilimsel çözümlemesini tam doğru olarak veren bir yazılım da mevcut değildi. Doğru çözümleme sonucu bulunabilse bile, genelde birden fazla sonuç elde edildiği için bunlar içinden uygun olanın da seçilmesi gerekiyordu. Bu nedenle algoritmalarda kullanılan bilgilerin çıkarılması oldukça uzun zaman almaktaydı.

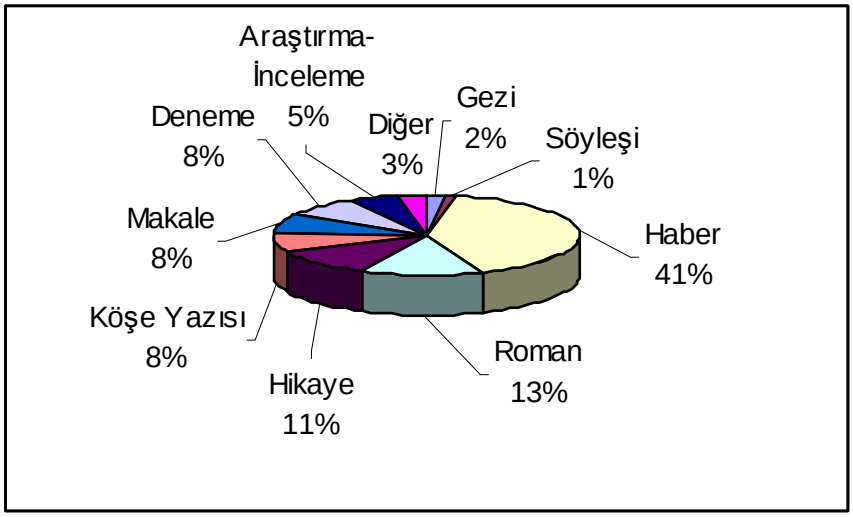
Bu zorluklar Türkçe DDİ çalışmalarının genelinde ortaya çıktığı için, bu amaçla geliştirilen ODTÜ derleme metninin kullanıma açılması ile birlikte çalışma kapsamında da bu derleme metinden yararlanılmasına karar verilmiştir.

İlk olarak Haziran 2003 tarihinde derleme metnin küçük bir bölümü kullanıma sunulmuştur. Önce bu küçük derleme metin üzerinde SAB çalışması yapılmıştır. Daha sonra Kasım 2003 sonunda derleme metnin tamamı akademik çalışmalar için kullanıma açılmıştır. Bu tarihten itibaren de tamamen bu derleme metin üzerinde çalışılmaktadır.

Kullanılan derleme metin, ODTÜ-BAP ve TÜBİTAK tarafından desteklenmiş ve ODTÜ-Sabancı Üniversiteleri işbirliği ile gerçekleştirilmiştir. Çalışmada bir ana derleme metin, bir de farklı kullanımlar için bu ana derleme metinden farklı bazı özellikleri olan ağaç bankası derleme metni geliştirilmiştir.

Derleme metinde kullanılan metinler 1990 yılı sonrası basılan eserlerden seçilmiştir. Derleme metinde yaklaşık olarak 2.000.000 sözcük bulunmaktadır. 201 kitap, 87 makale ve 3 tane günlük gazeteden seçilmiş haberlerden oluşan 999 farklı yazılı metin kullanılmıştır. Seçilen metinlerin dağılımı Şekil 3.1'de gösterilmektedir. Ortamdan bağımsızlık sağlanması için XML ve TEI (Text Encoding Initiative) uyumlu işaretleme benimsenmiş ve BNC gibi standartlaşmış derleme metin benzeri bir derleme metin

ortaya çıkarılması amaçlanmıştır. Çalışma akademik araştırmalar için de kullanıma açılmıştır.



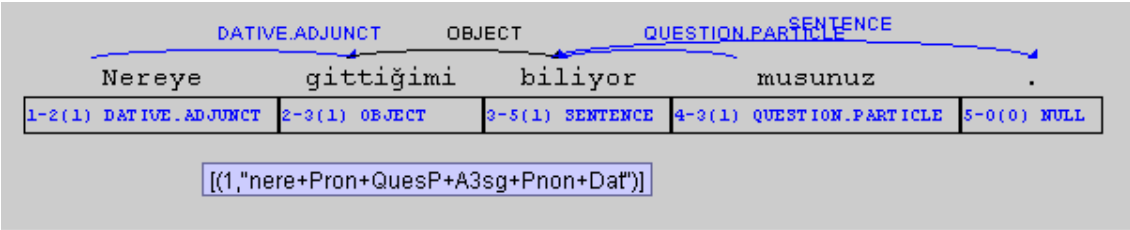
Şekil 3.1: ODTÜ-Sabancı derleme metninde kullanılan metinlerin konulara göre dağılımı

Ana derleme metnin alt derleme metni olan ağaç bankasında biçimbilimsel analiz ve yüzeysel çözümleme özellikleri bulunmaktadır (Atalay ve diğ., 2003, Oflazer ve diğ., 2003). Bu alt derleme metinde ana derleme metinden seçilmiş 10000 işaretlenmiş tümce olması planlanmıştır. Ancak şimdilik 6930 tümce bulunmaktadır. Metin dağılımı da ana derleme metinle orantılıdır. Biçimbilimsel çözümlemeler yapılmış ve birden fazla çözümlemesi olan sözcükler için doğru olanı seçilmiştir. Özel bazı sözcükler, deyimsel kullanışlar ve diğer özel durumlar için de bir ön işlem yapılmıştır.

Sözcük düzeyinde işaretlemede sözcükler çekimli gruplar, yani IG'lerden (inflectional groups) oluşan birimler olarak kabul edilmiştir. Örneğin, *şarklı* sözcüğü derleme metinde aşağıdaki IG'lere sahip olarak gösterilmiştir:

IG 1: (1,"şark+Noun+A3sg+Pnon+Nom")
IG 2: (2,"Adj+With")
IG 3:(3,"Noun+Zero+A3sg+Pnon+Nom")

Tümce düzeyinde işaretlemede ise tümce öğeleri arasındaki ilişkiler aşağıdaki gibi bağlı bir yapı ile gösterilmiştir:



Şekil 3.2: Tümce öğeleri ilişkisel gösterimi

ODTÜ derleme metni yapılandırılırken bazı biçimsel andaçlar kullanılmıştır. Bunlardan bazıları aşağıda verilmiştir:

1. Sentence (tümce),
2. Object (nesne),
3. Subject (özne),
4. Intensifier (vurgulayıcı),
5. Modifier (niteleyici),
6. Determiner (belirleyen),
7. Question-Particle (soru parçacığı),
8. Relativizer (ilişkilendirici),
9. Coordination (bağlaçlar),
10. Possessor (iye),
11. Classifier (sınıflandırıcı),
12. Ablative Adjunct (çıkma tümleci),
13. Dative Adjunct (yönelme tümleci),
14. LocativeAdjunct (kalma tümleci),
15. Instrumental Adjunct (ile tümleci)

olarak özetlenebilir.

Bilgisayar programları işaretleyicilere analiz, çözümleme ve belirsizlikleri gidermede yardımcı olmuş; ancak asıl işaretleme ve düzeltme işlemleri işaretleyiciler tarafından gerçekleştirilmiş ve gerekli düzeltmeler yapılmıştır. Sonuç olarak ortaya biçimbilimsel analizi ve çözümlemesi yapılmış XML belgesi biçiminde bir Türkçe ağaç bankası çıkmıştır (Tablo 3.4).

Tablo 3.4:ODTÜ ağaç bankası dosyalarının XML biçimi

```
<?xml version="1.0" encoding="windows-1254" ?>
- <Set sentences="2">
- <S No="1">
  <W IX="1" LEM="" MORPH="" IG="[(1,"ben+Pron+PersP+A1sg+Pnon+Nom")]"
     REL="[3,1,(SUBJECT)]">Ben</W>
  <W IX="2" LEM="" MORPH="" IG="[(1,"bir+Det")]" REL="[3,1,(DETERMINER)]">bir</W>
  <W IX="3" LEM="" MORPH="" IG="[(1,"tutsak+Noun+A3sg+P1sg+Nom")]"
     REL="[4,1,(SENTENCE)]">tutsağım</W>
  <W IX="4" LEM="" MORPH="" IG="[(1,",+Punc")]" REL="[5,1,(OBJECT)]">,</W>
  <W IX="5" LEM="" MORPH="" IG="[(1,"de+Verb+Pos+Past+A3sg")]"
     REL="[6,1,(SENTENCE)]">dedi</W>
  <W IX="6" LEM="" MORPH="" IG="[(1,".+Punc")]" REL="[,( )]">.</W>
  </S>
  +<S No="2">
 </Set>
```

Bu XML dosyalarında bir veya daha fazla sayıda tümce bulunmaktadır. Baştaki *<Set sentences="N">* andacındaki N tümce sayısını belirtmektedir. *<S No="K">* ile başlayan ve *</S>* ile biten andaç arasına yazılan bölüm bir tümceyi göstermektedir. *<W> </W>* arasında yazılan bilgiler o tümcede geçen bir sözcük için sağlanan veriyi ifade etmektedir. *IX="Z"* ile sözcüğün kaçıncı sözcük olarak kullanıldığı, *IG="*[Çözümleme]*"* ile sözcüğün biçimbilimsel çözümlemesi, *REL="**[i,j,(SUBJECT)]**"* ile de i. sözcüğün j. grubuyla bu sözcük arasındaki ilişki verilmiştir. *</W>* andacından hemen önce gelen bölümde ise sözcüğün tümcede kullanılışı yazılmıştır. *LEM="" ve MORPH=""* andaçları şu anda boş bırakılmıştır. Örneğin, Tablo 3.4'te Ben sözcüğü tümcenin birinci sözcüğüdür. ***ben+Pron+PersP+A1sg+Pnon+Nom*** biçimbilimsel analizi göstermektedir ve 3. sözcüğün 1. IG bölümü ile arasındaki ilişki ise *SUBJECT* yani öznedir.

İkinci tip derleme metin olan ODTÜ derleme metninin kullanılması ile birlikte çalışmadaki sözcük sayısı ve tipi arttırılmış ve bu sözcükler için yeni anlam sınıfları bulunmaya başlanmıştır. Derleme metnin ilk deneme sürümünde daha sonra kullanılan sözcüklerin bir alt kümesi elde edilmiştir (Tablo 3.5).

Tablo 3.5: ODTÜ ağaç bankasının deneme sürümünden seçilen sözcüklerin tümce ve anlam sayıları

Sözcük	**Metinlerdeki tümce sayısı**	**Anlam sayısı**
Yan	104	9
Git	189	10
Gör	133	9
Çık	231	15
Al	250	10
Gel	281	12
Yap	328	6
Ol	941	4

3.3. SÖZCÜK ANLAMLARININ OLUŞTURULMASI

Sözcük anlamını belirginleştirmede kullanılacak anlamların oluşturulmasında tüm dillerde oldukça uzun çalışmalar yapılmıştır. Türkçe sözcüklerin ortalama anlam sayısı bu alanda yoğun uygulamaları olan İngilizce gibi dillere göre çok daha fazladır.

Sözlüklerde bulunan anlamlar daha çok kullanımları sıralamakta, ancak bir anlam sınıflandırması yapmamaktadır. Sözcüklerin asıl kullanımları yanında deyimsel kullanımlar, bileşik sözcükler, atasözleri gibi farklı kullanımları da ele alınınca, anlam sayıları bunlara paralel olarak artmaktadır. Bu yan kullanımlar da ilave edildiğinde bazı sözcüklerin anlam sayısı SAB çalışmasında incelenemeyecek kadar fazla olmaktadır. Ayrıca, işlenmiş metinlerde bu anlamların bazıları ya hiç geçmemekte, ya da çok az geçtiği için daha sonraki sınıflandırma aşamaları için yeterli veri sağlanamamaktadır. Oysa SAB alanında diğer dillerde yapılan çalışmalara bakıldığında, özellikle bu

konudaki en kapsamlı değerlendirme projesi olan Senseval projesinde incelenen sözcüklerin veya diğer dillerdeki sözcüklerin ortalama anlamının yaklaşık olarak 8 olduğu görülür. Senseval 2 sözcüksel örnekler verisinde (isimler için kullanılan) test ve eğitim için ortalama anlam sayısı 7.93'tür. Eğitim verisinde seçilen sözcüğün farklı anlamlarının geçtiği 3587 tümce, test verisinde ise 1773 tümce bulunmaktadır. Ayrıca, deyimsel kullanışlar bu çalışmalarda değerlendirmeye alınmamıştır. Görüldüğü gibi kullanılan eğitim ve test verileri Türkçe için kullanılabileceklerden çok büyüktür. Bu koşullarda Türkçe SAB çalışmalarının sözcük anlamlarını oluşturma aşamasının zorluğu daha iyi anlaşılabilir.

Çalışmada sözcük anlamlarını sınıflandırmak için farklı yöntemler kullanılmıştır. Bunlardan birincisinde, sözcüklerin sözlük anlamları alınarak metinlerde geçen anlamlar için tüm anlamların bir alt kümesi oluşturulmuştur. Bu işlem sırasında bazı anlamlar genel bir başlık altında toplanmış, bazı anlamlar da hiç kullanılmadığı için göz ardı edilmiştir. Bu sınıflandırma sonucunda bile anlamlar alt kümesi bazı sözcükler için çok fazla sayıda anlam içerebiliyordu.

SAB için ikinci bir yol olarak çeviriler kullanılmıştır. Seçilen sözcüklerin anlam sınıflandırması diğer dillerdeki farklı karşılıklarına göre yapılmıştır. Ancak burada da bazı problemler ortaya çıkabilir. İlk olarak bu yaklaşım SAB'nın kullanılabileceği diğer alanlara uygun olmayabilir, ya da sözcüklerin farklı dillerde farklı karşılıkları olduğu için farklı sınıflandırmalar ortaya çıkabilirdi.

Aşağıda örneklendiği gibi bazı sözcüklerin kullanımı veya anlamı sadece Türkçe'ye özgü olabilir; diğer dillere çevirilse bile tam olarak karşılığı bulunamaz:

- Gelen ağamsa giden paşamdır.
- Gelirsem yanına sana gösteririm.
- Eli ekmek tuttu
- Elim ayağım buz kesti

Anlam sınıflandırması işlemi SAB çalışmalarının ilk ve en önemli basamağını oluşturur. Bu basamak başarı ile tamamlanabilirse, diğer basamaklardaki işlemler de çok daha kolaylaşacaktır. Elle yapılan anlam sınıflaması işleminde işaretleyiciler arasında zaman zaman anlaşmazlıklar olabilmekte, aynı sözcük için farklı anlam kümeleri

önerebilmektedirler. Hatta aynı tümcedeki bir sözcüğün anlamı iki farklı işaretleyici tarafından tamamen farklı iki anlam olarak sınıflandırılabilmektedir. Anlamları arasındaki fark belirgin olan sözcükler için bu durumla çok sık karşılaşılmaz. Fakat anlamları birbirine yakın olan bazı sözcükler için bu problem mutlaka ortaya çıkmaktadır. Anlam işaretlemesi sırasında ortaya çıkan bu problem, daha sonra SAB işleminde kullanılan özelliklerin ve algoritmaların belirlenmesinde de etkili olmaktadır. Bu nedenle anlam işaretlemesinin daha kolay ve daha net olarak yapılabileceği yöntemlere gereksinim duyulmuştur. Aşamalı anlam sınıflandırması ve yapay sözcüklerin kullanılması bu yöntemlerden ikisidir.

3.3.1. Aşamalı Anlam Sınıflandırması

Bu yöntemde seçilen sözcük için aşamalı bir anlam sınıflandırması yapılmıştır. Buna göre önce sözcüğün asıl anlamı ele alınmış; asıl anlamdaki kullanımlara birinci anlam, bunun dışındaki diğer tüm kullanımlara ise ikinci anlam denmiştir. Bu şekildeki sınıflandırma ile bazı algoritmalar denenerek, sonuçları daha sonraki aşamalarda sınıflandırma için kullanılmıştır. Sonraki aşamalarda diğer kullanımlar kendi arasında tekrar gruplandırılmış ve anlam sınıflandırması bu şekilde bölünerek devam etmiştir. Aşamalı anlam sınıflandırmasının anlama etki eden pek çok faktörü ayırt etmede oldukça yararlı olacağı düşünülmüştür. Bölüm 4.2.1'de *gel* sözcüğü için aşamalı anlam sınıflandırması çalışmasının sonuçları görülebilir.

3.3.2. Yapay Sözcükler ile Anlam Sınıflandırması

Aslında bir sözcüğün farklı anlamlara sahip olması ile, birden fazla sözcüğün farklı anlamlar için kullanılması arasında bir benzerlik bulunmaktadır. Örneğin, *kar* sözcüğünün *bir doğa olayı* olma anlamı ile *kazanç* anlamı arasındakine benzer bir ilişki *haber* ve *ağaç* sözcüklerinin iki farklı anlam için kullanılmasına benzetilebilir. Bir başka deyişle *haber* ve *ağaç* sözcükleri *haberağaç* şeklinde tek bir sözcük gibi ele alınırsa bu yeni sözcüğün bir *haber* bir de *ağaç* anlamı olmak üzere iki anlamı vardır denilebilir.

Bu şekilde bir düşünce ile *haber* ve *ağaç* sözcüklerinin geçtiği bütün tümceler bulunup *haber* olarak geçen tümceler birinci anlam, *ağaç* olarak geçen tümceler ikinci anlam olarak çok kolay bir şekilde işaretlenebilir. Çünkü hangi tümcelerde *haber* sözcüğünün kullanıldığı, hangilerinde de *ağaç* sözcüğünün kullanıldığı bilinmektedir.

Araştırmalardan yapay sözcüklerle ilgili pek fazla uygulama yapılmadığı görülebilir. Bu tür uygulamaların fazla olmamasının nedeni, incelenen dillerin çoğunun anlam bulma ve işaretleme işlemi için gereksinmelerini karşılamış olmasıdır. WordNet gibi kaynaklar sözcüklerin anlamlarını belirleme işlemini tamamlamış, anlam işaretlemesi yapılmış büyük metinleri kullanıcılara zaten sunmuştur. Böylece Türkçe'deki metin sorunu, diğer dillerde pek yaşanmamaktadır.

Anlam işaretlemesinin böyle bir yaklaşımla gerçekleştirilmesi için bazı durumların göz önüne alınması gerekir. Bunlar yapay olmayan sözcükleri iki sınıfa ayırarak açıklanabilir:

- Anlamları birbirine yakın olan sözcükler
- Anlamları birbirine uzak olan sözcükler

Bu iki sınıfın benzetimini yapay sözcüklerle yapabilmek için, seçilen yapay sözcüklerden bazılarının anlam olarak birbirine yakın olması ve birinci gruptaki sözcüklerin benzetimini yapması gerekmektedir. Aynı şekilde anlam olarak birbirine uzak olan yapay sözcükler seçilerek, ikinci gruptaki sözcükler olarak kullanılabilir. Ayrıca sözcük türüne göre anlam farklılaşması olan sözcükler için de, farklı türden seçilen yapay sözcüklerle gösterime gerek vardır. Örneğin, *kız* sözcük kökü ad ve eylem olarak kullanıldığında farklı anlamlara gelmektedir. Sadece bu özelliği ile bile anlam belirginleştirmesi yapılabilir. Yapay sözcüklerle ilgili olarak da bu tür durumlar oluşturulmalıdır.

Yapay sözcüklerde diğer bir konu anlam sayısıdır. Her sözcüğün farklı sayıda anlamı bulunabilir. Tek anlama sahip sözcükler olduğu gibi, 50-60 değişik anlamda kullanılabilecek sözcükler de bulunmaktadır. Yapay sözcüklerdeki anlam sayısı da kullanılan sözcük sayısı arttırılarak bu durumlara paralel olarak ayarlanabilir.

3.4. SÖZCÜK ANLAMLARINA ETKİ EDEN ÖZELLİKLERİN SEÇİLMESİ

Bir SAB sistemi tasarımının ilk aşamalarında ele alınması gereken önemli konulardan birisi, kullanılacak olan özellikler kümesine karar vermektir. Uygun özelliklerin seçilmesi son derece önemlidir. YZ uygulamalarının pek çoğunda da özellik seçimi üzerine çalışmalar yapılmakta ve bir dizi özellik başarılı bir şekilde kullanılmaktadır.

SAB çalışmalarında kullanılabilecek özellik adayları arasında hedef sözcük etrafında verilen bir pencere içerisindeki sözcükler ve bunların türleri (Bruce ve Wiebe, 1999, Ng ve diğ., 1997), anahtar sözcükler (Ng ve Lee, 1996) veya bağlamdaki ikili sözcükler (bigrams in the context) (Pedersen, 2001), yerel sözdizimleri (collocations) (Kelly ve Stone, 1975; Yarowsky, 1993; Ng ve Lee, 1996) ve çeşitli biçimsel özellikler (Fellbaum ve diğ., 1998; Ng ve Lee, 1996; Bruce ve Wiebe, 1994) sayılabilir. Yerel sözdiziminden anlaşılması gereken anlam belirsizliği olan hedef sözcük komşuluğunda bulunan, sözcük sırasının önemli olduğu kısa sözcük dizisidir. Bu dizinin yerel sözdizimi olarak kabul edilmesi için bir deyim olması şartı yoktur. Bu gruptaki elemanların hedef sözcük etrafındaki diğer komşu sözcüklerden ayrılan yönü, yerel sözdizimlerinde sözcük sıralamasının dikkate alınmasıdır. Kullanılabilecek özellikler konusunda en kapsamlı özellik kümeleri olarak aşağıdaki gibi sıralanabilir (Mihalcea, 2002):

- Anlam belirsizliği olan sözcüğün kendisi ve türü
- Sözcüğün bağlamsal özellikleri: Etrafındaki sözcükler ve bunların türleri
- Yerel sözdizimler (en fazla k tane sözcükle sınırlandırılmış)
- Ad öbeğinin başı
- Anlama özgü anahtar sözcükler
- İkili sözcükler
- Anlam belirsizliği olan sözcükten önce ve sonra bulunan eylemler, adlar, özel adlar, ilgeçler, adıllar, belirteçler.

Yukarıda sıralanan bu geniş örnek kümesi içinden en etkin olarak gözlemlenenler ise, anlam belirsizliği olan sözcüğün kendisi ve türü, bağlamsal özellikleri ve yerel sözdizimleri olarak seçilmiştir.

Sözcük anlamına etki eden faktörler gerçekten çok çeşitlilik göstermektedir. Weaver (1949) ünlü Memorandum'unda BÇ'de SAB'ın gerekliliğini ve diğer çalışmalarda temel olacak ilk SAB yaklaşımını şöyle özetlemiştir:

Bir kişi bir kitapta bulunan sözcükleri, her bir sözcüğü gösteren opak(donuk) bir maske altında bir defada sözcük sözcük incelerse, sözcüklerin anlamlarını tahmin etmesi imkansızdır.[...] Ancak bu opak maskede alanı, sadece incelenen ana sözcüğü değil, her iki taraftan N tane sözcüğü gösterecek şekilde genişletilirse ve N de yeterince büyükse, sözcüğün anlamına doğru olarak karar verilebilir. [...] Buradaki asıl soru: "Ana sözcüğün doğru anlamını en azından kabul edilebilir bir oranda seçmeyi sağlayacak minimum N değeri ne olmalıdır?"

Burada verilen N farklı sözcüğün anlamı belirlemede etkisi çok büyüktür. Ancak anlamın sadece N sözcüğe bağlı bir fonksiyon olarak çözülebilmesi mümkün değildir. Bu N sözcükle beraber kişinin daha önceden öğrenmiş olduğu bilginin de anlam belirginleştirmesinde etkisi vardır. Bunların neler olduğu şu anki araştırmalar çerçevesinde tam olarak belirlenememektedir. İnsanın öğrenme mekanizmasının nasıl çalıştığı konusunda henüz cevaplanamayan pek çok karanlık nokta vardır. Bu mekanizmanın işleyişi kısmen veya tamamen çözülebilirse, tam anlamıyla başarılı sistemlerin geliştirilmesi mümkün olabilecektir.

Çalışmada öncelikle sözcük anlamına etki eden biçimsel faktörler ele alınmaktadır. Çünkü bu faktörlerin çıkarılması ve kullanılması hem daha kolaydır; hem de daha az zaman almaktadır. Bu çalışmada da biçimsel faktörler olarak XML dosyalarından alınabilecek bütün bilgilerden yararlanılmış ve algoritmalarda kullanılmıştır.

Çalışmanın sözcük anlamlarına etki eden faktörlerin incelendiği Dördüncü Bölüm'de sözcüklerin sınıflandırması yapılmaktadır. Bu amaçla, anlamlarla ilgili olarak birtakım kalıplar oluşturulmuştur ve kalıbın bazı kısımları değişken olarak ifade edilebilmektedir. Bu değişken kısımlara aynı gruptan sözcükler getirilerek yeni tümceler elde edilebilmekte ve aynı anlamda kullanımlar korunmaktadır. Tablo 3.6 bu konudaki bazı örnekleri vermektedir:

Tablo 3.6: Sözcüklerin sınıflandırılmasının anlam belirginleştirmesine etkisi

Ali geldi. Veli geldi. Ben geldim. Kedi geldi. Adam geldi.	➔ **Biri veya bir şey geldi** şeklinde genelleştirebilir
Aklıma geldi. Hayalime geldi	➔ **Soyut kavrama** geldi şeklinde genelleştirilebilir
Yüzde beş zam geldi Yüzde on arttı. Yüzde yirmi azaldı	➔ **% sayı** olarak kullanım şeklinde genelleştirilebilir
Denizde yüzdü. Havuzda yüzdü. Suda yüzdü.	➔ **Su kaynağında** yüzdü şeklinde genelleştirilebilir

Hedef sözcükten önce veya sonra gelen sözcüklerin kökleri ele alınırsa, bu sözcüklerin arasındaki bağlantılar algılanamayacağından çıkarım yapılması zor olacak ve algoritmaların başarım oranları düşecektir. Ancak sözcük kökleri ile birlikte sözcüklere ait daha genel sınıflar verilebilirse, genel çıkarımlar yapılarak kalıplar oluşturulabilir. Bu nedenle çalışmaya daha önce Türkçe için yapılmamış olan sınıflandırma işlemi de dahil edilmiştir. Ancak Türkçe'de kullanılan bütün sözcüklerin sınıflandırması, hatta bu sınıflandırmanın bir düzeyden daha ileri devam ettirilmesi oldukça uzun bir süreç gerektireceğinden, sınıflandırma işlemi metinlerde belli bir oranda geçen sözcükler için başlatılmıştır. Ayrıca, birden fazla anlama sahip olan sözcükler için birden fazla uygun sınıflandırma bulunabilmekte ve uygun sınıfın seçilebilmesi için sınıf belirginleştirmesi işlemi gerekmektedir ki; bu da SAB probleminin başka bir türünü ortaya çıkarır.

