

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

**AKRABA VE BİTİŞKEN DİLLER ARASINDA
BİLGİSAYARLI ÇEVİRİ İÇİN
KARMA BİR MODEL**

**DOKTORA TEZİ
Y. Müh. Ahmet Cüneyd TANTUĞ**

Anabilim Dalı : BİLGİSAYAR MÜHENDİSLİĞİ

Programı : BİLGİSAYAR MÜHENDİSLİĞİ

MAYIS 2007

**AKRABA VE BİTİŞKEN DİLLER ARASINDA
BİLGİSAYARLI ÇEVİRİ İÇİN
KARMA BİR MODEL**

**DOKTORA TEZİ
Y. Müh. Ahmet Cüneyd TANTUĞ
504022200**

**Tezin Enstitüye Verildiği Tarih : 20 Şubat 2007
Tezin Savunulduğu Tarih : 9 Mayıs 2007**

Tez Danışmanı : Prof. Dr. Eşref ADALI

Diğer Jüri Üyeleri : Prof.Dr. Kemal OFLAZER (Sabancı Ü.)

Prof.Dr. A. Coşkun SÖNMEZ (Y.T.Ü.)

Y. Doç. Dr. Şule Gündüz ÖĞÜDÜCÜ (İTÜ)

Y. Doç. Dr. Zehra ÇATALTEPE (İTÜ)

MAYIS 2007

ÖNSÖZ

Bu tez çalışması boyunca, desteğini ve yardımlarını benden esirgemeyen tez danışmanım Sayın Prof. Dr. Eşref ADALI'ya, değerli görüşlerini benimle paylaşarak birçok konuda beni yönlendiren Sayın Prof. Dr. Kemal OFLAZER'e, tez süresince çalışmalarımı izleme ve değerlendirme nezaketini gösteren Sayın Prof. Dr. Coşkun SÖNMEZ'e, aileme ve Begüm'e sonsuz teşekkürler.

Mayıs 2007

Ahmet Cüneyd TANTUĞ

İÇİNDEKİLER

KISALTMALAR	vi
TABLO LİSTESİ	vii
ŞEKİL LİSTESİ	viii
ÖZET	ix
SUMMARY	xiii
1. GİRİŞ	1
1.1 Bilgisayarla Dil İşleme ve Bilgisayarlı Çeviri	1
1.2 Çalışmanın Amacı	3
1.3 Önceki Çalışmalar	3
1.4 Tezin Bölümleri	8
2. DOĞAL DİL İŞLEME TEKNİKLERİ	9
2.1 Sonlu Durumlu Makineler	9
2.1.1 Düzenli İfadeler ve Düzenli Diller	10
2.1.2 Sonlu Durumlu Tanıyıcılar	11
2.1.3 Sonlu Durumlu Dönüştürücüler	12
2.2 Biçimbilim	13
2.2.1 İki Düzeyli Biçimbilimsel Çözümleme	15
2.2.1.1 Yazım Kuralları	16
2.2.1.2 Bitiştirme Kuralları	17
2.2.1.3 Sonlu Durum Dönüştürücülerinin Birleştirilmesi	18
2.3 İstatistiksel Dil Modelleri	19
2.3.1 Yumuşatma	20
2.3.2 İDM Değerlendirmesi	21
3. BİLGİSAYARLI METİN ÇEVİRİSİ	24
3.1 Bilgi Tabanlı Çeviri Yöntemleri	26
3.1.1 Doğrudan Aktarım	26
3.1.2 Sözdizimsel Gösterimin Aktarımı	28
3.1.3 Anlamsal Gösterimin Aktarımı	28
3.1.4 Dilden Bağımsız Anlamsal Gösterimin Aktarımı	28
3.2 İstatistiksel Yöntemler	29
3.3 Örnek Tabanlı Yöntemler	32
3.4 Çeviri Kalitesinin Değerlendirilmesi	34
3.4.1 BLEU/NIST	34
3.4.2 F Ölçütü	36
3.4.3 Meteor	37

4. AKRABA ve BİTİŞKEN DİLLER ARASINDA ÇEVİRİ	39
4.1 Giriş	39
4.2 Aktarım Fonksiyonu Modelleri	41
4.2.1 Aktarım Modeli 0 – Temel Model	41
4.2.2 Aktarım Modeli I	44
4.2.3 Aktarım Modeli II	47
4.3 Bitişken Diller İçin İDM Oluşturulması	48
4.3.1 İDM Tip-I – Kök	49
4.3.2 İDM Tip-II – Son Sözcük Türü	50
4.3.3 İDM Tip-III – Son Çekim Grubu	50
4.3.4 İDM Tip-IV – Kök + Son Sözcük Türü	51
4.3.5 İDM Tip-V – Kök + Son Çekim Grubu	52
4.3.6 İDM Tip-VI – Tüm Etiketler	52
4.3.7 Farklı Dil Modeli Tiplerinin Entropi Değerleri	53
5. TÜRK DİL AİLESİ	54
5.1 Türk Dilleri Arasındaki Benzerlikler	56
5.2 Türk Dilleri Arasındaki Farklılıklar	58
5.3 Türk Dilleri Hakkında Özet Bilgiler	59
5.3.1 Azerice	59
5.3.2 Türkmence	61
5.3.3 Kazakça	62
5.3.4 Kırgızca	65
5.3.5 Uygurca	68
5.3.6 Özbekçe	70
6. TÜRK DİLLERİ ARASINDA BİLGİSAYARLI ÇEVİRİ	72
6.1 Kaynak Dilde Biçimbilimsel Çözümleme	74
6.1.1 Kaynak Dilde Biçimbilimsel Belirsizliğin Giderilmesi	75
6.2 Sözcük Köklerinin Kaynak Dilden Hedef Dile Aktarımı	76
6.3 Biçimbilimsel Yapıların Kaynak Dilden Hedef Dile Aktarımı	76
6.4 İDM Bileşeni	77
6.5 Hedef Dilde Biçimbilimsel Üretici	79
7. TÜRKMECEDEN TÜRKÇEYE BİLGİSAYARLI ÇEVİRİ SİSTEMİ	80
7.1 Aktarım Modeli 0 Gerçeklemesi	80
7.1.1 Türkmence Biçimbilimsel Çözümleyicinin Geliştirilmesi	81
7.1.1.1 Türkmen Dilinin Biçimbilimsel Özellikleri	81
7.1.1.2 İki Düzeyli Kurallar	85
7.1.1.3 Morfotaktik Kurallar	91
7.1.2 Kök Sözcük Aktarım Kuralları	94
7.1.2.1 Birden Fazla Sözcükten Oluşan Karşılıklar	94
7.1.2.2 Sözcüksel Aktarım Kuralları	95
7.1.3 Biçimbilimsel Aktarım Kuralları	96
7.1.4 İstatistiksel Dil Modeli Bileşeni	97
7.1.5 Türkçe Biçimbilimsel Üretici	97
7.2 Aktarım Modeli I Gerçeklemesi	98
7.3 Aktarım Modeli II Gerçeklemesi	101
7.4 İki Seviyeli İDM Kullanılması	104

8. UYGULAMA VE SONUÇLAR	106
8.1 Eğitim ve Sınama Verisi	106
8.2 Değerlendirme Ölçütleri	107
8.3 Sonuçlar	110
8.3.1 Temel Modelin Başarımı	110
8.3.2 Aktarım Modeli I'ın Başarımı	111
8.3.3 Aktarım Modeli II'ın Başarımı	111
8.3.4 Farklı Türde Dil Modellerinin Başarımı	112
8.3.5 İki Seviyeli İDM Başarımı	119
8.4 Hatalı Durumların İncelenmesi	121
8.5 Çeviri Örnekleri	122
8.6 Aktarım Süreleri	124
9. DEĞERLENDİRME VE TARTIŞMA	126
10. KAYNAKLAR	131
EK A – BİÇİMBİLİMSEL ETİKETLERİN AÇIKLAMALARI	138
EK B – İNGİLİZCE TERİMLERİN TÜRKÇE KARŞILIKLARI	142
ÖZGEÇMİŞ	145

KISALTMALAR

DDİ	: Doğal Dil İşleme
BÇ	: Bilgisayarlı Çeviri
HMM	: Hidden Markov Model
MM	: Markov Model (Visible Markov Model)
ASCII	: American Standard Code for Information Interchange
İDM	: İstatistiksel Dil Modeli
ÇG	: Çekim Grubu (Inflectional Group)
SDD	: Sonlu Durumlu Dönüştürücü
SDT	: Sonlu Durumlu Tanıyıcı
BLEU	: Bilingual Evaluation Understudy
METEOR	: Method for Evaluation of Translation with Explicit Reordering

TABLO LİSTESİ

	<u>Sayfa No</u>
Tablo 2.1	İki Düzeyli Biçimbilimde Kural Türleri 17
Tablo 4.1	Derlemde Gözlenen Tam Çözümleme ve Çekim Grubu Sayıları..... 49
Tablo 4.2	İDM Tip-I İçin Örnek Eğitim Tümcesi..... 49
Tablo 4.3	İDM Tip-II İçin Örnek Eğitim Tümcesi..... 50
Tablo 4.4	İDM Tip-III İçin Örnek Eğitim Tümcesi 51
Tablo 4.5	İDM Tip-IV İçin Örnek Eğitim Tümcesi 51
Tablo 4.6	İDM Tip-V İçin Örnek Eğitim Tümcesi 52
Tablo 4.7	İDM Tip-VI İçin Örnek Eğitim Tümcesi 52
Tablo 4.8	Farklı Tipte ve Derecede İDM’lerin Entropi Değerleri 53
Tablo 5.1	Türk Dilleri İle İlgili Bazı Bilgiler 55
Tablo 5.2	Bazı Türk Dilleri İçin İsim Durum Ekleri..... 57
Tablo 5.3	Azericenin Abecesi 59
Tablo 5.4	Türkmencenin Abecesi 61
Tablo 5.5	Kazakçanın Abecesi 62
Tablo 5.6	Kırgızcanın Abecesi..... 65
Tablo 5.7	Uygurcanın Abecesi 68
Tablo 5.8	Özbekçenin Abecesi..... 70
Tablo 6.1	Geliştirilmiş Doğrudan Aktarım Yöntemi Aşamaları 73
Tablo 6.2	İDM İle En Olası Tümcenin Bulunması 79
Tablo 7.1	Türkmencede Çatı Eklerinin Sıralanışı 84
Tablo 7.2	ASCII Olmayan Karakterler Yerine Kullanılan Karşılıklar..... 86
Tablo 8.1	Eğitim Derlemi İle İstatistikler 106
Tablo 8.2	Sınama Kümesinde Ölçülen Belirsizlik Oranları 107
Tablo 8.3	Aktarım Modeli 0’ın (Temel Model) Başarımı..... 110
Tablo 8.4	Aktarım Modeli I Başarımı 111
Tablo 8.5	Aktarım Modeli II'nin Başarımı 111
Tablo 8.6	İDM Tip-I Başarımı 112
Tablo 8.7	İDM Tip-II Başarımı 113
Tablo 8.8	İDM Tip-III Başarımı..... 114
Tablo 8.9	İDM Tip-IV Başarımı 114
Tablo 8.10	İDM Tip-V Başarımı..... 115
Tablo 8.11	İDM Tip-VI Başarımı 116
Tablo 8.12	İki Seviyeli İDM İndirgemesinin <i>BLEU</i> Puanları 119
Tablo 8.13	İki Seviyeli İDM İndirgemesinin <i>BLEUr</i> Puanları 120
Tablo 8.14	Aktarım Süreleri..... 124

ŞEKİL LİSTESİ

Sayfa No

Şekil 2.1	: Örnek Bir Sonlu Durumlu Tanıyıcı.....	11
Şekil 2.2	: Düzenli İfade-Düzenli Dil-Sonlu Durumlu Tanıyıcı Arasındaki İlişki..	12
Şekil 2.3	: Sonlu Durumlu Dönüştürücü Örneği	12
Şekil 2.4	: SSD'lerin Birleştirilmesi.....	13
Şekil 2.5	: Türkçe Biçimbilimsel Dönüştürücü	13
Şekil 2.6	: "Xerox Lexc" İçin Sözlük Yapısı	18
Şekil 2.7	: İki Düzeyli Biçimbilimsel Çözümleyici/Üretici.....	18
Şekil 3.1	: Bilgi Tabanlı Yöntemlerin Sınıflandırılması-Vauquouis Üçgeni	26
Şekil 3.2	: İngilizce-Fransızca Arası Sözcüksel Belirsizlik Örneği.....	27
Şekil 3.3	: Doğrudan Aktarım Yöntemleri İçin Örnek Aktarım Süreci.....	27
Şekil 3.4	: Sözdizimsel Gösterimin Aktarımı	28
Şekil 3.5	: Anlamsal Gösterim Örneği.....	29
Şekil 3.6	: Gürültü Kanal Modeli Uyarınca Çeviri İşlemi	30
Şekil 5.1	: Türk Dil Ailesinin Sınıflandırması.....	54
Şekil 5.2	: Türk Dillerinin Konuşulduğu Başlıca Coğrafyaların Haritası	55
Şekil 5.3	: Türkçe-Özbekçe Tümcelerde Sözcük Sıraları Farklılığı Örneği.....	58
Şekil 6.1	: Örnek Türkmence Sözcüğün Türkçe Karşılığının Oluşturulması.....	74
Şekil 6.2	: Temel Modeli Gerçekleyen Örnek Bir Çeviri Sistemi.....	75
Şekil 6.3	: Örnek Bir Tümcenin HMM İle Çözümleme Süreci.....	78
Şekil 7.1	: Aktarım Modeli 0 Temelinde Oluşturulan Sistemin Bileşenleri.....	80
Şekil 7.2	: İsim Soylu Sözcükler İçin Morfotaktik Kuralların SDMSi	92
Şekil 7.3	: Eylem Soylu Sözcükler İçin Morfotaktik Kuralların SDMSi	92
Şekil 7.4	: Eylem Soylu Sözcükler İçin Morfotaktik Kuralların SDMSi (Devamı)	93
Şekil 7.5	: Kök Aktarım Bileşeni.....	94
Şekil 7.6	: ÇSG'lerin Aktarılması.....	95
Şekil 7.7	: Aktarım Modeli I Temelinde Oluşturulan Sistemin Bileşenleri	98
Şekil 7.8	: Çekimlenen ÇSG'lerin Bulunması.....	100
Şekil 7.9	: Aktarım Modeli II Temelinde Oluşturulan Sistemin Bileşenleri	101
Şekil 7.10	: İki Seviyeli İDM Uygulaması	105
Şekil 8.1	: Aktarım Modeli 0'ın Farklı İDM Tipleri İle Başarım Grafiği	116
Şekil 8.2	: Aktarım Modeli I'ın Farklı İDM Tipleri İle Başarım Grafiği.....	117
Şekil 8-3	: Aktarım Modeli II'nin Farklı İDM Tipleri İle Başarım Grafiği	118
Şekil 8-4	: İki Seviyeli İDM Kullanımında Başarım Karşılaştırmaları	120

AKRABA VE BİTİŞKEN DİLLER ARASINDA BİLGİSAYARLI ÇEVİRİ İÇİN KARMA BİR MODEL

ÖZET

Diller arası çeviride bilgisayarların kullanılması fikrinin başlangıcı 1950'lerin ilk yıllarına kadar uzanmaktadır. O tarihten günümüze kadar teknolojik alanda yaşanan gelişmelere, konunun çekiciliği yüzünden akademik ve ticari çevrelerde bu konu ile uğraşan araştırmacı sayısının fazlalığına, siyasi, ekonomik ve sosyal nedenlerden dolayı devletler tarafından desteklenmesine ve yüklü miktarlarda bütçeler ayrılmasına karşın, günümüzde dahi bilgisayarlı çeviri (BÇ) tam anlamıyla hayata geçirilememiştir. Elbette ki bütün bu emeklerin sonucunda BÇ alanında belirli bir noktaya gelinmiştir. Ancak hedeflenen, şu üç özelliği bir arada sağlayan sistemlerin ortaya çıkmasıdır:

1. İnsan etkisine gerek duymadan tam otomatik çeviri yapabilme yeteneği,
2. Belirli bir konuya bağlı olmaksızın genel amaçlı metinler üzerinde çeviri yapabilme yeteneği,
3. Üretilen sonuçların, insanların hazırladığı çeviriler gibi anlaşılır olması ve aslına uygun olan anlamı taşıması

BÇ'nin zorluğunun en temel nedeni, diller arasındaki büyük farklılıklardır. Aynı dil ailesinde bulunan diller arasında çeviri yapılırken bile karşılaşılan zorluklar, tamamen farklı dil ailelerinden gelen ve bütünü ile farklı dilbilgisel yapıya sahip diller arası çevirilerde yaşanan zorluklarının boyutunu göz önüne sermektedir. Günümüzde BÇ çalışmalarının çoğu İngilizce, Almanca, Fransızca, Çince ve Arapça gibi yaygın kullanıma sahip diller arasında gerçekleştirilmektedir. Son yıllarda ise Çinceden İngilizceye ve Arapçadan İngilizceye çeviri çalışmaları, ABD'nin sağladığı desteklerle yoğunluk kazanmıştır.

