

EGE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

(DOKTORA TEZİ)

**TÜRKÇE İÇİN
TÜMLEŞİK BİR BİÇİMBİRİM ÇÖZÜMLEME
VE
SÖZCÜK TÜRÜ TESPİT YÖNTEMİ
Tarık KIŞLA**

Uluslararası Bilgisayar Anabilim Dalı

Bilim Dalı Kodu: 619.02.04

Sunuş Tarihi: 25.06.2009

Tez Danışmanı: Prof. Dr. Bahar KARAOĞLAN

BORNOVA - İZMİR

III

Tarık KIŞLA tarafından Doktora tezi olarak sunulan “Türkçe İçin Tümeleşik Bir Biçimbirim Çözümleme ve Sözcük Türü Tespit Yöntemi” başlıklı bu çalışma E.Ü. Lisansüstü Eğitim ve Öğretim Yönetmeliği ile E.Ü. Fen Bilimleri Enstitüsü Eğitim ve Öğretim Yönergesi’nin ilgili hükümleri uyarınca tarafımızdan değerlendirilerek savunmaya değer bulunmuş ve 25.06.2009 tarihinde yapılan tez savunma sınavında aday oybirliği ile başarılı bulunmuştur.

-

Jüri Üyeleri:

İmza

Jüri Başkanı	:		
Raportör Üye	:		
Üye	:		
Üye	:		
Üye	:		

ÖZET

TÜRKÇE İÇİN TÜMLEŞİK BİR BİÇİMBİRİM ÇÖZÜMLEME VE SÖZCÜK TÜRÜ TESPİT YÖNTEMİ

KIŞLA, Tarık

Doktora Tezi, Uluslararası Bilgisayar Enstitüsü

Tez Yöneticisi: Prof. Dr. Bahar KARAOĞLAN

Haziran 2009, 185 sayfa

Dillerin sayısal ortamlarda etkili bir şekilde temsilinin ve işlenmesinin sosyal, politik, kültürel ve bilimsel amaçlarla dilleri kullanan insanların desteklemesinde ve dillerin yaşamlarını sürdürmesinde oldukça önemli bir faktör olacağı şüphesizdir. Bu nedenle, çalışmamızın ana motivasyonu, Türkçe'nin sayısal ortamlarda temsilini ve işlenmesini güçlendirmek için doğal dil işleme uygulamalarına hizmet edebilecek temel yöntemler geliştirmektir.

Çalışmamızda, doğal dil işleme sahasının ana problemlerinden olan biçimbirimsel çözümleme ve sözcük türü tespiti problemleri için özgün çözümler sunulmuştur. Eklemeli diller için karmaşıklığı teorik olarak NP-complete olarak bilinen biçimbirimsel çözümleme için, Türkçe'nin dilbilgisel özellikleri göz önünde tutularak basitleştirilmiş bir yöntem ortaya konmuştur. İstatistiksel ve kural tabanlı yaklaşımları bir arada kullanan yöntemimiz çözümleme sonucunda tek bir sonuç sunması ile belirsizliğin giderilmesini de sağlamaktadır. Buna ek olarak, önerilen yöntemin kapalı ve sınırlı bir sözlük kullanması, diğer yöntemlerden farklılaşan önemli bir özelliğidir. Tez kapsamında geliştirilen yöntemler deneysel çalışmalar ile test edilmiş ve elde edilen sonuçların önceki çalışmalardan daha yüksek bir başarı yüzdesine sahip olduğu gösterilmiştir.

Anahtar Kelimeler: Doğal dil işleme, sözcük türü tespiti, biçimbirimsel çözümleme, Gövdeleme

ABSTRACT**AN INTEGRATED METHOD FOR MORPHOLOGICAL ANALYSE
AND PART OF SPEECH TAGGING IN TURKISH**

KIŞLA, Tarık

Phd. in International Computer Institute

Supervisor: Prof. Dr. Bahar KARAOĞLAN

June 2009, 185 pages

There is no doubt that the languages which are effectively represented and processed digitally will sustain their existence more firmly than the others and strengthen the people who are using them in social, political, cultural and scientific platforms. The major motivation behind this study is to develop basic methods that will serve and enhance the natural language processing tasks for Turkish.

Here, we present original solutions for the morphological analysis and part of speech tagging for Turkish which are fundamental tasks for natural language processing applications. The morphological analysis which has NP-complete complexity for agglutinative languages is simplified in methods for rooting, stemming and finding derivational and inflectional boundaries by taking into account some dependencies and features of the Turkish language. The methods employ both statistical and rule based approaches, and present a single result for the analysis avoiding the disambiguation problem. The reliability of the methods are high due to the fact that a closed and restricted vocabulary is used and unsuccessful analysis resulting from not being in the vocabulary are thwarted. The empirical tests of the methods on Turkish corpora have shown higher success ratios than the current methods that we have met in the literature.

Keywords: Natural language processing, part of speech tagging, morphological analysis, stemming

TEŞEKKÜR

İlk olarak tez çalışmam sırasında deneyimlerini, fikir ve önerilerini bana aktaran, bitmek bilmeyen çalışma arzusu ile beni sürekli destekleyen, bana bir akademisyenin nasıl olması gerektiğini gösteren çok değerli tez danışmanım Prof. Dr. Bahar KARAOĞLAN'a saygı ve sevgilerimi sunarak teşekkürü bir borç bilirim.

Tez izleme jürimde yer alarak değerli önerileri ve katkılarını benden esirgemeyen Prof. Dr. Aydın ÖZTÜRK ve Yrd. Doç. Dr. Adil ALPKOÇAK'a teşekkür ederim.

Tez konusunun ortaya çıkışında fikirleri ile bana yön veren değerli hocam Yrd. Doç. Dr. Taner DİNÇER'e katkıları için teşekkürlerimi sunarım.

Türkçe konusundaki sorularıma sabırla cevap vererek bu konudaki gelişimime katkıda bulunan değerli arkadaşım Yrd. Doç. Dr. Soner AKŞEHİRLİ'ye teşekkürü bir borç bilirim.

Ayrıca tez kapsamındaki deneylerim sırasında yardımını esirgemeyen değerli hocam Hayri SEVER'e katkılarından dolayı teşekkür ederim.

Bana çalışmalarım sırasında manevi olarak destek veren Ege Üniversitesi Eğitim Fakültesinde görevli hocalarım ve mesai arkadaşlarımdan hepsine tek tek teşekkürlerimi sunarım. Ayrıca Ege Üniversitesi Uluslararası Bilgisayar Enstitüsünde görevli değerli hocalarım, araştırma görevlisi arkadaşlarım ve idari personele teşekkür ederim.

Ayrıca, beni dünyaya getiren ve her zaman beni destekleyen sevgili anne ve babama, yaşıntım boyunca sürekli yanımda olan kardeşime

X

teşekkürler ederim. Beni kendi evlatları gibi gören ve seven kayınvalidem Ayfer HUYUGÜZEL, kayınpederim Prof. Dr. Ömer Faruk HUYUGÜZEL ve ablam Pınar ÇAVAŞ'a verdikleri destekten dolayı teşekkürü bir borç bilirim.

Tezim boyunca her türlü sıkıntımı paylaşan, desteğini benden esirgemeyen biricik eşim Gül Şerife HUYUGÜZEL KİŞLA'ya sonsuza kadar beraber olmak dileğiyle teşekkürlerimi sunarım.

İÇİNDEKİLER

Sayfa

ÖZET	V
ABSTRACT	VII
TEŞEKKÜR	IX
1. GİRİŞ	1
2. TÜRKÇE.....	9
2.1. ALFABE (ABECE)	11
2.2. HECE	11
2.3. KELİME (SÖZCÜK)	12
2.4. KÖK VE GÖVDE	13
2.5. EKLER	14
2.5.1. <i>Yapım Ekleri</i>	14
2.5.1.1. İsimden İsim Yapma Ekleri	15
2.5.1.2. İsimden fiil yapma ekleri	16
2.5.1.3. Fiilden isim yapma ekleri	16
2.5.1.4. Sıfat Fiil (Partisip) Ekleri.....	17
2.5.1.5. Zarf Fiil (Gerundium) Ekleri	17
2.5.1.6. Fiilden Fiil Yapma Ekleri	18
2.5.2. <i>Çekim ekleri</i>	18
2.5.2.1. İsim Çekim Ekleri.....	19
2.5.2.2. Fiil Çekim Ekleri	21
2.6. SÖZCÜK TÜRLERİ	23
2.7. TÜRKÇENİN EKLEME KURALLARI	24
3. TÜRKÇE İÇİN YAPILAN DOĞAL DİL İŞLEME ÇALIŞMALARI.27	
3.1. METİN ÖN-İŞLEME	28
3.2. TEMEL ÇALIŞMALAR	34

İÇİNDEKİLER (Devam)

	<u>Sayfa</u>
3.2.1. <i>Dil Analizi Üzerine Yapılan Çalışmalar</i>	34
3.2.1.1. Biçimbirimsel Çözümleme	34
3.2.1.2. Sözlük Anlamsal Çözümleme.....	37
3.2.1.3. Sözdizimsel Çözümleme	38
3.2.1.4. Anlamsal Çözümleme ve Bakış-Açısal (Pragmatic) Çözümleme....	39
3.2.2. <i>Dil İstatistikleri Üzerine Yapılan Çalışmalar</i>	40
3.3. UYGULAMAYA YÖNELİK ÇALIŞMALAR	40
4. SÖZCÜK TÜRÜ TESPİT YÖNTEMİ.....	45
4.1. GİZLİ MARKOV MODELİ	47
4.2. YÖNTEM	49
4.2.1. <i>Sözcük Türü Etiketleme Kurgusu</i>	54
4.3. YÖNTEMİN BAŞARIM ORANLARININ TESPİTİ	57
4.3.1. <i>Kullanılan Derlemler</i>	57
4.3.2. <i>DeneySEL Sonuçlar</i>	59
5. GÖVDELEME.....	67
5.1. YÖNTEM	69
5.1.1. <i>Çekim Ekleri İçin Ek Kalıpları</i>	71
5.1.2. <i>Eklerin Stırılması</i>	77
5.1.3. <i>Uygun Gövdenin Belirlenmesi</i>	79
5.2. YÖNTEMİN BAŞARIM ORANLARININ TESPİTİ	89
5.2.1. <i>Kullanılan Derlemler</i>	89
5.2.2. <i>DeneySEL Sonuçlar</i>	90
6. TÜMLEŞİK BİÇİMBİRİMSEL ÇÖZÜMLEME VE SÖZCÜK TÜRÜ TESPİTİ YÖNTEMİ	99
6.1. YÖNTEM	101

İÇİNDEKİLER (Devam)

	<u>Sayfa</u>
6.1.1. Yapım ekleri	103
6.1.2. Yapım Eklerinin Stırılması.....	105
6.1.3. Uygun Çözümün Belirlenmesi.....	106
6.1.3.1. Kökteki Dilbilgisel Değişimler	112
6.2. YÖNTEMİN BAŞARIM ORANLARININ TESPİTİ.....	115
6.2.1. Kullanılan Derlemler	115
6.2.2. Deneyisel Sonuçlar.....	116
7. SONUÇ VE TARTIŞMA	127
KAYNAKLAR.....	131
8. EKLER.....	141
Ek 1. TÜRKÇEDE BULUNAN YAPIM EKLERİ	143
Ek 2. TÜRKÇEDE EK KURALLARI	157
Ek 3. SÖZCÜK TÜRÜ TESPİTİ BAŞARIM ORANLARI	165
Ek 4. GÖVDELEME YÖNTEMİ BAŞARIM ORANLARI.....	173
Ek 5. KÖKLEME VE GÖVDELEME YÖNTEMİ BAŞARIM ORANLARI.....	177
ÖZGEÇMİŞ.....	185

ŞEKİLLER DİZİNİ

Sekil

Sayfa

Şekil 4.1. Sözcük türü tespit yönteminin akışı (Hesaplama birimi: son hece, Model: 1. dereceden GMM)	56
Şekil 4.2. Hesaplama birimi kelime alındığında elde edilen başarımlar grafiği	59
Şekil 4.3. Hesaplama birimi son hece alındığında elde edilen başarımlar grafiği .	60
Şekil 4.4. 1. dereceden GMM ile hesaplama birimi kelimenin sonundan alınan harf birlikleri olduğunda elde edilen başarımlar grafiği	61
Şekil 4.5. 1. dereceden geçmişe bir adım GMM ile hesaplama birimi kelimenin sonundan alınan harf birlikleri olduğunda elde edilen başarımlar grafiği	61
Şekil 4.6. 1. dereceden GMM ile hesaplama birimi kelimenin başından alınan harf birlikleri olduğunda elde edilen başarımlar grafiği	62
Şekil 4.7. 1. dereceden geçmişe bir adım GMM ile hesaplama birimi kelimenin başından alınan harf birlikleri olduğunda elde edilen başarımlar grafiği	62
Şekil 4.8. 1. dereceden GMM ile hesaplama birimi kelimenin başından ve sonundan alınan harf birlikleri olduğunda elde edilen başarımlar grafiği	63
Şekil 4.9. 1. dereceden geçmişe bir adım GMM ile hesaplama birimi kelimenin başından ve sonundan alınan harf birlikleri olduğunda elde edilen başarımlar grafiği	64
Şekil 4.10. Hesaplama birimi kelime alındığında elde edilen başarımlar grafiği ..	64
Şekil 4.11. Hesaplama birimi çekim ekleri alındığında elde edilen başarımlar grafiği	65
Şekil 5.1. Fiil çekim ek kalıpları için SDM	76
Şekil 5.2. İsim çekim ek kalıpları için SDM	76

ŞEKİLLER DİZİNİ (Devam)

<u>Sekil</u>	<u>Sayfa</u>
Şekil 5.3. “istedim” kelimesinin analizi	78
Şekil 5.4. ODTÜ-Sabancı Türkçe derleminin elle gövdelenmesine yardımcı araç	89
Şekil 5.5. Sözcük türü bilinen kelimeler üzerinden gerçekleştirilen deney sonuçlarının grafiği	91
Şekil 5.6. Sözcük türü belirlenen kelimeler üzerinden gerçekleştirilen deney sonuçlarının grafiği	92
Şekil 6.1. Yöntemin genel görünüşü	102
Şekil 6.2. Yapım ekleri için tasarlanan SDM.....	105
Şekil 6.3. Gövdeleme işleminden elde edilen tüm olası gövdelerin kök tespitinde kullanımı	110
Şekil 6.4. Gövdeleme işleminden elde edilen gövdenin kök tespitinde kullanımı	111
Şekil 6.5. ODTÜ-Sabancı Türkçe derleminin köklerinin elle belirlenmesine yardımcı araç.....	116
Şekil 6.6. Sözcük türü bilinen tüm olası gövdeler üzerinden gerçekleştirilen deney sonuçları	118
Şekil 6.7. Sözcük türü bilinen en uygun gövde üzerinden gerçekleştirilen deney sonuçları	118
Şekil 6.8. Sözcük türü bilinmeyen tüm olası gövdeler üzerinden gerçekleştirilen deney sonuçları	119
Şekil 6.9. Sözcük türü bilinmeyen en uygun gövde üzerinden gerçekleştirilen deney sonuçları.....	120

ÇİZELGELER DİZİNİ

Çizelge

Sayfa

Çizelge 2.1. Bazı harflerin simgesel gösterimi	15
Çizelge 3.1. DDİ üzerine yapılmış çalışmaların sınıflandırılması	27
Çizelge 3.2. ODTÜ-Sabancı Türkçe derleminden elde edilen gövde uzunluğu istatistikleri	31
Çizelge 4.1. Bir milyon ve on milyon kelimelik Türkçe derlemelerin kelime dağılımları (Hakkani-Tür ve ark., 2002).....	52
Çizelge 4.2. Talim derleminden elde edilen sözcük türü oranları.....	54
Çizelge 4.3. Tüm hesaplama birimlerinden elde edilen en yüksek başarımlar	66
Çizelge 5.1. Türkçe gövdeleme çalışmalarının sınıflandırılması.....	69
Çizelge 5.2. Fiil çekim ekleri için örnek ek kalıpları	72
Çizelge 5.3. Fiil çekim ekleri ek kalıpları için alt gruplar	73
Çizelge 5.4. İsim çekim ekleri için örnek ek kalıpları	74
Çizelge 5.5. İsim çekim ekleri ek kalıpları için alt gruplar	75
Çizelge 5.6. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon I ile elde edilen olasılık sonuçları	82
Çizelge 5.7. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon II ile elde edilen olasılık sonuçları	82
Çizelge 5.8. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon III ile elde edilen olasılık sonuçları	83
Çizelge 5.9. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon IV ile elde edilen olasılık sonuçları.....	84

ÇİZELGELER DİZİNİ (Devam)

Çizelge

Sayfa

Çizelge 5.10. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon V ile elde edilen olasılık sonuçları	85
Çizelge 5.11. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon VI ile elde edilen olasılık sonuçları	85
Çizelge 5.12. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon VII ile elde edilen olasılık sonuçları.....	86
Çizelge 5.13. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon VIII ile elde edilen olasılık sonuçları	87
Çizelge 5.14. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon IX ile elde edilen olasılık sonuçları	88
Çizelge 5.15. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon X ile elde edilen olasılık sonuçları	88
Çizelge 5.16. Gövdeleme yöntemi başarımları (Bootstrapping yöntemi ile yapılan test).....	93
Çizelge 5.17. Sözcük türü bilinen kelimeler üzerinden gerçekleştirilen gövdeleme yönteminin başarımları (k-parça çapraz doğrulama yöntemi ile yapılan test).....	94
Çizelge 5.18. Sözcük türü bilinen kelimeler üzerinden gerçekleştirilen gövdeleme yönteminin başarımları (k-parça çapraz doğrulama yöntemi ile yapılan test).....	94
Çizelge 5.19. Sözcük türü belirlenen kelimeler üzerinden gerçekleştirilen gövdeleme yönteminin başarımları (k-parça çapraz doğrulama yöntemi ile yapılan test).....	95

ÇİZELGELER DİZİNİ (Devam)

Çizelge

Sayfa

Çizelge 5.20. Sözcük türü otomatik belirlenen kelimeler üzerinden gerçekleştirilen gövdeleme yönteminin başarımları (k-parça çapraz doğrulama yöntemi ile yapılan test).....	96
Çizelge 6.1. Yapım ekleri için oluşturulan gruplar.....	104
Çizelge 6.2. “gözlükçüler” kelimesi için 10 farklı varyasyon üzerinden kök olma olasılığı.....	109
Çizelge 6.3. Kök ve Gövdenin Tespit Deney Sonuçları	120
Çizelge 6.4. Sözcük türü bilinen olası tüm gövdeler üzerinden gerçekleştirilen deney sonuçları.....	121
Çizelge 6.5. Sözcük türü bilinen en uygun gövde üzerinden gerçekleştirilen deney sonuçları.....	122
Çizelge 6.6. Sözcük türü bilinmeyen tüm olası gövdeler üzerinden gerçekleştirilen deney sonuçları.....	123
Çizelge 6.7. Sözcük türü bilinmeyen en uygun gövde üzerinden gerçekleştirilen deney sonuçları.....	123
Çizelge 6.8. Biçimbirimsel çözümleme test sonuçları.....	124
Çizelge 8.1. Hesaplama birimi kelime alındığında elde edilen başarımları oranları tablosu.....	165
Çizelge 8.2. Hesaplama birimi son hece alındığında elde edilen başarımları oranları tablosu.....	166
Çizelge 8.3. Hesaplama birimi kelimenin sonundan alınan harf birlikleri olduğunda elde edilen başarımları oranları tablosu.....	167

ÇİZELGELER DİZİNİ (Devam)

Çizelge

Sayfa

Çizelge 8.4. Hesaplama birimi kelimenin başından alınan harf birlikleri olduğunda elde edilen başarımlar oranları tablosu.	168
Çizelge 8.5. Hesaplama birimi kelimenin başından ve sonundan alınan harf birlikleri olduğunda elde edilen başarımlar oranları tablosu.	169
Çizelge 8.6. Hesaplama birimi kelimenin gövdesi olarak alındığında elde edilen başarımlar oranları tablosu.	170
Çizelge 8.7. Hesaplama birimi kelimenin çekim ekleri olarak alındığında elde edilen başarımlar oranları tablosu.	171
Çizelge 8.8. Sözcük türü bilinen kelimeler üzerinden gerçekleştirilen deney sonuçları	173
Çizelge 8.9. Sözcük türü bilinen kelimeler üzerinden gerçekleştirilen deney sonuçları	174
Çizelge 8.10. Sözcük türü belirlenen kelimeler üzerinden gerçekleştirilen deney sonuçları	175
Çizelge 8.11. Sözcük türü belirlenen kelimeler üzerinden gerçekleştirilen deney sonuçları	176
Çizelge 8.12. Sözcük türü bilinen olası tüm gövdeler üzerinden gerçekleştirilen deney sonuçları (Varyasyon 1-5)	177
Çizelge 8.13. Sözcük türü bilinen olası tüm gövdeler üzerinden gerçekleştirilen deney sonuçları (Varyasyon 6-10)	178
Çizelge 8.14. Sözcük türü bilinen en uygun gövde üzerinden gerçekleştirilen deney sonuçları (Varyasyon 1-5)	179

ÇİZELGELER DİZİNİ (Devam)

Çizelge

Sayfa

Çizelge 8.15. Sözcük türü bilinen en uygun gövde üzerinden gerçekleştirilen deney sonuçları (Varyasyon 6-10).....	180
Çizelge 8.16.Sözcük türü bilinmeyen tüm olası gövdeler üzerinden gerçekleştirilen deney sonuçları (Varyasyon 1-5).....	181
Çizelge 8.17. Sözcük türü bilinmeyen tüm olası gövdeler üzerinden gerçekleştirilen deney sonuçları (Varyasyon 6-10).....	182
Çizelge 8.18. Sözcük türü bilinmeyen en uygun gövde üzerinden gerçekleştirilen deney sonuçları (Varyasyon 1-5).....	183
Çizelge 8.19. Sözcük türü bilinmeyen en uygun gövde üzerinden gerçekleştirilen deney sonuçları (Varyasyon 6-10).....	184

KISALTMALAR

BGG	: Bilgi Geri Getirim
CG	: Categorical Gramer
DD	: Doğal Dil
DDİ	: Doğal Dil İşleme
FSM	: Finite State Machine
GMM	: Gizli Markov Modelleri
GPSG	: General Phrase Structure Grammars
HMM	: Hidden Markov Models
HPSG	: Head Driven Phrase Structure Grammars
LFG	: Lexical Functional Grammars
NLP	: Natural Language Processing
ODTÜ	: Orta Doğu Teknik Üniversitesi
POS	: Part of Speech Tagging
PSG	: Phrase Structure Grammars
SDM	: Sonlu Durum Makinesi
SFG	: Systemic-Functional Grammars
TG	: Transformational Grammars

1. GİRİŞ

Doğal dil işleme (DDİ), belli kurallar çerçevesinde oluşmuş ve gelişmiş doğal dillerin yapısının sayısal olarak çözümlenip anlaşılmasını ve yeniden üretilmesini amaçlayan yapay zeka ve dilbilim sahalarının alt uğraş alanı olarak tanımlanır. DDİ ilerleyen yıllarda bilgisayar ve insan arasındaki etkileşimin yapısını değiştirebilecek teknolojilere altyapı sağlamasından dolayı oldukça fazla çalışılan ve önemli bir saha olma özelliği göstermektedir. Bu nedenle DDİ, bilgisayar bilimlerinin temel bir uğraş alanı olarak, dil uzmanları ve bilgisayar bilimcilerinin ortak çalışmasını gerektirmektedir.

Covington'a (1994) göre, DDİ üzerinde gerçekleştirilen çalışmalar temel olarak beş ana seviyede incelenebilmektedir: sesbilim (*ing.* Phonetic), biçimbilim (*ing.* Morphology), sözdizim bilim (*ing.* Syntactic), anlambilim (*ing.* Semantic) ve kullanımbilim (*ing.* Pragmatic).

Covington' un çalışmasında belirttiği beş ana seviyeden birincisi olan sesbilim harflerin seslerini ve bunların dil içinde nasıl kullanıldığını incelemektedir. Sesbilim konuşma anlama, konuşmayı yazıya çevirmek için gerekmektedir. Biçimbilim sözcüklerin meydana gelme durumlarını incelemektedir. Temel olarak iki tür sözcük oluşturma yöntemi (üretme ve ses değişmesi) mevcuttur. Sözdizim bilim, cümle içerisindeki veya cümle oluştururken kelimelerin ne şekilde sıralandığı incelerken anlambilim ise dili ifade eden kelimelerinden oluşan cümlelere verilebilecek anlamları incelemektedir. Bir sözcük tek başınayken ya da bir cümle içindeyken farklı anlamlar ifade edebilmektedir. Dilin zaman içerisindeki gelişimini inceleyen kullanımbilimin ana konusu doğal dildeki kelimelerin zaman içerisindeki kullanım ve anlam değişikliklerini incelemektir.

Sesbilim, biçimbilim ve sözdizim bilim alanında neredeyse tüm dillerde yoğun olarak çalışılırken bu seviyelerin üzerinde yer alan anlambilim ve kullanım bilim alanları diğerlerine göre daha bakir alanlar olarak görülmektedir. Türkçe dili ele alındığında ise yaklaşık son 20 yıldır yürütülen çalışmaların henüz tam olarak olgunlaşmadığı görülmektedir.

Bilgisayar ile DDİ yazılı metinlerde diller arası otomatik çevirim, soru cevap makineleri, otomatik konuşma, otomatik metin düzeltme, cümle üretme, komut anlama, konuşma sentezi, bilgi geri getirim ve otomatik metin indeksleme gibi birçok alanda olanaklar sağlamıştır. Belirtilen çalışma sonuçlarından bazıları oldukça başarılı bir şekilde hayatımıza girerken (otomatik metin düzeltme gibi) bilgi geri getirim, konuşma anlama, sentezleme ve dil üretme gibi alanlar üzerine yapılan çalışmalar henüz istenilen seviyelere gelememiştir.

DDİ çalışmalarına bakıldığında birbirlerinin üzerine bina edildikleri söylenebilir. Konuşma tanıma, konuşma sentezi, DD üretilmesi, doğal dil anlama, diller arası çevirim ve bilgi geri getirim gibi birçok çalışma alanı cümle sonu tespiti, gövdeleme, biçimbirimsel çözümleme, sözcük türü tespiti vb. çalışmaları temel alarak gerçekleştirilmektedir. Bu nedenle cümle sonu tespiti, gövdeleme, biçimbirimsel analiz, sözcük türü tespiti gibi çalışma alanlarını temel çalışmalar olarak nitelendirebilmekteyiz. Temel çalışmalar daha üst seviyede yer alan hedef çalışmalara hizmet ettiğinden dolayı iyi bir şekilde tasarlanmaları ve başarımları oldukça önem kazanmaktadır.

DDİ sahasında, işlenen dilin yapısından bağımsız olma iddiasına sahip kuramlar geliştirilmeye çalışılmasına rağmen, yapılan çalışmalar dünya üzerinde en fazla geçerli olan dil olarak kabul edilen İngilizce üzerine yoğunlaşmaktadır. Ancak kendi dilimiz olan Türkçe içinde benzeri çalışmaları yapmak dilimiz için oldukça büyük önem

taşımaktadır. Zira, gelecekte sayısal ortamlarda temsili az olan ve bilgisayarlarca işlenemeyen diller geçerlilikleri kaybetmeye mahkum kalacaklardır. Türkçe kökeni ve gramer yapısı itibari ile İngilizceden oldukça farklıdır. Bu nedenle İngilizce veya diğer diller için yapılan çalışmaları Türkçeye uygunluğuna bakmadan uyarlamaya çalışmak doğru bir yöntem olarak görülmemektedir. Türkçenin sayısal ortamlarda daha etkin işlenebilmesi ve anlaşılması için dilbilgisel yapısının dikkatlice incelenip dilbilimcilerle ortak bir çalışma ortamında sistemlerin geliştirilmesi gerekmektedir.

Bu tez çalışmasında dil bilgisi konusunda Türkçe uzmanları ile birlikte çalışarak doğal dil işleme ve anlamının temelinde bulunan gövdeleme, kökleme, biçimbirimsel çözümleme ve sözcük türü tespitinde dil özellikleri göz önünde bulundurularak ve cümle bazında bir yaklaşım ile belirsizliği ve karmaşıklığı indirgenmiş istatistiksel yöntemlerle birleştirilmiş bir yaklaşım ortaya konmaktadır.

Biçimbirimsel çözümleme, kelimenin kök ve eklerinin geçiş hallerinin belirlenmesi şeklinde tanımlanabilir. Bilgimiz dahilinde Türkçe için gerçekleştirilen biçimbirimsel çözümleme yöntemlerinin hemen hemen hepsinde sonlu durum makineleri (SDM) kullanılmaktadır. Sonlu durum makinelerini temel alan tüm yöntemlerde de bir sözlük bulunmaktadır. Bu sözlük genel itibarı ile belirli sayıda kelimeden ve her kelime için atanmış bir grup özellikten oluşmaktadır. Biçimbirimsel çözümleyicinin temel hesaplama biriminin kelime olarak alınması nedeniyle Türkçe gibi sondan eklemeli çekimli dillerde bir kelime şekli için çoğunlukla birden fazla biçimbirimsel çözümleme ortaya çıkmaktadır. Biçimbirimsel üretkenliğin yüksek olması yüksek bir hesaplama karmaşıklığına neden olmaktadır. Genelde yapılan karmaşıklık tespiti hesaplamalarında, birden fazla sunulan biçimbirimsel çözümün hangisinin doğru olduğunun tespiti problemi dahil

edilmemektedir. Yani, bahsedilen hesaplama karmaşıklığının tamamı, biçimbirimsel çözümlemelerden gelen karmaşıklıklaştır.

Şimdiye kadar Türkçe için geliştirilmiş biçimbirimsel çözümleme yöntemlerindeki hesaplama karmaşıklığının diğer bir sebebi, ek dizim kurallarını gerçekleyen SDM' lerde her alternatif ek geçişinin eşit şansa sahip olmasıdır. Ancak, dizim kuralları içinde eklerin birbiri ardına gelişleri, en azından Türkçe için rastsal değildir, yani bir ekten sonra gelebilecek olası ekler kümesindeki tüm elemanlar eşit şansa sahip değildir. Bir başka ifadeyle, Türkçe'de kök veya gövdelerin sonuna getirilen eklerin herhangi bir karışıklığa meydan vermeyecek şekilde belirli bir sırayla dizilmekte ve bu düzenli sıra Türkçe'nin ek yapısına sadelik, açıklık ve kolaylık getirmektedir (Ergin, 2003).

Belirsizliğe yol açan diğer bir sebep de, bir ek şeklinin sadece tek bir dilbilgisi özelliği için kullanılmamasıdır. Örneğin –acak eki hem yapım eki¹ hem de çekim eki olarak kullanılabilir. Fakat en azından Türkçe için, kelimenin cümle içinde bulunduğu yer dikkate alındığında bu tür eş şekilli farklı göreve sahip olan eklerin de ayrımı yapılabilir. Daha doğrusu bir ek hem yapım eki hem de çekim eki olabilir. Ayrıca hem isim için hem de fiil için eş ek şekline sahip olabilir; ancak, kelimenin kullanıldığı yer itibari ile bu seçeneklerden sadece bir tanesi geçerlidir. Bu durumu Ergin (2003), Türkçe bir cümlede kelimelerin dizilişi ile ek sistemi arasında sıkı sıkıya bir ilişkinin var olduğunu söyleyerek vurgulamıştır.

Yukarıda bahsedilen sebepler göz önüne alındığında, mevcut biçimbirimsel çözümleme yöntemlerinin kelimeleri tek başlarına çözümlemelerinden dolayı, geçersiz alternatifleri de hesapladığı söylenebilir.

¹ Esasen –acak eki partisip eki olup yapım eki sınıfında ele alınabilen bir ektir.

Sözcük türü tespiti, kelimenin metin içerisinde hangi dilbilgisi işlevinde kullanıldığının belirlenmesi şeklinde tanımlanabilir. Sözlük-anlamsal seviyede kelimenin hangi dilbilgisi işlevinde kullanıldığı, anlamının tespitinden önce belirlenmelidir. Bunun nedeni, bir kelimenin dilbilgisi işlevinin, (örneğin *isim*, *sıfat* v.b.) anlamı destekleyen bir kavram olmasıdır. Ayrıca yukarıda da bahsedildiği gibi bir kelimenin sözcük türü onun biçimbirimsel çözümlemesini etkileyen önemli faktörlerden biridir.

Ardışık iki süreç olan biçimbirimsel ve sözcük türü tespiti çözümlemeyi tümleşik süreçler olarak gerçeklemeyi hedeflediğimizde ortaya iki alt çözümleme problemi çıkmaktadır:

- ✓ Yazılı bir kelimenin olası biçimbirimsel çözümlemelerinin yapılması (*ing.* Morphological Analysis),
- ✓ Belirlenmiş olası biçimbirimsel çözümlemelerden hangisinin, kelimenin cümle içindeki dilbilgisi işlevini karşıladığının tespiti (*ing.* Part-of-Speech Tagging and Disambiguation)

Bu tez çalışmasında ortaya koyulan iddia, problemin dil çözümlemelerinin doğruluğunun artırılması ve belirsizliğinin indirgenmesi için kelime değil cümle bazında ele alınması ve Türkçenin dil özellikleri göz önünde bulundurularak, kelimelerin dilbilgisi işlevlerinin önceden belirlenmesi sayesinde, basitleştirilmiş ve doğruluk oranı arttırılmış tümleşik bir biçimbirimsel çözümleme ve sözcük türü tespit yönteminin geliştirilebileceğidir.

Bu iddia doğrultusunda geliştirilen yöntem ile kelimeler için tek bir sonuç sunularak belirsizliğin giderilmesi de sağlanmıştır. Bunlara ek olarak yöntemimizde yine alanda kabul görmüş çalışmalardan farklı olarak bir kelime sözlüğü yerine ek ve ek dizilimlerinden oluşan kapalı bir sözlük kullanılmıştır. Bu da kelime dağarcığının kapalı olduğu

söylenemeyecek diller için etkin bir çözüm olarak görülmektedir. Bilgimiz dahilinde, doğal dil işleme alanında etkileşimli tümleşik çözümlemeye yönelik bir başka çalışma bulunmamaktadır.

Bu yaklaşım, mevcut bir biçimbirimsel çözümlemenin, olasılık tabanında bir sözcük türü tespit yönteminin sağladığı bilgiyi kullanabilecek hale getirilmesi şeklinde tanımlanabilir. Tez kapsamında gerçekleşen çalışma ile biçimbirimsel çözümleme ve sözcük türü tespiti problemlerinin içinde önemli bir problem olan gövdeleme ve kökleme için de çözüm sunulmuştur. Ayrıca yaptığımız çalışmalar sonucunda araştırmacıların kullanımına açık, bilgi geri getirim, bilgi çıkarma, konuşma anlama vb. gibi uygulama alanları için oldukça önemli olan biçimbirimsel çözümleme ve sözcük türü tespit yordamı sağlanmıştır.

Tezin akışı içerisinde toplam 7 bölüm bulunmaktadır. Giriş bölümü dışında, Türkçe ek yapıları ve ekleme kuralları hakkında bilgilerin sunulduğu ikinci bölüm tezin geliştirilmesinde temel olarak kabul edeceğimiz teorik bilgileri barındırmaktadır. Tezin üçüncü bölümünde doğal dil işleme alanında daha önce gerçekleştirilmiş çalışmalar bir sınıflama içerisinde sunulmaktadır.

Tez kapsamında geliştirilen kapalı bir dağarcık üzerinde çalışan sözcük türü tespit yönteminin tanıtımı ve başarımlar sonuçları dördüncü bölümde sunulurken, tezimizin alt problemlerinden biri olan gövdeleme işleminin tanıtımı ve başarımlar sonuçları ise beşinci bölümde yer almaktadır.

Bir sonraki bölümde ise tezin ana gövdesini oluşturan tümleşik bir biçimbirimsel analiz ve sözcük türü tespit yönteminin tanıtımı ve başarımlar sonuçları yer almaktadır.

Tezin son bölümünde ise sonuç ve çalışmayı daha ileriye götürebilecek tartışma ve öneriler yer almaktadır.

2. TÜRKÇE

Bu bölümde tez kapsamında geliştirilen yöntemlerin üzerinde çalıştığı dil olan Türkçe hakkında kısa bir bilgi verildikten sonra Türkçenin ek yapısından bahsedilecektir.

Dil temel olarak insanların düşündüklerini ve hissettiklerini belirtmek için kullandıkları kelimeler ve kurallar bütünü olarak tanımlanabilir. Diller yapıları itibarı ile üç ana sınıf altında toplanabilirler. *Tek heceli diller* sınıfında yer alan dillerde tüm kelimeler tek heceden oluşmakta ve kelimeler cümleler içerisindeki kullandıkları yerlere göre anlam kazanmaktadırlar. *Eklemeli ya da bitişken dillerde* ise kelimelerin kökleri değişmeden aldıkları eklerle gövdeleri oluşturur veya anlam kazanırlar. Bu tip dillere örnek olarak Türkçe, Moğolca ve Macarca gibi diller verilebilmektedir. Son olarak da *çekimli veya bükümlü diller* sınıfında yer alan dillerde kullanılan kelimeler ise kullanım sırasında görevine ve yerine göre değişikliklere uğramaktadır. Arapça, Farsça, İngilizce gibi diller bu sınıfa örnek olarak verilebilir (Ergin, 2003; Banguoğlu, 2004).

Diller yapısal olarak sınıflandırılmalarının yanı sıra doğdukları bölge ve yayılımlarına göre belli başlı gruplar altında toplanmaktadırlar. Bu gruplama ile birbirleri ile ilişkili ve yakın olan diller bir aile kabul edilmektedir. Dünya üzerindeki diller için her ne kadar böyle bir sınıflama yapılsa da bazı dillerin hiçbir yaşayan akrabası bulunmamaktadır. Dillerin kökenlerine göre aileleri: Hint – Avrupa dilleri ailesi, Hami-Sami dilleri ailesi, Çin-Tibet dilleri ailesi, Bantu dilleri ailesi, Ural dilleri ailesi ve Altay dilleri ailesi şeklinde sıralanabilir. Bazı kaynaklarda Ural ve Altay dilleri ailelerinin diğer dil ailelerinde olduğu gibi sağlam bir aile özelliği göstermedikleri belirtilerek bu sınıflandırma içerisinde dahil edilmemektedir (Voegelin & Voegelin, 1977; Ergin, 2003; Raymond, 2005).

İçerisinde yüzlerce dil bulunan Hint – Avrupa dilleri ailesi, dünya üzerindeki en büyük dil ailesidir. Dünya üzerinde yaklaşık 2,5 milyar kişi ana dil olarak bu dil ailesinde yer alan bir dili kullanmaktadır. Bu dillere örnek olarak İngilizce, Almanca, Fransızca, İspanyolca, Portekizce, Hintçe, Rusça ve İtalyanca örnek olarak verilebilir. Bu dil ailesi kendi içerisinde birçok kola ayrılmaktadır.

Yaklaşık 250 dil içeren Hami-Sami Dil Ailesi, Orta Doğu ve Kuzey Afrika'ya yayılmış çeşitli topluluklarca konuşulmaktadır. Bu dil ailesine mensup dillere örnek olarak Arapça, İbranice, Aramice, Habeşçe, Amhara, Tigrece ve Tigrina verilebilir.

Bir başka dil ailesi olan Çin-Tibet dil ailesi yaklaşık 300 adet Doğu Asya dilinden oluşur. Çince, Tibetçe, Dzongka ve Birmanca bu dil ailesine giren dillere örnek olarak verilebilir.

Bantu dilleri ailesi, Afrika kullanılan en büyük dil ailesidir. Orta ve güney Afrika'da konuşulan Bantu dillerine örnek olarak Kikongo, Lingala, Kikenya, Swahili ve Kirundi verilebilir.

Kendi içerisinde iki kola ayrılan Ural dillerine örnek olarak Fince, Estonca, Mari Dili, Komice, Karelce. Macarca, Khansi dili, Samoyet ve Nenetçe verilebilir.

Altay dilleri Avrupa'dan, Orta Doğu'ya ve Orta Asya'dan Uzak Doğu'ya kadar uzanan büyük bir coğrafyada konuşulan dilleri kapsayan bir dil ailesidir. Altay dil ailesinin kolları arasında Türk dilleri, Moğol, Japon ve Kore dilleri bulunmaktadır.

Ural ve Altay dil aileleri arasındaki yapısal yakınlık bu iki dil ailesinin bazı kaynaklarda tek bir aile olarak gösterilmesine neden olmaktadır. Ural – Altay dil ailelerine mensup diller yapı olarak eklemeli dillerdir. Bunun yanı sıra dillerin hemen hemen hepsinde benzer bir sesli uyumu göze çarpmaktadır.

Özetlemek gerekirse Türkçe, Altay dil ailesi grubunun Türk dilleri koluna bağlı, sondan eklemeli bir dildir. Altay dilleri içinde Türkçeye en yakın olan Moğolcadır. Bunun yanı sıra Sinanoğlu, 2002 yılındaki çalışmasında Türkçe ve Japonca'nın da oldukça benzer diller olduğunu belirtmiştir.

2.1. Alfabe (Abece)

Alfabe (ya da öz Türkçesi ile Abece), dildeki bir sese karşılık gelen harfler dizisi olarak tanımlanır. Alfabeler dilin yazıya geçirilmesinde kullanılan sembollerdir. Alfabe (Alphabet) sözcüğü, eski Yunanca kullanılan ilk iki harf olan "alpha" ile "beta"nın birleşiminden oluşurken "Abece" sözcüğü ise Türkçe'deki ilk üç harfin okunuşundan meydana gelmektedir.

Türkler tarih boyunca dillerini yazıya geçirmek için çok farklı alfabeler kullanmışlardır (Göktürk alfabesi, Uygur alfabesi, Kiril alfabesi, Arap alfabesi, Latin alfabesi). Günümüzde Türkler arasında en yaygın alfabe Latin ve Kiril alfabeleridir.

Türkiye Türkçesinde kullanılan alfabede 29 harf bulunmaktadır: A, B, C, Ç, D, E, F, G, Ğ, H, I, İ, J, K, L, M, N, O, Ö, P, R, S, Ş, T, U, Ü, V, Y, Z. Temel olarak bu harfler ses özellikleri bakımından sesli harfler ve sessiz harfler olmak üzere iki gruba ayrılırlar.

2.2. Hece

Hece, kelimeleri oluşturan ağzın tek hareketinde sarf edilebilen ses modeli olarak tanımlanabilir. Türkçede heceler bir sesli harf olmak şartı ile en az bir harf, en çok dört harften oluşmakta ve genel olarak 6 farklı hece tipi bulunmaktadır. Bu tipler hece içerisindeki harf sayısı ve dizilişlerine göre gruplanmaktadır:

- ✓ Sadece bir sesli harften oluşan heceler: a, e, ı vb.

- ✓ Bir sesli, bir sessiz harften oluşan heceler: ab, ay, ol vb.
- ✓ Bir sesli, iki sessiz harften oluşan heceler: üst, alt vb.
- ✓ Bir sessiz, bir sesli harften oluşan heceler: ma, fa, ta vb.
- ✓ Bir sessiz, bir sesli, bir sessiz harften oluşan heceler: kat, tak, sak vb.
- ✓ Bir sessiz, bir sesli, iki sessiz harften oluşan hece: park, kalp vb.

Bu hece tipleri dışında Türkçede oldukça sık kullanılan yabancı dillerden gelen kelimelerde rastlanılan farklı tip heceler de bulunabilmektedir. Örneğin iki sessiz bir sesli harften oluşan heceler: pro, sta vb. ve iki sessiz, bir sesli, bir sessiz harften oluşan heceler: spor, trak vb.

2.3. Kelime (Sözcük)

Kelime (sözcük), genellikle bir anlama sahip ve tek başına kullanılabilen birbirine sıkı sıkı bağlı bir ya da daha fazla biçim birimden oluşan dil birimi olarak tanımlanmaktadır (Uzun, 1998). Bir kök ve eklerden oluşan kelimeler, bir varlığı, bir nesneyi veya bir eylemi karşılamaktadır. Fakat, anlamı olmayan, hiçbir varlığı, nesneyi veya eylemi karşılamayan kelimeler de bulunmaktadır (ve, ile, fakat vb.).

Türkçede sözcük türleri isim, sıfat, zamir, zarf, edat, bağlaç, ünlem ve fiil şeklinde belirlenmektedir. Her ne kadar sekiz farklı sözcük türü belirlense de bu sözcük türlerinden ilk yedisi isim soylu olarak bilinmektedir. Bir başka deyişle Türkçede kelimeler iki temel sözcük türü vardır: İsim ve Fiil. Bu yapı dilin anlatım yapısına uygundur. Çünkü anlatım yapısında bir yapılan iş ve bu işi yapan kişi bulunmaktadır. Bu da fiil ve isimi karşılamaktadır. Diğer sözcükler ise yalnızca isim ve fiilin anlamını tamamlamaktadır. Bu yapıdan farklı olarak bazı kaynaklara

göre edatlar da bir başka temel sözcük türü olarak gösterilmektedir (Deny, 1955).

Kelimelerin ve özellikle isim soylu kelimelerin sözcük türlerini belirlemek oldukça zor bir iştir. Çünkü, Türkçede isim soylu sözcükler için sözcük türleri birbirinden kesin çizgilerle ayrılmamaktadır. Bir kelime isim, sıfat, zarf, ünlem gibi görevleri üstlenebilir. Bu nedenle sözcük türlerinin tespiti meselesinde kelimelerin cümle içerisindeki anlam ve işlevlerine bakmak gerekmektedir (Hengirmen, 2005).

2.4. Kök ve Gövde

Kelime yapısında tek başına anlamlı olan en küçük ve daha fazla parçalanamayan sesbirimsel şekillere *kök* adı verilir. Bir başka deyişle kök kelimenin çekirdeğidir. Kökler eskiden beri var olan ve sonradan yapılamayan dil birlikleridir. Türkçede tüm köklerin tek heceden oluştuğu ve zaman içerisinde farklılaştığı düşünülmektedir. Yeni bir kökün dile eklenmesi oldukça seyrek rastlanan bir durumdur. Yeni kökler genellikle yabancı dillerden gelmektedir (örneğin internet, radyo vb.).