3.5. BİLGİSAYARLA ÖĞRENME VE DDİ ÇALIŞMALARI ARASINDAKİ ETKİLEŞİM

BÖ yöntemlerinin ortaya atılması ve çok farklı alanlarda uygulanabilirliğinin gösterilmesi bilgisayar bilimlerinde önemli ve faydalı bir gelişme olmuştur. Bu sayede çok büyük miktarlardaki verilerin analiz edilebilmesi mümkün olmuş ve hangi bilgilerin ne oranda probleme etki ettiği konuları daha kolay araştırılabilir konuma gelmiştir. Yapılan bu çıkarımlar sayesinde de daha sonra kullanılacak olan veriler üzerinde

otomatik olarak hesaplamalar yapma veya daha hızlı ve daha güvenilir tahmin yapma konusunda ayrıntılı bilgiler sunma olanağı sağlanmıştır.

Bilgisayarla öğrenme (BÖ) algoritmaları genel olarak denetimli ve denetimsiz olarak iki kategoriye ayrılabilir. Denetimli algoritmalarda yapılacak sınıflandırma için gerekli olan kategoriler önceden bilinmektedir ve eğitim verisi kullanılarak bu sınıflar için gereken bilgiler elde edilir. Test verileri de önceden elde edilen bu bilgiler kullanılarak yine aynı sınıflara eşlenir. Denetimsiz algoritmalarda ise kategoriler önceden bilinmez ve eğitim verisine göre farklılık gösteren örnekler farklı sınıflara atanır. Daha önceki verilerden çok farklı olan her veri için yeni sınıf oluşturularak kategoriler eğitim aşamasında tamamlanır.

DDİ tabanlı uygulamalara gereksinimin sürekli artmasının doğal bir sonucu olarak, çok büyük derleme metinlerin oluşturulması ve bunlara devamlı yenilerinin eklenmesi mecburiyeti, daha güçlü hesaplama araçlarının ortaya çıkmasına ve de BÖ ve DDİ alanındaki çalışmaların son yıllarda birlikte yürütülmesine neden olmuştur.

DDİ uygulamalarının pek çoğunda elle çıkarılması gereken çok büyük verilere gereksinim duyulmaktadır. Ancak bu çıkarımların otomatik olarak yapılmasını sağlamak, elle yapılan çıkarımlara bağımlı kalmaktan araştırmacıları kurtaracağı gibi, mevcut verilerden gerekli bilginin daha kısa ve daha güvenilir bir şekilde elde edilmesi sonucunu da doğuracaktır.

BÖ'de veri sınıflandırması oldukça aktif biçimde çalışılan bir konudur ve BÖ ile DDİ çalışmalarının birleşme noktalarında da DDİ' deki ortak bir problemin bir sınıflandırma problemi haline dönüştürülmesi bulunmaktadır. Bu nedenle de çok bilinen BÖ tekniklerinin hemen hemen tüm DDİ problemlerine uygulandığı görülmektedir. Bu iki alanın birlikte kullanımı ile ilgili etkin bir inceleme olarak Kazakov'un (1996) çalışması verilebilir.

Özelliklerin seçiminden ve bağlamdan elde edilen bilgilerin kullanılarak örneklerin oluşturulmasından sonra hangi metodun kullanılacağına karar vermek SAB çalışmalarındaki önemli bir basamağı oluşturmakatdır. SAB çalışmaları için önerilen

pek çok yaklaşım bulunmaktadır. Bayes türü olasılıksal (Bayesian probabilistic) algoritmalar (Bruce ve Wiebe, 1994; Gale ve diğ., 1992; Mooney, 1996; Leacock ve diğ., 1993; Pedersen ve Bruce, 1997; Yarowsky, 1992), sinir ağları (neural networks) (Leacock ve diğ., 1993; Mooney, 1996), karar ağaçları (decision trees-(DTr)) (Mooney, 1996; Pedersen, 2001; Yarowsky, 2000), bellek tabanlı öğrenme (BTÖ-memory based learning (MBL)) (Ng, 1997; Mooney, 1996; Ng ve Lee, 1996; Cardie, 1993; Veenstra ve diğ., 2000) gibi yöntemler bu alanda en sık başvurulanlardır. BTÖ için seçenek olarak Exemplar-based (EB), instance based (IB), case-based(CB), similarity-based(SB) ve lazy learning (LL) terimleri de kullanılmaktadır.

Bundan sonraki bölümlerde, bu yöntemler arasından seçilen ve çalışmada kullanılan yöntemler hakkında daha ayrıntılı bilgiler verilmiştir.

3.5.1. Karar Ağaçları

Karar ağaçları (decision trees) büyük veri örneklerini hiyerarşik bir yapıya dönüştürmede kullanılmaktadır. Bu algoritmalar sınıflandırma için kullanılabildiği gibi, kullanılan eğitim verilerinin kural kümelerine dönüştürülmesinde de yararlıdır. Karar ağaçları, öğe bulma (Marquez, 1999), makine çevirisi (Tanaka, 1996), çözümleme (Haruno ve diğ., 1998), metin sınıflandırma (Weiss ve diğ., 1999) ve SAB (Brown ve diğ., 1991) gibi DDİ'nin pek çok alt alanındaki uygulamalarda yer almıştır.

Karar ağaçları sembolik tümevarım algoritmalarıdır. Temelde ID3 (Induction Decision Tree) algoritmasından geliştirilmiş pek çok karar ağacı algoritması vardır. ID3 Ross Quinlan tarafından geliştirilen ilk karar ağacı algoritmasıdır ve denetimli bir yöntemdir (Quinlan, 1986). C4.5 ve CART (Classification and Regression Trees) gibi karar ağacı algoritmaları SAB çalışmalarında da kullanılmaktadır (Breiman ve diğ., 1984). CART seksenli yılların başlarında Breiman ve diğer bazı istatistikçiler tarafından geliştirilmiş olan bir karar ağacı algoritmasıdır. Makine öğrenme algoritmaları içerisinde çok sık rastlanan bir yöntem C4.5 tümevarım makine öğrenme yöntemidir (Quinlan, 1993). C4.5, Quinlan tarafından ID3 algoritması baz alınarak geliştirilmiştir. Sürekli ve ayrık veriyi işleyerek sınıflandırma yapabilmektedir. Ayırma işleminde bilgi kazanımını esas

alır ve işlemler bittikten sonra oluşturulan karar ağacı üzerinde budama yapar. 1990'lı yıllardan bu yana yaygın olarak kullanılmaktadır. C4.5 şu şekilde özetlenebilir:

- C4.5 sınıflandırma yapan bir tümevarım tekniğidir. Sınıflandırma yapılabilmesi için modelin gereksinimleri şunlardır:
    - Analizi yapılacak verinin bazı özellikleri bulunmalıdır. Bu özellikler sayısal veya ayrık değerlere sahip olmalıdır. Verinin bu özelliklere karşılık gelen değerleri bulunmalıdır.
    - Verinin doğru sınıflandırılması önceden yapılmalıdır. Sınıflar arasında net ayrımlar bulunmalıdır. Yani bir verinin hangi sınıfa ait olduğu kesin olarak bilinebilmelidir.
    - İstatistiksel testlerin etkin olabilmesi için yeterince veri bulunmalıdır.
- Sonuçta program karar ağaçları veya üretim kuralları çıkarmaktadır.

Karar ağaçlarında düğümler ve bağlantı okları vardır. Karar ağacındaki düğüm yapraksa bir sınıfa ulaşılmış demektir. Ara düğümlerde ise bir özelliğin değerine göre düğüm altında alt dallar bulunmaktadır. Ara düğümde gereken özelliğe göre özelliğin değeri test edilir ve uygun olan daldan devam edilir. Ara düğümlerde yapılan bu testlerden sonra bir yaprağa gelindiğinde elimizdeki verinin sınıfı bulunmuş olur.

3.5.2. İstatistiksel Yöntemler

İstatistiksel yöntemler de sembolik yöntemlere benzerler. Veri üzerinde istatistiksel analizler yaparak istatistiksel sınıflandırıcı modelleri oluştururlar. Bu alanda kullanılan istatistiksel yöntemlerden en sık kullanılanı NB'dir. Sınıfların şartlı olasılıklarını hesaplayan basit bir tümevarım algoritmasıdır. Bir örnek verildiği zaman o örnek için en yüksek olasılığa sahip sınıfı seçer (Domingos ve Pazzani, 1997). Yöntemlerde kullanılan bazı notasyonları Tablo 3.7 vermektedir.

Bu yaklaşımda belirginleştirilecek sözcüğün etrafındaki geniş bir alana bakılır. Her bir bileşenin hangi anlamın seçileceğine dair verilecek kararda etkisi vardır. Herhangi bir özellik seçimi yapmaz. Bunun yerine bütün özelliklerden gelen bilginin kombinasyonunu kullanır. Her bir belirginleştirilecek sözcük için doğru anlamının işaretlendiği bir derleme metnin kullanıldığını varsayar. Bayes sınıflandırıcı Bayes karar kuralını uygular ve hatayı en aza indirgemeye çalışır.

Tablo 3.7:Algoritmalarda kullanılan notasyon

Kısaltma	**Anlamı**
W	belirginleştirilecek sözcük
$s_1, s_2, \ldots s_k, \ldots s_K$	belirginleştirilecek sözcük w'nin anlamları
$c_1, c_2, \ldots c_i, \ldots c_I$	w'nin çevresindeki sözcükler
$v_1, v_2, \ldots v_j, \ldots v_J$	w'nin anlamını belirginleştirmek için çevresindeki sözcüklerin özellikleri olarak kullanılan sözcükler

Bayes karar kuralı şu şekildedir: Eğer bir $s' \neq s_k$ için $P(s'|c)>P(s_k|c)$ ise anlam olarak s' seçilir. Genelde $P(s_k|c)$'yi bilmeyiz ancak Bayes kuralı ile hesaplayabiliriz:

$$P(s_k \mid c) = \frac{P(c \mid s_k)}{P(c)}(P(s_K) \tag{3.1}$$

P(c) tüm anlamlar için sabittir ve elenebilir

$P(s_k)$ etrafındaki sözcükler bilinmeden s_k anlamının olasılığıdır

Yapılmak istenen ise w için

$$\begin{aligned} s' &= \arg\max_{s_k} P(s_k \mid c) \\ &= \arg\max_{s_k} \frac{P(c \mid s_k)}{P(c)} P(s_K) \\ &= \arg\max_{s_k} P(c \mid s_k) P(s_K) \\ &= \arg\max_{s_k} [\log P(c \mid s_k) + \log P(s_K)] \end{aligned} \tag{3.2}$$

NB'de kullanılan özelliklerin birbirinden bağımsız olduğu varsayılır ve yapı ile doğrusal sıralama önemsenmez. Buna göre NB varsayımı:

$$P(c \mid s_k) = P(\{v_j \mid v_j \in c\} \mid s_k) = \coprod_{v_j \in c} P(v_j \mid s_k) \tag{3.3}$$

NB karar kuralı ise eğer

$$s' = \arg\max[\log P(s_k) + \sum_{v_j \in c} \log P(v_j \mid s_k)] \quad (3.4)$$

ise anlam olarak s' seçilir.

$P(v_j \mid s_k)$ ve $P(s_k)$ en büyük olabilirlik (Maksimum-Likelihood) tahmini kullanılarak hesaplanır. Tablo 3.8 NB algoritmasını göstermektedir.

$$P(v_j \mid s_k) = \frac{C(v_j, s_k)}{C(s_k)} \text{ ve } P(s_k) = \frac{C(s_k)}{C(w)} \quad (3.5)$$

$C(v_j, s_k)$ v_j 'nin s_k anlamıyla beraber derleme metinde geçiş sayısı.

$C(s_k)$ s_k anlamının derleme metinde geçiş sayısı

C(w) w sözcüğünün derleme metinde geçiş sayısı

NB'de kullanılan özelliklerin birbirinden bağımsız olduğu varsayılır ve yapı ile doğrusal sıralama önemsenmez. NB yöntemini geliştirmek üzere pek çok yöntem önerilmiştir. Bu yöntemler genel olarak NB'deki özelliklerin birbirinden bağımsız olduğu varsayımının zayıflatılması üzerinde durmaktadır (LBR-Lazy Bayesian Rules ve TAN-Tree Augmented Naive Bayes) (Friedman ve diğ., 1997; Wang ve Webb, 2002). Ancak bu yöntemlerde hata oranı azaltılmasına karşılık hesaplama zamanı çok artmıştır. AODE (Aggregating one-dependence estimators) NB'deki özelliklerin birbirinden bağımsız olduğu varsayımının getirdiği olumsuzlukları, özelliklerin sadece sınıfa ve diğer bir özelliğe bağımlı olduğu bütün modellerin ortalamasını alarak ortadan kaldırır. Sonuçta ortaya çıkan sınıflandırma öğrenme algoritması hem hesaplama açısından çok etkin olmakta hem de hata oranı düşmektedir (Webb ve diğ., 2002).

Tablo 3.8:Naive Bayes Algoritması

1. **açıklama:** Öğrenme
2. **for** $\forall$ s_k **do**
3. **for** $\forall$ v_j **do**
4. $$P(v_j \mid s_k) = \frac{C(v_j, s_k)}{C(s_k)}$$
5. **end**
6. **end**
7. **for** $\forall$ s_k **do**
8. $$P(s_k) = \frac{C(s_k)}{C(w)}$$
9. **end**
10. **açıklama: Belirginleştirme**
11. **for** $\forall$ s_k **do**
12. skor(s_k) =log$P(s_k)$
13. **for** $\forall$ v_j **do**
14. skor(s_k) = skor(s_k) + log$P(v_j \mid s_k)$
15. **end**
16. **end**
17. **seç**
$$s' = \arg\max_{s_k} skor(s_k)$$

3.5.3. Örnek Tabanlı Sınıflandırıcılar

Örnek tabanlı sınıflandırıcılar veriyi genellemeye çalışmaz, bunun yerine veriyi belleğinde uygun bir biçimde saklayarak bunları yeni gelen ve sınıfı bilinmeyen örnekleri sınıflandırmakta kullanır. Bu işlem sırasında belleğinde saklanan örneklerden yeni gelen örneğe en yakın olanının sınıfı işlem sonucu olarak atanır.

Burada temel varsayım benzer örneklerin benzer sınıflara ait olacağıdır. Temel varsayımdaki anahtar sözcükler olan benzer örnek ve benzer sınıf kavramının belirlenmesi bu yöntemlerdeki en kritik noktadır. Bu yöntemlerde iki ana bileşen vardır:

- Birincisi; iki örneğin birbirine ne kadar benzediğini ya da diğer bir deyişle birbirine olan uzaklığı ölçen uzaklık fonksiyonudur.
- İkincisi ise; bu uzaklık fonksiyonunu kullanarak yeni gelen örneğin sınıfını atayan sınıflandırma fonksiyonudur.

Bunlara ek olarak, örnek tabanlı yöntemlerde yeni yapılan sınıflandırma doğru ise, bunu da veritabanına ekleyen ve hangi örneklerin sınıflandırmada kullanılacağına karar veren sınıf güncelleyicisi bulunur.

En yakın komşu algoritmaları (Cover ve Hart, 1967) örnek tabanlı yöntemlerin en basit şeklidir. Alana özgü uzaklık fonksiyonları kullanarak en benzer örneği eğitim verisinden elde ederler ve elde edilen örneğin sınıfını sonuç olarak atarlar.

k-nn (Hart, 1968) algoritmaları biraz daha karmaşıktır. Sınıfı bulunacak örneğe en yakın olan k tane örnek seçilir ve bunlar arasından en baskın olanı yeni örneğin sınıfı olarak atanır. KStar yönteminde iki örnek arasındaki uzaklığın hesaplanması için enformasyon teorisi kullanılmıştır. İki örnek arası uzaklık bir örneğin diğer bir örneğin diğer örneğe dönüştürülebilmesi için gereken hesaplama olarak kabul edilmiştir. Bu hesaplama iki adımda gerçekleştirilmiştir. Birinci basamakta örnekleri örneklere eşleyen bir dizi sonlu sayıda dönüşüm tanımlanmıştır. Bir *a* örneğini *b* örneğine dönüştürecek olan program *a* ile başlayıp *b* ilen biten bir dizi dönüşüm izlemektedir. Bu dönüşümlerde en kısa olan yolun seçilmesi sağlanmaktadır. Sonuçta ortaya çıkan uzaklık ölçütü örnek uzayındaki en küçük değişikliklere bile çok duyarlı olmaktadır (Cleary ve Trigg, 1995).

Yapılan çalışmada uygulanan ilk örnek tabanlı yöntemde Ng ve Lee'nin (1996) çalışması örnek alınmıştır. Bu çalışmada her bir örnek tümce için bazı özellikler seçilmekte ve kaydedilmektedir. Daha sonra test edilecek her bir tümce için de aynı özellik kümesi bulunmakta ve daha önce tespit edilen örneklerin özellikleriyle karşılaştırılmaktadır. Özellikleri en yakın olan örnek tümcedeki anlam test edilen tümcedeki sözcüğün de anlamı olarak seçilmektedir. Algoritmanın en önemli kısmı özelliklerin seçilmesi ve özellikler arası uzaklığın hesaplanmasıdır.

Bir f özelliğinin iki sembolik değeri v_1 ve v_2 arasındaki uzaklık için şu formül kullanılmıştır:

$$d(v_1, v_2) = \sum_{i=1}^{n} \left| \frac{C_{1,i}}{C_1} - \frac{C_{2,i}}{C_2} \right| \tag{3.6}$$

$C_{1,i}$ = f özelliği v_1 olan ve s_1 anlamında olan tümce sayısı
$C_{2,i}$ = f özelliği v_2 olan ve s_1 anlamında olan tümce sayısı
C_1 = f özelliği v_1 olan herhangi bir anlamda olan tümce sayısı
C_2 = f özelliği v_2 olan herhangi bir anlamda olan tümce sayısı
n = Toplam anlam sayısı

İki tümce arasındaki uzaklık bütün özellikler arasındaki uzaklığın toplamına eşittir. Eğer uzaklıkları eşit birden fazla örnek bulunursa herhangi biri rastgele seçilmektedir

3.5.4. Diğer Yöntemler

Doğrusal sınıflandırıcılar (linear classifiers) genelde iki sınıfı bulunan problemler için basit ve etkin hesaplama yöntemleridir. Örneğin, doğrusal eşik algoritmaları verilen özelliklerden faydalanarak ağırlıklı toplam hesaplarlar ve elde edilen toplamın belirli bir eşik değerinden büyüklük/küçüklüğüne göre verinin sınıfı hakkında bir karara varırlar. Eşik değerleri eğitim verisi kullanılarak oluşturulur. Bu tip sınıflandırıcılar metin sınıflandırma (Dagan ve diğ.,1997; Levis ve diğ., 1996), yüzeysel çözümleme (Muonoz ve diğ., 1999) ve öğe bulma (Roth ve Zelenko, 1998) gibi DDİ alanlarında kullanılmaktadır.

Kümeleme (clustering) tamamen denetimsiz bir yöntem olduğu için güçlü bir tekniktir. Veri içerisindeki doğal bölümleri verilen elemanlar arasındaki benzerlikleri kullanarak gruplandırmaya çalışır. Biçimbilimsel (Hughes, 1994) ve anlamsal sınıflandırma, bilgi çıkarımı (Ibrahimov ve diğ., 2001) alanlarında yapılan çalışmalar bulunmaktadır.

Bellek tabanlı (memory-based) yöntemler de isminden de anlaşılacağı gibi daha önce karşılaşılan bütün örneklerin kullanımının bellekte saklanmasıyla gerçekleştirilen bir sınıflandırma yöntemidir. Bu tekniklerde en önemli nokta, kural tabanlı sistemlerin aksine yeni gelen örneklerle geçmişteki örneklerin benzerlik oranına göre öğrenmenin gerçekleşmesidir. Bu sistemlerin performansı sınıflandırma aşamasında ortaya

çıkmaktadır. Bu tekniklerin temelinde k-nn (k-nearest neighbour) algoritması (Cover ve Hart, 1967) yatmaktadır.

Tümevarımsal Mantık Programlama (Inductive Logic Programming) yöntemleri ise BÖ kullanarak verilen alanda bilgi edinir ve elde edilen bilgiyi birinci dereceden mantık kurallarına dönüştürür.

Yapay sinir ağları (neural networks-YSA) yapay zeka alanında kullanılan çok karşılaşılan ve oturmuş yöntemlerden birisidir. DDİ çalışmalarında da sıkça kullanılmaktadır. YSA'lar birbirlerine bağlanmış birimlerden oluşur. Bu bağlantılara değerler atanmıştır. Bu birimler gelen veri üzerinde basit hesaplamalar yaparak bunları çıkışlara yönlendirir. Eldeki probleme göre bu birimler ağ içerisinde uygun olan bir topoloji ile bağlanabilir. Gizli birimlerin katmanları oluşturulduktan sonra daha karmaşık problemlerin ifade edilmesi kolaylaşır. Öğrenme aşaması geriye doğru yayılma algoritması gibi bir yöntemin kullanıldığı eğitimle olmaktadır. Başarımı arttırmak için bağlantıların ağırlıkları güncellenebilir. Konuşma işleme ve çözümleme alanlarında örnekler vardır.

Arttırma (boosting) algoritmalarında ise ana fikir pek çok basit ve kısmen kesin olan hipotezleri (zayıf sınıflandırıcıları) tek ve kesinliği daha fazla bir sınıflandırıcı olacak şekilde birleştirmektir. Bu algoritmalar temelde AdaBoost.MH (Freund ve Schapire, 1996) algoritmasını kullanır. LazyBoosting algoritmaları ise bir zayıf sınıflandırıcı öğrenildiği zaman özellik uzayının boyutu azaltmaya çalışan Ada.Boost üzerine kurulmuş yöntemlerdir. Bunun için de küçük bir özellik kümesi rastgele seçilir ve bunlar için en uygun olan zayıf sınıflandırıcı bulunur. Öğe bulma, (Abney, 1999), metin sınıflama (Schapire ve Singer, 2000) gibi DDİ alanlarında kullanılmaktadır.

3.5.5. Bilgisayarla Öğrenme Algoritmaları ile Karşılaştırmalı Çalışmalar

DDİ alanında yapılan araştırmalarda genellikle makine öğrenme algoritmalarının farklı şekillerde karşılaştırmaları yapılmaktadır. Bu çalışmalarda, değişik bilgi kaynaklarının öğrenmeye etkisi veya bir yöntemin yaklaşımının DDİ alanında seçenek olabilecek diğer yöntemlere göre daha uygun olup olmadığı araştırılmaktadır. Birinci amaç için,

sonuçlarda ortaya çıkacak istatistiksel olarak önemli bir artma veya azalma olup olmadığının ölçülebilmesi amacıyla belirli bir bilgi kaynağıyla ve bu kaynak olmadan yapılan deney sonuçları karşılaştırılmaktadır. Örneğin, bu amaçla SAB için eylemlerde önce gelen sözcüğün aldığı ekin, eylemin anlamını belirginleştirmedeki etkisini ölçmek için bu bilgi kaynağı kullanılarak NB yöntemiyle elde edilecek sonuçlar, yine bu kaynak kullanılmadığı durumda NB kullanılarak elde edilecek sonuçlarla karşılaştırılarak doğruluk oranındaki değişim gözlemlenebilir.

İkinci amaç için ise farklı metodolojiler benimseyen iki veya daha fazla öğrenme yöntemi aynı veri üzerinde denenerek yöntemlerin başarım oranları karşılaştırılır. SAB çalışması için örnek verecek olursak anlam belirsizliği olan farklı sözcükler için oluşturulan öğrenme verileri NB ve EBL iki yöntemde denenerek bu iki yöntemden hangisinin belirginleştirmede daha etkin sonuçlar verdiği ölçülür. Bu karşılaştırmaların nasıl yapılacağı ile ilgili ayrıntılı bir inceleme Weiss ve Indurkhya'nın (1998) çalışmasında bulunabilir.

Karşılaştırmalı olarak başarımın test edilmesi sadece elimizdeki problemde uygulanabilmesi yönüyle değil, uluslararası alanda yapılan karşılaştırmalı değerlendirme yöntemleriyle ilgili Senseval ve CoNLL gibi projelerde de giderek önem kazanan bir konu olmuştur. Bu çalışmalarda sistemler aynı eğitim ve test verileriyle çalıştırılarak birbirleriyle karşılaştırılmakta ve değerlendirme yöntemleri için standartlaşma sağlanmaya çalışılmaktadır.