Yöntembilim açısından incelendiğinde, BÇ alanında iki farklı dönem olduğu görülmektedir: 1950'lerden 1990'ların başına kadar geliştirilen sistemler çoğunlukla

çeşitli dilbilgisel düzeylerde (biçimbilimsel, sözdizimsel, anlamsal) çalışan kural tabanlı yöntemler kullanılarak gerçekleştirilmişlerdi. Bunlara örnek olarak Meteo, Systran, METAL, Logos gibi sistemler sayılabilir. 1990'lı yıllardan başlayarak, ses işleme ve tanıma gibi alanlarda başarısı kanıtlanmış olan istatistiksel yöntemlerin BÇ alanında da denenmesi yoluna gidilmiştir. IBM'in önderliğini yaptığı bu çalışmalar sonucunda elde edilen başarılar, kural tabanlı sistemlerde tıkanan ve ileri gidemeyen çalışmaların büyük bir bölümünü istatistiksel BÇ'ye yöneltmiştir. Yine son 20 yılda ortaya çıkan ve çevrilecek metni bir veri tabanındaki örneklerle benzeterek çıktı üreten üçüncü yöntem, örnek tabanlı BÇ yöntemi, kullanılmaya başlanılmıştır. Günümüzde de çalışmaların çoğunluğu istatistiksel BÇ ve örnek tabanlı BÇ ekseninde gerçekleştirilmektedir.

BÇ önünde en büyük engel olan dillerin arasındaki farklılıklar, akraba ya da yakın dil çiftleri için daha az görüleceğinden, bu diller arasında BÇ'nin daha kolay olacağı sezgisel olarak anlaşılmaktadır. Hatta sözdizimsel açıdan aynı olan birçok dil çifti için sözcük temelinde yapılacak doğrudan aktarım yöntemlerinin bile başarılı olabileceği söylenebilir. Nitekim 2000'li yılların başından itibaren yayınlanan bazı çalışmalarda, benzer ya da yakın akraba dil çiftleri arasında BÇ yapmanın, birbirlerinden çok farklı dil çiftleri arasında BÇ yapmaktan daha kolay olması gerektiği ileri sürülmüş ve bu fikirden hareketle çeşitli uygulamalar gerçekleştirilmiştir. Çekçe ve İspanyolca olmak üzere iki merkezde toplanan çalışmalarda Çekçe-Slovakça, Çekçe-Lehçe, Çekçe-Litvanyaca, İspanyolca-Katalanca gibi birbirlerine yakın diller arasında BÇ ile başarılı sonuçlar elde edilmiştir. Anılan bu dil çiftleri birbirlerine çok benzemektedirler (Litvanyaca hariç). Hatta, buradaki hedef diller, kaynak dil ile neredeyse tamamen aynı sözdizimsel yapıya sahiptir. Bütün çeviriler, sözcük sıraları değiştirilmeden sözcük temelinde gerçekleştirilmektedir. Üstelik eşadlı (homonym) sözcükler dahi hedef dile çevrildiği zaman genellikle aynı biçimde eşadlı olarak kaldığından, çeviri sözlükleri bire bir çıktı üretecek şekilde tasarlanmıştır.

Yapılan çalışmaların diğer benzer ya da akraba diller için genişletilmesi noktasında birtakım sorunlar ortaya çıkmaktadır. Bunlardan en önemlisi, sözcük seçiminde anlamsal farklılıkların olduğu dil çiftlerinde görülmektedir. Her ne kadar seçilen dil çifti yakın olsa bile, kaynak dildeki her sözcüğün hedef dilde sadece bir karşılığının olması, her dil çifti için geçerli değildir. Seçilecek hedef dil karşılıklarının, bağlama

bağlı olarak belirlenmesinin gerekli olduğu durumlarda bu sistemler yetersiz kalmaktadır. Bir başka sorun ise, örneğin Türkçe gibi bitişken dillerin sahip olduğu karmaşık biçimbilimsel yapıların aktarılması noktasında ortaya çıkmaktadır. Söz konusu sistemler, kaynak dil için biçimbilimsel çözümleyici, sözcük türü etiketleyici ve hedef dil için de biçimbilimsel üretici araçlara gerek duymaktadır.

Tez çalışması kapsamında, akraba ve bitişken olan dil çiftleri arasında BÇ için bazı çeviri modelleri önerilmiş, bu modellerle Türk Dil ailesi için BÇ konusu incelenmiştir. Önerilen çeviri modelleri, kural tabanlı çeviri ile istatistiksel çeviri yöntemlerinin birleşiminden oluşan karma modellerdir. Önerdiğimiz modellerin matematiksel tanımı verildikten sonra modeller, Türk Dil ailesi içerisinde BÇ açısından değerlendirilmiştir. Ancak önerilen çeviri modelleri Türk Dillerine özgü olmayıp, bitişken ve akraba ya da yakın iki dil çifti arasında BÇ için kullanılabilir niteliktedir. Modelin uygulaması için, Türkmenceden Türkçeye bir aktarım sistemi gerçekleştirilmiş ve sonuçları değerlendirilmiştir.

Türk dilleri, Azerice, Türkmence, Özbekçe, Tatarca, Kazakça, Kırgızca, Uygurca vb. gibi dillerden oluşan, toplamda yaklaşık 180 milyon insan tarafından geniş bir coğrafyada kullanılmakta olan dillerden oluşmaktadır. Bu dillerin benzerlik ve farklılıkları değişmekle beraber, çoğu Türk dili, sözdizimsel açıdan hemen hemen aynıdır. Bütün Türk dilleri üretken türetme ve çekim özellikleri içerdiğinden, diğer dillere oranla daha karmaşık biçimbilimsel yapılara sahiptirler. Türk Dilleri üzerinde çalışmanın en büyük zorluğu, doğal dil işleme (DDİ) açısından gerekli araçların henüz oluşturulmamış olmasından kaynaklanmıştır. Türkçe, DDİ alanında günümüze kadar yapılan çalışmalar sonucunda ortaya konulan çeşitli araçlar ve veri kaynakları ile diğer Türk Dillerinden bu konuda ayrılmaktadır.

Önerilen modellerin başarımlarını ölçmek ve gerçekleştirmede ortaya çıkan pratik sorunları görmek için sonlu durumlu dönüştürücüler, Hidden Markov Modelleri (HMM), Viterbi algoritması gibi yöntemleri kullanan bir uygulama gerçekleştirilmiştir. Uygulama sonucunda önerilen modellerin başarıları karşılaştırılmış, sistemin zayıf ve kuvvetli yönleri irdelenmiştir. Gerçeklenen uygulama, sadece Türkmence - Türkçe dil çifti için değil, önerilen modelleri kullanarak tüm Türk Dilleri arasında çeviri yapılabilmesini sağlayacak çerçeve bir sistem geliştirilmiştir.

Özet olarak tez çalışması kapsamında bitişken yapıya sahip akraba ya da benzer diller arasında BÇ için, kural tabanlı ve istatistiksel tabanlı bileşenlerden oluşan karma çeviri modelleri tanımlanmış, bu modeller Türk Dil Ailesi'ndeki diller için incelenmiş, Türkmenceden Türkçeye bir aktarım sistemi gerçekleştirilerek önerilen modellerin gerçekten başarılı çeviri yapabildiği gösterilmiştir.

A HYBRID MODEL FOR MACHINE TRANSLATION BETWEEN AGGLUTINATIVE AND RELATED LANGUAGES

SUMMARY

The idea of using computers for natural language translation has a very long history which goes back to the beginnings of 1950s. Unfortunately, the ideal machine translation (MT) could not be achieved truly despite of the huge development in technology, the enormous amount of researchers all over the world and tons of projects mainly supported by governments with very high budgets. An ideal MT should have the following three properties:

1. Ability to process without any human intervention
2. Ability to process text without any domain/topic restriction
3. Ability to produce high quality outputs, can easily be understood by human and preserving the meaning of the original sentence

MT is a hard task because of the differences between languages like typological, morphological, structural and lexical divergences. Even MT between languages belonging to the same language family suffer from these differences.

Most of the work in the MT field is concentrated around the major languages like English, German, French, Chinese and Arabic. Recently, US Government supports the efforts of MT from Arabic or Chinese to English.

From the point of view of the methodologies used, MT history can be separated into two sections. The first part took place between the early 1950s and the late 1980s. During this period, MT was primarily performed using rule based transfers of some representation levels like morphological, syntactical or semantic representations. Some of the famous systems from this period are Meteo, Systran, METAL and Logos. In the late 1980s, statistical methods appeared in MT arena. Statistical methods were proven techniques in other fields like speech recognition and pattern

processing. Statistical MT (SMT), led by IBM, encouraged researchers to shift their motivation from complex and hard to improve rule based systems to brand-new SMT based methods. Meanwhile, a third approach for MT emerged: Example Based MT (EBMT). This approach is based on translation by making an analogy between the input and previously seen training sentence pairs. Today, most of the work in MT field has concentrated around SMT and EBMT methods.

Intuitively, translation between related languages seems to be accomplished easier than the translation between distinct language pairs. Besides, it can be said that word-by-word direct transfer of source language (SL) to the target language (TL) can produce successful translations if both the source and target language has the same or similar word order. MT between related languages is recently a new topic where the first studies are carried out mostly after 2000. Studies on this topic have been done around these two languages: Czech and Spanish. MT systems between very close language pairs like Czech-Slovak, Czech-Polish, Czech-Lithuanian and Spanish-Catalan are implemented within the context of direct transfer fashion. The results of these MT systems showed that successful translations can be generated by employing relatively less effort. All of the language pairs above are very close languages (except Czech-Lithuanian which is only close) and the syntactical structures of both of the languages are almost the same. Even homonyms preserve their homonymy after translation, so it is usually enough to transfer the root words in one-to-one manner.

While expanding these studies to other cognate language pairs, some serious problems can emerge. The most important problem is the lexical ambiguity problem; which means existence of multiple TL words for one SL word. Generally, the selection of the right TL equivalent requires the usage of context. However the systems above do not process polysemous words because of the assumption stating that there is always one-to-one mapping between SL and TL words. Another problem can be observed in MT between close and agglutinative language pairs which require a complex transformation of both root words and the morphology.

In this work, we present some new MT models for translation between agglutinative and related languages. The proposed MT models are hybrid models which have both statistical and rule-based components. The mathematical explanations of these models are presented in this thesis. We have investigated the usability of these

models for Turkic Languages and have implemented a real MT system from Turkmen Language to Turkish in order to evaluate our models. Although most of the examples and the implementation are on Turkic Languages, the proposed models are not designed only for Turkic Languages. In fact, these models are generic models that can be applicable to all other related language pairs that have agglutinative morphologies (like Finnish-Estonian).

Turkic Language family includes cognate languages like Turkish, Azeri Language, Turkmen Language, Uzbek Language, Uighur Language, Kazakh Language and etc. These languages are used more than 180 million people especially in Central Asia and West Asia. All of these languages share lots of akin properties as well as being agglutinative and having the same word order (SOV). Having productive inflectional and derivational morphology is the major common property of Turkic Languages. The most serious problem for MT between Turkic Languages is the lack of NLP related tools such as morphological analyzers, morphological disambiguators and training corpus for most of these languages (except Turkish). At this point, it must be asserted that these languages are not mutually intelligible despite these similarities. Otherwise there would be no need for translation.

In this thesis, we implement an MT tool employing our models by using finite state methods. As the statistical component, we have used Hidden Markov Models (HMM) and a modified Viterbi Algorithm for the decoding process. For performance evaluation of the proposed methods, test translations of our MT implementation are used. The framework used in the implementation can be used for almost all Turkic Languages.

In summary, we present a hybrid translation model for MT between related and agglutinative language pairs. This model is based on both statistical and rule-based transfer approaches. MT between Turkic Languages is investigated and a test implementation is performed from Turkmen to Turkish. The MT implementation has succeed in generating satisfactory translations. Even though the *BLEU* evaluation method underestimates the quality of translations in agglutinative and word order free languages, we get remarkable *BLEU* scores with our simple but efficient implementation.

1. GİRİŞ

“Dil, insanların meramlarını anlatmak için kullandıkları bir sesli işaretler dizisidir.” [1]. En kısa ve temel tanımlarından biri yukarıdaki gibi verilmiş olan dil, insanların birbirleri ile anlaşmalarını sağlayan en güçlü araçtır. Ancak dili oluşturan bu sesli işaretler dizisi, dünya üzerinde farklı coğrafyalarda çok değişik biçimlerde kullanılmaktadır. Dillerdeki bu farklılık, dilin bilgisayarla işlenmesi noktasında genel geçer kuralların ve yöntemlerin kullanılmasını sınırlamış, her dilin kendine özgü sorunlarının çözümü için özel çaba harcanmasını ve yeni yöntemlerin geliştirilmesini zorunlu kılmıştır. Türkçemiz üzerine şimdiye dek yapılan çalışmalar, İngilizce, Fransızca, Almanca gibi diğer başlıca (hatta göreceli olarak çok daha az insanın kullandığı Danimarkaca, Hollandaca gibi diğer) diller üzerine yapılan çalışmalar yanında sayısal olarak oldukça yetersiz kalmaktadır. Üstelik Türkçe, yapısal ve dilbilgisel zenginlikleri ve kendine has farklılıkları ile diğer dillerden çok daha değişik bir konumdadır. Bu tez ile Türkçe ve diğer Türk dilleri üzerine yapılan çalışmalara katkıda bulunulması hedeflenmiştir.

1.1 Bilgisayarla Dil İşleme ve Bilgisayarlı Çeviri

Yapay zeka ve dilbilimi disiplinlerinin bir alt dalı olan doğal dil işleme konusu, temel amaç olarak insanların iletişim amaçlı kullandıkları dilleri (doğal diller) otomatik olarak çözümlemeyi, anlamayı, yorumlamayı ve üretmeyi gütmektedir.

Günümüzde insan-bilgisayar iletişimi (human-computer interaction), çoğunlukla insanların bilgisayarları çeşitli araçlarla (tuş takımı, fare gibi) yönetmeleri ile sağlanmaktadır. Doğal dil işleme araştırmalarının amacı insanların bilgisayarlar ve genel olarak makinelerle iletişimlerini çok daha doğal yollardan yani konuşarak ya da yazarak gerçekleştirebilmelerini sağlamaktır. Bu kapsamda konuşma tanıma, konuşma üretme, yazılı metinden anlam çıkarma, veri tabanında bulunan bilgileri tümceye dökme, bilgisayarlı çeviri yapma, soru-cevap sistemleri geliştirme, özet çıkartma, yazım hatası düzeltme gibi bir çok uygulama hayat bulmaktadır.

Doğal dil işleme yöntemleri çeşitli bilgi düzeylerinde çalışırlar:

1. Sesbilim (Phonetics & Phonology)

Konuřmaları, sesleri ve harflerin okunuřlarını inceler.

2. Biçimbilim (Morphology)

Sözcüklerin yapı taşlarını (anlamalı alt birimlerini) inceler.

3. Sözdizimi Düzeyi (Syntax)

Dilbilgisi yapısını ve tümcenin sözcükleri arasındaki ilişkileri inceler.

4. Anlambilim (Semantics)

Dildeki her türlü anlam konusunu inceler.

5. Bağlam Düzeyi (Pragmatics)

Dilin, farklı bağlamlarda kullanımının getirdiğı farklılıkları inceler.

6. Konuşma Düzeyi (Discourse)

Birbirini takip eden konuşmaları (diyalog, monolog gibi) inceler.

Hiç kuřkusuz, doğal dil işleme alanındaki bütün uygulamaların en büyük sıkıntısı dillerdeki karmařıklık ve belirsizliktir. İnsanlar arasında dahi iletişim güçlüklerine ve yanlış anlaşılmalara yol açan dildeki bu belirsizlikler ve karışıklıklar, bilgisayar ortamında doğal dillerin modellenmesinin önündeki en büyük engeli oluşturmaktadır.

Bilgisayarlı çeviri (BÇ), doğal dil işlemenin en güncel konuları arasında gelmektedir. Her geçen gün daha da küreselleřen dünyada, farklı dilleri konuşan insanların iletişim kurmaları, geçmiş yıllara göre çok daha önem kazanmıştır. Bu yüzden dil engelini kaldıracak her türlü çalışmaya gereksinim her zamankinden fazladır. Günümüzde, bilgisayarlı çeviri alanında başarılı sayılabilecek sonuçlar üretebilen sistemler bulunmakla beraber genel amaçlı kullanılabilecek yüksek başarılı BÇ sistemleri henüz geliştirilememiştir. Ayrıca geliştirilen sistemler az sayıda dil çifti için gerçekleştirildiğinden, bir çok dil çifti arasında BÇ sistemi henüz bulunmamaktadır. BÇ tarihi incelendiğinde, 1980'lerin sonlarına kadar insan emeğı ile oluşturulan kural tabanlı sistemlerin kullanıldığı görülürken, 1990'ların başlarından itibaren elektronik ortamdaki metinlerin miktarlarının ve kullanılabilirliğinin artması ile istatistiksel ve örnek tabanlı yöntemlerin yaygınlaştığı görülür. Ancak bütün çalışmalara ve teknolojik gelişmelere karşın arzulanan BÇ sistemlerinin geliştirilmesi řu an için gerçekleştirilebilir görünmemektedir.