Türkçede temel olarak isim kökü ve fiil kökü olmak üzere iki çeşit kök bulunmaktadır. İsim kökleri evrendeki varlıkları, nesneleri ve kavramları karşılayan köklerdir. Fiil kökleri ise, bu varlık ve kavramların eylemlerini karşılamaktadır. İsim kökleri tek başlarına kullanılabilirken, fiil kökleri anlam yapıları açısından bir isme (nesne, şahıs) ve zamana bağlı olarak kullanılabilirler. Fiilleri tek başlarına kullanmak isteğimizde ise genellikle bu fiillere verilen isimlerini kullanırız. Örneğin “koş_” fiilini “koşma” veya “koşmak” gibi isim halleri ile kullanırız.

Türkçede her varlık, nesne veya eylem için bir kök bulunmamaktadır. Birbirine yakın anlamlı birçok kelime aynı kökten bazı ekler kullanılarak türetilmiştir. Örneğin “göz” kökü kullanılarak gözlük, gözcü, gözlükçü vb. birçok kelime türetilmektedir. Köklerde türeyen

bu yapılar “gövde” veya “geniş kök” olarak tanımlanmaktadır. Gövdeler, köklere getirilen yapım ekleri ile oluşturulmaktadır. Gövde dilbilgisel olarak kök ile aynıdır. Tek farkı yapım eki almış olmasıdır.

Bir Türkçe sözcüğün kökünün veya gövdesinin bilinmesi, sözcüğün temel anlamını vermektedir. Her ek, sözcüğe yeni ancak herkesin anlayacağı bir anlam veya ifade kazandırmaktadır. Bu nedenle, bir sözcüğün kökünün veya gövdesinin elde edilmesi ve eklerinin belirlenmesi önemlidir. Ancak, biçimsel olarak bir sözcüğün çözümlenmesi oldukça zahmetli bir iştir.

2.5. Ekler ¹

Tek başına manası olmayan ve kullanılmayan, ancak köklerle birleşmek suretiyle mana veya çekim ile ilgili bir vazife gören birimlere ise ek denir. Köklere gelen ekler Türkçenin gramer kuralları doğrultusunda şekil değiştirebilirler. Bu nedenle eklerin büyük bir kısmı çok şekilli olmaktadır (Örneğin koş-an, gel-en, al-ıp, ver-ip, koş-up, gör-üp, yaz-dı, sev-di, oku-du, gör-dü, at-tı, git-ti, sus-tu, düş-tü). Ekler genel olarak bir veya iki heceden oluşmasına rağmen üç heceli eklerle de rastlanmaktadır.

Türkçede ekler temel olarak ikiye ayrılmaktadır: yapım ve çekim ekleri.

2.5.1. Yapım Ekleri

Eklendiği kökün veya kelimenin anlamında değişiklik yaparak eklendikleri kökten yeni bir gövde meydana getiren eklerle verilen genel isimdir. Yapım ekleri isim veya fiil soylu köklere veya gövdelere

¹ Bu bölüm Muharrem Ergin’in “Türk Dili” kitabı ve Zeynep Korkmaz ve arkadaşlarının “Türk Dili ve Kompozisyon Bilgileri” kitaplarından derlenmiştir.

eklenerek isim veya fiil gövdesi oluşturabilirler. Dolayısı ile dört çeşit yapım eki mevcuttur.

- ✓ İsimden isim yapma ekleri.
- ✓ İsimden fiil yapma ekleri.
- ✓ Fiilden isim yapma ekleri.
- ✓ Fiilden fiil yapma ekleri.

Eklerin birçoğu harf değişimlerine göre farklı şekiller almaktadır (Örneğin araba-cı, göz-cü, kitap-çı vb.). Çok şekilli ekleri tek bir biçimde göstermek amacı ile belirli simgeler kullanılmıştır. Bu simgeler Çizelge 2.1’de verilmektedir.

Çizelge 2.1. Bazı harflerin simgesel gösterimi

Gösterim	Karşılıkları
A	a, e
I	ı, i
U	u, ü
K	k, g
C	c, ç
D	d, t

Tez çalışması kapsamında derlenen Türkçe’deki yapım eklerinin bazıları açıklamaları ile birlikte örnek olarak aşağıda verilirken tamamı ise ek 1’de sunulmaktadır.

2.5.1.1. İsimden İsim Yapma Ekleri

-lUk Eki: (-lık, -lik, -luk, -lük) Yer, topluluk, alet, vasıf (sıfat), ve soyut isimleri yapar. Bu ekin bir özelliği de bazen çekim eklerinden

sonra gelerek isim yapmasıdır: gün-de-lik, on-da-lık gibi. Örnekler: taşlık, zeytinlik, koruluk, gözlük

-CU Eki: (-cı, -ci, -cu, -cü, -çı, -çi, -çu, çü) Bu eklerin başlıca fonksiyonu meslek isimleri yapmaktır. Örnekler: avcı, eskici, yolcu, gözcü, aşçı, simitçi, okçu, çöpçü

-IU Eki: (-lı, -li, -lu, -lü) Ek ve sahiplik ya da bağlılık ifade ederek sıfat olarak kullanılan vasıf isimleri yapmaktır. Örnekler: kanlı, düşünceli, sulu, köpüklü

-sUz Eki: (-sız, -siz, -suz,- süz) Bu ek -IU ekinin olumsuzudur. Örnekler: ağaçsız, içkisiz, duygusuz, ölçüsüz

2.5.1.2. İsimden fiil yapma ekleri

-IA Eki: (-la, -le) İşlek bir isimden fiil yapma ekidir. Olma, yapma ifade eden fiiller yapmaktadır.. Örnekler: sula, iğnele

-I Eki: (-l) Sıfatlardan olma ifade eden fiiller yapmaktadır. Örnekler: kısal

-Ar Eki: (-ar, -er) Daha çok renk isimlerinden fiil yapmaktadır. Örnekler: morar, yeşer

-DA Eki: (-da, -de, -ta, -te) Ses taklidi isimlerden fiil yapmaktadır. Örnekler: mırılda, gürülde

2.5.1.3. Fiilden isim yapma ekleri

-mAk Eki: (-mak, -mek) En çok kullanılan eklerdendir. Bütün fiil kök ve gövdelerine getirilebilir. Görevi eylemin ismini oluşturmaktır. Örnekler: Okumak, yürümek, yüzmek, konuşmak

-mA Eki: (-ma, -me) Tüm fiil kök ve gövdelerinde kullanılabilir. Fonksiyonu iş isimleri yapmaktır. Örnekler: Okuma, yürüme, yüzmeye, konuşma

-Uş Eki: (-ış, -iş, -uş, -üş) İş ismi yapmakta kullanılan bu ek çok kullanılan eklerden biridir. Örnekler: alış, yürüyüş, çekiliş, oturuş

-m Eki: (-m) Kalıcı nesne ismi yapan bu ek çok kullanılan eklerdendir. Örnekler: uçur-u-m, geç-im

2.5.1.4. Sıfat Fiil (Partisip) Ekleri

Sıfat fiil (partisip) ekleri genel görüşe göre yapım ekleri ile çekim ekleri arasında yer almaktadır. Fakat bu ekler kalıcı isim yapma görevini icra edince tam anlamı ile fiilden isim yapma eklerine dahil edilebilmektedir. Tez kapsamımızda kabulümüz sıfat fiil eklerinin fiilden isim yapma ekleri grubuna dahil olduğudur.

-An Eki: (-an, -en) Oldukça sık kullanılan ek kalıcı isim yapmaktadır. Örnekler: seven, koşan

-Ar Eki: (-ar, -er) Geçici ve kalıcı isimleri yapmaktadır. Örnekler: gider, keser, koşar

-r Eki: (-r) Geçici ve kalıcı isim yapmaktadır. Örnekler: yürür, gelir, yatır

-mUş Eki: (-mış, -miş, -muş, -müş) Geçici ve kalıcı isim yapmaktadır. Örnekler: koşmuş, okumuş geçmiş, yemiş

2.5.1.5. Zarf Fiil (Gerundium) Ekleri

Zarf fiiller (gerundiumlar) hareket, devamlılık, zaman ve sebep gibi zarf fonksiyonlarını ifade eden fiil şekilleri olarak tanımlanmaktadır. Hiçbir çekim eki almamaktadır. Fiilleri zarfa çevirmeleri dolayısı ile tez kapsamında fiilden isim yapma grubuna dahil alınarak incelenmektedir.

-A Eki: (-a, -e) Genellikle ikilemelerde kullanılan bu ek tek kullanıldığı zaman bileşik fiil yapmaktadır. Örnekler: koşa koşa, gele gele, çekebil, düşeyaz

-ArAk Eki: (**-arak, -erek**) Oldukça kullanılan bir ektir.

Örnekler: koş-arak, başla-y-arak, gül-erek

-UncA Eki: (**-ınca, -ince, -unca, -ünce**) Oldukça kullanılan bir ektir. Örnekler: yapınca, gidince, okuyunca

-AlI Eki: (**-alı, -eli**) Devamlılık ifade eden zarflar oluşturur. Örnekler: geleli, görelî, oturalı

2.5.1.6. Fiilden Fiil Yapma Ekleri

-mA Eki: (**-ma, -me**) Olumlu fiillerden olumsuz fiiller yapmaktadır. Örnekler: koşma, yürüme, okuma

-n Eki: (**-n**) Edilgen fiiller oluşturmaktadır. Örnekler: Aran, sevin, okun

-s Eki: (**-ş**) Ortaklaşma ve oluş ifade eden fiiller türetmektedir. Örnekler: vuruş, atış, dayanış, gülüş

-r Eki: (**-r**) Yaptırma ve oldurma ifade eden fiiller türetmektedir. Sadece ç,s,t,ğ,p,y ile biten tek heceli fiillere getirilmektedir. Örnekler: aşır, yatır

2.5.2. Çekim ekleri

Kelimelerin cümle içerisinde kullanılmasını sağlayan çekim ekleri, eklendiği kök veya gövdenin anlamında değişiklik yapmadan diğer kelimelerle bağ kurmasını, cümlede bir görev almasını, hâlini, sayısını, zamanını, şahsını belirten eklerdir.

Çekim ekleri getirildikleri kelimenin türüne göre ikiye ayrılır: isim çekim ekleri ve fiil çekim ekleri.

2.5.2.1. İsim Çekim Ekleri

Çokluk Ekleri

Çokluk ekleri, eklendiği ismin birden fazla olduğunu ifade etmektedir. Türkçede çokluk eki -lAr'dır. Örnekler: kalemler, evler, balıklar.

İyelik Ekleri

İyelik ekleri, eklendiği ismin bir şahsa veya bir nesneye ait olduğunu ifade eden eklerdir. Türkçede bulunan iyelik ekleri şunlardır:

Teklik 1. şahıs: -m (evim, kalemim)

Teklik 2. şahıs: -n (evin, kelemin)

Teklik 3. şahıs: -U, -sU (evi, kalemi, arabası, paltosu)

Çokluk 1. şahıs: -mUz (evimiz, topumuz)

Çokluk 2. şahıs: -nUz (eviniz, arabanız)

Çokluk 3. şahıs: -lArI (evleri, topları)

Tekil 3. şahıs iyelik eki -U ve -sU olmasına rağmen ne ve su kelimeleri için tekil 3. şahıs iyelik eki eklendiğinde su-yu, ne-yi şeklini almaktadır.

Aitlik Eki

Bağlantı kavramı veren bu ek büyük ve küçük ünlü uyumuna uymamaktadır. Türkçede aitlik eki -ki'dir.

Durum (Hal) ekleri

Hal ekleri, ismin başka bir kelimeyle ilgi kurmak için aldığı eklerdir. Bu ekler ismi, kendisine bağlı olamayan, kendisinin bağlı olduğu öğelerle ilgi kurmaktadır. Bu ekler isimleri isimlere, edatlara ve fiillere bağlamaktadır. Türkçede isimlerin halleri şunlardır:

Yalın hal: İsimlerin herhangi bir öğeye bağlı olmadığı durumdur.

İlgi Hali (Genitif Hali): İsmın başka bir isimle ilişkilendirmektedir. İlgi hali ekleri –Un, –nUn’dır. (evin, kolun, devenin)

Belirtme Hali (Akuzatif Hali): İsmın, geçişli fiillerin etkisi altında kaldığını ifade eden halidir. Belirtme hali eki –(y)U’dur. (evi, sürüyü)

Yaklaşma Hali (Datif Hali): İsmın kendisine yaklaşma ifade eden fiillerle ilişkilendiğini belirtmektedir. Yaklaşma halinde –(y)A eki kullanılmaktadır. (eve, arabaya)

Bulunma Hali (Lokatif Hali): İsmın kendisinde bulunma ifade eden fiillerle ilişkilendiğini belirtmektedir. Bulunma hali eki –DA’dır. (evde, arabada)

Ayrılma Hali (Ablatif Hali): İsmın kendisinden ayrılma ifade eden fiillerle ilişkilendiğini belirtmektedir. Ayrılma hali eki –DA’dır. (evden, arabadan, attan)

Vasıta Hali (Instrumental Hali): İsmın fiile bir vasıta olduğunu ifade etmektedir. Vasıta hali eki –Un’dır. (yazın, kışın)

Eşitlik Hali (Ekvatif Hali): İsmın eşitlik, benzerlik, gibilik ifade ettiği durumdur. –CA eki ile yapılmaktadır. (bence, insanca)

Yön gösterme Hali (Direktif Hali): Fiilin kendi önünde yapıldığını göstermek için ismin girdiği haldir. –rA ve –ArI ekleri ile yapılır. (sonra, üzere, ileri, yukarı)

Soru Eki

İsmın soru şeklini yapan bir ektir. İsmın yalın haline, çokluk, iyelik ve hal şekillerine eklenebilir. Soru eki –mU’dur. (evde mi, kitap mı, büyük mü)

Ek Fiil Ekleri

Ek fiil ekleri, isimlerin sonuna gelerek onların yüklem olmasını sağlayan, ek halindeki bir fiildir. Esasen “imek” fiilinin ek olarak kullanımıdır. Dört kipe göre çekimi vardır: Geniş zaman, -di’li geçmiş zaman, -miş’li geçmiş zaman, şart kipi.

Geniş zaman: İsimlere iyelik ekleri getirilerek yapılır. Bunlar geniş zaman eklerinin yerini tutar. Üçüncü şahıslar için “-dIr” eki kullanılır. (insanım, insansın, insandır, insanız, insansınız, insandırlar)

-di’li geçmiş zaman: İsimlerin bilinen geçmişteki durumuna şahit olunduğunu göstermektedir. (hayranıydım, hayranıydın, hayranıydı, hayranıydık, hayranıydınız, hayranıydılar)

-miş’li geçmiş zaman: İsimlerin öğrenilen geçmişteki durumunun başkasından duyulduğunu anlatmaktadır. (çalışkanmışım, çalışkanmışsın, çalışkanmış, çalışkanmışsınız, çalışkanmışlar)

Şart kipi: Eklendiği isimlerle yancümlecik kurarak ve temel cümlenin şartını bildirmektedir. (güzelsem, güzelsen, güzelse, güzelsek, güzelseniz, güzelseler)

2.5.2.2. Fiil Çekim Ekleri

Eylemleri karşılayan kelimeler olan fiiller Türkçede muhakkak çekimli halde kullanılmaktadır. İkinci tekil şahıs emir kipi haricinde tüm fiiller çekim eki almaktadır. Fiil çekim ekleri, fiil kök veya gövdelerine eklenerek, fiillerin zamanını, yapılış şeklini ve şahsını belirtmektedir. Dilimizdeki fiil çekim ekleri şu şekilde sıralanabilir: Şahıs Ekleri, şekil ve zaman ekleri ve soru ekleri. Fiil çekim ekleri belli kurallar eşliğinde farklı şekillerde arka arkaya eklenerek fiil çekimi sağlarlar. Bu kurallar ilerleyen bölümlerde sunulacaktır.

Şahıs Ekleri

Fiillerde şahısları gösteren eklere denir. Türkçede 3 çeşit şahıs eki vardır.

1. grup şahıs ekleri (görülen geçmiş zaman, dilek ve şart kiplerinde kullanılır.): tekil 1. şahıs: -m, tekil 2. şahıs: -n, tekil 3. şahıs: - (eksizdir), çoğul 1. şahıs: -k, çoğul 2. şahıs: -nUz, çoğul 3. şahıs: -lAr.

2. grup şahıs ekleri (şimdiki zaman, geniş zaman, gelecek zaman, öğrenilen geçmiş zaman, dilek, istek, gereklilik kiplerinde kullanılır.): tekil 1. şahıs: -Um, tekil 2. şahıs: -sUn, tekil 3. şahıs: - (eksizdir), çoğul 1. şahıs: -Uz, çoğul 2. şahıs: -sUnUz, çoğul 3. şahıs: -lAr.

3. grup şahıs ekleri (Emir kiplerinde kullanılır.): tekil 1. şahıs: -AyIm, tekil 2. şahıs: - (eksizdir), tekil 3. şahıs: -sUn, çoğul 1. şahıs: -Allm, çoğul 2. şahıs: -Un, UnUz çoğul 3. şahıs: -sUnlAr.

Şekil ve Zaman Ekleri

Şekil ekleri fiilin ne şekilde yapıldığını belirtirken zaman ekleri fiilin ne zaman yapıldığını belirtmektedirler.

Bildirme Kipleri: Belli bir zaman kavramı ile birlikte eylemi bildiren fiil şekillerine denir. Zamanlar şu şekilde sıralanabilir: Belirli geçmiş zaman (-dU), anlatılan geçmiş zaman (-mUş), şimdiki zaman (-yor), geniş zaman (-r, -mez), gelecek zaman (-AcAk).

Tasarlama Kipleri: Zaman kavramı olmayan tasarlama kipleri, şart (-sA), istek (-A), gereklilik (-mAll) veya emir (-sUn) şeklinde görülmektedir.

Soru eki

Soru eki (-mU) fiillerden sonra da kullanılabilir. Eklendiği kelimeden ayrı yazılmasına rağmen ünlü uyumlarına uymaktadır.

2.6. Sözcük Türleri

Türkçede genellikle kabul edilen 8 farklı sözcük türü bulunmaktadır: İsim (Ad), Sıfat (Önad), Zamir (Adıl), Zarf (Belirteç), Edat (İlgeç), Bağlaç (Rabıt), Ünlem (Nida) ve Fiil (Eylem). Bu sekiz sözcük türünden ilk yedisi isim soylu yani isimlerden türetilmiş ya da isimlerin farklı görevlerde kullanılmaya başlanmasıyla meydana gelmiştir. Yine bu sekiz sözcük türünden sonuncusu olan fiiller ise kendi içinde önemli bir sözcük türünü oluşturmaktadır. Bir başka deyişle sözcük türleri isim ve fiil olmak üzere 2 ana grup altında toplanmaktadır. Bazı kaynaklarda ise edatlar da farklı bir alt grup olarak tanımlanarak sözcük türleri 3 ana grup altına toplanmaktadır.

- ✓ İsimler, canlı ve cansız varlıkları, duygu ve düşünceleri, çeşitli durumları bildiren kelimelerdir.
- ✓ Sıfatlar, bir ismin önüne gelerek o ismi niteleme ya da belirtme görevini üstlenen kelimelerdir.
- ✓ Zamirler, gerçek bir isim olmadıkları halde cümlede isimlerin yerine kullanılabilen, isimlerin yerini tutan kelimelerdir.
- ✓ Zarflar, bir fiilin, sıfatın veya başka bir zarfın anlamını yer, zaman, nicelik, soru bakımından belirten veya etkileyen kelimelerdir.
- ✓ Edatlar, farklı türdeki kelimelerin veya kavramlar arasında anlam ilgisi kurmaya yarayan tek başlarına anlamları olmayan ancak diğer sözcüklerle birlikte, cümle içinde işlev gören kelimelerdir.
- ✓ Bağlaçlar, kelimeleri, kelime gruplarını veya cümleleri birbirine bağlayan kelimelerdir.
- ✓ Ünlemler, sevinç, üzüntü, korku vb. duyguları belirtmede kullanılan genelde doğa seslerini yansıtan veya bir kimseyi

çağırarak için kullanılan, tek başlarına anlam ifade etmeyen kelimelerdir.

- ✓ Fiiller, iş, durum ve hareketleri karşılayan kelimelerdir.

2.7. Türkçenin Ekleme Kuralları

Tez kapsamında, farklı Türk dili kitaplarının incelenmesi sonucunda özellikle eklerin köke veya gövdeye eklenmeleri konusunda bir takım kurallar derlenmiştir. Derlenen bu kurallar yöntemlerin geliştirilmesinde, SDM tasarlanmasında ve ek dizilimleri (tez kapsamında ek kalıpları diye isimlendirdiğimiz) sözlüğünün oluşturulmasında kullanılmıştır. Elde edilen kuralların tamamı ek 2’de verilmiş aşağıda ise çıkarılan kurallardan bazı örnekler sunulmuştur.

- 1- Yapım ekleri kökten hemen sonra gelir. Çekim ekleri yapım eklerinden önce gelmez. –lUk –cUk, -gil ve –ki ekleri istisnadır.
 - i. On-da-lık, gün-de-lik gibi. Fakat burada –dA eki çekim eki işlevi kaybolmuş ek sanki –dAlUk şeklini almıştır.
 - ii. –ki ekide benzer şekilde ismin -DA, -Un, -nUn eki almış çekimli şekilleri ile (yani lokatif ve genitif halleri ile) birleşmektedir. Bu durumun en önemli nedeni –ki ekinin edat geçmişinin olmasıdır. Aynı zamanda –ki eki çekim eki (aitlik eki) olarak da sınıflandırılmaktadır.
 - iii. –cUk eki de –CA çekim ekinden sonra kullanılabilir. Örneğin, usulcacık, yavaşcacık vb.
 - iv. Babamgil örneğinde olduğu gibi –gil yapım eki –m şahıs eklerinden sonra gelebilmektedir.
- 2- Yapım eki sıralanışı aşağıdaki gibi olmaktadır:
 - a. İsimden fiil yapma => Fiilden fiil yapma

- b. İsimden fiil yapma => fiilden isim yapma
 - c. İsimden isim yapma => isimden fiil yapma
 - d. İsimden isim yapma => isimden isim yapma
 - e. Fiilden fiil yapma => fiilden fiil yapma
 - f. Fiilden fiil yapma => fiilden isim yapma
 - g. Fiilden isim yapma => isimden fiil yapma
 - h. Fiilden isim yapma => isimden isim yapma
- 3- Ekte “o” veya “ö” harfi bulunmaz. Sadece -yor eki istisna oluşturmaktadır.
- 4- İsim ve fiillere gelen çekim ekleri birbirinden farklıdır.
- 5- Aynı cins çekim eki üst üste gelmez.
- 6- Genetif ekinden sonra soru ve -ki eki gelir.
- 7- Akuzatif ekinden sonra sadece soru eki gelebilir.
- 8- Datif ekinden sonra ancak soru eki ve -dır, -dir bildirme eki gelebilir.

3. Türkçe İçin Yapılan Doğal Dil İşleme Çalışmaları

Covington (1994)'un gruplamasında yer alan 5 ana seviye (sesbilim, biçimbirim, sözdizim bilim, anlambilim ve kullanımbilim) üzerinde DDİ' nin uğraşı alanlarına bakıldığında konuşma tanıma, konuşma sentezi, DD üretilmesi, DD anlama, bilgi geri getirim gibi belli başlı konular karşımıza çıkmaktadır.

Diğer dillerde olduğu gibi Türkçe üzerinde de DDİ çalışmaları yukarıda belirtilen seviyelerde ve uygulama alanlarında yürütülmektedir. Tezin bu bölümünde Türkçe üzerinde gerçekleştirilen DDİ araştırmaları ve uygulamaları derlenmiş ve belli bir sınıflandırma içerisine alınmaya çalışılmıştır.

Çizelge 3.1. DDİ üzerine yapılmış çalışmaların sınıflandırılması

Uygulamaya Yönelik Çalışmalar	Enformasyon Bilim (Bilgi Geri Getirim, Sıkıştırma vb.), Doğal Dil İşleme Uygulamaları (Metin anlama, konuşma Anlama, Dil Üretme vb.) (Darcan (1991), Say (1998), Çetinoğlu (2001), Maden ve ark. (2003), Cebiroğlu ve ark. (2003), Bozşahin, Oflazer (1994), Oflazer (1996), Tür (2000), Dinçer (2004), WordNet (2006))
Temel Çalışmalar	Söylev (Discourse) ve Bakış-Açısal (Pragmatic) Çözümleme (Kural tabanlı ve istatistiksel)
	Anlamsal Çözümleme (Semantic Analysis) (Kural tabanlı ve istatistiksel)
	Söz-dizimsel Çözümleme (Syntactic Analysis) (Kural tabanlı ve istatistiksel) (Şehitoğlu (1996), Çiçekli, Korkmaz (1998), Bozşahin (1994), Güngördü (1995), Birtürk (1998))
	Sözlük-Anlamsal Çözümleme (Sözcük türü tespiti ve anlamsal belirsizliğinin giderilmesi) (Lexical Analysis) (Kural tabanlı ve istatistiksel) (Oflazer -Kuruöz (1994), Oflazer-Tür (1996, 1997), Tür ve ark. (2002), Dinçer (2004))
	Biçimbirimsel Çözümleme (Morphologic Analysis) (Kural tabanlı ve istatistiksel) (Oflazer (1994), Sever (2003), İçer (2004), Güngör, (1995))
Metin Ön-işleme (Köksal (1979), Solak ve Can (1994), Alpkoçak ve ark. (1995), Tür (2000), Dalkılıç ve Dalkılıç (2001), Hakkani-Tür ve ark. (2002), Oflazer ve ark. (2003), Atalay ve ark. (2003), Sever ve Bitirim (2003), Dinçer (2004), Özcan (2004), Akın (2004), Eryiğit ve Adalı (2004), Çilden (2006))	

DDİ üzerinde birçok teorik ve uygulamaya dönük çalışma gerçekleştirilmiştir ve halen de birçok çalışma yürütülmektedir. Türkçe için tamamlanmış veya devam etmekte olan çalışmalar Çizelge 3.1’de gösterilen şekilde sınıflandırılmıştır.

3.1. Metin Ön-İşleme

Çizelge 3.1’de belirtilen sınıflandırmanın tabanında yer alan metin-ön işlemeyle ilgili konular hakkında birçok çalışma gözlenmektedir.

Doğal dil işleme çalışmaların hemen hemen hepsinde öncelikle yazılı metnin belirli bir biçime getirilmesi gerekmektedir. Yazılı metnin noktalama işaretleri, boşluklar vb. anlamsal-bilgi açısından önemsiz olan öğelerden ayrıştırılma işlemine *metin normalleştirme* (ing. text normalization) yada *simgeleştirme* (ing. tokenization) denmektedir.

Metin normalleştirme işleminin en zor safhası cümle sonu tespiti işlemidir. Bunun nedeni noktanın, her zaman cümle sonu belirtmemesi bazı durumlarda bir kısaltmanın işareti (T.C. gibi) olabilmesidir. Bunun ayrıştırılması oldukça önemlidir.

Cümle sonu tespiti problemi için Türkçede birçok çalışma bulunmaktadır. Bunlardan ilki Tür’ün (2000) doktora çalışması sırasında geliştirdiği yöntemdir. Tür, Gizli markov modelini temel alarak geliştirdiği yöntemde olasılık dağılımları ve sözlük kullanarak yüksek sayılabilecek bir oranda başarı sağlamıştır. Yöntemin olumsuz yönü kelime dağarcığı açısından kapalı olmayan Türkçe üzerinde, sözlük kullanması ve çalışmanın biçimbirimsel çözümleme yapılmamış bir derlem üzerinde çalışmaması olarak gösterilebilir. Sistemde kullanılan biçimbirim çözümlemesi Oflazer (1993) tarafından geliştirilendir.

Diğer bir çalışmada ise Oflazer ve arkadaşlarının (2003) Siemens firması desteğinde yürütülmüş olan LC-STAR adlı bir projenin Türkçe konusundaki çalışmalarıdır. Bu proje aralarında Türkçenin de bulunduğu

13 dil arasında konuşmadan-konuşmaya çevirim yapabilecek bir sistem oluşturmayı hedeflemektedir. Projenin amacına uygun olarak düzenlenen alt görevlerden biri cümle sonu tespittir. Projede diğerinden farklı olarak cümle tespiti için kural tabanlı bir yöntem kullanılmıştır.

Dinçer (2004) doktora tez çalışmasında bu konu ile ilgili olarak farklı bir yöntem önermiştir. Dinçer bu yöntemde, “metin içerisindeki noktadan önce gelen kelime hecelenebiliyorsa, söz konusu noktanın cümle sonu olma olasılığı olmama olasılığından daha yüksektir” önermesinden yola çıkmış ve yöntemdeki başarı oranı %96.24 olarak hesaplamıştır.

Yazılı metinlerde dilin tespit edilmesine yönelik çalışmalar da belirtilen sınıflandırmada metin ön işleme kısmına dahil edebileceğimiz çalışmalardır. Bu konuda bahsedeceğimiz ilk çalışma Dalkılıç ve Dalkılıç (2001) tarafından yapılan ve n-gram istatistiklerini kullanan çalışmadır. Çalışmada sunulan iki öneriden birincisinde esas olarak kelime uzunlukları diğerinde ise kelime sonu ve başı harf dizilişleridir. Önerilerden ilkinin temelinde, Türkçe ve İngilizcede değişik kelime uzunlukları için, metinlerde görülme sıklığı oranlarının farkı yatmaktadır. İkinci öneride ise, kelime sonundaki harfin sesli veya sessiz oluşuna göre karar verilebileceğidir: İngilizcede sadece kelimelerin %28’i sesli harfle biterken, Türkçede kelimelerin yaklaşık %50’si sesli ile bitmektedir. Dalkılıç ve Dalkılıç yaptıkları bu çalışma ile esas olarak Türkçenin n-gram istatistiklerini çıkarmayı hedeflemişlerdir.

Yazılı metinlerde dilin tespit edilmesi için Dinçer (2004), Türkçenin belirli kurallar çerçevesinde heceleme özelliğinden yararlanmaktadır. Metin içerisinde hecelenebilen kelime oranına bakarak metnin (%70 ve üstü oran Türkçe metin olarak kabul edilmekte) Türkçe olup olmadığına karar vermektedir. Dinçer’in geliştirdiği yöntem daha

sonra Özcan (2004) tarafından internet üzerindeki metinlere uygulanmıştır.

Metin ön işleme çalışmaları kısmına dahil edebileceğimiz diğer bir konu ise derlem oluşturma çalışmalarıdır. Türkçe için oluşturulmuş derlemlerden ilki Bilkent Üniversitesinde Hakkani-Tür ve arkadaşları (2002) tarafından yapılan eklemeli dillerde biçimbirimsel ve sözlük anlamsal seviye üzerinde istatistiksel yöntemler kullanıldığı bir çalışmada ortaya çıkmıştır. Yaklaşık bir milyon kelimeden oluşan derlemin içeriğini internette sayısal olarak erişilebilen gazete makaleleri oluşturmuştur.

Türkçe dil çalışmalarında oldukça sık kullanılan diğer bir derlem ise, ODTÜ’de Atalay ve arkadaşları (2003) ile Oflazer ve arkadaşlarının (2003) çalışmaları sırasında oluşturdukları derlemidir. Bu derlem kontrollü olarak farklı konu başlıklarından meydana getirilmiş yaklaşık iki milyon kelimelik bir derlemidir.

Metin ön işleme çalışmalarına ekleyeceğimiz diğer bir işlem ise gövdelemedir. Gövdeleme esasen sınıflandırmamızda bulunan temel çalışmaların içerisinde de kendisine yer bulan bir alandır. Gövdeleme, kelimeden tüm çekim eklerinin atılma işi olarak tanımlanabilir. Türkçe için geliştirilmiş olan birkaç gövdeleme yöntemi bulunmaktadır. Gövdeleme hem doğal dil işleme hem de bilgi geri-getirim sahasında çalışılmasından dolayı en çok incelenen konu olmaktadır.

Türkçe için ise geliştirilmiş olan birkaç gövdeleme yöntemi bulunmaktadır. Bunlar içinde en eski olanı Aydın Köksal’ ın (1979) ortaya koyduğu yöntemidir. Bu yöntemle göre gelen her kelimenin ilk 5 harfi gövde olarak kabul edilmektedir. Bu yöntemin başarısını test için gövdelerini elle işaretlediğimiz ODTÜ-Sabancı Türkçe derlemi üzerinden çekilen gövde uzunluğu istatistikleri Çizelge 3.2’de verilmiştir.

Buna göre 5 harf uzunluğundaki gövdeler toplam kelime sayısının %23 ünü oluşturmaktadır. 3, 4, 5 ve 6 harf uzunluklarına toplam olarak baktığımızda ise kelime sayısının %69' unu oluşturmaktadır.

Çizelge 3.2. ODTÜ-Sabancı Türkçe derleminden elde edilen gövde uzunluğu istatistikleri¹

<i>Gövde Uzunluğu (Harf cinsinden)</i>	<i>Kelime Sayısı</i>	<i>%</i>	<i>Gövde Uzunluğu (Harf cinsinden)</i>	<i>Kelime Sayısı</i>	<i>%</i>
1	776	1	6	5410	10
2	6019	11	7	3910	7
3	10582	20	8	2681	5
4	7833	15	9	1209	2
5	12154	23	10 ve üstü	1572	3

Bir diğer çalışmada ise Alpkoçak ve arkadaşları (1995), L-M (*ing.* Longest Match) veya en uzun eşleşme adıyla bir yöntem ortaya konmuştur. Yöntemde, kelime gövdelerinin ve olası biçimlerinin yer aldığı bir sözlükten yararlanılarak, gövdelenen kelime bu sözlükte yer alan kelimelerle harf esasında soldan başlayarak eşleştirilmektedir. En uzun eşleşmenin yakalandığı kelime, gövde olarak değerlendirilmektedir.

Solak ve Can (1994), Türkçe gövdelerin ve her gövde için, gövde üretme yöntemlerini destekleyen 64 özelliğin bulunduğu bir sözlük kullanmışlardır. Gövdeleme için, kelimenin soldan sağa doğru harf birlikleri sırasıyla, kök sözlüğünden eşleştirilmekte ve eşlenen kök için olası gövdeler türetilmektedir. Türetilen gövdelerin kelime içerisinde var olma durumuna göre olası gövdeler belirlenmektedir. Bu yöntemle, bir kelime için olası birkaç gövde belirlenebilmektedir. Bu çalışma temel olarak, Oflazer (1993) tarafından geliştirilmiş olan biçimbirimsel çözümleme yönteminin, gövdeleme işlemine uygulanmış halidir.

¹ ODTÜ-Sabancı Türkçe Derleminden çekilen istatistiklere noktalama işaretleri dahil edilmemiştir. Noktalama hariç 52179 adet kelime istatistiklere dahil edilmiştir.

Sever ve Bitirim (2003) tarafından geliştirilmiş olan gövde-bul (*ing.* FindStem) adlı gövdeleme yöntemi ise temel olarak üç bölümden meydana gelmektedir: kökün tespiti, biçimbirimsel çözümleme ve gövdenin belirlenmesi. Yöntemde, Türkçe kelime köklerinin, bu kelimelerin biçimbirimsel özelliklerinin, sözcük türlerinin ve dizim kurallarının bulunduğu bir sözlük kullanılmaktadır. Kök tespitinden sonra gerçekleşen biçimbirimsel çözümleme işleminde ise, yapım ekleri ve çekim ekleri ayrı olarak ele alınmakta ve analiz edilip gövde belirlenmektedir.

Akın (2004) tarafından Türkçe ve diğer Türki diller için ortaya konulan “zemberek” isimli projede ise Türkçe dil işleme sahasındaki temel problemlere cevap verebilecek çözümler sunulmaktadır. Geliştiriciler bunu Türkçe doğal dil işleme kütüphanesi olarak belirtmişler ve açık kaynak kodlu olarak kullanıma sunmuşlardır. Konumuz itibari ile ilgilendiğimiz kelime çözümleme yönteminde Türkçede yer alan tüm kök ve gövdeler bir sözlük içerisine yerleştirilmiş ve her sözcük için harf bazında ağaç yapıları kullanılmıştır. Buna göre gövdelerde veya kökte oluşabilecek her tür ses değişimi de sözlükte yer bulmaktadır. Kelime incelenirken soldan harf harf ilerlenerek sözlük içerisinde aranmış ve kelimenin olası kök ve ya gövdesi bulununca eklerin ayrıştırılmasına geçilmiştir. Yapılan bu çalışmanın en büyük avantajı yapısı değişmeden sadece sözlük ve bazı değişkenleri adaptasyonu sayesinde diğer Türki diller içinde uygulanabilir olmasıdır.

Bir başka çalışmada ise Eryiğit ve Adalı (2004) ek sıyırma yöntemi ile gövdeleme yapmışlardır. Bu amaç için ek sözlüğü ve geliştirdikleri SDM’lerini kullanmışlardır. Kelime sağdan başlayarak, harf harf alınarak belirlenen harf birlikleri ek sözlüğünde aranmış ve kelimenin olası gövdeleri belirlenmiştir. Çilden (2006), Eryiğit ve Adalı’nın çalışmasında

ortaya konan SDM'lerini, Porter tarafından geliştirilen “snowball” uygulamasına uyarlamıştır.

İncelenen son çalışmada ise Dinçer (2004), kelimelerin gövdesini n-gram istatistikleri kullanarak tespit etmeyi önermiştir. Bu yöntemle göre iki harfin ard arda gövdede, ekte ve gövde ek geçişinde olma olasılıkları kullanılmıştır. Gövdesi tespit edilecek kelime soldan başlayarak ikişer harf alınarak incelemeye tabi tutulmaktadır. İncelemede iki harfin ekte bulunma olasılığı gövdede bulunma olasılığından yüksekse ve aynı zamanda iki harfin gövde ek geçişinde olma olasılığı belli bir parametre değerinin üstünde ise bu noktadan önce bulunan harf birlikleri gövde olarak kabul edilmektedir. Buna göre bir kelimedenden birden fazla olası gövde türetilmektedir. Çalışmada yöntemin gövdeleri üretebilme doğruluğu %95,8 olarak bildirilmiştir. Yöntemin en büyük getirisi sözlük kullanmaması olarak belirtilmiştir.

Gövdeleme üzerinde yapılan çalışmalara genel olarak bakıldığında, neredeyse tamamında belli bir kök veya gövde sözlüğünün kullanıldığı ve sonuç olarak olası gövdeleri sunduğu gözlemlenmektedir.

Metin ön işleme sınıfında belirtilen bu çalışmalar sınıflandırmanın bir üst seviyesi olan temel çalışmalar için alt veriyi sağlamaktadır. Genelde metin ön işlemeden gelen veriler ile beslenen temel çalışmalar, yapılan sınıflandırmada “Dil Analizi” ve “Dil İstatistikleri” üzerine yapılan çalışmalar olarak iki grup altında incelenmiştir. Türkçede DDI üzerine yapılan çalışmalara bakıldığında genelde biçimbirimsel çözümleme ve bu çözümlemelerin üzerinde çalışan uygulamalar üzerine yoğunlaşıldığı söylenebilir.

3.2. Temel Çalışmalar

3.2.1. Dil Analizi Üzerine Yapılan Çalışmalar

Doğal dil işlemede dil analizi üzerine yapılan çalışmalar Çizelge 3.1’de de görüldüğü gibi beş ana başlık altında toplanabilir:

- ✓ Biçimbirimsel Çözümleme (*ing.* Morphologic Analysis)
- ✓ Sözlük-Anlamsal Çözümleme (*ing.* Lexical Analysis)
- ✓ Söz-dizimsel Çözümleme (*ing.* Syntactic Analysis)
- ✓ Anlamsal Çözümleme (*ing.* Semantic Analysis)
- ✓ Söylev (*ing.* Discourse) ve Bakış-Açısal (*ing.* Pragmatic) Çözümleme

3.2.1.1. Biçimbirimsel Çözümleme

Bu başlıklardan ilki olan biçimbirimsel çözümleme, yazılı bir kelimenin kök, gövde, yapım ve çekim eklerine ayrıştırılması şeklinde tanımlanabilir. Biçimbirimsel çözümleme, yazılı metinlerden anlam çözümleme sürecinin temel alt işlemlerinin ilkidir. Daha üstte bulunan çözümlemelerin kullandığı temel bilgiyi sağlamaktadır.

Türkçe için en bilinen ve birçok çalışmada da kullanılan biçimbirim çözümleme yöntemi Oflazer (1994) tarafından yapılan çalışmadır. Bu yordam, sonlu-durum makinelerinin (Finite State Machine) sondan eklemeli, çekimli dillerde kelimelerin yazım şekillerinde gösterdikleri farklılıkları kapsayacak şekilde geliştirilmesi ile meydana getirilen ve Koskeniemi (1983) tarafından iki-seviyeli dil tanımlaması (*ing.* Two-level Language Description) olarak adlandırılan bir kurguya dayanır. Aslen bu kurgu Fince için geliştirilmiştir ve Barton (1986) tarafından Türkçe için kullanılmıştır. Daha sonra bu problemin NP-Complete bir problem olduğu gösterilmiştir (Sever, 2003). Fakat

pratikte bu problemin karmaşıklığının daha makul seviyelerde olduğu belirtilmiştir. Oflazer, Türkçede yapım ve çekim eklerinin kelime kök ve gövdelerine getiriliş kurallarını ve oluşan gövde veya çekimli halin hangi dilbilgisi özelliğine ait olduğunu sonlu-durum makineleri kullanarak ortaya koymuştur.

İki-seviyeli biçimbirim, kelime yapılarının biçimbirimsel çözümlemesi için genel bir model olarak görülebilir. İki-seviyeli tanımlamalar, kelime yapısının sözcüksel ve yüzeysel gösterimlerini temel almaktadır. Sözcüksel gösterim kelimenin işlevsel parçalarının yapılarını belirtir. Yüzeysel gösterim ise verilen sözcüksel yapıya uygun olarak imla kurallarının işletilmesi sonucu ortaya çıkan kelimeyi belirtir. Bu modelde, kelimenin yazım (*ing.* surface) şekli için biçimsel (*ing.* lexical) bir şekil tanımlanmakta ve yazımsal şekil ile biçimsel şekil arasında harf harf eşleştirme yapılmaktadır.

Bu yapıda eklerin eklenmesi sırasındaki ses kısıtlamaları ve değişimlerini göstermek için 4 farklı kural tipi kullanılmıştır: Sağ-Yön Kuralı (*ing.* Context Restriction Rules), Sol-Yön Kuralı (*ing.* Surface Coercion Rule), Çift-Yön Kuralı (*ing.* Composite Rule), Yasaklama Kuralı (*ing.* Exclusion Rule).

Yukarıda verilen kurallar için kullanılan harfler sözcüksel ve yüzey formları kelimeler için farklı şekillerde ifade edilmiştir. Burada Türkçe’de harflerin eklenme sırasında uymak zorunda oldukları uyum kuralları dikkate alınmıştır. Sözcüksel form olarak tanımlanmış bir harf, yüzey formu olarak birden fazla harfle eşleşebildiği gibi yüzey formunda kullanımı söz konusu değildir. Örneğin, ‘H’ harfi esasen dar seslileri temsil etmekte ve yazımsal olarak hiçbir yerde kullanılmamaktadır.

Belirtilen 4 farklı kural tipi sınıflaması göz önünde bulundurularak tanımlanmış kurallar, yani dilin biçimbirimlerinin tam sıralamasını

belirleyen “morfortaktikler” (biçim taktikler), kök kelimeler ve ekler için sözlük kullanılarak sonlu-durum makinesi şeklinde tasarlanmıştır.

Oflazer’ in biçimbirimsel çözümleme çalışmasında (Oflazer,1994) iki farklı SDM modeli önerilmektedir. Bunlar Türkçenin isim çekim eklerinin dizim kuralları (biçim taktikleri) ve Türkçenin fiil çekim eklerinin dizim kurallarıdır. Bu SDM’ler Türkçeden elde edilen Oflazer’ in oluşturduğu 22 adet iki seviyeli kuralı temel almaktadır. Bu iki seviyeli kurallar dizisi kelimeleri sözcüksel ve yüzey formları ile tanımlar. Çalışmada SDM kullanıcıdan kurallara uygun verileri alan bir ayrıştırıcı olarak tasarlanmıştır. Bu tasarım PC-KIMMO (Antworth, 1990) ortamı kullanılarak SparcStation ELC makinelerde yaklaşık 23.000 kelime temel alınarak oluşturulmuştur. İcer (2004) yüksek lisans tezinde Oflazer’ in tanımladığı yapıyı kullanarak FSM yapıları yerine daha basit olan FST (*ing.* Finite state transducer) yapılarıyla biçimbirimsel çözümlemeciyi tasarlamıştır.

Oflazer’ in çalışması dışında sayısı çok az olmakla beraber biçimbirimsel çözümleme yapan çalışmalar bulunmaktadır. Bunlardan biri Boğaziçi Üniversitesi Bilgisayar Mühendisliği bölümünde doktora tez çalışması olarak gerçekleştirilmiştir. Bu uygulama esas olarak yazım (imla) kontrolü ve düzeltmesi yapmaktadır. (Güngör, 1995) Çalışmada Türkçenin biçimbirimsel yapısı, son eklerin dizim kuralları ve son eklerin biçim durumlarının sesbilim kuralları açısından araştırılmış ve tanımlanmıştır. Bu analizin sonuçları ATN (*ing.* Augmented Transition Network) yardımı ile Türkçenin biçimbirimsel yapısının tasarımında kullanılmıştır. Ayrıca geniş bir Türkçe sözlük oluşturulmuş ve oluşturulan çalışmanın imla kontrol kısmında kullanılmıştır. Bu sözlük kök kelimeler, ek kelimeler (ekler) ve özel isimler (*ing.* proper noun) olarak bölünmüştür. İmla denetleyicinin, çözümleyici kısmı kök-eşleme algoritmasını esas almaktadır. İmla denetleyici eğer bir girdiyi

parçalayamaz ise sözcük düzeltici kısım bu yanlış yazılmış girdi için alternatif adaylar üretir. Proje “Pascal” programlama dili ile kodlanmış ve uygulama aşamasında 21.727 kök kelime ve 9.528 özel isim kullanılmıştır.