Son yıllarda SAB sistemlerinin başarımını farklı sözcüklerde, farklı kriterlere ve farklı dillere göre ölçmek üzere kontrollü ortamlarda karşılaştırmalı yarışmalar düzenlenmekte (Edmonds ve Kilgarriff, 2003; Kilgarriff ve Palmer, 1999) ve ortaya çok fazla sayıda bahsedilen iki temel karşılaştırmanın yapıldığı çalışmalar çıkmaktadır (Escudero ve diğ., 2000; Lee ve Ng, 2002; Mooney, 1996; Ng ve Lee, 1996). Sınıflandırıcının başarımını ölçmek veya bilgi kaynaklarının katkısını test etmek için k-katlı çapraz doğrulama (Aha ve Goldstone, 1992; Dietterich, 1998; Kohavi ve John, 1997), istatistiksel bir yöntem olan McNemar (Dietterich, 1998) veya ikili çapraz doğrulamalı t-testleri (paired cross-validation t-tests) bazı problemlerine rağmen (Leacock ve diğ., 1993) standart olarak bu alanda kullanılmaktadır.

Karşılaştırma yapılırken eğer yöntemler karşılaştırılacaksa yukarıda bahsedilen faktörler sabit tutularak yöntemler değiştirilmekte ve elde edilen istatistiksel sonuçlardaki farklılıklar gözetilerek hangi yöntemin ve neden eldeki problem için daha uygun olduğu yorumu yapılmaya çalışılmaktadır. Aynı şekilde bilgi kaynakları karşılaştırılırken de bu kez kullanılan yöntem sabit tutulmakta ve özellik kümesinin elemanları değiştirilerek bu değişimin sonuç üzerindeki etkisi ve bunun nedenleri iredelenmektedir. Bunun dışında belki de asıl incelenmesi gereken konu her iki karşılaştırmanın birlikte yapılarak birbiri üzerindeki etkisini görmek de olabilir.

Hesaplamalı dilbilim alanındaki karşılaştırmalı yöntemler genelde algoritmalarda olası bir veya bir kaç noktayı ele almaktadır. Doğruluk oranlarının karşılaştırıldığı bu çalışmalarda ele alınan faktörlere göre sonuçlar ve yapılabilecek yorumlar oldukça farklılık gösterebilmektedir. SAB üzerine yapılan çalışmalara (Veenstra ve diğ., 2000; Daeleman ve diğ., 2002) ait bazı sonuçlar bellek tabanlı bir yöntemin parametrelerinin ayarlanması ile iyileştirilmiştir (Daeleman ve diğ., 2003). Bunun sadece bu yöntemle sınırlı kalmayıp başka yöntemlerde de iyileştirme sağlayacağı ve özellik seçimi ile algoritmalarda kullanılan parametrelerin çok yüksek düzeyde etkileşime sahip olduğu ve birlikte ele alınması gerektiği de aynı makalede gösterilmiştir. Karşılaştırmalı sonuçların göreceli olduğu ve literatürde yayınlanan sonuçların çok da güvenilir olmadığı savı ortaya konmuştur. Bu görüşü destekleyen ve literatürde önemli etkisi bulunan başka bir çalışmada (Banko ve Brill, 2001) *"Bir milyon sözcükten oluşan bir veriden elde edilen karşılaştırma sonuçlarının daha büyük bir derleme metin kullanıldığında elde edilecek sonuçlarla uyumlu olmasını beklemek için elimizde herhangi bir dayanak bulunmamaktadır"* görüşüne yer verilmektedir. Bu görüşlerini desteklemek için de bir milyon sözcükten bir milyar sözcüğe kadar içeriği bulunan farklı boyuttaki derleme metinler kullanılmış, farklı makine öğrenme algoritmaları tipik bir DDİ uygulaması olan SAB konusunda denenmiş ve elde edilen sonuçlar belirtilen görüşü destekler nitelikte olmuştur. Karşılaştırma sonuçlarını sadece derleme metin boyutunun değil, pek çok faktörün etkilediği deneylerle de gösterilmiştir (Daeleman ve diğ., 2003). Bu çalışmalar, karşılaştırmalı çalışmaların güvenilirliğine önemli şüpheler getirmekle birlikte, doğruluk oranlarının iyileştirilmesi için hem algoritmaların

parametrelerinin, hem bilgi kaynaklarının, hem de her ikisinin birlikte iyileştirilebileceğini göstermesi açısından umut verici olmuştur.

Yöntemlerin karşılaştırılması ile ilgili de çalışmalar yapılmıştır. Bu karşılaştırmalardan ilginç sonuçlar elde edilmiştir (Daeleman, 2002). Bu yöntemlerden bazılarının başarımları için bir çalışmada varılan sonuçlara göre NB ve sinir ağları en başarılı olurken, bunları karar ağaçları takip etmiş ve en başarısız olanı ise BTÖ olmuştur (Mooney, 1996). Buna karşın, başka bir çalışmada BTÖ yönteminin NB yönteminden daha başarılı olduğu gözlemi yapılmıştır (Escudero ve diğ., 2000) ki, bu sonuçlar birbiriyle çelişki içinde bulunmaktadır. Bunun dışında başka bir karşılaştırma çalışmasında ise yedi farklı yöntem için elde edilen sonuçlarda, Support Vector Machines (SVM) yöntemi, içinde NB ve karar ağaçlarının da bulunduğu diğer yöntemlerden daha başarılı olarak gösterilmiştir (Lee ve Ng, 2002).

Birbiriyle çok da uyumlu olmayan bu sonuçlar göz önüne alındığında, asıl önemli olan konunun seçilen yöntem değil, bu yöntemlerde kullanılan ve etkili olan parametreler olduğu sonucuna varılabilir. Bu parametrelerden bazıları da özet olarak şu şekilde sıralanmıştır (Daeleman ve diğ., 2003):

- Bilgi kaynakları: Özellik seçimi ve gösterimi
- Algoritmanın kendi parametreleri: Varsayılan ve iyileştirilen değerler
- Eğitim verisi: Seçimi ve büyüklüğü
- Yöntem kombinasyonları
    - Birarada kullanım (bagging)
    - Birinden elde edilen sonuçları diğerine aktarma (boosting)
    - Çıktıyı kodlama (output coding)
- Test için kullanılan yöntem
    - k-katlı çapraz doğrulama
    - McNemar
    - İkili çapraz doğrulama t-testi
    - Öğrenme eğrileri

Buradan çıkarılabilecek sonuç olarak karşılaştırmalı çalışmaların, ancak ve ancak sabit özellikler ve algoritma parametreleri kullanıması durumunda güvenilebilir olacağı vurgulanmış ve bir algoritma veya bir alandaki problem için en iyi özellik veya en iyi

parametre ayarlarını bulmanın çok da kolay olmadığının altı çizilmiştir (Daeleman, 2002). Bu çalışmada da yöntemler arası karşılaştırmalar Dördüncü Bölüm'de yapılmıştır.

3.5.6. WEKA Projesi

Mevcut standart BÖ teknikleri WEKA (Waikato Environment for Knowledge Analysis) adı verilen bir projede Java programlama dili ile yazılarak bir kütüphane oluşturulmuştur (Witten ve Frank, 1999). Bu proje sayesinde bilinen pek çok BÖ yöntemi kullanıcılar tarafından farklı problemlerde rahatlıkla kullanılabilecek bir konuma getirilmiştir. Özellikle analiz yapmanın zor olduğu çok büyük verilerin incelenmesi gereken uygulamalarda WEKA projesi büyük kolaylıklar sağlamaktadır.

Projenin Java'da yazılması ortamdan bağımsızlık sağlamıştır. Yazılan kodlar kullanıma açıktır, bu kodlar kullanılarak yeni teknikler yazılmasına da olanak sağlayan pek çok destek kodu da proje kapsamında bulunmaktadır. Algoritmalar doğrudan verilere uygulanabilmekte veya başka Java programları içinden de çağrılabilmektedir. Proje içerisinde veri üzerinde ön işlem yapma, sınıflandırma, bağlanım (regression), gruplama, birleştirme, kurallar oluşturma ve görselleştirme sağlayan araçlara da yer verilmiştir. Ayrıca elde edilen sonuçlar, hata oranları ve uygulanan algoritmanın türüne göre yararlı pek çok bilgi de çıktı olarak edilebilmektedir (Tablo 3.9). WEKA projesi içerisinde kullanıma sunulan algoritmaların listesini Şekil 3.3 göstermektedir.

Çalışmanın bundan sonraki aşamalarında ODTÜ-Sabancı derleme metninin ve WEKA'nın sağladığı avantajlar kullanılarak Türkçe SAB için uygun olan algoritma ve diğer kriterlerin seçilmesine çalışılmıştır. WEKA'nın sağladığı uygun olan bütün algoritmalar elde edilen metinler üzerinde uygulanarak elde edilen sonuçların değerlendirilmesi amaçlanmıştır. Değerlendirmeler yapılırken farklı derleme metinlerin vereceği sonuçlar, değişik sözcük türlerinin (isim, eylem, sıfat, vb.) SAB işlemi için etkin olan faktörlerin belirlenmesi, belirlenen faktörlerin etki oranları, algoritmalardan uygun olanları ve uygunluk dereceleri gibi çok kapsamlı bir çalışma yapılarak Türkçe için ilk defa yapılmış bu incelemede bu konuda yapılacak daha sonraki çalışmalar için de kaynak olabilecek çıkarımlar ortaya konmaya çalışılmıştır.

Tablo 3.9:WEKA j48 algoritmasının örnek veriye uygulanışı ve elde edilen sonuçlar

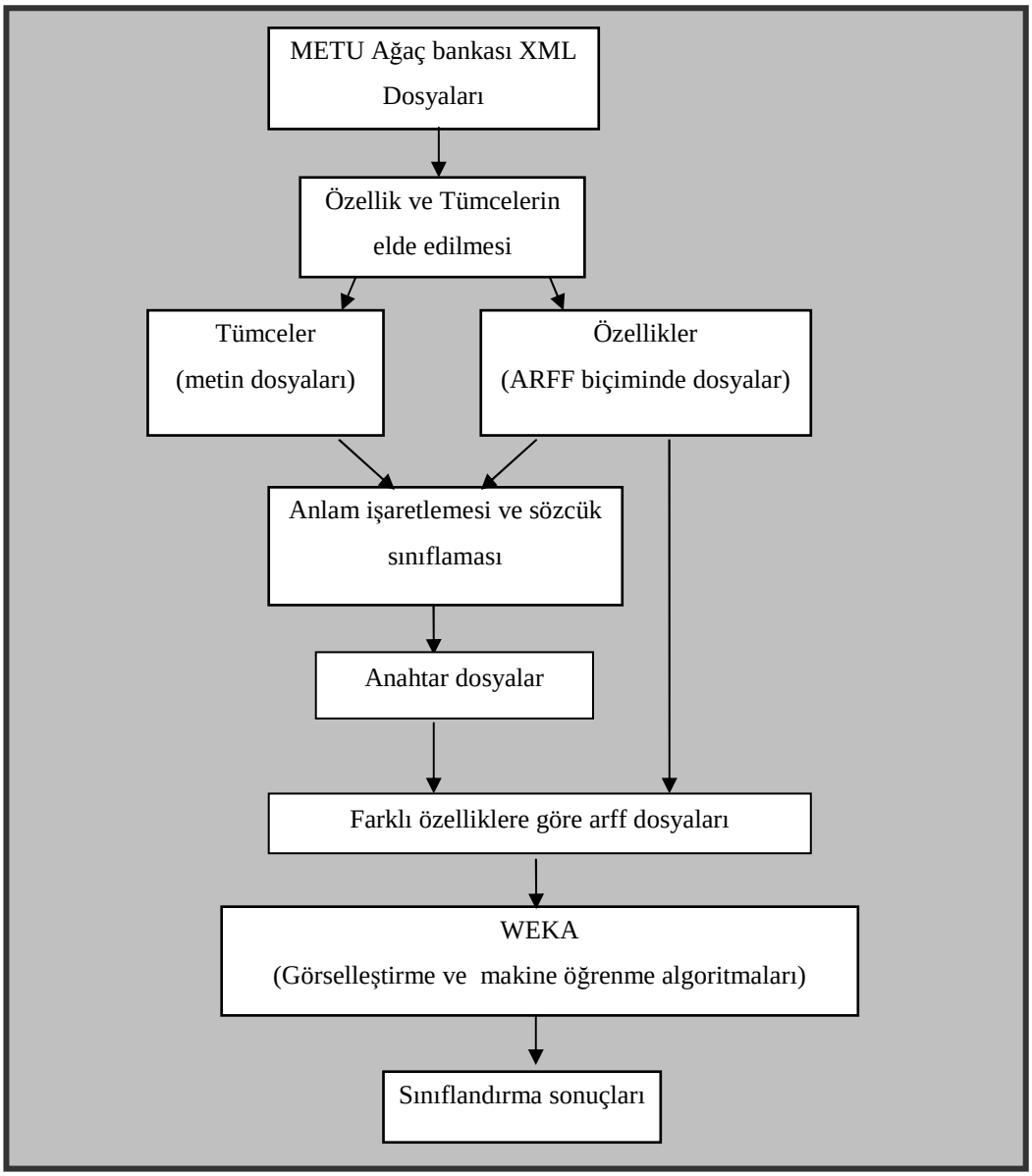
```
=== Run information ===
Scheme:
weka.classifiers.trees.j48.J48 -C 0.25 -M 2
Relation:     weather
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
Test mode:    10-fold cross-validation
=== Classifier model (full training set)
J48 pruned tree
outlook = sunny
|   humidity <= 75: yes (2.0)
|   humidity > 75: no (3.0)
outlook = overcast: yes (4.0)
outlook = rainy
|   windy = TRUE: no (2.0)
|   windy = FALSE: yes (3.0)
Number of Leaves  :           5
Size of the tree :  8
```

```
Time taken to build model: 0.14 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances          9               64.2857 %
Incorrectly Classified Instances        5               35.7143 %
Kappa statistic                          0.186
Mean absolute error                      0.2857
Root mean squared error                  0.4818
Relative absolute error                 61.5385 %
Root relative squared error            100.4843 %
Total Number of Instances               14

=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   Class
                 0.778    0.6        0.7       0.778    0.737      yes
                 0.4      0.222      0.5       0.4      0.444      no

=== Confusion Matrix ===
 a b   <-- classified as
 7 2 | a = yes
 3 2 | b = no
```

Şekil 3.3:WEKA algoritmaları

3.5.7. Sistem Mimarisi

Türkçe SAB çalışmada kullanılan sistem şu an için bir kaç bölümden oluşmaktadır. Sistem mimarisini Şekil 3.4 göstermektedir.



Şekil 3.4: Sistem mimarisi

ODTÜ'den alınan Türkçe ağaç bankası XML dosyaları şeklindedir. Bu dosyalarda farklı dökümanlardan alınmış tümceler bulunmaktadır. Bu XML dosyaları SAB çalışmasının girdi verilerini oluşturmaktadır. Bu veri Java programı ile alınıp tüm tümceler veya verilen kök sözcüğü içeren tümcelerle birlikte (Tablo 3.10) bu tümcelere ait bazı özellikler de elde edilebilmektedir. Sistemin kullandığı özellikler arasında tümce numarası, incelenen sözcükle ilişkili olan diğer sözcüklerin kökleri, ekleri, tipi, iki sözcük arasındaki ilişki, incelenen sözcüğün tipi, kendisinin ilişkide olduğu diğer sözcükle aralarındaki ilişki vb. bulunmaktadır (Tablo 3.11). Bu özellikler kullanılarak WEKA sisteminde kullanılabilecek girdi dosyaları ARFF biçimine göre hazırlanmaktadır (Tablo 3.12). Bu dosyadaki *?* ile gösterilen değerler metinden bu

değerlerin (problemli durumlar veya bu değerlerin olmaması nedeniyle) alınamadığını göstermektedir. ARFF biçiminde incelenecek yapıya bir isim verilmekte (@relation gel: gel kökü inceleniyor), daha sonra seçilen bütün özelliklerin alabilecekleri değerler kümesi aşağıdaki gibi sıralanmaktadır.

Tablo 3.10: Aranan kökü içeren tümcelerin listesi

00002213112.xml--1) Neden geldiniz .
0000221313.xml--2) Bağışlayın , koşa koşa geldim buraya .
0000221314.xml--3) Hepimiz kıraathaneye gelen bu yabancıya dikkatle bakıyorduk .
00002213172.xml--4) Geldik ! diye bağırdı Kerem .
0000221319.xml--5) Yeni gelen , masanın bir kenarındaki boş sandalyeye atmıştı kendini .
0000221322.xml--6) Gözleri kıraathanenin loş ışığına alışmış , sanki biraz kendine gelmişti .
0000221323.xml--7) Erkekler Parkı'ndan geliyorum .
0000221333.xml--8) Parktaki tüm erkekler onun gelmesini bekliyorlardı .
0000221333.xml--9) Birden , geldiğini duydum , parktaki kalabalık karıştı .
0000221333.xml--10) Galiba gelmemiş , öyle söylediler , dedim .
0000221334.xml--11) Öyle kolay gelmez ki o .
0000221335.xml--12) Sanırım o zaman da gelmemişti .
0000221337.xml--13) Bir insanın ruhundan geliyorum ben , dedi .
0000221337.xml--14) Kurtulup buraya gelmeyi başardım .
0000221339.xml--15) O , her an gelecek gibidir ; her an .
.
.
.

Tablo 3.11: Algoritmalarda kullanılan özellikler

ARFF Doyasındaki Değişken Adı	Alan Adı	Örnekteki Değeri
Vectors0	Tümce no	2
Vectors1	Önceki sözcük kökü	BURA
Vectors2	Önceki sözcük tipi	NOUN
Vectors3	Önceki sözcük eki	DAT
Vectors4	Önceki sözcük-hedef sözcük arası ilişki tipi	OBJECT
Vectors5	Hedef sözcük tipi	VERB
Vectors6	Sonraki sözcük-hedef sözcük arası ilişki tipi	SENTENCE
….	….	…
Vectorsk-4	Sonraki sözcük kökü	?
Vectorsk-3	Sonraki sözcük tipi	?
Vectorsk-2	Sonraki sözcük eki	?
Vectorsk-1	Sonraki sözcük ilişki tipi	?
Vectorsk	Anlam No	1

(*@attribute vectors*$_1$ {değer$_1$,değer$_2$,..., değer$_x$}

@attribute vectors$_2$ {değer$_1$,değer$_2$,..., değer$_y$}

.

.

.

@attribute vectors$_k$ {değer$_1$,değer$_2$,..., değer$_z$}

Son olarak da *@data* ile başlayan kısımda elde edilen veriler yazılmaktadır. *gel* sözcük kökünün ikinci geçtiği *Bağışlayın , koşa koşa geldim buraya .* tümcesinde *koşa* ve *buraya* sözcükleri *gel* sözcüğü ile ilişkilidir. İki sözcük ilişkili olduğu için özellikler kısmında ikinci tümce için iki veri bulunmaktadır.

2,KOŞA,ADJ,?,MODIFIER,VERB,SENTENCE,…,?,?,?,?,1

2,BURA,NOUN,DAT,OBJECT,VERB,SENTENCE,…,?,?,?,?,1

Bu şekilde bir yazım ile her sözcük için farklı sayıda ilişkili sözcüklerin özellikleri sabit olarak algoritmalara verilebilmektedir. Tümce numarası bir özellik olarak kullanılıp yerel bağlamın anlama etkisi test edilmeye çalışılmıştır. Ancak kullanılan yöntemler sınıflandırmada sadece bu özelliğe dayalı karar vererek diğer özellikleri göz ardı ettiği için bu özellik sadece anlam işaretlemesine yardımcı olması ve okunabilirliği arttırması için bırakılmış, algoritmalar uygulanırken bu özellik çıkarılmıştır.

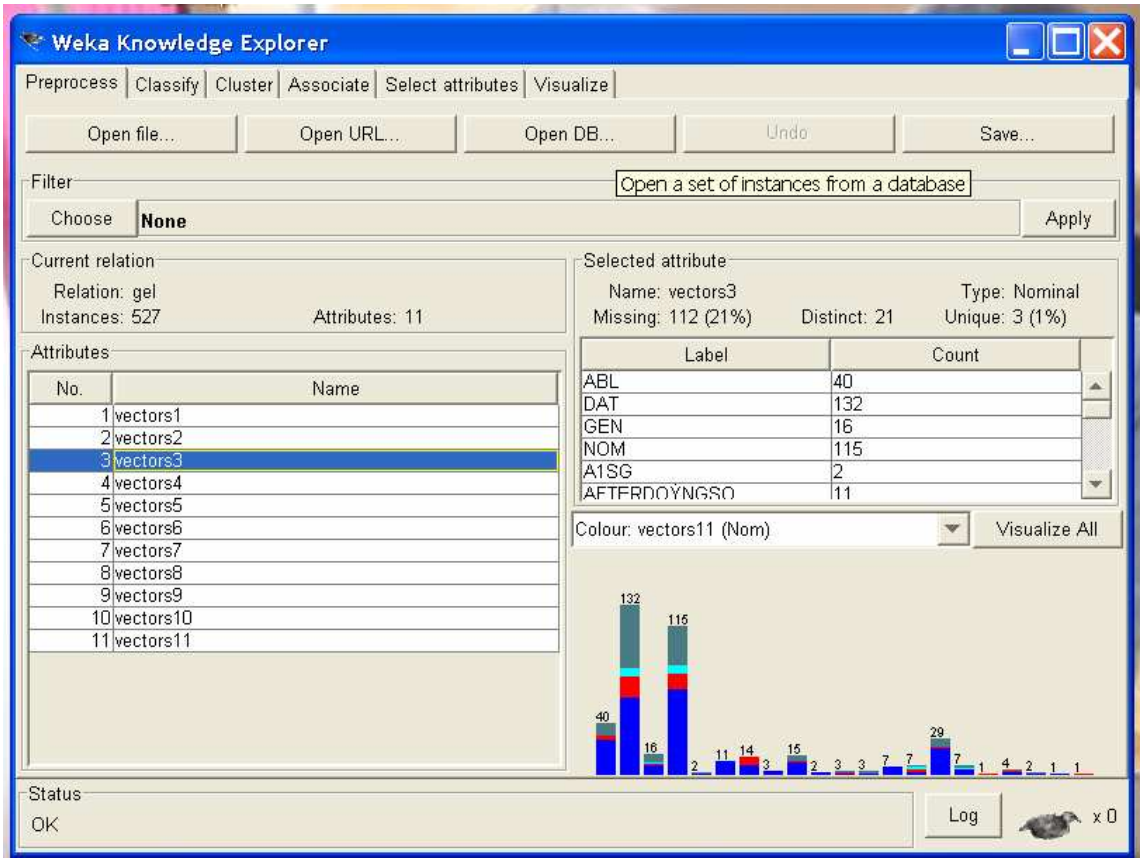
Seçilen bu özelliklerin hepsi anlamı belirlemede eşit ağırlıkta etki etmemektedir. Hatta hedef sözcüğün tipine göre bu özellikler farklılık gösterebilir. Örneğin, eylemlerde genellikle sonraki gelen sözcük anlam üzerinde çok etkili olmazken, isimlerde sonra gelen sözcükler daha etkili olabilmektedir. Eylemler genelde tümce sonunda bulundukları için sonra gelen sözcük ile ilgili özellikler boş olabilmektedir. Yukarıdaki *2* numaralı tümcede olduğu gibi sonraki sözcük özellikleri eylemlerde genelde bilinmeyen veya bulunmayan özellik olarak kullandığımız *?* olarak çıkmıştır. Aşağıdaki tümceler anlama etki eden farklı özelliklerin olduğu durumları göstermektedir:

- Ali okul**a gelecek**. *(Önceki sözcük yönelme hali ve sözcük türü gel anlamını belirlemiş)*
- **Gelecek**te bunlar olmayacak. *(Sözcük türü gel anlamını belirlemiş)*
- **Elini alnı**na koydu. *(Sonraki sözcük anlamda etkili)*
- **Kasaptan et** aldı. *(Sözcük türü ve önceki sözcük anlamı belirlemiş)*

- **Ettiğini** buldu. *(Sözcük türü ve aldığı ek anlamı etkilemiş)*
- Sonunda **kabul etti**. (*Önceki sözcük ve sözcük türü anlamı belirlemiş)*

Tablo 3.12: Tümcelerden elde edilen bilgilerin ARFF biçimi

```
@relation gel
@attribute vectors0 {1, 2, 3,…, 286}
@ attribute vectors1 {NE, KOŞA, BURA, KIRAATHANE, … , BİRLİKTE}
@attribute vectors2 {PRON,ADJ,NOUN,ADV,CONJ,VERB,POSTP,QUES,DET,INTERJ}
@attribute vectors3 {ABL,DAT,GEN,NOM,A1SG,AFTERDOİNGSO,PCNOM,A3SG,INS,A2PL,WHİLE,WHEN,
BYDOİNGSO,ACC, LOC,PCABL,PNON,PCDAT,A2SG,LY,EQU}
@attribute vectors4 {MODIFIER,OBJECT,COORDINATION,COLLOCATION,SUBJECT,INTENSIFIER,
ABLATIVE.ADJUNCT, DATIVE.ADJUNCT,VOCATIVE,QUESTION,POSSESSOR, QUESTION-
PARTICLE,DISCARD,INSTRUMENTAL.ADJUNCT,LOCATIVE.ADJUNCT,
QUESTION.PARTICLE,EQU.ADJUNCT,RELATIVIZER}
@attribute vectors5 {VERB,ADJ,NOUN,ADV}
@attribute vectors6 {SENTENCE,MODIFIER,SUBJECT,OBJECT,COLLOCATION,DISCARD,POSSESSOR,
ABLATIVE.ADJUNCT,DETERMINER,INSTRUMENTAL.ADJUNCT}
…
@attribute vectorsk-4 {YABANCI,KEREM,AT,… ,BELİR}
@attribute vectorsk-3 {NOUN,VERB,POSTP,ADV,ADJ,CONJ}
@attribute vectorsk-2 {DAT,NOM,A3SG,PAST,A1SG,ACC,PCNOM,BYDOINGSO,WHEN,A1PL,
P2SG,LOC,A3PL,A2PL,A2SG, ABL,GEN,PRESPART,PNON,PCDAT,P3PL,P3SG,INS}
@attribute vectorsk-1 {OBJECT,SUBJECT,SENTENCE,CLASSIFIER,MODIFIER,COORDINATION,
LOCATIVE.ADJUNCT, DATIVE.ADJUNCT,PROBLEM,POSSESSOR,NEGATIVE.PARTICLE}
@attribute vectorsk{1,2}
@data
1,NE,PRON,ABL,MODIFIER,VERB,SENTENCE,?,?,?,?,1
2,KOŞA,ADJ,?,MODIFIER,VERB,SENTENCE,?,?,?,?,1
2,BURA,NOUN,DAT,OBJECT,VERB,SENTENCE,?,?,?,?,1
3,KIRAATHANE,NOUN,DAT,OBJECT,ADJ,MODIFIER,YABANCI,NOUN,DAT,OBJECT,1
4,?,?,?,?,VERB,SENTENCE,KEREM,NOUN,NOM,SUBJECT,1
5,YENİ,ADJ,?,MODIFIER,ADJ,SUBJECT,AT,VERB,A3SG,SENTENCE,1
6,?,?,?,COORDINATION,VERB,SENTENCE,?,?,?,?,2
6,SANKİ,ADV,?,MODIFIER,VERB,SENTENCE,?,?,?,?,2
6,BİRAZ,ADV,?,MODIFIER,VERB,SENTENCE,?,?,?,?,2
6,KENDİ,PRON,DAT,COLLOCATION,VERB,SENTENCE,?,?,?,?,2
7,PARK,NOUN,ABL,OBJECT,VERB,SENTENCE,?,?,?,?,1
8,O,PRON,GEN,SUBJECT,NOUN,OBJECT,BEKLE,VERB,PAST,SENTENCE,1
9,?,?,?,?,NOUN,OBJECT,DUY,VERB,A1SG,SENTENCE,1
10,GALİBA,ADV,?,MODIFIER,VERB,SENTENCE,?,?,?,?,1
.
.
```



Şekil 3.5: ARFF biçimindeki dosyaların bir özelliğe göre WEKA ortamında görselleştirilmesi

WEKA sisteminde işaretlenen dosyaların dağılımları gözlenmiş (Şekil 3.5) ve anlam sınıflandırması çeşitli makine öğrenme algoritmaları ile denenmiş; elde edilen sonuçların hata analizi yapılmıştır. Şekil 3.5 incelenen sözcüğün öncesinde gelen diğer sözcüklerin aldıkları eklere göre dağılımlarını göstermektedir. Sağda, gelen ekler ve sayıları, altında ise grafiksel olarak dağılım bulunmaktadır.