1.2 Çalışmanın Amacı

Bu tez çalışmasının amacı, bilgisayarlı çeviri konusunda bilinen eski tekniklerle son yıllarda ortaya çıkan bilgi ve istatistik tabanlı yeni teknikleri beraber kullanarak, akraba ve bitişken yapıları diller arasında çeviri yapabilecek bir çeviri modelini geliştirmektir. Tarihsel, yapısal ve dilbilgisel açıdan farklılıklar gösteren dil çiftleri arasında çevirinin yapılabilirliği halen uzak olmasına karşın, akraba dil çiftleri arasında gözlenen yapısal ve dilbilgisel benzerliklerin getirdiği kolaylıklar sayesinde yakın dil çiftleri arasında bilgisayarlı çeviri sistemlerinin geliştirilmesi daha akla yakın gelmektedir.

Türkçe, Azerice, Türkmence, Özbekçe, Kırgızca, Kazakça gibi Türk Dilleri, ortak tarih ve coğrafya hamuru içerisinde oluştukları için bugün dahi İngilizce, Almanca gibi batı kökenli diğer dillerle kıyaslanmayacak derecede birbirlerine benzemektedirler. Her ne kadar uzun yıllar boyu süren Sovyet yönetiminin getirdiği olumsuz etkiler bu diller üzerinde önemli tahribatlara yol açmış olsa da Türk dilleri arasındaki yakınlıklar ve benzerlikler belirli ölçeklerde günümüze kadar korunmuştur.

Bu çalışmada, Türk dillerinin özellikleri ile bilinen bilgisayarlı çeviri yöntemleri incelenmiş, Türk dilleri arasında bilgisayarlı çeviri yapılabilmesi için yeni matematiksel modeller önerilmiştir. Bu modellere göre Türkmence-Türkçe dilleri arasında çeviri yapabilecek bir sistem gerçekleştirilerek uygulamada ortaya çıkan aksaklıklar irdelenmiş, bu sorunlara yeni çözüm yöntemleri önerilerek bu yöntemlerinin etkileri araştırılmıştır.

1.3 Önceki Çalışmalar

Bilgisayarların, farklı diller arasında çeviri yaptırmak amacıyla kullanılması fikri yaklaşık 50 yıllık bir geçmişe sahiptir. Bu süre boyunca dillerin doğasındaki karmaşıklığı ve değişik diller arasındaki farklılıkları gidererek çeviri yapabilmek için çeşitli bilgi seviyelerinde birçok yöntem denenmiştir. Günümüzde iki dil arasında, konudan bağımsız olarak yüksek kaliteli çevirileri otomatik üretebilen sistemler henüz geliştirilememiştir. Ancak geliştirilen sistemler içerisinde, aralarında derin ayrılıklar içermeyen dil çiftleri arasında çeviri yapan sistemlerin sonuçlarının, yapısal olarak çok farklı diğer dil çiftleri arası çevirilerden daha kaliteli olduğu

görülmektedir. Örneğin aynı dil ailesinde sınıflandırılan İngilizce ve Almanca arasında gerçekleşen çeviri sistemlerinin başarısı, farklı dil ailelerinde olan İngilizce ve Japonca arasındaki bilgisayarlı çeviri sistemlerinin başarısından daha yüksektir [2]. Bu tespit, diller arasındaki benzerlik oranı arttıkça bilgisayarlı çeviri sistemlerinin geliştirilmesinin daha kolay olacağı ve yüksek başarımlı çevirilerin elde edilebileceği izlenimini uyandırmaktadır.

Benzer diller arasında yapılan ilk bilgisayarlı çeviri çalışması, Çekçe ile Rusça arasında bilgisayarlı çeviri sistemi olan RUSLAN sistemidir [3]. Eski Doğu Bloku ülkelerinde, anaçatı (mainframe) işletim sistemleri ile ilgili her türlü belgenin Rusçaya çevrilmesinin zorunluluğu bulunduğundan, 1985'te geliştirilmeye başlanmış olan bu sistem 1990'larda ödenek yetersizliğinden durdurulmuştur. Gerçeklenen sistem sadece Çekçeden Rusçaya çeviri yapabilmekteydi.

Bu sistemde, kural tabanlı biçimbilimsel çözümleyici, çeviri ve biçimbilimsel üretici modülleri bulunmakla beraber, sözdizimsel ayrıştırıcı (Çekçe için) ve birleştirici (Rusça için) modülleri de bulunmaktadır.

Çalışmanın sonuçlarında, hedef dilde oluşturulan tümcelerın %40'nın doğru olarak çevrildiğı, %40'nın bir kullanıcı tarafından düzeltilmesi gereken ufak hatalar içerdiğini, geri kalan %20'lik bölümün ise tamamen baştan çevrilmesi gerektiğı bildirilmiştir. Başarısız çevirmelerin nedenleri olarak, çeviri sözlüğünün küçük olması (10.000 sözcük) ve bu sözlükte Çekçenin, kural tabanlı bir çözümleyici ile çözümlenemeyen bir çok sözdizimsel durumun varlığı gösterilmiştir.

Çekçe üzerine yapılan bir başka çalışma ile RUSLAN sistemine benzer bir çeviri sistemi, birbirlerine çok benzeyen iki dil, Çekçe ve Slovakça arasında geliştirilmiştir [4]. ČEŠILKO adı verilen bu çalışmada, kaynak ve hedef diller birbirlerine çok benzediğı için daha basit yöntemler kullanılmıştır. Bu yöntemlerin temelinde sözcük bazında çeviri yapılmaktadır. Aşağıda bu sistemi oluşturan bileşenler tanıtılmıştır:

1. Çekçe biçimbilimsel çözümleme
2. Çekçe biçimbilimsel belirsizliğin giderilmesi
3. Konuya özel aktarım sözlüğü (tek ve birden fazla sözcükten oluşan girdiler)
4. Genel amaçlı aktarım sözlüğü
5. Slovakça biçimbilimsel üretim

Çekçenin biçimbilimsel çözümlemesi için, başarısı %99 olarak belirtilen bir çözümleyici (700.000 kök girdili) kullanılmıştır. Sözcük türü etiketlemesi için istatistiksel yöntemlere dayanan bir araç kullanılmıştır. Bu yöntemde, gazetelerin haberlerinden oluşan bir eğitim kümesi üzerinde eğitilen ve başarı oranının %98 olduğu belirtilen “log-lineer” olasılık dağılımı modeli kullanılmıştır. Sözcüklerin ve biçimbilimsel bilgilerin aktarılması için çift dilli aktarım sözlükleri kullanılmıştır. Eşsesli sözcükler çevrildikten sonra da gene eşsesli kaldıkları için iki dil arası çeviri de herhangi bir anlam belirsizliği oluşmadığı belirtildiğinden, aktarım sözlüğü, bire-bir aktarım yapacak şekilde geliştirilmiştir.

ČESILKO sistemin başarımı Çekçe-Slovakça arasında %90 civarında rapor edilmiştir. Daha sonra bu çalışma diğer Slav dilleri için de genişletilmiş, ilk olarak yine Çekçeye benzer bir dil olan ve Hint-Avrupa dil ailesinin Slav kolunun batı grubunda Çekçe ile aynı yerde sınıflandırılan Lehçe hedef dil olarak seçilmiştir. Bir sonraki adımda ise Çekçe ile diğerlerine göre daha az benzerlik gösteren ancak yine Hint-Avrupa dil ailesinin Baltık kolunda yer alan Litvanyaca hedef dil olarak kullanılmıştır [5]. Çalışma sonucunda Çekçe-Lehçe başarımı %71, Çekçe-Litvanyaca başarımı ise %69 olarak belirtilmiştir. Litvanyaca dilinin farklı yapısından dolayı, çeviri sistemine çok kapsamlı olmayan bir Çekçe sözdizimsel ayrıştırıcı eklenmiştir. Yapılan çalışmaların başarımlarını ölçütü, TRADOS isimli, bilgisayar destekli çeviri için geliştirilmiş bir yazılımın eşleştirme bileşeni ile ölçülmüştür. Çevirmenler, sistem çıktılarını düzelterek ya da olduğu gibi kullanarak referans çeviriler üretmiş, üretilen bu referans ile sistem çıktısı çevirilerin benzerliği, TRADOS yazılımının eşleştirme bileşeni yardımı ile hesaplanmıştır. ČESILKO sistemi çatısında yapılan bir başka çalışma da çok az kişinin (20.000’den az) kullandığı Sırpçanın bir türü (Lower Serbian) üzerinde gerçekleştirilmiş ve aynı ölçüm yöntemi ile %93 başarı elde edildiği ileri sürülmüştür [6].

Benzer diller arasında gerçekleşen bir diğer çalışma ise Hint-Avrupa dil ailesinin Romans kolunun İber-Romans alt grubunda beraber sınıflandırılan İspanyolca ile Katalanca dilleri arasında yapılmıştır [7]. Söz konusu sistemde sekiz temel bileşen bulunmaktadır:

1. Biçim Ayıklayıcı: RTF ve HTML etiketlerini ayıklar
2. Biçimbilimsel Çözümleyici
3. Sözcük Türü Etiketleyici

4. Çeviri Sözlüğü
5. Örüntü İşleme
6. Biçimbilimsel Üretici
7. Son İşleme
8. Biçim Yapıştırıcı : RTF ve HTML etiketlerini geri yapıştırır.

Bu modüllerden temel işlevlere sahip olan biçimbilimsel çözümleyici, çeviri sözlüğü, biçimbilimsel üretici ve son işleme modülleri sonlu durumlu makineler kullanılarak gerçekleştirilmiştir. Çeviri, temelde sözcük bazında yapılırken bazı durumlarda sözcük öbeklerinin çevirisi yapılmaktadır. Sözcük türü etiketleyici modülünde, 2-gram ve 3-gramlardan oluşturulmuş HMM kullanılmıştır. Sözcüklerin çevrilmesi aşamasında kullanılan çift dilli aktarım sözlüğü, kaynak dildeki her bir sözcük için hedef dilde sadece bir sözcük karşılık düşürmektedir. Bunun nedeni olarak da, farklı anlamlara sahip eşadlı (homonym) sözcüklerin her iki dilde de aynı şekilde ifade edilmesi gösterilmiştir.

Bu projenin sonuçlarında İspanyolcadan Katalancaya çeviride (üretilen metnin kabul edilebilir olması için gereken sözcük ekleme, silme ya da değiştirme adetleri bazında ölçülen) hata oranı %5 iken ters yönde (Katalanca-İspanyolca) çevirilerde daha kötü bir başarı sağlandığı belirtilmiştir.

Yakın tarihte, bu proje açık kaynak kodlu hale getirilerek Apertium ismi ile genel kullanıma açılmıştır [8]. Bu sistem ile İspanyada çok az kişinin konuştuğu bazı dillerle (Aranese Occitian gibi) İspanyolca arasında bilgisayarlı çeviri çalışmaları gerçekleştirilmiştir [9].

Devamlılık sağlayan ve pratik anlamda kullanılan bu iki temel çalışma dışında, benzer ya da akraba diller arasında bilgisayarlı çeviri için birkaç çalışma daha bulunmaktadır. Bunlardan bir tanesi, gene İspanyolca-Katalanca arasında gerçekleşen, istatistik tabanlı bir bilgisayarlı çeviri amacıyla karşılıklı tümcelerdeki sözcük öbeklerinin hizalanması (alignment) için dillerin benzerliğinin kullanıldığı bir çalışmadır [10]. Benzer diller dikkate alındığında, eğitim derlemindeki karşılıklı tümcelerde, hem kaynak hem de hedef dildeki sözcük öbeklerinin aynı sıralarda bulunduğu ortaya konmuş ve istatistiksel çeviri modeli buna uygun olarak yeniden şekillendirilmiştir. Klasik anlamdaki bir istatistiksel çeviri modeli (IBM-2) ile %22.3 hata yapılırken bu yöntemle %12.4 hata yapıldığı sonucuna ulaşılmıştır.

Türk dilleri arasında bilgisayarlı çeviri ile ilgili bilinen ilk çalışma Azeri ile Türkçe arasında bir çeviri sisteminin gerçekleştirildiği çalışmadır [11]. Bu tez çalışmasında, Türkçeye en çok benzeyen dil olan Azerice hedef dil olarak seçilerek sözlük tabanlı bir yöntem izlenmiştir. Önerilen yöntem, sözdizimsel bir değişikliğe gitmeden sözcük bazında çeviri yapılmasını kapsamaktadır. Türkçe sözcüğün biçimbilimsel çözümlemesi yapıldıktan sonra kök sözcüğün ve ilgili eklerin Azerice karşılıkları bulunarak Azerice sözcük oluşturulur. Biçimbilimsel çözümlemede ve birden fazla karşılığı olan kök sözcüklerin çevrilmesinde ortaya çıkan belirsizliğin çözülmesi için çeviri sözlüğüne anlamsal bilgiler (örneğin isim türlü sözcüklere soyut, somut, insan, hayvan, sayılabilir vb. gibi ek bilgiler) eklenmiştir. Ek olarak, sözcüklere kavramsal özellikler iliştilmiş, bu sayede cümle içerisinde etkin olan kavramla ilişkili karşılığın seçilmesi sağlanmıştır. Belirsizliklerin giderilmesinde bir başka yöntem olarak birliktelik (collocation) tablolarının hazırlanarak sık geçen yapıların belirlenmesi yoluna gidilmiştir. Sözkonusu çalışmada önerilen yöntem, ek özellikler içeren sözlükler ve belirsizliklerin giderilmesi için kapsamlı listelerin hazırlanmasını gerektirmektedir. Bütün bu gereksinimler elle hazırlanmak zorunda olduğundan, hedef dil olarak Türkçeye benzerliği Azericeden daha az olan diğer Türk dilleri için bu yöntemi kullanmanın maliyeti çok yüksek olacaktır.

Türk dilleri arasında bilgisayarlı çeviri amaçlı bir başka çalışma ise Kırım Tatarcası ve Türkçe arasındadır [12]. Bu çalışmada 5300 adet sözcük içeren bir çeviri sözlüğü hazırlanmış ve bu sözlük kullanılarak sözcük bazında çeviri yöntemi gerçekleştirilmiştir. Çalışmada sadece Türkçeden Tatarcaya çeviri yapılmış, Tatarcadan Türkçeye çeviri gerçekleştirilmemiştir. Türkçe için Oflazer'in biçimbilimsel çözümleyicisi kullanılmış [13] ve Kırım Tatarcası için biçimbilimsel çözümleyici tasarlanmıştır [14]. Kullanılan biçimbilimsel çözümleyiciler kural tabanlıdır ve sonlu durum makineleri kullanılarak tasarlanmışlardır. Çeviri sözlüğünden, kaynak dilde eklerine ayrılmış sözcük kökünün karşılığı bulunmakta ve bu karşılık ile ilgili eklerin hedef dildeki karşılıkları birleştirilerek sözcük Tatarca olarak yeniden oluşturulmaktadır.

Biçimbilimsel çözümleme sonucu ortaya çıkabilecek çözümlemeler birden fazla olabilir. Altıntaş'ın çalışmasında, üretilen birden fazla sonucun teke indirilmesi (morphological disambiguation) için herhangi bir yöntem kullanılmadığı için her sözcüğün olası her çözümlemesi için giriş tümcesi Tatarcaya çevrilmektedir. Bu da bir giriş tümcesi için birden fazla karşılığın çıkması anlamına gelmektedir. Anılan

alıřmada ikinci bir olumsuz yn ise szck kklerinin aktarılması srecinde yařanmaktadır. Trke szcğn anlamına gre Tatarcada birden fazla farklı szck olabilir ve bu szcklerden hangisinin seileceğinin belirlenmesi gerekmektedir. Ancak alıřmada alıřmada bu belirsizlik de giderilmemiřtir.