3.2.1.2. Sözlük Anlamsal Çözümleme

Biçimbirimsel çözümlemenin bir üst aşamasında kelimelerin sözlük-anlamsal çözümlemeleri yer almaktadır. Sözlük-anlamsal çözümleme ise en genel halinde kelimenin karşıladığı anlamlar arasından, kullanıldığı yer itibari ile hangi anlama geldiğinin belirlenmesi olarak tanımlanabilir. Yani bir kelimenin kullanıldığı cümle içerisinde hangi sözcük türünde olduğunun tespiti ve ardından kelimenin varsa anlamsal belirsizliğinin giderilmesi şeklinde tanımlanabilir. Sözlük-anlamsal çözümlemede, kelimenin cümle içinde kullanıldığı yer itibari ile dilbilgisi işlevinin tespiti (*ing.* Part-of-Speech Tagging) ve ardından varsa anlamsal belirsizliğin giderilmesi (*ing.* Word-Sense Disambiguation) gerçekleştirilmektedir.

Türkçe için ise sözcük türü tespitine yönelik kural tabanlı çalışmalara örnek olarak Oflazer -Kuruöz (1994) ve Oflazer-Tür (1996, 1997) tarafından yapılan çalışmaları verebiliriz. Çalışmalardan ilkinde Oflazer ve Kuruöz, temelde Oflazer’in morfolojik çözümlemesini kullanan ve Türkçe metinleri dilbilgisi işlevlerine göre etiketleyen bir araç geliştirmişlerdir. Diğer çalışmada ise Oflazer ve Tür, cümle içerisindeki kelimelerin anlamlarında oluşabilecek belirsizliği gidermek üzere bir yöntem önermişlerdir. Bu yöntemde temelde kullanılan biçimbirimsel çözümlemenin yanında kelime istatistikleri kullanılarak kelimelerdeki anlam belirsizliği giderilmeye çalışılmıştır.

İstatistiğe dayanan çalışmalar ise Türkçe için oldukça azdır. Bu çalışmalara örnek olarak Tür ve arkadaşlarının (2002) yaptığı çalışmayı

verebiliriz. Çalışmada Hakkani-Tür ve arkadaşları, biçimbirimsel çözümlemesi tamamlanmış metinler üzerinde Gizli Markov Modellerinde (*ing.* Hidden Markov Model) hesaplama birimi olarak kelimeleri kullanıp %93.95 lik bir başarı sağlamışlardır.

İstatistiki yöntemlere dayanan diğer bir çalışmada Dinçer (2004) biçimbirimsel çözümlemeye gerek kalmadan hesaplama birimini kelime sonlarından alınan birkaç harf olarak belirlemiş ve %85' lik bir başarı elde etmiştir. Bu yöntemde Dinçer bir önce anlatılan çalışmada olduğu gibi gizli Markov modelini esas almış, farklı olarak hesaplama biriminde kelime yerine kelimenin sonundan alınan belirli sayıda harf kümesini kullanmıştır.

3.2.1.3. Sözdizimsel Çözümleme

Sözdizimsel çözümleme, cümle içerisindeki kelimelerin dizilimlerini bu dizilimlerin değişmesi ile oluşan anlam kaymalarını incelemektedir. Dilbilim uygulamalarında kullanılmak üzere gerçekleştirilmiş Türkçe cümlelerin sözdizimsel çözümlemesi üzerine de pek çok çalışma bulunmaktadır.

Yapılan çalışmalarda, farklı gramer yapıları üzerine çalışmalar yoğunlaşmıştır. Araştırmalarda sözdizimsel çözümleme için kullanılan ve en bilinen yaklaşımlar, LFG (*ing.* Lexical Functional Grammars), PSG (*ing.* Phrase Structure Grammars), SFG (*ing.* Systemic-Functional Grammars) ve CG (*ing.* Categorical Gramer)'de değiştirilerek çoğaltılan TG (*ing.* Transformational Grammars), GPSG (*ing.* General Phrase Structure Grammars) ve HPSG (*ing.* Head Driven Phrase Structure Grammars) şeklinde sıralanabilir. Bu sözdizimsel çözümlemeler ve benzeri türevleri çeşitli araştırmalar ile Türkçe için yapılmıştır (Şehitoğlu,1996; Çiçekli ve Korkmaz, 1998; Bozşahin, 1994; Güngördü, 1995; Birtürk, 1998).

Şehitoğlu (1996) yüksek lisans tezinde, HPSG yaklaşımı kullanılarak sözdizimsel çözümleme yapmıştır. Bu yaklaşım dildeki nesneleri bilgisel içerikleri ile gösterimine dayanan bir yaklaşımdır. Gerçekleştirilen uygulama HPSG uygulamaları için tasarlanmış bir mantık programlama dili olan ALE ile yapılmıştır.

Çiçekli ve Korkmaz 1998’de gerçekleştirdikleri çalışmalarında, Türkçenin sözdizimini SFG kullanarak basit Türkçe cümleler oluşturan bir yapı tasarlamışlardır. Proje girilen anlamsal gösterimlerden Türkçe cümleler oluşturmaktadır. Projede Oflazer tarafından geliştirilen biçimbirimsel çözümleme yöntemi kullanılmıştır.

Sözdizimsel çözümleme ile ilgili bir diğer çalışma ise Güngördü’nün (1994, 1995) gerçekleştirdiği çalışmadır. Carnegie-Mellon üniversitesinde geliştirilen “Tomita” ayrıştırıcısını temel alan uygulamada, LFG yaklaşımı kullanılarak sözdizimsel çözümleme gerçekleştirilmeye çalışılmıştır. Sistemin biçimbirimsel çözümleme temelinde ise Oflazer’ in biçimbirimsel çözümlemecisi kullanılmıştır.

DDI’ nin biçimbirimsel, sözlük anlamsal ve sözdizimsel seviyelerinde çözümleme, ilgili seviyedeki bileşenlere ait kurallar yardımı ile doğal dilin bileşenlerini inceleme manasına gelmektedir. Biçimbirimsel seviyede kelimeler, morfotaktik ve morphophonomic kurallara göre; sözlük-anlamsal seviyede ise kelimeler karşıladığı anlam ve kullanıldığı yer itibari ile, sözdizimsel seviyede ise sözcük grupları, dilbilgisel ilişkiler kullanılarak incelenmektedir.

3.2.1.4. Anlamsal Çözümleme ve Bakış-Açısız (Pragmatic) Çözümleme

Araştırmamız sonucunda bilginiz dahilinde sınıflandırmamızda yer alan anlamsal ve pragmatik çözümleme konusunda Türkçe üzerine gerçekleştirilen çalışmalar alt alandaki çalışmaların tam olarak

tamamlanmamasından dolayı belli olgunluğa erişememiştir. Varolan birkaç çalışma uygulamaya yönelik çalışmalar bölümünde tanıtılacaktır.

3.2.2. Dil İstatistikleri Üzerine Yapılan Çalışmalar

Türkçe dil istatistikleri üzerine şu ana kadar yapılmış en kapsamlı araştırmalardan biri Dalkılıç ve Dalkılıç (2002) tarafından gerçekleştirilen ve basılı Türkçenin bazı istatistiklerini çıkaran çalışmadır. Çalışma kapsamında, tekli, ikili, ..., beşli harf gruplarının sıklık dağılımı, ilk-son harf çözümlemeleri, harf başına belirsizlik (entropi) ve fazlalık, rasgelelik endeksi, sözcük uzunluk dağılımı, sesli-sessiz harf oranı içeren bazı istatistik değerleri belirlenmiştir. Araştırma bir gazetenin internet arşivinden elde edilen derlem (*ing. corpus*) üzerinde gerçekleştirilmiştir.

Bu bölümde bahsedilebilecek diğer bir konu ise diğer dillerin büyük bir kısmı için incelemeleri yapılmış Zipf kanunlarıdır (Zipf, 1949). Türkçe’de konu ile ilgili çok fazla araştırma yapılmamıştır. Zipf kanunları ile ilgili olarak Türkçe için bilinen tamamlanmış ilk çalışma Dinçer (2004) tarafından yapılandır. Dinçer çalışmasında Zipf kanunları ile ilgili temel istatistikler çıkarmıştır. Bir diğer çalışmada ise Dalkılıç ve Çebi (2004) bir Türkçe derlemi olan TurCo’yu kullanarak Zipf kanunlarının Türkçe için geçerliliği ve Mandelbrot sabitini bulmuştur. Bunun dışında Karaoğlu ve Dinçer (2008)’nin çalışmasında Zipf kanunları esasında Türkçe yazılı derlemlerin değerlendirilmesini hedef alan bir çalışma yapılmıştır.

3.3. Uygulamaya Yönelik Çalışmalar

Türkçe, diğer dillere özellikle de İngilizce’ye göre üzerinde az çalışılmış bir dil olduğundan yapılan çalışmalar daha çok temel nitelikte kalmaktadır. Sayıları çok fazla olmamasına rağmen Türkçe için uygulamaya yönelik çalışmalardan bu bölümde bahsedilecektir.

Darcan' ın (1991) yüksek lisans tez çalışması anlamsal araştırmalar açısından önemli bir örnektir. Yapılan uygulamada kullanıcı tarafından girilen Türkçe sorgular sistem tarafından ara gösterim diline çevrildikten sonra SQL diline çevrilir. Bu SQL sorgusu kullanılarak veri tabanı sorgulanır. Alınan sonuçlar analiz edilir ve kullanıcıya geri çevrilir. Türkçe sorguları SQL sorgusuna çeviren ve cevabını veren benzer bir çalışma da NALAN-TS adındaki çalışmadır (Maden ve Ark., 2003).

Çetinoğlu'nun 2001 yılındaki tez çalışmasında TOY (Türkçe Okur Yazar) isimli bir uygulama geliştirmiştir. TOY, insan makine iletişim programı olarak tanımlanmaktadır. Uygulamada kullanıcı Türkçe cümleler ile soru sorabilmekte veya yeni bilgiler vererek bilgisayar ile etkileşime girebilmektedir. Çalışma DDİ' nin biçimbirimsel, sözdizimsel ve anlamsal seviyelerini kapsamaktadır. Biçimbirimsel seviyede çift yönlü bir sistemle kelimeler hem çözümlenip hem de üretilmektedir. Sözdizimsel seviyede Türkçe cümle yapıları öbek yapısı kuralları şeklinde tanımlanırken anlamsal seviyede ise sözdizimsel öğeler *prolog* yüklemeleri olarak tanımlanmıştır. Böylelikle cümle sözdizimsel olarak çözüldüğünde anlamsal gösterim de ortaya çıkmaktadır. Çetinoğlu'nun Ofłazer'in çalışmasını temel alan biçimbirimsel çözümleyicisi ve Türkçe fiil çekim programı gibi uygulamaları da bulunmaktadır.

Türkçenin anlamsal yapısı ile uğraşan diğer bir çalışma ise ALİ (Aritmetik Lisan İşleyici)' dir (Say,1998). ALİ, ilköğretim seviyesinde Türkçe yazılmış aritmetik problemlerinin çözümlerini hesaplayıp cevabı yine Türkçe olarak veren bir çalışmadır. Uygulama kullanıcı tarafından girilen problemi kelime bazında morfolojik analize tabi tutar ve kelimenin varsa birden fazla çözümlemesi uygulamanın sözdizimsel analiz yapan kısmına gönderilir. Sözdizimsel analiz kısmı tüm hedef problemleri kapsayan basit cümlelerin bulunduğu küme kullanılarak cümleyi analiz eder. Bu analiz sona erdiğinde anlamsal analiz kısmı

cümleleri tanımlı şablonlarla eşleme yöntemini kullanarak çözümler. Anlamsal olarak da çözümlenen problem matematiksel olarak çözüme ulaştırılıp kullanıcıya sunulur.

Doğal dil türetme alanında is Cebiroğlu ve arkadaşlarının 2003 yılında yaptığı çalışmayı örnek olarak gösterebiliriz. Çalışma kapsamında Türkçenin dilbilgisi kuralları kullanılarak geliştirilen bir bilgisayar yazılımı ile olası sözcük kökleri üretilmiştir. Çalışmada Oflazer'in çalışmasında kullandığı sözlük ile üretilen köklerin Türkçede kullanılıp kullanılmadığı kontrol edilmiştir.

NATO destekli, Türkçe doğal dil işleme çalışmasında ise (*ing.* The Turkish Natural Language Processing Initiative, TNLP) (Bozşahin, Oflazer, 1994), Türkçe için DDİ uygulamaları geliştirilmeye ve Türkçe metinlerin hesaplamalı analizinin yapılmasına çalışılmıştır. Bu projede Bilkent Üniversitesi Bilgisayar mühendisliği bölümü, ODTÜ Bilgisayar mühendisliği bölümü ve Halıcı yazılım şirketi yer almış ve proje kapsamında birçok dil işleme ürünü ortaya çıkarılmıştır.

TNLP projesinin uzantısı olarak ortaya çıkan uygulamalardan biri Online Türkçe Biçimbirimsel analizcisidir (*ing.* Online Turkish Morphological Analyser, TMA). Uygulamada kullanıcının web üzerinden girdiği Türkçe kelimeler analiz edilir ve kullanıcıya geri verilir. Verilen bu listede tüm olası çıktılar bulunmaktadır. Bu da uygulamanın cümleleri girdi olarak alıp kelime kelime analizine izin vermektedir.

TNLP projesinin uzantısında geliştirilen bir başka uygulama ise Türkçe yazım hatası düzelticisidir (Oflazer, 1996) (*ing.* Turkish Spelling Corrector). Bu uygulamada kullanıcının girdiği kelime eğer hatalıysa kelimenin doğru olabilecek örnekleri gösterilir. Bu düzeltme işlemi kelimede yazılan eksik bir harfi tamamlamak, fazla yazılan harfi silmek,

yanlış yazılan harfi deęiřtirmek ve pozisyonu yanlış olan harfin yerini deęiřtirmekten ibarettir.

Doęal dil iřleme konularıyla da ilgilenen enformasyon bilim altında yer alan bilgi geri getirim uygulamalarına örnek olarak Tür' ün (2000) doktora tezini gösterebiliriz. Bu tezde istatistiksel dil iřleme yöntemleri kullanılarak Türkçe metinlerden bilgi çıkarımı üzerine yapılan bir dizi çalışma mevcuttur. Tür tezinde kullandığı istatistiksel yöntemlerle dil iřlemenin birçok problemine (Türkçe metin düzenleme, sözcüklere ayırma, cümlelere ayırma, konulara ayırma vb.) çözüm önerileri sunmaktadır.

Dinçer (2004) doktora tezinde Türkçe dilbilgisel özelliklerin göz önünde bulundurularak istatistiksel verileri ve yöntemleri kullanmış ve bilgi geri getirim sistemlerinin başarısında artış sağlamıştır. Bilgi geri getirim sisteminin performansını arttırmak amacıyla doęal dil iřleme sahasına giren yöntemler de (Türkçe yazılı belgelerin tespiti, cümle sonlarının tespiti, kelime gövdeleme, kelimelerin sözcük türü tespiti) bu tez kapsamında geliştirilmiştir.

Alanında önemli örnek çalışmalardan biri olan WordNet, (Princeton WordNet Web Page) farklı sözdizimsel kategorilerdeki kelimelerin benzer kümeler (*ing.* synsets) içerisinde organize edildiğı ve farklı anlamsal ilişkiler kullanılarak bağlandığı sözcüksel referans sistemidir. Bu sistem tasarımıda insanın sözcüksel hafızası temel alınmıştır. Tüm dünya üzerinde birçok arařtırmacı Türkçe, İngilizce vb. farklı diller için wordNets sürecini geliřtirmektedir. Kullanıcılar sistemin online versiyonunu kullanarak istedikleri dilin kelime taksonomisini ve kelimeler arası ilişkilerini kolayca öğrenebilmektedirler. Balkan dilleri için de çok dilli sözcük veritabanı geliřtirmek amacı ile Balkan WordNet (BalkaNet) projesi yürütölmektedir. Bu projenin bir parçası olarak

Türkçe için yapılan çalışmalar da Sabancı Üniversitesi tarafından sürdürülmektedir (Turkish WordNet Web Page).

4. Sözcük Türü Tespit Yöntemi

Sözcük türü tespiti işlemi temel olarak kelimelerin cümle içerisindeki diziliminde, hangi sözcük türünden olduğunun belirlenmesi olarak tanımlanabilir. Daha önceki bölümlerde de bahsedildiği gibi Türkçe’de kelimeler sözcük türlerine göre: *isim* (ing. noun), *özel isim* (ing. proper noun), *fiil* (ing. verb), *sıfat* (ing. adjective), *zarf/belirteç* (ing. adverb), *zamir/adıl* (ing. pronoun), *bağlaç/rabıt* (ing. conjunction), *soru* (ing. Question) ve *edat/ilgeç* (ing. preposition) şeklinde sınıflandırılır. Bir sözcüğün türünün belirlenmesi kelimenin cümle içerisindeki yerine bağlı olarak değişebilmektedir. Örneğin:

“Gelecek kişiler bizim için çok önemlidir.”

“Gelecek şimdiden düşünülmelidir.”

“Misafirimiz bu akşam gelecek.”

cümlelerinde “gelecek” kelimesi yukarıda verilen üç cümlede de farklı sözcük türüne sahip olmaktadır. İlk cümlede sıfat görevindeyken ikinci cümlede isim, son cümlede ise fiil görevindedir.

Bu alanda yapılan ilk çalışmalar İngilizce için oluşturulmuş ilk derlem olan Brown derlemi (Kucera, Frances, 1967) üzerinde gerçekleştirilmiştir. Alanda yapılan çalışmalara genel olarak bakıldığında teorik açıdan iki ana grup etrafında toplanmaktadır: kural tabanlı ve istatistiksel. Sözcük türü tespiti alanında yapılan istatistikî çalışmalar genellikle bir derlem kullanılarak gerçekleştirilmektedir.

Kural tabanlı çalışmalar 1960’lı yıllarda başlarken istatistiksel çalışmaların 1980’li yılların ortalarında yoğunlaştığı görülmektedir. Kural tabanlı çalışmaların öncüsü olarak Klein ve Simmons (1963) görülmektedir. Bu çalışmayı takiben gelen önemli çalışmalardan biri de Greene ve Rubin (1971)’in gerçekleştirdikleridir. TAGGIT ismi ile

anılan program %71 oranında bir başarı ile Brown derlemini etiketlemiştir. Bir başka önemli çalışmada ise Hindle (1989) kelimelerdeki belirsizliğin giderilmesi ve sözcük türü tespiti için kural tabanlı bir yöntem geliştirmiştir. Bahsedeceğimiz son kural tabanlı çalışma ise istatistiksel çalışmaların oldukça yoğunlaştığı yıllarda Brill (1995) tarafından gerçekleştirilmiştir. Brill çalışmasındaki başarı oranını %95-97 olarak rapor etmiştir.

Türkçe üzerinde özellikle sözcük türü tespiti çalışmalarının sayısı oldukça az olmasına rağmen yüksek başarı oranları elde edilmiştir. Oflazer ve Kuruöz (1994) ve Oflazer ve Tür'ün (1996, 1997) çalışmaları kural tabanlı sözcük türü tespiti çalışmalarının önemli örneklerindendir.

İstatistiksel çalışmalara bakıldığında ise birçoğunun Gizli Markov Modellerini (GMM- Hidden Markov Models- HMMs) kullandığı görülmektedir. GMM, belirli bir derlem içerisinde kelimelerin sıralanış durumlarından elde edilen olasılıkları kullanmaktadır. Bu olasılıklar sayesinde bir kelime ile karşılaştığımızda o kelimeden sonra gelebilecek kelime türünün ne olabileceği istatistiksel olarak tahmin edilebilmektedir. Bu işlem bir cümle içerisinde yer alan tüm kelime çiftleri için gerçekleştirildiğinde oluşturulabilecek sözcük türü çözüm kümelerinde en yüksek olasılığa sahip zincirin seçilmesi işlemi GMM'inin temelini oluşturmaktadır. GMM'inde kullanılan istatistikler sadece kelime çiftlerinden elde edilmeyebilir. Bu verilerde kelimelerin üçlü, dörtlü veya daha fazla sayıda ard arda gelme durumlarını kullanılırsa buna “yüksek dereceli” Gizli Markov Modelleri ismi verilmektedir. Tez kapsamında önerdiğimiz yönteminde temelinde yer alan GMM hakkında daha geniş bilgi ve matematiksel temel ilerleyen kısımlarda verilmektedir.

GMM'inin sözcük türü tespiti çalışmaları üzerine uygulandığı ilk olarak Garside (1987) tarafından gerçekleştirilmiştir. CLAWS (the Constituent Likelihood Automatic Word-tagging System) adı verilen ve

günümüzde UCREL (*ing.* University Centre for Computer Corpus Research on Language) çalışma grubu tarafından çalışmaları sürdürülen sistemin Lancaster-Oslo-Bergen derlemi üzerindeki başarı oranı %96-97 olarak belirtilmektedir (<http://ucrel.lancs.ac.uk/claws>).

Markov modelleri içerisinde kelime sıralanış olasılıklarının yanı sıra kelime görülme olasılıklarının da kullanılması ilk olarak DeRose (1988) ve Church (1988) tarafından gerçekleştirilmiştir. Bu çalışmalarda dinamik programlama algoritması olan Viterbi algoritması kullanılmıştır (Viterbi, 1967, Forney, 1973).

Sözcük türü tespit çalışmalarında kural tabanlı ve ya istatistiksel olarak başarı oranları %90-96 aralığında değişmektedir. Bu başarı oranlarını yakalayan diğer belli başlı algoritmalar ise şunlardır: Viterbi algoritması (Gizli veya görünür markov modellerinde kullanılmaktadır.), Brill Tagger (Brill, 1992), Constraint Grammar (Karlsson, 1990, Karlsson ve ark. 1995)

Türkçe için istatistiğe dayalı sözcük türü tespiti çalışmalarına örnek olarak ise GMM'inin kullanıldığı Tür ve arkadaşlarının (2002) ve Dinçer'in (2004) yaptığı çalışmaları verebiliriz.

4.1. Gizli Markov Modeli

Markov zincirleri olarakta bilinen Markov modelleri istatistik sahasının en fazla kullanılan kavramlarından birisidir. Andrey Markov tarafından ortaya konan model temelde stokastik (rastsal) bir süreçtir. Stokastik süreç (stochastic process) zaman indeksli bir rastsal değişkenler dizisidir ve birbirinden bağımsız olasılık değişkenlerinden (X_t) meydana gelmektedir ($t=0,1,2,\dots$). Bir stokastik sürecin Markov zinciri olabilmesi için Markov özelliğinin olması gerekmektedir. Markov özelliği, şimdiki durum verildiğinde gelecekteki durumun geçmişten bağımsız olmasını

gerektirir. İstatistikte kullanılan markov modelleri görünür ve gizli olmak üzere iki çeşittir.

Gizli Markov modeli, stokastik süreçleri modelleyebilen sonlu durum ağlarıdır. Gizli Markov modelinde durum uzayı, çıktı uzayı ve durumlar arası geçişler ele alınmaktadır. Bu model, çıktı uzayındaki elemanlar ile durum uzayındaki elemanlar ve bunların geçişlerinden oluşan uygun stokastik zincirin eşleşmesini amaçlamaktadır. Görünür Markov modelinde ise, durum uzayı ve çıktı uzayının belli olmasının yanında elemanlarının birebir şekilde eşlenmeleri de bellidir.

Markov modeli parametrelerinin üzerinde çalıştığımız sözcük tespit problemindeki karşılıklarını ise şu şekilde belirleyebiliriz:

- ✓ Çıktı uzayı: kelimeler
- ✓ Durum uzayı: sözcük türleri
- ✓ Stokastik süreç: cümle

Burada Markov modelinin görünür olması cümlelerin ve cümle içerisinde bulunan kelimelerin sözcük türlerinin belirli olma durumudur. Markov modelinin gizli olma durumunda ise, cümlelerin ve sözcük türleri kümesinin belli fakat cümle içerisindeki kelimelerin hangi sözcük türünde oldukları belli değildir. Gizli Markov modeli bu eşleşmeyi yaparak olasılığı en yüksek zincirin belirlenmesi prensibine dayanarak cümledeki kelimelerin sözcük türlerinin belirlenmesinde kullanılmaktadır.

Gizli Markov modelinin günümüzde birçok kullanım sahası bulunmaktadır: Konuşma tanıma, el yazısı tanıma, el hareketleri ve yüz

mimiklerini tanıma, sözcük türü belirleme, müzik sesleri tanıma, biyoinformatik, ekonomi, hava tahminleme vb.

Daha önce de bahsedildiği gibi bir sürecin stokastik süreci olarak tanımlanabilmesi için “kısıtlı ufuk” (*ing.* limited horizon) ve “zaman içinde değişmeme” (*ing.* Time invariance) gibi özellikleri sağlaması gerekmektedir.

Markov özelliği olarak da bilinen kısıtlı ufuk özelliği, şimdiki durum verildiğinde gelecekteki durumun geçmişten bağımsız olması anlamına gelmektedir. Başka bir deyişle kısıtlı ufuk, X_t olasılık değişkeninin, sadece kendisinden bir önceki (bu sayı daha fazla da olabilir.), yani X_{t-1} olasılık değişkeninin değerinden etkilendiği, daha öncekilerden etkilenmediği şeklinde ifade edilebilir. Bu özellik, “Markov modelinin derecesi” ifadesi ile belirtilmektedir. Yani, bir olasılık değişkeni sadece kendisinden bir önce gelen olasılık değişkeninin değerinden etkileniyorsa buna 1. dereceden Markov modeli denmektedir. Bu derecenin “n” olarak kabul edildiği Markov modelinde kısıtlı ufuk özelliği aşağıdaki şekilde ifade edilebilir.

$$P(X_t = i \mid X_{t-1}, X_{t-2}, \dots, X_0) = P(X_t = i \mid X_{t-1}, X_{t-2}, \dots, X_{t-n}) \quad (4.1)$$

Zaman içinde değişmeme özelliği ise kısıtlı ufuk özelliğinin her zaman diliminde geçerli olduğunu belirtmektedir. Bu özellik 1. dereceden Markov modeli için denklem 4.2 de verilmiştir.

$$P(X_t = i \mid X_{t-1} = j) = P(X_1 = i \mid X_0 = j) \quad (4.2)$$

4.2. Yöntem

Bu tez çalışmasında istatistiksel yöntemlere dayalı olarak geliştirilen yöntem gizli Markov modellerini kullanmaktadır. Bu nedenle yöntemin kurgusu açıklanmadan önce sözcük türü tespit problemi Markov süreci olarak tanımlanacaktır. Daha öncede belirtildiği gibi Markov modelinde kullanılan stokastik süreç sözcük türü tespiti

işleminde cümleye karşılık gelmektedir. “ w_i ” ($i:0...n$) kelimeleri temsil etmek üzere n kelimeden oluşan bir cümle $w_1, w_2, \dots, w_n$ şeklinde ifade edilmektedir. Cümledeki kelimelerin sözcük türleri de $t_1, t_2, \dots, t_n$ şeklinde kabul edilirse sözcük türü tespiti problemi bir cümlede bulunan kelime dizinine karşılık gelebilecek maksimum olasılıkta sözcük türü eşleşmesinin bulunmasıdır ve aşağıdaki şekilde ifade edilebilir.

$$\operatorname{argmax}_{t_1 \dots t_n} P(t_1 \dots t_n | w_1 \dots w_n) \quad (4.3)$$

Problemde cümlelerin, yani çıktı uzayının, ne olduğu belli fakat bu cümledeki kelimelerin sözcük türlerinin dizilişinin ne olduğu belli değildir. Denklem ile cümle $(w_1, \dots, w_n)$ için, olası tüm $t_1, \dots, t_n$ sözcük türü dizilimleri içerisinde en yüksek olan sözcük türü dizilimi seçilmesi hedeflenmektedir. Eşitlikte Bayes olasılık dönüşümü uygulandığında aşağıdaki belirtilen denklem elde edilmektedir.

$$\operatorname{argmax}_{t_1 \dots t_n} \frac{P(w_1 \dots w_n | t_1 \dots t_n) \cdot P(t_1 \dots t_n)}{P(w_1 \dots w_n)} \quad (4.4)$$

Eşitlikte yer alan $P(w_1 \dots w_n)$ olasılığı daima sabit olması nedeni ile denklemden çıkarıldığında problem tanımının son hali olan aşağıdaki denklem elde edilmektedir.

$$\operatorname{argmax}_{t_1 \dots t_n} P(w_1 \dots w_n | t_1 \dots t_n) \cdot P(t_1 \dots t_n) \quad (4.5)$$

Problemde, 1. dereceden ve 2. dereceden (1.dereceden geçmişe 1 adım) Markov modeli kullanılarak oluşturulan ifade sırası ile aşağıdaki şekilde belirlenmektedir.

1. dereceden Markov modeli,

$$\operatorname{argmax}_{t_1 \dots t_n} \prod_{i=1}^n P(w_i | t_i) \cdot P(t_i | t_{i-1}) \quad (4.6)$$

2. dereceden (1.dereceden geçmişe 1 adım) Markov modeli ,

$$\operatorname{argmax}_{t_1 \dots t_n} \prod_{i=1}^n P(w_i | t_i) \cdot P(t_i | t_{i-2}, t_{i-1}) \quad (4.7)$$

Bu denklemlerde verilen olasılıkların hesaplanması için,

$$\begin{array}{l} \text{Sözcük türünün verilen} \\ \text{kelime ile eşleşme} \\ \text{olasılığı} \end{array} \quad \hat{P}(w_i | t_i) = \frac{f(t_i, w_i)}{f(t_i)} \quad (4.8)$$

$$\begin{array}{l} \text{İki sözcük türünün ard} \\ \text{arda gelme olasılığı} \end{array} \quad \hat{P}(t_i | t_{i-1}) = \frac{f(t_{i-1}, t_i)}{f(t_{i-1})} \quad (4.9)$$

$$\begin{array}{l} \text{Üç sözcük türünün ard} \\ \text{arda gelme olasılığı} \end{array} \quad \hat{P}(t_i | t_{i-2}, t_{i-1}) = \frac{f(t_{i-2}, t_{i-1}, t_i)}{f(t_{i-2}, t_{i-1})} \quad (4.10)$$

ifadeleri kullanılmaktadır. Hesaplamalarda kullanılan üç sözcük türünün ard arda gelme ($f(t_{i-2}, t_{i-1}, t_i)$), iki sözcük türünün ard arda gelme ($f(t_{i-2}, t_{i-1})$), sözcük türünün $f(t_i)$, ve kelimenin belli bir sözcük türünde bulunmasının ($f(t_i, w_i)$) gözlenme sıklık değerleri talim için kullanılan derlemlerden elde edilmektedir. Sözcük türü tespiti probleminin Markov modelleri ile ifade edilmesi dinamik bir programlama tekniği olan Viterbi algoritmasının kullanımına yol açmaktadır.

Sözcük türü tespiti işleminde verilen denklemlerin hesaplaması sırasında derlemden bulunmayan bir kelime veya sözcük türü dizilimi için olasılık değerleri 0 olması ile karşımıza çıkmaktadır.

$$P(w_i | t_i) = 0, \quad P(t_i | t_{i-2}, t_{i-1}) = 0$$

Türkçe gibi, sadece bir kelimedenden yazım şekli farklı biçimlerin sayısı binleri bulabilecek bir dil için bu problemin çözümü oldukça büyük önem arz etmektedir. Çünkü yüzeysel biçim olarak Türkçenin kelime dağarcığının teorik olarak kapalı olmasından söz edilemez. Örnek olarak Çizelge 4.1’de 1 milyon ve 10 milyon kelimedenden oluşan iki derlemin sözcük dağarcıklarının sayıları verilmiştir (Hakkani-Tür, Oflazer and Tür, 2002).

Çizelge 4.1. Bir milyon ve on milyon kelimelik Türkçe derlemlerin kelime dağarcıkları (Hakkani-Tür ve ark., 2002).

Derlem boyutu	Kelime dağarcığı
1M kelime	106,547
10M kelime	417,775

Olasılığı “0” olan değerlerin olasılıklarının sıfırdan farklı bir hale getirme işlemine olasılık yumuşatma (ya da düzleme) yöntemleri ismi verilmektedir. Günümüze kadar bir çok yumuşatma yöntemi geliştirilmiştir. Bunlardan başlıcaları, “Laplace yumuşatma yöntemi” (bir ekleme yumuşatma yöntemi) (Lidstone, 1920), “Good-Turing yumuşatma yöntemi” (Good, 1953; church and Gale, 1991), “Linear interpolation yumuşatma yöntemi” (Jelinek-Mercer yumuşatma yöntemi) (Jelinek ve Mercer, 1980), “Katz Backoff yumuşatma yöntemi” (Katz, 1987), “Class-based yumuşatma yöntemi” (Resnik, 1993), “Absolute discounting yumuşatma yöntemi” (Simons et al., 1997) ve “Kneser-Ney yumuşatma yöntemi”dir (Ney, Kneser, 1994).

Yukarıda da belirtildiği gibi çalışmamızda verilen cümlede geçen kelimenin ve/veya sözcük türü geçişinin talim derleminde rastlanmaması (yani olasılık değerlerinin 0 olması) olmak üzere yumuşatma gerektiren iki durum vardır.

Geliştirilen yöntemde söz konusu iki problemin ortadan kaldırılabilmesi amacı ile farklı yumuşatma yöntemleri kullanılmaktadır.

Buna göre sözcük türü dizilimlerindeki talim derleminde görülmeyen geçiş olasılıkları için Kneser-Ney yumuşatma yöntemi kullanılmıştır. Bu işlem için gerekli ifadeler 2 dereceden GMM’inde kullanılacak geçiş olasılık hesabı üzerinden aşağıda sunulmuştur.

$$\begin{aligned} \hat{P}_{kn}(t_i | t_{i-2}, t_{i-1}) &= \frac{\max\{f(t_{i-2}, t_{i-1}, t_i) - D, 0\}}{\sum_{t_i} f(t_{i-2}, t_{i-1}, t_i)} \\ &+ \frac{D}{\sum_{t_i} f(t_{i-2}, t_{i-1}, t_i)} \cdot N_{1+}(t_{i-2}, t_{i-1}, \bullet) \\ &\times \frac{N_{1+}(\bullet, t_{i-1}, t_i)}{N_{1+}(\bullet, t_{i-1}, \bullet)} \end{aligned} \quad (4.11)$$

Eşitlikte yer alan $N_{1+}(t_{i-2}, t_{i-1}, \bullet)$, $N_{1+}(\bullet, t_{i-1}, t_i)$ ve $N_{1+}(\bullet, t_{i-1}, \bullet)$ ifadeleri aşağıdaki şekilde hesaplanmaktadır.

$$N_{1+}(t_{i-2}, t_{i-1}, \bullet) = \left| \left\{ t_i \mid f(t_{i-2}, t_{i-1}, t_i) \geq 1 \right\} \right| \quad (4.12)$$

$$N_{1+}(\bullet, t_{i-1}, t_i) = \left| \left\{ t_{i-2} \mid f(t_{i-2}, t_{i-1}, t_i) \geq 1 \right\} \right| \quad (4.13)$$

$$N_{1+}(\bullet, t_{i-1}, \bullet) = \left| \left\{ (t_{i-2}, t_i) \mid f(t_{i-2}, t_{i-1}, t_i) \geq 1 \right\} \right| \quad (4.14)$$

Yine eşitlik 4.11’de kullanılan D değerinin hesaplanması için de,

$$D = n_1 / (n_1 + 2n_2) \quad (4.15)$$

kullanılmaktadır. Buradaki n_1 ve n_2 değerleri toplam sayısı bir ve iki olan 3-gramların sayısı olmaktadır.

Diğer problem olan derlemde yer almayan kelimelerin kelime-etiket olasılıklarının sıfır değerini alması durumunda ise yumuşatma değeri yerine talim derleminden elde edilen istatistikler temel alınmıştır (Çizelge 4.2). Buna göre derlemde yer almayan bir kelime ile

karşılaştığında Çizelge 4.2’de verilen bulunan tüm olasılık değerleri kelime-etiket ($P(w_i|t_i)$) istatistiği olarak kullanılmaktadır.

Çizelge 4.2. Talim derleminden elde edilen sözcük türü oranları

Sözcük Türü	Olasılık değerleri	Sözcük Türü	Olasılık değerleri
İsim	0,4663	Belirtgeç	0,0338
Noktalama	0,18	Sayı	0,0268
Sıfat	0,0963	Edat	0,0173
Fiil	0,0789	Zamir	0,0162
Bağlaç	0,0396	Soru	0,0022
Zarf	0,0352	Ünlem	0,0005

4.2.1. Sözcük Türü Etiketleme Kurgusu

Literatür bölümünde de belirtildiği gibi Türkçe kelimelerin istatistiksel yöntemlerle sözcük türü tespiti çalışmalarının en bilineni ve en yüksek sonuç üreteni Tür ve arkadaşlarının (2002) yaptığı çalışmadır. Çalışmada Tür ve arkadaşları, Gizli Markov Modellerinde hesaplama birimi olarak kelimeleri kullanıp %93.95 lik bir başarı sağlamışlardır. Elde edilen bu başarı Türkçe için yüksek sayılabilecek bir başarıdır. Ancak, yöntem Türkçe sözcük dağarcının kapalı olduğu kabulü üzerine kurulmuştur. Oysaki Türkçenin sözcük dağarcığının kapalı kabul edilmesi doğru bir yaklaşım değildir. Tez kapsamında önerdiğimiz yöntemde ise HMM de hesaplama birimi olarak kelimeler yerine;

- ✓ kelimelerin son heceleri,
- ✓ kelimenin sonundan alınacak harf birlikleri (1 ile 7 harf),
- ✓ kelimenin başından alınacak harf birlikleri (1 ile 7 harf)

- ✓ kelimelerin başından ve sonundan alınan harf birlikleri,
- ✓ Kelimede bulunan çekim ekleri kullanılmaktadır.

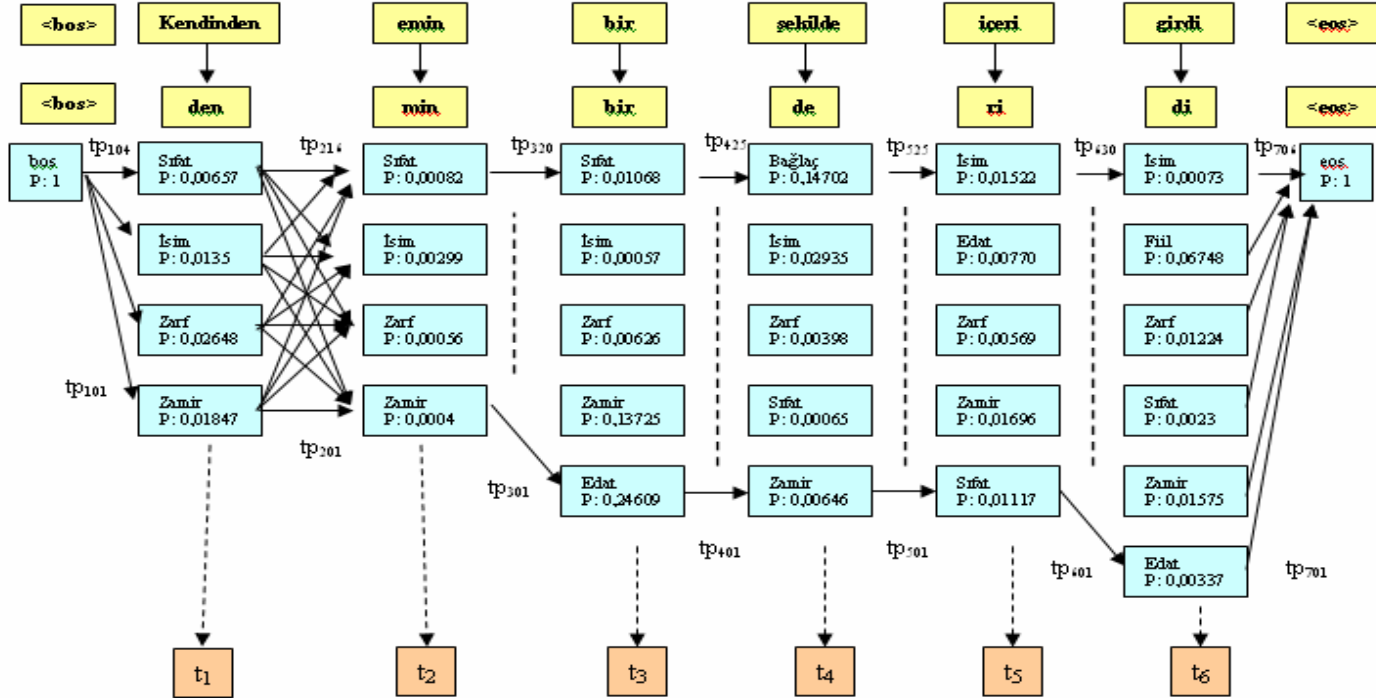
Böylece GMM kullanılan birimlerin kapalı bir sözlüğe sahip olmaları sağlanmıştır.

Önerilen sözcük türü tespit yöntemi daha önce yapılan çalışmalara göre istatistiklerin toplandığı hesaplama birimleri ve kullanılan yumuşatma yöntemleri ile farklılaşmaktadır. İlk olarak yöntemin kurgusu örnek bir cümle üzerinde anlatılacaktır. Örnekte hesaplama birimi olarak son heceyi kullanan 1. dereceden HMM kullanılmıştır.

Örnek Cümle: “Kendinden emin bir şekilde içeri girdi.”

Örnekte, cümle stokastik süreci, kelimelerin son heceleri (hesaplama birimi) çıktı uzayını, sözcük türleri dizisi ise durum uzayını oluşturmaktadır. Her bir hesaplama birimi mevcut tüm sözcük türleri ile etiketlenme olasılığına sahiptir. Böyle bir durumda 8 temel sözcük türü için 6 kelimedenden oluşan örneğimizde 6^8 adet sözcük türü dizisi (durum uzayı elemanı) mevcuttur. Hesaplanması gereken bu dizilerden hangisinin en yüksek olasılığa sahip olduğudur.

Örnekte yer alan “kendinden”, “emin”, “bir”, “şekilde”, “içeri” ve “girdi” kelimeleri cümle içinde kullanım sahasına bağlı olarak farklı dilbilimsel işlevlere sahip olabilmektedir. Örnek olarak gösterdiğimiz akışta hesaplama birimi olarak “kelimenin son hecesi” birimi kullanılmaktadır. Kelimelerin son hecelerinin elde edilebilmesi amacı ile Dinçer (2004) tarafından önerilen Türkçe kelimeler için heceleme algoritması kullanılmıştır. Cümlede yer alan kelimelerin son heceleri göz önüne alınarak problem incelendiğinde ise, emin sözcüğünün son hecesi olan “min” 4 farklı sözcük tipinde karşımıza çıkmaktadır: isim, sıfat, zarf ve zamir. Kurgu ile ilgili akış Şekil 4.1’de sunulmuştur.



Şekil 4.1. Sözcük türü tespit yönteminin akışı (Hesaplama birimi: son hece, Model: 1. dereceden GMM)

Yöntem kurgusunda ilk önce cümledeki tüm kelimelerin son hecelerine ait olası sözcük türleri, bu türlere ait hesaplama birimi-etiket ($p = P(w_i | t_i)$) ve geçiş istatistikleri ($tp_{ij} = P(t_i | t_{i-1})$) çıkarılmaktadır. Bu sayede cümle sözcük türü zincirleri de ortaya çıkmaktadır. Verilen örnekte 12000 ($4*4*5*5*5*6$) adet farklı sözcük türü zinciri bulunmaktadır. Gerçekleştirdiğimiz dinamik algoritma önceki bölümde anlatılan formülleri kullanarak, oluşan bu zincirlerden en yüksek olasılığa sahip zinciri ($t_1, t_2, t_3, \dots t_7$) bulup ortaya çıkarmaktadır.

4.3. Yöntemin Başarım Oranlarının Tespiti

4.3.1. Kullanılan Derlemler

Önerilen çözüm yöntemlerin başarımının tespitinde dilbilimsel özellikleri işaretlenmiş bir derleme ihtiyaç duyulmuştur. Türkçe’de üzerinde dilbilim çalışmalarının yürütüldüğü az sayıda derlem bulunmaktadır. Bunlardan en yaygın kullanılanları Bilkent Türkçe Derlemi, ODTÜ-Sabancı Türkçe derlemi ve ODTÜ Türkçe Derlemi derlemleridir. Bilkent Türkçe Derlemi ve ODTÜ-Sabancı Türkçe derleminin dilbilimsel özellikleri işaretlenmişken, ODTÜ Türkçe derleminin dilbilimsel özellik işaretlenmiş değildir. ODTÜ Türkçe derlemi boyut olarak en büyük derlem olmasına rağmen sadece derlem içindeki belgeler birbirinden ayrıştırılmış ve her belgenin genel yapısal özellikleri işaretlenmiştir. ODTÜ-Sabancı Türkçe derlemi, ODTÜ Türkçe derleminden seçilen bazı belgeler ve bu belgelerin içerisinden seçilen bazı paragraflardan oluşturulmuştur. ODTÜ-Sabancı Türkçe derlemi ve Bilkent Türkçe derlemi dilbilimsel olarak işaretli olmalarına rağmen aralarındaki en önemli fark ODTÜ-Sabancı Türkçe derleminin elle Bilkent Türkçe derleminin otomatik olarak işaretlenmiş olmasıdır. Araştırma ODTÜ-Sabancı ve Bilkent Türkçe derlemleri üzerinde yapılmıştır.

ODTÜ-Sabancı derlemi içerisinde yaklaşık olarak 64.900 kelime bulunmaktadır. Bu kelimelerin 51.209' u harf birliği iken 13.700' ü rakam ve noktalama birlikleridir. Ayrıca ODTÜ-Sabancı derleminde 17.487 farklı yüzeysel form, 5.201 farklı gövde, 7.462 cümle, 2,457 paragraf, 33 belge bulunmaktadır. Bilkent Türkçe derleminde ise yaklaşık olarak 712.100 kelime bulunmaktadır. Bu kelimelerin 587.200' u harf birliği iken 124.800' ü rakam ve noktalama birlikleridir. Bunun yanında Bilkent Türkçe derleminde 94.230 farklı yüzeysel form, 20.268 farklı gövde, 48.070 cümle, 2.283 belge bulunmaktadır (Dinçer, 2004).