3.5.8. Değerlendirme Yöntemleri

Bu çalışmalar sırasında değerlendirme için kullanılan iki yöntem vardır. Test için bir kez eğitimde kullanılan örnekler, bir kez de k-katlı çapraz doğrulama (k-fold cross validation- CV) yöntemi kullanılmıştır. CV bir modelin doğruluk değerini test etmek için kullanılan tekrar örnekleme tekniğidir. CV veri satırlarını yaklaşık eşit büyüklükte olacak şekilde k parçaya (fold) ayırır (F_1..F_k). Bu ayırma işlemi gerçekleştikten sonra k tane deney gerçekleştirilir. Her bir deneyde F_i test kümesi olarak kullanılırken geriye kalan k-1 parça eğitim kümesi olarak kullanılır. Son olarak da bu teknikle elde edilen k

deneyin ortalama sonucu ve standart sapması hesaplanır. k değerinin çok küçük olması doğruluk değerlerinin güvenilirliğini düşürebileceği gibi, çok büyük olması da varyansın ve hesaplama zamanının artmasına neden olur. Deneysel çalışmalar k=10 olması durumunda en uygun sonuçların elde edilebileceğini göstermektedir.

Bu analizler sırasında kullanılan *hata matrisi (Confussion Matrix-CM)* yapılan tahminlerden elde edilen olası tüm sonuçları göstermektedir. CM bir sınıflandırıcının ne kadar iyi tahmin edebileceğini gösteren bir tablodur. Yanlış sınıflandırma sonucu ortaya çıkan hata oranlarını basit bir şekilde sunmak yerine, modelin hangi durumlarda zorlandığını göstermesi bakımından daha faydalıdır. Her bir sınıf için modelin hangi sınıfı seçebileceği hakkında fikir verir. Elde edilebilecek olası sonuçlar Doğru pozitif (TP), Doğru Negatif (TN), Yanlış Pozitif (FP) ve Yanlış Negatif (FN) elemanlarından oluşmaktadır.

Tablo 3.13:Hata matrislerinde gösterilen ve hesaplamalarda kullanılan terimlerin açıklaması

TP: x sınıfına ait x olarak seçilen örnekler
TN: x sınıfına ait olmayan ve x sınıfı olarak seçilmeyen örnekler
FP: x sınıfına ait olmayan, fakat x olarak seçilen örnekler
FN: x sınıfına ait olan, fakat x olarak seçilmeyen örnekler

Değerlendirme kriterlerinde doğru sınıflandırılan örnek sayısı yanında farklı bazı kriterler de kullanılmıştır. Bunlar:

- Precision (P): x sınıfına ait seçilen doğru örneklerin x sınıfına ait olarak seçilen örneklere oranıdır.

$$P = (TP) / (TP + FP) \tag{3.7}$$

- Recall (R): x sınıfına ait olarak seçilen örneklerin x sınıfına ait bütün örneklere oranıdır.

$$R = (TP) / (TP + FN) \tag{3.8}$$

P değeri herhangi bir sınıf için seçilen örneklerin doğruluk oranının yüksek veya düşük olmasını gösterir. Buna karşılık R değeri seçilen örneklerin seçilmesi gereken örneklerden ne kadarını kapsadığını gösterir. Yapılmak istenen hem doğruluk oranının hem de kapsamın yüksek olmasını sağlamaktır.

4. BULGULAR

4.1. BİRİNCİ DERLEME METİNDEN ELDE EDİLEN SONUÇLAR

Çalışmanın ilk aşamalarında Bölüm 3'te açıklanan işlenmemiş metinlerden oluşan birinci derleme metin kullanılmış ve üzerinde uzun bir süre yoğun olarak çalışılmıştır. Bu ham metinler üzerinde yapılan elle işaretlemeler, farklı gruplar tarafından oldukça uzun bir sürede gerçekleştirilmiştir. Elde edilen derleme metin üzerinde yapılan testler, ikinci derleme metine göre daha sınırlı olarak gerçekleştirilmiştir. Çünkü seçilen sözcüklerin anlam işaretlemelerinin yanında, kullanılacak özelliklerin de elle işaretlenmesi gerekmekteydi. Birinci derleme metin kullanılarak *git* sözcüğü için sınırlı sayıda örneğe WEKA'daki yöntemler değil, Bölüm 3'te anlatılan istatistiksel NB ve örnek tabanlı EB yöntemleri uygulanarak çalışmanın ilk sonuçları elde edilmiştir. Bu uygulamada örnek büyüklüğünün küçüklüğü nedeni ile bu sonuçları etkin şekilde değerlendirmek oldukça zor olmasına rağmen, ikinci derleme metin üzerinde yapılacak çalışmalara ışık tutması ve çalışmanın temelini oluşturması nedeni ile oldukça faydalı bilgiler sunmuştur.

Tablo 4.1: Algoritmalarda kullanılan özellikler

Özellik	Açıklama
L1P	Hedef sözcüğün solundaki birinci tümce öğesi
L2P	Hedef sözcüğün solundaki ikinci tümce öğesi
R1P	Hedef sözcüğün sağındaki birinci tümce öğesi
L1M	Hedef sözcüğün solundaki birinci tümce öğesinin hal eki
L2M	Hedef sözcüğün solundaki ikinci tümce öğesinin hal eki
R1M	Hedef sözcüğün sağındaki birinci tümce öğesinin hal eki
HK	Hedef sözcüğün kökü
CL	Soldan birinci kalıp sözcük
CR	Sağdan birinci kalıp sözcük

Yöntemlerde kullanılan özellikleri Tablo 4.1 açıklamaktadır. Görüldüğü gibi incelenen sözcüğün solundaki ve sağındaki sözcükler sadece ikili olarak öğelerine ayrılabilmiştir. Küçük bir veri üzerinde uygulanan (Altan ve Orhan, 2003) bu algoritmalardan EB NB'ye göre biraz daha iyi sonuç vermiştir (Tablo 4.2). Elde edilen sonuçlar Tablo

3.3'teki tümceler kullanılarak çıkarılmıştır. Tabloda birinci sütun tümce numaralarını, 2. sütunla 11. sütun arasındaki alanlar SAB için kullanılan özellikleri, son iki sütun ise NB ve EB yöntemlerinden elde edilen anlam sınıfını göstermektedir. Tablodaki anlam sütununun altındaki sayılar tümcede geçen git sözcüğünün asıl anlam sınıfını belirtmektedir.

Tablo 4.2:Programdan elde edilen sonuçlar

No	L1P	L2P	R1P	L1M	L2M	R1M	HK	Anlam	CL	CR	NB	EB
1	duz	-	-	-	-	-	git	1	birlikte	-	1	1
2	yez	-	-	e	-	-	git	1	suratya	-	1	1
3	duz	öz	-	-	-	-	git	3	iyi	-	3	3
4	yez	-	zaz	e	-	de	git	1	buzdenizine	olasılığı	1	1
5	keyl	-	-	i	-	-	git	5	merakımı	-	5	5
6	zaz	duz	yez	-	-	e	git	1	önce	yana	1	1
7	keyl	öz	eyl	e	-	-	git	4	hoşuna	gibiydi	4	4
8	-	-	dot	-	-	-	git	12	-	artan	1	12
9	keyl	miz	-	e	-	-	git	4	hoşlarına	-	4	4
10	-	-	-	-	-	-	git	1	-	-	1	1

4.2. İKİNCİ DERLEME METİNDEN ELDE EDİLEN SONUÇLAR

4.2.1. İlk Çalışma ve Gel Sözcüğünün İncelenmesi

Bu bölümde *gel* sözcüğünün anlam incelemesi detaylı olarak anlatılacaktır. *gel* sözcüğünün sırasıyla 2, 3 ve 4 anlam sınıfı kullanıldığı zaman elde edilen sonuçları çıkarılmıştır. Tablolarda birinci sütün kullanılan yöntemin adıdır. Çalışmanın bu kısmında J48, IBK, KStar ve AODE yöntemleri kullanılmıştır. Algoritmalardan IBk ve KStar genelde birbirine yakın sonuçlar vermektedir. Bu durum her ikisinin de örnek tabanlı yöntemler olmasından ve temelde aynı esasları kullanmalarından kaynaklanmaktadır. Bu nedenle sadece birine ait sonuçlar verilmiştir. Test aşamasında ise iki farklı yöntem tercih edilmiştir. Birinci yöntemde test ve öğretim verisi aynıdır. İkinci test yönteminde ise, k-katlı CV kullanılmıştır. Her iki test yöntemi karşılaştırıldığında k-katlı CV ile elde edilen sonuçların daha fazla hata içerdikleri görülmüştür. Bunun nedeni örnek verinin az olması ve eğitimde kullanılan veri ile testte

kullanılan verinin farklı anlamlar içerebilmesidir. Bölme işlemi farklı k değerleri için denenmiş; ancak sonuçlar çok farklılık göstermemiştir.

Örneğin, gel sözcüğü iki, üç ve dört farklı anlamlı alınarak bunların dağılımları incelenmiş ve bu dağılımlar Tablo 4.3'te özetlenmiştir. Dağılımlarda birinci anlam olarak gösterilen anlam, gel sözcüğünün *gitmek, ulaşmak, varmak*, gibi genel olarak kullanılan anlamını göstermektedir. İlk anlam sınıflamasında ikinci anlam ise gel sözcüğünün bu genel anlam dışında kalan kullanımlarını kapsamaktadır. Bu ikinci anlam sonraki aşamada kendi arasında tekrar gruplanmıştır. Deyimsel kullanışlar on birinci anlam olarak kabul edilmiş ve diğerleri ikinci anlam olarak bırakılmıştır. Bir sonraki aşamada ise ikinci anlam içerisinden *bir şeyin zamanı, yeri, sırası gelmek* anlamı 5 numaralı anlam olarak yeniden gruplandırılmış böylece ikinci anlam kendi arasında tekrar bölünmüştür.

Tablo 4.3: Gel sözcüğünün aşamalı anlam belirginleştirmesinde kullanılan anlam sayılarına göre dağılımları (%)

AS (Anlam Sayısı)	**ANo (Anlam No)**	**Dağılımı**
2	1	59
	2	41
3	1	59
	2	14
	11	27
4	1	59
	2	10
	5	4
	11	27

Değerlendirme kriterlerinde hata matrisleri (Tablo 4.4, Tablo 4.5, Tablo 4.6) ve doğru sınıflandırılan örnek sayısı (Tablo 4.7) yanında P ve R değerleri gibi farklı bazı kriterler de kullanılmıştır (Tablo 4.8, Tablo 4.9) (Orhan ve Altan, 2004). Tablolarda anlamlar için *a* ve *b* gibi değerler kullanılmaktadır. Satırlarda gösterilen *a=1* ifadesi *1* numaralı anlam için *a* sembolik değerinin kullanıldığını göstermektedir. Satırlar o sınıfa ait verilerin toplamını vermektedir. Sütunlar ise yöntemlerce bulunan değerleri göstermektedir. Tablo 4.4'te birinci anlam için 313 ve ikinci anlam için 214 örnek bulunmaktadır. J48 yönteminde k-katlı CV kullanılarak birinci anlama ait olan 313

örnekten 288 tanesi birinci anlam, 25 tanesi de ikinci anlam olarak bulunmuştur. İkinci anlama ait olan 214 örneğin ise 68 tanesi ikinci anlam olarak bulunurken, 146 tanesi birinci anlam olarak hesaplanmıştır. Başarım oranı doğru sınıflandırılan örneklerin tüm örneklere oranı olan (288+68)/(288+25+146+68) ifadesinden elde edilmişitr. P ve R değerleri her anlam için ayrı ayrı hesaplanmıştır. Örneğin, birinci anlamın P değeri; birinci anlam olarak doğru bulunan örneklerin, birinci anlam olarak bulunan toplam örnek sayısına oranını veren 288/(288+146) ifadesinden ve R değeri de benzer şekilde birinci anlam olarak doğru bulunan örneklerin, birinci anlama ait toplam örnek sayısına oranını veren 288/313 ifadesinden bulunmuştur.

Tablo 4.4:Gel sözcüğünün iki anlamlı sınıflandırmasında elde edilen hata matrisi değerleri

	AS=2					
Algoritma	**Ö=T (Öğretim verisi= Test verisi)**			**CV**		
J48	**a**	**b**		**a**	**b**	
	312	1	**a = 1**	288	25	**a = 1**
	20	194	**b = 2**	146	68	**b = 2**
IBK	**a**	**b**		**a**	**b**	
	310	3	**a = 1**	275	38	**a = 1**
	8	206	**b = 2**	99	115	**b = 2**
AODE	**a**	**b**		**a**	**b**	
	299	14	**a = 1**	265	48	**a = 1**
	29	185	**b = 2**	92	122	**b = 2**

Tablo 4.5:Gel sözcüğünün üç anlamlı sınıflandırmasında elde edilen hata matrisi değerleri

	AS=3											
Algoritma	**Ö=T**						**CV**					
J48	**a**	**b**	**c**				**a**	**b**	**c**			
	309	0	1	**a**	=	**1**	286	6	18	**a**	=	**1**
	7	68	0	**b**	=	**2**	49	22	4	**b**	=	**2**
	13	0	129	**c**	=	**11**	97	3	42	**c**	=	**11**
IBK	**a**	**b**	**c**				**a**	**b**	**c**			
	309	0	1	**a**	=	**1**	278	7	25	**a**	=	**1**
	5	70	0	**b**	=	**2**	36	26	13	**b**	=	**2**
	12	1	129	**c**	=	**11**	80	7	55	**c**	=	**11**
AODE	**a**	**b**	**c**				**a**	**b**	**c**			
	303	0	7	**a**	=	**1**	281	4	25	**a**	=	**1**
	28	38	9	**b**	=	**2**	39	18	18	**b**	=	**2**
	41	1	100	**c**	=	**11**	72	3	67	**c**	=	**11**

Tablo 4.6:Gel sözcüğünün dört anlamlı sınıflandırmasında elde edilen hata matrisi değerleri

	AS=4	
Algoritma	**Ö=T**	**CV**
J48	a b c d 309 0 0 1 a = 1 4 50 0 0 b = 2 3 0 18 0 c = 5 13 0 0 129 d = 11	a b c d 287 4 1 18 a = 1 32 17 1 4 b = 2 15 1 5 0 c = 5 98 3 0 41 d = 11
IBK	a b c d 309 0 0 1 a = 1 2 52 0 0 b = 2 3 0 18 0 c = 5 12 0 1 129 d = 11	a b c d 277 6 1 26 a = 1 23 24 1 6 b = 2 11 1 5 4 c = 5 80 3 3 56 d = 11
AODE	a b c d 301 0 0 9 a = 1 23 26 0 5 b = 2 12 0 3 6 c = 5 42 0 0 100 d = 11	a b c d 279 3 1 27 a = 1 30 13 1 10 b = 2 13 0 0 8 c = 5 73 2 0 67 d = 11

Tablo 4.7:Algoritmaların başarım oranları (%)

	AS=2		AS=3		AS=4	
Algoritma	**Ö=T**	**CV**	**Ö=T**	**CV**	**Ö=T**	**CV**
J48	96.0	67.5	95.0	65.0	96.0	66.0
IBK	97.9	74.0	96.3	69.4	96.3	68.7
AODE	91.8	73.4	83.6	69.2	81.6	68.1

Elde edilen sonuçlarda öğretim ve test verileri aynı olduğu durumda algoritmalar %90 ve üzeri gibi oldukça iyi bir başarım göstermektedir. Ancak k-katlı CV ile değerlendirme yapıldığında oranlar %65 değerlerine düşmektedir. Özellikle örnekler içerisinde dağılımı az olan anlamların P ve R değerleri önemli oranda düşüş göstermektedir. IBk ve J48 yöntemlerinde P ve R değerleri sınıf dağılımına duyarlıdır ve sonuçlar dağılımla orantılı olmaktadır. AODE ise dağılımlar arasında fark büyük olduğu zaman az sayıda örnek içeren sınıfa ait sonuçları bulmakta başarısız olmaktadır. Anlam sayısının dört olduğu testte bu durum çok açık olarak görülebilir. Dağılımı %4 olan 5 numaralı anlam için J48 ve IBk düşük bir P veR değeri verirken, AODE değerleri

0 olmaktadır. k-katlı CV değerlendirme yöntemi örnek sayısı fazla olduğu durumlarda daha etkilidir. Örnek sayısı az olduğu zaman verinin k parçaya ayrılması sırasında içerisinde az olan örnekler bulunmayabilir bu da öğretim aşamasını olumsuz etkiler. Başarım oranları karşılaştırıldığında yöntemlerin birbirine yakın sonuçlar verdiği, ancak IBk'nın genel olarak daha iyi sonuçlar verdiği gözlenmektedir. Bu durum örnek tabanlı yöntemlerin diğer pek çok alanda da başarılı olmasını ve birinci derleme metinle elde edilen sonuçları da desteklemektedir.

Tablo 4.8:Algoritmalardan elde edilen P değerleri (%)

	AS=2		AS=3		AS=4	
Algoritma	**Ö=T**	**CV**	**Ö=T**	**CV**	**Ö=T**	**CV**
J48	94 99	66 73	94 100 99	66 71 66	94 100 100 99	66 68 71 65
IBK	97 99	73 75	95 99 99	70 65 59	95 100 95 99	71 71 50 61
AODE	91 93	74 72	81 97 86	72 72 61	80 100 100 83	71 72 0 60

Tablo 4.9:Algoritmalardan elde edilen R değerleri (%)

Algoritma	AS=2		AS=3		AS=4	
	Ö=T	**CV**	**Ö=T**	**CV**	**Ö=T**	**CV**
J48	99 91	92 32	99 91 91	92 29 30	99 93 86 91	93 31 24 29
IBK	99 96	88 54	99 93 91	90 35 39	99 96 86 91	89 44 24 39
AODE	95 86	85 57	98 51 70	91 24 47	97 48 14 70	90 24 0 47

4.2.2. Yapay Sözcükler ve Aşamalı Anlam Sınıflaması Çalışmasının Geliştirilmesi

Bu bölümde yapay sözcükler ve aşamalı anlam sınıflaması için oluşturulan 28 tane test kümesi kullanılarak sonuçlar elde edilmiştir. Kullanılan algoritmalarımız AODE, KStar ve J48 olarak seçilmiştir. 28 test kümesi beş grupta toplanmaktadır. Bu gruplar daha önce belirtilen aşamalı anlam kullanımı ve yapay sözcükler yaklaşımlarını da kapsamaktadır (Orhan ve Altan, 2005). Gruplar şu şekilde seçilmiştir:

- Yapay Sözcükler (anlamları uzak olanlar):G1
- Yapay Sözcükler (anlamları yakın olanlar): G2
- Anlamlarının Belirginleştirilmesi Kolay Olan Sözcükler: G3
- Anlamlarının Belirginleştirilmesi Zor Olan Sözcükler: G4
- Aşamalı Anlam Sınıflandırması: G5

Tablo 4.10: Yapay uzak ve yakın anlamlı sözcüklerin dağılımları (%)

Yapay Uzak Anlamlı Sözcükler		**Yapay Yakın Anlamlı Sözcükler**			
G1		**G2**			
Test	**Sözcük/Oran**	**Test**	**Sözcük/Oran**	**Test**	**Sözcük/Oran**
T1	Acı-kız(1-2)/ 32-56-12	**T6**	Anne-baba-çocuk/31-29-40	**T12**	Eski-yeni/36-64
T2	Anne-ses-gece/35-34-30	**T7**	Anne-baba/52-48	**T13**	Ev-okul-park/71-12-17
T3	Yıl(1-3)-insan-göz/ 34-34-2-30	**T8**	Kadın-erkek/67-33	**T14**	Köpek-geyik/49-51
T4	Bugün-geyik-ülke/35-3-35	**T9**	Dün-yarın-bugün/29-21-50	**T15**	Vergi-para-faiz/30-44-27
T5	Kardeş-oda-sabah/34-36-31	**T10**	Sabah-akşam-gece/24-20-55	**T16**	Peynir-ekmek-yemek/43-13-44
		T11	Süre-saniye-saat/47-4-49		

Tablo 4.11: Gerçek sözcüklerin dağılımları (%)

Yapay Olmayan Uzak Anlamlı Sözcükler		**YapayOlmayan Yakın Anlamlı Sözcükler**	
G3		**G4**	
Test No	**Sözcük/Oran**	**Test No**	**Sözcük/Oran**
T17	Kar(1-2-3)/ 24-57-19	**T20**	Çal(1-2-3)/ 61-11-28
T18	Yüz(1-2-3-4)/ 28-43-26-3	**T21**	Hareket(1-2)/ 86-14
T19	Kız(1-2) /90-10	**T22**	Hava(1-2-3-4)/ 37-21-26-16
		T23	Yan(1.1-1.2-2-3-4-5-7)/ 7-4-44-23-4-13-6

Bu gruplardan G1 için beş (T1-T5), G2 için on bir (T6-T16) , G3 için üç (T17-T19), G4 için dört (T20-T23) ve G5 için beş (T24-T28) test kümesi kullanılmıştır. Bu sayılar mevcut metinlerden istenilenleri çıkarabilme durumları göz önüne alınarak belirlenmiştir. Tablo 4.10, Tablo 4.11 ve Tablo 4.12 bu grupları, kullanılan sözcükleri

ve derleme metinde geçiş sıklıklarını vermektedir. Örneğin, yapay uzak anlamlı sözcükler için oluşturulan T1 kümesini *acı* ve *kız* sözcükleri oluşturmaktadır. Kız sözcüğü birbirinden çok farklı iki anlama sahiptir. Bu nedenle T1 kümesinde 3 farklı anlama karşılık gelen bir test hazırlanmıştır. Acı sözcüğü %32, kız sözcüğünün birinci anlamı %56 ve ikinci anlamı %12 oranında örneğe sahiptir.

Tablo 4.12: Aşamalı anlam sınıflamasında sözcüklerin dağılımları (%)

Aşamalı	
G5	
Test No	**Sözcük/Oran**
T24	Yan1-2/13-87
T25	Yan1.1-1.2-2/7-7-86
T26	Yan1-2-3/13-64-23
T27	Yan1-2-5-7/13-68-14-6
T28	Yan1-2-3-5-7/13-45-23-14-6

Tablo 4.13:Test kümeleri için algoritmalardan elde edilen başarım oranları (%)

Test No	**AODE**	**KStar**	**J48**	**Test No**	**AODE**	**KStar**	**J48**
T1	77.97	80.35	78.57	**T 15**	71.43	71.43	72.45
T2	70.65	71.01	65.21	**T 16**	66.97	60.55	60.55
T3	72.77	73.51	75.49	**T 17**	72.22	72.22	72.22
T4	73.43	75.78	69.53	**T 18**	87.06	87.06	97.41
T5	73.72	70.33	66.94	**T 19**	94.4	92.00	97.6
T6	58.14	57.51	57.51	**T 20**	72.09	74.41	60.46
T 7	63.49	60.31	55.55	**T 21**	100	96.97	100
T 8	68.86	65.35	67.54	**T 22**	58.82	58.82	55.88
T 9	61.80	64.04	59.55	**T 23**	69.63	68.89	68.89
T 10	58.38	57.04	61.07	**T 24**	94.81	95.55	100
T 11	81.41	76.10	76.10	**T 25**	94.07	91.85	97.77
T 12	70.25	70.25	71.07	**T 26**	79.26	81.48	84.44
T 13	75.46	77.28	77.66	**T 27**	83.70	88.14	88.14
T 14	81.08	85.13	72.97	**T 28**	71.11	75.55	71.11

Bu sonuçlara bakıldığında test verilerinin sayısı anlamlar arasında eşit oranlarda dağılmışsa önceki sonuçlarla uyumlu olarak AODE yönteminin daha başarılı olduğu görülmektedir. Dağılımın dengesiz olduğu durumlarda diğer yöntemler daha başarılı olmaktadır. Bunun dışında bütün algoritmalar benzer sonuçlar vermekle birlikte genel olarak örnek tabanlı Kstar yöntemi biraz daha iyi sonuçlar vermektedir.