1.4 Tezin Blmleri

Bu tezin blmleri řu řekilde dzenlenmiřtir: 2. Blm’de, doėal dil iřleme alanında kullanılan ve tez alıřmasında faydanılan temel teknikler tanıtılmıřtır. 3. Blm, bilgisayarlı eviri ařamalarını, otomatik eviri sistemlerinin karřılařtıkları zorlukları ve kullanılan yaklařımların ayrıntılarını iermektedir. 4. Blm’de, akraba ve bitiřken diller iin nerilen B modelleri tanıtılmıř ve bu modellerin matematiksel ayrıntıları sunulmuřtur. 5. Blm’de, Trk Dil Ailesi ierisinde yer alan dillerin ortak ve farklı ynleri ile bu dillerle ilgili zet bilgiler verilmiřtir. 6. Blm’de, Trk Dilleri arasında bilgisayarlı eviri konusu incelenmiř ve tez alıřmasında kullanılan yntemlerin kuramsal ayrıntıları aıklanmıřtır. 7. Blm’de, kuramsal olarak tanıtılan yntemin, Trkmenceden Trkeye gereklenen bir uygulaması ve yntem zerinde uygulama ařamasında gerekli grlen deėiřiklikler anlatılmıřtır. 8. Blm’de ise gereklenen uygulamanın sonuları ve bařarım deėerlendirmeleri karřılařtırmalı olarak verilmiřtir. 9. Blm, alıřma ile ilgili son deėerlendirmeleri ve yorumları iermektedir.

2. DOĞAL DİL İŞLEME TEKNİKLERİ

Doğal dili işlemek için değişik bilgi düzeylerinde işlem yapan farklı biçimsel modeller ve kuramlar kullanılmaktadır. Bunlar çoğunlukla bilgisayar bilimleri, matematik ve dilbilimi alanlarında yaygın olarak kullanılan tekniklerdir. Bunlar arasında sonlu durumlu yöntemler (sonlu durumlu tanıyıcılar, sonlu durumlu dönüştürücüler, Markov Modelleri, Hidden Markov Modelleri vb.), biçimsel kural sistemleri (düzenli ifadeler, bağlamdan bağımsız gramer, vb.), arama yöntemleri (derinliğine arama, ilk en iyiyle arama, A^*) ve olasılık kuramına dayanan makine öğrenmesi yöntemleri sayılabilir. Doğal dil işlemede kullanılan teknikler, çok geniş bir konu yelpazesine yayılmasına karşın bu bölümde sadece yapılan çalışma ile ilintili yöntemler tanıtılacaktır.

2.1 Sonlu Durumlu Makineler

Sonlu sayıda durum, bu durumlar arası geçişler ve bu geçişler sırasında yapılan işlemlerle modellenen davranış şeklini gerçekleyen yapıya “sonlu durumlu makine” adı verilmektedir. Sonlu durumlu makineleri tanımlamak için aşağıdaki kavramlar kullanılmaktadır:

1. $Q : \{q_1, q_2, \dots, q_n\}$ gibi sonlu sayıda durum içeren durumlar kümesi
2. Σ : sonlu sayıda elemandan oluşan giriş abecesi
3. q_0 : başlangıç durumu
4. $\delta(q, a)$: durum geçiş fonksiyonu. Sistem $q \in Q$ durumunda iken $a \in \Sigma$ girişi gelirse, sistemin geçeceği bir sonraki $\bar{q} \in Q$ durumunu döndüren fonksiyondur. Bu yüzden $Q \times \Sigma \rightarrow Q$ üzerinde bir bağıntı tanımlar.
5. $F : \{f_1, f_2, \dots, f_m\}$ gibi sonlu sayıda kabul durumunu içeren küme. $F \subseteq Q$

Bu kavramlar ışığında sonlu durumlu bir M makinesi aşağıdaki gibi tanımlanır:

$$M=(Q,\Sigma,q_0,\delta,F) \quad (2.1)$$

2.1.1 Düzenli İfadeler ve Düzenli Diller

Düzenli ifadeler, Σ abecesi üzerinde tümevarımla tanımlanır:

- Boş katar ε ve Σ abecesinin elemanları a ($\forall a \in \Sigma$) bir düzenli ifadedir.
- A ve B düzenli ifadeler olmak koşulu ile $A \bullet B$ (bitiştirme), $A \mid B$ (veya) ve $A^* = \varepsilon \mid A \mid A^2 \mid A^3 \mid \dots$ (Kleene yıldızı) işlemleri sonucu oluşan anlatımlar da birer düzenli anlatımdır.

Belirli bir Σ abecesindeki harflerin oluşturdukları harf dizilerini içeren kümeye **dil** (language) adı verilir. Dil, boş küme olabileceği gibi sonlu ya da sonsuz sayıda elemana da sahip olabilir.

Belirli bir Σ abecesi üzerinde tanımlı **düzenli dillerin** (regular language) özyinelemeli tanımı aşağıdaki gibidir:

- Hiçbir eleman içermeyen boş dil $\emptyset$, bir düzenli dildir.
- Boş katar $\{\varepsilon\}$ ve Σ abecesinin elemanları $\{a\}$ ($\forall a \in \Sigma$) bir düzenli dildir.
- A ve B düzenli diller olmak koşulu ile $A \cup B$ (birleşim), $A \bullet B$ (bitiştirme) ve A^* işlemleri sonucu oluşan diller de birer düzenli dildir.
- Σ üzerinde tanımlı diğer diller içerisinde bu şekilde üretilmeyen dillerin hiçbirisi düzenli dil olamaz.

Düzenli ifadeler ile ona ilişkin diller arasındaki ilişki aşağıdaki gibidir:

- $L(\varepsilon) = \{\varepsilon\}$; $L(a) = \{a\}$; ($\forall a \in \Sigma$)
- A ve B düzenli ifadeler ise $L(AB) = L(A)L(B)$
- A ve B düzenli ifadeler ise $L(A \mid B) = L(A) \cup L(B)$
- A bir düzenli ifade ise $L(A^*) = L^*(A)$

Düzenli ifadeler ve onları temsil eden düzenli dil genelde karıştırılmaktadır. Düzenli dil, grameri düzenli olan bir dildir. Ayrıca düzenli dil, bir küme olarak tanımlanır.

Örnek olarak aşağıda çift sayıda **a** harfi içeren harf dizilerini gösteren düzenli ifade ve bu dilin tanımladığı düzenli dil gösterilmiştir:

$$R = [b^*ab^*a]^*b^*$$

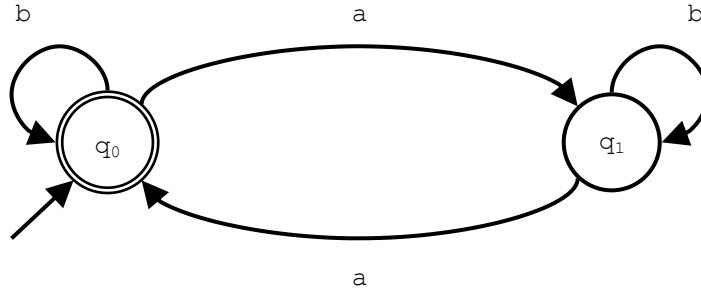
$$L = \{\varepsilon, aa, baa, baba, babab, bbaba, bbabba, babababa, babababbb, \dots\}$$

Görüldüğü gibi basit bir düzenli ifade R ile sonsuz sayıda harf dizisini içeren bir dil L (küme) tanımlanmıştır.

2.1.2 Sonlu Durumlu Tanıyıcılar

Sonlu durumlu bir makine, kendisine giriş olarak verilen bir $w \in \Sigma^*$ harf dizisinin, Σ abecesi üzerinde tanımlanan belirli bir düzenli dilde olup olmadığına karar verebilir (tanıyabilir). Bu tür bir kullanıma sonlu durumlu makinelerin **tanıma** amaçlı kullanımı adı verilir.

Tanıma süreci şu şekilde işler: başlangıç q_0 durumundan itibaren, girişteki diziden bir harf alınır ve δ fonksiyonu uyarınca bir sonraki duruma geçilir. Giriş dizisindeki harfler bitene kadar bu işlem tekrarlanır. Son harf de tüketildikten sonra M makinesinin bulunduğu durum, F yani kabul kümesi durumlarından bir tanesi ise giriş olarak verilen harf dizisi kabul edilmiş olur.



Şekil 2.1 : Örnek Bir Sonlu Durumlu Tanıyıcı

Şekil 2.1’de görülen örnek sonlu durumlu tanıyıcı aşağıdaki biçimde tanımlanmıştır:

$$Q = \{q_0, q_1\}$$

$$\Sigma = \{a, b\}$$

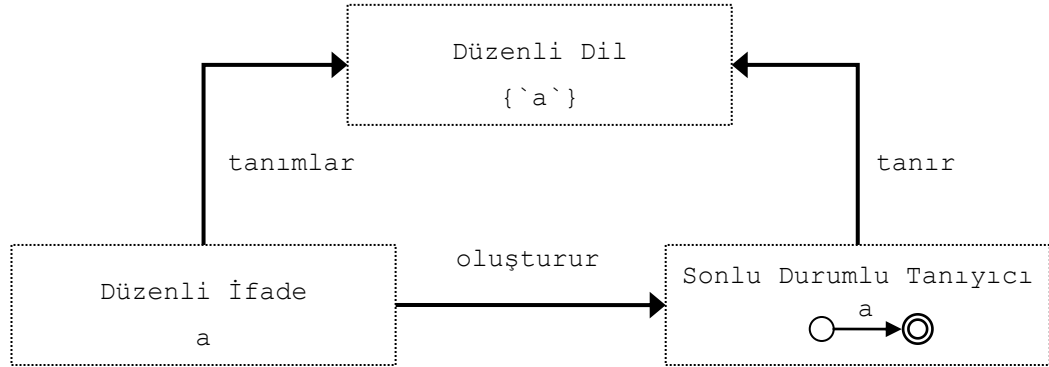
$$F = \{q_0\}$$

$$\delta = \{((q_0, a), q_1), ((q_0, b), q_0), ((q_1, a), q_0), ((q_1, b), q_1)\}$$

Çift daire içerisine alınan durumlar genellikle kabul durumlarını göstermektedir. Bu tanımlamalara göre bu makine $w=b$, $w=bbb$, $w=baba$, $w=abab$... gibi birçok girişi kabul etmekte iken $w=a$, $w=ab$, $w=abb$, $w=bababa$, ... gibi girişleri tanımamaktadır. Sezgisel olarak bu makinenin, “içerisinde çift sayıda a geçen girişleri tanıdığını, tek sayıda a geçen girişleri ise red ettiğini” söyleyebiliriz.

Sonlu durumlu tanıyıcıların en önemli özelliklerinden bir tanesi sonlu sayıda eleman ile (durum, abece ve geçiş fonksiyonu) sonsuz sayıda girişin işlenebilmesidir, yani giriş dizisinin üzerinde herhangi bir sınırlama bulunmamaktadır.

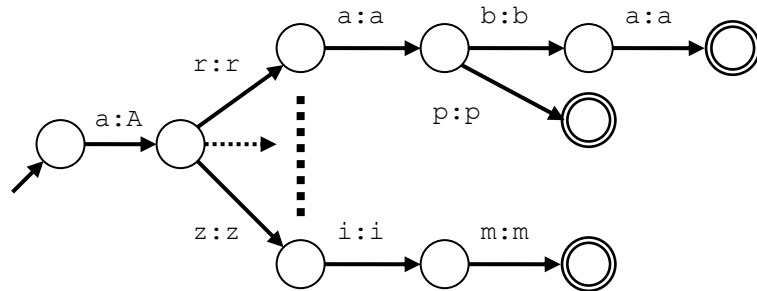
Sonlu durumlu makinelerin kabul ettiği dile $L \subseteq \Sigma^*$ düzenli dil adı verilmektedir. Düzenli ifadeler ile tanımlanan düzenli diller, sonlu durumlu tanıyıcılarla tanınırlar.



Şekil 2.2 : Düzenli İfade - Düzenli Dil - Sonlu Durumlu Tanıyıcı Arasındaki İlişki

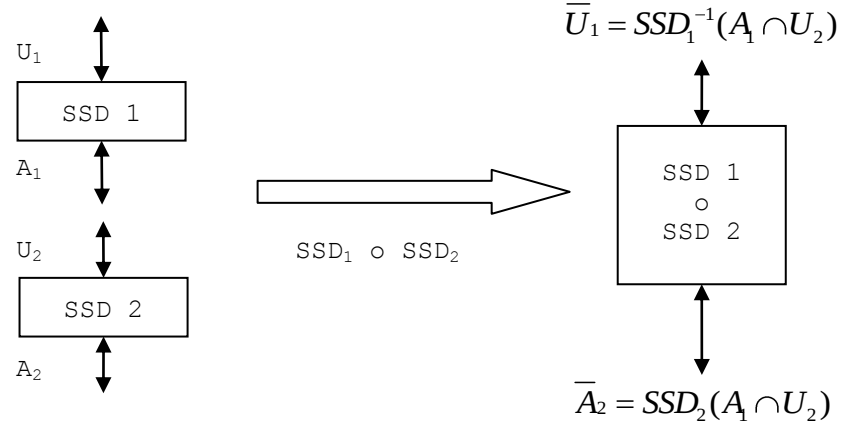
2.1.3 Sonlu Durumlu Dönüştürücüler

Sonlu durumlu dönüştürücülerin (SDD) sonlu durumlu makinelerden tek farkı, girişlere göre durum geçişi yaparken çıkış üretmeleri olarak gösterilebilir. Düzenli bağıntılar, $a:u$ sembol çiftlerinden (a :alt, u :üst) oluşan bir abece üzerinde bir düzenli ifade ile tanımlanabilir. Bu semboller boş katar olabilir. Örneğin a harfi ile başlayan sözcüklerin baş harfini büyük A yapan bir SDD aşağıda verilmiştir:



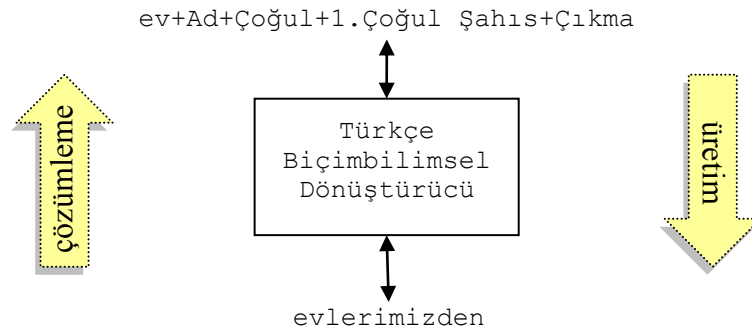
Şekil 2.3 : Sonlu Durumlu Dönüştürücü Örneği

Sonlu durumlu dönüştürücüler, birleştirilip kullanılabilirler. Bu durumda birisinin çıkışı, diğerinin girişi olmaktadır:



Şekil 2.4 : SSD'lerin Birleştirilmesi

Şekil 2.5'te SSD'nin doğal dil işlemede örnek kullanımlarından bir tanesi verilmiştir. Şekildeki dönüştürücü aşağıdan yukarıya doğru biçimbilimsel çözümleyici, yukarıdan aşağıya doğru ise biçimbilimsel üretici olarak çalışmaktadır.



Şekil 2.5 : Türkçe Biçimbilimsel Dönüştürücü

2.2 Biçimbilim

Sözcükleri, sözcük türlerini, kökler ve eklerle ilgili her konuyu, yani dilin yapı özelliklerini inceleyen dilbilgisi dalına **biçimbilim** (morphology) adı verilir. Doğal dil işleme alanında her türlü işlemin ön koşulu, sözcüklerin incelenmesi ve bu sözcüklerden gerekli bilgilerin çıkartılmasıdır. Teknik anlamda biçimbilim çözümlemesi de sözcüklerin bilgi taşıyan alt yapılarına ayrılması sürecini tanımlamaktadır. Bu alt yapılara **biçimbirim** (morpheme) adı verilir. Biçimbirimler sözcük kökü olabileceği gibi teklik/çokluk bilgisi, durum bilgisi ya da kip bilgisini taşıyan ekler de olabilir. Kök sözcükler gibi tek başına kullanılabilen biçimbirlere **serbest biçimbirim** (free morpheme), yalnızca serbest biçimbirlere eklenerek

kullanılabilen biçimbirimlere ise **ek biçimbirimi** (bound morpheme) denir. Diller biçimbirimlerine göre dörde ayrılır:

1. Yalıtımlı Diller (Isolated Languages)

Sözcüklere ek biçimbirimlerinin eklenemediği dillerdir.

Örnekler: Mandarin Çincesi

2. Bitişken Diller (Agglutinative Languages)

Bu tür dillerde, serbest biçimbirimlerden sonra bir ya da birkaç ek biçimbirimi bitleştirilerek yeni sözcükler oluşturulur.

Örnekler: Türkçe, Macarca, Fince

3. Çekimli Diller (Inflectional Languages)

Bir ek biçimbiriminin birden çok bilgiyi içerdiği dillerdir.

Örnekler: Latince ve Çekçe

4. Çokbireşimli Diller (Polysynthetic Languages)

Yalıtımlı dillerin tam tersine, bir sözcük içerisinde birden fazla sözcüğün (başka dillerde ayrı yazılan eylem, belirteç gibi birçok sözcüğün) bulunduğu dillerdir.

Örnekler : Eskimo Dili, Sibirya’da ve Kafkasya’da konuşulan bazı diller bu sınıftandır.