Yapılan çalışmalarda ODTÜ-Sabancı ve Bilkent Türkçe derlemlerinden elde edilen alt derlemler kullanılmıştır. Talim derlemi boyutunun başarıya olan etkisinin gösterilmesi amacı ile Bilkent Türkçe derleminde boyutu 250, 500, 750, 1.000, 1.250, 1.500, 1.750, 2.000, 2.500, 3.000, 3.500, 4.000, 4.500, 5.000, 7.500, 10.000, 12.500, 15.000, 17.500, 20.000, 22.500, 25.000, 27.500, 30.000, 32.500, 35.000, 37.500, 40.000, 42.500 ve 45.000 cümle olan 30 adet alt derlem seçilmiştir. Deneylerde bu alt derlemlerden elde edilen istatistikler kullanılmaktadır.

Yöntemin başarı oranının tespit edilmesi amacı "bootstrapping" yöntemi kullanılmıştır. Bu yöntemle göre çalışma evreninin dağılımı, evrenden rasgele seçilen belirli sayıda örneklemin dağılımlarının ortalaması ile belirlenmektedir. Bu amaç doğrultusunda deney derlemi olarak ODTÜ-Sabancı derleminde boyutu 1000 cümle olan 30 adet alt derlem rastsal bir biçimde seçilmiştir. Elde edilen başarı oranları bu 30 derlemin başarı ortalamalarıdır.

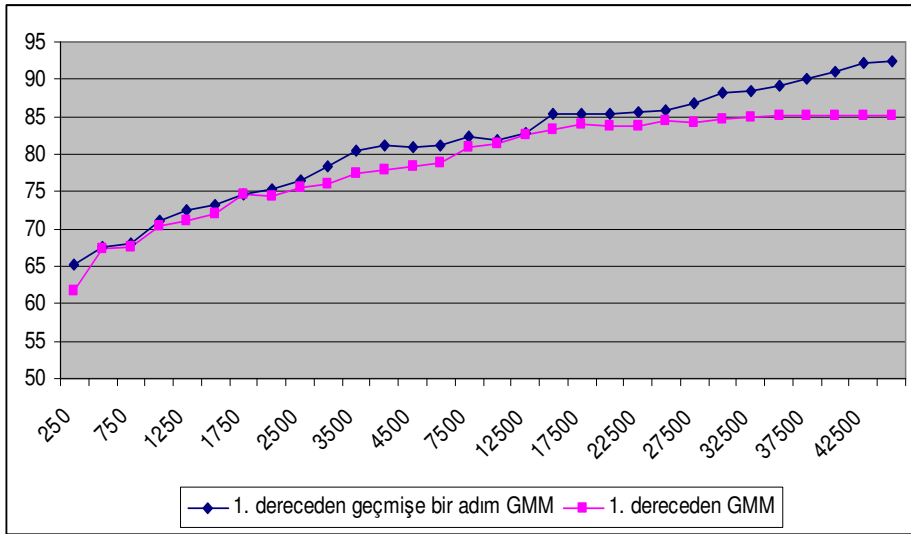
Gerçekleştirilen deney sayısı 25 farklı hesaplama birimi ve 2 farklı metod kullandığımız göz önünde bulundurulursa 45.000 ($2 \times 25 \times 30 \times 30$)' dir. Başarı oranları, doğru olarak tespit edilen sözcük türü sayısının, tüm kelime sayısına bölünmesi ile elde edilmiştir.

$$\text{Başarı Oranı} = \frac{\text{Dogru _ Etiketlenen _ Kelime _ Sayisi}}{\text{Toplam _ Kelime _ Sayisi}} \quad (4.16)$$

4.3.2. Deneysel Sonuçlar

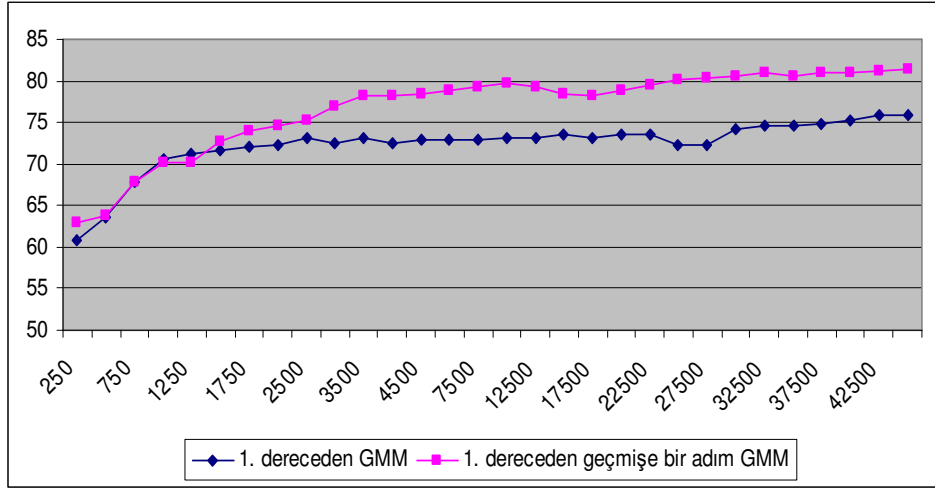
Deneylerde hesaplama birimi olarak alınan kelime, son hece, kelime sonundan alınan harf birlikleri, kelime başından alınan harf birlikleri, kelime başından ve sonundan alınan harf birlikleri ve çekim ekleri kullanılarak elde edilen istatistikler 1. ve 2. derece Markov sürecine sokularak sonuçlar değerlendirilmiştir. Tüm deney sonuçları ek 3'te tablolar halinde verilmiştir.

Hesaplama birimi olarak kelimenin alındığı deney sonuçları grafiksel olarak Şekil 4.2'de verilmektedir. Sonuçlardan da görüldüğü üzere, talim derleminin tüm boyutları için, en yüksek başarımlar 1. dereceden GMM için %85,3; 2. derece GMM için ise %92,4 olmuştur.



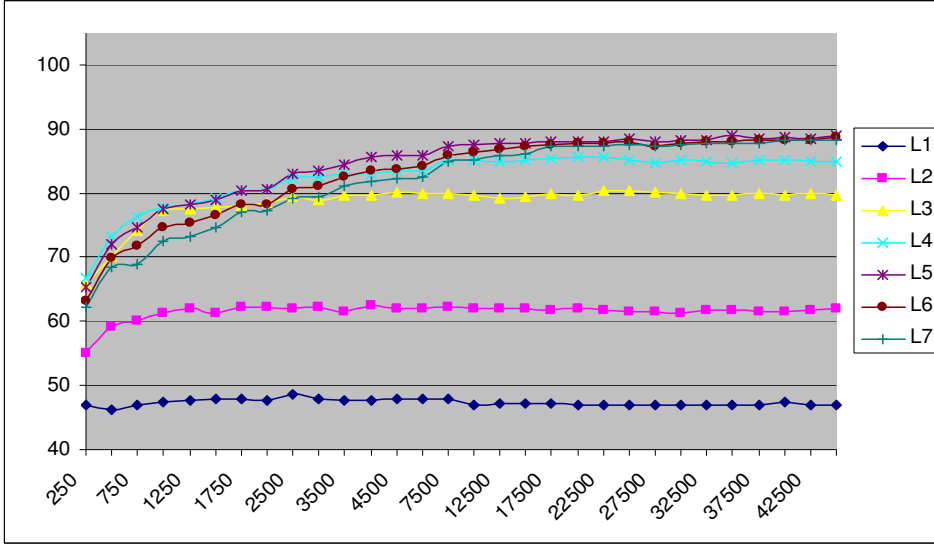
Şekil 4.2. Hesaplama birimi kelime alındığında elde edilen başarımlar grafiği

Hesaplama birimi son hece olarak alındığında talim derleminin tüm boyutları için, en yüksek başarımlar 1. dereceden GMM için %75,9; 2. derece GMM için ise %81,4 olmuştur.

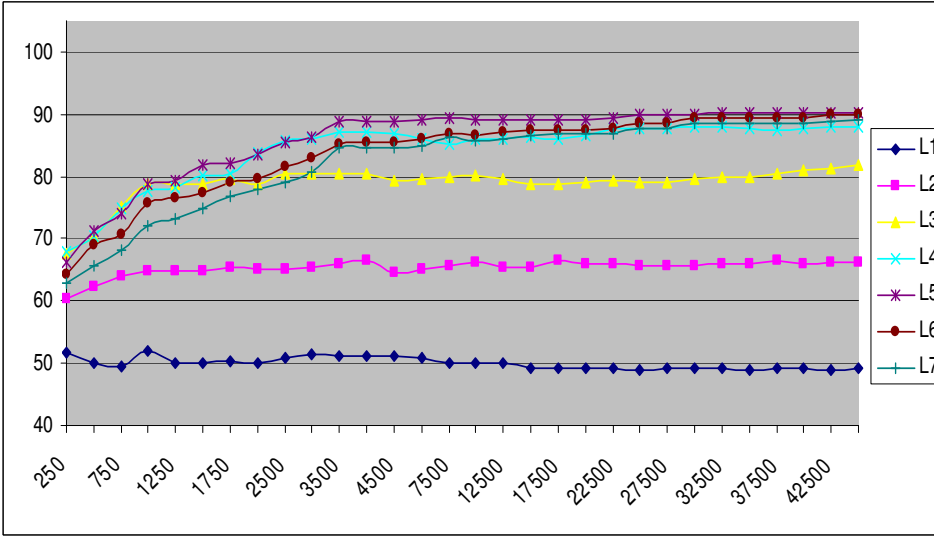


Şekil 4.3. Hesaplama birimi son hece alındığında elde edilen başarımların grafiği

Kelime sonundan harf birlikleri (1, 2 ,... 7 adet harf) alınarak yapılan deneyde talim derleminin tüm boyutları için, en yüksek başarımların değerleri (kelimenin son 5 harfi kullanılarak) 1. dereceden GMM için %88,9, 2. derece GMM için ise %90,2 olmuştur. Benzer çalışma Dinçer (2005), tarafından da yapılmış ve en yüksek başarımların değerleri 1. dereceden GMM için %60-70 arasında, 1. dereceden geçmişe bir adım GMM için ise %85,9 (kelimenin son 4 harfi kullanılarak) bulunmuştur. Çalışmamızın başarı oranının Dinçer' in çalışmasından daha yüksek çıkmasının nedenini kullandığımız yumuşatma (smoothing) yöntemine bağlayabiliriz.



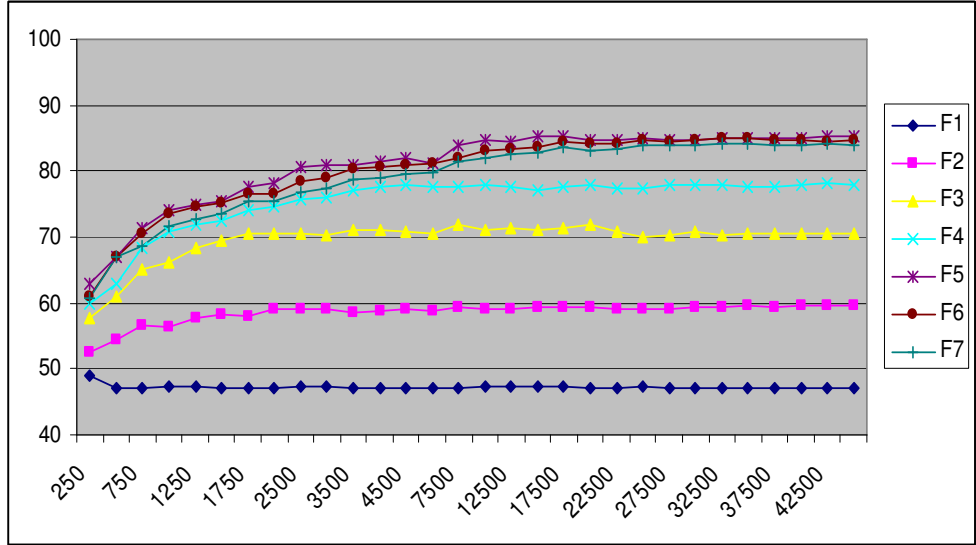
Şekil 4.4. 1. dereceden GMM ile hesaplama birimi kelimenin sonundan alınan harf birlikleri olduğunda elde edilen başarımlar grafiği



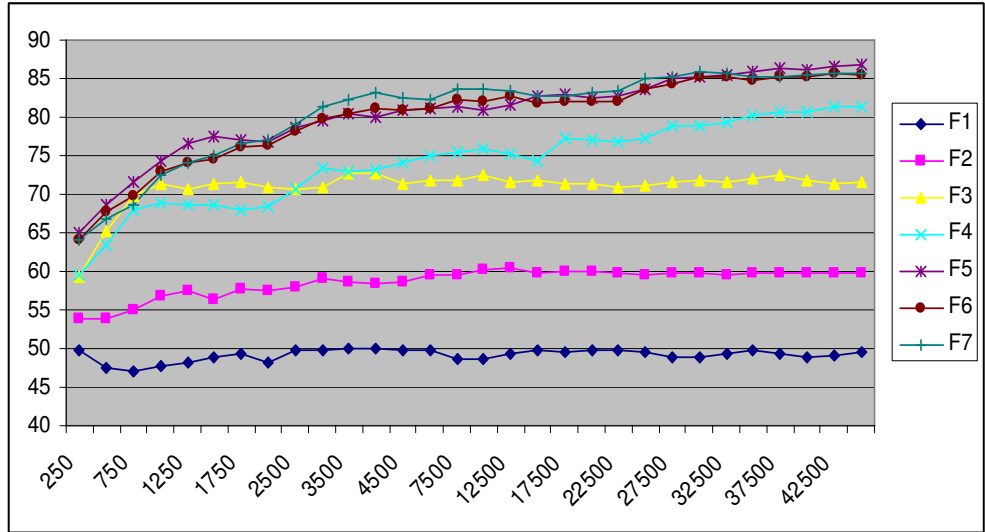
Şekil 4.5. 1. dereceden geçmişe bir adım GMM ile hesaplama birimi kelimenin sonundan alınan harf birlikleri olduğunda elde edilen başarımlar grafiği

Hesaplama birimi olarak kelimenin başından alınan harf birlikleri (1, 2 ,... 7 adet harf) kullanıldığında talim derleminin tüm değerleri için

en yüksek başarımlar değerleri (kelimenin son 5 harfi kullanılarak) 1. dereceden GMM için %85,3, 2. derece GMM için ise %86,8 olmuştur.



Şekil 4.6. 1. dereceden GMM ile hesaplama birimi kelimenin başından alınan harf birlikleri olduğunda elde edilen başarımlar grafiği

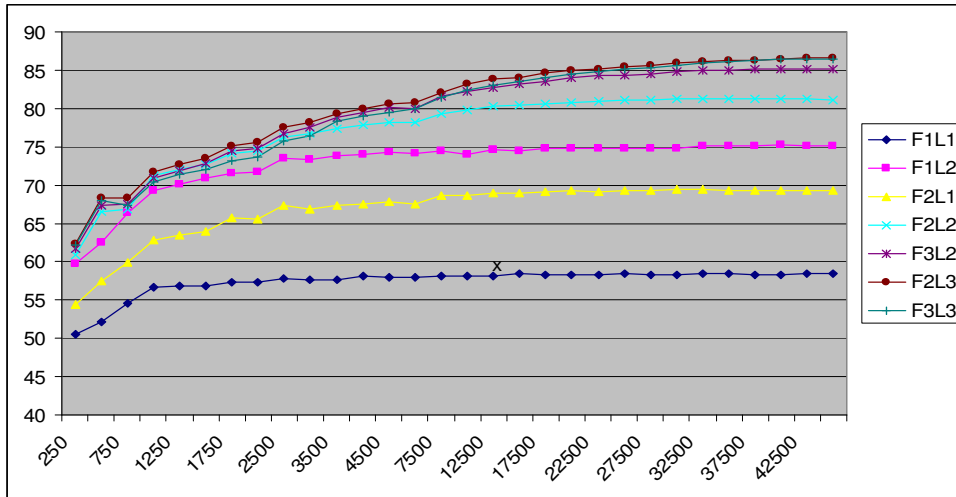


Şekil 4.7. 1. dereceden geçmişe bir adım GMM ile hesaplama birimi kelimenin başından alınan harf birlikleri olduğunda elde edilen başarımlar grafiği

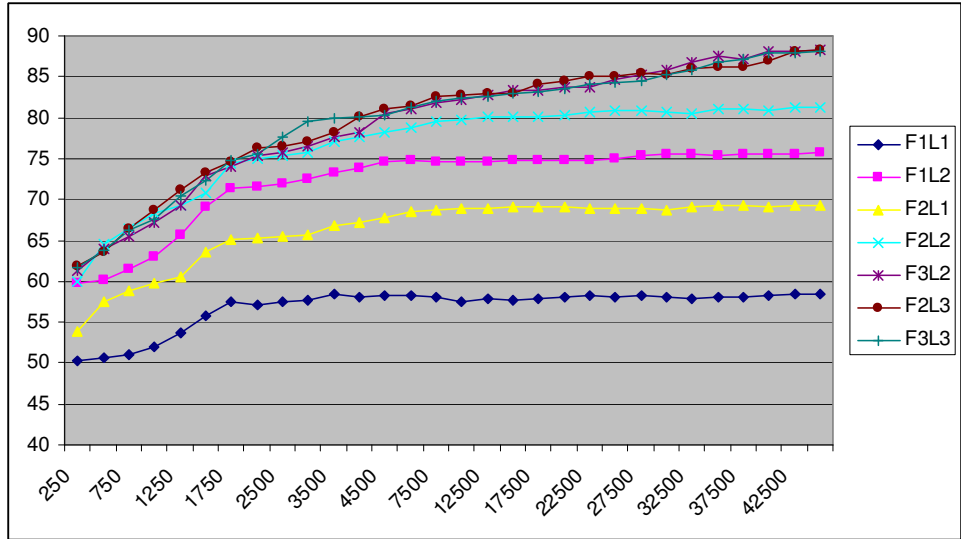
Kelimenin başından ve sonundan aşağıda belirtilen harf birlikleri kullanılarak yapılan deney sonuçları Şekil 4.8 ve Şekil 4.9’ da verilmiştir.

- ✓ kelime başından 1 harf, kelime sonundan 1 harf (F1L1)
- ✓ kelime başından 2 harf, kelime sonundan 1 harf (F2L1)
- ✓ kelime başından 1 harf, kelime sonundan 2 harf (F1L2)
- ✓ kelime başından 2 harf, kelime sonundan 2 harf (F2L2)
- ✓ kelime başından 3 harf, kelime sonundan 1 harf (F3L1)
- ✓ kelime başından 1 harf, kelime sonundan 3 harf (F1L3)
- ✓ kelime başından 3 harf, kelime sonundan 2 harf (F3L2)
- ✓ kelime başından 2 harf, kelime sonundan 3 harf (F2L3)
- ✓ kelime başından 3 harf, kelime sonundan 3 harf (F3L3)

Talim derleminin tüm boyutları için, en yüksek başarımların değerleri (kelimenin başından 2, sonundan 3 harf) 1. dereceden GMM için %86,5, 2. derece GMM için ise %88,3 olmuştur.

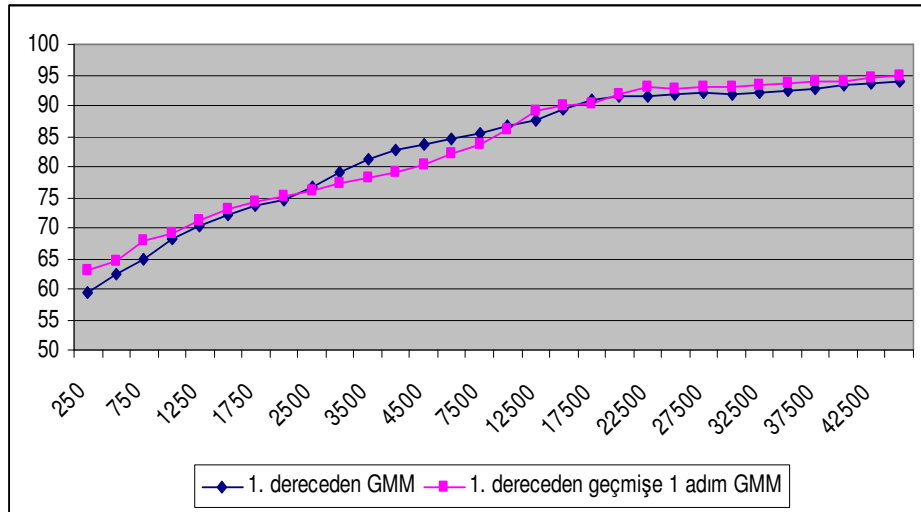


Şekil 4.8. 1. dereceden GMM ile hesaplama birimi kelimenin başından ve sonundan alınan harf birlikleri olduğunda elde edilen başarımların grafiği



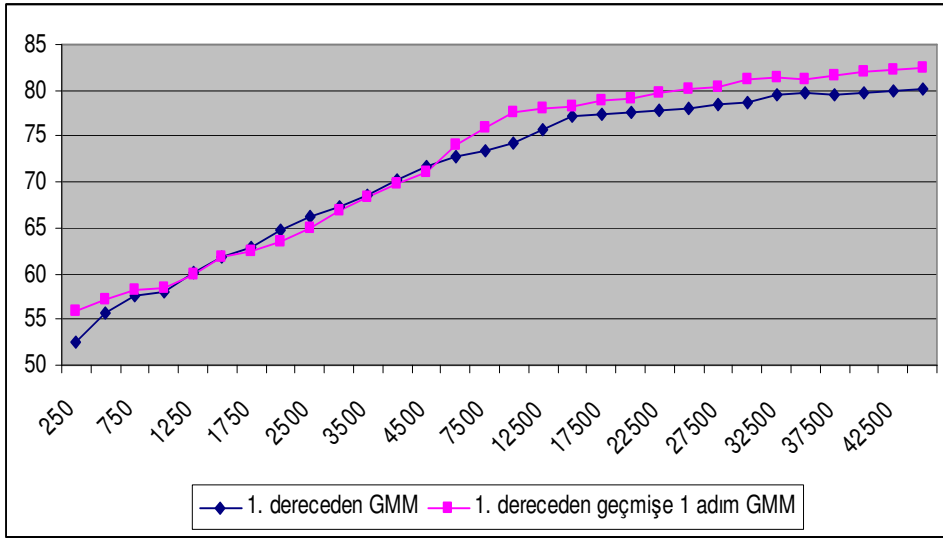
Şekil 4.9. 1. dereceden geçmişe bir adım GMM ile hesaplama birimi kelimenin başından ve sonundan alınan harf birlikleri olduğunda elde edilen başarıım grafiği

Hesaplama birimi olarak gövdenin alınması ile yapılan deneylerden elde edilen sonuçlar, Şekil 4.10'da verilmektedir. En yüksek başarıım değerleri 1. dereceden GMM için %93,9, 2. derece GMM için ise %94,9, olmuştur. Sonuçlardan da görüleceği üzere hesaplama birimi olarak gövdenin kullanılması başarıyı arttırmıştır.



Şekil 4.10. Hesaplama birimi kelime alındığında elde edilen başarıım grafiği

Deneyde hesaplama birimi olarak “kelimenin aldığı çekim ekleri” kullanılmıştır (Şekil 4.11). Talim derleminin tüm boyutları için, en yüksek başarımlar 1. dereceden GMM için %80,07; 2. derece GMM için ise %82,47 olmuştur. Görüldüğü üzere başarımlar beklendiği kadar yüksek elde edilmemiştir. Bunun en önemli nedeni olarak birçok yüzeysel formun çekim eki almaması gösterilebilir.



Şekil 4.11. Hesaplama birimi çekim ekleri alındığında elde edilen başarımlar grafiği

Sonuç olarak tüm hesaplama birimlerinde elde edilen en yüksek başarımların oranları Çizelge 4.3’ de özetlenmektedir.

Çizelge 4.3. Tüm hesaplama birimlerinden elde edilen en yüksek başarımların değerleri

Hesaplama Birimi	En Yüksek Başarımların Değerleri (%)	
	1. derece GMM	2. Derece GMM
Kelime	85,28	92,38
Son Hece	75,98	81,39
Son 1 harf	48,67	51,19
Son 2 harf	62,56	66,47
Son 3 harf	80,43	81,91
Son 4 harf	85,75	87,97
Son 5 harf	88,95	90,23
Son 6 harf	88,77	89,82
Son 7 harf	88,29	89,06
İlk 1 harf	47,34	49,84
İlk 2 harf	59,18	60,55
İlk 3 harf	71,98	71,88
İlk 4 harf	78,14	81,28
İlk 5 harf	85,37	86,88
İlk 6 harf	84,78	85,64
İlk 7 harf	84,14	85,94
F1L1	58,47	58,45
F1L2	75,27	75,65
F2L1	69,24	69,32
F2L2	81,34	81,24
F3L2	85,20	88,25
F2L3	86,63	88,37
F3L3	86,52	88,10
Gövde	93,94	94,91
Çekim Ekleri	80,07	82,47

Kelimenin son 5 harfini hesaplama birimi olarak kullanılması ile önerdiğimiz sözcük türü tespit yönteminden elde ettiğimiz başarı oranı (%90,2) tez kapsamında geliştireceğimiz tümleşik yöntemin başlangıç değeri için yeterli görülmektedir. Özellikle tezimizin önemli iddialarından biri olan kapalı sözlük kullanma amacımızı da desteklemektedir. Bunun yanı sıra hesaplama birimi olarak gövdelerin kullanılmasının hesaplama birimi kelime alınmasına göre daha yüksek bir başarı göstermesi de tezde elde edilen önemli verilerden bir tanesidir.

5. Gövdeleme

Bilgi geri getirim (BGG) ve doğal dil işlemenin alanlarının en eski ve temel problemlerinden biri olan gövdeleme işlemi, İngilizce başta olmak üzere birçok dil için üzerinde çözüm üretilmiş bir konudur. Gövdeleme, aynı anlamı taşıyan yüzeysel biçimlerden çekim eklerinin ayrılması ile ortak bir biçime getirilme işlemi olarak tanımlanabilir.

Gövdeleme, biçimbirimsel yapısı çok karışık olmayan analitik dillerde bitişken dillere nazaran daha kolay gerçekleştirilebilmektedir. Analitik dillerde kelimeye getirilen ekler sadece basit işlevler için kullanılır. Örnek olarak İngilizcede, sona gelen –s çoğul takısı veya –ed geçmiş zaman takısını verebiliriz (Cat+s, Publish+ed, Swim+ing, Typical+ly, Fish+er).

Fakat, Türkçe gibi dillerde ek sistemi cümle içerisinde anlamsal ve yapısal birçok görevi yerine getirmektedir. Ayrıca analitik dillerden farklı olarak bir kelimedenden türetilcek yüzeysel form sayısı oldukça fazla (gel-irim, gel-miştim, gel-seydim, gel-ecekmişim, gel-irsek, gel-diysek, gel-mekteyim ...) olması nedeniyle gövdeleme işlemi eklemeli dillerde daha karmaşık bir hal almaktadır.

Diğer önemli bir problem ise bir kelimedede birden fazla gövde alternatifinin bulunmasıdır. Örneğin aşağıdaki cümlelerde kullanılan “gidilecek” kelimesinin gövdesi “gidil-” ve ya “gidilecek” olabilmektedir.

Cümle1: Bu şehirde *gidilecek* yerler oldukça sınırlıdır.

Cümle 2: Bugün o müzeye muhakkak gidilecek.

Bu belirsizlik kelimenin cümle içerisindeki görevi (sözcük türü) göz önüne alındığında büyük oranda ortadan kalkmaktadır. Bir başka

deyişle örnek olarak verilen “gidilecek” kelimesi sıfat ise gövdesi “gidilecek”, fiil ise “gidil-” olacaktır.

Gövdeleme, doğal dil işleme ve bilgi geri-getirim alanlarında üzerinde en fazla çalışılan konulardan birisidir. Bunun nedeni gövdelemenin doğal dil işleme alanının bir çok uygulama sahasında (bilgi geri-getirim, sıkıştırma, konuşma anlama vb.) başarımı ve performansı artırıcı bir unsur olarak görülmesidir. (Solak ve Can, 1994; Ekmekçioğlu, Lynch, ve Willett, 1996; Sever ve Bitirim, 2003). Eklemeli dillerde gövdelemenin bilgi geri getirim sahası üzerindeki olumlu etkisinden söz edilirken analitik diller için böyle bir etkiden söz etmek çok doğru olmamaktadır.

Gövdeleme konusunda İngilizce için ilk çalışma 1968 yılında Lovins tarafından gerçekleştirilmiştir. Bu alanda günümüze kadar birçok çalışma yapılmıştır. Ama bu çalışmalar arasında en kabul göreni Porter’ın 1980 yılındaki çalışmasıdır. Porter’ın ortaya koyduğu ve üzerinde birçok uygulamanın geliştirildiği bu algoritma, İngilizce kelimelerde gövdeleme işlemi için standart gibidir. Bunun dışında da farklı yöntemleri kullanan farklı performans ve doğruluk oranlarına sahip birçok algoritma ortaya konmuştur (Brute Force Algorithms, Suffix Stripping Algorithms, Lemmatisation Algorithms, Stochastic Algorithms, Matching Algorithms vb.).

Gövdeleme için İngilizce dışında diğer dillerde de yapılan birçok araştırma mevcuttur (Dolamic, 2007, Savoy, 2006, Popovic & Willett, 1992, Viera & Virgil, 2007, Tordai & Rijke, 2005, Korenius et al, 2005).

Çizelge 5.1’de Türkçe için şimdiye kadar yapılan gövdeleme çalışmaları 3 ana başlıkta (yöntem, sözlük kullanımı, belirsizliğin giderilmesine ihtiyaç olma durumu) değerlendirilmiştir.

Çizelge 5.1. Türkçe gövdeleme çalışmalarının sınıflandırılması

Çalışmalar	Yöntem	Kök veya Gövde Sözlüğü Kullanımı	Belirsizliğin Giderilmesine İhtiyaç Olma Durumu
İlk 5 harf (Koksal, 1979)	Doğrudan	Hayır	Hayır
En uzun eşleme (Alpkocak et al., 1995)	Doğrudan	Evet	Hayır
A-F Algoritması (Solak and Can, 1994)	Analitik	Evet	Evet
Gövde-bul (Sever and Bitirim, 2003)	Analitik	Evet	Hayır
Zemberek (Akin, 2004)	Analitik	Evet	Evet
Ek soyma (Eryigit and Adali, 2004)	Analitik	Hayır	Evet
n-gram istatistikleri (Dincer, 2004)	İstatistiksel	Hayır	Evet
Melez Gövdeleme (Kisla, 2009)	Melez	Hayır	Hayır

Yöntem kolonunda yer alan “Doğrudan” kelimesi gövdeleme için dilbilgisel özelliklerin kullanılmadığını ifade etmektedir. “Analitik” kelimesi ise biçimbirimsel analiz ve dilbilgisel özellikler kullanıldığını belirtmektedir. “İstatistiksel” kelimesi ise çalışmanın istatistik olasılık temelinde olduğunu ifade etmektedir. Yöntemleri açısından sınıflandırmanın yanı sıra çalışmalar kök veya gövde sözlüğü kullanım durumları ve tek gövde alternatifi üretmelerine göre de sınıflandırılmıştır.

Literatür taraması sonucunda bulunan çalışmalara genel olarak bakıldığında, neredeyse tamamında belli bir kök veya gövde sözlüğünün kullanıldığı ve kelimelerin kendi başlarına, cümleden bağımsız olarak irdelenmesi neticesinde olabilecek tüm gövdelerin çıktı olarak verildiği görülmektedir.

5.1. Yöntem

Genellikle bütün dillerde ama özellikle sondan eklemeli dillerde, dile sürekli olarak yeni kelimelerin girmesi söz konusudur. Yani kelime dağarcığı kapalı değildir, sürekli genişlemektedir. Halbuki bir dilde

kullanılabilecek ekler ve bunların dizilişleri sınırlıdır. Ekler bağlandıkları sözcük türüne ve takip ettikleri ek dizisine göre farklılık gösterirler.

Sözlük kullanan yöntemlerde, sözlükte bulunmayan bir kelime ile karşılaşıldığında çözüm üretilmemektedir. Ayrıca alan yazında karşılaştığımız analitik yöntemler oldukça karmaşık sonlu durum makinelerine (SDM) dayanmaktadır. Halbuki makul zamanda sonuç alınabilmesi için daha az karmaşık yöntemlere ihtiyaç vardır. Bu durumlar göz önünde bulundurularak önerdiğimiz yöntemde kelime yerine kapalı olan eklerin ve ek dizilimlerinin bulunduğu bir sözlük ve karmaşıklığı indirgenmiş, sözcük türlerine göre süreçleri modelleyen bir yaklaşım benimsenmiştir. Yöntem hem istatistiksel hem de analitik yöntemlerden oluştuğundan melez bir yöntem olarak değerlendirilmektedir.

Türkçede kelimeler için 8 farklı sözcük türü (Fiil, isim, sıfat, zarf, bağlaç, zamir, edat, ünlem) olmasına rağmen temel olarak isim ve fiil olarak sınıflandırılabilir. Bu iki sınıfa gövdeleme problemi açısından baktığımızda farklı ekler aldıkları görülmektedir. Örneğin “arabanın” kelimesinde yer alan “-nın” çekim eki birliği bir fiilde görülmemektedir. Buradan hareketle bir kelimenin gövdesinin belirlenme işi yüksek oranda sözcük türüne bağlı olmaktadır. Bu ifadeyi temel alan yöntemimiz öncelikle kelimenin sözcük türünün belirli olmasını gerektirmektedir.

Yöntemde, isim ve fiillerin dilbilgisel özelliklerine (ek alma özellikleri) göre isim ve fiiller için iki adet SDM tasarlanmış ve çekim ekleri için ek kalıp kümeleri derlenmiştir. İki aşamadan oluşan sistemin birinci aşamasında tasarlanan SDM’ler ve derlenen çekim eki kalıpları kullanılarak kelimedeki ek sıyırma yöntemi ile çekim ekleri ayrıştırılmıştır. Bu işlem sonucunda verilen kelime için olası gövdeler

belirlenmiştir. İkinci aşamada ise önerilen gövdelerden hangisinin en uygun olduğunun belirlenmesi için istatistiksel yöntemler kullanılmıştır.

Yöntemimizin önemli bir katkısı da bu alandaki çalışmalardan farklı olarak kelimenin cümle içerisindeki yerine göre tek bir gövde önermesi ile belirsizliğin giderilmesi (disambiguation) problemini ortadan kaldırmasıdır. Ayrıca bir kök veya gövde sözlüğü yerine kapalı bir ek sözlüğü kullanması da yöntemimizin bir diğer katkısı olarak belirtilebilir.

5.1.1. Çekim Ekleri İçin Ek Kalıpları

Alan yazın çalışması sonucu analitik yöntemlerle gerçekleştirilen gövdeleme veya biçimbirimsel çözümleme çalışmalarının hemen hemen hepsinde ekler tek başlarına ele alınmıştır. Bunun yanı sıra tasarlanan SDM' leri Türkçe için olanaksız olan ek dizilimlerini de kapsamaktadır. Oysaki Türkçe'de gövdelerin sonuna getirilen ekler belirli bir dizilişe sahiptir ve bu diziliş rastsal değildir. Bu noktadan hareketle tezimizde yer alan gövdeleme probleminin çözümünün temelinde yatan iddia şu şekilde belirtilebilir:

“Türkçede bulunan çekim eklerinin dizilişleri göz önünde bulundurularak ek sıyırma işlemini basitleştirecek ek kalıpları belirlenebilir.”

Çekim ek kalıplarının belirlenmesi Türk dili kitaplarında yaygın olarak yer alan dilbilgisi kuralları incelenerek gerçekleştirilmiştir. Örneğin “geleceğim” kelimesindeki “-im” eki hiçbir zaman “-eceğ” ekinden sonra gelmemektedir. Yani “gelimcek” şeklinde bir kelime ile karşılaşılması ancak bu kelimenin tamamının gövde olma durumunda mümkün olmaktadır. Ek kalıplarının derlenmesinde kullanılan kurallar tezin “Türkçe” başlıklı bölümünde verilmiştir.

Çekim eklerinin Türkçede isim ve fiiller için farklı olmasından hareketle kalıplar iki ayrı küme altında toplanmıştır:

- ✓ Fiil çekim ekleri için ek kalıpları
- ✓ İsim çekim ekleri için ek kalıpları

Fiil çekim ekleri için 588 adet ek kalıbı belirlenmiştir. Çizelge 5.2’de fiil çekim ekleri için örnek ek kalıpları verilmiştir.

Çizelge 5.2. Fiil çekim ekleri için örnek ek kalıpları¹

Fiil çekim ekleri için ek kalıbı	Tanım
mUştUm	Öğrenilen Geçmiş zamanın hikayesi 1. tekil şahıs
sAydIm	Şartın Hikayesi 1. tekil şahıs
AcAkmİşIm	Gelecek zamanın rivayeti 1. tekil şahıs
AcAkmİşsInIz	Gelecek zamanın rivayeti 2. çoğul şahıs
rsAk	Geniş zamanın Şartı 1. çoğul Şahıs
DUysAk	Görülen geçmiş zamanın şartı 1. çoğul şahıs

Ek kalıplarının Türkçede yer alan büyük sesli uyumu, hece düşmesi, yardımcı ses eklemesi gibi bazı kurallar işletilerek elde edilen tüm formları hesaba katıldığında 2623 adet farklı form elde edilmiştir. Bir ek kalıbının tüm formlarına aşağıda sunulan kalıplar örnek olarak verilebilir.

$$sAydIm = \{saydım, seydım\}$$

$$DUysAk = \{dıysak, duysak, diysek, düysek, tıysak, tuysak, tiysek, tüysek\}$$

¹ Burada görülen (A, U, I, D) gibi harflerin karşılıkları tablo 2.1’ de verilmiştir.

Belirlenen 588 ek kalıbı, çözümleme sürecinin basitleştirilmesi amacıyla yapılarının benzerliği göz önünde bulundurularak 5 grupta toplanmaktadır (Çizelge 5.3).

Çizelge 5.3. Fiil çekim ekleri ek kalıpları için alt gruplar

Grup Numarası	Tanım	Ek Kalıbı Sayıları
1	Bileşik Soru eki	23
2	İsim-fiil ekleri	24
3	Birleşik fiil çekimleri	480
4	Kuvvetlendirme ihtimal- fiil çekim ekleri	42
5	İsme gelen bileşik soru isim fiil ekleri	19

Fiil çekim ek kalıpları için belirlenen bu gruplar Çizelge 5.3’de verilen numaralarına göre şu şekilde açıklanabilir:

Grup Numarası 1: Türkçede yer alan soru eki ve sonra gelen zaman ve şahıs çekimlerini içermektedir. Soru ekinden sonra gelen zaman çekimleri esasında i- fiilinin çekimleridir. Örnek: sever miydim?, sevecekler miymiş?

Grup Numarası 2: İsim fiil ekleri: Bu ekler aslında i- fiilinin çekimleri olup, tüm isim ve fiil şekillerinin sonuna gelebilir. İsimlerin sonuna geldiğinde isimleri fiilleştirir. Örnek: “işte-ydi”, “güzel-miş”

Grup Numarası 3: En fazla ek kalıbı içeren bu grupta bir fiilin alabileceği tüm zaman ve şahıs çekimleri yer almaktadır. Örnek: koş-uyorlarmış, sevmi-yormuşsunuz.

Grup Numarası 4: Kuvvetlendirme-ihtimal fiil çekim ekleri: Türkçede fiil çekiminin sonuna gelerek onları kuvvetlendirme anlamı

veya ihtimal anlamı katan “-dir” ekinin tüm çekimlerini içermektedir.
Örnek: bil-irimdir, yap-ıyorumdur.

Grup Numarası 5: Bu ekler bir ismin sonuna gelip soru anlamı katarak, ismin cümle içerisinde fiil görevi görmesini sağlamaktadır.
Örnek: evde miydiniz?, evde misin?

İsim çekim ekleri için ise toplam 133 adet ek kalıbı belirlenmiştir.
Çizelge 5.4’de isim çekim ekleri için örnek ek kalıpları verilmiştir.

Çizelge 5.4. İsim çekim ekleri için örnek ek kalıpları

İsim çekim ekleri için ek kalıbı	Tanım
UnlArIn	2. şahıs İyelik + çokluk + genetik
nIzA	2. çoğul şahıs iyelik + datif
lArImIzIA	Çokluk + 1. çoğul şahıs iyelik + instrumental (-IA)
Larıda	çokluk + 3. çoğul şahıs iyelik + lokatif
mIzI	1. çoğul şahıs iyelik + akuzatif
UndAn	2. tekil şahıs iyelik + ablatif

İsim çekim ekleri için belirlenen 130 adet ek kalıbı tüm formları hesaba katıldığında 811 adet olmaktadır. Bu ek kalıpları da Çizelge 5.5.’de belirtilen 3 alt grupta sınıflandırılmıştır.

Çizelge 5.5. İsim çekim ekleri ek kalıpları için alt gruplar

Grup Numarası	Tanım	Ek Kalıbı Sayıları
6	Birleşik isim çekimleri	127
7	Aitlik -ki eki	1
8	-cesine ve -ken Eki	2

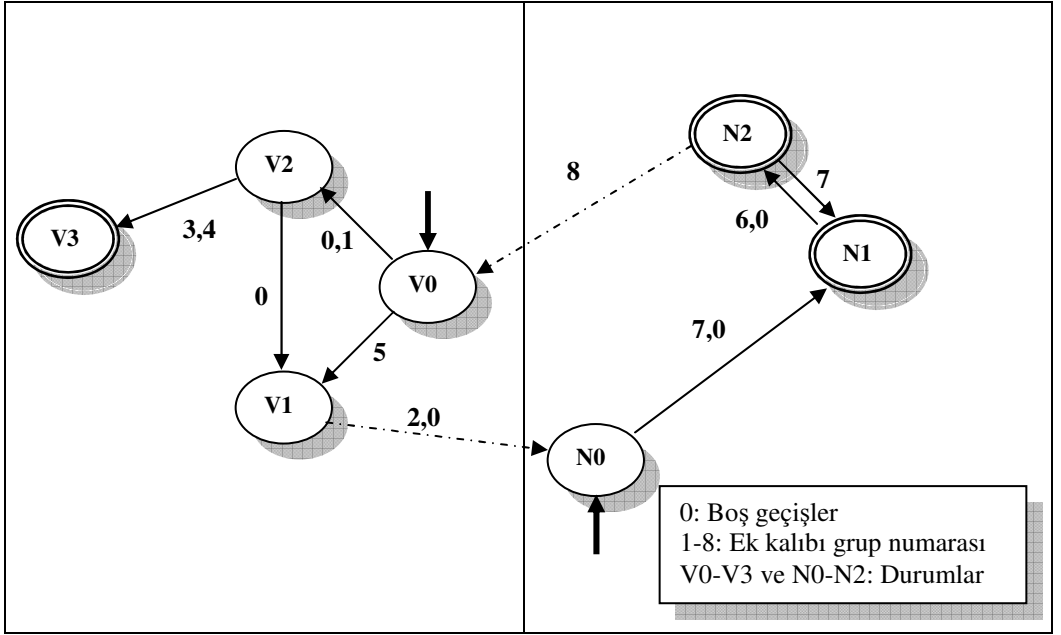
İsim çekim ek kalıpları için belirlenen bu gruplar Çizelge 5.5’de verilen numaralarına göre şu şekilde açıklanabilir:

Grup Numarası 6: Bir ismin tüm çekimlerini içermektedir. Örnek: kitap-ların, ev-lerinizden, araba-dayım.

Grup Numarası 7: Türkçede aitlik eki olarak kullanılan “-ki” eki kelime içerisindeki bağlaç olarak nitelendirilebilir. Yani bu ek eklendiği kelimeyi tekrardan isim çekim ekleri alabilecek vaziyete getirmektedir. Örnek kitap-larımda-ki-lerden

Grup Numarası 8: Gerendium eki olarak da bilinen “-ken” ve “-cesine” eki çekimli fiillerden zarf yapmaktadır. Örneğin “-erek” gerendium eki yalın haldeki fiillere eklenirken (“gel-erek”), “-ken” ve “-cesine” ekleri zaman çekim eki almış fiillere eklenmektedirler (gel-ir-ken, gel-miş-cesine). Bu sebepten dolayı bu ekler çekim ekleri kapsamında ele alınmıştır.

Ek kalıplarının kelimedenden ayrılması amacıyla isimler ve fiiller için ayrı ayrı sonlu durum makineleri tasarlanmıştır (Şekil 5.1, Şekil 5.2). Şekilden de görüleceği gibi tasarlanan SDM’lar aynı problemin çözümü için geliştirilmiş SDM’larına göre oldukça basit bir yapıdadır. Bu basit yapının nedeni olarak karmaşıklığı ortaya çıkaran ek dizilimlerinin çözümünün büyük bir kısmının ek kalıpları kullanılarak ortadan kaldırılması gösterilebilir.