AODE çok karmaşık bir yöntem olmamakla beraber başarı oranı diğer iki yöntemle benzerlik göstermektedir. KStar ve J48 hesaplama karmaşıklığı daha fazla olan

yöntemlerdir; çünkü sınıflandırılacak örnek bilinen bütün örneklerle tek tek karşılaştırılmakta ve diğer örneklere uzaklığına göre sınıfına karar verilmektedir. Ancak AODE önceden elde edilen istatistiksel bilgiyi kullanarak olasılığı en fazla olan örneği sınıf olarak seçmektedir. Yöntemlerin başarısı verilen eğitim kümesindeki örneklerin sayısına, anlamlara, örneklerin dağılımına ve sözcüklerin özelliklerine göre değişebilmektedir. Anlamları kolay belirginleştirebilen ya da bir veya iki biçimsel özelliğe göre anlamı belirgin bir şekilde belirginleştirebilen sözcükler de algoritmalar da daha başarılı olmaktadır.

4.2.3. Özellik Seçimi ile İlgili İlk Testler

Türkçe SAB için etkili özelliklerin bulunması amacıyla başlangıçta bir dizi test yapılmıştır. İlk aşamada bu amaçla bir grup sözcük (Tablo 4.14) ve bir grup özellik (Tablo 4.15) seçilmiştir. Tablo 4.14 seçilen sözcüklerin kaç anlamı olduğunu, bu anlamların neler olduğunu ve dağılımlarını göstermektedir. Seçilen sözcükler farklı sözcük türlerinden alınmıştır. Her bir sözcük için 17 test yapılmış ve bu 17 farklı testte farklı özellik kümeleri denenmiştir (Tablo 4.16) (Orhan ve Altan, 2005).

Tablo 4.14: Sözcükler, bunların anlamları ve dağılımları (%)

Sözcük	Anlam 1	Anlam 2	Anlam 3	Anlam 4	Anlam 5
yan	Ateşle hal değiştirme (eylem) / 7	Parlamak, açılmak (eylem) / 8	Yön,kısım (ad) / 47	birlikte (belirteç) / 24	taraf (ad) / 14
kız	bayan (ad) / 87	sinirlenmek (eylem) / 13			
kap	Çekip almak (eylem) / 79	konteyner (ad) / 21			
art	arka (ad) / 14	Arkasından gelen, takip eden (belirteç) / 18	yükselmek (eylem) / 69		
yüz	Organ, bir şeyin ön kısmı (ad) / 28	sayı (sıfat) / 43	sebep (belirteç) / 26	Suda hareket etmek (eylem) / 3	
iç	Sıvı almak(eylem) / 26	Bir şeyin içinde kalan kısmı (ad, belirteç,sıfat) / 74			

Tablo 4.15: Seçilen özellik kümeleri

Özellik Adı	**Özellik Kümesi**	**Özellik Adı**	**Özellik Kümesi**
Tümce no		hedefHalEki	F6
Dosya no		hedefİyelik	F7
onceKok	F1	hedefİliski	
onceTur	F2	sonraKok	F8
onceTuretme		sonraTur	
onceHalEki	F3	sonraTuretme	
onceİyelik	F4	sonraHalEki	
onceİliski		sonraİyelik	
hedefKok		sonraİliski	
hedefTur	F5	anlam	
hedefTuretme			

Tablo 4.16:Oluşturulan test kümelerinde kullanılan özellikler

TEST	**F1**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**TEST**	**F1**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**
1	+	-	-	-	-	-	-	**10**	-	-	-	+	+	-	-
2	-	+	-	-	-	-	-	**11**	-	-	-	+	-	+	-
3	-	-	+	-	-	-	-	**12**	-	-	-	+	+	+	-
4	+	+	-	-	-	-	-	**13**	-	-	-	-	-	-	+
5	+	-	+	-	-	-	-	**14**	+	-	-	+	-	-	-
6	+	+	+	-	-	-	-	**15**	+	+	-	+	-	-	-
7	-	-	-	+	-	-	-	**16**	+	+	+	+	-	-	-
8	-	-	-	-	+	-	-	**17**	+	+	+	+	+	+	+
9	-	-	-	-	-	+	-								

Farklı algoritmalar kullanmak yapılan test sonuçlarında önemli bir değişim sağlamazken, özelliklerde durum çok farklı olmuştur. Yapılan testler elde edilen doğruluk oranlarının kullanılan özelliklere çok duyarlı olduğunu göstermiştir. Bazı özellikler daha önce de bahsedildiği gibi sınıflandırmada etkisiz kalırken, bazıları da tek başına bile oldukça iyi sonuçlar vermiştir. Bazı özellikler de diğer özelliklerle birlikte kullanıldığında başarı oranına pozitif etki etmiştir. Örneğin, art sözcüğü önceki sözcük kökü ile %37.66 oranında belirginleştirilebilirken önceki kök sözcük ve hal eki birlikte kullanıldığında bu oran %72.73 olmaktadır. Aynı sözcük için sadece hedef sözcüğün hal eki kullanılarak %87.01 başarım sağlanabilmektedir. Kız sözcük kökü için ise sadece hedef sözcüğün türü %100 gibi bir başarım vermektedir.

Tablo 4.17: Özellik Seçimine bağlı başarım oranları (%)

Test	Art	İç	Kap	Kiz	Yan	Yüz	Test	Art	İç	Kap	Kiz	Yan	Yüz
1	37.66	58.48	75	88.8	37.5	60.34	**10**	87.01	100	100	99.2	57.81	96.55
2	72.73	76.79	80.55	91.2	53.12	54.31	**11**	87.01	100	100	100	56.25	75.86
3	70.13	74.55	80.55	88	46.88	47.41	**12**	87.01	99.55	100	99.2	62.25	96.55
4	72.73	79.02	75	88.8	53.12	63.79	**13**	76.62	61.16	86.11	90.4	52.34	75
5	72.73	76.34	80.55	85.6	49.22	60.34	**14**	89.61	100	100	100	57.03	83.62
6	67.53	79.02	80.55	88	51.56	64.65	**15**	89.61	99.55	100	100	57.03	87.93
7	87.01	100	100	100	55.47	73.28	**16**	87.01	100	100	100	53.12	87.93
8	76.62	83.04	77.78	95.2	52.34	87.93	**17**	90.90	92.85	91.67	95.2	71.18	87.07
9	70.13	82.14	80.55	88	46.09	62.93							

4.3. SENSEVAL YAKLAŞIMININ TÜRKÇE'YE UYGULANMASI

Türkçe sözcükler içinden *biçim (syntax)* ile *anlam (semantics)* arasındaki ilişkilerin sadece basit olanlarını seçmek mümkün değildir. Zira tüm sözcükler için biçimbilim Türkçe'nin sondan eklemeli bir dil olması özelliği nedeni ile karmaşıktır. Yapılan analizlerle anlamlı sonuçlar elde edebilmek için, ilk hesaplamalı dilbilim çalışmaları olarak kabul edilen yüksek frekanslı sözcük seçiminin yapıldığı Zipf (1949) Kanunu'ndan yararlanılmıştır. Penn Tree Bank derleme metninde tümcelerin POS'leri incelenirken *açık sınıf sözcükler* ve *kapalı sınıf sözcükler* şeklinde iki sınıflandırmadan sonra, diğer temel sınıflandırmalara devam edilir. Sözdizimsel öneme sahip kapalı sınıf sözcüklere yeni sözcük ilavesi çok seyrek gerçekleşir; zira bu sözcükler ilgeç, bağlaç, soru sözcükleri, tanımlık ve adıllardır. Oysa isim, eylem, sıfat ve belirteç tipindeki sözcük kümeleri sürekli olarak yeni sözcüklerin eklenmesine açıktır ve anlamın belirlenmesinde çok daha etkindirler. Dilin iskeletini oluşturan ve sık kullanılan ancak sadece biçimsel işlevi olan kapalı sınıf sözcükler bilgi çıkarımı, indeksleme gibi pek çok DDİ çalışmasında ihmal edilmiştir. Fakat bu sözcüklerin dilin modellenmesinde olasılıkları hesaplamada önemli katkıları olabilir. Bu çalışmadaki Türkçe derleme metinde açık sınıf sözcükler ve kapalı sınıf sözcükler şeklinde bir ön sınıflandırma yapılmamıştır. Fakat tüm metinler içindeki sözcük frekansları incelendiğinde *bir* sözcüğü 1134 defa işlenmiştir ve en yüksek frekanslı sözcüktür. *ve* sözcüğü ise ikinci en sık frekanslı kapalı sınıf sözcük özelliğine sahip olan 658 tekrarlı bir bağlaçtır. Buna karşılık düşük frekanslı kapalı sınıf sözcükler çok azdır ve genellikle bunlar türlerinin

oldukça özel örnekleridir. Ekte verilen sözcük sıklıklarından derleme metnin Zipf Kanunu'nu sağladığı kolaylıkla görülebilir.

Derleme metinde tümcelerin sözdizimsel olarak işaretlenmesi gerçekleştirildiği için hem açık sınıf sözcükler hem de kapalı sınıf sözcükler teker teker ve ayırt edilmeksizin kategorilendirilmiştir.

Tablo 4.18: Metinlerde 55 ve daha fazla geçen sözcüklerin geçiş sıklıkları ve anlam sayıları

Sözcük	GS	AS	Sözcük	GS	AS	Sözcük	GS	AS	Sözcük	GS	AS	Sözcük	GS	AS
Eylem			**bul**	98	13	**yap**	354	22	**uçak**	77	1	**gün**	158	11
ara	55	6	**düşün**	98	8	**et**	426	10	**bilim**	79	3	**ev**	179	5
belir	55	3	**var**	100	7	**de**	543	10	**kapı**	79	7	**şey**	191	2
at	56	37	**çalış**	101	6	**İsim**			**gece**	82	4	**yer**	193	16
tut	56	50	**konuş**	101	15	**yaş**	55	8	**adam**	84	11	**Diğer**		
oku	57	9	**yaşa**	101	11	**neden**	56	3	**baba**	87	14	**bütün**	55	5
bekle	59	7	**kal**	113	20	**durum**	58	4	**ses**	88	5	**doğru**	55	9
bırak	61	24	**anlat**	129	3	**masa**	59	5	**yol**	88	12	**böyle**	57	4
bit	67	4	**başla**	131	6	**yüz**	61	15	**kız**	89	6	**sonra**	57	5
yaz	67	11	**değil**	133	1	**hal**	62	5	**çocuk**	95	7	**artık**	58	5
düş	68	32	**gir**	134	17	**üzer**	62	0	**yan**	96	14	**çok**	58	2
göster	68	12	**gör**	143	10	**bura**	67	1	**anne**	97	2	**nasıl**	60	8
anla	74	8	**geç**	146	38	**dünya**	67	6	**baş**	102	14	**şimdi**	61	4
dön	75	14	**söyle**	155	8	**erkek**	67	7	**iş**	105	8	**güzel**	62	11
gerek	75	1	**iste**	168	5	**hayat**	67	13	**göz**	111	10	**başka**	71	4
sev	75	5	**bak**	185	19	**söz**	67	6	**kadın**	112	5	**aynı**	72	3
getir	78	9	**bil**	188	10	**an**	68	3	**insan**	117	4	**hiç**	77	4
çek	79	46	**git**	197	22	**kişi**	68	5	**ara**	136	9	**yeni**	87	8
dur	83	15	**çık**	238	57	**ön**	68	8	**yıl**	137	4	**bir**	88	12
sor	84	2	**ver**	247	23	**üst**	70	11	**zaman**	146	10	**yok**	92	7
aç	96	28	**al**	265	36	**konu**	71	2	**iç**	155	15	**büyük**	95	6
otur	96	14	**gel**	298	38	**ora**	73	1	**el**	157	11	**iyi**	101	9

Çalışmanın bu aşamasında Senseval Projesine benzer bir yaklaşım kullanılmıştır. Eylemler, isimler ve bunların dışında kalan sözcük türlerinden seçilen üç grup sözcük sınıfı kullanılmıştır. Sözcükler seçilirken geçiş sıklıklarına(GS) bakılmış ve derleme metinde belli oranda geçen bu sözcüklerin TDK sözlüğünden anlam sayıları (AS) çıkarılmıştır (Tablo 4.18).

Sözcüklerin anlam sınıflamasında hem kaba hem de ince anlamları kullanılmıştır. İnce ve kaba anlam sınıflamalarında kullanılan anlam sınıfları ve geçiş yüzdeleri bulunmuştur (Tablo 4.19). Tablodan da görüleceği gibi bazı anlamlar metinde hiç geçmemektedir. Ayrıca deyimsel kullanımlar için -1 anlam numarası kullanılmıştır. Kullanılan diğer anlam numaraları ince anlamlarda TDK sözlüğündeki anlam

numaralarıyla aynıdır. Kaba anlamlar ise ince anlamların birleştirilmesiyle elde edilmiştir.

Tablo 4.19: Eylemlerin ince ve kaba anlam sınıfları ve geçiş yüzdeleri

Kaba															
al		**bak**		**çalış**		**çık**		**geç**		**gel**		**gir**		**git**	
ANo	**GS**	**ANo**	**GS**	**ANo**	**GS**	**ANo**	**GS**	**ANo**	**GS**	**ANo**	**GS**	**ANo**	**GS**	**ANo**	**GS**
2	0,47	1	0,64	1	0,66	1	0,47	2	0,34	1	0,67	1	0,58	4	0,12
1	0,17	9	0,22	3	0,34	2	0,36	3	0,09	3	0,26	5	0,03	1	0,74
3	0,21	3	0,02			28	0,04	1	0,35	2	0,07	3	0,18	2	0,05
-1	0	4	0,09			5	0,07	-1	0,07			6	0,21	5	0,05
4	0,14	-1	0,03			4	0,03	4	0,07					-1	0,03
31	0,01					21	0,02	5	0,03					3	0,02
						-1	0,01	9	0,01						
								6	0,01						
İnce															
al		**bak**		**çalış**		**çık**		**geç**		**gel**		**gir**		**git**	
ANo	**GS**	**ANo**	**GS**	**ANo**	**GS**	**ANo**	**GS**	**ANo**	**GS**	**ANo**	**GS**	**ANo**	**GS**	**ANo**	**GS**
9	0,01	1	0,6	6	0,03	1	0,15	22	0,21	1	0,4	1	0,46	14	0,08
1	0,09	17	0,15	1	0,13	52	0	18	0,07	21	0,04	14	0,08	10	0,03
21	0,02	3	0,02	5	0,31	4	0,04	8	0,08	11	0,05	5	0,03	1	0,55
7	0,14	9	0,07	3	0,3	10	0,05	1	0,24	3	0,04	3	0,14	5	0,04
8	0,02	2	0,03	2	0,19	54	0,01	36	0,05	38	0,02	10	0,06	2	0,06
28	0,01	5	0,02	4	0,04	35	0,14	-1	0,07	27	0,04	13	0,01	21	0,04
-1	0	11	0,02			17	0,1	17	0,07	7	0,03	15	0,02	15	0,02
4	0,14	19	0,04			2	0,1	28	0,01	39	0,06	8	0,02	4	0,06
3	0,02	12	0,01			19	0,03	39	0,01	31	0,03	9	0,01	-1	0,03
12	0,1	-1	0,03			53	0,01	2	0,01	28	0,01	12	0,02	16	0
30	0,04	13	0,01			13	0,06	11	0,07	24	0,12	6	0,11	6	0,02
37	0,06					9	0,05	30	0,02	10	0,01	7	0,01	3	0,02
6	0,01					28	0,04	16	0,01	6	0,01	16	0,01	12	0,01
33	0,01					14	0,02	33	0,01	16	0,01	4	0,01	11	0
38	0,09					44	0,02	40	0,02	4	0	2	0,01	20	0,01
25	0,01					47	0,02	12	0,01	17	0,03			7	0,02
14	0,03					16	0	34	0	32	0			8	0,02
19	0,01					24	0,02	15	0,01	22	0,02				
26	0,02					18	0,02	26	0,01	26	0,02				
16	0,01					27	0,01			9	0,02				
13	0,03					21	0,02			15	0,01				
2	0,01					45	0,01			12	0,01				
32	0,04					38	0			19	0,01				
27	0,02					49	0			8	0,01				
5	0,03					37	0,01			13	0,01				
11	0					43	0,05			29	0				
31	0,01					-1	0,01								
17	0,01					46	0,01								
20	0,01														
35	0,01														

Tablo 4.20: Sözcüklerin kaba/ince anlam sayıları ve taban başarım oranları(%)

	Kaba		İnce			Kaba		İnce	
Sözcük	**AS**	**Taban**	**AS**	**Taban**	**Sözcük**	**AS**	**Taban**	**AS**	**Taban**
Eylem					**yol**	5	64	10	43
al	6	47	30	14	**yüz**	6	63	6	63
bak	5	64	11	60	**İsim Ort.**	**4,17**	**68**	**7,25**	**52**
çalış	2	66	6	31	**Diğer**				
çık	7	47	28	15	**böyle**	3	61	5	36
geç	8	35	19	24	**büyük**	2	59	6	24
gel	3	67	26	40	**daha**	2	98	5	51
gir	4	58	15	46	**doğru**	3	58	8	45
git	6	74	17	55	**eski**	2	63	4	30
Eylem Ort.	**5.1**	**57.3**	**5.1**	**35.6**	**gerçek**	2	95	7	55
İsim					**küçük**	2	70	7	33
an	2	94	2	94	**nasıl**	3	73	6	43
ara	7	30	10	20	**öyle**	2	79	4	63
baş	5	57	9	27	**son**	2	91	6	63
el	5	69	6	67	**tek**	2	67	6	38
göz	6	76	8	64	**var**	3	64	3	64
kız	2	86	4	60	**yeni**	2	90	9	24
ön	3	83	10	45	**yok**	5	86	5	86
sıra	2	60	5	57	**Diğer Ort.**	**2,5**	**75**	**5,79**	**47**
üst	3	87	9	45	**Genel Ort**	**3.7**	**68.5**	**9.4**	**45.9**
yan	4	49	8	35					

Sözcüklerin derleme metinde geçen toplam ince ve kaba anlamları hesaplanmış ve bu anlamlardan en sık kullanılanın geçiş oranı taban puan olarak kullanılmak üzere belirlenmiştir (Tablo 4.20). Taban puan başarım için alt sınırı göstermektedir; çünkü yöntemlerde herhangi bir hesaplama yapılmadan doğrudan en sık geçen anlam bütün örneklere atandığında elde edilecek başarım taban puan kadar olacaktır. Bundan daha düşük bir başarım yöntemin başarısız olduğunu göstermektedir.

Farklı özelliklerin (Tablo 4.21) ve farklı algoritmaların bu sözcük gruplarının anlam belirginleştirmesine etkileri incelenmiştir ve başarım oranları hesaplanmıştır. Sonuçlar irdelenirken eylemler ayrıntılı olarak ele alınmış; diğer türlerin ise sadece ortalama değerleri üzerinde durulmuştur. Son aşamada gerçekleştirilen deneylerde 34 sözcük seçilmiş ve her bir sözcük için ince ve kaba anlamlarda 3 (AODE, IBk ve J48) algoritma kullanılmış ve 23 özellik için de işlemler yinelenmiştir. Toplam olarak 34x2x3x23=4692 test yapılmıştır. Tablo 4.22 ve Tablo 4.23 eylemler için bu testlerin ayrıntılı sonuçlarını içermektedir.

Tablo 4.21: Tablolarda kullanılan özellikler ve kısaltmaları

Özellikler	**Kısaltmalar**
hepsi	hepsi
onceKok	ÖK
onceTur	ÖT
onceHalEki	ÖHE
onceIyelik	Öİ
onceIliski	ÖŞ
onceKokTur	ÖKT
onceKokTurHalEki	ÖKTHE
onceKokTurHalEkiIyelik	ÖKTHEİ
hedefTur	HT
hedefHalEki	HHE
hedefIyelik	Hİ
hedefIliski	HŞ
sonraKok	SK
sonraTur	ST
sonraHalEki	SHE
sonraIyelik	Sİ
sonraIliski	SŞ
sonraKokTur	SKT
sonraKokTurHalEki	SKTHE
sonraKokTurHalEkiIyelik	SKTHEİ
onceKokSonraKok	ÖKSK
onceKokHalEkiSonraKokHalEki	ÖKHESKHE

Eylemler için elde edilen sonuçlarda yöntemler arasındaki farkın çok fazla olmadığı görülebilir. Ancak kullanılan özelliklere göre sonuçlarda önemli değişiklikler olmaktadır. Örneğin, al sözcüğünün ince anlamları için ÖK özelliğinde yöntemler %28-%33-%34, ÖT özelliğinde %16-%17-%17 gibi yakın sonuçlar vermiştir. Ancak IBk yöntemi ile sadece ÖK özelliği kullanılarak %33 başarı elde edilirken ÖT özelliği ile bu oran %17'ye düşmektedir. Hİ, SHE ve Sİ gibi bazı özellikler SAB etkisiz kalırken, ÖK, ÖT, HHE gibi özellikler de genelde SAB'da etkin olmuştur.

Değerlendirmeler yapılırken taban puanlar önem kazanmaktadır. Herhangi bir özellik ve yöntemle elde edilen sonuç taban puandan daha düşük veya çok yaklaşık ise bu yöntem veya özeliğin SAB için etkin olmadığı sonucu çıkarılabilir; çünkü herhangi bir işlem yapılmadan bütün anlamlar en sık geçen anlam olarak işaretlense, taban puan kadar başarım elde edilmesi sağlanır. Örneğin *al* sözcüğünün ince ve kaba anlamlarda taban puanı sırasıyla 14 ve 57 olarak verilmiştir (Tablo 4.20). Bu durumda IBk yöntemiyle ÖT, ÖHE, Öİ, HT, HHE, Hİ, HŞ, ST, SHE ve Sİ kaba anlamlarda, ÖT, Öİ, HT, HHE,

Hİ, HŞ, SHE ve Sİ ise benzer şekilde ince anlamlarda etkisiz olmaktadır. Bunların dışındaki özellikler belli oranda sonuca olumlu katkı sağlamaktadır.

Tablo 4.22: Kaba anlamlar için AODE, IBk ve J48 algoritmalarının farklı özellikler için seçilen eylemlerdeki başarım oranları

Özellik	Yöntem	Sözcük								Özellik	Yöntem	Sözcük							
		al	bak	çalış	çık	geç	gel	gir	git			al	bak	çalış	çık	geç	gel	gir	git
hepsi	**AODE**	52	71	77	61	50	71	65	75	**HŞ**	**AODE**	48	63	69	52	43	67	59	76
	IBk	57	71	78	58	52	74	63	74		**IBk**	48	62	69	52	43	67	59	76
	j48	46	69	70	64	53	67	60	75		**j48**	46	64	63	52	43	67	58	75
ÖK	**AODE**	60	64	66	51	44	71	63	75	**SK**	**AODE**	55	68	73	57	50	68	71	75
	IBk	58	61	66	53	48	72	63	73		**IBk**	57	72	73	60	53	71	71	79
	j48	46	64	65	47	36	67	58	75		**j48**	53	64	65	60	53	67	61	75
ÖT	**AODE**	46	65	67	50	37	67	57	74	**ST**	**AODE**	46	67	67	49	41	67	57	76
	IBk	46	64	67	50	36	67	57	74		**IBk**	46	67	67	49	43	67	57	76
	j48	46	64	65	48	34	67	58	75		**j48**	46	64	65	48	43	67	58	75
ÖHE	**AODE**	47	64	76	46	43	67	58	75	**SHE**	**AODE**	47	65	69	49	37	66	60	75
	IBk	47	64	76	46	44	67	58	75		**IBk**	47	67	69	49	38	67	60	74
	j48	46	64	76	45	44	67	58	75		**j48**	47	67	67	49	35	67	58	75
Öİ	**AODE**	46	64	69	47	39	67	58	75	**Sİ**	**AODE**	47	64	65	47	36	67	58	75
	IBk	46	64	69	47	39	67	58	75		**IBk**	47	64	65	47	36	67	58	75
	j48	46	64	69	47	39	67	58	75		**j48**	46	64	65	47	36	67	58	75
ÖKT	**AODE**	56	64	67	55	44	71	58	73	**SKT**	**AODE**	56	68	72	58	50	68	67	75
	IBk	60	64	68	54	49	72	63	74		**IBk**	57	72	71	60	51	71	69	79
	j48	46	64	65	47	36	67	58	75		**j48**	53	64	65	60	53	67	61	75
ÖKTHE	**AODE**	55	65	78	54	40	70	59	74	**SKTHE**	**AODE**	55	69	71	58	48	67	65	74
	IBk	59	66	76	52	47	76	63	75		**IBk**	57	72	76	61	53	71	71	79
	j48	46	64	73	47	36	67	58	75		**j48**	57	65	66	60	53	67	61	75
ÖKT HEİ	**AODE**	52	64	73	56	48	71	55	74	**SKTHEİ**	**AODE**	55	68	72	58	48	67	63	73
	IBk	55	62	70	54	48	74	61	73		**IBk**	56	72	74	61	52	71	71	79
	j48	46	64	73	47	36	67	58	75		**j48**	57	65	66	60	53	67	61	75
HT	**AODE**	47	64	65	47	36	67	58	75	**ÖKSK**	**AODE**	60	70	71	57	50	72	71	78
	IBk	47	64	65	47	36	67	58	75		**IBk**	64	73	69	59	54	73	71	74
	j48	46	64	65	47	36	67	58	75		**j48**	46	64	65	60	36	67	61	75
HHE	**AODE**	47	65	65	50	41	67	63	75	**ÖKHE SKHE**	**AODE**	60	65	79	58	51	71	68	74
	IBk	47	64	65	50	42	67	63	74		**IBk**	62	71	79	56	53	72	69	74
	j48	46	64	65	50	43	67	63	75		**j48**	46	67	73	60	53	67	61	75
Hİ	**AODE**	46	64	65	52	38	67	60	75										
	IBk	46	64	65	52	38	67	60	75										
	j48	46	64	65	52	38	67	60	75										

Taban puanın yüksek olması da sonuçları olumsuz etkilemektedir. Bunun en önemli nedeni ise belli bir anlamın fazlaca kullanılması ve diğer anlamlara ait verilerin çok az veya hiç bulunmamasıdır. Başka bir ifadeyle, anlamlar arasındaki dağılımın dengesizliği

belli bir anlamın seçilmesini sağlamaktadır. Taban puan, anlam sayısı ile orantılı olduğu zaman hem başarım oranı artmakta, hem de farklı anlamlar için yeterince örnek veri bulunduğu için elde edilen sonuçlar daha sağlıklı olmaktadır. Bu nedenle kaba anlamlarda taban puan daha yüksek olduğu için elde edilen sonuçlar ince anlamlara göre daha az artış sağlamaktadır. Kaba anlamlarda *bak, çalış, gel, gir* ve *git* taban puanları %58-74 arasında bulunduğu için sonuçlardaki katkılar da düşük gerçekleşmiştir. *Geç* sözcüğünün kaba anlamlarda taban puanı 35 olduğu için elde edilen sonuçlardaki artış %18'e kadar çıkmıştır. Benzer durumlar ince anlamlarda da gözlenebilir.