Biçimbirimlerin sınıflandırılması ise üç türlüdür:

1. **Türetme**

Biçimbirimlerin eklenmesi yolu ile ilk sözcükten farklı bir sözcüğün (genellikle farklı bir sözcük türünden) elde edilmesidir. Örneğin Türkçede “**yasak**” isim türlü sözcüğünden “**+lı**” biçimbirimi ile “**yasaklı**” sıfatı türetilmektedir. Sözcük türünün değişmediği bir örnek olarak da “**kapi**” isminden türetilen “**kapıcı**” ve ikinci kez türetilen “**kapıcılık**” örnekleri verilebilir. Burada sözcük türü değişikliği değil sadece anlamsal bir değişiklik gözlenir.

2. **Çekim**

Sözcük türünü değiştirmeyen, sadece tekillik/çokluk, kişi, iyelik, durum, kip gibi ek bilgiler kazandıran biçimbirimlerin eklenmesidir. Örnek olarak Türkçede

“*adam*” sözcüğünün çekimlenerek çoğul ve çıkma durumu belirten “*adamlarda*” sözcüğünün üretilmesi bu tür bir çekim olayıdır.

3. Birleşik Sözcükler

Özellikle serbest biçimbirimlerin iki veya daha fazlasının birleştirilerek yeni sözcüklerin üretilmesidir. Örneğin Almanca, Hollandaca ve Danimarkaca gibi dillerde aşırı derecede birleşik sözcük kullanımı görülür. Buna örnek olarak Almanca’daki “*lebensversicherungsgesellschaftangesteller*” (sigorta şirketi çalışanı) sözcüğü gösterilebilir. Bu sözcük “*versicherung*” (sigorta), “*gesellschaft*” (şirket), “*angesteller*” (çalışan) sözcüklerinin birleştirilmesinden oluşturulmuştur.

Biçimbilimsel çözümlemenin bilgisayar yardımı ile yapılmasına ilişkin çalışmalar, Koskeniemi tarafından ortaya konulan “iki düzeyli biçimbilimsel çözümleme” isimli çalışmadan [15] sonra önemli ölçüde hızlanmış ve bu amaçla çeşitli araçlar hazırlanmıştır.

2.2.1 İki Düzeyli Biçimbilimsel Çözümleme

İki düzeyli biçimbilim (two-level morphology) yöntemi, biçimbilimsel çözümlemede en sık kullanılan yöntemlerden bir tanesidir [15-18]. Bu yöntem ile Türkçe [13], Türkçe gibi bitişken bir dil olan Fince ve daha farklı biçimbilimsel özelliklere sahip İngilizce, Japonca, Fransızca, Rumence gibi birçok dil için biçimbilimsel çözümleyiciler geliştirilmiştir [19-22].

Adından da anlaşılacağı gibi bu yöntem, bir sözcüğün **yüzeysel biçim** (surface representation) ve **yapısal biçim** (lexical representation) şeklinde iki türlü gösterimine dayanmaktadır. Yüzeysel biçim, sözcüğün çekim ve/veya türetme ekleri alındıktan sonra metin içerisinde gözlemlenen yazımsal biçimidir. Yapısal biçim ise, sözcüğün hangi biçimbilimsel yapılardan (sözcüğe eklenmiş çekim/yapım eklerinin özellikleri ve türleri, eklenme sıraları vb.) oluştuğunun belirtildiği biçimdir.

Yapısal Biçim: ev + AD + ÇOĞUL + 1.TEKİL ŞAHİS İYELİK + BULUNMA HALİ

Yüzeysel Biçim: evlerimde

Yapısal gösterimde belirtilen bu biçimbilimsel özelliklerin ekleri, çeşitli ses ve harf değişikliklerine uğrayarak sözcüğe eklenmekte ve sonuçta sözcüğün yazımda kullanılan *yüzeysel biçimi* ortaya çıkmaktadır.

İki düzeyli biçimbilim yöntemi, iki bileşenden oluşur:

1. Yazım Kuralları

Bu bileşen ile yapısal biçimden yüzeysel biçime geçerken, yani sözcüğün yazıdaki hali oluşurken meydana gelen harf değişiklikleri ve ses olayları modellenir.

2. Bitiştirme Kuralları

Sözlükte yer alan sözcük köklerine hangi eklerin gelebileceği ve geçerli bir sözcüğün oluşması için bu eklerin geliş sıraları bu bileşen yardımı ile modellenir.

Yapısal biçim ile yüzeysel biçim arasında dönüşümler, sonlu durumlu dönüştürücülerle SDD gerçekleştirilmektedir.

2.2.1.1 Yazım Kuralları

Yapısal biçimden yüzeysel biçime geçerken yazım kuralları (ortographic rules) uyarınca harflerde değişimler olur. Bu kuralların modellenmesini sağlayan bileşene “yazım kuralları bileşeni” adı verilir. Örneğin Türkçedeki “*karın*” sözcüğünün sesli ile başlayan bir ek alması durumunda, “*ı*” harfinin düşerek “*karnı*” şeklinde yazılması bu tür bir değişikliktir. Benzer şekilde sesli uyumlarına göre eklerin sesli harflerinin değişmesi (ör. +*ler*/*lar*), sessiz benzeşmesi (*yaz+dım/aç+tım*), kaynaştırma harflerinin oluşması (*iste+y+ecek / döv+ecek*) gibi bir çok ses olayı da bu değişimlere örnek olarak sayılabilir.

Bu ses değişiklikleri, Tablo 2.1’deki 4 temel kural türü ile ifade edilebilmektedir [18].

Bu kurallar kullanılarak, dildeki ses olaylarını modelleyen SDD gerçekleştirilir. Daha sonra bu dönüştürücüler, bitiştirme kurallarını modelleyen SDD ile birleştirilerek kullanılır.

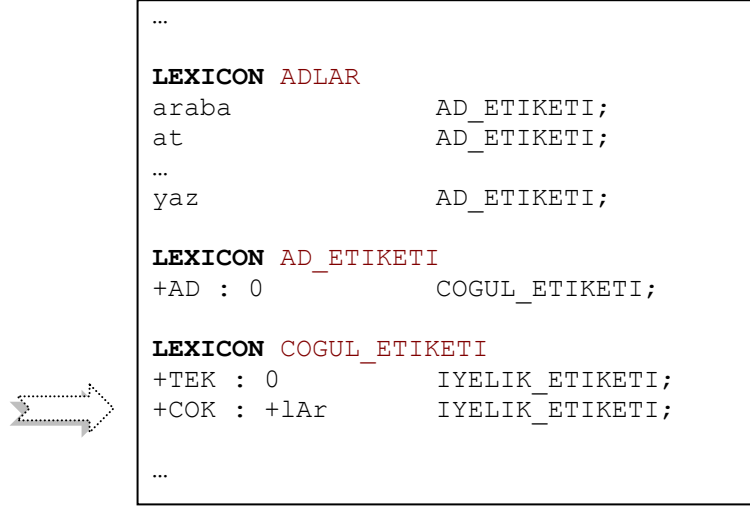
Tablo 2.1 : İki Düzeyli Biçimbilimde Kural Türleri

a:b => LC _ RC	Yapısal gösterimdeki bir 'a' sesi, kendinden önce ve sonra belirtilen bağlamlar varsa (LC - left context ve RC - right context) yüzeysel biçimde 'b' sesine <u>dönüştürülebilir</u> , (ancak bu dönüşüm zorunlu değildir).
a:b <= LC _ RC	Yapısal gösterimdeki bir 'a' sesi, kendinden önce ve sonra belirtilen bağlamlar varsa (LC - left context ve RC - right context) yüzeysel biçimde 'b' sesine <u>mutlaka dönüşür</u> (bu dönüşüm zorunludur, koşullar sağlandığında gerçekleşmelidir).
a:b <=> LC _ RC	Yapısal gösterimdeki bir 'a' sesi, kendinden önce ve sonra belirtilen bağlamlar varsa (LC - left context ve RC - right context) yüzeysel biçimde 'b' sesine <u>mutlaka dönüşür</u> , (bu dönüşüm zorunludur) ve <u>başka hiçbir bağlamda bu dönüşüm olmaz</u> .
a:b /<= LC _ RC	Yapısal gösterimdeki bir 'a' sesi, kendinden önce ve sonra belirtilen bağlamlar varsa (LC - left context ve RC - right context) yüzeysel biçimde 'b' sesine <u>kesinlikle dönüşmez</u> .

2.2.1.2 Bitiştirme Kuralları

Bu bileşen kapsamında köklerden oluşan bir sözcük listesi ve bu sözcüklerin türleri bulunur. Bileşen ayrıca hangi eklerin, hangi sözcüklerden sonra hangi sıralarla gelebileceği bilgilerini de içermektedir. Bu bileşen kök sözcükler, sözcük türleri ve ekleri içeren bir SDD ile gerçekleştirilebilir. Bu SDD, eklerin köklere bitleştirildiği varsayımı ile çalışmaktadır. İki düzeyli biçimbilimsel araçların çoğunda, eklerin bitiştirme modeli, sözlüğün alt parçaları arasında geçiş yapma ilkesine dayanır [21, 23]. Örneğin kök sözcüğe bir ek geldiğinde hangi **alt sözlükten** (lexicon) devam edileceği belirtilir. XEROX araçları için örnek bir sözlük düzeni Şekil 2.6'da gösterilmiştir [24].

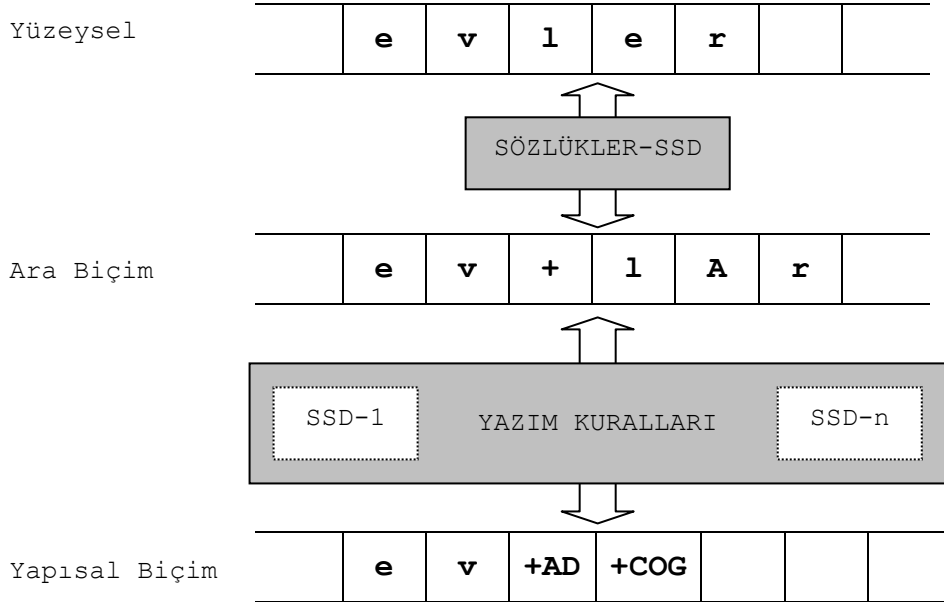
Bu sözlük girdilerinin her bir satırı sırasıyla üst simge, alt simge ve bir sonraki adımda geçilecek alt sözlüğün isminden oluşur. Örneğin işaretli girdi ile, alt simge olarak **+IAR** eki görüldüğü zaman üst simge olarak **+COK** etiketinin oluşturulması ve bir sonraki adımda **IYELIK_ETIKETI** isimli alt sözlükten devam edilmesi gerektiği belirtilir [23].



Şekil 2.6 : "Xerox Lexc" İçin Sözlük Yapısı

2.2.1.3 Sonlu Durum Dönüştürücülerinin Birleştirilmesi

Yazım kurallarını modelleyen SDD ile bitişirme kurallarını modelleyen SDD birleştirilerek tek bir SDD haline getirilir ve iki-düzeyle biçimbilimsel çözümleyici elde edilir. Oluşan SDD, her iki yönde, çözümleme ve üretim amaçlı çalıştırılabilir.



Şekil 2.7 : İki Düzeyli Biçimbilimsel Çözümleyici/Üretici

2.3 İstatistiksel Dil Modelleri

Doğal dil işlemenin konuşma tanıma, el yazısı tanıma, istatistiksel bilgisayarlı çeviri gibi birçok alt kolunda sıklıkla istatistiksel dil modelleri (statistical language model) kullanılmaktadır [25].

İstatistiksel dil modellerinin (İDM ya da DM) temel ilkesi, herhangi bir tümce içerisindeki bir sözcüğün, bu sözcükten önceki sözcüklere bakılarak öngörülmesidir. Tanım olarak istatistiksel dil modelleri, tümceleri oluşturan sözcük dizilerinin olasılık dağılımlarını ifade etmektedir.

$s_1 s_2 s_3 \dots s_K$ sözcük dizisinden oluşan bir C tümcesinin olasılığı, zincir kuralına (chain rule) uygun olarak açılarak aşağıdaki şekilde hesaplanabilir:

$$P(C)=P(s_1)P(s_2|s_1)P(s_3|s_1s_2)...P(s_K|s_1...s_{K-1}) \quad (2.2)$$

Denklem (2.2)'de yer alan olasılıklar, bir eğitim derlemindeki (corpus) sözcük dizilerinin sayılması ile hesaplanmaktadır. Ancak uygulamada, C tümcesinde yer alan tüm sözcüklerin, birebir aynı sırada eğitim derleminde de geçmesi olası değildir. Örneğin tümcenin son sözcüğü s_K 'ya ait $P(s_K|s_1...s_{K-1})$ olasılığının hesaplanması için eğitim derleminde s_K 'nın tüm geçmişinin (history) yani $s_1...s_{K-1}$ bulunması zorunludur. Olasılıkların bu şekilde hesaplanması durumunda çoğu olasılık çarpanı, seyrek veri sorunundan dolayı sıfır olacaktır. Bunu engellemek için, bir tümcenin olasılığı belirlenirken sözcüklerin belirli sayıda (N) geçmişi kullanılarak yaklaşık bir olasılık hesaplanır. Bu bağlamda bir tümcenin olasılığı aşağıdaki şekilde hesaplanır:

$$P(C) \approx P(s_1)P(s_2|s_1)P(s_3|s_1s_2)...P(s_K|s_{K-N+1}...s_{K-1}) \quad (2.3)$$

Örneğin, her sözcüğün kendisinden önce geçen sadece bir sözcüğe bağlı olduğu bir modelde (1. dereceden İDM) olasılık denklemi şu şekilde ifade edilir:

$$P(C)=P(s_1)P(s_2|s_1)P(s_3|s_2)...P(s_K|s_{K-1}) \quad (2.4)$$

N -gram yapıları da İDM ile aynı olasılık dağılımını ifade etmektedir. N -gram modellerinde olasılıklar, sözcüklerin $N-1$ adet geçmişi kullanılarak hesaplanır. Örneğin 3-gram modelinde bir sözcüğün olasılığı, kendisinden önceki 2 sözcük tarafından belirlenir.

N. dereceden İDM, aslında (N-1). dereceden bir Markov modelidir. (N-1). dereceden Markov modeli, herhangi bir s_n sözcüğünün, kendisinden önceki (N-1) adet sözcüğe bağlı olduğu varsayımını kullanmaktadır.

DM'deki olasılıkların çıkarılması ise **En Büyük Olabilirlik Kestirimi** (Maximum Likelihood Estimation) yöntemine göre eğitim derlemindeki sözcüklerin sayılması esasına dayanmaktadır [26] :

$$P(s_i | s_{i-n+1}, \dots, s_{i-1}) \approx \frac{C(s_{i-n+1}, \dots, s_i)}{C(s_{i-n+1}, \dots, s_{i-1})} \quad (2.5)$$

Bu denklemde $C(s_i, \dots, s_j)$ ifadesi, derlemde $s_i, \dots, s_j$ sözcük dizisinin kaç defa geçtiğini göstermektedir.

2.3.1 Yumuşatma

İDM'lerin uygulamasındaki sorunlardan en önemlisi, olasılıkların hesaplandığı eğitim derleminin sınırlı olmasından kaynaklanmaktadır. Olabilecek her türlü sözcük sıralaması, bu derlem içerisinde olmayabilir. Bu durumda bazı olasılıklar sıfır olarak hesaplanacaktır. Aslında genellikle, olasılığı sıfır olan n-gramların sayısının diğerlerinden çok daha fazla olduğu görülmüştür. Eğitim derleminin sınırlı olmasından dolayı olasılığı sıfır olarak hesaplanan ancak gerçekte sıfır olmayan sözcük sıralamaları için bir iyileştirme yapılmalıdır. Bu iyileştirme ise, olasılığı sıfırdan farklı sözcük sıralamalarından bir miktar olasılık dağılımını “çalarak”, olasılığı sıfır olanlara “dağıtmak” şeklinde gerçekleştirilebilir. Bu tür yöntemler, olasılığı sıfırdan farklı olan sözcük sıralamalarında yapılan azaltmaya, indirmeye atfen “indirim” (discounting) olarak adlandırılmaktadır.