Şekil 5.1. Fiil çekim ek kalıpları için SDM

Şekil 5.2. İsim çekim ek kalıpları için SDM

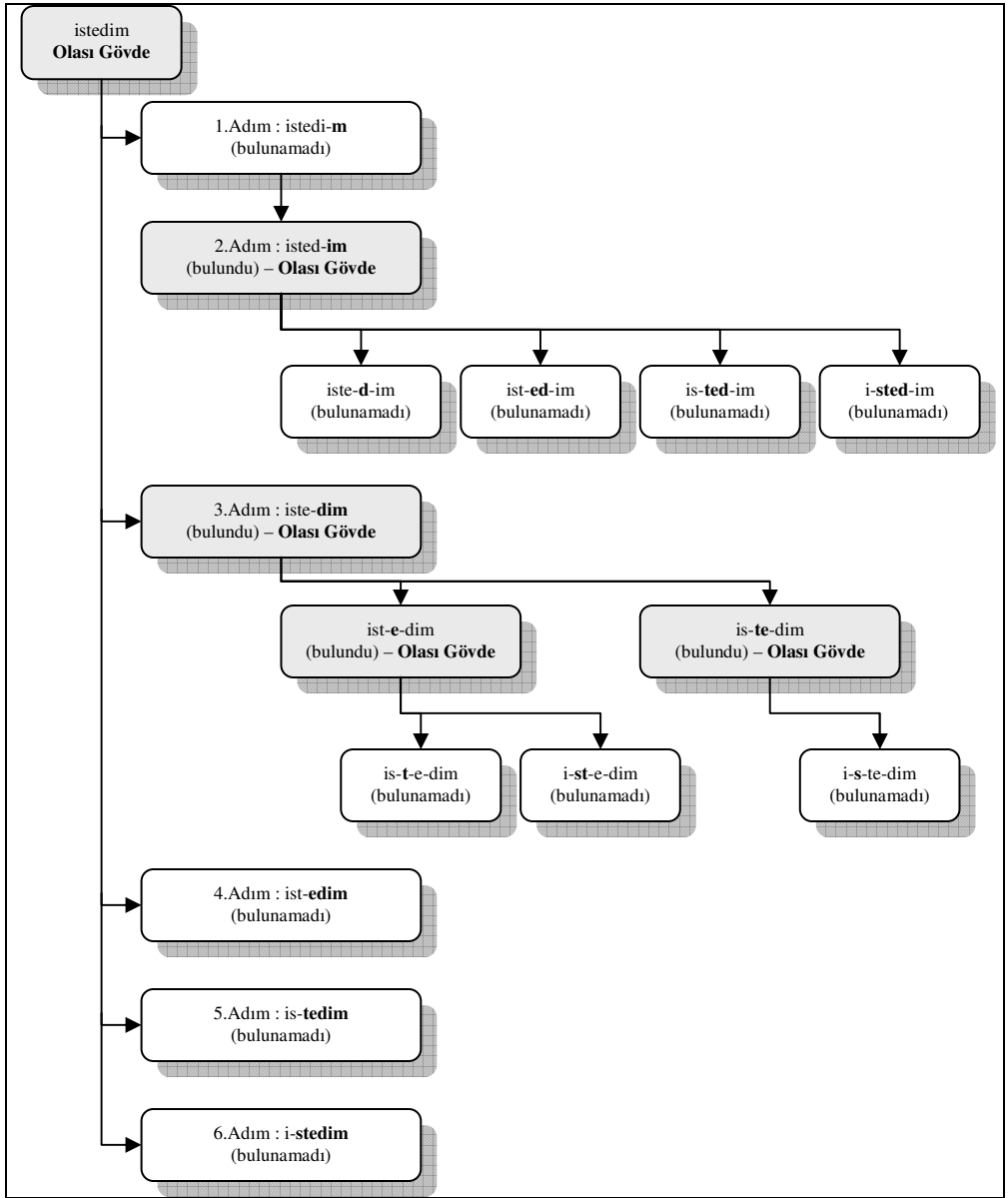
Sonlu durum makinelerinde okların üzerinde bulunan ek kalıp grup numaraları (Çizelge 5.3 ve Çizelge 5.5) geçiş koşulları olarak belirlenmiştir. “0” olarak belirlenen geçiş koşulu ise boş geçişleri temsil etmektedir. V0,V1,V2 ve V3 fiil çekim ek kalıpları SDM’si için durumları gösterirken N0, N1 ve N2 ise isim çekim ek kalıpları SDM’si için durumları göstermektedir. N0 ve V0 sırası ile isim ve çekim ek kalıpları SDM’leri için başlangıç durumları olup gövdelenecek kelime bu durumlardan işleme başlamaktadır. SDM’lerinin sonlanma durumları ise çift çizgili elipsler ile belirlenmiştir. Tasarlanan iki SDM arasındaki geçişler ise çizgili oklarla gösterilmiştir.

SDM’leri, çözümlenecek kelimedenden sıyrılan her bir çekim ek kalıbı ile durum değiştirmektedir. Örneğin “geleceğim” kelimesinde fiil çekim ek kalıbı olarak “-eceğim”in (grup numarası: 3) belirlenmesi halinde, başlangıçta V0 durumunda olan SDM, 3 geçiş koşulu için çalışmaya başlayacaktır. Öncelikle V0 durumundan V2 durumuna geçen SDM daha

sonra V3 durumuna geçecektir. Başka bir fiil çekim ek kalıbının bulunmaması halinde sonlanma durumunda bulunan SDM başarı ile bitecektir ve bulunan gövde olası çözümler arasına eklenecektir.

5.1.2. Eklerin Sıyırılması

Bu aşamada eklerin soyulma işlemi için kelime sağdan başlayarak harf harf ilerlenerek her bir harf birliği ek kalıpları içerisinde aranmaktadır. Arama sonucunda harf birliği ek kalıpları arasında bulunursa bulunan ek kalıbının grubuna göre sonlu durum makinesinde durum geçişleri gerçekleştirilmektedir. Bu işlem kelimedeki harfler bitene kadar devam ettirilmektedir. İşlemlerin sonucunda analiz edilen kelimeye ait olası gövdeler belirlenmektedir. Sürecin işleyişi Şekil5.3'de verilen örnekte açıklanmaktadır.



Şekil 5.3. “istedim” kelimesinin analizi

Yöntemde sözlük kullanılmadığından dolayı analiz edilen kelimenin tek başına gövde olma olasılığı daima mevcuttur. Bu nedenle kelimenin bütünü daima olası gövdeler listesinde yer almaktadır. Diğer

olası gövdelerin bulunması amacı ile örnek olarak gösterilen “istedim” kelimesi sağdan başlayarak analiz sürecine alınır. İlk olarak “-m” harf birliği belirlenerek fiil çekim ekleri için hazırlanan ek kalıpları içerisinde aranır. Arama sonucunda eşleşme bulunamadığı için bir sonraki harf birliğe dahil edilerek aranılacak birlik “-im” şeklini alır. “-im” birliği ek kalıp listesinde mevcut olduğundan eşleşme bulunduktan sonra ek kalıp grup numarası tespit edilerek sonlu durum makinesi işletilir. Bu durumda sonlu durum makinesi sonlandığı için bulunan gövde olası gövde listesine dahil edilmektedir. Bu süreç kelimenin tüm harfleri işleninceye kadar devam etmektedir.

Sonuç olarak “istedim” kelimesi için 5 olası gövde tespit edilmiş olur: isted-im, ist-e-dim, is-te-dim, iste-dim, istedim (ilk tireden önce yer alan birlik gövde olarak belirtilmiştir). Belirlenen bu gövdelerden hangisinin en uygun gövde olduğu bir sonraki aşamada belirlenmektedir.

5.1.3. Uygun Gövdenin Belirlenmesi

En uygun gövdenin belirlenmesi aşaması istatistik/olasılık tabanlı bir yöntem olup daha önceden belirlenen olası gövdelerden en uygununun seçilmesini hedefler. Anlatımda kullanılan simge gösteriminin açıklaması şöyledir:

$$\begin{array}{ll} n \text{ harften oluşan} & k_n = h_1 h_2 \cdots h_n \\ \text{bir kelimenin} & \\ \text{gösterimi:} & n: \text{kelimenin uzunluğu, } n > 0, h: \text{harf gösterimi} \end{array} \quad (5.1)$$

$$\begin{array}{ll} m \text{ harften} & g_m = h_1 h_2 \cdots h_m \\ \text{oluşan gövdenin} & \\ \text{gösterimi:} & m: \text{gövdenin uzunluğu, } n \geq m > 0 \end{array} \quad (5.2)$$

$$\begin{array}{ll} p \text{ harften oluşan} & e_p = h_{m+1} h_{m+2} \cdots h_n \\ \text{eklerin} & \\ \text{gösterimi:} & p: \text{gövdenin uzunluğu, } n > p \geq 0, n = m + p \end{array} \quad (5.3)$$

*Kelimenin
gösterimi*

$$k_n = g_m e_p \quad (5.4)$$

$$k_n = h_1 h_2 \cdots h_m h_{m+1} h_{m+2} \cdots h_n$$

*İki harf çiftinin
gövdede olma
durumu:*

$$G = \{ (h_i, h_{i+1}) \mid h_i \in g_m \wedge h_{i+1} \in g_m \wedge 0 < m \leq n \} \quad (5.5)$$

*Bir harfin
gövdede diğer
harfin ekte olma
durumu:*

$$B = \{ (h_i, h_{i+1}) \mid h_i \in g_m \wedge h_{i+1} \in e_p \wedge 0 < p < n \} \quad (5.6)$$

Yukarıdaki tanımlamalar kullanılarak, verilen bir kelimenin ($k_n = h_1 h_2 \cdots h_n$) içerisinde yer alan herhangi bir harf çiftinin (h_i, h_{i+1}) gövdenin bir parçası olma olasılığı, ek sisteminin bir parçası olma olasılığı ve gövde ile ek sisteminin geçişinde olma olasılığının hesabında kullanacak simgeler ve olasılık hesapları aşağıda tanımlanmıştır.

$$P((h_i, h_{i+1}) \in G) = P_G((h_i, h_{i+1})) = w_{g,i} / N \quad (5.7)$$

$$P((h_i, h_{i+1}) \in B) = P_B((h_i, h_{i+1})) = w_{b,i} / N \quad (5.8)$$

Hesaplama kullanılan semboller:

$f_{g,i}$: (h_i, h_{i+1}) harf çiftinin istatistiklerin çekileceği derlemdeki kelimelerde i . sırada ve gövdede görülme sayısı

$f_{b,i}$: (h_i, h_{i+1}) harf çiftinin istatistiklerin çekileceği derlemdeki kelimelerde i . sırada ve gövde ve ek geçişinde görülme sayısı

$$w_{g,i} : f_{g,i} / (f_{g,i} + f_{e,i} + f_{b,i}) \quad (5.9)$$

$$w_{b,i} : f_{b,i} / (f_{g,i} + f_{e,i} + f_{b,i}) \quad (5.10)$$

N : Tüm (h_i, h_{i+1}) harf çiftlerinin toplam sayısı

Tanımlanan bu simgeler ile olası tüm gövde sonuçlarına bir olasılık değeri atanmaktadır.

$$P_K(j) \quad \begin{array}{l} \text{j. olası gövde indisi,} \\ 1 \leq j \leq ng \quad (ng: \text{kelimenin tespit edilen olası gövde sayısı.}) \end{array}$$

Bu olasılık değeri gövdenin son elemanın sırası i olmak üzere şu şekilde hesaplanmaktadır:

$$P_K(j) = P_G((h_{i-1}, h_i)) * P_B((h_i, h_{i+1})) \quad (5.11)$$

Her olası gövde için hesaplanan bu olasılık değerleri arasında en yüksek değere sahip gövde sonuç olarak belirlenmektedir:

$$\operatorname{argmax}_{1 \leq j \leq ng} (P_K(j)) \quad (5.12)$$

Yöntemde belirlenen gövdelerin olasılıklarının belirlenmesi amacıyla 10 farklı varyasyon kullanılmıştır. Bunlar; istatistiklerin çekildiği verilere ve hesaplama birimi olarak kullanılan harf sayılarına göre belirlenmiştir.

I. Hesaplama Birimi: İkili harf dizileri

İstatistik/ olasılık verileri: Kelimenin sözcük türleri sadece fiil ve isim olarak gruplayarak çekilen istatistikler

II. Hesaplama Birimi: İkili harf dizileri

İstatistik/ olasılık verileri: Kelimenin tüm sözcük türlerine göre (isim, fiil, sıfat, zarf ...) gruplayarak çekilen istatistikler

Örnek kelime ('istedim') için belirlenen olası gövdelere atanan olasılık değerleri her iki varyasyon için şu şekilde olmaktadır:

Çizelge 5.6. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon I ile elde edilen olasılık sonuçları

Olası Gövdeler	Varyasyon I (istatistikler isim/fiil gruplarına göre)	Olasılık Değeri
istedim	$P_K(1) = P_G(('i', 'm')) * P_B(('m', '_'))$	2,7821E-12
isted-im	$P_K(2) = P_G(('e', 'd')) * P_B(('d', 'i'))$	1,2302E-10
ist-e-dim	$P_K(3) = P_G(('s', 't')) * P_B(('t', 'e'))$	1,4588E-8
is-te-dim	$P_K(4) = P_G(('i', 's')) * P_B(('s', 't'))$	5,1051E-10
iste-dim	$P_K(5) = P_G(('t', 'e')) * P_B(('e', 'd'))$	2,1189E-8

Çizelge 5.7. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon II ile elde edilen olasılık sonuçları

Olası Gövdeler	Varyasyon II (istatistikler tüm sözcük türü gruplarına göre)	Olasılık Değeri
istedim	$P_K(1) = P_G(('i', 'm')) * P_B(('m', '_'))$	2,9223E-12
isted-im	$P_K(2) = P_G(('e', 'd')) * P_B(('d', 'i'))$	1,3102E-10
ist-e-dim	$P_K(3) = P_G(('s', 't')) * P_B(('t', 'e'))$	1,5033E-8
is-te-dim	$P_K(4) = P_G(('i', 's')) * P_B(('s', 't'))$	5,5941E-10
iste-dim	$P_K(5) = P_G(('t', 'e')) * P_B(('e', 'd'))$	2,1835E-8

Bu iki varyasyonda yukarıda verilen simge sistemi ve hesaplamalar kullanılmaktadır. İlerleyen iki varyasyonda ise hesaplamalar biraz farklılaşmaktadır.

Dört harf çiftinin gövdede olma durumu:

$$G = \left\{ (h_{i-1}, h_i, h_{i+1}, h_{i+2}) \mid h_{i-1}, h_i, h_{i+1}, h_{i+2} \in g_m \wedge 0 < m \leq n \right\} \quad (5.13)$$

İki harfin gövdede diğer iki harfin ekte olma durumu:

$$B = \left\{ (h_{i-1}, h_i, h_{i+1}, h_{i+2}) \mid h_{i-1}, h_i \in g_m \wedge h_{i+1}, h_{i+2} \in e_p \wedge 0 < p < n \right\} \quad (5.14)$$

III. Hesaplama Birimi: Dörtlü harf dizileri

İstatistik/ olasılık verileri: Kelimenin sözcük türleri sadece fiil ve isim olarak gruplayarak çekilen istatistikler

IV. Hesaplama Birimi: Dörtlü harf dizileri

İstatistik/ olasılık verileri: Kelimenin tüm sözcük türlerine göre (isim, fiil, sıfat, zarf ...) gruplayarak çekilen istatistikler

Örnek kelime ('istedim') için belirlenen olası gövdelere atanan olasılık değerleri her iki varyasyon için şu şekilde olmaktadır:

Çizelge 5.8. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon III ile elde edilen olasılık sonuçları

Olası Gövdeler	Varyasyon III (istatistikler isim/fiil gruplarına göre)	Olasılık Değeri
istedim	$P_K(1) = P_G(('e', 'd', 'i', 'm')) * P_B(('i', 'm', ' ', ' '))$	0
isted-im	$P_K(2) = P_G(('s', 't', 'e', 'd')) * P_B(('e', 'd', 'i', 'm'))$	0
ist-e-dim	$P_K(3) = P_G((' ', 'i', 's', 't')) * P_B(('s', 't', 'e', 'd'))$	0
is-te-dim	$P_K(4) = P_G((' ', ' ', 'i', 's')) * P_B(('i', 's', 't', 'e'))$	1,3829E-10
iste-dim	$P_K(5) = P_G(('i', 's', 't', 'e')) * P_B(('t', 'e', 'd', 'i'))$	1,1299E-8

Çizelge 5.9. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon IV ile elde edilen olasılık sonuçları

Olası Gövdeler	Varyasyon IV (istatistikler tüm sözcük türü gruplarına göre)	Olasılık Değeri
istedim	$P_K(1) = P_G(('e', 'd', 'i', 'm')) * P_B(('i', 'm', ' ', ' '))$	0
isted-im	$P_K(2) = P_G(('s', 't', 'e', 'd')) * P_B(('e', 'd', 'i', 'm'))$	0
ist-e-dim	$P_K(3) = P_G((' ', 'i', 's', 't')) * P_B(('s', 't', 'e', 'd'))$	0
is-te-dim	$P_K(4) = P_G((' ', ' ', 'i', 's')) * P_B(('i', 's', 't', 'e'))$	2,4198E-13
iste-dim	$P_K(5) = P_G(('i', 's', 't', 'e')) * P_B(('t', 'e', 'd', 'i'))$	1,17232E-8

V.Hesaplama Birimi: İkili harf dizileri (gövdede bulunma olasılıkları için), Dörtlü harf dizileri (gövde ve ek geçişinde bulunma olasılıkları için)

İstatistik/ olasılık verileri: Kelimenin sözcük türleri sadece fiil ve isim olarak gruplayarak çekilen istatistikler

VI.Hesaplama Birimi: İkili harf dizileri (gövdede bulunma olasılıkları için), Dörtlü harf dizileri (gövde ve ek geçişinde bulunma olasılıkları için)

İstatistik/ olasılık verileri: Kelimenin tüm sözcük türlerine göre (isim, fiil, sıfat, zarf ...) gruplayarak çekilen istatistikler

Örnek kelime ('istedim') için belirlenen olası gövdelere atanan olasılık değerleri her iki varyasyon için şu şekilde olmaktadır:

Çizelge 5.10. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon V ile elde edilen olasılık sonuçları

Olası Gövdeler	Varyasyon V (istatistikler isim/fiil gruplarına göre)	Olasılık Değeri
istedim	$P_K(1) = P_G(('i', 'm')) * P_B(('i', 'm', ' ', ' '))$	0
isted-im	$P_K(2) = P_G(('e', 'd')) * P_B(('e', 'd', 'i', 'm'))$	0
ist-e-dim	$P_K(3) = P_G(('s', 't')) * P_B(('s', 't', 'e', 'd'))$	0
is-te-dim	$P_K(4) = P_G(('i', 's')) * P_B(('i', 's', 't', 'e'))$	2,8329E-10
iste-dim	$P_K(5) = P_G(('t', 'e')) * P_B(('t', 'e', 'd', 'i'))$	7,3574E-8

Çizelge 5.11. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon VI ile elde edilen olasılık sonuçları

Olası Gövdeler	Varyasyon VI (istatistikler tüm sözcük türü gruplarına göre)	Olasılık Değeri
istedim	$P_K(1) = P_G(('i', 'm')) * P_B(('i', 'm', ' ', ' '))$	0
isted-im	$P_K(2) = P_G(('e', 'd')) * P_B(('e', 'd', 'i', 'm'))$	0
ist-e-dim	$P_K(3) = P_G(('s', 't')) * P_B(('s', 't', 'e', 'd'))$	0
is-te-dim	$P_K(4) = P_G(('i', 's')) * P_B(('i', 's', 't', 'e'))$	2,9236E-10
iste-dim	$P_K(5) = P_G(('t', 'e')) * P_B(('t', 'e', 'd', 'i'))$	7,6075E-8

Varyasyon 7, 8, 9 ve 10'da ise olası gövdeler için atanan olasılık değerlerinin hesaplamasında sadece harflerin gövde-ek geçişinde olma olasılıkları kullanılmıştır.

VII. *Hesaplama Birimi:* İkili harf dizileri (gövde ve ek geçişinde bulunma olasılıkları için)

İstatistik/ olasılık verileri: Kelimenin sözcük türleri sadece fiil ve isim olarak gruplayarak çekilen istatistikler

VIII. *Hesaplama Birimi:* İkili harf dizileri (gövde ve ek geçişinde bulunma olasılıkları için)

İstatistik/ olasılık verileri: Kelimenin tüm sözcük türlerine göre (isim, fiil, sıfat, zarf ...) gruplayarak çekilen istatistikler

Örnek kelime ('istedim') için belirlenen olası gövdelere atanan olasılık değerleri her iki varyasyon için şu şekilde olmaktadır:

Çizelge 5.12. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon VII ile elde edilen olasılık sonuçları

Olası Gövdeler	Varyasyon VII (istatistikler isim/fiil gruplarına göre)	Olasılık Değeri
istedim	$P_K(1) = P_B('m', ' _')$	1,23423E-20
isted-im	$P_K(2) = P_B('d', 'i')$	1,62034E-21
ist-e-dim	$P_K(3) = P_B('t', 'e')$	1,36914E-20
is-te-dim	$P_K(4) = P_B('s', 't')$	1,50337E-20
iste-dim	$P_K(5) = P_B('e', 'd')$	1,17232E-7

Çizelge 5.13. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon VIII ile elde edilen olasılık sonuçları

Olası Gövdeler	Varyasyon VIII (istatistikler tüm sözcük türü gruplarına göre)	Olasılık Değeri
istedim	$P_K(1) = P_B('m', ' _')$	1,2324E-20
isted-im	$P_K(2) = P_B('d', 'i')$	1,2324E-20
ist-e-dim	$P_K(3) = P_B('t', 'e')$	1,26807E-5
is-te-dim	$P_K(4) = P_B('s', 't')$	3,84878E-8
iste-dim	$P_K(5) = P_B('e', 'd')$	3,25555E-5

IX.Hesaplama Birimi: Dörtlü harf dizileri (gövde ve ek geçişinde bulunma olasılıkları için)

İstatistik/ olasılık verileri: Kelimenin sözcük türleri sadece fiil ve isim olarak gruplayarak çekilen istatistikler

X.Hesaplama Birimi: Dörtlü harf dizileri (gövde ve ek geçişinde bulunma olasılıkları için)

İstatistik/ olasılık verileri: Kelimenin tüm sözcük türlerine göre (isim, fiil, sıfat, zarf ...) gruplayarak çekilen istatistikler

Örnek kelime (‘istedim’) için belirlenen olası gövdelere atanan olasılık değerleri her iki varyasyon için şu şekilde olmaktadır:

Çizelge 5.14. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon IX ile elde edilen olasılık sonuçları

Olası Gövdeler	Varyasyon IX (istatistikler isim/fiil gruplarına göre)	Olasılık Değeri
istedim	$P_K(1) = P_B (('i', 'm', ' ', ' '))$	1,43434E-20
isted-im	$P_K(2) = P_B (('e', 'd', 'i', 'm'))$	1,59077E-21
ist-e-dim	$P_K(3) = P_B (('s', 't', 'e', 'd'))$	1,34415E-21
is-te-dim	$P_K(4) = P_B (('i', 's', 't', 'e'))$	1,47593E-20
iste-dim	$P_K(5) = P_B (('t', 'e', 'd', 'i'))$	1,12991E-7

Çizelge 5.15. “istedim” kelimesinin gövde tespitinde kullanılan varyasyon X ile elde edilen olasılık sonuçları

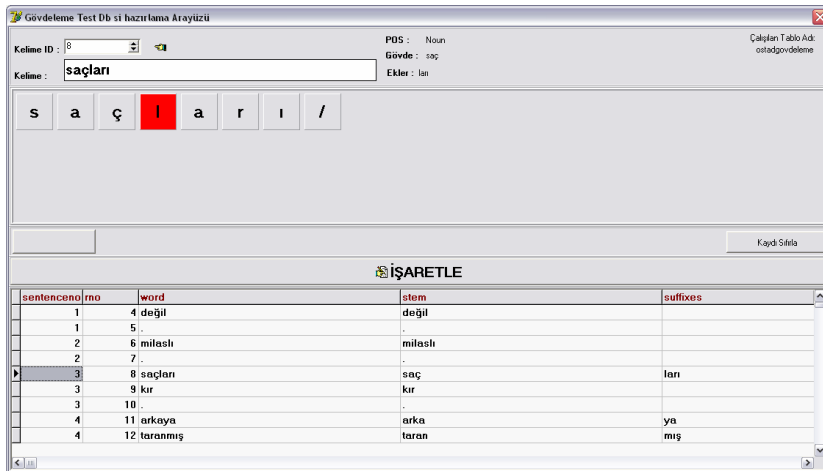
Olası Gövdeler	Varyasyon X (istatistikler tüm sözcük türü gruplarına göre)	Olasılık Değeri
istedim	$P_K(1) = P_B (('i', 'm', ' ', ' '))$	1,34343E-20
isted-im	$P_K(2) = P_B (('e', 'd', 'i', 'm'))$	1,54545E-14
ist-e-dim	$P_K(3) = P_B (('s', 't', 'e', 'd'))$	1,65447E-10
is-te-dim	$P_K(4) = P_B (('i', 's', 't', 'e'))$	3,44642E-11
iste-dim	$P_K(5) = P_B (('t', 'e', 'd', 'i'))$	1,13424E-5

5.2. Yöntemin Başarım Oranlarının Tespiti

5.2.1. Kullanılan Derlemler

Tez kapsamında geliştirilen gövdeleme yönteminin birinci aşamasında kural tabanlı bir yöntem kullanılırken, ikinci aşamasında ise istatistiksel yöntemlere yer verilmektedir. Bu nedenle ikinci aşamada kullanılacak istatistiklerin çekilebilmesi amacıyla kelimelerin sözcük türlerinin ve gövdelerinin işaretlenmiş olduğu bir talim derlemine, aynı zamanda üzerinde yöntemimizin başarısının sınanması amacıyla kullanacağımız test derlemine ihtiyaç duyulmaktadır. Bu amaca yönelik olarak kelimelerin sözcük türleri ve dilbilgisi özellikleri el ile işaretlenmiş ODTÜ-Sabancı Türkçe derlemi kullanılmıştır. Fakat bu derlem içerisinde net ve doğru bir gövdeleme işaretlemesinin olmaması nedeni ile derlem elle tekrar gövdelenmiştir. Gövdeleme işlemi, geliştirilen bir araç (Şekil 5.4) yardımı ile Türk dili konusunda bir uzman denetiminde gerçekleştirilmiştir.

Gövdeleme işlemi sonucunda ODTÜ-Sabancı Türkçe derlemi içerisinde bulunan 7494 cümle içerisinde 52179 (Noktalama işaretleri hariç) kelimenin gövdeleri işaretlenmiştir.



The screenshot shows a software interface for preparing a morphological database. At the top, there's a search bar with 'saçları' entered. Below it, a list of words is shown with their stems and suffixes. The table below the list shows the results of the morphological analysis.

sentenceno	no	word	stem	suffixes
1	4	değil	değil	
1	5	.	.	
2	6	milaslı	milaslı	
2	7	.	.	
3	8	saçları	saç	ları
3	9	kir	kir	
3	10	.	.	
4	11	arkaya	arka	ya
4	12	teranmış	teran	mış

Şekil 5.4. ODTÜ-Sabancı Türkçe derleminin elle gövdelenmesine yardımcı araç

5.2.2. Deneysel Sonuçlar

Gövdeleme yönteminin başarısı iki farklı test yöntemi ile sınanmıştır: bootstrapping ve çapraz doğrulama (cross validation).

Bootstrapping yönteminde ODTÜ-Sabancı Türkçe derlemi 6000 cümle talim derlemi için 1462 cümle olmak üzere iki alt derleme ayrılmıştır.

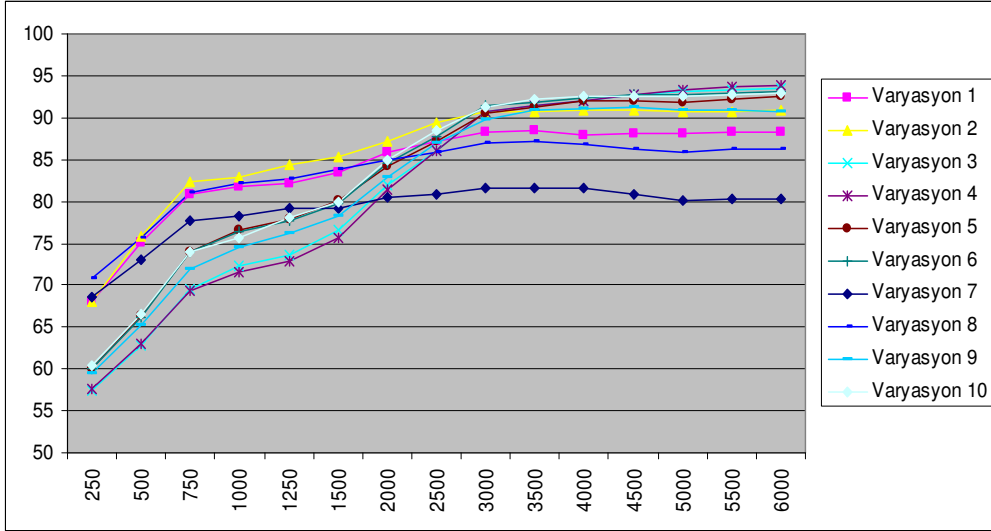
6000 cümleden oluşan talim derleminden, istatistiklerin çekildiği derlemlerin boyutunun başarıya olan katkısının belirlenmesi amacı ile 15 adet farklı boyutta (250, 500, 750, 1.000, 1.250, 1.500, 2.000, 2.500, 3.000, 3.500, 4.000, 4.500, 5.000, 5500 ve 6000) derlem elde edilmiştir.

Yöntemin testlerinin yapılacağı 1462 cümlelik test derlemi boyutu 200 cümle olan (yaklaşık 1000 kelime) 30 adet alt derlem rastsal bir biçimde seçilmiş ve farklı boyuttaki 15 talim derleminden elde edilen istatistikler ile test edilmiştir. Elde edilen bu başarı oranları bu 30 derlemin başarı ortalamalarıdır.

Gerçekleştirilen deney sayısı 10 farklı varyasyon kullandığımız göz önünde bulundurulursa 4.500 (10x15x30)' dür. Her bir deneydeki başarı oranları, doğru olarak tespit edilen sözcük gövdesinin, tüm kelime sayısına bölünmesi ile elde edilmiştir.

$$basari_oranı = \frac{Dogru_sekilde_gövdelenen_kelime_sayisi}{Toplam_kelime_sayisi} \quad (5.15)$$

Gövdeleme yöntemimiz sözcük türüne bağımlı çalışmaktadır. Yapılan deneylerde ilk olarak kelimelerin ODTÜ-Sabancı derleminde daha önceden işaretlenmiş sözcük türleri kullanılmıştır. İstatistiklerin çekildiği verilere ve hesaplama birimi olarak kullanılan harf birliklerine göre belirlenmiş 10 farklı varyasyon için farklı talim derlemi boyutlarından elde edilen sonuçların tablosu ek 4'te ve grafiği Şekil 5.5'te verilmiştir.



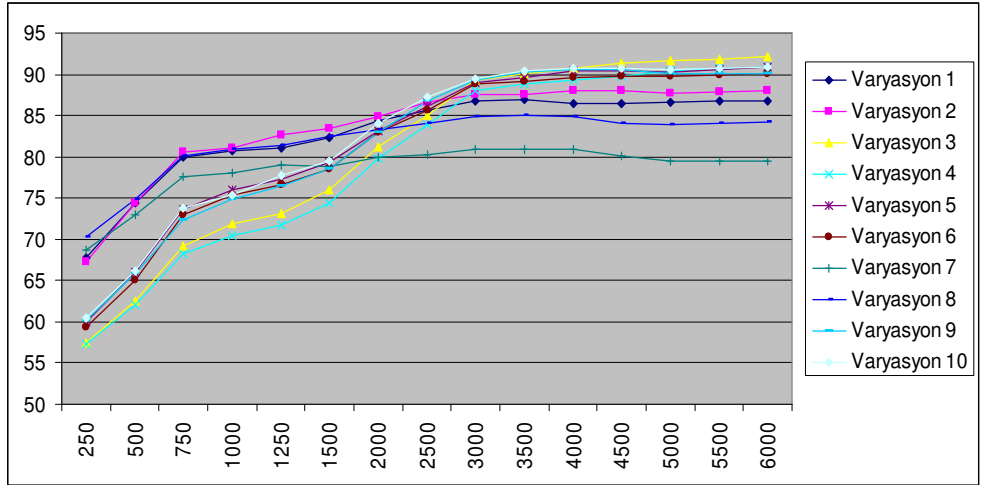
Şekil 5.5. Sözcük türü bilinen kelimeler üzerinden gerçekleştirilen deney sonuçlarının grafiği

En iyi başarıyı (%93,83) 4 numaralı varyasyon (Hesaplama Birimi: Dörtlü harf dizileri, İstatistik/ olasılık verileri: Kelimenin sözcük türleri sadece fiil ve isim olarak gruplayarak çekilen istatistikler) göstermiştir. Şekile genel olarak bakıldığında ise 1, 2, 7 ve 8 numaralı yöntemin başarı oranlarının 3000 cümlelik derlemde çekilen istatistiklerden sonra durağan bir hal aldığını görebilmekteyiz. Bunun nedeni olarak bu yöntemlerde harf çiftlerinin kullanılmasını gösterebiliriz. Başka bir deyişle başarı oranlarının sabitlenmesi 3000 cümlede ikili harf birlikleri için tüm olasılıklarının gerçekleştiği şeklinde yorumlanabilir. Ayrıca 10 numaralı varyasyon deneyimizde kullanılan istatistiklerin sadece gövde-ek geçiş istatistikleri olmasına rağmen elde ettiği başarı oranının kayda değer olduğu söylenebilir.

Deneyimizin bir sonraki bölümünde ise deney derlemindeki kelimelerin sözcük türleri bilinmediği varsayılmış ve gövdeleme işlemine başlanmadan önce kelimelerin sözcük türleri belirlenmiştir. Bunun için tez kapsamında geliştirilen, başarı oranı yaklaşık %90 olan, kelimenin

son 5 harfini ve 2. dereceden GMM kullanan sözcük türü tespit yöntemi kullanılmıştır. Başarım oranı %90 olan sözcük türü tespit yönteminin sadece isim ve fiil ayrıştırmadaki başarı oranına bakıldığında ise bu oran %96 olarak tespit edilmiştir. Bu başarı oranı gövdeleme yöntemimizde kullanmak için oldukça iyi bir yüzdendir.

İstatistiklerin çekildiği verilere ve hesaplama birimi olarak kullanılan harf sayılarına göre belirlenmiş 10 farklı varyasyon için farklı talim derlemi boyutlarından elde edilen sonuçlar aşağıdaki grafikte verilmiştir.



Şekil 5.6. Sözcük türü belirlenen kelimeler üzerinden gerçekleştirilen deney sonuçlarının grafiği

Yukarıda verilen grafiğe göre en iyi başarıyı 3 numaralı varyasyon göstermiştir. Sonuçların başarı oranlarının bir önceki deneyden düşük olmasının en büyük nedeni sözcük türü tespitindeki %90 lık başarının sonuçlar üzerine etkisidir. Sonuçlarda dikkat edilmesi gereken farklı bir noktada 4 numaralı yöntem yerine 3 numaralı yöntemin başarılı çıkmasıdır. Bu durum 3 numaralı yöntemin tüm sözcük türleri yerine kelimenin isim ve fiil gruplarını temel alarak çalışması şeklinde açıklanabilir. Bir önceki deneyimizle benzer olarak 10 numaralı

varyasyon deneyimizde kullanılan istatistiklerin sadece gövde-ek geçiş istatistikleri olmasına rağmen elde ettiği başarıım oranının kayda değer olduğu söylenebilir.

Yapılan deneylerin sonucu olarak gövdeleme yönteminin %93,83 oranında bir doğruluk ile çalıştığı belirlenirken bütünleşik olarak çalışan sözcük türü tespiti ve gövdeleme yönteminin %92,1 oranında başarıya sahip olduğu tespit edilmiştir.

Çizelge 5.16. Gövdeleme yöntemi başarıım oranları (Bootstrapping yöntemi ile yapılan test)

	Sözcük Türü Elle İşaretlenmiş	Sözcük Türü Otomatik Olarak İşaretlenmiş
Gövde Tespiti	%93,83 (Varyasyon 4)	%92,10 (Varyasyon 3)

Gövdelemenin başarısı k-parça çapraz doğrulama (k-fold cross-validation) yöntemiyle de test edilmiştir. Bunun için derlem 10 parçaya ayrılarak her seferinde farklı 1 parça test derlemi kalan 9 parça ise talim derlemi olarak kullanılmaktadır. Yani, yaklaşık 7500 cümleden oluşan ODTÜ-Sabancı derlemi 750 cümlelik 10 parçaya bölünmüştür.

Deneyde, kelimelerin ODTÜ-Sabancı derlemindeki daha önceden işaretlenmiş sözcük türleri kullanılmıştır. 10 farklı varyasyon için 10 farklı test derlemi üzerinden elde edilen sonuçlar aşağıdaki çizelgelerde verilmiştir (Çizelge 5.17 ve Çizelge 5.18).

Çizelge 5.17. Sözcük türü bilinen kelimeler üzerinden gerçekleştirilen gövdeleme yönteminin başarımlar oranları (k-parça çapraz doğrulama yöntemi ile yapılan test)

Parça No	Varyasyon 1	Varyasyon 2	Varyasyon 3	Varyasyon 4	Varyasyon 5
1	%88,11	%89,46	%93,24	%93,91	%91,77
2	%87,84	%90,51	%93,38	%93,84	%92,08
3	%89,01	%90,27	%93,91	%93,77	%92,45
4	%89,32	%90,44	%93,67	%93,65	%92,87
5	%88,44	%90,67	%93,58	%92,98	%92,42
6	%87,77	%87,66	%92,67	%92,82	%91,94
7	%89,47	%90,02	%92,94	%93,41	%92,06
8	%88,36	%90,61	%93,98	%93,87	%91,79
9	%88,18	%90,53	%93,57	%94,15	%92,47
10	%88,39	%90,91	%93,86	%94,33	%92,98
Ortalama	%88,49	%90,11	%93,48	%93,67	%92,28

Çizelge 5.18. Sözcük türü bilinen kelimeler üzerinden gerçekleştirilen gövdeleme yönteminin başarımlar oranları (k-parça çapraz doğrulama yöntemi ile yapılan test)

Parça No	Varyasyon 6	Varyasyon 7	Varyasyon 8	Varyasyon 9	Varyasyon 10
1	%93,23	%80,01	%86,77	%90,89	%92,84
2	%93,18	%80,34	%86,03	%90,71	%92,93
3	%92,98	%80,44	%86,37	%90,48	%93,71
4	%93,45	%79,89	%85,58	%90,55	%93,20
5	%93,21	%80,02	%85,91	%91,13	%92,87
6	%91,94	%80,31	%85,87	%90,21	%89,69
7	%92,82	%81,51	%86,73	%91,02	%91,84
8	%92,57	%81,19	%86,95	%91,20	%92,91
9	%92,94	%80,82	%86,38	%90,74	%92,89
10	%93,37	%80,07	%86,61	%90,36	%93,29
Ortalama	%92,97	%80,46	%86,32	%90,73	%92,62

Diğer bir deneyde ise kelimelerin sözcük türleri otomatik olarak belirlenmiştir. Gövdeleme işleminde 10 farklı varyasyon için elde edilen başarımlar aşağıdaki çizelgelerde verilmektedir (Çizelge 5.19 ve Çizelge 5.20).

Çizelge 5.19. Sözcük türü belirlenen kelimeler üzerinden gerçekleştirilen gövdeleme yönteminin başarımlar oranları (k-parça çapraz doğrulama yöntemi ile yapılan test)

Parça No	Varyasyon 1	Varyasyon 2	Varyasyon 3	Varyasyon 4	Varyasyon 5
1	%87,45	%88,12	%92,27	%90,55	%91,02
2	%86,91	%87,99	%92,31	%91,15	%90,91
3	%86,44	%87,33	%91,95	%90,94	%90,13
4	%86,32	%87,81	%92,30	%90,66	%90,09
5	%86,15	%87,48	%92,24	%91,46	%90,33
6	%85,74	%86,06	%90,88	%89,92	%89,67
7	%86,65	%87,34	%91,96	%90,86	%90,72
8	%86,85	%87,78	%91,89	%90,75	%90,69
9	%86,76	%88,17	%92,23	%90,77	%90,34
10	%86,88	%87,94	%92,05	%91,42	%90,48
Ortalama	%86,62	%87,60	%92,01	%90,85	%90,44

Çizelge 5.20. Sözcük türü otomatik belirlenen kelimeler üzerinden gerçekleştirilen gövdeleme yönteminin başarımlar oranları (k-parça çapraz doğrulama yöntemi ile yapılan test)

Parça No	Varyasyon 6	Varyasyon 7	Varyasyon 8	Varyasyon 9	Varyasyon 10
1	%89,43	%80,01	%84,09	%90,13	%91,02
2	%90,56	%79,55	%84,55	%89,66	%90,36
3	%89,78	%78,83	%84,81	%90,64	%89,95
4	%90,55	%79,92	%83,40	%89,67	%90,72
5	%90,39	%80,51	%84,36	%90,11	%90,41
6	%89,25	%79,77	%83,34	%89,25	%89,79
7	%89,53	%79,83	%84,41	%90,21	%90,44
8	%89,78	%80,33	%84,53	%90,42	%90,93
9	%89,91	%80,41	%85,34	%90,63	%91,25
10	%90,18	%80,06	%84,66	%90,23	%91,44
Ortalama	%89,94	%79,92	%84,35	%90,10	%90,63

Her iki test yönteminde de elde edilen başarımlar oranları sözcük türü bilinen kelimeler ve sözcük türü otomatik olarak belirlenen kelimeler üzerinden sırasıyla yaklaşık %93 ve %92’ dir.

Gerçekleştirdiğimiz deneyler sonrası elde ettiğimiz başarı oranının diğer çalışmalar ile kıyaslamasını yapabilmek amacıyla literatüre bölümünde de bahsettiğimiz Sever ve Bitirim (2003)’ in gövde-bul (FindStem) yöntemi kullanarak test derleminde gövdeleme işaretlemesi yaptırılmıştır. Gövde-bul yönteminin, test derleminden rasgele olarak elde edilen 30 adet 200 cümlelik alt derlemden elde edilen başarı oranlarının ortalamasının hesaplanması sonucunda yaklaşık %80 ’lık bir başarı elde ettiği görülmüştür. Bu başarı oranı geliştirdiğimiz yöntem ile karşılaştırıldığında yaklaşık %13 oranında daha düşüktür. Birçok doğal dil işleme ve bilgi geri getirim konusunun temelinde yer alan gövdeleme

probleminde elde edilen bu gelişimin oldukça önemli bir fark olduğu söylenebilir.

6. Tümleşik Biçimbirimsel Çözümleme ve Sözcük Türü Tespiti Yöntemi

Biçimbirimsel analiz, kelimenin kök, gövde ve eklerinin belirlenmesi şeklinde tanımlanabilmektedir. Doğal dil işleme sahasının en temel problemlerinden olan biçimbirimsel çözümleme, üzerinde çalışan alanlara sağladığı bilgi açısından oldukça önemlidir. İyi tasarlanmış ve ihtiyaçlara cevap verecek doğrulukta ve hızda çalışan bir biçimbirimsel çözümleme mekanizmasının bilgi geri getirim, konuşma anlama, diller arası otomatik çevirim gibi alanların başarısında doğrudan katkısı bulunmaktadır.

Türkçe için bilimiz dahilinde çok fazla sayıda biçimbirimsel çözümleme çalışması bulunmamaktadır. Varolan biçimbirimsel çözümleme çalışmalarının hemen hemen hepsinde sonlu durum makineleri (SDM) ve bir kök/gövde sözlüğü kullanılmaktadır. Yöntemlerin kurgularında yer alan bu sözlükler, eşleme yapılarak çözümlenecek kelimelerin gövde veya köklerinin belirlenmesi amacı ile kullanılmaktadır. Bu nedenle sözlükte bulunmayan bir kök veya gövde ile karşılaşıldığında biçimbirimsel çözümleme gerçekleştirilememektedir. Belirtilenlere ek olarak sözlük kullanımının mümkün olması için yapılan çalışmalarda genelde kuramsal olarak kelime dağarcığının kapalı olduğu kabullenmesi yapılmaktadır. Bir dilin kelime dağarcığının kapalı olduğundan söz edilebilmesi için o dilde kullanılabilecek tüm kelime biçimlerinin bilinmesi veya ulaşılabilmesi gerekmektedir. Bu kabullenme Türkçe gibi sadece bir fiilden bile binlerce yüzeysel biçimin türetilbildiği diller için çok doğru olmamaktadır.

Bunlara ek olarak biçimbirimsel çözümleyicinin temel hesaplama biriminin kelime olması nedeniyle yazıda geçen bir kelime şekli çoğunlukla birden fazla biçimbirimsel çözümleme seçeneğine sahiptir.

Bunun yanı sıra geliştirilen biçimbirimsel çözümleme yöntemleri yüksek sayılabilecek bir karmaşıklığa sahiptir.

Biçimbirimsel çözümlemede içerisinde yer alan başka bir belirsizlikte, bir ek şeklinin sadece tek bir dilbilgisi özelliği için kullanılmamasıdır. Örneğin –a eki hem yapım eki hem de çekim eki olarak kullanılabilmektedir; ayrıca –a eki çekim eki olarak hem isimlere hem de fiillere gelebilmektedir.

Örnek:

Koşa koşa okula yetiştim.

(koş (fiil) + a (Gerendium eki))

Taşa yalın ayak basmazsan hasta olmazsın.

(Taş (isim) + a (isim çekim eki))

Örnekte de görüldüğü üzere, kelimenin cümle içinde bulunduğu yer (veya kelimenin sözcük türü) dikkate alındığında bu tür eş şekilli farklı göreve sahip olan eklerin de ayrımı yapılabilir. Daha doğrusu bir ek hem yapım eki hem de çekim eki olabilir, ayrıca hem isim için hem de fiil için aynı ek şekline sahip olabilir, ancak kelimenin kullanıldığı yer itibari ile bu seçeneklerden sadece bir tanesi geçerlidir.

Yukarıda bahsedilen sebepler göz önünde bulundurularak, mevcut biçimbirimsel çözümleme yöntemlerinin kelimeleri cümle içerisinde değil de tek başlarına çözümlemelerinden, olanaksız alternatifleri de değerlendirdikleri söylenebilir.