Ayrıca fazla özellikle başarım oranı doğru orantılı olarak da artmamaktadır. Bazı durumlarda tek bir özellik veya bir iki özelliğin birlikte kullanılması bütün özelliklerin veya daha büyük özellik kümelerinin kullanılmasından daha etkili olabilmektedir. Örneğin *gel* sözcüğünün ince anlamlarında btün özelliklerin kullanılması ile IBk yöntemi %48 başarılı olurken, sadece ÖK özelliği kullanılarak da aynı başarı elde edilebilmektedir. Etkin özelliklerin bulunması ve bunların kullanılması hesaplama zamanını da düşürmektedir.

Eylemler dışındaki diğer tür sözcüklerle yapılan testlerin ortalama sonuçları Tablo 4.24'te eylemlerle karşılaştırılmıştır. Bu sonuçlara göre isimlerin ve diğer sözcük türlerinin anlamlarının eylemlerden daha kolay belirginleştirildiği söylenebilir. Ayrıca ortalama sonuçlar da eylemlerde olduğu gibi diğer sözcük türlerinde de yöntemler arası farkın fazla olmadığını göstermektedir. Kullanılan özelliklerin etkisi yöntemlerden daha fazla olmaktadır. Eylemlerden farklı olarak diğer sözcük türleri önceki ve sonraki sözcüklerin ek ve köklerinden, hedef sözcüğün ek ve tür özelliklerinden de önemli ölçüde etkilenmektedir.

Tablo 4.23: İnce anlamlar için AODE, IBk ve J48 algoritmalarının farklı özellikler için seçilen eylemlerdeki başarım oranları

Özellik	Yöntem	Sözcük								Özellik	Yöntem	Sözcük							
		al	bak	çalış	çık	geç	gel	gir	git			al	bak	çalış	çık	geç	gel	gir	git
hepsi	**AODE**	30	65	57	31	38	43	57	60	**HŞ**	**AODE**	16	63	51	23	27	42	45	58
	IBk	37	67	55	39	43	48	53	55		**IBk**	16	63	51	23	28	42	44	58
	j48	35	68	54	37	43	41	65	57		**j48**	17	62	51	23	28	41	48	57
ÖK	**AODE**	28	61	39	25	40	46	50	57	**SK**	**AODE**	26	62	52	29	39	42	63	57
	IBk	33	58	33	28	37	48	52	53		**IBk**	32	69	52	33	41	48	65	63
	j48	34	62	30	28	40	41	48	57		**j48**	32	63	52	33	41	41	65	57
ÖT	**AODE**	16	61	50	16	28	42	47	58	**ST**	**AODE**	20	62	47	18	29	42	52	58
	IBk	17	61	49	16	27	41	48	58		**IBk**	21	62	47	18	28	42	52	56
	j48	17	62	49	16	24	41	48	57		**j48**	21	62	47	18	28	41	52	57
ÖHE	**AODE**	20	61	52	16	27	41	48	56	**SHE**	**AODE**	16	63	36	16	27	41	49	58
	IBk	20	61	52	16	27	41	49	57		**IBk**	16	64	36	16	28	41	49	58
	j48	20	62	52	16	25	41	48	57		**j48**	16	64	37	16	27	41	48	56
Öİ	**AODE**	16	62	38	15	29	41	48	57	**Sİ**	**AODE**	17	62	30	15	24	41	49	57
	IBk	16	62	38	15	29	41	48	57		**IBk**	17	62	30	15	24	41	48	57
	j48	16	62	38	14	29	41	48	57		**j48**	17	61	28	15	24	41	48	57
ÖKT	**AODE**	29	62	48	24	39	46	49	58	**SKT**	**AODE**	28	67	54	29	38	43	62	57
	IBk	36	60	51	27	40	47	51	56		**IBk**	30	69	55	33	39	48	63	63
	j48	34	62	30	28	39	41	48	57		**j48**	32	63	52	33	41	41	65	57
ÖKTHE	**AODE**	30	61	54	25	33	46	47	58	**SKTHE**	**AODE**	28	67	52	30	37	43	59	57
	IBk	38	60	53	27	37	50	51	56		**IBk**	32	70	55	35	39	47	64	62
	j48	34	62	30	28	41	41	48	57		**j48**	32	67	52	33	41	41	65	57
ÖKTHEİ	**AODE**	30	61	51	25	35	47	44	58	**SKTHEİ**	**AODE**	28	66	50	29	38	42	59	58
	IBk	35	57	48	28	33	50	45	52		**IBk**	33	71	55	36	39	47	64	62
	j48	35	62	30	29	40	41	48	57		**j48**	32	66	52	33	41	41	65	57
HT	**AODE**	16	62	31	15	25	41	48	57	**ÖKSK**	**AODE**	36	65	54	31	43	47	63	61
	IBk	16	62	31	15	25	41	48	57		**IBk**	39	70	53	36	44	51	62	56
	j48	16	62	31	15	25	41	48	57		**j48**	34	62	30	31	40	41	65	57
HHE	**AODE**	15	62	39	16	28	41	55	57	**ÖKHESKHE**	**AODE**	39	62	61	27	40	46	58	57
	IBk	15	62	39	16	29	40	55	57		**IBk**	39	65	62	29	40	50	60	56
	j48	15	62	39	16	27	41	55	57		**j48**	35	65	52	33	43	41	65	57
Hİ	**AODE**	16	62	26	16	24	41	51	57										
	IBk	16	62	26	16	24	41	51	57										
	j48	16	62	26	16	23	41	51	57										

Tablo 4.24: İnce ve kaba anlamlar için AODE, IBk ve J48 algoritmalarının farklı özellikler için seçilen sözcüklerdeki ortalama başarım oranları (%)

	Kaba			İnce		
Tür	**AODE**	**IBK**	**J48**	**AODE**	**IBK**	**J48**
Eylem	61	63	60	43	46	45
İsim	81	82	75	70	73	66
Diğer	85	88	74	57	65	61
Genel	75	77	69	57	61	57

4.4. EYLEMLER ÇALIŞMASINA KAVRAMSAL SÖZLÜK EKLENMESİ

Çalışmanın bu aşamasında örnek olarak oluşturulan küçük bir kavramsal sözlük ek bir özellik olarak SAB'da kullanılmıştır. Diğer dillerde yapılan SAB çalışmalarında sıkça kullanılan WordNet benzeri kavramsal bir sözlük Türkçe için henüz tam olarak geliştirilmemiştir. Bu nedenle kavramsal sözlük kullanımının SAB üzerindeki etkisinin incelenmesi amacıyla eylemler için sınırlı bir sözlük oluşturulmuştur. Seçilen eylemlerle birlikte kullanılan ilişkili sözcük köklerinin hem eylemden önce, hem de eylemden sonra geçenler için üç düzeyden oluşan bir işaretleme yapılmıştır. Bu üç farklı düzeyde kullanılan sınıf andaçlarının WordNet'te kullanılan andaçlarla uyumlu olarak tanımlanmasına özen gösterilmiştir. Ancak WordNet içinde sadece isimler ve eylemler ayrıntılı olarak sınıflandırırken, diğer sözcük türleri için benzeri bir sınıflandırma yapılmamıştır. Ayrıca kullanılan sınıflandırmada eylemler ve isimler arasındaki benzerlik, yakınlık ve ilişkileri yansıtacak bir sınıflandırma yapılmamış; her iki tür için farklı andaçlar kullanılmıştır.

Türkçe için oluşturulan kavramsal sözlük, eylem ve isim dışında kalan sözcük türlerini de sınıflandırmıştır. Sözcük türlerini arasındaki benzerlik, yakınlık ve ilişkileri yansıtmak amacıyla mümkün olduğunca ortak sınıf andaçları kullanılmıştır. Bu sınıflandırma işleminde her üç düzeyde kullanılan andaçlar ve bunların hangi sözcükleri kapsadığı Tablo 4.25 ve Tablo 4.26'da gösterilmiştir. Birinci düzeyde sadece soyut ve somut kavramlar ayırdedilmiştir. İkinci ve üçüncü düzeyde daha kapsamlı bir sınıflama yapılmıştır. Birinci düzey en genel, üçüncü düzey ise en özel sınıflamayı göstermektedir.

Toplam olarak 1082 farklı sözcük için sınıflandırma işlemi tamamlanmıştır. Bunların çoğu incelenen eylemden önce gelen sözcüklerdir; çünkü Türkçe'de genelde eylemden sonra tümce bitmekte ve başka sözcük bulunmamaktadır. Örneklerde de sonra gelen sözcüklerin çoğu noktalama işareti olmuştur. Örneğin "*Şimdi tekrar eve girdiler.*" tümcesinde *gir* sözcüğünün SAB'ı için ÖK özellikleri arasında *ev, tekrar* ve *şimdi* sözcükleri ele alınmıştır. SK özelliği "*punc*" yani noktadır. Bu nedenle sınıflandırma

yapılmamıştır. ÖK özellikler için sınıflandırmada ise aşağıdaki sınıflandırmalar yapılmıştır.

- Ev: noun physical entity location region
- Tekrar: adverb abstraction quantity measurement unit
- Şimdi: adverb abstraction quantity time

Tablo 4.25: Kavramsal sözlükte birinci düzeyde kullanılan sınıf andaçları

Düzey 1	
Andaç	**Anlamı**
abstract entity	Soyut varlık
physical entity	Somut varlık

Tablo 4.26: Kavramsal sözlükte ikinci ve üçüncü düzeyde kullanılan sınıf andaçları

Düzey 2		**Düzey 3**			
Andaç	**Anlamı**	**Andaç**	**Anlamı**	**Andaç**	**Anlamı**
Artifact	İnsan eliyle Yapılan ürün	**Animal**	Hayvan	**Measurement Unit**	Ölçü birimi
Change	Hal/durum değişimi	**Attitude**	Tavır/tutum	**Part**	Parça/bölüm
Cognition	Bilişsel etkinlik	**Attribute**	Özellik	**Perception**	Algı
Grouping	Grup	**Body**	Vücut bölümü	**Person**	Kişi
Location	Yer	**Business**	İş/ekonomi terimi	**Plant**	Bitki
Motion	Hareket	**Cause**	Sebep	**Possession**	Mal/mülk/iyelik
Organism	Organizma	**Communication**	İletişim etkinliği	**Process**	İşlem/süreç
Quality	Nitelik	**Container**	Kap	**Region**	Bölge
Quantity	Nicelik	**Delight**	Zevk verici madde	**Society**	Toplum/topluluk
Relation	İlişki	**Dress**	Kıyafet	**State**	Durum
Stative	Durum/oluş Bildiren etkinlik	**Drink**	İçecek	**Time**	Zaman
Substance	Madde	**Emotion**	Duygu	**Tool**	Alet/araç
Thing	Diğer	**Food**	Yiyecek	**Vehicle**	Taşıt
		Material	Materyal		

Tablo 4.27: İnce ve kaba anlamlar için kavramsal sözlük kullanılarak elde edilen başarım oranları (%)

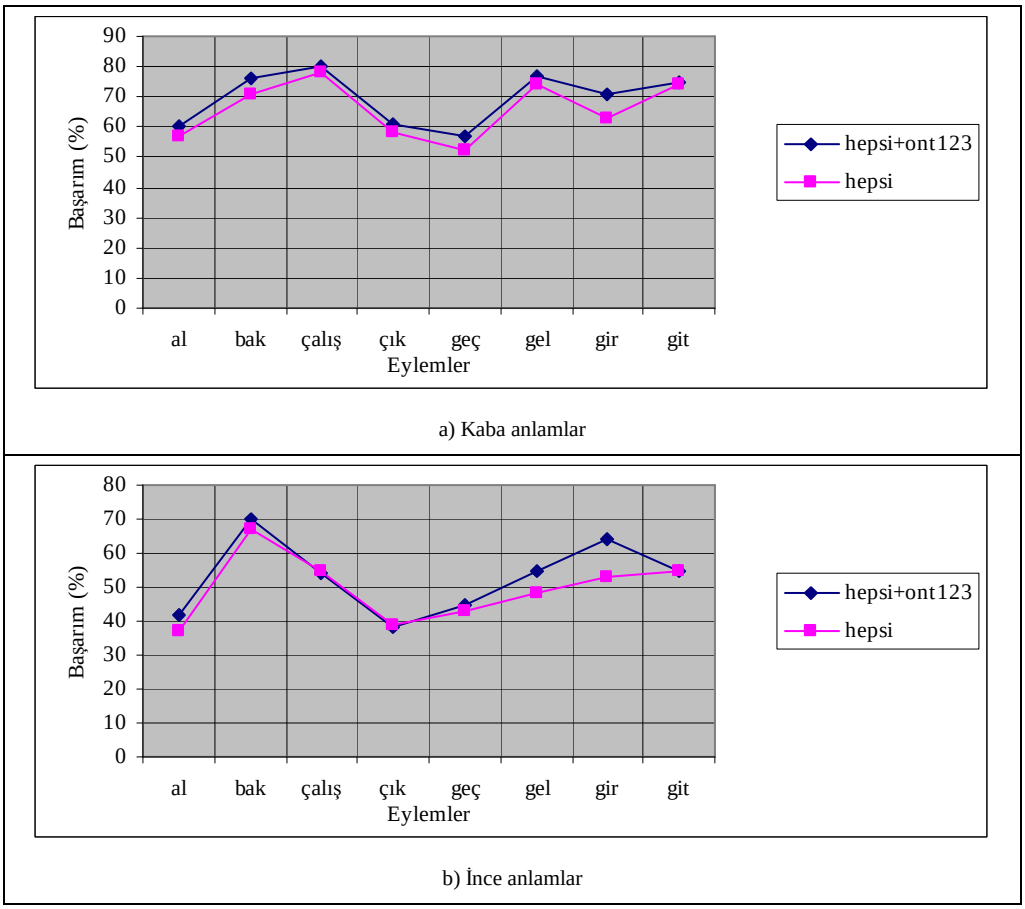
Kaba	hepsi+ ont123	ÖKTHEİ+ ont123	ÖKTHEİ+ ont12	ÖKTHEİ+ ont1
al	60	59	57	57
bak	76	67	68	66
çalış	80	76	75	74
çık	61	60	59	61
geç	57	53	53	49
gel	77	75	75	75
gir	71	65	63	59
git	75	75	73	73
İnce				
al	42	37	36	37
bak	70	61	62	60
çalış	54	46	48	49
çık	38	32	33	34
geç	45	41	41	36
gel	55	53	53	52
gir	64	55	52	50
git	55	54	55	53

Tablo 4.28: Kavramsal sözlük kullanımıyla elde edilen kazanım

Kaba	hepsi+ ont123	ÖKTHEİ+ ont123	ÖKTHEİ+ ont12	ÖKTHEİ+ ont1
al	3	1	-1	-1
bak	5	6	7	5
çalış	2	10	9	8
çık	3	7	6	8
geç	5	5	5	1
gel	3	3	3	3
gir	8	2	0	-4
git	1	2	0	0
İnce				
al	5	4	3	4
bak	3	3	4	2
çalış	-1	13	15	16
çık	-1	4	5	6
geç	2	4	4	-1
gel	7	5	5	4
gir	11	3	0	-2
git	0	1	2	0

Sözcüklerin sınıflandırılması işlemi tamamlanıp, kısıtlı da olsa sınırlı bir kavramsal sözlük elde edildikten sonra, eylemler için yapılan SAB testleri, bu yeni bilgiler eklenerek tekrarlanmış ve Tablo 4.27'deki sonuçlar elde edilmiştir. IBk yönteminde daha önce kullanılan özelliklere ek olarak üç düzeyden oluşan sınıflama, hem önceki

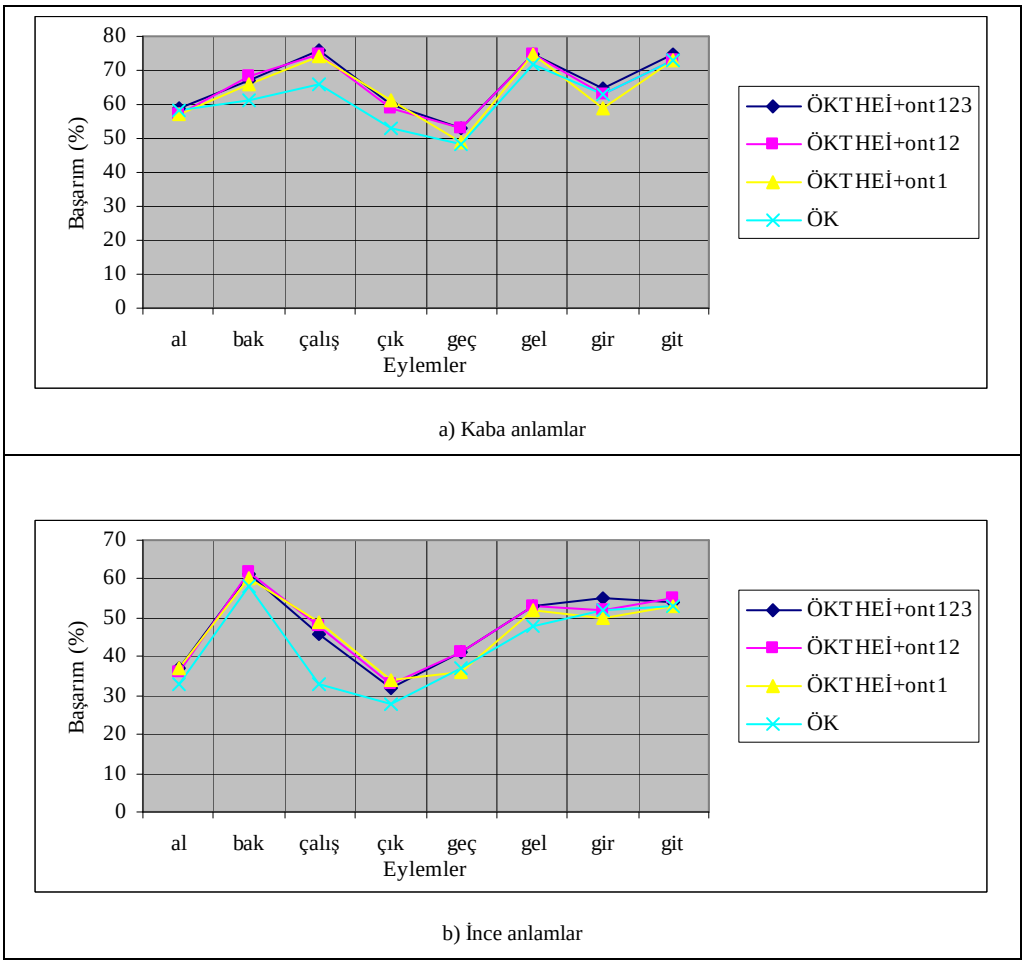
hem de sonraki sözcük kökü için alınmış ve tabloda ikinci sütunda verilen yeni sonuçlar elde edilmiştir. Tabloda üç, dört ve beşinci sütunlarda SAB yapılan eylemden önce gelen sözcüğün kökü, türü, hal ve iyelik eki bilgileri ile birlikte farklı düzeylerdeki sınıflama ile hesaplanan sonuçlar verilmiştir. Üçüncü sütunda her üç düzey, dördüncü sütunda birinci ve ikinci düzey, son sütunda ise sadece bir düzey sınıflama kullanılmıştır.



Şekil 4.1: İnce ve kaba anlamlar için eylemlerde kavramsal sözlüğün bütün özelliklerle kullanılmasının başarıma etkisi

Kavramsal sözlük ve tüm özellikler kullanıldığında elde edilen kazanımlar Tablo 4.28'de verilmiştir. Kaba anlamlarda önceki sonuçlara göre %1-8 arasında değişen olumlu bir ilerleme görülürken, ince anlamlarda üç eylem dışında %2-11 arasında değişen olumlu sonuçlar gözlenmiştir (Şekil 4.1). İnce anlamlarda *git* eylemi için sınıflamanın sonuçlar üzerinde bir katkısı olmazken, *çalış* ve *çık* eylemlerinde olumsuz etkisi olmuştur. Bunun nedeni ise, örnek sayısının çok sınırlı olması ve bunların sınıflandırmada sözcükler arasında yeterince ilişki ve genelleme sağlayacak yeterlilikte

olmamasıdır. Önceki kök sözcük ve ilgili özellikleri ile sınflandırma birlikte kullanıldığında kaba anlamlarda %1-10, ince anlamlarda ise %1-16 arasında iyileşme görülmektedir (Şekil 4.2). Bazı durumlarda sonuçların kötüleşmesi, genel sınıflandırmadan özel sınıflandırmaya gidildikçe sonuçların olumsuz etkilenmesi örnek saysının azlığı ile ilişkilendirilebilir. Örnekler çoğaldıkça ve kavramsal sözlük tam olarak ortaya çıkmaya başladıkça sınıflandırmanın da olumlu etkisinin artacağı düşünülmektedir.



Şekil 4.2: İnce ve kaba anlamlar için eylemlerde kavramsal sözlüğün önceki kök sözcük ve diğer özellikleriyle kullanılmasının başarıma etkisi

5. TARTIŞMA VE SONUÇ

Yapılan çalışmada elde edilen sonuçlar üzerinde etkili olan bazı faktörler ile gelecekte yapılabilecek çalışmalar bu bölümde ele alınmıştır.

5.1. SONUÇLARDAKİ HATALARIN ANALİZİ

Türkçe SAB için yapılan çalışmada şu ana kadar elde edilen sonuçlar, tümcelerin biçimsel bilgileri ve eylemler için oluşturulan sınırlı bir kavramsal sözlük kullanılarak ortaya çıkarılmıştır. Ancak bu işlemler anlam belirginleştirmede hata payını arttıran bazı çok önemli faktörleri de ortaya çıkarmaktadır. Bunların bir kısmı kullanılan veri ve yazılımlardan kaynaklanan fiziksel faktörlerken, diğerleri Türkçe'nin yapısından ve DDİ çalışmalarındaki temel problemlerdir.

5.1.1. Veri ve Yazılımlardan Kaynaklanan Hatalar

- *XML dosyalarındaki hatalar:* Bu veriler ODTÜ tarafından kullanıma yeni sunulduğu için, içerisinde hem XML yapısına uymayan, hem de yanlış ya da eksik işaretlemelerin yapıldığı dosyalar olduğu görülmüştür. İşaretlemede problem oluşturan bazı tümceler tümü ile boş bırakılmış veya *DISCARD*, *PROBLEM* gibi andaçlar konulmuştur. Bazı durumlarda ise yanlış işaretlemeler nedeniyle seçilen örneklerin özellikleri hatalı veya eksik olmuştur. Bunların bir kısmı düzeltilmiş, bir kısmı da düzeltilmeye çalışılmaktadır.
- *İlişkisel gösterim:* XML dosyalarındaki sözcükler arasındaki bağlar, ilişkisel bir yaklaşımla gösterilmiştir. Bu yaklaşımın bazı noktalarda olumlu, bazı noktalarda ise olumsuz etki yaptığı söylenebilir. Türkçe'de tümce öğeleri yer değiştirebilmekte, devrik tümceler bulunabilmekte ve sözcüklerin bağlantılı olduğu diğer sözcüklerle aralarına pek çok sözcük girebilmektedir. İlişkisel gösterim Türkçe'nin yapısına uygunluk göstermekte ve ilişkili sözcükler arasında yer değişkliği olsa bile incelemeye alınabilmektedir. Ancak, sözcükler arasında doğrudan ilişki olmadığı halde anlama etki etme söz konusu olabilir. Bu nedenle bazı anlam belirginleştirmesi işlemleri için ilişkisel yapıdan alınabilenden daha fazla sözcüğe gereksinim olduğu durumlar vardır. Örneğin, *Akla hayale gelmez işler* ve *Akla gelmez işler* ifadelerinde *akla gelmek* ve *akla*

hayale gelmek aynı anlamdadır; ancak bir tümcede *akla* sözcüğü, diğerinde ise *hayale* sözcüğü ilişkisel yapıdan alındığı için iki farklı durum gibi algılanmaktadır.

- *ARFF dosyalarındaki fazla bilgiler:* ARFF dosyalarında her ilişkili sözcük bir veri oluşturur, ama X anlamı bu ilişkilerden sadece birinden dolayı oluşmaktadır. Diğer ilişkili sözcükler başka anlamda olabilmesine rağmen, X anlamı olarak sınıflandırılmaktadır. Hata payını azaltmak için diğer anlamlarda da olabilecek özellikler çıkarılmalıdır.