Anılan iyileştirme yöntemleri arasında birkaç yöntem bulunur:

- Bir Fazla Sayma Yöntemi [27]
- Witten-Bell İyileştirme Yöntemi [28]
- Good-Turing İyileştirme Yöntemi [29]

Bir başka iyileştirme yöntemi de derece düşürme (back-off) yöntemidir [30]. Bu yöntemde, eğer derlemde bir sözcük sıralamasına hiç rastlanmamışsa, bu sözcük sıralamasının olasılığı hesaplanırken bir merteye daha düşük dil modeline geçilir. Örneğin aranılan 3-gram'a, $P(w_3|w_1w_2)$, ilişkin bir örnek derlemde yoksa, bir derece

daha düşük olan 2-gram modeline geçilir ve burada $P(w_2|w_1)$ olasılığı hesaplanır. Sonuçta aşağıdaki n . dereceden bir sözcük sıralamasının olasılığı, aşağıdaki formül ile hesaplanabilir:

$$\hat{P}(w_i | w_{i-2}w_{i-1}) = \begin{cases} \tilde{P}(w_i | w_{i-2}w_{i-1}) & , eğer C(w_{i-2}w_{i-1}w_i) > 0 \\ \alpha(w_{n-2}^{n-1})\tilde{P}(w_i | w_{i-1}) & , eğer C(w_{i-2}w_{i-1}w_i) = 0 \text{ ve} \\ & C(w_{i-1}w_i) > 0 \\ \alpha(w_{n-1})\tilde{P}(w_i) & , diğer durumlarda \end{cases} \quad (2.6)$$

Bu denklemdeki α katsayısı ve düzeltilmiş olasılıklar $\tilde{P}$, olasılık dağılımı toplamının 1 olması için yeniden hesaplanmış olasılıkları göstermektedir.

Anılan iyileştirme yöntemlerinin ayrıntıları burada ele alınmayacaktır; ancak bu konularda ayrıntılı bilgi için [26, 31] kaynakları incelenebilir. Ek olarak, farklı türdeki iyileştirme yöntemlerinin karşılaştırmalarının ve başarımlarının ayrıntılı biçimde incelenmesi için [32] kaynağına başvurulabilir.

2.3.2 İDM Değerlendirmesi

İDM'lerin değerlendirilmesinde en sık kullanılan ölçütler, **entropi** ve **perplexity** ölçütleridir. Bu iki ölçütün tanımlaması bilişim kuramı (information theory) kapsamında yapılmaktadır.

Bilginin bir ölçüsü olan entropi, öngörü yapılacak veri kümesi V üzerinde değer alan bir X serbest olasılık değişkeni ve buna ait $p(x)$ fonksiyonu ile hesaplanır:

$$H(X) = - \sum_{x \in V} p(x) \log_2 p(x) \quad (2.7)$$

Hesaplanan entropi değerinin biriminin ikili değer (bit) olması için logaritma tabanı olarak genellikle 2 kullanılır. Entropinin sezgisel tanımı, bir bilginin alıcı tarafa en uygun kodlama yöntemi ile kodlanarak aktarılması için gereken en düşük bit sayısıdır.

$s_1, s_2, \dots, s_n$ gibi bir sözcük dizisi üzerinde tanımlanan **entropi oranı**, bir anlamda sözcük başına düşen ortalama entropi değerini belirtir. Bu değeri hesaplamak için sözcük dizileri üzerine tanımlanmış bir serbest değişken tanımlanabilir. Bu sayede belirli bir D dilinde tanımlı bütün n uzunluklu sözcük dizileri üzerinde aşağıdaki entropi hesabı yapılabilir:

$$\begin{aligned}
H(D) &= -\lim_{n \rightarrow \infty} \frac{1}{n} H(s_1 s_2 \dots s_n) \\
&= -\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{s \in D} p(s_1 s_2 \dots s_n) \log_2 p(s_1 s_2 \dots s_n)
\end{aligned} \tag{2.8}$$

Ancak bir dile ait gerçek entropi değerinin hesaplanması için, bu sözcük dizisi uzunluğunun sınırsız olması gerekir. Eğer D dili, belirli koşulları sağlarsa (stationary ergodic), D dilinin entropi değeri aşağıdaki gibi hesaplanabilir:

$$H(D) = \lim_{n \rightarrow \infty} -\frac{1}{n} \log_2 p(s_1 \dots s_n) \tag{2.9}$$

Fakat doğal dil söz konusu olduğunda, sözcük sıralarını üreten gerçek p olasılık dağılımının belirlenmesi mümkün değildir. Bu durumda İDM'leri değerlendirmek için **çapraz entropi** (cross entropy) kullanılır.

Bilinmeyen bir p olasılık dağılımı üzerine tanımlanan m modeline (p olasılığına yaklaşıklık sağlayan bir model) çapraz entropi değeri şu şekilde hesaplanır:

$$H(p, m) = \lim_{n \rightarrow \infty} -\frac{1}{n} \sum_{s \in D} p(s_1 \dots s_n) \log_2 m(s_1 \dots s_n) \tag{2.10}$$

D dili, yine belirli koşulları sağlarsa (stationary ergodic), denklem (2.10) aşağıdaki biçime dönüşür:

$$H(p, m) = \lim_{n \rightarrow \infty} -\frac{1}{n} \log_2 m(s_1 \dots s_n) \tag{2.11}$$

Çapraz entropinin kullanışlı olan özelliği ise $H(p, m)$ değerinin $H(p)$ değeri ile alttan sınırlı olmasıdır:

$$H(p) \leq H(p, m) \tag{2.12}$$

Bu özellik, p olasılık dağılımını modelleyen iki ayrı m_1 ve m_2 modeli arasında karşılaştırma yapılmasını sağlamaktadır. Hangi modelin çapraz entropi değeri daha küçükse, bu model gerçek p dağılımına en yakın modeldir ve daha doğrudur. $H(p) = H(p, m)$ değeri sağlandığında, p dağılımı tam olarak modellenmiş olur. Son olarak çapraz entropi değerinin hiçbir zaman gerçek dağılımın entropisinin altına düşmeyeceğini de belirtmekte yarar vardır.

Preplexity değeri ise $2^{H(p)}$ [25, 33] biçiminde tanımlanır. S sözcük dizisi üzerinde tanımlanmış P modelinin perplexity değeri ise aşağıdaki gibi hesaplanır:

$$Perplexity(S) = 2^{H(S)} = \sqrt[n]{\prod_{i=1}^n \frac{1}{P(s_i | s_1 \dots s_{i-1})}} \quad (2.13)$$

3. BİLGİSAYARLI METİN ÇEVİRİSİ

İnsanlar arasındaki iletişimi kolaylaştırmak adına, bilgisayarları kullanarak diller arası çevirilerin yapılması fikri, bilgisayarın kullanılmaya başlandığı ilk yıllardan bu yana birçok araştırmacının ilgisini çekmiştir. Ancak günümüz teknolojisi ve teknikleri ile bile yetkin bir çeviri sisteminin gerçekleşmesi çok zordur [26]. Yetkin bir bilgisayarlı çeviri sisteminin temelde şu üç özelliği barındırması beklenir:

1. **Otomatiklik** : İnsan müdahalesine gerek kalmadan sonuç üretebilmeli
2. **Kaliteli Çeviri Yapabilme** : Sistemin ürettiği çıktılar anlaşılabilir ve asıllarına uygun olmalı
3. **Geniş Kapsamlılık** : Çeviri sistemi her türlü konuyu içeren genel metinler (makale, haber, hikaye, mektup vs.) üzerinde işlem görebilmeli

Bu üç özellik İngilizcede FAHQT (Fully Automatic - High Quality output - unrestricted Text) olarak geçmektedir.

Diller arasında yetkin bilgisayarlı çeviri gerçeklemenin başlıca zorluğu, toplumlar arası kültür ve yaşayış farklılıklarının dillerine yansımış olmasıdır. Dili, bir anlaşma aracı olarak kullanan insanların hayat görüşleri, toplumsal yargıları, kültürleri, yaşayış şekilleri, yaşadıkları ortamların doğa koşulları gibi birçok farklı etken dillerini etkilemiştir. Bu etkiler nesnelere, kavramlara, eylemlere verilen isimlerin farklılaşmasına yol açtığı gibi tümce kuruluşlarını, vurguları ve anlatım biçimlerini de değiştirmiştir.

Bilgisayarlı çeviri her ne kadar tam anlamıyla başarılmış değilse de bu konuda sayısız çalışma bulunmaktadır. Kuşkusuz böyle bir konunun sahip olduğu ticari potansiyel, üniversite çevrelerinin dışında da bu konuyla ilgilenenlerin sayısını arttırmıştır.

Şimdiye dek yapılan çalışmalar sonucunda, eksik ve hatalı da olsa çeviri yapan uygulamalar gerçekleşmiş ve bu tür uygulamaların çeşitli alanlarda işe yaradığı görülmüştür. Gerçeklenen sistemlerin özellikleri incelendiğinde, yetkin çeviri sisteminin temel özelliklerinden bazılarından vazgeçilmesi yoluyla ancak

gerçeklenebilir sistemler üretildiği görülmektedir. Üretilen sistemler genel olarak üç ana amaçla kullanılmaktadır:

1. Yüzeysel Çeviri

Bazı uygulamalarda, bir metnin yüzeysel bir çevirisi dahi iş görmektedir. Özellikle internet ortamından bilgi toplama (information acquisition) sistemleri, kötü de olsa çeviri sonuçlarını kullanarak farklı dillerde içeriklere erişebilmeyi sağlamaktadır. Bu uygulama türlerinde yetkin sistemin “yüksek kaliteli çıktılar” üretebilme özelliği göz ardı edilmiştir.

2. Bilgisayar Destekli Çeviri (BDÇ)

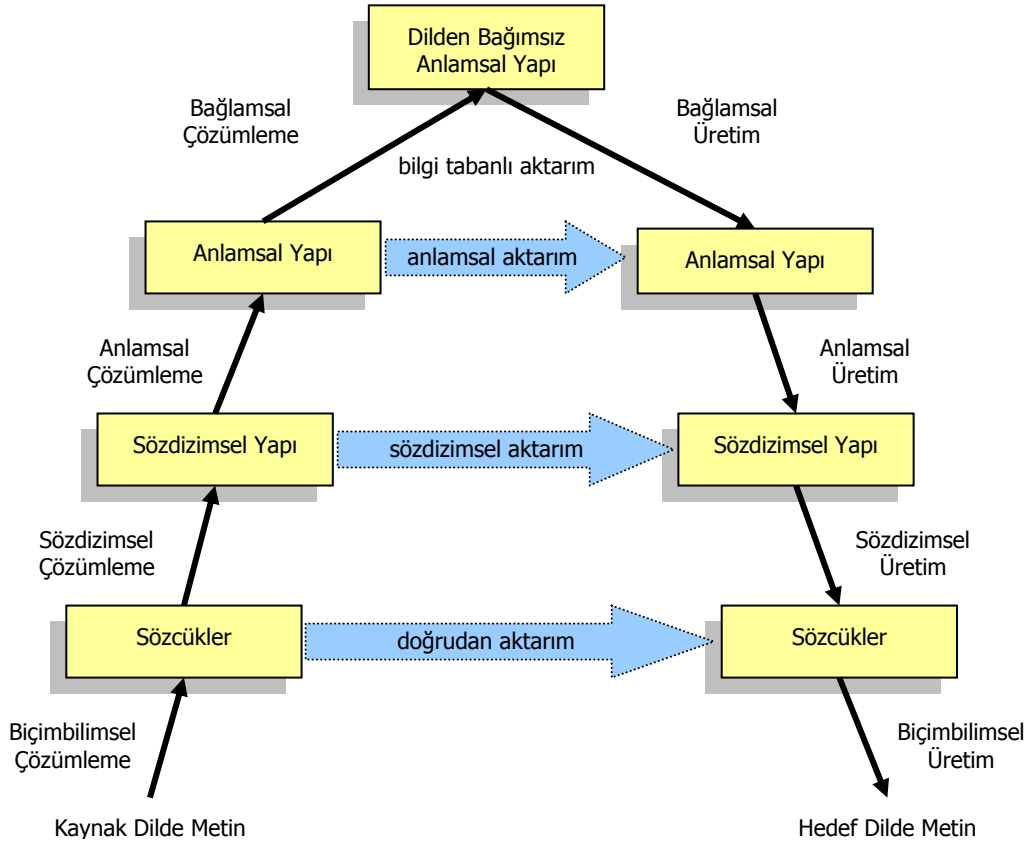
Bilgisayarlı çeviri sistemlerinin bir diğer kullanım alanı ise insan emeği ile yapılan klasik anlamdaki çeviri işlemlerini kolaylaştıran bir araç olarak kullanılmasıdır. Bu tür uygulamalarda bilgisayarlı çeviri sisteminin ürettiği sonuçlar, doğrudan kullanılmak yerine çevirmenler tarafından düzeltilerek kullanılır. Çevirmenlerin yapacakları bu değişiklikler, çoğu kez sıfırdan çeviri yapmaktan çok daha kolay olmaktadır. Yetkin çeviri sisteminin “otomatiklik” özelliğinden vazgeçilerek gerçekleşmiş bu sistemler özellikle yüksek hacimli ve hızlı yapılması gereken çeviri işlerinde tercih edilir.

3. Sadece Belirli Konuları İçeren Metinlerde

Çevrilecek metin türleri ve konuları kısıtlanarak gerçekleştirilen sistemlerde ise yetkin sistemin “geniş kapsamlılık” özelliği kullanılmamış olur. Bu tür sistemlerde konular ve hatta çevrilecek metinlerin dilbilgisi yapılarında bile bazı kısıtlamalara gidilir. Böyle bir sistemin en eski örneği olarak İngilizce-Fransızca arasında hava tahminlerini çeviren Météo sistemi gösterilir [34]. Hava tahmin raporları, kullanılan dil itibarı ile sabit kalıplardan ve hatta sabit sözcüklerden oluşmaktadır. Bu tür bir dil kullanımına **alt dil** (sublanguage) adı verilmektedir.

3.1 Bilgi Tabanlı Çeviri Yöntemleri

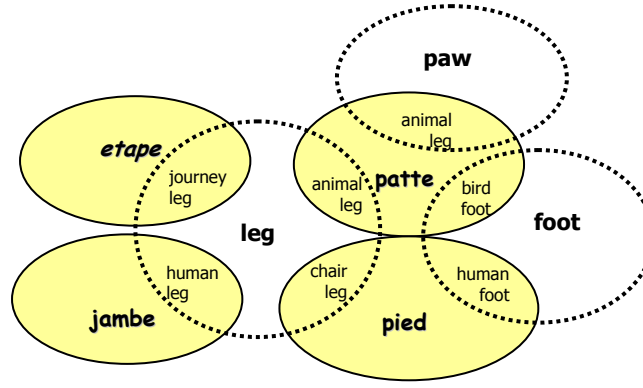
Bilgi tabanlı çeviri yöntemleri, çeviri yapmak için giriş tümcesinin çeşitli bilgi seviyelerinde gösterimlerini oluşturduktan sonra bu bilgi seviyesinde aktarım yapılmasını öngören bir dizi yöntem kullanır. Bu yöntemleri görselleştirmek için Vauquois Üçgeni yaygın olarak kullanılır.



Şekil 3.1 : Bilgi Tabanlı Yöntemlerin Sınıflandırılması-Vauquois Üçgeni

3.1.1 Doğrudan Aktarım

En temel çeviri türü, kaynak dildeki sözcüklerinin karşılıklarının bulunarak hedef dile çevrilmesidir. Ancak bu basit yöntemde dahi birçok sorunla karşılaşmaktadır. Bunların en önemlisi çevrilecek sözcüğün birebir karşılığının bulunmadığı durumlardır. Bu sorunlara **sözlüksel belirsizlik** (lexical ambiguity) denilmektedir. Örneğin aşağıda İngilizce “*leg*” sözcüğünün Fransızca karşılıklarının verildiği bir örnek bulunmaktadır:



Şekil 3.2 : İngilizce-Fransızca Arası Sözcüksel Belirsizlik Örneği

Şekil 3.2’de görüldüğü gibi İngilizcede “**leg**” sözcüğü dört farklı anlamda kullanılabilmektedir; (1) bir yarışmanın, yolculuğun vb.nin ayağı, etap, (2) insan bacağı, (3) hayvan bacağı, (4) sandalye, masa bacağı. Ancak bu 4 farklı anlam için Fransızcada 4 farklı sözcük “**etape**”, “**jambe**”, “**patte**” ve “**pied**” bulunmaktadır. Sözcük bazında çeviri yapan sistemlerin üstesinden gelmesi gereken en büyük zorluk, bu sözcüksel belirsizliğin giderilerek doğru karşılığın seçilmesidir.