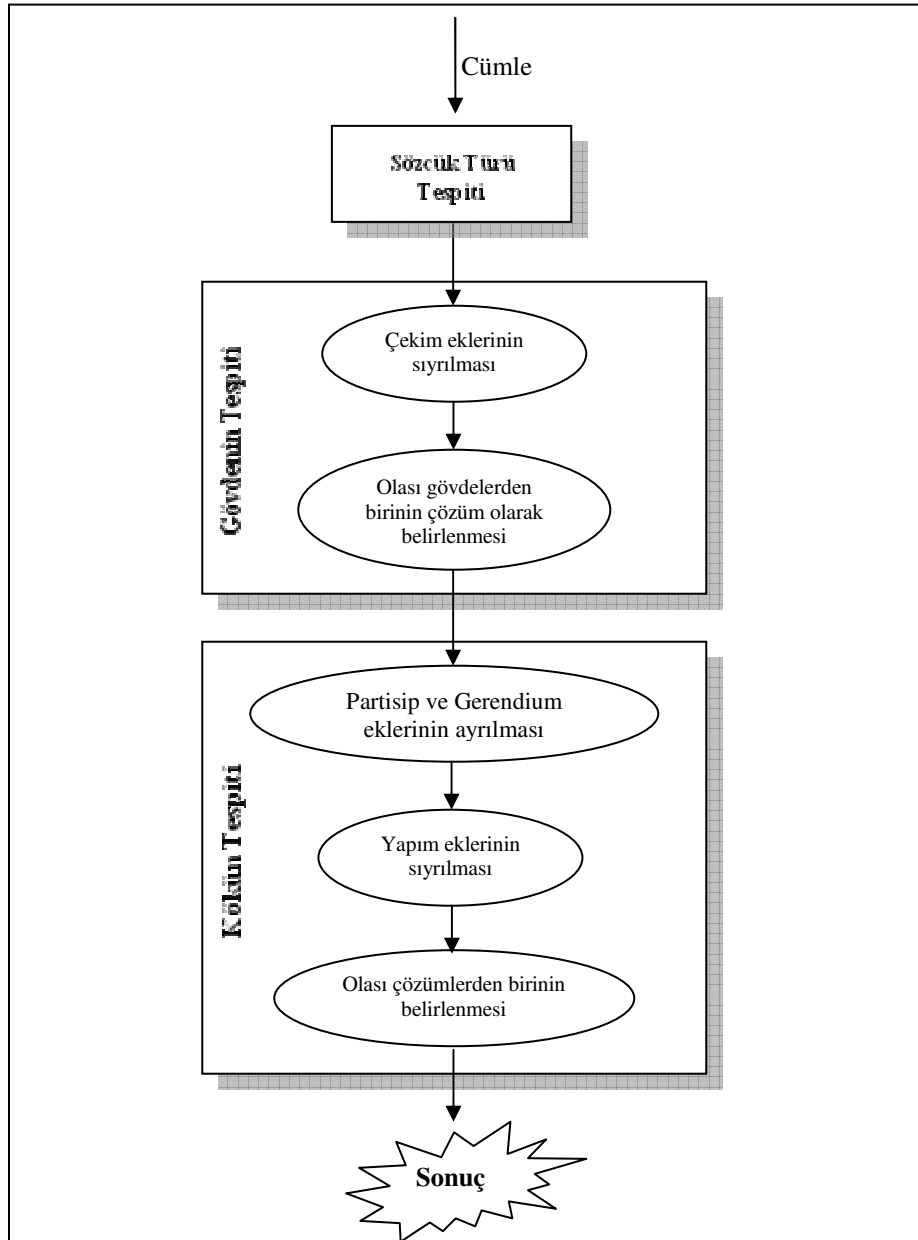
Bu tez çalışmasında ortaya konan iddia, kelimelerin dilbilgisi işlevlerinin önceden belirlenmesi sayesinde, basitleştirilmiş ve doğruluk oranı artmış tümleşik bir biçimbirimsel çözümleme ve sözcük türü tespit yönteminin geliştirilebileceğidir.

Türkçe için geliştirilmiş ilk ve neredeyse tek biçimbirimsel çözümleme yöntemi Oflazer (1993) tarafından gerçekleştirmiştir. Bu yöntem tam bir biçimbirim çözümleme gerçekleştirerek Türkçe kelimelerin olası tüm kök ve gövdelerini bulabilmektedir. Oflazer Türkçe'nin biçimbirimsel analizi üzerine yaptığı çalışmalarda Koskenniemi'nin doktora çalışmasında ortaya koyduğu iki seviyeli biçimbirim çözümleme yapısından yararlanmıştır. Oflazer, Türkçenin biçimbirimsel analizini gerçekleştirmek için yaklaşık 23.000 kök veya gövdelerden oluşan bir sözlük, Türkçenin ek sistemi ile ilgili dil bilgisi kurallarını içeren 22 iki-seviyeli kural ve bu kuralların üzerinde işleyecek fiiller ve isimler için sonlu-durum makinelerini oluşturmuştur. Oflazer Türkçe üzerine yaptığı çalışmaların dilbilgisi temelini Lewis ve Underhill'in Türkçe grameri üzerine yazdıkları kitaplar üzerine oturtmuştur.

Oflazer'in biçimbirimsel çözümleyicisi, bir sözlük kullanması, tek bir sonuç üretmemesi yani belirsizlik giderilmesine ihtiyaç duyulması ve kullandığı, iki-seviyeli dil modellemesinin teorik olarak hesaplanabilme açısından "NP-complete" bir mesele olması (Barton, 1986) nedeni ile Sever (2003)'in ve Dinçer (2004)'in de belirttiği gibi bilgi geri getirim gibi sahalarda kullanılması uygun değildir.

6.1. Yöntem

Türkçe kelimelerin biçimbirimsel çözümlemesi için geliştirdiğimiz yöntem, sadece belirtilen ek kalıplarını içeren kapalı bir sözlük, kelimenin sözcük türünü dikkate alan, ek sıyırma yöntemi ile çalışan ve sonuç olarak sadece bir çözüm sunan bir yöntem olarak tarif edilebilir. Sistemin genel görünümü Şekil 6.1'de verilmektedir.



Şekil 6.1. Geliştirilen yöntemin genel görünüşü

Yöntem içerisinde yer alan sözcük türü tespiti ve çekim eklerinin ayrıştırılması önceki bölümlerde detaylı bir şekilde anlatıldığından bu bölümde sadece kökün ve yapım eklerinin tespitine yer verilmiştir.

6.1.1. Yapım ekleri

Eklendiği kökün veya gövdenin anlamında değişiklik yaparak yeni bir gövde meydana getiren eklere verilen genel isimdir. Yapım ekleri isim veya fiil soylu köklere veya gövdelere eklenerek isim veya fiil gövdesi oluşturabilirler. Dolayısı ile dört çeşit yapım eki mevcuttur.

- ✓ İsimden isim yapma ekleri.
- ✓ İsimden fiil yapma ekleri.
- ✓ Fiilden isim yapma ekleri.
- ✓ Fiilden fiil yapma ekleri.

Yapım eklerinin birçoğu belli bir sıra izlemeden kök veya gövdeye eklenebilmesi nedeniyle yapım ekleri için çekim eklerinde olduğu gibi ek kalıpları oluşturma yoluna gidilmemiştir. Bunun nedeni olarak yapım ekleri için ek kalıplarını oluşturma zor olması gösterilebilir.

Örnek: göz-lük-çü ve göz-cü-lük kelimelerinde olduğu gibi.

Bu nedenle yapım eklerinin sıyrılma işleminde ek kalıpları oluşturulmamış ekler tek tek sıyrılmıştır. Türk dili kitaplarından derlenen 146 yapım eki 6 grup içerisinde sınıflandırılmıştır. Eklerin Türkçede yer alan büyük sesli uyumu, hece düşmesi, yardımcı ses eklemesi gibi bazı kurallar işletilerek elde edilen tüm formları hesaba katıldığında 575 adet farklı form elde edilmiştir.

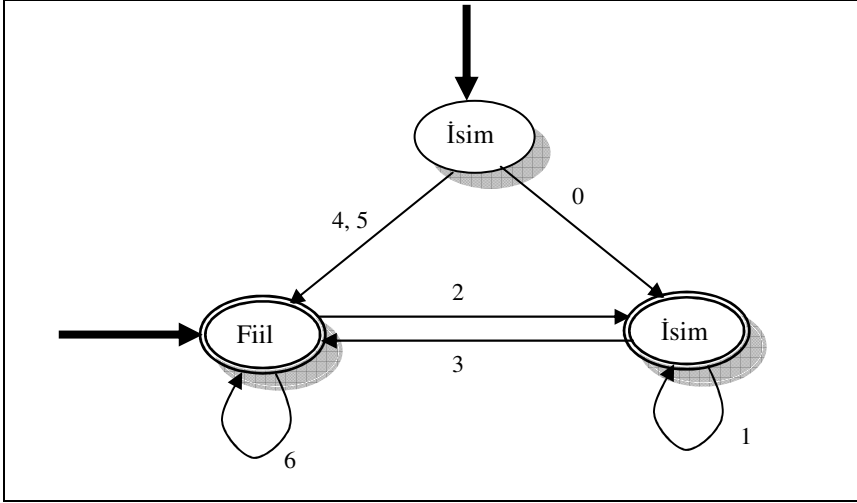
Çizelge 6.1. Yapım ekleri için oluşturulan gruplar

Grup numarası	Yapım ek çeşitleri	Ek Sayıları
1	İsimden İsim Yapma Ekleri	44
2	İsimden Fiil Yapma Ekleri	12
3	Fiilden İsim Yapma Ekleri	42
4	Partisip ekleri (isim fiil ekleri) – Fiilden İsim Yapma Ekleri	9
5	Gerendium ekleri (zarf fiil ekleri) – Fiilden İsim Yapma Ekleri	20
6	Fiilden Fiil Yapma Ekleri	19

4 ve 5 numaralı grupta bulunan partisip ve gerendium eklerinin bir yapım eki mi yoksa bir çekim eki mi olduğu tartışmaları hala sürmektedir. Tez kapsamında bu ekler kelimenin sözcük türünü ve cümle içerisinde görevini değiştirdiği için yapım ekleri kapsamında ele alınmıştır. Ayrıca yine Türkçede oldukça sık kullanılan birleşik fiiller (-Abil-, -Uver-, -Akal-, -Ayaz- vb.) “fiilden fiil yapma ekleri” kategorisine dahil edilmiştir.

Yapım eklerinin gövdeden ayrılması amacıyla tasarlanan sonlu durum makinesi Şekil 6.2’de gösterilmiştir. Şekilden de görüleceği gibi tasarlanan SDM’lar aynı problemi çözen mevcut SDM’larına göre oldukça basit bir yapıdadır.

Şekil 6.2. Yapım ekleri için tasarlanan SDM



Sonlu durum makinesinde okların üzerinde bulunan ek grup numaraları (Çizelge 6.1) geçiş koşulları olarak belirlenmiştir. “0” olarak belirlenen geçiş koşulu ise boş geçişleri temsil etmektedir. Elips içerisinde yer alan isim ve fiil ifadeleri SDM için durumları göstermektedir. Kalın oklar ile gösterilen bölümler SDM’ye girdi olarak verilen kelimenin sözcük türüne göre farklı durumlarını göstermektedir. SDM’lerinin sona ereceği durumlar ise çift çizgili elipsler ile belirlenen durumlardır.

6.1.2. Yapım Eklerinin Sıyırılması

Yapım eklerinin sıyırılma işleminde, gövde sağdan harf harf ilerlenerek her bir harf birliği ek listesi içerisinde aranmaktadır. Arama sonucunda harf birliği ek listesi içerisinde bulunursa bulunan ekin grubuna göre sonlu durum makinesinde durum geçişleri gerçekleştirilmektedir. Bu işlem kelimedeki harfler bitene kadar devam ettirilmektedir. İşlemlerin sonucunda analiz edilen kelimeye ait olası kökler ve bu köklere gelen yapım ekleri belirlenmektedir. Belirlenen bu

analizlerden hangisinin en uygun çözüm olduğu bir sonraki aşamada belirlenmektedir.

6.1.3. Uygun Çözümün Belirlenmesi

En uygun çözümün belirlenmesi aşaması istatistik/olasılık tabanlı bir yöntem olup daha önceden belirlenen olası çözümlerden en uygununun seçilmesini hedefler. Anlatımda kullanılan simge gösterimleri ve açıklamaları şöyledir:

$$k_n = h_1 h_2 \cdots h_n \quad (6.1)$$

n harften oluşan bir kelimenin gösterimi: n : kelimenin uzunluğu, $n > 0$, h : harf gösterimi

$$g_m = h_1 h_2 \cdots h_m \quad (6.2)$$

m harften oluşan gövdenin gösterimi: m : gövdenin uzunluğu, $n \geq m > 0$

$$e_p = h_{m+1} h_{m+2} \cdots h_n \quad (6.3)$$

p harften oluşan çekim eklerin gösterimi: p : gövdenin uzunluğu, $n > p \geq 0$, $n = m + p$

$$r_s = h_1 h_2 \cdots h_s \quad (6.4)$$

s harften oluşan kökün gösterimi: s : kökün uzunluğu, $n \geq m \geq s > 0$

$$y_t = h_{s+1} h_{s+2} \cdots h_{s+t} \quad (6.5)$$

t harften oluşan yapım eklerin gösterimi: t : kökün uzunluğu, $n > t \geq 0$

$$k_n = g_m e_p, \quad k_n = h_1 h_2 \cdots h_m h_{m+1} h_{m+2} \cdots h_n$$

Kelimenin gösterimi veya $k_n = r_s y_t e_p, \quad (6.6)$

$$k_n = h_1 h_2 \cdots h_s h_{s+1} h_{s+2} \cdots h_m h_{m+1} h_{m+2} \cdots h_n$$

$$R = \{ (h_i, h_{i+1}) \mid h_i \in r_s \wedge h_{i+1} \in r_s \wedge 0 < s \leq n \} \quad (6.7)$$

İki harf çiftinin kökte olma durumu:

*Bir harfin kökte
diğer harfin
ekte olma
durumu:*

$$D = \left\{ (h_i, h_{i+1}) \mid h_i \in r_s \wedge h_{i+1} \in y_t, e_p \wedge 0 < t < n \right\} \quad (6.8)$$

Yukarıdaki tanımlamalar kullanılarak, verilen bir kelimenin $(k_n = h_1 h_2 \cdots h_n)$ içerisinde yer alan herhangi bir harf çiftinin (h_i, h_{i+1}) kökün bir parçası olma olasılığı, ek sisteminin bir parçası olma olasılığı ve kök ile ek sisteminin geçişinde olma olasılığının hesabında kullanılacak simgeler ve olasılık hesapları aşağıda tanımlanmıştır.

$$P((h_i, h_{i+1}) \in R) = P_R((h_i, h_{i+1})) = w_{r,i} / N \quad (6.9)$$

$$P((h_i, h_{i+1}) \in D) = P_D((h_i, h_{i+1})) = w_{d,i} / N \quad (6.10)$$

Hesaplama kullanılan semboller:

$$f_{R,i} : (h_i, h_{i+1}) \text{ harf çiftinin istatistiklerin çekileceği derlemdeki kelimelerde } i. \text{ sırada ve kökte görülme sayısı} \quad (6.11)$$

$$f_{D,i} : (h_i, h_{i+1}) \text{ harf çiftinin istatistiklerin çekileceği derlemdeki kelimelerde } i. \text{ sırada ve kök-ek geçişinde görülme sayısı} \quad (6.12)$$

$$w_{R,i} : f_{R,i} / (f_{R,i} + f_{e,i} + f_{D,i}) \quad (6.13)$$

$$w_{D,i} : f_{D,i} / (f_{R,i} + f_{e,i} + f_{D,i}) \quad (6.14)$$

$$N : \text{Tüm } (h_i, h_{i+1}) \text{ harf çiftlerinin toplam sayısı} \quad (6.15)$$

Tanımlanan bu simgeler ile olası tüm kök sonuçlarına bir olasılık değeri atanmaktadır.

$$P_{\text{Kök}}(j) \quad j. \text{ olası kök indisi,} \\ 1 \leq j \leq nk \quad (nk: \text{kelimenin tespit edilen olası kök sayısı.})$$

Bu olasılık değeri kökün son elemanın sırası i olmak üzere şu şekilde hesaplanmaktadır:

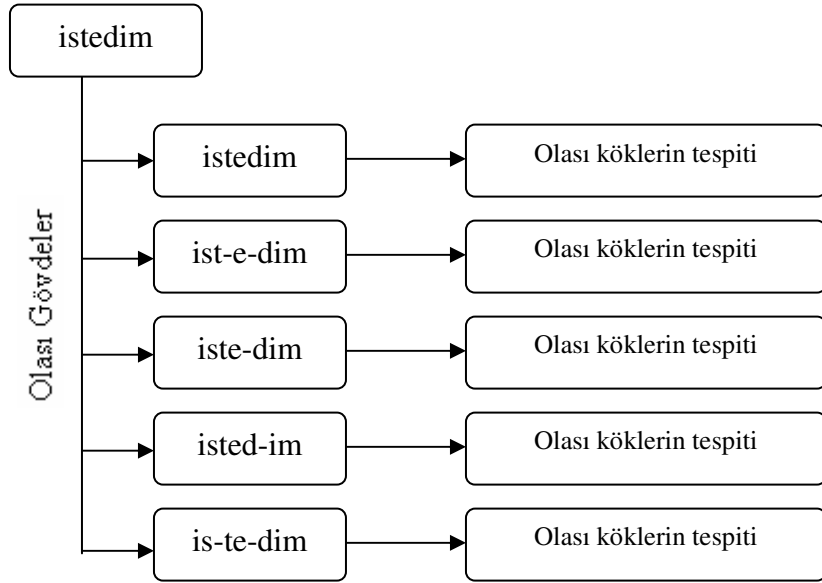
$$P_{\text{Kök}}(j) = P_R((h_{i-1}, h_i)) * P_D((h_i, h_{i+1})) \quad (6.16)$$

Her olası kök için bu olasılık değerleri hesaplanmaktadır. Yöntemin bir sonraki aşamasında kök olma olasılıklarının belirlenmesi için 10 farklı varyasyon kullanılmıştır. Belirtilen bu varyasyonlar gövdeleme için yapılanların aynısıdır. Bunlar; istatistiklerin çekildiği verilere ve hesaplama birimi olarak kullanılan harf sayılarına göre belirlenmiştir. Aşağıdaki çizelgede örnek olarak “gözlükçüler” kelimesinde olası kökün “göz” olma durumu ile ilgili hesaplamalar verilmiştir.

Çizelge 6.2. “gözlükçüler” kelimesi için 10 farklı varyasyon üzerinden kök olma olasılığı

Varyasyon No	Hesaplama Birimi	İstatistik/olasılık verileri	Olasılık Hesaplaması ($P_{Kök}(j)$)
1	İkili harf dizileri	Fiil ve isme göre	$P_R(('ö', 'z')) * P_D(('z', 'l'))$
2	İkili harf dizileri	Tüm sözcük türlerine göre	$P_R(('ö', 'z')) * P_D(('z', 'l'))$
3	Dörtlü harf dizileri	Fiil ve isme göre	$P_R((' ', 'g', 'ö', 'z')) * P_D(('ö', 'z', 'l', 'ü'))$
4	Dörtlü harf dizileri	Tüm sözcük türlerine göre	$P_R((' ', 'g', 'ö', 'z')) * P_D(('ö', 'z', 'l', 'ü'))$
5	İkili ve dörtlü harf dizileri	Fiil ve isme göre	$P_R(('ö', 'z')) * P_B(('ö', 'z', 'l', 'ü'))$
6	İkili ve dörtlü harf dizileri	Tüm sözcük türlerine göre	$P_R(('ö', 'z')) * P_B(('ö', 'z', 'l', 'ü'))$
7	İkili harf dizileri	Fiil ve isme göre	$P_D(('z', 'l'))$
8	İkili harf dizileri	Tüm sözcük türlerine göre	$P_D(('z', 'l'))$
9	Dörtlü harf dizileri	Fiil ve isme göre	$P_D(('ö', 'z', 'l', 'ü'))$
10	Dörtlü harf dizileri	Tüm sözcük türlerine göre	$P_D(('ö', 'z', 'l', 'ü'))$

Biçimbirimsel analiz yapılırken iki farklı yol izlenmiştir. Bunlardan ilkinde gövdeleme işleminden gelen tüm olası gövdelerin yapım ekleri sıyrılmıştır (Şekil 6.3).



Şekil 6.3. Gövdeleme işleminden elde edilen tüm olası gövdelerin kök tespitinde kullanımı

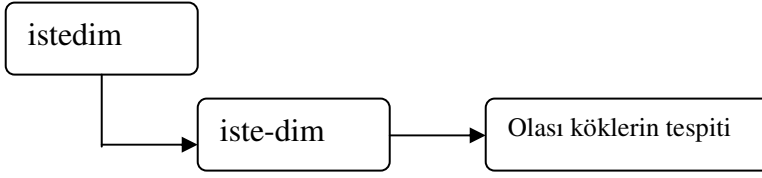
Böyle bir sistemden çıkan tüm olası çözümler arasından en uygununu bulmak için;

$$\operatorname{argmax}_{1 \leq j \leq n} (P_{Kok}(j) * P_K(j)), \quad (6.17)$$

n: olası çözüm sayısı

kullanılmaktadır.

Diğer yolda ise sadece gövdeleme yönteminden sonuç olarak ortaya çıkan gövde işleme tabi tutulmuştur (Şekil 6.4).



Şekil 6.4. Gövdeleme işleminden elde edilen gövdenin kök tespitinde kullanımı

Bu yol tercih edildiğinde ise sistemden çıkan tüm olası çözümler arasından en uygununu bulmak için;

$$\operatorname{argmax}_{1 \leq j \leq n_k} (P_{Kok}(j)) \quad (6.18)$$

kullanılmaktadır.

Yöntemde şimdiye kadar sunulanlarda bir kelimenin gövdesinin ve kökünün doğru bulunması hedeflenmiştir. Fakat kök ve gövde doğru tahmin edilse dahi yapım eklerinin ayrıştırılması başka bir problem olarak karşımıza çıkmaktadır. Bu problem bir örnek üzerinde şu şekilde açıklanabilir.

Örnek: “Osman-lı-la-ş-tır”

“Osman-lı-la-ş-tı-r”

Örnek: “Yaz-ı-lı”

“Yaz-ıl-ı”

“Yaz-ı-l-ı”

İki örnekte de sunulan çözümler Türkçe dilbilgisi kurallarına göre mümkün olabilmektedir. Fakat, esas çözüm sadece bir tanedir. Ek sıralanışlarından hangisinin doğru olduğunun tespiti bu noktada önem kazanmaktadır.

Tez kapsamında bu problemin çözümü için istatistik/olasılık tabanlı bir çözüm önerilmektedir. Çözümleme sonucu bulunan kök ve gövde arasında kalan yapım ekleri bir harf dizisi olarak belirlenip bu dizinin

talim derlemi içerisindeki olası tüm ek sıralanışları kullanılarak bir olasılık değeri elde edilmektedir. Bu olasılık değeri en yüksek olan sıralanış uygun çözüm olarak belirlenmektedir. Yöntemde kullanılan simgeler aşağıda verilmiştir.

$$P_{yD}(y_t) = w / N \quad (6.19)$$

f_{yD} : İstatistiklerin çekileceği derlemdeki, t harften oluşan yapım eklerinin (y_t), istenilen yapım eki dizilişinde görülme sayısı,

f_{yT} : İstatistiklerin çekileceği derlemdeki t harften oluşan yapım eklerinin (y_t) görülme sayısı

N : Derlemde elde edilen tüm yapım ekleri dizilerinin sayısı (6.20)

$$w = f_{yD} / f_{yT}$$

6.1.3.1. Kökteki Dilbilgisel Değişimler

Türkçe’de eklerin köklere veya gövdelere eklenmesi sırasında büyük ünlü uyumu, seslerin yumuşaması, harf düşmesi, kaynaşma harfinin eklenmesi, ünlü daralması, sessiz benzeşmesi gibi bazı durumlar söz konusu olmaktadır.

sev- + mAk	=>	sevmek
sokak + DA	=>	sokakta
gel- + yor	=>	gel(i)yor
kitap + I	=>	kitabı
burun + U	=>	burnu
ara + yor	=>	arıyor

Bu gibi durumlar ekte veya kökte meydana gelebilmektedir. Örnekte verilen son üç kelimedede oluşan değişimler köke etki ederken ilk üç değişiklik ekte meydana gelmektedir. Eklerde meydana gelen değişimler, kullanılan çekim ek kalıpları veya yapım ek listelerinin içerisine eklenmiştir. Bu yapı sayesinde eklerde oluşan değişimlerin tespiti kolaylıkla gerçekleştirilebilmektedir.

Fakat, yöntemimiz içerisinde bir kök veya gövde sözlüğü kullanılmadığından, kökte meydana gelen değişimlerin tespiti için farklı yaklaşımlar kullanılmaktadır.

Genelde yabancı kökenli (özellikle Arapça ve Farsça) kelimelerde karşılaşılan harf düşmesinin tespiti amacı ile sayıları yüzlerle ifade edilebilecek kelimelerin bir listesi oluşturulmuştur. Bu aşamadan sonra değişim sonrası oluşan formların son iki harfleri tespit edilmiştir (br,cm, ğl, ğr, ğs, ğz, hn, hr, hs, kf, kl, kr, kt, ln, mr, mz, nl, rn, ps, sm, sl, th, tn, vr, yb, yd, yf, yn, yr).

cisim => cism - i

burun => burn - u

göğüs => göğs - ü

ağız => ağız - ı

akıl => akıl - ı

Sonuç olarak 29 harf ikilisinden oluşan liste, kökte gerçekleşen harf düşmelerinin tespiti amacı ile kullanılmaktadır. Biçimbirimsel çözümleme sonucu elde edilen kökün son iki harfinin listede olma durumuna göre harf düşmesi tespit edilmektedir.

Kökte yaşanan bir başka değişim ise kökün –yor eki alması ile meydana gelmektedir. Son hecesi geniş sesliler (a-e) ile biten fiil

köklerine –yor eki getirildiğinde geniş sesliler daralarak (ı-i-u-ü)’ye dönüşmektedir.

ara + yor => arıyor

uza + yor => uzuyor

Biçimbirimsel çözümleme sonucunda –yor ekinin köke eklendiğinin tespit edilmesi ile kökte gerekli harf değişimi uygulanarak bu problem çözüme kavuşturulmaktadır.

Kökte yaşanabilecek diğer bir değişim ise sert sessizle (p, ç, t, k) biten köklere, sesli harfle başlayan bir ek eklenmek istendiğinde kökün son harfi yumuşayarak b, c, d, g harflerinden birine dönmesidir. Bu kurala sert sessizlerin yumuşaması denir.

kitap**p**+I => Kitab**ı**

ağaç**ç**+I => Ağac**ı**

tabak**k**+A => Tabağ**a**

tat**t**+A => Tada**ı**

Bu kurala uymayan bazı kelimeler de vardır. Bunların büyük bir bölümü sonu sert sessiz ile biten tek heceli kelimeler ve fiil kökleridir.

süt+U => Sütü

unut+AcAk => Unutacak

kalk+DU => Kalktı

Bu problemin çözümü için sunulan yaklaşımda biçimbirimsel çözümleme sonucu bulunan kökün,

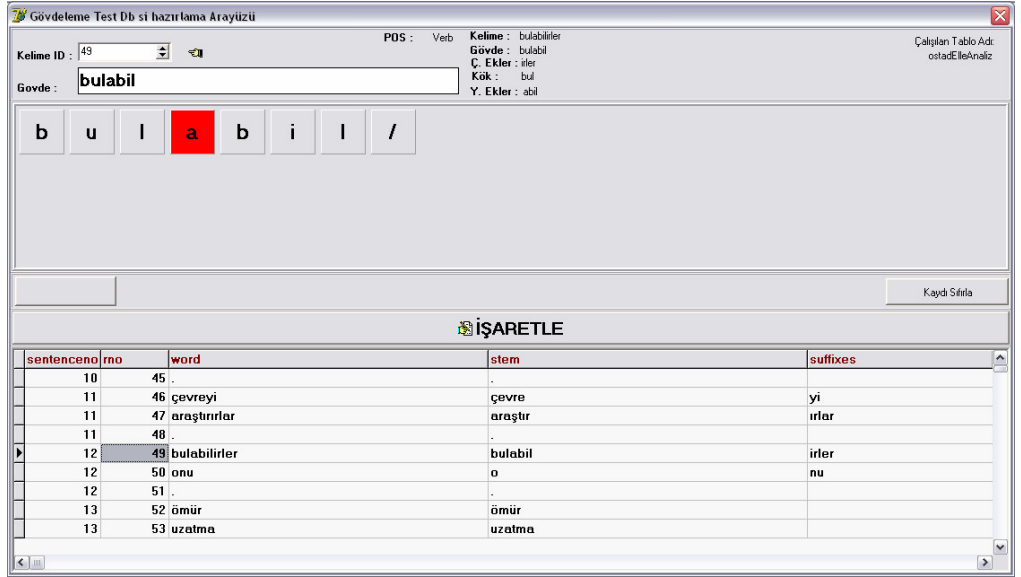
- ✓ tek heceli olmaması,
- ✓ son harfinin b, c, d, ğ olması durumunda

kelimenin son harfinin yerine p, ç, t, k harflerinden bir yerleştirilmiştir. Önerilen çözüme gerekçe olarak, genellikle Türkçe kelimelerin sonlarında yumuşak sessizlerden b, c, d, ğ bulunmaması gösterilebilir. Bunun ispatı için yapılan incelemede Türk Dil Kurumu sözlüğü içerisinde yaklaşık 120 adet son harfi yumuşak sessizle biten kelimeye rastlanmıştır. Bu kelimelerin de büyük bir kısmının da tek heceli olduğu görülmüştür.

6.2. Yöntemin Başarım Oranlarının Tespiti

6.2.1. Kullanılan Derlemler

Tez kapsamında geliştirilen biçimbirimsel analiz yönteminin başarısının tespiti için kelimelerin sözcük türlerinin, gövdelerinin, köklerinin ve eklerinin işaretlenmiş olduğu bir talim derlemine, aynı zamanda üzerinde yöntemimizin başarımının sınanması amacıyla kullanacağımız test derlemine ihtiyaç duyulmaktadır. Bu amaca yönelik olarak ODTÜ-Sabancı Türkçe derlemi kullanılmıştır. Gövdeleme çalışması sırasında elle gövdelenen derlemdeki kelimelerin kökleri ve ekleri de yine elle işaretlenmiştir. İşaretleme için kullanılan aracın ekran görüntüsü şekil 5.6'da verilmektedir.



Şekil 6.5. ODTÜ-Sabancı Türkçe derleminin köklerinin elle belirlenmesine yardımcı araç

6.2.2. Deneysel Sonuçlar

Yöntemin başarısının tespiti için yapılacak deneylerde iki farklı test yöntemi kullanılmıştır: bootstrapping ve çapraz doğrulama (cross validation).

İlk test yönteminde elle gövdelenilen ODTÜ-Sabancı Türkçe derleminden elde edilen alt derlemler kullanılmıştır. ODTÜ-Sabancı Türkçe derlemi bu amaç için ikiye bölünmüş, 6000 cümle talim derlemi için 1462 cümle ise deney derlemi için ayrılmıştır.

6000 cümleden oluşan talim derleminden, istatistiklerin çekildiği derlemlerin boyutunun başarıya olan katkısının belirlenmesi amacı ile 15 adet farklı boyutta (250, 500, 750, 1.000, 1.250, 1.500, 2.000, 2.500, 3.000, 3.500, 4.000, 4.500, 5.000, 5500 ve 6000) alt derlem elde edilmiştir.

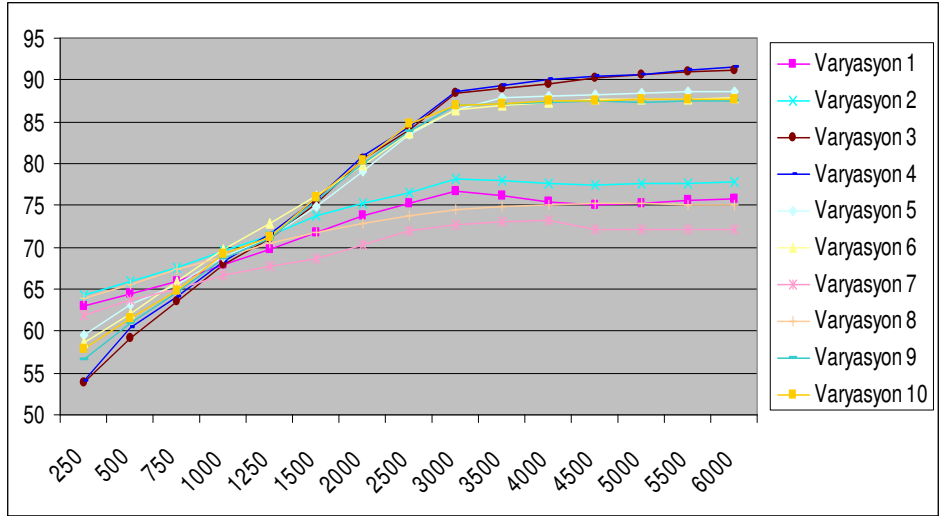
Yöntemin testlerinin yapılacağı ODTÜ-Sabancı Türkçe derleminin 1462 cümlelik alt derleminden boyutu 200 cümle olan (yaklaşık 1000 kelime) 30 adet alt derlem rastsal bir biçimde seçilmiş ve farklı boyuttaki 15 talim derleminden elde edilen istatistikler ile test edilmiştir. Elde edilen bu başarı oranları bu 30 derlemin başarı ortalamalarıdır.

Gerçekleştirilen deney sayısı 10 farklı varyasyon kullandığımız göz önünde bulundurulursa 4.500 (10x15x30)' dür. Her bir deneydeki başarı oranları, doğru olarak tespit edilen sözcük gövdesinin, tüm kelime sayısına bölünmesi ile elde edilmiştir.

$$basari_orani = \frac{Dogru_sekilde_koklenen_ve_gövdelenen_kelime_sayisi}{Toplam_kelime_sayisi}$$

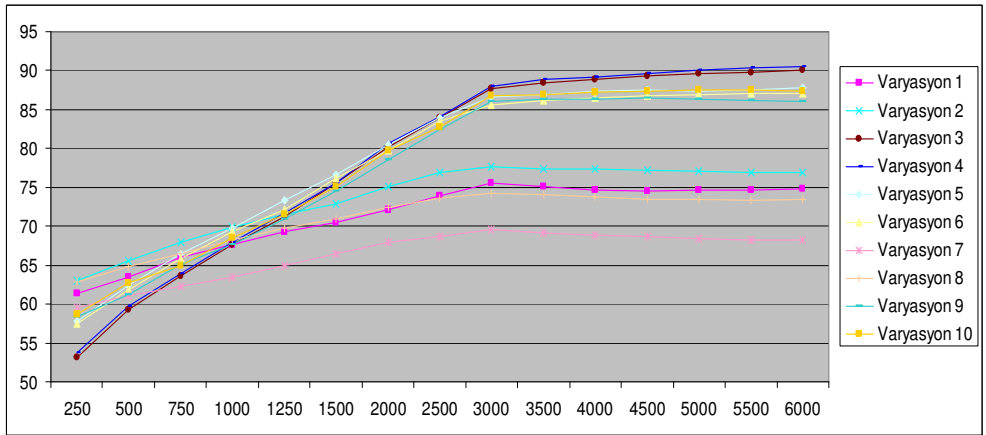
(6.21)

Deneyler gövdeleme işlemi sonucu bulunan tüm olası gövdeler ve en uygun tek gövde için tekrarlanmıştır. Yapılan deneylerde verilen bir kelimenin hem kökünün hem de gövdesinin tespiti incelenmektedir. Biçimbirimsel analiz yöntemimiz önceden de bahsedildiği üzere kelimenin sözcük türünün bilinmesi ile çalışmaktadır. Yapılan deneylerin ilk etabında kelimelerin OTDÜ-Sabancı derleminde daha önceden işaretlenmiş sözcük türleri kullanılmıştır. İstatistiklerin çekildiği verilere ve hesaplama birimi olarak kullanılan harf sayılarına göre belirlenmiş 10 farklı varyasyon için farklı boyutlardaki talim derlemlerinden elde edilen sonuçlar aşağıdaki grafikte verilmiştir (Tablolar Ek 5' te yer almaktadır). Sunulan değerler, gövdeleme işleminden çıkan “tüm olası gövdeler” için yapılan kökleme işleminden gelen çözümler arasından en uygunun tespit edilme işlemi sonucunda elde edilmiştir.



Şekil 6.6. Sözcük türü bilinen tüm olası gövdeler üzerinden gerçekleştirilen deney sonuçları

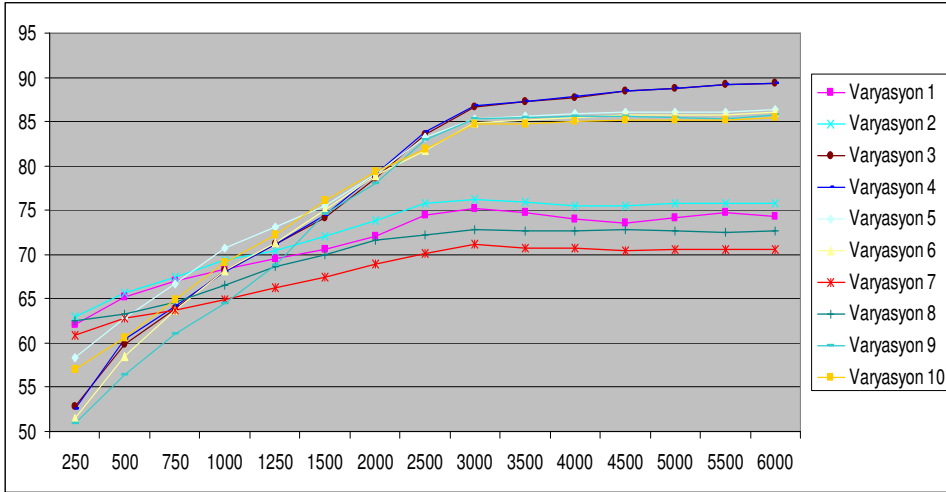
İstatistiklerin çekildiği verilere ve hesaplama birimi olarak kullanılan harf sayılarına göre belirlenmiş 10 farklı varyasyon için farklı boyutlardaki talim derlemelerinden elde edilen sonuçlar Şekil 6.7’de verilmiştir. Sunulan değerler, gövdeleme işleminden sonuç olarak çıkan “*en uygun gövde*” için yapılan kökleme işleminden gelen çözümler arasından en uygunun tespit edilme işlemi sonucunda elde edilmiştir.



Şekil 6.7. Sözcük türü bilinen en uygun gövde üzerinden gerçekleştirilen deney sonuçları

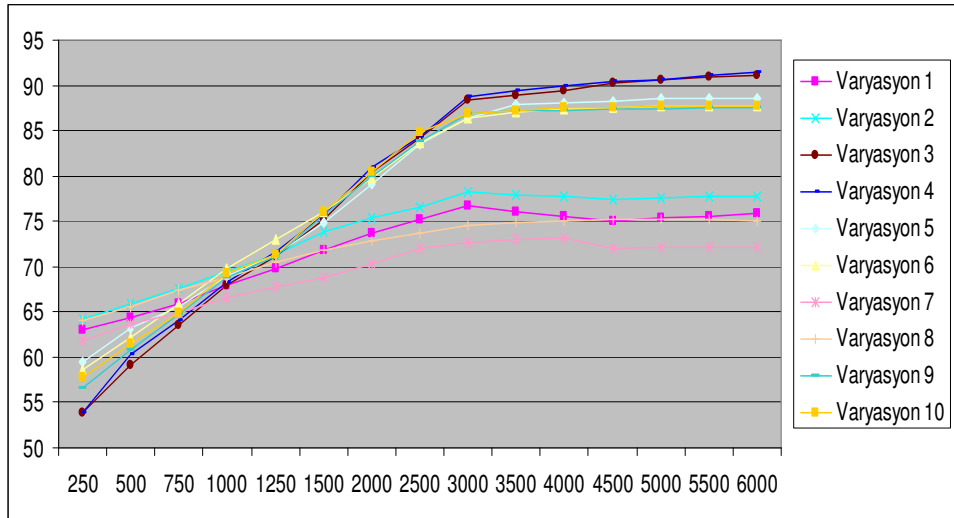
Deney derlemindeki kelimelerin sözcük türleri bilinmediği durumları test etmek amacı ile gövdeleme işlemine başlanmadan önce kelimelerin sözcük türleri belirlenmiştir. Bunun için tez kapsamında geliştirilen, başarımları oranı yaklaşık %90 olan, “kelimenin son 5 harfini” ve “2. dereceden GMM” kullanan sözcük türü tespit yöntemi kullanılmıştır.

10 farklı varyasyon için farklı boyutlardaki talim derlemlerinden elde edilen sonuçlar Şekil 6.8’de verilmiştir. Sunulan değerler, gövdeleme işleminden çıkan tüm olası gövdeler için yapılan kökleme işleminden gelen çözümler arasından en uygunun tespit edilme işlemi sonucunda elde edilmiştir.



Şekil 6.8. Sözcük türü bilinmeyen tüm olası gövdeler üzerinden gerçekleştirilen deney sonuçları

Farklı boyutlardaki talim derlemlerinden 10 farklı varyasyon için elde edilen sonuçlar Şekil 6.9’da verilmiştir. Sunulan değerler, gövdeleme işleminden sonuç olarak çıkan gövde için yapılan kökleme işleminden gelen çözümler arasından en uygunun tespit edilme işlemi sonucunda elde edilmiştir.



Şekil 6.9. Sözcük türü bilinmeyen en uygun gövde üzerinden gerçekleştirilen deney sonuçları

Sonuç olarak yapılan deneylerde bir kelimenin kök ve gövdesinin doğru belirlenmesi incelenmiştir. İnceleme sonucunda farklı girdiler için elde edilen en iyi başarımlar oranları Çizelge 6.3’de belirlenmiştir.

Çizelge 6.3. Kök ve Gövdenin Tespit Deney Sonuçları

	Sözcük Türü Elle İşaretlenmiş	Sözcük Türü Otomatik Olarak İşaretlenmiş
Tüm Olası Gövdeler Üzerinden	%91,46 (Varyasyon 4)	%89,36 (Varyasyon 3)
En Uygun Gövde Üzerinden	%90,53 (Varyasyon 4)	%88,38 (Varyasyon 3)

Gövde ve kökün doğru tespitinin belirlenmesi için kullanılan diğer bir test yöntemi olan k-parça çapraz doğrulama (k-fold cross-validation) yönteminde, derlem 10 parçaya ayrılarak her seferinde farklı 1 parça test derlemi kalan 9 parça ise talim derlemi olarak kullanılmaktadır. Bu amaç

doğrultusunda, yaklaşık 7500 cümleden oluşan ODTÜ-Sabancı derlemi 750 cümlelik 10 parçaya bölünmüştür.

Deneyde, kelimelerin ODTÜ-Sabancı derleminde daha önceden işaretlenmiş sözcük türleri kullanılmıştır. 10 farklı varyasyon için 10 farklı test derlemi üzerinden elde edilen sonuçlar aşağıdaki çizelgelerde verilmiştir (Çizelge 6.4 ve Çizelge 6.5).

Çizelge 6.4. Sözcük türü bilinen olası tüm gövdeler üzerinden gerçekleştirilen deney sonuçları

Parça No	Varyasyon 1	Varyasyon 2	Varyasyon 3	Varyasyon 4	Varyasyon 5	Varyasyon 6	Varyasyon 7	Varyasyon 8	Varyasyon 9	Varyasyon 10
1	%75,38	%77,53	%91,03	%91,45	%87,98	%87,22	%72,12	%75,41	%87,44	%87,51
2	%75,24	%77,32	%91,24	%91,32	%88,13	%87,56	%72,23	%75,02	%87,56	%87,62
3	%75,43	%77,54	%91,53	%91,85	%88,33	%87,12	%72,44	%74,83	%87,77	%87,73
4	%75,12	%77,65	%91,25	%91,42	%88,27	%87,76	%72,38	%74,88	%87,82	%87,44
5	%75,72	%77,44	%90,98	%91,29	%88,97	%87,43	%72,47	%75,05	%87,42	%87,55
6	%74,88	%76,50	%90,85	%90,63	%87,83	%86,75	%71,78	%73,35	%86,58	%86,85
7	%75,06	%78,24	%91,22	%90,98	%88,03	%87,57	%73,08	%74,31	%87,14	%87,20
8	%75,80	%77,95	%91,32	%91,39	%88,28	%87,32	%73,17	%74,79	%87,26	%87,42
9	%75,83	%77,69	%91,02	%91,23	%88,83	%87,68	%72,05	%75,23	%87,45	%87,53
10	%76,43	%77,05	%91,27	%91,58	%88,64	%87,82	%72,53	%75,62	%87,86	%87,79
Ort.	%74,49	%77,49	%91,16	%91,31	%88,33	%87,42	%72,43	%74,85	%87,43	%87,46

Çizelge 6.5. Sözcük türü bilinen en uygun gövde üzerinden gerçekleştirilen deney sonuçları

Parça No	Varyasyon 1	Varyasyon 2	Varyasyon 3	Varyasyon 4	Varyasyon 5	Varyasyon 6	Varyasyon 7	Varyasyon 8	Varyasyon 9	Varyasyon 10
1	%74,22	%76,54	%89,54	%90,55	%87,44	%86,58	%67,93	%73,13	%85,86	%87,12
2	%74,43	%76,12	%89,64	%90,25	%87,52	%86,39	%68,03	%73,31	%85,58	%87,33
3	%74,26	%76,45	%89,73	%90,47	%87,34	%86,70	%68,23	%73,37	%85,33	%87,47
4	%74,36	%76,67	%89,21	%90,62	%87,71	%86,45	%68,10	%73,48	%85,59	%87,34
5	%74,45	%76,88	%89,98	%90,03	%87,41	%86,65	%67,98	%73,61	%85,78	%87,41
6	%73,95	%75,07	%89,01	%89,45	%86,77	%85,77	%67,63	%72,86	%85,12	%86,92
7	%74,51	%76,44	%89,64	%90,50	%87,46	%86,28	%68,40	%73,61	%85,77	%87,13
8	%74,02	%77,01	%89,72	%90,38	%87,58	%86,55	%68,07	%73,06	%85,77	%87,35
9	%74,67	%76,92	%90,01	%91,21	%87,96	%87,08	%68,39	%73,28	%86,12	%87,22
10	%74,48	%76,64	%90,07	%91,38	%87,75	%87,11	%68,65	%72,66	%86,07	%87,60
Ort.	%74,34	%76,47	%89,66	%90,48	%87,49	%86,56	%68,14	%73,24	%85,70	%87,29

Diğer bir deneyde ise kelimelerin sözcük türleri otomatik olarak belirlenmiştir. 10 farklı varyasyon için elde edilen başarımlar aşağıdaki çizelgelerde verilmektedir (Çizelge 6.6 ve Çizelge 6.7).