5.1.2. Dilin Yapısından Kaynaklanan Problemler

Türkçe'nin biçimsel yapısından kaynaklanan ve sözcük anlamlarına etki eden pek çok farklı etken vardır. Bunlar aşağıdaki örneklerde görüldüğü gibi sözcüğün mecaz veya gerçek anlamda kullanılmasına, sözcüğün tipine (isim, eylem, adıl) göre değişiklik gösterebilmektedir.

- **Aklına** bir soru **geldi**
- **Elimizden geleni** yaptık.
- Bebek artık **ele gelir** oldu.
- Bağlama**yı eline aldı**.
- Konu**yu ele aldı**.
- Bir yar sevdim **el aldı**.

Birinci tümcede *gel* sözcüğünün anlamına önceki sözcük *akıl*, bu sözcüğün aldığı *iyelik* ve *-e hal eki;* ikinci tümcede önceki sözcük *el*, *iyelik* ve *-den hal ek;* üçüncü tümcede ise önceki sözcük *el* ve *-e hal eki* etki etmektedir. Dördüncü tümceden altıncı tümceye kadar *al* sözcüğünün anlamı yine önceki sözcük olan *el* ile *el* sözcüğünün aldığı *hal* ve *iyelik eklerine* bağlı olarak belirginleştirilmektedir. Ayrıca dördüncü ve beşinci tümcelerde ilişkili diğer sözcüklerin aldığı *–i hal eki* de önem kazanmaktadır. Biçimsel ipuçlarının SAB'da ancak belli bir noktaya kadar etkin olabildiğini kabul etmek gerekir. Özellikle eylem türündeki sözcüklerin anlam belirginleştirmesi için biçimsel yapının dışında en az bir önceki sözcüğün de incelenmesi gerekebilir. Aşağıdaki tümcelerde *gel* sözcüğünün anlamını bulmak için önceki sözcüklerin bu tip özellikleri göz önüne alınmalıdır:

- Taş yukarıdan geldi.
- Emir yukarıdan geldi.

- Adam yukarıdan geldi.

Bu örneklerde geçen *gel* sözcüğünden önceki sözcük olan *yukarıdan* hem kök, hem aldığı ekler, hem de öğe olarak aynıdır. Ondan önceki sözcükler de işlevsel olarak aynı sınıftadır. Ancak her bir tümcedeki anlam tamamen farklıdır. Birinci tümcedeki taşın yukarıdan gelmesi tamamen taşın *düşme* özelliğine sahip olmasından kaynaklanan bir anlam içermektedir ki, bunun biçimsel öğelerden çıkarılabilmesi mümkün değildir. İkinci tümcedeki yukarıdan gelmek anlamını ise emir sözcüğü belirlemektedir. Burada anlam *bir üst makamdan gelme, buyurulma* şeklinde olmaktadır. Üçüncü tümcedeki anlam ise sadece bu tümceyle çözülemeyecek kadar karmaşıktır. Adam *yukarıda bulunan bir yerden aşağıya geldi* şeklinde algılanabileceği gibi, önceki bağlama göre *adamın üst bir makamın torpili ile geldiği* de anlaşılabilir.

Yukarıdakilere benzer örnekler sadece biçimsel bilgilere dayanarak anlam belirginleştirmesi yapılamayacak durumlardan sadece bir kaçıdır. Bu üst basamakta sözcükler arası ve sözcüğün kendine özgü bazı özelliklerini işleme dahil etmek yararlı olmaktadır. Belirginleştirmedeki hata payını düşürmek amacı ile, sözcükleri genel anlamlarına göre gruplandırıp, grup ismini analiz sırasında özellikler kümesi içerisinde kullanmak bir çözüm olabilir. Aynı anlam grubundaki sözcükler öğeleri göz önüne alınmaksızın sınıflandırılabilir. Ancak gruplandırma iyi yapılmalı, birden fazla gruba girebilecek sözcükler dikkatli olarak sınıflandırılmalıdır. Tümcede geçen ve hedef sözcükle birinci dereceden ilişkili olmayan diğer sözcükler ve bu sözcüklerin genel sınıfları alınıp, sözcüğün herbir anlamıyla geçiş sıklıklarına göre birlikte kullanılan bu sözcükler sıralanarak, anlam sınıflandırmasına etkisi ölçülebilir. Sıralamada alt düzeylerde yer alan veya birden fazla anlamla birlikte kullanılan sözcük ve sözcük sınıfları anlam belirginleştirmede etkisiz kabul edilebilir. Ayrıca bu şekilde bir sınıflama yapılabilirse anlama etki eden özellikler bulunurken de bu sınıflardaki sözcükler için genellemeler yapılması mümkün olabilir. Çalışmaya bu yaklaşım eklenmemiştir, çünkü hedef sözcükle birlikte geçen bütün sözcükleri gözönüne almak ve bunların tamamını sınıflandırmak oldukça uzun bir zaman ve uzman dilbilimci çalışması gerektirmektedir. Kısıtlı da olsa sadece birinci dereceden ilişkili sözcüklerin sınıflandırılması işlemi bile uzun bir süreç sonucunda elde edilmiştir. Bu işlem her ne kadar başarıma genel olarak olumlu katkı yapmış olsa da hesaplama zamanınını da arttırmıştır.

Yukarıda belirtilenlerin dışında, aynı sözcüğün önceki kullanımlardaki anlamı daha sonraki kullanımlardaki anlamlarını da etkileyebilir. Örneğin,

- **Korktuğun başına gelebilirdi**, ama **gelmedi.**
- Bu at **hoşuna gitti** mi? **Gitmez** olur mu hem de çok **gitti.**

Bu tümcelerden de görülebileceği gibi, anlama etki eden özellikler ikinci kullanımda bulunmamaktadır. Sözcük anlamının belirginleştirilmesi için eksik özellikler önceki kullanımdan alınmakta ve otomatik olarak insan beyni tarafından daha sonraki kullanımlar için ilave edilmektedir. Bu nedenle önceki tümcelerdeki anlamı da incelemek yararlı olacaktır. Yukarıdaki örneklerde ilk tümcede incelenen *gel* sözcüğünün bir kullanım öncesi, ikinci tümcede ise *git* sözcüğünün iki kullanım öncesi anlamı etkilemektedir. Türkçe'deki bu özellik ile farklı tümcelerde daha da ileri düzeylerde de karşılaşılmaktadır.

Sadece daha önceki kullanımda geçtiği için değil, başka sebeplerle de eksik sözcükler bulunabilir. Özellikle gazete haberleri ve günlük konuşma dilinde bu tip durumlara sık rastlanır. Örneğin,

- Galatasaray Fenerbahçe'ye 5 attı.
- Bu maddede kırkaltı ret, dört yüz otuz dört kabul çıktı.

Herhangi bir bildirideki bir sözcüğün atlanması, verilmek istenen anlamı değiştirebilir. Yukarıdaki birinci örnekte, *gol* sözcüğünün atlanmasıyla belirsizlik oluşmuştur. İkinci örnekte ise *oy* sözcüğü eksik bırakılmıştır.

Sözcük anlamlarına etki eden diğer önemli bir konu da konuşmacının maksadının ne olduğudur. Bazı durumlarda olumlu tümceler olumsuz anlamda da kullanılabilmektedir.

- Sen her işin üstesinden gelirsin zaten!
- İçerisi ne kadar da sessiz değil mi?

Gibi ifadeler bu tür iki örnektir. İlk tümcede, bir işi başaramayacağı düşünülen veya sürekli olarak her işi yapabileceğini düşünen kişiye, bu düşüncesinin yanlış olduğunu ifade etmek için kullanılmış imalı bir ifade yer almaktadır. Benzer şekilde ikinci tümcede de, gürültülü bir ortamı ifade etmek için olumlu bir tümce, olumsuz anlamda kullanılmıştır.

Bunlardan başka, tümcelerin içinde yanlış anlamda kullanılan sözcüklerin sayısı hiç de az değildir. Anlamları farklı olan bazı sözcüklerin, sözcükler arasındaki anlam farkına dikkat edilmeksizin birbirinin yerine kullanılması veya sözcüklerin taşıdıkları anlam dışında kullanılmaları, yanlış kullanımlara ve anlatım bozukluklarına yol açmaktadır. Aşağıdaki tümcelerde tümcelerde *tepki, borçlu* ve *dehşet* sözcükleri anlamları dışında kullanılmıştır:

- Benzin fiyatlarının düşmesine halk olumlu *tepki* gösterdi.

→ *tepki* Karşı çıkma, geri tepme anlamına gelen tepki sözcüğü olumlu olamaz.

- Başarısızlığını düzensiz çalışmasına *borçludur*.

→ *borçlu* sözcüğü, taşıdığı anlam dışında kullanılmıştır.

- Bu film *dehşet* güzel

→ *dehşet* taşıdığı anlam dışında kullanılmıştır.

Kısaca özetlenecek olursa, bu çalışma kapsamında anlamların elde edilmesi, anlam işaretleme ve kavramsal sözlük ilk örneğinin oluşturulması ile ilgili çalışmaların tümü bilgisayar bilimi bakış açısıyla gerçekleştirilmiştir. Çalışmanın her bir aşamasında profesyonel dilbilimcilerin katılımı mümkün olmamıştır. Oysa Senseval çerçevesinde veya bu konuda yapılan diğer çalışmalarda uzman dilbilimciler bu aşamalara aktif olarak katılmakta ve daha sağlıklı derleme metinler ortaya konmaktadır. DDİ kapsamında elektronik ortamda hazırlanan kaynaklar da bu çalışmaların başarısında etkin olmaktadır. WordNet gibi önemli kaynaklar yıllar süren uzman çabalarıyla elde edilmiştir. Türkçe kavramsal sözlük SAB çalışmasında sonuçları olumlu yönde etkilemiştir. Türkçe için de hem biçimbilimsel hem de anlamsal olarak uzmanlar tarafından işaretlemesi ve işlenmesi gerçekleştirilmiş derleme metinlerin hazırlanması gerektiği ortadadır.

Bu bağlamda 2007 yılında gerçekleşecek olan *Senseval 4/Semeval 1* çalıştayına Türkçe SAB çalışmasının dahil edilmesi için başvuru yapılmış ve ilk defa Türkçe *Turkish Lexical Sample Task* adı altında bir işle uluslarası önemli bir çalışmada yer almıştır. Çalıştaya katılım için uzman dilbilimci ve bilgisayar bilimcilerden destek alınması için bir ekip oluşturulmuştur. Türkçe ile ilgili çalışmalar çalıştay takvimine bağlı olarak sürdürülmektedir. Bu alanda Türkçe için eksikliği duyulan böyle bir çalışmanın

oluşturulması hem Türkçe DDİ çalışmalarını olumlu yönde etkileyecek hem de çok yeni bilimsel çalışmalara kapı açacaktır.

5.2. GELECEKTE YAPILACAK ÇALIŞMALAR VE SONUÇLAR

SAB ile ilgili olarak çalışmanın başından itibaren verilen bilgilerden de anlaşılacağı gibi, bu konu oldukça uzun bir süredir incelenmesine rağmen, pek çok problem henüz tam olarak çözümlenememiştir. Anlamların bulunması ve işaretlenmesinden, kullanılacak kaynaklara, uygulanacak yöntemlerden değerlendirme ve karşılaştırma şekillerine kadar pek çok konuda henüz tam bir standart geliştirilememiştir. Ancak genel olarak bazı çıkarımların yapılabilmesi, bazı değerlendirmelerin de sözlü olarak ifade edilmesine rağmen, bu çalışmalar yeterli düzeyde değildir. Örneğin SAB için sözcüğün geçtiği paragraf, bölüm, konu gibi bağlama duyarlı bazı özelliklerin etkisi olduğu iddia edilmekle birlikte, bununla ilgili olarak kanıtlanmış bilimsel bir veri yoktur.

Ayrıca SAB, çalışmanın başlangıcında da açıklandığı gibi tek başına bir uygulama olmayıp, DDİ uygulamaları için gerekli bir ara basamaktır. Bu nedenle yapılan çalışmalarda değerlendirmeler her ne kadar belli bir uygulamadan bağımsız olarak ele alınıyorsa da, asıl önemli olan SAB işleminin ana uygulamalarda ne oranda etkili olduğunun belirlenmesidir. Örneğin BÇ veya İnternet üzerinde yapılan arama işlemlerinde SAB kullanılarak ve kullanılmadan elde edilecek sonuçlar karşılaştırılmalı ve buradan elde edilecek sonuçlara göre SAB'a ne kadar gereksinim olduğu çıkarılmalıdır. Bu tür uygulamaların başında da çeviriler gelmektedir. Çeviriler içinde SAB kullanımı iyi bir başlangıç noktası olacaktır.

Bu nedenle, bu alandaki önemli bir çalışma olan Senseval projesindeki araştırmaların önemli bir dezavantajının da gerçek uygulamalardan bağımsız olarak SAB sistemleri arasında değerlendirme yapması olduğu söylenebilir. Belirli bir alanda veya uygulamada bu sistemlerin ne kadar başarılı olacağı, kendi başına bir SAB sisteminin etkili olabileceği bir alanın bulunup bulunamayacağı ya da daha önemlisi ne kadar ayrıntılı anlam sınıflaması yapmanın gerekeceği cevap bekleyen sorular arasındadır.

Senseval 3 çalışmaları hazırlık aşamasındayken bu nedenlerden dolayı SAB sistemlerini değerlendirebilecek uygulamaya bağımlı veya uygulamadan bağımsız farklı işlerin önerilmesi için çağrı yapılmıştır. Bu bağlamda, farklı dillerin çalışmaya dahil edilmesi, çok dilli uygulamalara ağırlık verilmesi, BÇ veya bilgi çıkarımı gibi özel bir DDİ alanında kullanılması gibi fikirler üzerinde yoğunlaşılmıştır. SAB ile ilgili olan anlamsal andaçlama ve alan sınıflandırması gibi konulardaki işlemler de yan ürünler olarak ortaya çıkabilecek çalışmalardır.

Önemli diğer bir konu ise kullanılan sözcük ve anlamlarına bağlı olarak sonuçlarda ortaya çıkan önemli sapmalardır. Bu durumda yapılan çalışmalardaki sonuçların ne kadar güvenilir olduğu konusu da tartışılmaktadır. Örneğin, son yapılan testler içinde kullanılan *an* sözcüğü derleme metinde %95 oranında bir tek anlamında kullanılmıştır. Dolayısıyla uygulanan yöntem ve seçilen özelliklerin çoğu için yapılan testlerden elde edilen sonuçlar çok iyi çıkmaktadır. Sadece bu tür sözcüklerle elde edilecek sonuçlar çok yanıltıcı olabilir. Hatta bu tür sözcükler ortalama değerlerini de olduğundan yüksek çıkarabilir. Bu nedenle çalışmada yapıldığı gibi, net başarım oranları ile dağılımların da dikkatlice incelenmesi gerekmektedir.

Sözcük türlerine göre çalışmalarda kullanılan özellikler, anlamlar, yöntemler ve değerlendirmeler de farklılık gösterebilmektedir. Genelde SAB çalışmalarında ele alınan sözcük grupları eylemler ve isimlerdir. Ortalama anlam sayısı en fazla olan sözcük türü eylemlerken, bunları isimler izlemektedir. Diğer sözcük türleri bu iki türe göre çok daha az anlama sahiptir. Eylemlerde ince anlamlar alındığında anlam sayısı çok fazla olmasına rağmen kaba anlamlar için sınıflandırma yapıldığında anlam sayısı önemli oranda düşebilmektedir. Bunun en önemli nedeni ise, eylemlerde anlamların genelde birbirine çok yakın olması ve sözlüklerde bu anlamların hepsinin ayrı bir anlam olarak sınıflandırılmasıdır. Ancak isimlerde verilen anlamlar daha zor birleşmekte ve sözlüklerde eylemlere göre daha kolay ayırdedilebilen anlamlar bulunmaktadır. Bu nedenle ince anlamlarda anlam sayısı eylemler gibi çok belirgin bir düşüş göstermemektedir.

Kullanılacak anlamlar için üzerinde uzlaşma sağlanacak ve bu iş için hazırlanmış anlam kaynakları gerekmektedir. Özellikle WordNet anlamlarına geçişte işaretleyiciler arası

uzlaşmada görülen düşüş bu düşünceyi desteklemektedir. Senseval 2'den çıkan önemli bir sonuç ise iyi bir anlam sınıflandırmasının ve incelenen sözcüğe ait uygun anlamlar listesinin elde edilmesinin öneminin kavranmasıdır. İnsanların bile ayırt etmekte zorlandığı anlamlarda bilgisayarların başarılı olmasını beklemenin ya da bu işlemi yapan yöntemleri değerlendirmenin olası olmadığı vurgulanmıştır. Bu nedenle uygun anlam sınıflandırmaları oluşturmak ve de bu anlamlarla derleme metinlerde işaretlemeler yapmak için yeni yöntemlerin geliştirilmesi için sözlükbilimciler ve anlambilimcilerle işbirliği içinde çalışmalar yapılması gerekmektedir.

Anlam işaretlemesinde kullanılan anlam kümesinin sonuçlara etkisi kaba ve ince anlamlardan elde edilen sonuçlarda, Senseval'deki kaba, ince ve karışık anlamlardan elde edilen sonuçlara benzer olarak çok net bir şekilde görülebilir. Ancak kaba ve ince taneli değerlendirmelerin karşılaştırılması hiç de kolay gerçekleşmemektedir. Sonuç olarak; bazı çıkarımlar elde edilmesine rağmen, genellemeler yapmak için yeterince veri oluşturulamamıştır.

Bugüne kadar yapılan SAB çalışmalarında en başarılı sistemler denetimli sistemler olmuştur. Ayrıca anlamlara etki eden en önemli faktörlerden birisi kullanılan özelliklerdir. Bu nedenle algoritmaların farklı çok anlamlılık durumlarındaki başarımlarında etkili olan özelliklerin seçimi üzerine araştırmalar yapmak önemli ilerlemeler sağlayacaktır. Farklı yöntem ve özelliklerin çokanlamlılık hakkında daha ayrıntılı bilgiler verebileceği görüşü yaygındır ve özellik seçimi ile ilgili çalışmalara ağırlık verilmesi gerekmektedir.

Ayrıca kavramsal sözlüklerin SAB çalışmalarında en sık başvurulan kaynaklar olması ve bu tür kaynakların kullanılması ile elde edilen olumlu sonuçlar; bilgi kaynaklarının geliştirilmesini zorunlu kılmaktadır. Türkçe için bu kaynakların sınırlı olması bu alanda yapılacak çalışmaların önünde bulunan en önemli engellerden biridir. Türkçe SAB çalışmasında, oluşturulan kısıtlı kavramsal sözlük ve metinlerle elde edilen sonuçlar oldukça ümit vericidir.

Kavramsal sözlüğün kapsamının geliştirilmesi ve kullanılan tümcelerin arttırılması ile çok daha başarılı sonuçlar elde edilecektir, ancak bu çalışmaların uzun bir sürede

gerçekleştirilebileceği gözardı edilmemelidir. Bu çalışma, bu sürece başlangıç olacak ve ilerideki çalışmalara ışık tutacak bir ilk adım olmuştur.

KAYNAKLAR

ABNEY, S., SCHAPIRE, R. E., SINGER, Y., 1999, Boosting Applied to Tagging and PP–attachment, *In Proceedings of EMNLP–VLC'99*, 100-108.

ADRIAENS, G., SMALL, S. L., 1988, Word Expert Revisited In A Cognitive Science Perspective, In Small, S.; Cottrell, G. W., Tanenhaus, M. K. (Eds.) *Lexical Ambiguity Resolution: Perspectives From Psycholinguistics, Neuropsychology, And Artificial Intelligence*, Morgan Kaufman, San Mateo, California, 13-43.

AGIRRE, E., ANSA, O., MARTINEZ, D., HOVY, E., 2001, Enriching Wordnet Concepts With Topic Signatures, *Proceedings Of The NAACL Workshop On Wordnet And Other Lexical Resources: Applications, Extensions And Customizations*, Pittsburg, USA, 123-132.

AHA, D.W., GOLDSTONE, R.L., 1992, Concept Learning And Flexible Weighting, *In Proceedings Of The Fourteenth Annual Conference Of The Cognitive Science Society*, Illinois: Lawrence Erlbaum, 534-539.

AKIN, H.L., KURU, S., GÜNGÖR, T., HAMZAOGLU, I., ARBATLI, D., 1993, A Spelling Checker and Corrector for Turkish, *International Second Turkish Symposium on Artificial Intelligence and Artificial Neural Networks (TAINN 1993)*, Istanbul, 113-120.

ALTAN, Z., ORHAN Z., 2003, Disambiguation Of Turkish Word Senses By Supervised Statistical Methods, International XII. Turkish Symposium On Artificial Intelligence And Neural Networks - TAINN 2003, Çanakkale,Turkey, *International Journal Of Computational Intelligence*, Volume: 1, Number :1, 16-21, July 2003.

ARNOLD, D., AARTS, B., BUCKLEY, J., BERGLUND, Y., NELSON G., RONDELL, M., 1999, *Corpora And Grammars On The Web: The W3Corpora/IGE Project*, Final Report, JTAP-2/247.

ATALAY, N. B., OFLAZER, K., SAY, B., 2003, The Annotation Process In The Turkish Treebank, *In Proceedings Of The EACL Workshop On Linguistically Interpreted Corpora-LINC*, April 13-14, Budapest, 100-110.

AYTEKIN, Ç., SAY, A. C. C., AKÇOK, E., 1994, ELIZA speaks Turkish: A conversation program for an agglutinative language, *Üçüncü Türk Yapay Zeka ve Yapay Sinir Ağları Sempozyumu*, Ankara, 435.

BANKO, M., BRILL, E., 2001, Scaling To Very Very Large Corpora For Natural Language Disambiguation, *In Proceedings Of The 39th Annual Meeting Of The*

Association For Computational Linguistics, Association For Computational Linguistics, 26–33.

BERGER, A., BROWN, P., PIETRA, S. D., PIETRA, V. D., LAFFERTY, J., PRINTZ, H., URES, L., 1994, The Candide system for machine translation. *In Proceedings of the ARPA Conference on Human Language Technology*, 89-103.

BERKER, İ., SAY, A. C. C., 1993, A crossword puzzle generator for Turkish, *Proceedings of the Eighth International Symposium on Computer and Information Sciences (ISCIS-VIII)*, İstanbul, 474-477.

BİLGİN, O., ÇETİNOĞLU, Ö., OFLAZER, K., 2004, Building A Wordnet For Turkish, *Romanian Journal Of Information Science And Technology*, Volume 7, Numbers 1–2, 163–172

BIRD, M., 1996, System Overload. Excess Information Is Clogging The Pipes Of Commerce - And Making People Ill, In *Time Magazine*, December 9th, 1996, 46-47.

BLACK, E., 1988, An Experiment In Computational Discrimination Of English Word Senses., *BM Journal Of Research And Development*, 32(2), 185-194.

BOZŞAHİN, C., 2002, The Combinatory Morphemic Lexicon, *Computational Linguistics*, 28, 145-186.

BREIMAN, L., FRIEDMAN, L., OLSHEN, R., STONE, C., 1984, *Classification And Regression Trees*, Wadsworth Inc., Belmont, California.

BROWN, P.F., DELLA PIETRA, S., DELLA PIETRA, V., MERCER, R.L., 1991, Word Sense Disambiguation Using Statistical Methods. *In Proceedings Of The 29th Annual Meeting Of The Association For Computational Linguistics (ACL)*, 264-270.

BRUCE, R., WIEBE., J., 1994, Word-Sense Disambiguation Using Decomposable Models, *In Proceedings Of The 32nd Annual Meeting Of The Association For Computational Linguistics*, 139-146.

BRUCE, R., WIEBE, J., 1999. Decomposable Modeling In Natural Language Processing, *Computational Linguistics*, 25(2):195-207.

CARDIE, C., 1993, A Case-Based Approach To Knowledge Acquisition For Domain-Specific Sentence Analysis, *In Proceedings Of The Eleventh National Conference On Artificial Intelligence*, Washington, D.C., 798-803.

CEBIROGLU, G., 2002, Sözlüksüz Köke Ulaşma Yöntemi, Yüksek Lisans Tezi, I.T.Ü. Fen Bilimleri Enstitüsü, İstanbul.

CHOMSKY, N., 1957, *Syntactic Structures*, Mouton, The Hague.

CLEARY, J. G., TRIGG, L. E., 1995, K*: An Instance-Based Learner Using An Entropic Distance Measure, *Proceedings Of The 12th International Conference On Machine Learning*, 108-114.

COLLINS, A. M., LOFTUS, E. F., 1975, A Spreading Activation Theory Of Semantic Processing, *Psychological Review*, 82(6), 407-428.

COTTRELL, G. W., 1989, A Connectionist Approach To Word Sense Disambiguation, *Research Notes In Artificial Intelligence*, London: Pitman.

COVER, T.M., HART, P.E., 1967, Nearest Neighbor Pattern Classifcation, *IEEE Transactions On Information Theory*, 13:21-27.

DAELEMAN, W., HOSTE, V., HENDRICKX, I., BOSCH, A.V.D., 2002 Parameter Optimization For Machine-Learning Of Word Sense Disambiguation. Natural Language Engineering, Pages 311–325.

DAELEMAN, W., HOSTE, V., MEULDER, F. D., NAUDTS, B., 2003, Combined Optimization Of Feature Selection And Algorithm Parameters in Machine Learning Of Language, *Proceedings Of The 14th European Conference On Machine Learning (ECML-2003), Lecture Notes in Computer Science 2837*, Springer-Verlag, Cavtat-Dubrovnik, Croatia, 84-95.

DAGAN, I., KAROV, Y., ROTH, D., 1997, Mistake-Driven Learning in Text Categorization, *In Proceedings Of The 2nd Conference On Empirical Methods in Natural Language Processing (EMNLP,*. Brown University, Providence, Rhode Island, 78-90.

DAHLGREN, K. G., 1988, *Naive Semantics For Natural Language Understanding.* Kluwer Academic Publishers, Boston.