Sözcük bazında çeviri yapan doğrudan aktarım yönteminin en önemli bileşeni kullanılan gelişmiş aktarım sözlüğüdür. Aslında bu sözlüğün her girdisi, belirli bir sözcüğü aktarmak için çalışan küçük bir program olarak düşünülebilir. Şekil 3.3’te İngilizce-Rusça arasında doğrudan aktarım ilkesine göre çeviri yapan bir sistemden alınan örnek bir süreç gösterilmiştir. Bu süreç ile İngilizcedeki **much** ve **many** sözcükleri Rusçaya aktarılmaktadır.

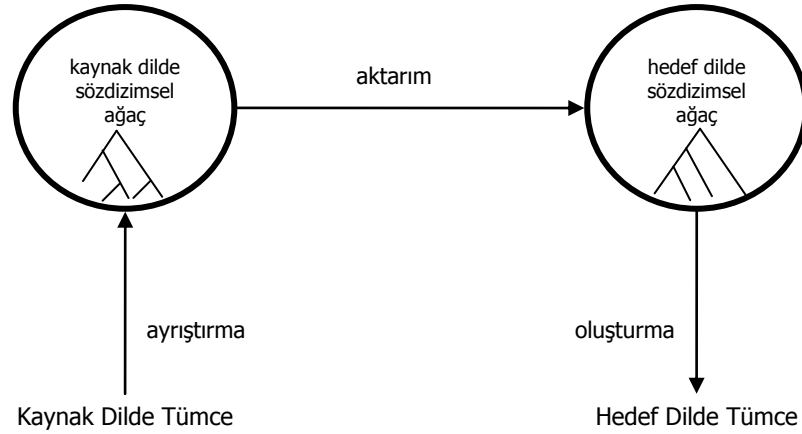
```
function Direct_Translate_Much/Many(word) returns RussianTranslation
if preceding word is "how" return "skol'ko"
else if preceding word is "as" return "skol'ko zhe"
else if word = "much"
    if preceding word is = "very" return nil
    else if following word is a noun return "mnogo"
else /* word = many */
    if preceding word is a preposition and
        following word is a noun return mnogii
    else return mnogo
```

Şekil 3.3 : Doğrudan Aktarım Yöntemleri İçin Örnek Aktarım Süreci

Her ne kadar doğrudan aktarım yönteminde tümce üzerinde çözümleme yapılması gerekmeseyse de birçok uygulamada biçimbilimsel çözümleme yapılır.

3.1.2 Sözdizimsel Gösterimin Aktarımı

Bilgisayarlı çeviri yöntemleri arasında diğer bir yöntem de sözdizimsel temelde çeviri yapmaktır. Buna göre kaynak dildeki sözcük öncelikle sözdizimsel olarak ayrıştırılır ve elde edilen ağaç yapısı, hedef dilde aynı anlamı taşıyan ağaç yapısına çevrilmeye çalışılır.



Şekil 3.4 : Sözdizimsel Gösterimin Aktarımı

Sözdizimsel yapının aktarılmasından sonraki süreç ise sözcüklerin aktarılmasıdır. Tıpkı doğrudan aktarım yönteminde olduğu gibi bu aşamada da her iki dilde sözcükleri içeren bir aktarım sözlüğü kullanılır. Bazı sistemlerde, bu aşamada ortaya çıkan sözcüksel belirsizliklerin giderilmesi için kaynak tümce çözümlemeleri sırasında anlamsal belirsizlik giderici yöntemler uygulanmaktadır.

3.1.3 Anlamsal Gösterimin Aktarımı

Anlamsal çeviri (semantic transfer), önce kaynak dildeki tümcenin sözdizimsel ayrıştırması yapıldıktan sonra ayrıştırılan yapılar anlamsal görevlerin yüklenmesi ve aktarımın bu görevler çerçevesinde yapılması temeline dayanmaktadır. Örnek bir tümcenin anlamsal gösterimi Şekil 3.5’de verilmiştir. Yöntem, sözdizimsel çeviride karşılaşılan yapı uyumsuzluklarının bazıları çözebilmektedir.

3.1.4 Dilden Bağımsız Anlamsal Gösterimin Aktarımı

Bilgisayarlı dil çevirisi yöntemlerinin sonuncusu ise “interlingua” adı verilen ve tümcenin taşıdığı anlamı, dilden bağımsız bir yapıda ifade eden yapıları kullanılmasıdır. Bu yöntemin temel dayanak noktası, farklı dillerde, anlamların ifade edilme biçimlerinden bağımsız bir anlam temsidir.

Mehmet, bu güzel çöreği yemedi.

<i>Olay</i>	<i>yemek</i>	
<i>Etmen</i>	<i>Mehmet</i>	
<i>Kip</i>	<i>geçmiş</i>	
<i>Olumluluk</i>	<i>olumsuz</i>	
<i>Tema</i>	<i>çörek</i>	
	<i>İşaret</i>	<i>bu</i>
	<i>Özellikler</i>	<i>tatlı</i>

Şekil 3.5 : Anlamsal Gösterim Örneği

Bu yöntemin en yararlı yönü, ikiden fazla dil arasında çeviri yapılacağı zaman ($1 \rightarrow N$) ortaya çıkmaktadır. Diğer yöntemlerde bu tür bir işlem, toplam N^2 çeviri yapılması anlamına gelir. Halbuki kaynak tümcenin dilden bağımsız anlamsal gösterimi elde edildikten sonra, bu gösterime ilişkin tümcenin N adet dil için üretilmesi yeterlidir. Bu tür çeviri yöntemi, Avrupa Birliği gibi birçok dilin kullanıldığı ortamlar için kullanışlı olmaktadır.

Ancak “interlingua” adı verilen bu yapının kullanılmasında çok büyük bir sorun bulunmaktadır: doğal dil ile ifade edilen anlamı, başka bir biçimde sunacak olan “bilgi temsil diline” çevirmedeki güçlük. Her dilin, belirttiği özellikler farklılık göstermektedir. Örneğin Türkçede 3. tekil şahıslar için cinsiyet bilgisi yokken İngilizcede bulunur. Benzer şekilde Türkçede *amca* ve *dayı* ayrı sözcüklerle ifade edilirken İngilizcede her ikisi de *uncle* sözcüğü ile belirtilir. Anlamsal gösterimin dilden bağımsız olabilmesi için her dilde farklı ifade edilen kavramları içermek için de bir yol bulunması gerekmektedir. Bu farklılıklardan dolayı dilden bağımsız bir gösterimin tam olarak nasıl sağlanabileceği konusunda halen büyük boşluklar bulunmaktadır.

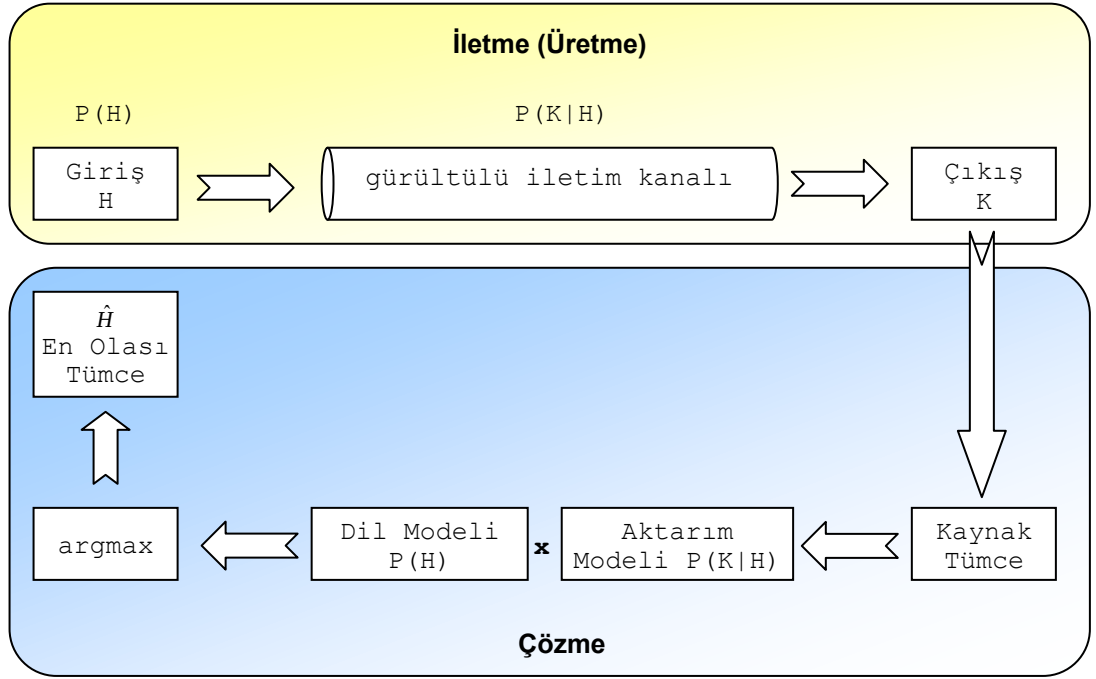
3.2 İstatistiksel Yöntemler

Bilgi tabanlı bilgisayarlı çeviri yöntemlerinin ana teması, kaynak dildeki tümcelerin hangi bilgi seviyesinde (sözcük, sözdizimsel yapı, anlamsal yapı gibi) işlem göreceğini belirlemek ve seçilen gösterimin hedef dile nasıl aktarılacağını yollarını araştırmak üzerine yoğunlaşır. İstatistik tabanlı yöntemler ise tamamen sonuç odaklı

çalışır ve aktarma işleminin nasıl yapılması gerektiğinden çok nasıl sonuçlanması gerektiği üzerinde durur.

İstatistiksel çeviri yöntemlerinin fikirleri 1950’li yıllarda ortaya atılmış olsa da gerçek anlamdaki çalışmalar 1990’lı yıllarda başlamıştır [35, 36]. Elektronik ortama aktarılmış, karşılıklı çevirilerden oluşan metinlerin sayısının giderek çoğalması ve bilgisayarların yeteneklerinin hızla artması, bilgi tabanlı aktarım için kural karmaşasında boğulmakta olan çevrelerin, istatistiksel çeviri yöntemlerine doğru hızlı bir kayma yaşamasına neden olmuştur.

İstatistiksel yöntemlerin çalışma mantığı, çeviri işlemini, Shannon’un Gürültü Kanal Modeli (Noisy Channel Model) uyarınca bozulmuş bir işareti düzeltme olarak değerlendirir:



Şekil 3.6 : Gürültü Kanal Modeli Uyarınca Çeviri İşlemi

Bu yaklaşıma göre hedef dildeki tümce H , iletim kanalından geçerken kanaldaki gürültü nedeniyle değişmiş ve çıkışta kaynak dildeki tümce K oluşmuştur. Yöntemin ilkesi, iletim kanalının çıkış ucunda gözlenen kaynak dildeki tümce K ’dan yola çıkarak, gönderilen asıl metne yani “hedef dildeki” tümceye ulaşmaktır. Bunu sağlamak üzere aşağıdaki denklemin çözümünün bulunması yeterli olacaktır:

$$\hat{H} = \arg \max_{H \in \text{Hedef Dil}} P(H | K) \quad (3.1)$$

Denklem (3.1)'deki olasılık deęerini Bayes kuralına gre tekrar yazarsak:

$$\hat{H} = \arg \max_{H \in \text{Hedef Dil}} \frac{P(K | H)P(H)}{P(K)} \quad (3.2)$$

Bu denklemde $P(K)$ olasılıęı btn H tmceleri iin sabit olduęundan *argmax* iřleci iin sonucu deęiřtirmez. Bu durumda denklem (3.2)'yi ařaęıdaki gibi yazabiliriz:

$$\hat{H} = \arg \max_{H \in \text{Hedef Dil}} \underbrace{P(K | H)}_{\text{eviri modeli}} \underbrace{P(H)}_{\text{dil modeli}} \quad (3.3)$$

Bu denklemde iki bileřen gze arpmaktadır. Bunlardan $P(K|H)$, **eviri modeli** olarak adlandırılır ve hedef dildeki H tmcesinin, kaynak dildeki K tmcesinin evirisi olma olasılıęını belirtir. İkinci bileřen ise H tmcesinin, hedef dildeki olasılıęını belirtir. Bu iki olasılık deęerinin arpımını en oklayan $\hat{H}$ tmcesi sonu olarak retilir.

Denklem (3.3)'in sezgisel aıdan yorumu ise, kaynak dildeki K tmcesinin en yakın evirisi olma (eviri modeli bileřeni) ve aynı zamanda da hedef dil iin akıcı ve geerli bir tmce olma (dil modeli bileřeni) kořullarını birlikte saęlayan en olası $\hat{H}$ tmcesinin bulunmasıdır.

Sonu olarak istatistiksel yntemlerle bilgisayarlı eviri yapabilmek iin ařaęıdaki  bileřenin elde edilmesi gereklidir:

1. $P(H)$ 'nin hesaplanabilmesi iin hedef dil iin bir İDM
2. $P(K|H)$ 'nin hesaplanabilmesi iin bir eviri modeli
3. Btn bu olasılık deęerlerini kullanarak verilen bir K tmcesi iin en olası $\hat{H}$ tmcesini retebilen bir zc (decoder)

Bu bileřenlerden istatistiksel dil modeli bileřeni Blm 2.3'de anlatılmıřtır. Gerekli dil modelleri sadece hedef dil iin retilmek zorunda olduęundan bu dil modellerinin oluřturulması, eviri modelinin retilmesine gre daha kolaydır.

eviri modelinin oluřturulması iin, kaynak tmcedeki szck yada szck beklerinin, hedef dilde hangi szck yada szck beklerinin karřılıęı olduęu (rettięi) bilgisi gereklidir. Bu bilgileri ieren ok sayıda tmce zerinde eřitli tekniklerle gereklenen iřlemler sonucunda eviri modeli hesaplanmaktadır. Bu

amaçla, birbirlerinin çevirisi olan, hedef ve kaynak dildeki büyük miktarda (genellikle milyonlarca tümceden oluşan) metinler önce tümce bazında daha sonra da sözcük/sözcük öbeği bazında paralelleştirilir. Bu işlemlere **tümce hizalama** (sentence alignment), **sözcük hizalama** (word alignment), **sözcük öbeği hizalama** (phrase alignment) adı verilir.

İstatistiksel çevirinin son bileşeni ise çözücüdür. En olası çeviri olan $\hat{H}$ tümcesinin nasıl bulunacağı matematiksel olarak bilinse de, hedef dildeki olası bütün tümceleri üreterek bunlardan denklem (3.3)'e göre en olası $\hat{H}$ tümcesini belirlemek pratik olarak olanaksızdır. Bu amaçla demetli-arama, A^* gibi daha verimli arama yöntemleri kullanılabilir [37].

3.3 Örnek Tabanlı Yöntemler

Örnek tabanlı yöntemler (Example Based Translation), her iki dilde karşılıklı tümceler içeren tümcelerden oluşan bir derlemi kullanarak “örnekseme” yoluyla çeviri ilkesini kullanır [2]. Çalışma ilkesinden dolayı “örneksemeyle çeviri” (translation by analogy) olarak da isimlendirilir. Yöntemin çeviri üzerindeki temel varsayımları şunlardır:

- İnsanlar basit tümceleri derinlemesine dilbilgisi kurallarıyla çevirmezler.
- Aksine, çeviri yaparken insanların ilk adımları, kaynak tümceyi belirli alt öbeklere parçalamaktır. Daha sonra bu öbekleri hedef dile çevirir ve son adımda da bu öbekleri uygun biçimde birleştirerek daha uzun sonuç tümcesini üretir.
- Öbeklerin çevrilmesinde ise daha önceden “akılda kalan” örneklerle örnekseme yapılır.

Örneğin aşağıdaki iki çeviriyi ele alalım:

A man eats vegetables	↔	Hito wa yasai o taberu
Acid eats metal	↔	San wa kinzoku o okasu

Bu iki örnek tümceye benzetilerek aşağıdaki girdi tümcesi çevrilmek istensin:

He eats potatoes

Kuşkusuz çeviri işlemi için bir aktarım sözlüğü gereklidir. Ama buradaki asıl sorun İngilizce *eat* eylemi için olası iki Japonca karşılıktan (*taberu* ve *okasu*) hangisinin kullanılacağına karar vermektir. Yöntem, doğru karar vererek *taberu* eylemini seçer çünkü tümcenin diğer öğeleri *he* ve *potatoes* sözcükleri, örneklerden *man* ve *vegetables* sözcüklerine, *acid* ve *metal* sözcüklerinden anlamsal olarak daha yakındır. Benzer mantıkla aşağıdaki giriş tümcesi için de *okasu* eylemi seçilir:

Sulfuric acid eats iron.