Çizelge 6.6. Sözcük türü bilinmeyen tüm olası gövdeler üzerinden gerçekleştirilen deney sonuçları

Parça No	Varyasyon 1	Varyasyon 2	Varyasyon 3	Varyasyon 4	Varyasyon 5	Varyasyon 6	Varyasyon 7	Varyasyon 8	Varyasyon 9	Varyasyon 10
1	%74,53	%75,55	%89,65	%89,61	%86,41	%86,30	%70,23	%72,43	%85,72	%85,52
2	%74,09	%75,34	%89,38	%89,30	%86,52	%86,06	%70,14	%72,51	%85,64	%85,51
3	%74,43	%75,98	%89,42	%89,34	%86,12	%85,94	%70,60	%72,47	%85,49	%85,43
4	%74,31	%75,77	%89,75	%89,73	%86,15	%85,55	%70,12	%72,02	%85,63	%85,23
5	%73,92	%75,12	%89,81	%89,74	%86,02	%86,11	%70,91	%72,52	%85,21	%84,98
6	%73,89	%75,06	%87,73	%87,64	%85,83	%84,23	%70,05	%71,58	%85,14	%84,82
7	%74,76	%75,81	%88,65	%88,63	%86,02	%85,45	%70,12	%72,14	%85,57	%85,11
8	%74,02	%75,79	%88,95	%88,91	%86,25	%85,34	%71,07	%72,74	%85,48	%85,40
9	%73,89	%75,87	%89,69	%89,64	%86,24	%85,83	%70,32	%72,61	%85,61	%85,15
10	%74,23	%75,64	%89,78	%89,68	%86,45	%86,38	%70,69	%72,68	%85,82	%85,58
Ort.	%74,21	%75,59	%89,28	%89,22	%86,20	%85,72	%70,43	%72,37	%85,53	%85,27

Çizelge 6.7. Sözcük türü bilinmeyen en uygun gövde üzerinden gerçekleştirilen deney sonuçları

Parça No	Varyasyon 1	Varyasyon 2	Varyasyon 3	Varyasyon 4	Varyasyon 5	Varyasyon 6	Varyasyon 7	Varyasyon 8	Varyasyon 9	Varyasyon 10
1	%73,04	%74,45	%88,16	%88,34	%85,41	%85,46	%66,12	%70,06	%84,26	%84,31
2	%73,15	%74,79	%88,27	%88,41	%85,23	%85,29	%66,04	%70,61	%84,29	%84,36
3	%73,31	%74,61	%88,29	%88,05	%85,15	%85,26	%66,31	%70,32	%84,14	%84,23
4	%73,11	%74,18	%88,37	%88,56	%85,14	%85,29	%66,53	%70,44	%84,09	%84,15
5	%73,42	%74,38	%88,34	%88,32	%85,21	%85,34	%66,29	%70,12	%84,39	%84,39
6	%72,92	%74,03	%87,06	%86,89	%84,95	%84,91	%65,51	%69,67	%83,55	%83,86
7	%73,32	%74,41	%87,74	%87,94	%85,26	%85,37	%66,01	%70,46	%84,36	%84,42
8	%73,13	%74,64	%88,34	%88,08	%84,98	%84,96	%66,42	%70,62	%84,49	%84,45
9	%73,27	%74,86	%88,61	%88,14	%85,04	%85,34	%66,64	%70,96	%84,24	%84,37
10	%73,58	%74,92	%88,64	%88,39	%85,67	%85,88	%66,57	%71,08	%84,73	%84,83
Ort.	%73,23	%74,53	%88,18	%88,11	%85,20	%85,31	%66,24	%70,43	%84,25	%84,34

Her iki test yöntemi ile elde edilen sonuçlar daha önceden de bahsedildiği gibi çözümlemesi yapılacak kelimenin kök ve gövdesinin doğru tespiti ile ilgilidir. Ek dizilişlerinin de doğru işaretlenip işaretlenmediğinin, yani biçimbirimsel çözümlemenin doğruluk oranının tespiti için ise her bir deneyde en yüksek başarıyı sağlamış varyasyon kullanılarak işaretlenmiş test derlemi bir dilbilgisi uzmanı tarafından tek tek kontrol edilerek yapılmıştır. Test derlemi sonucu elde edilen veriler Çizelge 6.8’de verilmiştir.

Çizelge 6.8. Biçimbirimsel çözümleme test sonuçları

Deney No	Sözcük Türü Tespiti	Kullanılan Gövde	Kullanılan Varyasyon	Test Edilen Cümle Sayısı	Test Edilen Kelime Sayısı (Noktalama hariç)	Test Edilen Bölümün Kelime Dağılımı	Doğru Yapılan Biçimbirimsel Çözümleme Sayısı	Başarı Oranı
1	Elle	Olası tüm gövdeler üzerinden	4	1000	5142	2867	4664	%90,7
2	Elle	En uygun gövde üzerinden	4	1000	5142	2867	4627	%90,1
3	Otomatik	Olası tüm gövdeler üzerinden	3	1000	5142	2867	4566	%88,8
4	Otomatik	En uygun gövde üzerinden	3	1000	5142	2867	4524	%88,0

Özetlemek gerekirse, sözcük türü tespiti elle yapılmış cümleler üzerinden çalışan yöntemimiz yaklaşık %90’lık bir başarıya sahipken,

sözcük türlerinin otomatik olarak tespit edildiği yöntemimiz ise %88'lik bir başarı elde etmiştir. Bilgimiz dahilinde Türkçe üzerine daha önce geliştirilmiş biçimbirimsel çözümleme çalışmalarında başarımlar oranları verilmediğinden yöntemimizin sonuçları için bir karşılaştırma yapılamamıştır.

7. SONUÇ ve TARTIŞMA

Bu tezin amacı, doğal dil işleme ve enformasyon bilim çalışmalarının ihtiyaç duyduğu ve performanslarında doğrudan etki eden Türkçe dilbilgisi kurallarının üzerine bina edilmiş, tümleşik bir biçimbirimsel çözümleme ve sözcük türü tespit yönteminin geliştirilmesidir.

Bu amaç doğrultusunda Türkçe dil uzmanlarının danışmanlığında geliştirilen yöntem,

- ✓ Türkçe dilbilgisi kurallarını temel alması,
- ✓ kapalı bir sözlük kullanması,
- ✓ sözcük türü tespiti ve biçimbirimsel çözümleme süreçlerini beraber ele alması,
- ✓ tek bir sonuç üretmesi ile belirsizliği gidermesi,
- ✓ basitleştirilmiş yapısı,
- ✓ eklerde gerçekleşen dilbilgisel değişimlerin (harf düşmesi, harf yumuşaması vb.) tespitinin ek kalıpları kullanımı ile oldukça basitleştirmesi,
- ✓ kural tabanlı ve istatistiksel yöntemleri beraber kullanması ile melez bir yapıya sahip olması

nedeniyle özgün ve orjinal bir çalışmadır.

Tez kapsamında, yapılan DDİ çalışmaları şu şekilde sıralanabilir:

- a) Türkçe ek yapısının incelenmesi: Bu çalışma ile Türkçe’de bulunan yapım ve çekim ekleri incelenerek ekleri içeren bir liste derlenmiş, ayrıca eklerin dizilimleri ile ilgili kurallar çıkarılarak “Türkçede ek kuralları” başlığı ile sunulmuştur.

- b) Sözcük türü tespit yöntemi: Gizli Markov dil modeli kullanılarak yapılan sözcük türü tespit çalışmaları, kapalı bir sözlük kullanma iddiasının gerçekleştirilmesi amacı ile farklı hesaplama birimleri kullanılarak test edilmiştir. Yöntemin orjinalliği kapalı bir sözlük kullanmanın yanı sıra kullanılan yumuşatma yöntemlerinden de gelmektedir. Deneysel olarak yapılan çalışmalarda ulaşılan en iyi sonuç kelimenin son 5 harfi için, %90,2 bulunurken, gövdeler kullanılarak yapılan deneylerin sonucunda başarı %94,9 bulunmuştur. Yöntem tamamen istatistik/olasılık esasında olduğundan herhangi bir biçimbirimsel analize ihtiyaç duymamaktadır.
- c) Çekim ekleri için ek kalıpları: Türkçenin dilbilgisel özellikleri göz önünde bulundurularak elde edilen kurallar yardımı ile Türkçe’de yer alan çekim eklerinin tüm olası dizilimleri “Çekim ek kalıpları” adı altında belirlenmiştir.
- d) Gövdeleme: Kural tabanlı ve istatistiksel metodların beraber kullanıldığı yöntemde, cümle olarak ele alınan çözümleme süreci, kapalı bir sözlük kullanması, ek kalıpları sayesinde oldukça basitleştirilmiş SDM’leri ile diğer yöntemlerden farklılaşmaktadır. Deneysel olarak yapılan çalışmalarda sözcük türü bilinen kelimeler için başarı %93, sözcük türü otomatik işaretlenen kelimeler için ise %92 bulunmuştur. Şimdiye kadar yapılan ve tek bir sonuç döndüren gövdeleme çalışmalarında en iyi başarıya sahip olan Sever’in (2003) gövde-bul yöntemi ile karşılaştırılan yöntemimizin %12-13 civarında daha iyi performans sağladığı görülmüştür. Bu başarının yanı sıra yöntemimizin kapalı bir ek kalıp sözlüğü kullanması da oldukça önem taşımaktadır.

- e) Biçimbirimsel Çözümleme: Şimdiye kadar yapılan çalışmalardan farklı olarak cümle bazında çalışıp, biçimbirimsel çözümleme ve sözcük türü tespiti süreçlerini beraber ele alan yöntemimiz mevcut çalışmalardan oldukça farklılaşmaktadır. Deneysel çalışmaların sonucunda yöntemin sözcük türü bilinen kelimeler üzerinde kök ve gövdeyi aynı anda doğru tespit etme başarımı %91 olarak bulunurken, sözcük türü otomatik olarak bulunan kelimelerde başarıım %89 olarak bulunmuştur. Ayrıca biçimbirimsel çözümlemedeki başarı oranı ise benzer durumlar altında %90 ve %88 olarak tespit edilmiştir. Bilgimiz dahilinde biçimbirimsel çözümleme yapan çalışmaların hiçbirinde bir başarıım oranı sunulmamıştır.

Tez kapsamında yapılan çalışmaların sonuçlarından, Türkçe doğal dil işleme alanındaki birçok çalışma için kök veya gövde sözlüğü kullanılmadan problemlerin çözülebileceği, Türkçenin dilbilgisel özelliklerinin kullanılarak işlerin oldukça basitleştirilebileceği çıkarsamaları yapılabilir.

KAYNAKLAR

- Alpkoçak, A., Kut, A., and Özkarahan, E., 1995, Bilgi bulma sistemleri için otomatik Türkçe dizinleme yöntemi. Bilişim Bildirileri, Dokuz Eylül Üniversitesi, İzmir.
- Andrew J. Viterbi., 1967, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Transactions on Information Theory* 13(2):260–269, April 1967. (The Viterbi decoding algorithm is described in section IV.)
- Antworth, E., 1990, PC-KIMMO: A two-level processor for morphological analysis Dallas, TX: Summer Institute of Linguistics.
- Atalay, N.B., Oflazer, K., and 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, Hungary.
- Ballhysa, E., 2000, Albanian-Turkish-English Translaor, Y.L. Tez Projesi, Boğaziçi Üniversitesi.
- Banguoğlu, T., 2004, Türkçenin grameri, Türk dil Kurumu Yayınları, Ankara.
- Barton, G. E., 1986, Computational complexity in two-level morphology. In ACL Proceedings, 24th Annual Meeting.
- Birtürk, A., 1998, Türkçenin yönetim ve bağlama kuramı yaklaşımı ile bilgisayarla çözümlenmesi. Yayınlanmamış doktora tezi. Orta Doğu Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

KAYNAKLAR (Devam)

- Bozşahin, C. Oflazer, K, Turkish Natural Language Processing Initiative: An Overview, <http://www.nlp.cs.bilkent.edu.tr/turklang/> (son erişim: Ağustos, 2008)
- Bozşahin, C., Oflazer, K., 1994, Ulamsal dilbilgisi ve Türkçe, Dilbilim Araştırmaları.
- Brill, E., 1992, A simple rule-based part-of-speech tagger. Proceedings Third ACL Applied, Trento, Italy, 152-155.
- Brill, E., 1995, Transformation-based error driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, 21(4):543-566.
- Brill, E., 1995, Unsupervised learning of disambiguation rules for part of speech tagging. In Proceedings of the Third Workshop on Very Large Corpora, Cambridge, MA.
- Cebiroğlu G., Adalı E., 2006, Sözlüksüz Köke Ulaşma Yöntemi, yayınlanmamış yüksek lisans tezi, İTÜ.
- Çetinoğlu, Ö., 2001, A Prolog Based Natural Language Infrastructure for Turkish, Y.L. Tezi, Boğaziçi Üniversitesi.
- Charniak, E., 1997, Statistical Techniques for Natural Language Parsing. *AI Magazine* 18(4):33-44.
- Çiçekli, İ., Korkmaz, T., 1998 , Generation of simple Turkish sentences with systemic functional grammer, Proceedings of the third International Conferance on New Methods in Language Processing, Sydney, Australia.

KAYNAKLAR (Devam)

- Çiftçi, T., 2001, Multimedia Search engine for content based retrieval of images and texts, Y.L. Tezi, Boğaziçi Üniversitesi.
- Covington M. A., 1994, Natural Language Processing for Prolog Programmers, Prentice Hall, New Jersey.
- Church, K. and W. Gale, 1991, A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams, Computer Speech and Language, v. 5, pp. 19-54. Dale, R., Moisl, H. and Somers, H. (eds.), 2000, Handbook of Natural Language Processing. Marcel Dekker.
- Dalkılıç G., Çebi Y., 2004, Zipf Law and Madelbrot's constants for Turkish language using Turkish Corpus (TurCo), Advances in Information Systems, Springer Verlag, pp. 273-282.
- Dalkılıç, M.E. and Dalkılıç, G., 2001, Some measurable language characteristics of printed Turkish, Proc. of the XVI International Symposium on Computer and Inf. Sciences, pp. 217-224.
- Dalkılıç, M.E. and Dalkılıç, G., 2002, Some Important Statistical Properties of Printed Turkish, Journal of Statistical Research, Volume 1, Number 1, pp. 113-130.
- Darcan, O.,N, An Intelligent Database Interface for Turkish, 1991.
- Demir, C., 1993, Türkçe için bir ATN grameri. Yayınlanmamış yüksek lisans tezi. Bilkent Üniversitesi, Fen Bilimleri Enstitüsü, Ankara, 1993, 94s.

KAYNAKLAR (Devam)

- DeRose, Steven J., 1990, Stochastic Methods for Resolution of Grammatical Category Ambiguity in Inflected and Uninflected Languages, Ph.D. Dissertation. Providence, RI: Brown University Department of Cognitive and Linguistic Sciences.
- Dinçer, T., 2004, Türkçe için istatistiksel bir bilgi geri-getirim sistemi, yayımlanmamış doktora tezi, U.B.E., Ege Üniversitesi.
- Ergin, M., 2003, “Türk Dil Bilgisi”, ISBN: 975-7728-06-03, Bayrak Yayın evi, İstanbul.
- Forney, G. D., 1973, The Viterbi algorithm. Proceedings of the IEEE 61(3):268–278, March 1973.
- Fred K., Voutilainen A., Heikkilä J, and Anttila A. (eds), 1995, Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text. Natural Language Processing, No 4. Mouton de Gruyter, Berlin and New York. ISBN 3-11-014179-5.
- Garside, R. and Leech, F., 1987b, The UCREL probabilistic parsing system. In Roger Garside, Geoffrey Leech, and Geoffrey Sampson (eds.), The Computational Analysis of English: A Corpus-Based Approach, Longman, London, 66-81.
- Garside, R., 1987, The CLAWS Word-tagging System. In: R. Garside, G. Leech and G. Sampson (eds), The Computational Analysis of English: A Corpus-based Approach. London: Longman.
- Garside, R., 1987, The Claws word-tagging system, The Computational Analysis of English.
- Garside, R., 1987a, The CLAWS word-tagging system. In Garside R., Leech, F., and Sampson, G., editors, The Computational Analysis of English. Longman.

KAYNAKLAR (Devam)

- Garside, R., 1996, The robust tagging of unrestricted text: the BNC experience. In J. Thomas and M. Short (eds) Using corpora for language research: Studies in the Honour of Geoffrey Leech Longman, London, pp 167-180.
- Garside, R., and Smith, N., 1997, A hybrid grammatical tagger: CLAWS4, in Garside, R., Leech, G., and McEnery, A. (eds.) Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman, London, pp. 102-121.
- Good, I., 1953, The population frequencies of species and the estimation of population parameters, Biometrika, v. 40, pp. 237-264.
- Gordon, Raymond G., Jr. (Ed.), 2005, Ethnologue: Languages of the world (15th ed.). Dallas, TX: SIL International. ISBN 1-55671-159-X. (Online version: <http://www.ethnologue.com>).
- Greene B. and Rubin G., 1971, Automatic grammatical tagging of English, Technical Report, Department of Linguistics, Brown University.
- Gungor T, 1995, Computer Processing of Turkish: Morhological and lexial investigation, Yayınlanmamış doktora tezi, Boğaziçi Üniversitesi.
- Güngördü Z., Oflazer, K., 1995, Parsing Turkish using the Lexical-Functional Grammar Formalism, Machine Translation, Vol. 10, No 4.
- Güngördü, Z., 1993, Türkçe için bir sözcüksel-işlevsel gramer. Yayınlanmamış yüksek lisans tezi. Bilkent Üniversitesi, Fen Bilimleri Enstitüsü, Ankara, 138 sayfa.

KAYNAKLAR (Devam)

- Hakkani, D.Z., 1996, Serbest öge sıralı bir dil olan Türkçe için yüzeysel üretici tasarımı ve gerçekleştirilmesi. Yayımlanmamış yüksek lisans tezi. Bilkent Üniversitesi, Fen Bilimleri Enstitüsü, Ankara, 135 sayfa.
- Hakkani-Tür, D.Z., Oflazer, K. and Tür, G., 2002, Statistical morphological disambiguation for agglutinative languages. *Computers and the Humanities*.
- Hans van Halteren, Jakub Zavrel, Walter Daelemans. 2001. Improving Accuracy in NLP Through Combination of Machine Learning Systems. *Computational Linguistics*. 27(2): 199-229.
- Hengirmen M., 2005, Türkçe Dilbilgisi, Engin Yayınevi, İstanbul.
- Hindle, D., 1989, Acquiring disambiguation rules from test. Proceeding of the 27th Annual meeting of the association for computational Linguistics.
- Jelenik, F. and Mercer, R., 1980, Interpolated estimation of Markov source parameters from sparse data. In Proceedings of the Workshop on Pattern Recognition in Practice, Amsterdam, The Netherlands: North-Holland.
- Karaoğlu, B., Dincer T., 2008, Zipf kanunları esasında güncel yazılı Türkçenin nicel dilbilim ölçütleri, Tübitak Projesi sonuç raporu
- Karlsson, F.. 1990. Constraint Grammar as a Framework for Parsing Unrestricted Text. H. Karlgren, ed., Proceedings of the 13th International Conference of Computational Linguistics, Vol. 3. Helsinki 1990, 168-173.

KAYNAKLAR (Devam)

- Katz, S.M., 1987, Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Transactions on Acoustic, Speech and Signal Processing, ASSP-35(3), pp. 400-401.
- Klein, S. and Simmons, R. F., 1963 Computational approach to grammatical coding of English words. JACM (10):344-337.
- Klein, S. ve Simpson,R., 1963, Computational approach to grammatical coding of English words JACM (10).
- Korkmaz A., Ercilasun A., Zülfikar, H., Parlatır İ., Akalın M., Gülensoy T., Birinci N., 2003, Türk Dili ve Kompozisyon Bilgileri, ISBN: 975-8380-33-8. Yargı yayınevi, Ankara
- Korkmaz, T., 1996, Sistemik-fonksiyonel gramer yaklaşımı ile Türkçe metin üretimi. Yayımlanmamış Yüksek Lisans Tezi. Bilkent Üniversitesi, Fen Bilimleri Enstitüsü, Ankara, 132 sayfa.
- Koskenniemi, K., 1983, Two-level morphology: a general computational model for word-form recognition and production. Publication No. 11. University of Helsinki: Department of General Linguistics.
- Leech, G., Garside, R., and Bryant, M., 1994, CLAWS4: The tagging of the British National Corpus. In Proceedings of the 15th International Conference on Computational Linguistics (COLING 94) Kyoto, Japan, Sayfa 622-628.
- Lidstone, G.J., 1920, Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. Transactions of the Faculty of Actuaries, 8, Sayfa 182-192.

KAYNAKLAR (Devam)

- Markov, A.A., 1971, Extension of the limit theorems of probability theory to a sum of variables connected in a chain, volume 1: Markov Chains. John Wiley and Sons.
- Ney, H. Essen, U. and Kneser, R.,1994, On structuring probabilistic dependencies in stochastic language modeling. Computer, Speech and Language, 8, pp.1-38.
- Oflazer, K, Kuruöz, İ., 1994, Tagging morphology disambiguation of Turkish text, The forth Applied Natural Language Processing Conferance, ACL.
- Oflazer, K. and Tür, G., 1997, Morphological disambiguation by voting constraints, In Proceedings of the thirty-fifth Annual Meeting of the Association for Computational Linguistics (ACL'97/EACL'97), Madrid, Spain.
- Oflazer, K., 1996, Error-tolerant Finite State Recognition with Applications to Morphological Analysis and Spelling Correction, Computational Linguistics, Vol. 22 No 1.
- Oflazer, K., Say, B., Hakkani-Tür, D. and Tür, G., 2003, Building a Turkish Treebank, Chapter in Building and Using Parsed Corpora, Anne Abeillé (Ed.), Kluwer Academic Publishers.
- Oflazer, K., Tür, G., 1996, Combining hand-crafted rules and unsupervised learning in comstrained-based morphological disambiguation, ACL-Sıgdut Conferance on Emperical Methods in NLP.

KAYNAKLAR (Devam)

- Oflazer, K., 1994, Two-level Description of Turkish Morphology, *Literary and Linguistic Computing*, 9(2), 137-148.
- Özcan E., Demir Ş., Maden İ., 2003, Türkçe' den SQL Sorgularına Çeviri Yapan Bir Doğal Dil İşleme Uygulaması, TBD 20. Ulusal Bilisim Kurultayı.
- Özcan, N.O., 2004, Yazılı sayısal metinlerde Türkçe metinlerin tespiti ve internette uygulanması, Yayımlanmamış yüksek lisans tezi, Ege Üniversitesi, Uluslararası Bilgisayar Enstitüsü.
- Resnik P.S., 1993. Selection and Information: A Class-Based Approach to Lexical Relationships. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.
- Say, A., 1998, Bilgisayarla Türkçe Anlama, *Galatasaray Üniversitesi Mühendislik bilimleri Dergisi*, Vol.1.
- Say, B., Zeyrek, D., Oflazer, K. and Özge, U., 2002, Development of a Corpus and a Treebank for Present-day Written Turkish, in Proceedings of the Eleventh International Conference of Turkish Linguistics, August.
- Sever, H. and Bitirim, Y., 2003, The analysis and evaluation of stemming algorithms for Turkish, 10th International Symposium on String Processing and Information Retrieval (SPIRE'03), Manaus, Brazil, October 8-10. Lecture Notes in Computer Science (LNCS).
- Sezer, E., 1999, Smart bilgi erişim sistemi'nin Türkçe yerelleştirilmesi ve otomatik gömü üretimi, Yayımlanmamış yüksek lisans tezi, Hacettepe Üniversitesi.

KAYNAKLAR (Devam)

- Simons M., Ney H., Martin S . C, 1997. Distant Bigram Language Modelling using Maximum Entropy, IEEE Int. Con& on Acoustics, Speech, and Signal Processing, Mu nich, Vol. 11, pp. 787-790, April 1997.
- Sinanoğlu, O., 2002, Bye Bye Türkçe, ISBN: 975-8410-13-X , sf:228 – 249, Otopsi Yayınevi, İstanbul
- Tür, G., 2000, A statistical information extraction system, Yayımlanmamış Doktora Tezi, Bilkent Üniversitesi
- Turkish WordNet Web Page, 2006 Balkanet projesi <http://people.sabanciuniv.edu/~oflazer/balkanet/index.htm>, (son erişim: aralık 2008).
- Uzun, N., E., 1998, Dilbilgisinin Temel Kavramları (Türkçe Üzerine Tartışmalar), Ankara, s.28-29.
- Voegelin, C. F.; & Voegelin, F. M., 1977, Classification and index of the world's languages. New York: Elsevier.
- Zipf, G. K., 1949, Human Behaviour and the Principles of Least Effort. Cambridge, MA, Addison-Wesley.

EKLER

Ek 1. Türkçede bulunan yapım ekleri.

Ek 2. Türkçede Ek Kuralları

Ek 3. Sözcük Türü Tespiti Başarım Oranları

Ek 4. Gövdeleme Yöntemi Başarım Oranları

Ek 5. Kökleme ve Gövdeleme Yöntemi Başarım Oranları

Ek 1. Türkçede bulunan yapım ekleri.

İsimden İsim Yapma Ekleri

-Iuk Eki: (-lık, -lik, -luk, -lük) Bu ek yer, topluluk, alet, vasıf (sıfat), ve soyut isimler oluşturur. Bu ekin bir özelliği de bazen çekim eklerinden sonra gelebilmesidir. gün-de-lik, on-da-lık gibi. Örnekler: taşlık, zeytinlik, koruluk, gözlük

-CU Eki: (-cı, -ci, -cu, -cü, -çı, -çi, -çu, çü) Bu eklerin başlıca fonksiyonu meslek isimleri yapmaktır. Örnekler: avcı, eskici, yolcu, gözcü, aşçı, simitçi, okçu, çöpçü

-IU Eki: (-lı, -li, -lu, -lü) Ek ve sahiplik ya da bağlılık ifade ederek sıfat olarak kullanılan vasıf isimleri yapmaktır. Örnekler: kanlı, düşünceli, sulu, köpüklü

-sUz Eki: (-sız, -siz, -suz, -süz) Bu ek -IU ekinin olumsuzudur. Örnekler: ağaçsız, içkisiz, duygusuz, ölçüsüz

-ki Eki: (-ki) Aitlik eki olarak bilinen ek sıfat ve zamir yapar. Örnekler: Öteki, sonraki

-cUk Eki: (-cık, -cik, -cuk, -cük) Bu ek Türkçenin çok işlek küçültme ve sevgi ifadesi veren ekidir. Örnekler: kısacık, incecik, yavrucuk, köprücük

-cAk Eki: (-cak, -cek) Bu ek de bir sevgi ve küçültme ekidir. Örnekler: yavrucak, büyücek

-cUğAz Eki: (-cığaz, -ciğez, -cuğaz, -cüğez) Bu ek küçültme ve sevginin yanında zavallılık da ifade etmektedir. Örnekler: hanımcığazım, beyciğezim, yavrucuğazım

-cAğIz Eki: (-cağız, -ceğiz) Bu ek küçültme ve (sevginin yanında acıma da ifade etmektedir. Örnekler: kadıncağız, köyceğiz

-CA Eki: (-ca,-ce,-ça,-çe) Eskiden çekim eki olan bu ek sonradan fonksiyon değiştirerek yapım eki hâline gelmiştir. Yapım eki olarak fonksiyonu dil isimleri yapmaktır. Bazen yer isimleri türetmek içinde kullanılmaktadır. Örnekler: Almanca, İngilizce, Farsça, Türkçe, Kanlı-ca, Çamlı-ca,

-Daş Eki: (-daş, -taş) Başlıca fonksiyonları eşlik, ortaklık ve mensubiyet, bağlilik ifade etmektir. Örnekler: arkadaş, meslektaş

-ncU Eki: (-ncı, -nci, -ncu,- ncü) Asıl sayı isimlerinden sıra, derece ifade eden sayı isimleri yapmaktadır Örnekler: altıncı, ikinci, onuncu, yüzüncü

-Ar, -şAr Eki: (-ar, -er, -şar, -şer) Asıl sayı isimlerinden dağıtma, bölme, ayırma sayı isimleri yapmaktadır. Örnekler: onar, dörder, altışar, yedişer

-z Eki: (-z) Yakınlık, eşlik ifade eden topluluk isimleri yapmaktadır. Birden ona kadar olan sayılar arasında kullanılır. Örnekler: ikiz

-sU Eki: (-sı, -si, -su, -sü) Çok sık kullanılmayan ek benzerlik, gibilik ifade etmektedir. Örnekler: çocuksu

-msU Eki: (-msı, -msi, -msu, -msü) Renk ve tat isimlerinde kullanılan ek benzerlik ve gibilik ifade eden bir ektir. Örnekler: acımsı, mavimsi, morumsu, gölümsü

-mtırak Eki: (-mtırak) Ek benzerlik ve gibilik ifade eden bir ektir. Örnekler: acımtrak

-rAk Eki: (-rak, -rek) Bu ek karşılaştırma ekidir. Günümüzde çok fazla kullanımı yoktur. Örnekler: Eski Türkçe eki

-lı -lı Eki: (-lı -li -lu -lü.....-lı -li -lu -lü) Kısacası “ve” anlamı vermektedir. Örnekler: irili ufaklı

-lAyIn Eki: (-layın, -leyin) Vakit isimleri türetmekte kullanılır.

Örnekler: sabahleyin

-cIlayIn Eki: (-cılayın, -cileyin) Eskiden kullanılan bu ek günümüzde bazı zamirlerde görülür. Örnekler: Eski Türkçe eki

-An Eki: (-an, -en) Kullanımı oldukça az olan bu ek ancak bir iki kelimeye görülür. Örnekler: Eski Türkçe eki (oğlan)

-kek Eki: (-kek) Kullanımı oldukça az olan bu ek ancak bir iki kelimeye görülür. Örnekler: Eski Türkçe eki (erkek)

-kan Eki: (-kan) Kullanımı oldukça az olan bu ek ancak bir iki kelimeye görülür. Örnekler: Eski Türkçe eki

-ç Eki: (-ç) Kuvvetlendirme anlamı veren ve kullanımı oldukça az olan bu ek ancak bir iki kelimeye görülür fonksiyonu mevcuttur. Örnekler: Eski Türkçe eki (anaç)

-KA Eki: (-ka, -ge) Kullanımı oldukça az olan bu ek ancak bir iki kelimeye görülür Örnekler: Eski Türkçe eki (başka,özge)

-CUI Eki: (-cıl, -cil, -cul, -cül, -çıl, -çıl, -çıl, -çul, -çül) Benzetme ve abartma ifadesi taşır. Çok kullanılan bir ek değildir. Örnekler: evcil, balıkçıl, tavşancıl, ölümcül

-dUrUk Eki: (-dırk, -dirik, -duruk, -dürük) Alet isimleri yapmaktadır. Çok kullanılan bir ek değildir. Örnekler: Eski Türkçe eki

-mAn Eki: (-man, -men) Abartma ve benzerlik ifade eder. Çok kullanılan bir ek değildir. Örnekler: Eski Türkçe eki(kocaman, küçümen)

-Aç Eki: (-aç, -eç) Benzetme ve ilgi ifadesi taşımaktadır. Çok kullanılan bir ek değildir. Örnekler: Eski Türkçe eki (topaç)

-şIn Eki: (-şın, -şin) Yakınlık, benzerlik ifade etmektedir. Örnekler: Eski Türkçe eki (sarışın, gökşin)

-Ak Eki: (-ak, -ek) Benzerlik ifade etmektedir. Çok kullanılan bir ek değildir. Örnekler: Eski Türkçe eki (solak, benek)

-k Eki: (-k) Benzerlik ifade etmektedir. Çok kullanılan bir ek değildir. Örnekler: Eski Türkçe eki (topuk)

-z Eki: (-z) İlgi, benzerlik ifade etmektedir. Örnekler: Eski Türkçe eki (topuz)

-t Eki: (-t) Denklik ifade etmektedir. Örnekler: Eski Türkçe eki (yaşıt)

-tU Eki: (-tı, -ti, -tu, -tü) Yalnız doğayı taklit eden kelimelerde kullanılmaktadır. Örnekler: parıltı, gümbürtü

-Az Eki: (-az, -ez) İlgi ifade etmektedir. Çok kullanılan bir ek değildir. Örnekler: Eski Türkçe eki (ayaz)

Ay Eki: (-ay, -ey) İlgi ifade etmektedir. Çok kullanılan bir ek değildir. Örnekler: Eski Türkçe eki (güney)

-l Eki: (-l) Benzerlik ifade etmektedir. Çok kullanılan bir ek değildir. Örnekler: Eski Türkçe eki (kızıl)

-sUl Eki: (-sul, -sül, sıl, -sıl) Benzerlik ifade etmektedir. Çok kullanılan bir ek değildir. Örnekler: Eski Türkçe eki (yoksul)

-sal Eki: (-sal) Yer ifade etmektedir. Çok kullanılan bir ek değildir. Örnekler: Eski Türkçe eki (kumsal)

-Kul Eki: (-gıl, -gil, -gul, -göl, -kıl, -kil, -kul, -kül) İlgi ifade etmektedir. Çok kullanılan bir ek değildir. Örnekler: Eski Türkçe eki (içkil, Aligil, dayımgil)

-lA Eki: (-la, -le) Çok kullanılan bir ek değildir. Örnekler: Kışla

İsimden fiil yapma ekleri

-IA Eki: (-la, -le) İşlek bir isimden fiil yapma ekidir. Olma, yapma ifade eden fiiller yapmaktadır.. Örnekler: sula, iğnele

-Al Eki: (-al, -el) Genellikle sıfatlardan fiil yapmaktadır. Örnekler: daral, düzel

-I Eki: (-l) Sıfatlardan olma ifade eden fiiller yapmaktadır. Örnekler: kısal

-A Eki: (-a, -e) İşlek olmayan bir ektir. Yapma veya olma ifade eden fiiller yapmaktadır. Örnekler: yaşa, türe

-U Eki: (-ı, -i, -u, -ü) İşlek olmayan bir ektir. Yalnızca bir kaç kelimeye görülür. Örnekler: taşı

-Ar Eki: (-ar, -er) Daha çok renk isimlerinden fiil yapmaktadır. Örnekler: morar, yeşer

-DA Eki: (-da, -de, -ta, -te) Ses taklidi isimlerden fiil yapmaktadır. Örnekler: mırıl-da, gürül-de

-kUr Eki: (-kır, -kir, -kur, -kür) Ses taklidi kelimelerden fiil yapmakta kullanılmaktadır. Örnekler: haykır, püskür

-k Eki: (-k) İşlek bir ek değildir. Örnekler: Eski Türkçe eki (gözük)

-r Eki: (-r) Özellikle ses taklidi isimlerden fiil yapmaktadır. Örnekler: Eski Türkçe eki (çağır, üfür, bağır)

-sA Eki: (-sa, -se) İşlek olmayan bir ektir. Yalnızca bir kaç kelimeye görülür. Örnekler: Eski Türkçe eki (umursa, garipse)

-msA Eki: (-msa, -mse) İşlek olmayan bir ektir. Yalnızca bir kaç kelimeye görülür. Örnekler: azımsa, benimse

Fiilden isim yapma ekleri

-mAk Eki: (-mak, -mek) En çok kullanılan eklerdendir. Bütün fiil kök ve gövdelerine getirilebilir. Görevi eylemin ismini oluşturmaktır. Örnekler: Okumak, yürümek, yüzmek, konuşmak

-mA Eki: (-ma, -me) Tüm fiil kök ve gövdelerinde kullanılabilir. Fonksiyonu iş isimleri yapmaktır. Örnekler: Okuma, yürüme, yüzmek, konuşma

-Uş Eki: (-ış, -iş, -uş, -üş) İş ismi yapmakta kullanılan bu ek çok kullanılan eklerden biridir. Örnekler: alış, yürüyüş, çekiliş, oturuş

-m Eki: (-m) Kalıcı nesne ismi yapan bu ek çok kullanılan eklerdendir. Örnekler: uçur-u-m, geç-im

-k Eki: (-k (ka), -k (ke)) Oldukça işlek olan bu ek harekete uğramış, hareketten doğmuş veya hareketi yapmış çeşitli nesnelerin isimlerini meydana getirir. Örnekler: açık, yatık, buruşuk, elek

-Ak Eki: (-ak, -ek) İşlek olan bu ek fiilin tesirinde kalan çeşitli nesnelerin isimlerini yapar. Örnekler: döne-k, kaçak, dayanak, ürkek

-n Eki: (-n) Yapanı, olanı ve yapılanı ifadesi veren çok kullanılan bir ektir. Örnekler: tütün, ekin

-KU Eki: (-gı, -gi, -gu, -gü, -kı, -ki, -ku, -kü) Yapma ifade eden tek heceli fiillere gelen işlek bir ektir. Örnekler: sargı, bilgi, duygu, sorgu,

-gA Eki: (-ga, -ge) Örnekleri çok fazla değildir: bil-ge, dal-ga, yon-ga, böl-ge, süpür-ge, kavur-ga gibi. Örnekler:

-KUn Eki: (-gın, -gin, -gun, -gün, -kın, -kin, -kun, -kün) Tek heceli fiillere getirilen işlek bir ektir. Örnekler: az-gın, yetiş-kin, bay-gın, üz-gün

-GAn Eki: (-gan, -gen, -kan, -ken) Tek heceli fiillere getirilmeyen abartma veya aşırılık ifade eden işlek bir ektir. Örnekler: alingan, çekingen, girişken

-gUç Eki: (-gıç, -giç, -guç, -güç) Büyültme ifadesi veren bu ek çok kullanılmamaktadır. Örnekler: bilgiç, başlangıç

-GAç Eki: (-gaç, -geç, -kaç, -keç) -gUç ekine benzemektedir. Örnekler: kıs-kaç, utangaç

-AğAn Eki: (-ağan, -eğen) Aşırılık, devamlılık ifadesini kazandırmaktadır. Örnekler: olağan

-UcU Eki: (-ııcı, -ici, -ucu, -ücü) Fazlalık, devamlılık ifadesini kazandırmaktadır. Örnekler: alıcı, uçucu

-c Eki: (-ç) Aşırılık ifadesi getiren ek sadece -n ile biten fiillere getirilir. Örnekler: korkunç

-U Eki: (-ı, -i, -u, -ü) Oldukça sık kullanılan bir ektir. Örnekler: yazı, koku

-A Eki: (-a, -e) İşlek olmayan bir ektir. Örnekler: yara, oya

-tU Eki: (-tı, -ti, -tu, -tü) -n ile biten fiillere getirilir. Örnekler: akıntı, çöküntü

-t Eki: (-t) İşlek olmayan bir ektir. Örnekler: öğüt, yoğurt

-l Eki: (-l) İşlek olmayan bir ektir. Örnekler: ıslıl

-sU Eki: (-sı, -si, -su, -sü) İşlek olmayan bir ektir. Örnekler: yatsı, sinsi, tütsü

-AnAk Eki: (-anak, -enek) İşlek olmayan bir ektir. Örnekler: sağanak, gel-enek

-AmAk Eki: (-amak, -emek) İşlek olmayan bir ektir. Örnekler: kaçamak, basamak

-mUk Eki: (-mık, -mik, -muk, -mük) İşlek olmayan bir ektir. Örnekler: kıymık, kusmuk

-Aç Eki: (-aç, -eç) İşlek olmayan bir ektir. Örnekler: tıkaç

-em Eki: (-em) İşlek olmayan bir ektir. Örnekler: tutam, buram

-Al Eki: (-al, -el) İşlek olmayan bir ektir. Örnekler: Çatal

-AlAk Eki: (-alak, -elek) İşlek olmayan bir ektir. Örnekler: yatalak, asalak

-ArI Eki: (-arı, -eri) İşlek olmayan bir ektir. Örnekler: uçarı

-ArAk Eki: (-arak, -erek) Kullanılan bir ektir. Örnekler: tutarak

-AmAç Eki: (-amaç, -emeç) İşlek olmayan bir ektir. Örnekler: dönemeç

-mAç Eki: (-maç, -meç) İşlek olmayan bir ektir. Örnekler: bulamaç, tutmaç

-bAç Eki: (-baç, -beç) Çok fazla kullanımı yoktur. Eklendiği kelimenin –n olan son harfini –m ye çevirmektedir. Örnekler: saklambaç, dolambaç

-sAl Eki: (-sal, -sel) İşlek olmayan bir ektir. Örnekler: uysal

-mAn Eki: (-man, -men) İşlek olmayan bir ektir. Örnekler: seçmen, sayman

-sAk Eki: (-sak, -sek) İşlek olmayan bir ektir. Örnekler: tutsak

-pAk Eki: (-pak, -pek) İşlek olmayan bir ektir. Örnekler: kaypak

-vAn Eki: (-van, -ven) İşlek olmayan bir ektir. Örnekler: yayvan

-mUr Eki: (-mur, -mür) İşlek olmayan bir ektir. Örnekler: yağmur

-cA Eki: (-ca, -ce) İşlek olmayan bir ektir. Örnekler: eğlence, düşünce

-cAmA Eki: (-cama, -ceme) İşlek olmayan bir ektir. Örnekler: sürünceme

-mAca Eki: (-maca, -mece) İşlek olmayan bir ektir. Örnekler: bulmaca, bilmece

Sıfat Fiil (Partisip) Ekleri

-An Eki: (-an, -en) Oldukça sık kullanılan ek kalıcı isim yapmaktadır. Örnekler: seven, koşan

-Ar Eki: (-ar, -er) Geçici ve kalıcı isimleri yapmaktadır. Örnekler: gider, keser, koşar

-r Eki: (-r) Geçici ve kalıcı isim yapmaktadır. Örnekler: yürür, gelir, yatar

-mUş Eki: (-mış, -miş, -muş, -müştü) Geçici ve kalıcı isim yapmaktadır. Örnekler: koşmuş, okumuştur, geçmiş, yemiştür

-DU Eki: (-dı, -di, -du -dü, -tı, -ti, -tu, -tü) Çok kullanılmaya ek kalıcı isim yapmaktadır. Örnekler: külbastı, beğendi

-DUk Eki: (-dık, -dik, -duk, -dük, -tık, -tik, -tuk, -tük) Geçici ve kalıcı isimler oluşturmaktadır. Örnekler: söylenmedik, gittiği, bildik

-AcAk Eki: (-acak, -ecek) Daha çok geçici isim yapmaktadır. Örnekler: gelecek, sevecek, yakacak, yiyecek

-AsI Eki: (-ası, -esi) İşlek olmayan bir ektir. Örnekler: Görülesi, yıkılası

-mAz Eki: (-maz, -mez) Daha çok geçici isim yapmaktadır. Örnekler: gelmez, çıkmaz, solmaz, yılmaz

Zarf Fiil (Gerundium) Ekleri

-A Eki: (**-a, -e**) Genellikle ikilemelerde kullanılan bu ek tek kullanıldığı zaman bileşik fiil yapmaktadır. Örnekler: koşa koşa, gele gele, çekebil, düşeyaz

-U Eki: (**-ı, -i, -u, -ü**) Sadece birleşik fiil yapmakta kullanılmaktadır. Örnekler: alıver, çekiver

-Up Eki: (**-ıp, -ip, -up, -üp**) Oldukça kullanılan bir ektir. Örnekler: yürüyüp, uyuyup

-ArAk Eki: (**-arak, -erek**) Oldukça kullanılan bir ektir. Örnekler: koş-arak, başla-y-arak, gül-erek

-UncA Eki: (**-ınca, -ince, -unca, -ünce**) Oldukça kullanılan bir ektir. Örnekler: yapınca, gidince, okuyunca

-AlI Eki: (**-alı, -eli**) Devamlılık ifade eden zarflar oluşturur. Örnekler: geleli, görelî, oturalı

-mAdAn Eki: (**-madan, -meden**) Oldukça fazla kullanılan olumsuz bir gerundium ekidir. Örnekler: düşmeden, okumadan

-IcAk Eki: (**-ıcak, -icek**) Çok işlek olmayan farklı ağızlarda kullanılan bir ektir. Örnekler: alıcak, gelicek

-UbAn Eki: (**-uban, -üben**) Eskiden kullanılan günümüzde rastlanmayan bir ektir. Örnekler: duruban, gelüben

-UşUn Eki: (**-ışın, -işin, -uşun, -üşün**) Çok işlek olamayan bir ektir. Örnekler: gelişin, başlayışın

-ken Eki: (**-ken**) Fiillerin sonuna gelerek birleşik gerundium yapmaktadır. Örnekler: gelirken, yaparken

-DUKÇA Eki: (-dıkça, -dikçe, -dukça, -dükçe, -tıkça, -tikçe, -tukça, -tükçe) -dIk sıfat fiilinin eşitlik şeklidir. Örnekler: susadıkça, vurdukça

-AndA Eki: (-anda, -ende) Azeri dilinde sık görülmesine rağmen Türkçe için işlekliliğini kaybetmiştir. Örnekler: gelende, vuranda

-dİğInAn Eki: (-dığınan, -diğinen) -UncA ekinin farklı ağızlarda kullanılan karşılığıdır. İşlek değildir. Örnekler: geldiğinen, aldığınan

Fiilden Fiil Yapma Ekleri

-mA Eki: (-ma, -me) Olumlu fiillerden olumsuz fiiller yapmaktadır. Örnekler: koşma, yürüme, okuma

-n Eki: (-n) Edilgen fiiller oluşturmaktadır. Örnekler: Aran, sevin, okun

-I Eki: (-I) Edilgen fiiller oluşturmaktadır. Örnekler: görül, içil

-s Eki: (-ş) Ortaklaşma ve oluş ifade eden fiiller türetmektedir. Örnekler: vuruş, atış, dayanış, gülüş

-r Eki: (-r) Yaptırma ve oldurma ifade eden fiiller türetmektedir. Sadece ç,s,t,ğ,p,y ile biten tek heceli fiillere getirilmektedir. Örnekler: aşır, yatır

-t Eki: (-t) Çok fazla kullanılan bir ektir. Örnekler: diret, incit

-Dur Eki: (-dır, -dir, -dur, -dür, -tır, -tir, -tur, -tür) Çok fazla kullanılan bir ektir. Örnekler: yedir, yağdır

-Ar Eki: (-ar, -er) İşlek olmayan bir oldurma ekidir. Örnekler: kopar, gider

-dAr Eki: (-dar, -der) İşlek olmayan bir oldurma ekidir. Örnekler: dönder

-z Eki: (-z) İşlek olmayan bir ekidir. Örnekler: emzir

-A Eki: (-a, -e) İşlek olmayan bir ekidir. Örnekler: tıka

-U Eki: (-ı, -i, -u, -ü) İşlek olmayan bir ekidir. Örnekler: kazı, sürü

-k Eki: (-k (a), k (e)) İşlek olmayan bir ekidir. Örnekler: görük

-p Eki: (-p) İşlek olmayan bir ekidir. Örnekler: serp, kırp

-y Eki: (-y) İşlek olmayan bir ekidir. Örnekler: koy, doy

-sA Eki: (-sa, -se) İşlek olmayan bir ekidir. Örnekler: görset

Ek 2. Türkçede Ek kuralları

- 1- Yapım ekleri kökten hemen sonra gelir. Çekim eki yapım ekinden önce gelmez. –lUk –cUk ve –ki ekleri istisnadır.
 - i. On-da-lık, gün-de-lik gibi. Fakat burada –dA eki çekim eki işlevi kaybolmuş ek sanki –dAlUk şeklini almıştır.
 - ii. –ki ekide benzer şekilde ismin -DA , Un, nUn eki almış çekimli şekilleri ile (yani lokatif ve genitif halleri ile) birleşmektedir. Bu durumun en önemli nedeni –ki ekinin edat geçmişinin olmasıdır. Aynı zamanda –ki eki çekim eki (aitlik eki) olarak da sınıflandırılmaktadır.
 - iii. –cUk eki de –CA çekim ekinden sonra kullanılabilir. Örneğin, usulcacık, yavaşcacık vb.
 - iv. Babamgil örneğinde olduğu gibi –gil yapım eki –m şahıs eklerinden sonra gelebilmektedir.
- 2- Yapım eki sıralanışı aşağıdaki gibi olmaktadır:
 - a. İsimden fiil yapma => Fiilden fiil yapma
 - b. İsimden fiil yapma => Fiilden isim yapma
 - c. İsimden isim yapma => isimden fiil yapma
 - d. İsimden isim yapma => isimden isim yapma
 - e. Fiilden fiil yapma => Fiilden fiil yapma
 - f. Fiilden fiil yapma => fiilden isim yapma
 - g. Fiilden isim yapma => isimden fiil yapma
 - h. Fiilden isim yapma => isimden isim yapma
- 3- Aynı cins iki kök veya gövdeden birinin aldığı yapım ekini bir diğeri almayabilir.