DARCAN, O. N., 1991, *An Intelligent Database Interface for Turkish*, Yüksek Lisans Tezi, B.Ü. Fen Bilimleri Enstitüsü, Istanbul.

DEMİR, Ş., 2003, *Improved Treatment of Word Meaning in a Turkish Conversational Agent*, Yüksek Lisans Tezi, B.Ü. Fen Bilimleri Enstitüsü, Istanbul.

DIETTERICH, T. G., 1998, Approximate Statistical Tests For Comparing Supervised Classification Learning Algorithms, *Neural Computation*, 10(7):1895–1923.

DOMINGOS, P., PAZZANI, 1997, On The Optimality Of The Simple Bayesian Classifier Under Zero-One Loss, *Machine Learning*.

EDMONDS, P., KILGARRIFF, A., 2003, Editors. *Journal Of Natural Language Engineering Special Issue Based On Senseval-2*, Volume 9. Cambridge University Press, 2003.

EDMONDS, P., 2002, SENSEVAL: The evaluation of word sense disambiguation systems, *ELRA Newsletter*, Vol. 7 No. 3, 5-14.

ESCUDERO, G., MARQUEZ, L., RIGAU, G., 2000, Naive Bayes And Exemplar-Based Approaches To Word Sense Disambiguation Revisited, *In Proceedings Of The 14th European Conference On Artificial Intelligence*, ECAL, 58-69.

FELLBAUM, C., 1998, *WordNet: An Electronic Lexical Database*, The MIT Press.

FELLBAUM, C., GRABOWSKI, J., LANDES, S, 1998, Performance and confidence in a semantic annotation task, In Fellbaum, C. (ed.) (1998) *WordNet: An Electronic Lexical Database. Cambridge (Mass.)*, The MIT Press.

FREUND, Y., SCHAPIRE, R.E., 1996, Experiments With A New Boosting Algorithm, *Proc. International Conference On Machine Learning*, Morgan-Kaufmann, San Francisco, 148-156.

FRIEDMAN, N., GEIGER, D., GOLDSZMIDT, M., 1997, Bayesian Network Classifiers, *Machine Learning*, 29 (2), 131–163.

GALE, W., CHURCH, K., YAROWSKY, D., 1992, Work On Statistical Methods For Word Sense Disambiguation, *In Proceedings AAAI Fall Symposium On Probabilistic Approaches To Natural Language*, Cambridge, MA, 54-60.

GÜNGÖR, T, 2004, Generation of Sentence Parse Trees Using Parts of Speech, *Lecture Notes in Artificial Intelligence*, Vol.3238, 56-66, Springer-Verlag, Berlin Heidelberg.

GÜNGÖR, T., 2003, *Türkçe'nin Istatistiksel Incelenmesi*, Technical Report 02A107, Bogaziçi University Research Fund, Bogaziçi University, Istanbul.

GÜNGÖR, T., KURU, S., 1993, Representation of Turkish Morphology in ATN, *International Second Turkish Symposium on Artificial Intelligence and Artificial Neural Networks (TAINN 1993)*, Istanbul, 94-104.

HANKAMER, J., 1986, Finite State Morphology And Left-To-Right Morphology, *West Coast Conference On Formal Linguistics*.

HART, P.E., 1968, The Condensed Nearest Neighbour Rule, *IEEE Transactions On Information Theory*, 14, 515-516

HARUNO, M., SHIRAI, S., OOYAMA, Y., 1998, Using Decision Trees To Contruct A Practical Parser, *In Proceedings Of The Joint 17th International Conference On Computational Linguistics And 36th Annual Meeting Of The Association For Computational Linguistics (COLING-ACL)*, Montreal, Canada, 1136-1142

HEARST, M. A., 1994, Multiparagraph Segmentation Of Expository Text, *Proceedings Of The 32nd Annual Meeting Of The Association For Computational Linguistics*, Las Cruces, New Mexico, 9-16.

HIRSCHMAN, L, 1998, The evolution of evaluation: Lessons from the Message Understanding Conferences, *Computer Speech and Language*, 12, 281-305.

HIRST, G., 1987, *Semantic Interpretation And The Disambiguation Of Ambiguity*, Cambridge University Press, England

HUGHES, JOHN, 1994, *Automatically Acquiring A Classifcation Of Words*, Ph.D. Thesis, School Of Computing, University Of Leeds.

HUTCHINS, J., SOMMERS, H., 1992, *Introduction To Machine Translation*, Academic Press.

IBRAHIMOV, O., SETHI, I., DIMITROVA, N., 2001, Clustering Of Imperfect Transcripts Using A Novel Similarity Measure, *In Proceedings Of The SIGIR'01 Workshop On Information Retrieval Techniques For Speech Applications*.

IDE, N., VERONIS, J., 1998, Introduction To The Special Issue On Word Sense Disambiguation: The State Of The Art, *Computational Linguistics*, 24(1), 1-40

KARDEŞ, O., 2002, *Bir Uygulama Alanında Türkçe Metnin Anlambilimsel Gösterimi*, Yüksek Lisans Tezi, B.Ü. Fen Bilimleri Enstitüsü, Istanbul.

KAZAKOV, D., 1996, *Natural Language Processing Applications Of Machine Learning*, Ph.D. Thesis, Czech Technical University, Prague.

KEÇECİ, H., 1996, *Bir robot koluna kumanda eden dogal dil anlama sistemi*, Yüksek Lisans Tezi, I.T.Ü. Fen Bilimleri Enstitüsü, Istanbul.

KELLY, E., STONE, P., 1975, Computer Recognition of English Word Senses, North Holland, Amsterdam.

KİBAROĞLU, O., 1991, *Spell Checking in Agglutinative Languages and an Implementation for Turkish*, Yüksek Lisans Tezi, B.Ü. Fen Bilimleri Enstitüsü, Istanbul.

KILGARRIFF, A., 1999, I Don't Believe In Word Senses, *Computers And The Humanities. B. (Eds.)*, English Corpus Linguistics, Longman, London, 8-29.

KILGARRIFF, A., PALMER, M., 1999, *Computers And The Humanities Special Issue Based On Senseval-1*, Volume 34.

KILGARRIFF, A., ROSENZWEIG, J., 2000, Framework and results for English SENSEVAL, *Computers and the Humanities*, 34 (1-2) Special Issue on SENSEVAL, 15-48.

KOHAVI, R., JOHN, G., H., 1997, Wrappers For Feature Subset Selection. *Artificial Intelligence*, 97(1–2):273–323.

KÖKSAL, A., 1976, *Türkçe'nin Özdevimli Biçimbirim Çözümlemesi*, Doktora Tezi, Hacettepe Üniversitesi, Ankara.

LEACOCK, C., TOWELL, G., VOORHEES, E., 1993, Corpus-Based Statistical Sense Resolution, *In Proceedings Of The ARPA Workshop On Human Language Technology*, 260–265.

LEE, Y. K., NG, H. T., 2002, An Empirical Evaluation Of Knowledge Sources And Learning Algorithms For Word Sense Disambiguation, *Proceedings Of The 2002 Conference On Empirical Methods In Natural Language Processing (EMNLP-2002)*, 41-48.

LESK, M., 1986, Automated Sense Disambiguation Using Machine-Readable Dictionaries: How To Tell A Pine Cone From An Ice Cream Cone, *Proceedings Of The 1986 SIGDOC Conference*, Toronto, Canada, June 1986, 24-26.

LEVIS, D., SCHAPIRE, R.E., CALLAN, J.P., PAPKA, R., 1996, Training Algorithms For Linear Text Classiffiers. *In Proceedings Of The 19th International Conference On Research And Development İn Information Retrieval, SIGIR*, Zurich, Switzerland, 298-306.

MARQUEZ, L., 1999, *Part-Of-Speech Tagging: A Machine Learning Approach Based On Decision Trees*, Ph.D. Thesis, Dep. Llenguatges I Sistemes Informatics, Universitat Politecnica De Catalunya.

MARQUEZ, M., TAULE, L., PADRO, VILLAREJO, L., MARTI, M.A., 2004 On The Quality Of Lexical Resources For Word Sense Disambiguation, *In Proceedings of the EsTAL Conference*, Alicante, Spain, 291-303.

MASTERMAN, M., 1961, Semantic Message Detection For Machine Translation, Using An Interlingua, *International Conference On Machine Translation Of Languages And Applied Language Analysis*, London, 437-475.

MCCLELLAND, J., RUMELHART, L., DAVID E., 1981, An Interactive Activation Of Context Effects In Letter Perception: Part 1. An Account of Basic Findings, P*sychological Review*, 88, 375-407.

MCENERY, T., WILSON, A., 1996, *Corpus Linguistics*, Edinburgh University Press.

MIHALCEA, R., MOLDOVAN, D., 1998, Word Sense Disambiguation Based On Semantic Density, *Proceedings of COLING-ACL Workshop On Usage Of Wordnet In Natural Language Processing Systems* , Montreal, Canada, 123-130.

MIHALCEA, R., 2002, Instance Based Learning With Automatic Feature Selection Applied To Word Sense Disambiguation, *In Proceedings Of The 19th International Conference On Computational Linguistics (COLING 2002)*, Taiwan, 202-214.

MUÒNOZ, M., PUNYAKANOK, V., ROTH, D., ZIMAK, D., 1999, A Learning Approach To Shallow Parsing, *In Proceedings Of The Joint SIGDAT Conference On Emperical Methods In Natural Language Processing And Very Large Corpora (EMNLPVLC)*,37-45.

MOONEY, R. J., 1996, Comparative Experiments On Disambiguating Word Senses: An Illustration Of The Role Of Bias In Machine Learning, In Eric Brill, Kenneth Church, Editors, *Proceedings Of The Conference On Empirical Methods In Natural Language Processing*, Association For Computational Linguistics, Somerset, New Jersey, 82–91.

NG, H.T., LEE, H.B., 1996, Integrating Multiple Knowledge Sources To Disambiguate Word Sense: An Exemplar-Based Approach. *In Proceedings Of The 34th Annual Meeting Of The Association For Computational Linguistics (ACL-96)*, Santa Cruz, In Arivind Joshi ve Martha Palmer, Editors, *Proceedings Of The Thirty-Fourth Annual Meeting Of The Association For Computational Linguistics*, San Francisco, 1996, Morgan Kaufmann Publishers, 40–47

NG, H. T., ZELLE, J., WINTER, K. 1997, Corpus-Based Approaches to Semantic Interpretation in Natural Language Processing, *AI Magazine*, 45-64.

NG, H. T., 1997, Exemplar-Based Word Sense Disambiguation: Some Recent Improvements, *In Procs. Of The 2nd Conference On Empirical Methods In Natural Language Processing*, EMNLP, 37-49.

OFLAZER, K., 1993, Two-Level Specification of Turkish Morphology, *European Chapter Of The Association Of Computational Linguistics*, Utrecht, Hollanda, 45-56.

OFLAZER, K., SAY, B., TUR, D. Z. H., TUR, G., 2003, Building A Turkish Treebank, *Invited Chapter In Building And Exploiting Syntactically-Annotated Corpora*, Anne Abeille Editor, Kluwer Academic Publishers, 2003.

ORHAN, Z., ALTAN, Z., 2004, Makine Öğrenme Algoritmalarıyla Türkçe Sözcük Anlamı Açıklaştırma, ELECO'2004, Bursa/Türkiye, Dec. 2004, *Elektrik-Elektronik-Bilgisayar Mühendisliği Sempozyumu Bildiriler Kitabı*, 344-348, *MakinaTek*, No. 92, Jun. 2005, 100-104.

ORHAN, Z., ALTAN, Z., 2005, Determining Effective Features for Word Sense Disambiguation in Turkish, *Istanbul University - Journal of Electrical & Electronics Engineering (IU - JEEE)*, Vol. 5, No. 2, Jul. 2005, 1341-1352.

ORHAN, Z., ALTAN, Z., 2006, Determining Effective Features for Word Sense Disambiguation in Turkish, Impact of Feature Selection for Corpus-Based WSD in Turkish, *Lecture Notes in Artificial Intelligence, LNAI*, MICAI 2006, Mexico City, Mexica

ÖĞÜN, F., 2003, *Design and Implementation of an Improved Conversational Agent Infrastructure for Turkish,* Yüksek Lisans Tezi, B.Ü. Fen Bilimleri Enstitüsü, Istanbul.

ÖZGÜR, A., 2004, *Belge Sınıflandırma İçin Denetimli ve Denetimsiz Öğrenme Algoritmaları,* Yüksek Lisans Tezi, B.Ü. Fen Bilimleri Enstitüsü, Istanbul.

ÖZGÜR, L., 2003, *Türkçe Morfolojik Çözümleme, Yapay Sinir Ağları ve Bayes Filtreleme Tabanlı Uyarlamalı Spam-Önler Filtrelemesi*, Yüksek Lisans Tezi, B.Ü. Fen Bilimleri Enstitüsü, Istanbul.

PEDERSEN, T., BRUCE, R., 1997, A New Supervised Learning Algorithm For Word Sense Disambiguation, *In Proceedings Of The 14th National Conference On Artificial Intelligence (AAAI-97),* Providence, RI, 254-267.

PEDERSEN, T., 2001, A Decision Tree Of Bigrams Is An Accurate Predictor Of Word Sense, *In Proceedings Of The North American Chapter Of The Association For Computational Linguistics*, NAACL 2001, Pittsburg, 79-86.

PEMBE, F.C., 2004, *A Linguistically Motivated Information Retrieval System for Turkish*, Yüksek Lisans Tezi, B.Ü. Fen Bilimleri Enstitüsü, Istanbul.

QUILLIAN, M. R., 1961, A Design For An Understanding Machine Communication, *The Colloquium Semantic Problems In Natural Language, King's College*, Cambridge University, Cambridge, United Kingdom

QUINLAN, J. R., 1986, Induction of Decision Trees, *Machine Learning*, Vol.1, No.1, 81-106

QUINLAN, J.R., 1993, *C4.5: Programs For Machine Learning*, Morgan Kaufmann.

RESNIK, P., YAROWSKY, D., 1997, Evaluating Automatic Semantic Taggers, *ACL-SIGLEX Workshop "Tagging Text With Lexical Semantics: Why, What, And How?"*, April 4-5, 1997, Washington, D.C., 91.

ROTH, D., ZELENKO, D., 1998, Part Of Speech Tagging Using A Network Of Linear Separators, *In Proceedings Of The Joint 17th International Conference On Computational Linguistics And 36th Annual Meeting Of The Association For Computational Linguistics (COLING-ACL),* Montreal, Canada, 1136-1142.

SAGAY, Z., 1981, *A Computer Translation From English To Turkish*, Yüksek Lisans Tezi, Bilgisayar Mühendisliği Bölümü, ODTÜ, Ankara.

SAK, H., 2004, *Türkçe için Korpus Tabanlı Birleştirmeli Konuşma Sentezleme Sistemi*, Yüksek Lisans Tezi, B.Ü. Fen Bilimleri Enstitüsü, Istanbul.

SAY, A. C. C., 2001, Understanding arithmetic problems in Turkish, *International Journal of Pattern Recognition and Artificial Intelligence,* Vol. 15, 359-374.

SAY, A. C. C., SEN, S., BARENGI, R., 1993, Bir Kelime - Bir İşlem oynayan program, (in Turkish) *İkinci Türk Yapay Zeka ve Yapay Sinir Ağları Sempozyumu*, İstanbul, 349-355.

SCHAPIRE, R. E., SINGER, Y., 2000., BoosTexter: A Boosting-based System for Text Categorization, *Machine Learning*, 29(3/4):135–168

SCHÜTZE, H., PEDERSEN, J., 1995, Information Retrieval Based on Word Senses, *Proceedings of SDAIR'95*, April 1995, Las Vegas, Nevada, 98-110.

SEVEN, A., 1997, *Small Vocabulary Word and Speaker Recognition Using Artificial Neural Networks*, Yüksek Lisans Tezi, B.Ü. Fen Bilimleri Enstitüsü, Istanbul.

STAMOU, S., OFLAZER, K., PALA, K., CHRISTODOULAKIS, D., CRISTEA, D., TUFIS, D., KOEVA, S., TOTKOV, G., DUTOIT, D., GRIGORIADOU, M., 2002, Balkanet: A Multilingual Semantic Network For Balkan Languages, *In Proceedings Of The First International Wordnet Conference*, January 2002, Mysore India, 56-67.

STOOP, A., 1987, TRANSIT In The World Of Machine Translation: Towards An Automatic Translator For Dutch And Turkish, *3. Türkçe Dilbilimi Konferansı Bildiri Kitabi*, Tilburg, Hollanda, 78-85.

ŞENKAL, M., 2003, *Bir İspanyolca Türkçe Otomatik Çeviri Sistemi Yaklaşımı*, Yüksek Lisans Tezi, B.Ü. Fen Bilimleri Enstitüsü, Istanbul.

TANAKA, H., 1996, Decision Tree Learning Algorithm With Structural Attributes: Application To Eylemal Case Frame Acquisition, *In Proceedings Of The 16th International Conference On Computational Linguistics (COLING)*. Copenhagen, Denmark, 943-948.

USZKOREIT, H., 2000, Language Technology for Knowledge Management, Proceedings of *Japanese-German Workshop Computational Linguistics*, Yokohama, 26 May 2000, 1-10.

VEENSTRA, A. VAN DEN BOSCH, BUCHHOLZ, J., S., DAELEMANS, W., ZAVREL, J., 2000, Memory-Based Word Sense Disambiguation, *Computers And The Humanities*, 34:171-177.

WANG, Z., WEBB, G. I., 2002, Comparison Of Lazy Bayesian Rule And Tree-Augmented Bayesian Learning, *In Proceedings Of The IEEE International Conference On Data Mining*, ICDM-2002, Maebashi, Japan, 775–778.

WEAVER, W., 1949, *Translation*, Mimeographed, 12 Pp., July 15, 1949. Reprinted In Locke, William N. Ve Booth, A. Donald, 1955 (Eds.), *Machine Translation Of Languages*, John Wiley & Sons, New York, 15-23.

WEBB, G., BOUGHTON, J., WANG, Z., 2002, Averaged One-Dependence Estimators: Preliminary Results, *AI2002 Data Mining Workshop*, Canberra, 22-33.

WEISS, S., 1973, Learning To Disambiguate, *Information Storage And Retrieval*, 9.

WEISS, S., INDURKHYA, N., 1998, *Predictive Data Mining: A Practical Guide*, *Morgan Kaufmann*, San Francisco.

WEISS, S.M., APTE, C., DAMERAU, F.J., JOHNSON, D.E., OLES, F.J., GOETZ, T., HAMPP, T., 1999, Maximizing Text-Mining Performance, *IEEE Intelligent Systems*, 14(4), 63-69.

WEIZENBAUM, J., 1966, ELIZA - A Computer Program for The Study Of Natural Language Communication Between Man And Machine, *Commun. ACM*, 9(1), 36-45

WILKS, Y. A., FASS, D.,1990, *Preference Semantics: A Family History*, Report MCCS-90-194, Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico.

WILKS, Y., 1992, Systran: It obviously works but how much can it be improved? In John Newton, editor, *Computers in Translation: A Practical Appraisal*, Routledge, London, 166-188.

WITTEN, I.H., FRANK E., 1999, *DataMining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, San Francisco.

YAROWSKY, D., 1992, Word-Sense Disambiguation Using Statistical Models Of Roget's Categories Trained On Large Corpora, *In Proceedings, COLING-92*, Nantes, 454-460.

YAROWSKY, D., 1993, One Sense Per Collocation, *In Proceedings Of The ARPA Human-Language Technology Workshop*, Washington, D.C.: Advanced Research Projects Agency, 266-271.

YAROWSKY, D., 1995, Word Sense Disambiguation Rivaling Supervised Methods, *Proceedings Of COLING*, Cambridge, USA, 23-34.

YAROWSKY, D., 2000, Hierarchical Decision Lists For Word Sense Disambiguation, *Computers And The Humanities*, 34(2):179-186.

YÜKSEL, Ö., BOZŞAHIN, C., 2002, Contextually Appropriate Reference Generation, *Natural Language Engineering*, 8, 69-89.

ZIPF, G. K., 1949, *Human Behaviour and the Principle of Least-Effort*, Addison-Wesley, Cambridge MA.

EKLER

EK-A SÖZCÜK SIKLIKLARI

Metinlerde 25 ve daha fazla geçen sözcüklerin geçiş sıklıkları

GS	Sözcük	GS	Sözcük	GS	Sözcük	GS	Sözcük	GS	Sözcük
12791	Punc	95	çocuk	53	tarih	40	küçük	30	parti
3961	Özel ad	95	mi	52	başkan	40	madde	30	şekil
1153	num	92	mı	52	ise	40	son	30	tartış
1134	bir	92	yok	52	öl	40	yeniden	30	taşı
1128	ol	89	kız	52	sür	40	yürü	30	vergi
658	ve	88	bir	51	fark	39	gerçek	29	doğ
543	de	88	ses	51	orta	39	incele	29	faiz
529	bu	88	yol	50	ay	39	para	29	hava
481	de	87	baba	50	biri	39	sabah	29	kafa
454	da	87	yeni	50	ilk	39	taraf	29	korku
426	et	84	adam	50	kapa	39	ye	29	lira
354	yap	84	sor	50	uzun	38	ayır	29	mektup
298	gel	83	dur	50	zor	38	geyik	29	pilot
265	al	82	gece	49	saat	38	kaza	29	seçim
265	al	79	bilim	48	alan	38	oda	28	acı
255	bu	79	çek	48	belki	38	oluş	28	arkadaş
253	gibi	79	kapı	48	hak	38	öyle	28	aslında
250	için	78	getir	47	san	38	uza	28	bağımlı
247	ver	77	hiç	47	yarat	38	yok	28	dinle
238	çık	77	uçak	46	ad	37	doğru	28	eroin
214	daha	75	dön	46	birbiri	37	ilişki	28	hafta
213	o	75	gerek	46	bugün	37	kabul	28	ilişkin
212	ama	75	önce	46	genel	37	koş	28	köpek
198	ne	75	sev	46	görün	37	şu	28	makine
197	git	74	anla	46	kalk	36	art	28	say
197	kendi	73	ora	46	koy	36	güç	28	sokak
193	yer	72	aynı	46	nere	36	iç	28	üre
191	şey	71	başka	46	oğul	36	kardeş	28	yardım
188	bil	71	konu	46	önem	36	kim	27	belli
185	bak	70	üst	46	süre	36	yat	27	çal
179	diye	68	an	46	ulaş	35	aile	27	değişik
179	ev	68	düş	45	akıl	35	bilimsel	27	dün
168	iste	68	göster	45	biraz	35	birkaç	27	kes
158	gün	68	kişi	45	kullan	35	dikkat	27	kurtul
157	el	68	ön	45	su	35	hepsi	27	özel

155	iç	67	bit	44	haber	35	herkes	27	rakı
155	söyle	67	bura	44	hiçbir	35	sat	27	süreç
150	her	67	dünya	44	ifade	34	bir	27	terk
147	çok	67	erkek	44	olay	34	dol	27	üzere
146	geç	67	hayat	44	önce	34	isim	27	vur
146	zaman	67	söz	44	sağla	34	kez	27	yasa
143	gör	67	yaz	44	tüm	34	park	26	askeri
139	kadar	64	mu	44	ülke	34	sorun	26	genç
137	yıl	62	güzel	43	ancak	34	tanı	26	heyecan
136	ara	62	hal	43	bilgi	33	aşk	26	kavram
134	gir	62	üzer	43	değiş	33	düzenle	26	kazan
133	değil	61	bırak	43	hem	33	gerek	26	kolay
131	başla	61	şimdi	43	hep	33	hız	26	merak
131	ki	61	yüz	43	ilgi	33	inan	26	sanki
129	anlat	60	nasıl	43	işte	33	kahve	26	şöyle
126	en	59	bekle	43	rapor	33	şirket	26	toplum
117	insan	59	masa	42	arka	33	soru	25	abla
114	ile	58	artık	42	ayak	33	uzak	25	açık
113	kal	58	çok	42	can	33	yani	25	açıkla
112	kadın	58	durum	42	çünkü	32	ağız	25	belirle
111	göz	57	böyle	42	dönem	32	geliş	25	bey
107	sonra	57	oku	42	eski	32	kork	25	beyaz
105	iş	57	sonra	42	evet	32	okul	25	dil
102	baş	56	at	42	kur	31	bazı	25	düşünce
101	çalış	56	neden	42	yemek	31	diğer	25	hazır
101	iyi	56	tut	41	alt	31	dosya	25	hisset
101	konuş	55	ara	41	değil	31	eş	25	ihtiyaç
101	var	55	belir	41	gerçek	31	in	25	ışık
101	ya	55	bütün	41	görev	31	saç	25	kadar
101	yaşa	55	doğru	41	kimse	31	sahip	25	karşı
100	var	55	yaş	41	peynir	31	türk	25	paket
98	bul	54	bile	41	tek	31	yapı	25	sadece
98	düşün	54	göre	41	yaşam	30	akşam	25	sınıf
97	anne	54	karar	41	yine	30	hareket	25	şu
96	aç	54	sıra	40	bakan	30	hemen	25	veya
96	otur	54	son	40	bulun	30	izle		
96	yan	53	duy	40	hazırla	30	kaldır		
95	büyük	53	duygu	40	içeri	30	öğren		

ÖZGEÇMİŞ

Zeynep Orhan, 15.09.1974 tarihinde Konya'da doğdu. İlkokulu, Namık Kemal İlkokul'unda bitirdi. Ortaokulu ve liseyi Konya Meram Anadolu Lisesi'nde tamamladı. 1992 yılında Bilkent Üniversitesi Bilgisayar ve Enformatik Mühendisliği Bölümüne burslu olarak girdi ve lisans derecesini 1996 yılında aldı. 1996 yılında Bilkent Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda Yüksek Lisans eğitimine ve asistanlığa başladı. 1998 yılında Yüksek Lisansını tamamladı. Ekim 2001'de İstanbul Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda başladığı doktora çalışmalarını halen sürdürmektedir.