- 4- İki tane konsonant (ünsüz) yardımcı ses vardır: y, n. n sesi, yalnızca iyelik ekleri, -ki aitlik eki ve isim çekim ekleri arasında kullanılır.
- 5- Türkçede 9 çeşit kelime vardır:
 - a. İsim kökü
 - b. İsimden yapılmış isim gövdesi
 - c. Fiilden yapılmış isim gövdesi
 - d. İsim kökünün çekimli şekli
 - e. İsimden yapılmış isim gövdesinin çekimli şekli
 - f. Fiilden yapılmış isim gövdesinin çekimli şekli
 - g. Fiil kökünün çekimli şekli
 - h. İsimden yapılmış fiil gövdesinin çekimli şekli
 - i. Fiilden yapılmış fiil gövdesinin çekimli şekli
- 6- -yor, -ki, -daş, -ken, -leyin gibi ekler büyük ünlü uyumuna uymazlar.
- 7- Ekte o veya ö harfi bulunmaz. Sadece -yor eki istisna oluşturmaktadır.
- 8- İsim ve fiillere gelen çekim ekleri birbirinden farklıdır.
- 9- Çekim ekleri, kelimelerin kelime grupları ve cümle içerisinde görev almasını sağlamaktadır.
- 10- Aynı cins çekim eki üst üste gelmez.
- 11- Aynı cins iyelik çekim eki üst üste gelmez. Fakat zaman zaman Kim-i-si, Bir-i-si vb. istisna durumlar görülmektedir.
- 12- Soru ve çokluk ekleri isim ve fiil kök ve gövdeleri için ortakdır.

- 13- Çokluk eki yalnız kök veya gövdeden sonra gelir. İyelik hal ve soru ekinden sonra gelmez. Nadir olarak iyelik, birinci ve ikinci şahıs ekinden sonra gelebilir. (örn. bizimkiler)
- 14- İyelik ekleri çokluk ekinden sonra gelebilir fakat hal ve soru ekinden sonra gelemeyiz.
- 15- İyelik eklerinin kurduğu ilişki geriye doğrudur. Yani kelimeyi kendisinden önce gelen kelimeye bağlar.
- 16- İyelik eki sadece isim ile isim arasında ilişki kurar.
- 17- Hal ekleri isimleri isimlere bağlamalarına rağmen daha çok isimler ile fiiller arasında iletişim kurarlar.
- 18- Hal eklerinin kurduğu ilişki ileri doğrudur. Yani kelimeyi kendisinden sonra gelen kelimeye bağlarlar.
- 19- Hal ekleri çokluk ve iyelik eklerinden sonra soru ekinden ve -ki ekinden önce gelebilirler.
- 20- Hal ekleri 4 ana gruba ayrılır;
 - a. Genitif ekleri (ismi isme bağlar),
 - b. Akuzatif ekleri (İsmi doğrudan fiile bağlar),
 - c. Datif, lokatif, ablatif ekleri (ismi fiile bağlar, fiilin yerini gösterir),
 - d. Instrumental, yön ve eşitlik ekleri (ismi fiile bağlar, fiilin şartlarını gösterir, zarf yaparlar).
- 21- Genitif ekinden sonra soru ve ki eki gelir.
- 22- Akuzatif ekinden sonra sadece soru eki gelebilir.
- 23- Datif ekinden sonra ancak soru eki ve –dır, -dir bildirme eki gelebilir.

- 24- Ablatif eki isimlerin teklik çokluk ve iyelik şekillerini alır. Ablatif ekinden sonra sadece soru eki ve bildirme ekleri gelir.
- 25- Lokatif ekinden sonra soru, bildirme ve –ki aitlik eki gelmektedir.
- 26- -le –la instrumental eki, ekleşmiş ile edatıdır. Teklik, çokluk, iyelik ve genitif ekinden sonra gelebilir.
- 27- İsim çekim eki olarak soru ekinden sonra bildirme eki dışında başka çekim eki gelmez.
- 28- Soru eki ismi fiile bağlar.
- 29- Üçüncü şahıs iyelik eki sessiz harf ile biten kelimelerde ı,i,u,ü şeklinde olur. Eğer kelime sesli ile bitiyorsa ek si,sı,su,sü şeklini almaktadır. Sadece ne ve su isimleri istisnai durum oluşturur..
- 30- Zarflar esas olarak çekimsiz unsurlardır. Sadece hal ve yer çekim eki (eşitlik, instrumental ve yön ekleri) alırlar.
- 31- Zamirler iyelik eki almazlar.
- 32- Zamirler isim işletme eklerinden hal, çokluk ve soru eki alırlar.
- 33- Zamirlerden ben ve sen zamiri datif eki aldığıında kök değişimine uğrar. Bana, sana şeklini alır.
- 34- Zamirler instrumental eki almazlar. Sadece –le –la edattan dönüşme eki alırlar.
- 35- Şahıs zamirlerinden sonra yön eki getirilmez.
- 36- İşaret zamirlerinden sonra yön eki getirilmez.
- 37- Fiil kök veya gövdeleri tek başlarına kullanılamazlar.
- 38- Şahıs ekleri fiil kök ve gövdelerine şekil ve zaman ekleri ile bağlanırlar. Eğer şahıs eki varsa bu eklerde muhakkak bulunur.

- 39- Şekil ve zaman ekleri doğrudan kök veya gövdeye bağlanırlar. Önce başka çekim eki gelmez.
- 40- Şekil ve zaman eklerinden sonra şahıs ve soru ekleri gelir.
- 41- Şekil ve zaman eklerinden iki tanesi üst üste gelmez. Sadece iki kelimenin birleşip tek kelime haline gelmesi sonucundan ortaya çıkmış olan birleşik kiplerde iki hareket kipi eki bir kelimede üst üste kullanılabilir.
- 42- Şahıs ekinden sonra sadece soru eki gelebilir.
- 43- Fiil çekim eki olarak soru eki şahıs ekinden önce gelebilir.
- 44- Şahıs eklerin kurduğu ilişki geriye doğrudur.
- 45- 1. tipteki şahıs ekleri şimdiki zaman, geniş zaman, gelecek zaman, öğrenilen geçmiş zaman, istek, gereklilik çekimlerinde kullanılan eklerdir.
- 46- 2. tipteki şahıs ekleri görülen geçmiş zaman ve şart çekimlerinde kullanılan eklerdir.
- 47- Fiil çekiminde 5 adet zaman (bildirme kipleri) vardır: Geniş zaman, şimdiki zaman, görülen geçmiş zaman, öğrenilen geçmiş zaman, gelecek zaman.
- 48- Fiiller için 4 adet tasarlama kipi bulunmaktadır: Şart, isteklilik, gereklilik ve emir.
- 49- İsim- fiil ekleri aslında i- fiilinin çekimleri olup, tüm isim ve fiil şekillerinin sonuna gelebilir. İsimlerin sonuna geldiğinde isimleri fiilleştirir. İsim fiilin 4 adet kipi vardır: şimdiki zaman, görülen geçmiş zaman, öğrenilen geçmiş zaman, şart
- 50- Partisip ekleri görevleri ve kullanışları açısından yapım ekine oldukça yakın çekim ekleridir.

- 51- Dört tane vokal (ünlü) yardımcı ses vardır: ı, i ,u, ü.
- 52- -mak, -mek ekinden sonra iyelik eki getirilmemektedir.
- 53- -KAn eki tek heceli kök veya gövdelere gelmez.
- 54- -mA fiilden fiil yapma ekinden sonra başka bir fiilden fiil yapma eki getirilmez. Bu ekten sonra ancak fiilden isim veya çekim eki gelebilir. Daima gövdenin son ekidir.
- 55- -mA ekinden sonra y harfi gelirse A değişime uğrar ve (ı,i,u,ü) şekillerinden birine dönüşür.
- 56- Y sesi kendinden önce gelen a ve e harfini ı,i,u veya ü ye dönüştürebilir.
- 57- -n edilgenlik eki fiilden fiil yapma eklerinden sonra gelmez. Bu ek fiil kökleri ve isimden yapılmış fiilere gelir.
- 58- -l (pasiflik eki) ekinden sonra fiilden fiil yapma ekleri arasında sadece –mA- gelebilir. Ayrıca –l eki –mA- ekinden sonra gelmez.
- 59- -l eki sesli ile biten kök veya gövdelere gelmezler. Sadece sessiz harf ile biten kök ve gövdelere gelebilirler.
- 60- -r eki sesli ile biten fiil köklerine gelmez.
- 61- -r eki tek heceli fiil köklerine gelmektedir. Kendisinden önce –z eki hariç ek getirilmemektedir.
- 62- -r ekinden sonra fiilden fiil yapma eki olarak sadece –mA olumsuzluk eki –l- ve –t- eki gelebilir.
- 63- -t ekinden önce fiilden fiil yapma eki olarak faktitif ekleri gelebilir.
- 64- -DUr fiilden fiil yapma eki tek heceli fiillere ve sessiz ile biten fiil kök ve gövdelerine getirilir.

- 65- -DUr ekinden sonra fiilden fiil yapma eki olarak olumsuzluk eki ve -l- ve -t- ekleri getirilmektedir.
- 66- Ard arda en fazla dört adet faktitif eki kullanılır.
- 67- Gerundiumlar çekim eki almazlar.

Ek 3. Sözcük Türü Tespiti Başarım Oranları

Çizelge 8.1. Hesaplama birimi kelime alındığında elde edilen başarım oranları tablosu.

Talis Derlem Boyutu	Markov Süreci Başarım Oranları (%)	
	1. dereceden GMM	1. dereceden geçmişe bir adım GMM
250	61,83	65,13
500	67,39	67,47
750	67,63	67,98
1000	70,41	71,09
1250	71,13	72,39
1500	72,10	73,17
1750	74,51	74,72
2000	74,39	75,24
2500	75,48	76,53
3000	75,96	78,35
3500	77,53	80,42
4000	77,89	81,20
4500	78,38	80,94
5000	78,74	81,20
7500	80,91	82,23
10000	81,52	81,97
12500	82,48	82,75
15000	83,21	85,27
17500	83,93	85,27
20000	83,81	85,27
22500	83,81	85,53
25000	84,42	85,79
27500	84,29	86,79
30000	84,78	88,08
32500	85,02	88,34
35000	85,26	89,08
37500	85,14	90,12
40000	85,26	91,12
42500	85,26	92,12
45000	85,28	92,38

Çizelge 8.2. Hesaplama birimi son hece alındığında elde edilen başarımlar oranları tablosu.

Talim Derlem Boyutu	Markov Süreci Başarı Oranları (%)	
	1. dereceden GMM	1. dereceden geçmişe bir adım GMM
250	60,87	62,95
500	63,65	63,73
750	67,87	67,80
1000	70,65	70,17
1250	71,26	70,13
1500	71,74	72,62
1750	71,98	73,88
2000	72,34	74,62
2500	73,07	75,17
3000	72,58	76,95
3500	73,19	78,21
4000	72,58	78,21
4500	72,83	78,43
5000	72,83	78,91
7500	72,83	79,21
10000	73,18	79,69
12500	73,18	79,17
15000	73,51	78,39
17500	73,22	78,13
20000	73,51	78,91
22500	73,51	79,43
25000	72,34	80,19
27500	72,34	80,39
30000	74,10	80,61
32500	74,52	80,91
35000	74,52	80,65
37500	74,79	80,94
40000	75,25	81,04
42500	75,98	81,19
45000	75,98	81,39

Çizelge 8.3. Hesaplama birimi kelimenin sonundan alınan harf birlikleri olduğunda elde edilen başarımlar oranları tablosu.

Talim Derlem Boyutu	Markov Süreci Başarı Oranları (%)													
	1. dereceden GMM							1. dereceden geçmişe bir adım GMM						
	Harf Birlikleri							Harf Birlikleri						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
250	46,86	54,95	65,82	66,67	65,22	63,29	62,20	51,70	60,25	67,51	67,91	66,28	64,18	62,90
500	46,26	59,18	70,05	73,31	72,10	69,81	68,36	50,15	62,33	70,36	70,47	71,24	69,10	65,75
750	46,86	60,14	74,28	76,33	74,76	71,74	68,96	49,37	63,88	75,28	74,87	74,09	70,65	68,08
1000	47,34	61,23	77,17	77,78	77,54	74,64	72,46	51,96	64,92	78,65	77,76	78,76	75,58	71,97
1250	47,71	62,08	77,42	78,26	78,14	75,48	73,31	50,15	64,92	78,65	78,27	79,27	76,61	73,26
1500	47,83	61,23	77,66	79,23	78,86	76,45	74,64	50,15	64,92	78,91	80,31	81,83	77,39	74,82
1750	47,83	62,20	77,90	80,43	80,31	78,26	77,05	50,41	65,44	79,68	80,35	82,12	78,94	76,89
2000	47,71	62,20	78,50	80,56	80,56	78,26	77,17	50,15	65,18	78,91	83,87	83,38	79,72	77,92
2500	48,67	61,96	79,47	82,37	82,97	80,68	79,23	50,93	65,18	80,46	85,64	85,42	81,53	78,96
3000	47,95	62,32	78,99	82,61	83,45	81,04	79,47	51,45	65,44	80,46	86,16	86,20	82,83	80,77
3500	47,71	61,47	79,59	82,97	84,54	82,61	81,16	51,19	65,95	80,46	87,20	88,79	85,16	84,66
4000	47,71	62,56	79,71	83,09	85,75	83,45	81,76	51,19	66,47	80,46	87,20	88,79	85,42	84,66
4500	47,95	62,08	80,07	83,45	85,99	83,82	82,37	51,19	64,66	79,42	86,94	88,79	85,42	84,66
5000	47,95	61,96	79,95	83,57	85,87	84,18	82,61	50,93	65,18	79,68	86,16	89,05	85,94	84,92
7500	47,95	62,32	79,95	85,14	87,32	85,87	85,02	50,15	65,69	79,94	85,12	89,31	86,97	86,22
10000	46,98	61,96	79,59	85,27	87,56	86,47	85,27	50,15	66,21	80,20	85,90	89,05	86,72	85,70
12500	47,10	62,08	79,11	85,02	87,68	86,96	85,99	49,93	65,44	79,68	85,90	89,05	87,23	85,96
15000	47,10	61,96	79,47	85,27	87,80	87,21	86,23	49,33	65,44	78,91	86,42	89,05	87,49	86,47
17500	47,10	61,71	79,83	85,51	88,04	87,56	87,20	49,23	66,47	78,91	85,90	89,05	87,49	86,73
20000	46,98	61,96	79,59	85,75	88,04	87,68	87,20	49,12	65,95	79,17	86,68	89,05	87,49	86,73
22500	46,98	61,71	80,31	85,63	88,04	87,80	87,20	49,32	65,95	79,42	87,72	89,31	87,75	86,99
25000	46,98	61,52	80,43	85,14	88,41	88,11	87,56	48,97	65,69	79,17	87,72	89,82	88,53	87,77
27500	46,98	61,59	80,07	84,78	88,04	87,41	87,32	49,15	65,69	79,17	87,72	89,82	88,53	87,77
30000	46,98	61,35	79,83	85,27	88,16	87,89	87,44	49,25	65,69	79,68	87,97	90,04	89,31	88,55
32500	46,98	61,84	79,59	85,02	88,29	88,01	87,70	49,22	65,95	79,91	87,97	90,08	89,31	88,55
35000	46,98	61,71	79,71	84,78	88,95	88,13	87,80	48,92	65,95	79,91	87,72	90,08	89,31	88,55
37500	46,98	61,47	79,95	85,14	88,41	88,29	87,68	49,31	66,47	80,32	87,46	90,08	89,31	88,55
40000	47,34	61,59	79,71	85,27	88,77	88,33	88,16	49,14	65,95	80,91	87,72	90,08	89,31	88,55
42500	46,98	61,84	79,83	85,02	88,53	88,39	88,29	49,06	66,21	81,17	87,87	90,14	89,82	88,81
45000	46,98	61,96	79,59	84,90	88,90	88,77	88,16	49,32	66,21	81,91	87,97	90,23	89,82	89,06

Çizelge 8.4. Hesaplama birimi kelimenin başından alınan harf birlikleri olduğunda elde edilen başarımlar oranları tablosu.

Talim Derlem Boyutu	Markov Süreci Başarı Oranları (%)													
	1. dereceden GMM							1. dereceden geçmişe bir adım GMM						
	Harf Birlikleri							Harf Birlikleri						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
250	48,91	52,54	57,61	60,02	62,99	60,99	60,84	49,74	53,81	59,22	59,59	64,95	64,18	64,18
500	47,10	54,59	60,87	63,04	67,06	66,91	67,00	47,41	53,81	65,14	63,47	68,58	67,77	66,77
750	47,22	56,76	65,22	68,24	71,44	70,65	68,69	47,15	55,11	69,80	67,88	71,65	69,84	68,58
1000	47,34	56,40	66,18	70,89	74,14	73,67	71,71	47,67	56,92	71,36	68,91	74,24	72,95	72,47
1250	47,34	57,61	68,24	71,98	74,98	74,64	72,67	48,19	57,44	70,58	68,65	76,58	73,98	74,02
1500	47,22	58,33	69,32	72,34	75,46	75,12	73,52	48,96	56,40	71,36	68,65	77,58	74,50	75,06
1750	47,10	57,97	70,53	74,15	77,64	76,57	75,57	49,22	57,70	71,62	67,88	77,06	76,06	76,61
2000	47,10	59,18	70,65	74,76	78,24	76,57	75,57	48,19	57,44	70,84	68,39	76,80	76,32	77,13
2500	47,34	59,18	70,65	75,85	80,54	78,38	76,90	49,74	57,96	70,58	70,73	78,61	78,13	79,20
3000	47,34	59,06	70,17	75,97	80,78	78,99	77,26	49,74	58,99	70,84	73,50	79,65	79,68	81,27
3500	47,22	58,57	71,01	77,17	80,78	80,31	78,71	50,00	58,74	72,65	72,98	80,42	80,46	82,31
4000	47,22	58,70	71,14	77,54	81,50	80,68	78,95	50,00	58,48	72,65	73,24	79,91	81,24	83,09
4500	47,10	59,18	70,77	78,02	81,86	81,04	79,56	49,74	58,74	71,36	74,02	80,91	80,98	82,57
5000	47,10	58,94	70,65	77,54	81,26	81,16	79,80	49,74	59,51	71,88	75,02	81,17	81,24	82,31
7500	47,10	59,30	71,98	77,66	83,80	82,13	81,37	48,70	59,51	71,88	75,54	81,42	82,27	83,61
10000	47,34	59,18	71,14	77,90	84,76	83,09	82,09	48,70	60,29	72,39	75,80	80,91	82,02	83,61
12500	47,46	59,06	71,38	77,54	84,52	83,33	82,57	49,23	60,55	71,62	75,28	81,68	82,79	83,35
15000	47,34	59,30	71,01	77,17	85,25	83,57	82,82	49,67	59,77	71,88	74,24	82,68	81,76	82,83
17500	47,34	59,30	71,38	77,66	85,37	84,42	83,66	49,51	60,03	71,36	77,28	82,94	82,02	82,83
20000	47,22	59,30	71,98	77,78	84,76	84,06	83,18	49,73	60,03	71,36	77,02	82,42	82,02	83,09
22500	47,22	59,18	70,89	77,42	84,76	84,06	83,30	49,84	59,77	70,84	76,76	82,68	82,02	83,35
25000	47,34	59,18	70,05	77,42	85,00	84,78	84,02	49,51	59,51	71,10	77,28	83,72	83,57	84,90
27500	47,22	59,18	70,17	77,90	84,64	84,42	83,90	48,75	59,77	71,62	78,80	84,98	84,35	85,16
30000	47,22	59,30	70,89	77,90	84,76	84,66	83,78	48,85	59,77	71,88	78,80	85,21	85,12	85,94
32500	47,22	59,42	70,17	77,78	85,13	85,02	84,14	49,43	59,51	71,62	79,28	85,41	85,12	85,68
35000	47,10	59,54	70,53	77,66	85,13	85,02	84,16	49,69	59,77	72,13	80,28	85,90	84,87	85,16
37500	47,22	59,42	70,53	77,54	84,88	84,78	84,02	49,27	59,77	72,39	80,72	86,40	85,12	85,16
40000	47,10	59,54	70,53	77,78	84,88	84,66	84,02	48,82	59,77	71,88	80,72	86,14	85,12	85,42
42500	47,22	59,54	70,53	78,14	85,25	84,54	84,14	49,12	59,77	71,36	81,28	86,65	85,64	85,68
45000	47,22	59,59	70,65	78,02	85,37	84,66	84,02	49,61	59,77	71,62	81,28	86,88	85,38	85,68

Çizelge 8.5. Hesaplama birimi kelimenin başından ve sonundan alınan harf birlikleri olduğunda elde edilen başarımlar oranları tablosu.

Talis Derlem Boyutu	Markov Süreci Başarı Oranları (%)													
	1. dereceden GMM							1. dereceden geçmişe bir adım GMM						
	Harf Birlikleri							Harf Birlikleri						
	F1L1	F1L2	F2L1	F2L2	F3L2	F2L3	F3L3	F1L1	F1L2	F2L1	F2L2	F3L2	F2L3	F3L3
250	50,55	59,73	54,45	60,83	61,71	62,33	62,12	50,26	59,73	53,82	60,00	61,34	61,78	61,75
500	52,10	62,42	57,50	66,47	67,41	68,35	67,98	50,65	60,24	57,40	64,51	63,98	63,60	63,86
750	54,53	66,33	59,95	66,87	67,52	68,35	67,27	51,01	61,42	58,85	66,50	65,53	66,41	66,26
1000	56,70	69,29	62,84	71,22	70,85	71,78	70,44	52,03	63,04	59,86	68,41	67,27	68,65	67,52
1250	56,81	70,06	63,47	72,03	71,95	72,73	71,43	53,71	65,67	60,59	69,34	69,20	71,16	70,39
1500	56,88	70,88	63,97	72,69	72,88	73,50	72,05	55,75	69,07	63,57	70,84	72,96	73,27	72,27
1750	57,37	71,51	65,66	74,22	74,41	75,11	73,15	57,42	71,28	65,10	74,41	73,99	74,55	74,88
2000	57,32	71,75	65,53	74,53	74,84	75,65	73,69	57,20	71,60	65,20	74,95	75,27	76,40	75,46
2500	57,86	73,56	67,35	76,27	76,67	77,51	75,76	57,56	71,89	65,50	75,30	75,67	76,53	77,56
3000	57,56	73,37	66,87	76,70	77,50	78,23	76,43	57,75	72,43	65,67	75,68	76,51	77,04	79,51
3500	57,64	73,78	67,39	77,34	78,88	79,33	78,28	58,47	73,32	66,90	77,05	77,71	78,28	79,86
4000	58,12	74,02	67,54	77,82	79,43	79,92	78,92	58,14	73,82	67,27	77,66	78,12	80,19	80,05
4500	58,04	74,29	67,87	78,26	80,08	80,57	79,54	58,23	74,63	67,70	78,15	80,46	81,00	80,24
5000	58,01	74,18	67,54	78,17	80,03	80,84	79,94	58,17	74,74	68,54	78,77	81,12	81,47	81,16
7500	58,14	74,44	68,64	79,25	81,64	82,15	81,42	58,02	74,58	68,78	79,54	81,91	82,50	81,94
10000	58,20	74,04	68,65	79,86	82,26	83,19	82,41	57,53	74,61	68,94	79,70	82,28	82,84	82,37
12500	58,17	74,57	68,94	80,24	82,79	83,90	83,11	57,86	74,66	68,88	80,06	82,69	82,95	82,66
15000	58,37	74,44	68,97	80,51	83,20	83,95	83,55	57,71	74,71	69,08	80,19	83,41	82,93	82,88
17500	58,28	74,79	69,07	80,69	83,54	84,62	84,03	57,78	74,77	69,13	80,16	83,39	84,08	83,20
20000	58,33	74,84	69,29	80,84	83,98	84,96	84,45	58,06	74,87	69,02	80,35	83,65	84,48	83,59
22500	58,25	74,76	69,20	80,96	84,37	85,20	84,83	58,17	74,88	68,96	80,65	83,79	84,99	84,14
25000	58,45	74,84	69,34	81,02	84,37	85,42	85,07	58,12	74,92	68,92	80,84	84,59	85,13	84,30
27500	58,36	74,80	69,24	81,08	84,56	85,64	85,23	58,23	75,27	68,88	80,81	85,23	85,35	84,56
30000	58,34	74,87	69,48	81,21	84,86	85,91	85,64	58,02	75,59	68,72	80,65	85,85	85,24	85,16
32500	58,45	75,17	69,45	81,34	84,96	86,18	85,90	57,93	75,51	69,15	80,53	86,77	85,96	85,75
35000	58,42	75,14	69,35	81,26	85,04	86,20	86,14	58,02	75,31	69,21	81,00	87,47	86,25	86,74
37500	58,36	75,14	69,32	81,31	85,08	86,29	86,20	58,09	75,62	69,32	81,10	87,24	86,22	87,14
40000	58,36	75,27	69,29	81,29	85,13	86,49	86,37	58,31	75,52	69,13	80,92	88,14	87,01	87,87
42500	58,45	75,17	69,35	81,23	85,20	86,63	86,52	58,45	75,55	69,26	81,20	88,18	88,16	87,98
45000	58,47	75,09	69,24	81,16	85,20	86,55	86,49	58,41	75,65	69,35	81,24	88,25	88,37	88,10

Çizelge8.6. Hesaplama birimi kelimenin gövdesi olarak alındığında elde edilen başarımlar oranları tablosu.

Talim Derlem Boyutu	Markov Süreci Başarı Oranları (%)	
	1. dereceden GMM	1. dereceden geçmiş bir adım GMM
250	59,34	63,04
500	62,49	64,56
750	64,83	67,85
1000	68,17	69,11
1250	70,19	71,14
1500	72,24	73,16
1750	73,75	74,18
2000	74,43	75,19
2500	76,75	76,20
3000	79,08	77,22
3500	81,20	78,23
4000	82,60	78,99
4500	83,53	80,25
5000	84,69	82,22
7500	85,60	83,51
10000	86,70	86,08
12500	87,46	89,18
15000	89,49	89,95
17500	90,98	90,38
20000	91,40	91,90
22500	91,52	92,91
25000	91,93	92,66
27500	91,98	92,91
30000	91,89	93,16
32500	92,19	93,42
35000	92,49	93,67
37500	92,87	93,92
40000	93,29	94,08
42500	93,63	94,43
45000	93,94	94,91

Çizelge 8.7. Hesaplama birimi kelimenin çekim ekleri olarak alındığında elde edilen başarımlar oranları tablosu.

Talim Derlem Boyutu	Markov Süreci Başarı Oranları (%)	
	1. dereceden GMM	1. dereceden geçmişe bir adım GMM
250	52,54	55,95
500	55,68	57,22
750	57,61	58,23
1000	57,97	58,48
1250	60,14	60,00
1500	61,71	61,77
1750	62,80	62,53
2000	64,73	63,54
2500	66,18	65,06
3000	67,27	66,84
3500	68,48	68,35
4000	70,29	69,87
4500	71,74	71,14
5000	72,83	73,97
7500	73,43	76,03
10000	74,28	77,58
12500	75,72	78,09
15000	77,17	78,35
17500	77,42	78,87
20000	77,66	79,12
22500	77,90	79,64
25000	78,14	80,15
27500	78,38	80,41
30000	78,74	81,26
32500	79,59	81,44
35000	79,71	81,19
37500	79,59	81,70
40000	79,83	81,96
42500	79,95	82,22
45000	80,07	82,47

Ek 4. Gvdeleme Yntemi Bařarım Oranları

Çizelge 8.8. Szck tr bilinen kelimeler zerinden gerekleřtirilen deney sonuları

Talim Derlemi (Cmle Sayısı)	Varyasyon 1	Varyasyon 2	Varyasyon 3	Varyasyon 4	Varyasyon 5
250	%67,99	%68,07	%57,48	%57,68	%60,01
500	%75,07	%75,77	%62,81	%62,94	%66,30
750	%80,89	%82,29	%69,51	%69,25	%74,04
1000	%81,69	%82,91	%72,29	%71,54	%76,60
1250	%82,14	%84,46	%73,67	%72,79	%77,94
1500	%83,49	%85,29	%76,50	%75,64	%80,09
2000	%85,85	%87,26	%82,04	%81,45	%84,23
2500	%87,21	%89,41	%86,07	%86,00	%87,24
3000	%88,38	%90,61	%90,62	%90,71	%90,52
3500	%88,52	%90,67	%91,43	%91,44	%91,17
4000	%87,99	%90,84	%92,08	%92,02	%91,95
4500	%88,04	%90,86	%92,81	%92,71	%91,97
5000	%88,19	%90,65	%93,04	%93,27	%91,86
5500	%88,33	%90,76	%93,29	%93,63	%92,18
6000	%88,34	%90,90	%93,57	%93,83	%92,48

Çizelge 8.9. Sözcük türü bilinen kelimeler üzerinden gerçekleştirilen deney sonuçları

Talim Derlemi (Cümle Sayısı)	Varyasyon 6	Varyasyon 7	Varyasyon 8	Varyasyon 9	Varyasyon 10
250	%60,01	%68,60	%70,81	%59,50	%60,47
500	%66,16	%73,04	%75,61	%65,28	%66,47
750	%74,00	%77,78	%81,04	%71,91	%73,91
1000	%76,47	%78,31	%82,15	%74,52	%75,58
1250	%77,78	%79,16	%82,72	%76,12	%78,06
1500	%79,80	%79,13	%83,81	%78,28	%80,01
2000	%84,55	%80,50	%84,88	%82,97	%84,87
2500	%87,80	%80,81	%85,91	%87,05	%88,50
3000	%91,49	%81,62	%86,99	%89,81	%91,24
3500	%91,82	%81,62	%87,22	%90,89	%92,24
4000	%92,43	%81,61	%86,84	%91,12	%92,63
4500	%92,68	%80,83	%86,22	%91,22	%92,64
5000	%92,71	%80,19	%85,97	%90,81	%92,60
5500	%92,94	%80,32	%86,18	%90,81	%92,76
6000	%93,17	%80,27	%86,26	%90,75	%92,91

Çizelge 8.10. Sözcük türü belirlenen kelimeler üzerinden gerçekleştirilen deney sonuçları

Talim Derlemi (Cümle Sayısı)	Varyasyon 1	Varyasyon 2	Varyasyon 3	Varyasyon 4	Varyasyon 5
250	%67,68	%67,23	%57,46	%57,23	%59,93
500	%74,39	%74,35	%62,54	%62,02	%65,96
750	%80,01	%80,62	%69,12	%68,23	%73,55
1000	%80,72	%81,06	%71,82	%70,48	%75,95
1250	%81,13	%82,57	%73,15	%71,66	%77,23
1500	%82,40	%83,36	%75,92	%74,34	%79,27
2000	%84,46	%84,93	%81,17	%79,77	%83,19
2500	%85,70	%86,60	%85,05	%83,83	%86,09
3000	%86,75	%87,53	%89,30	%88,03	%89,00
3500	%86,93	%87,57	%90,04	%88,77	%89,60
4000	%86,43	%87,97	%90,73	%89,27	%90,41
4500	%86,52	%87,98	%91,37	%89,85	%90,39
5000	%86,64	%87,78	%91,60	%90,38	%90,23
5500	%86,77	%87,85	%91,83	%90,71	%90,54
6000	%86,75	%88,03	%92,10	%90,92	%90,85

Çizelge 8.11. Sözcük türü belirlenen kelimeler üzerinden gerçekleştirilen deney sonuçları

Talim Derlemi (Cümle Sayısı)	Varyasyon 6	Varyasyon 7	Varyasyon 8	Varyasyon 9	Varyasyon 10
250	%59,42	%68,74	%70,24	%60,11	%60,50
500	%65,13	%73,05	%74,89	%65,78	%66,18
750	%72,93	%77,58	%80,04	%72,35	%73,74
1000	%75,34	%78,06	%80,93	%74,85	%75,33
1250	%76,62	%78,99	%81,45	%76,40	%77,74
1500	%78,51	%78,91	%82,50	%78,52	%79,52
2000	%82,91	%79,91	%83,23	%82,95	%84,07
2500	%85,69	%80,20	%84,03	%86,80	%87,23
3000	%88,84	%80,93	%84,85	%89,34	%89,49
3500	%89,15	%80,90	%85,07	%90,30	%90,34
4000	%89,65	%80,89	%84,81	%90,54	%90,72
4500	%89,72	%80,07	%84,10	%90,62	%90,65
5000	%89,76	%79,43	%83,88	%90,16	%90,54
5500	%89,93	%79,51	%84,11	%90,14	%90,74
6000	%90,15	%79,51	%84,22	%90,10	%90,88

Ek 5. Kök ve Gövdenin Doğru Tespiti Başarım Oranları

Çizelge 8.12. Sözcük türü bilinen olası tüm gövdeler üzerinden gerçekleştirilen deney sonuçları (Varyasyon 1-5)

Talim Derlemi (Cümle Sayısı)	Varyasyon 1	Varyasyon 2	Varyasyon 3	Varyasyon 4	Varyasyon 5
250	%63,07	%64,26	%53,81	%53,97	%59,45
500	%64,39	%65,86	%59,16	%60,35	%63,25
750	%65,98	%67,51	%63,51	%64,05	%65,60
1000	%67,90	%69,48	%67,93	%68,34	%68,89
1250	%69,81	%71,37	%71,12	%71,67	%71,27
1500	%71,83	%73,77	%75,25	%75,79	%74,86
2000	%73,73	%75,32	%80,49	%80,88	%79,12
2500	%75,23	%76,55	%84,10	%84,35	%83,50
3000	%76,70	%78,26	%88,33	%88,67	%86,33
3500	%76,09	%77,91	%88,96	%89,36	%87,82
4000	%75,49	%77,69	%89,47	%89,99	%88,01
4500	%75,07	%77,38	%90,24	%90,40	%88,29
5000	%75,32	%77,57	%90,54	%90,68	%88,50
5500	%75,58	%77,70	%90,95	%91,16	%88,56
6000	%75,83	%77,82	%91,19	%91,46	%88,62

Çizelge 8.13. Sözcük türü bilinen olası tüm gövdeler üzerinden gerçekleştirilen deney sonuçları (Varyasyon 6-10)

Talim Derlemi (Cümle Sayısı)	Varyasyon 6	Varyasyon 7	Varyasyon 8	Varyasyon 9	Varyasyon 10
250	%58,64	%61,84	%63,99	%56,62	%57,82
500	%62,11	%63,69	%65,51	%60,90	%61,46
750	%65,86	%64,97	%67,37	%64,72	%64,88
1000	%69,78	%66,58	%68,84	%68,73	%69,21
1250	%72,93	%67,71	%70,45	%71,11	%71,27
1500	%76,10	%68,71	%71,83	%75,63	%76,04
2000	%79,79	%70,22	%72,87	%80,03	%80,45
2500	%83,72	%71,95	%73,73	%83,76	%84,79
3000	%86,40	%72,71	%74,53	%86,95	%86,89
3500	%86,98	%73,01	%74,79	%87,15	%87,20
4000	%87,36	%73,22	%75,12	%87,25	%87,49
4500	%87,63	%72,05	%75,23	%87,42	%87,58
5000	%87,70	%72,12	%75,25	%87,34	%87,65
5500	%87,75	%72,17	%75,13	%87,54	%87,71
6000	%87,78	%72,22	%75,07	%87,51	%87,75

Çizelge 8.14. Sözcük türü bilinen en uygun gövde üzerinden gerçekleştirilen deney sonuçları (Varyasyon 1-5)

Talim Derlemi (Cümle Sayısı)	Varyasyon 1	Varyasyon 2	Varyasyon 3	Varyasyon 4	Varyasyon 5
250	%61,42	%62,99	%53,21	%53,73	%57,84
500	%63,40	%65,60	%59,32	%59,75	%62,35
750	%66,01	%67,95	%63,58	%63,91	%66,39
1000	%67,59	%69,88	%67,58	%67,96	%69,67
1250	%69,29	%71,57	%71,17	%71,63	%73,32
1500	%70,50	%72,94	%75,51	%75,53	%76,66
2000	%72,16	%75,06	%80,16	%80,58	%80,55
2500	%73,91	%76,94	%83,91	%84,16	%83,92
3000	%75,55	%77,60	%87,65	%87,97	%86,69
3500	%75,04	%77,40	%88,45	%88,83	%86,98
4000	%74,68	%77,30	%88,80	%89,10	%87,44
4500	%74,49	%77,15	%89,27	%89,64	%87,53
5000	%74,60	%77,06	%89,60	%90,09	%87,40
5500	%74,66	%76,98	%89,80	%90,32	%87,51
6000	%74,83	%76,96	%90,05	%90,53	%87,84

Çizelge 8.15. Sözcük türü bilinen en uygun gövde üzerinden gerçekleştirilen deney sonuçları (Varyasyon 6-10)

Talim Derlemi (Cümle Sayısı)	Varyasyon 6	Varyasyon 7	Varyasyon 8	Varyasyon 9	Varyasyon 10
250	%57,51	%59,62	%62,86	%58,31	%58,65
500	%61,80	%61,10	%64,83	%61,27	%62,70
750	%65,89	%62,31	%66,37	%65,13	%65,00
1000	%69,31	%63,51	%68,22	%67,69	%68,51
1250	%72,17	%64,98	%69,72	%70,86	%71,59
1500	%76,12	%66,47	%70,94	%74,58	%75,12
2000	%79,78	%67,90	%72,43	%78,61	%79,69
2500	%83,44	%68,73	%73,64	%82,47	%82,81
3000	%85,57	%69,55	%74,24	%85,98	%86,77
3500	%86,20	%69,08	%74,14	%86,28	%86,99
4000	%86,53	%68,84	%73,77	%86,37	%87,17
4500	%86,73	%68,70	%73,51	%86,52	%87,32
5000	%86,91	%68,46	%73,46	%86,33	%87,51
5500	%87,02	%68,29	%73,39	%86,18	%87,46
6000	%87,11	%68,17	%73,40	%86,01	%87,44

Çizelge 8.16.Sözcük türü bilinmeyen tüm olası gövdeler üzerinden gerçekleştirilen deney sonuçları (Varyasyon 1-5)

Talim Derlemi (Cümle Sayısı)	Varyasyon 1	Varyasyon 2	Varyasyon 3	Varyasyon 4	Varyasyon 5
250	%62,14	%63,02	%52,87	%52,59	%58,30
500	%65,27	%65,72	%59,88	%60,37	%62,81
750	%66,92	%67,45	%63,91	%64,17	%66,66
1000	%68,37	%69,40	%68,14	%68,07	%70,66
1250	%69,57	%70,47	%71,11	%71,16	%73,04
1500	%70,58	%72,08	%74,18	%74,50	%75,54
2000	%72,09	%73,86	%78,60	%78,99	%79,04
2500	%74,39	%75,79	%83,46	%83,87	%83,17
3000	%75,20	%76,20	%86,66	%86,82	%85,35
3500	%74,77	%75,86	%87,20	%87,30	%85,56
4000	%74,03	%75,55	%87,67	%87,78	%85,86
4500	%73,59	%75,41	%88,41	%88,49	%86,03
5000	%74,14	%75,78	%88,70	%88,76	%86,09
5500	%74,79	%75,83	%89,19	%89,11	%86,12
6000	%74,33	%75,82	%89,36	%89,34	%86,33

Çizelge 8.17. Sözcük türü bilinmeyen tüm olası gövdeler üzerinden gerçekleştirilen deney sonuçları (Varyasyon 6-10)

Talim Derlemi (Cümle Sayısı)	Varyasyon 6	Varyasyon 7	Varyasyon 8	Varyasyon 9	Varyasyon 10
250	%51,50	%60,88	%62,50	%50,86	%56,98
500	%58,47	%62,77	%63,23	%56,34	%60,56
750	%63,67	%63,65	%64,55	%61,05	%64,84
1000	%68,24	%64,94	%66,57	%64,49	%69,10
1250	%71,33	%66,18	%68,63	%68,77	%72,17
1500	%74,98	%67,45	%70,03	%74,56	%76,01
2000	%78,96	%68,90	%71,58	%77,94	%79,32
2500	%81,76	%70,06	%72,28	%82,91	%81,95
3000	%84,83	%71,10	%72,75	%85,25	%84,68
3500	%85,21	%70,75	%72,72	%85,46	%84,79
4000	%85,47	%70,66	%72,64	%85,57	%84,96
4500	%85,84	%70,44	%72,80	%85,56	%85,11
5000	%85,76	%70,49	%72,68	%85,50	%85,14
5500	%85,81	%70,56	%72,54	%85,35	%85,18
6000	%86,03	%70,60	%72,60	%85,74	%85,50

Çizelge 8.18. Sözcük türü bilinmeyen en uygun gövde üzerinden gerçekleştirilen deney sonuçları (Varyasyon 1-5)

Talim Derlemi (Cümle Sayısı)	Varyasyon 1	Varyasyon 2	Varyasyon 3	Varyasyon 4	Varyasyon 5
250	%60,58	%61,68	%52,46	%52,33	%57,09
500	%62,83	%64,47	%57,68	%57,42	%61,31
750	%64,83	%66,68	%62,06	%61,61	%64,84
1000	%67,19	%68,34	%65,40	%64,87	%68,37
1250	%68,38	%69,52	%69,31	%68,85	%71,54
1500	%69,99	%71,11	%73,57	%73,15	%74,11
2000	%72,04	%72,37	%78,99	%78,20	%77,08
2500	%73,21	%74,53	%82,90	%81,89	%80,37
3000	%74,10	%75,27	%86,09	%85,85	%84,39
3500	%73,81	%75,11	%86,58	%86,35	%84,67
4000	%73,32	%74,92	%87,30	%86,90	%84,81
4500	%73,02	%74,87	%87,54	%87,23	%84,96
5000	%73,23	%74,65	%87,88	%87,58	%85,05
5500	%73,47	%74,57	%88,24	%88,01	%85,10
6000	%73,40	%74,75	%88,38	%88,22	%85,25

Çizelge 8.19. Sözcük türü bilinmeyen en uygun gövde üzerinden gerçekleştirilen deney sonuçları (Varyasyon 6-10)

Talim Derlemi (Cümle Sayısı)	Varyasyon 6	Varyasyon 7	Varyasyon 8	Varyasyon 9	Varyasyon 10
250	%57,00	%58,82	%61,36	%58,06	%58,07
500	%61,39	%60,93	%63,10	%62,64	%62,96
750	%65,00	%62,44	%65,02	%65,68	%65,80
1000	%68,59	%63,41	%66,58	%68,37	%68,93
1250	%71,79	%64,84	%67,58	%71,67	%71,96
1500	%74,50	%65,35	%69,05	%74,43	%74,81
2000	%77,31	%66,61	%69,97	%78,78	%78,96
2500	%80,68	%67,72	%71,44	%83,36	%83,44
3000	%83,92	%67,96	%71,62	%84,37	%84,49
3500	%84,26	%67,69	%71,57	%84,52	%84,60
4000	%84,75	%67,36	%71,20	%84,63	%84,75
4500	%85,01	%67,04	%70,96	%84,73	%84,84
5000	%85,10	%66,90	%70,89	%84,35	%84,76
5500	%85,17	%66,54	%70,92	%84,12	%84,64
6000	%85,39	%66,50	%70,91	%84,33	%84,90

ÖZGEÇMİŞ

Haziran 1978’ de İzmir’de doğan Tarık KIŞLA, lise öğrenimini İzmir Eşrefpaşa Lisesinde 1994 yılında bitirdikten sonra Ege Üniversitesi Fen Fakültesi Matematik bölümünde lisans öğrenimine başlamıştır. 1998 yılında bilgisayar opsiyonlu matematik bölümünden mezun olan Kışla, yüksek lisans eğitimini Uluslararası Bilgisayar Enstitüsünde 2003 yılında tamamlamış ve aynı yıl içerisinde doktora eğitimine başlamıştır. Doktora eğitimi süresince 2002-2005 yılları arasında Eğitim Fakültesi Eğitim Programları ve Öğretim anabilimdalında bir yüksek lisans programı daha tamamlamıştır. Kışla, lisans eğitimi ardından başladığı meslek hayatında, 3 sene özel sektörde bilgisayar programcısı olarak çalıştıktan sonra 2001-2005 yılları arasında Eğitim Fakültesi Bilgisayar ve Öğretim Teknolojileri Bölümünde araştırma görevlisi olarak görev yapmıştır. 2005 yılından itibaren aynı bölümde öğretim görevlisi olarak çalışmaktadır.

Akademik olarak Türkçe, doğal dil işleme, yazılım, veritabanları, bilgisayar ağları, bilgisayar destekli eğitim ve uzaktan eğitim konularında çalışmalar yapan Tarık Kışla’nın, belirtilen konularda dergi yayınları, çeşitli ulusal ve uluslararası konferanslarda sunduğu bildirileri ve kitapları bulunmaktadır.