Sözcüklerin anlamsal olarak birbirlerine yakınlık ve uzaklıkları, bir sözlük ve kavramlar dizini (thesaurus) kullanılarak bulunur. Kavramlar dizini, sözcüklerin eş/zıt anlamlılarını (synonym/antonym), alt/üst kavramlarını (upper/lower concept) , parça/bütün (part/whole) ilişkilerini de içeren geniş kapsamlı bir sözlük olarak değerlendirilebilir.

Eğitim derlemindeki tümceler çoklukla birbirinden sadece tek sözcük farklı olacak şekilde seçilir. Bu sayede yöntemin tümcelerinin alt parçalarını daha kolay öğrenmesi sağlanır.

How much is that red umbrella? ↔ Ano akai kasa wa ikura desu ka?

How much is that small camera? ↔ Ano chiisai kamera wa ikura desu ka?

Bu örneklerden aşağıdaki bilgiler çıkartılır:

1. How much is that X? ↔ Ano X wa ikura desu ka?
2. red umbrella ↔ akai kasa
3. small camera ↔ chiisai kamera

Öğrenilen bu bilgiler daha sonraki çevirilerde kullanılmak üzere depolanır. Son gelişmelerle, bu kuralların depolanmadan çalışma anında çıkartılarak kullanılması yoluna gidilmiştir.

Bu yöntem yayınlandıktan sonra Türkçe dahil birçok dilde çalışmalar yapılmıştır [38-40].

3.4 Çeviri Kalitesinin Değerlendirilmesi

Geliştirilen bilgisayarlı çeviri yöntemlerinin ve yöntemler üzerinde yapılan değişikliklerinin sonuçlarının değerlendirilmesi için, üretilen sonuçların yani çevirilerin doğruluğu ve başarısı ölçülmelidir.

Çeviri kalitesinin ölçülmesi için en basit yol, sistem çıktılarının insanlar tarafından çeşitli yönlerden (üretilen tümcenin akıcılığı, kaynak tümcedeki anlamın aktarılmasındaki doğruluk gibi) puanlanmasıdır. Üretilen çıktıları insanlar kullanacağı için en uygun değerlendirme yöntemi aslında bu olmasına rağmen, hem maliyet açısından çok pahalı hem de hız açısından oldukça yavaştır. Ayrıca aktarım sistemlerinin sürekli geliştirildiği ve her yapılan değişikliğin etkilerini görmek için böyle bir değerlendirmeye gereksinim duyulduğu göz önüne alınırsa bu yöntemin yapılabilirliği oldukça azalmaktadır.

Bazı değerlendirme sistemleri ise başarımlar ölçütü olarak, sistem tarafından üretilen çevirinin bir çevirmen tarafından düzeltilmesi sürecinin karmaşıklığını ölçme esasına dayanmaktadır. Bu tür yöntemlerin çıkış fikri, bilgisayarlı çeviri sistemlerinin çıktılarının genel olarak insan emeği ile düzeltilerek kullanıldığını dolayısı ile sistemin başarısının, çevirmenin harcadığı çaba ile ölçülebileceği görüşüdür. Bu tür ölçme yöntemleri, çevirmenin harcadığı çabayı, aday çeviri üzerinde tüm düzeltmeleri yapmak için, tuş takımında kaç defa tuşa basıldığı ya da çevirmen tarafından düzeltilen son sürümü ile aday arasındaki en kısa değişim uzaklığı (minimum edit distance) değeri ile orantılı olarak hesaplar.

Değerlendirmede izlenebilecek bir başka yol da otomatik yöntemlerle değerlendirme yapmaktır. Sonraki bölümlerde bu yöntemler kısaca tanıtılacaktır.

3.4.1 BLEU/NIST

BLEU yöntemi, IBM tarafından 2002 yılında geliştirilmiştir [41]. Değerlendirme mantığı, sistem çıktısı aday tümcelerinin, çevirmenler tarafından elle çevrilmiş k adet referans çeviri ile olan benzerliğinin ölçülmesine dayanır. Benzerliğin ölçülmesi ise, sistem çıktısındaki sözcüklerin (1-gram) ve sözcük dizilerinin (2,3,4,...-gram), referans çevirilerdeki sözcük ve sözcük dizileri ile eşleştirilmesiyle yapılır. Uygulamada dörtten uzun sözcük dizilerinin eşleştirilmesinin gereksiz olduğu görülmüştür.

Çevirinin doğası gereği bir tümcenin, aynı anlamı taşıyan birden fazla çevirisi olabilir. Sözcük ve ifade seçimlerindeki bu serbestlik derecesi, değerlendirme aşamasında birden fazla referans çeviri kullanılarak çözülmeye çalışılmıştır.

Her n-gram mertebesi için, C derlemi içindeki her S aday tümcesi için hesaplanan **değiştirilmiş kesinlik (modified precision)** değeri p_n şu şekilde bulunur:

$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} \text{Adet}_{eşleşen}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} \text{Adet}(ngram)} \quad (3.4)$$

Bu denklemde, çeviri aday tümcesinde yer alan ngram (yani sözcük yada sözcük dizisi), referans çevirilerde birden fazla defa geçse de bir eşleşme olarak sayılır.

BLEU yöntemi ağırlıklı olarak kesinlik (precision) ölçütüne dayanmaktadır. Birden fazla referans çeviri kullanılabildiği için gerigetirim (recall) değerini hesaplamak zordur. Bu nedenle, referans çevirilerden çok daha kısa bir aday çevirinin, yüksek kesinlik değeri sayesinde yüksek BLEU puanları almasını engellemek amacıyla bir **kısalık cezası** (Brevity Penalty) tanımlanmıştır:

$$BP = \begin{cases} 1 & \text{eğer } c > r \\ e^{1-r/c} & \text{eğer } c \leq r \end{cases} \quad (3.5)$$

Bu denklemde c derlemdeki aday çevirilerin tamamının toplam uzunluğunu, r ise etkin (effective) referans uzunluğunu göstermektedir. Etkin referans uzunluğu, referans tümceler derlemi içerisinde, kendi aday tümcesinin uzunluğuna en yakın olan referansların uzunlukları toplamıdır.

Bu tanımlamalardan sonra BLEU puanı aşağıdaki gibi hesaplanır:

$$BLEU = BP \times \exp\left(\sum_{n=1}^N w_n \log p_n\right) \quad (3.6)$$

BLEU değeri temel olarak eşleşen n-gram oranlarının geometrik ortalamasının bulunmasıyla hesaplanır ve 0 ile 1 aralığındadır. BLEU puanının 1 olması, aday çevirilerin, referanslardan en az 1 tanesi ile birebir aynı olduğunu göstermektedir.

Yapılan çalışmada, bir deneme kümesindeki tümcelerın çevirilerinin BLEU puanları ile seçilen hakemlerin çevirilere verdikleri puanlar karşılaştırılmış ve BLEU puanları ile bu kişilerin değerlendirmeleri arasında ilinti olduğu gösterilmiştir [41].

NIST yaklaşımı da temel olarak BLEU ile aynı değerlendirme adımlarını izlemesine karşın n-gram eşleşmelerinin geometrik ortalaması yerine aritmetik ortalamasını kullanır ve hesaplanan p_n değerlerini n-gramların sıklıkları ile ilişkilendirilir (daha az sıklığa sahip eşleşme daha önemlidir) [42].

Son yıllarda yapılan çalışmalar sonucunda BLEU değerlendirme sisteminin bazı olumsuz yanları ortaya çıkartılmıştır. Örneğin yüksek BLEU puanlarının, her zaman çevirilerin daha kaliteli olduğunun bir göstergesi olmadığı, tersine BLEU puanlarında artış elde edilerek üretilen çevirilerin kalitesinin yükseltilemeyebileceği ortaya çıkartılmıştır [43]. Ancak BLEU yönteminin otomatik olması ve insan emeği gerektiren değerlendirmelere oranla çok daha ucuz ve hızlı olması gibi nedenlerden dolayı günümüzde BLEU yöntemi yaygın olarak kullanılmaktadır.

Anılan nedenlerden ötürü BLEU yönteminin farklı (en azından istatistiksel-kural tabanlı gibi farklı aktarım yaklaşımları kullanan) sistemlerin başarılarının karşılaştırılmasında kullanılmaması gerektiği, BLEU'nun daha çok tek bir sistem üzerinde yapılan ardışık değişikliklerin etkilerini değerlendirmede kullanılmasının uygun olacağı görüşü ağırlık kazanmıştır [43].

3.4.2 F Ölçütü

F-ölçütü, bilgi getiriminde (information retrieval) kullanılan kesinlik ve gerigetirim metriklerinin harmonik ortalaması olarak tanımlanabilir [44]. Bu yöntem, aday tümce ile referans tümce arasında, daha uzun sözcük dizisi eşleşmelerini kayıracak biçimde “en uzun eşleşmeyi” (maximum matching) bulmak üzerine yoğunlaşır. Bu eşleşmenin bulunmasından sonra kesinlik ve gerigetirim değerleri, bulunan bu en uzun eşleşme *EUE* sözcük sayısı ile, sırasıyla aday *A* ve referans *R* tümcelerindeki sözcük sayılarına bölünerek bulunur:

$$\text{Precision}(A|R) = \frac{|EUE(A,R)|}{|A|} \quad (3.7)$$

$$\text{Recall}(A|R) = \frac{|EUE(A,R)|}{|R|} \quad (3.8)$$

3.4.3 Meteor

Meteor ölçütü, F-ölçütü'nü birkaç yönden değiştirerek kullanmaktadır [45]. Meteor değerlendirme sisteminde bazı dilbilimsel süreçler¹ değerlendirme aşamasına dahil edilerek doğrudan sözcük eşleşmeleri yerine sözcük köklerinin de eşleşmesine olanak tanınmıştır. Ayrıca Meteor yönteminde, gerigetirim değeri üzerinde ağırlaştırma yapan bir harmonik ortalama kullanılır :

$$F_{ort} = \frac{10PR}{R + 9P} \quad (3.9)$$

Kesinlik ve gerigetirim değerlerinin sözcük eşleşmesine (1-gram) bağlı olmasından dolayı daha uzun eşleşmeler değerlendirmeye alınmamaktadır. Yöntem, bu açığı kapatmak amacıyla bir ceza katsayısı içermektedir. Bu katsayı hesaplanmadan önce, aday ve referans tümcede birbirlerinin karşılığı olan sözcük grupları (chunks) oluşturulur. Bu sözcük gruplarını oluşturulmasında tek kısıt, hem aday hem de referans tümcede birbirlerinin karşılığı olan sözcüklerin sıralarının grup içinde de aynı kalmasıdır. Örneğin “daha sonra beraber eve gittiler” aday çevirisi ile “daha sonra hep beraber eve gittiler” referans çevirisinde iki grup oluşur : (1) “daha sonra” (2) “beraber eve gittiler”. Bu gruptan sonra ceza katsayısı ve sonuç puanı aşağıda formüllere göre hesaplanır:

$$Ceza = 0,5 \times \left(\frac{|gruplar|}{|eşleşen\ sözcükler|} \right)^3 \quad (3.10)$$

$$METEOR = F_{ort} \times (1 - Ceza) \quad (3.11)$$

Meteor yönteminin en önemli olumsuz yönü, birden fazla referans çeviri olduğunda ortaya çıkmaktadır. Her referans çeviri için bir puan hesaplanarak en yüksek puan çıktı olarak kullanılır. Oysa bazı durumlarda çevirinin bir bölümü (örneğin özneyi oluşturan ad öbeği) referanslardan bir tanesinin bir bölümü ile eşleşirken, çevirinin başka bir bölümü (örneğin eylem öbeği) başka bir referans ile eşleşebilir. Yöntemin bir başka bir olumsuz tarafı ise puanlamada kullanılan katsayıların değerlerinin

¹ İngilizce için Porter Stemmer yöntemi ile eklerin atılarak köklerin elde edilmesi süreci eklenmiştir.

belirlenmesidir. Geçerli katsayılar deneme-yanılma yöntemi ile bulunduğundan en uygun katsayılar olup olmadığı şüphelidir.

4. AKRABA ve BİTİŞKEN DİLLER ARASINDA ÇEVİRİ

4.1 Giriş

Akraba diller arasındaki yapısal benzerlikler yardımı ile, bu diller arasında bilgisayarlı çevirinin gerçekleşmesi, farklı dil aileleri arasında çeviri yapmaktan, en azından sezgisel olarak, daha kolay görünmektedir. Tamamen farklı dil ailelerinde sınıflandırılan diller arasında çeviri yapmanın zorluğu, M. Nagao tarafından aşağıdaki sözlerle ifade edilmiştir [2] :

“Avrupa dilleri arasında belirli bir ortak taban bulunur, dolayısıyla bu diller arasında çeviri yaparken, tümcedeki ifadeler yapısal olarak büyük değişikliklere uğramazlar. Oysa İngilizce ve Japonca gibi tamamen farklı iki dil arasında çeviri yapmanın birçok zorluğu vardır. Bazen aynı anlam tamamen farklı yapılarla ifade edilir. Her iki dilde bu yapıların karşılıklarının doğru bir eşleşmesi her zaman bulunmayabilir.”

Bu noktadan hareketle, akraba diller arasında çeviri yapmanın daha kolay olacağı sonucu kendiliğinden doğmaktadır.

Bu tez çalışması kapsamında akraba diller arasında bilgisayarlı metin çevirisi için istatistiksel ve bilgi tabanlı yöntemlerin beraber kullanıldığı karma modeller önerilmiştir. Önerilen bu karma modeller sayesinde hem istatistiksel yöntemlerin en maliyetli koşulu olan hizalanmış eğitim kümesi gereksinimi ortadan kaldırılmış olmakta hem de yüksek başarılar elde eden istatistiksel yöntemlerin sunduğu getirilerden faydanılmış olunmaktadır.

Geliştirilen modeller, temelde bitişken yapıdaki akraba diller için tasarlanmıştır. Çalışma kapsamında önerilen modeller Türk dil ailesindeki dillerin birbirine çevrilmesi amacıyla incelenmiş, seçilen bir dil çifti için uygulama gerçekleştirilerek modellerin başarımı ortaya konulmuştur. Ancak geliştirilen modeller Türk dillerine özgü olmayıp, dilden bağımsız düşünülmüş, akraba ve bitişken olan tüm dil çiftleri için kullanılabilir niteliktedir.

Modellerin çıkış noktası, Bölüm 3.2’de tanıtılan istatistiksel çeviri yöntemlerinin temel denklemidir. Bu denklem aşağıda tekrar verilmiştir :

$$\hat{H} = \arg \max_{H \in \text{Hedef Dil}} \underbrace{P(K | H)}_{\text{çeviri modeli}} \underbrace{P(H)}_{\text{dil modeli}} \quad (4.1)$$

İki bileşenden oluşan bu denklemin dil modeli bileşeni, sadece hedef dil için hazırlandığından, çeviri modeline göre daha kolay oluşturulabilir. Çeviri modelinin oluşturulması için çok sayıda hizalanmış tümce çiftlerine gereksinim duyulmaktadır. Ancak günümüzde dahi her dil çifti için birbirlerinin çevirisi olan hizalanmış tümcelerin yeterli miktarlarda bulunması mümkün olmamaktadır. Amacımız, bu olumsuz koşullarda da istatistiksel çevirinin kullanılabilir hale gelmesi için, istatistiksel çevirideki bu “çeviri modeli olasılık dağılımı” yerine kullanılabilecek bir modelin oluşturulmasıdır.

Kuramsal açıdan denklem (4.1), hedef dildeki bütün tümceler içerisinden, çeviri ve dil modellerine göre en yüksek olasılığa sahip tümcenin bulunması anlamına gelmektedir.

Ancak bir dildeki olası tümcelerin sayısının sonsuz olması nedeni ile uygulamada çeviriyi üreten çözücü, hedef dildeki H tümcesini adım adım (sözcük ya da sözcük öbeği adımları ile) üretmeye başlar [37]. Her adımda, çeviri modeline ve dil modeline göre en yüksek olasılığa sahip seçenek ya da seçeneklerden devam ederek sonunda tüm çeviri tümcesini oluşturur. Bu çalışma düzeninde çeviri modeli, kaynak tümcenin sözcüklerinin ya da sözcük öbeklerinin karşılığı olabilecek tümceleri (olasılıklarına göre) oluştururken, dil modeli bileşeni de oluşan bu tümceler içinden hedef dil için en uygununu bulmaya çalışır. Bir anlamda “çeviri modeli”, hedef dildeki tüm tümcelerde arama yapmak yerine, kaynak tümcenin çevirisi olabilecek tümceleri olasılıklarına göre seçerek aramayı yönlendirmektedir.

Bu tez çalışmasında önerilen yaklaşım, akraba diller örneğin Türk dilleri, arasında çeviri söz konusu olduğunda, olasılık dağılımı esasına göre çalışan “çeviri bileşeninin”, bilgi tabanlı çalışan “aktarım fonksiyonu” ile değiştirilerek istatistiksel dil modeli ile beraber kullanılması yönündedir. Bu yaklaşım sezgisel olarak, dil modeline göre en uygun tümceyi, “çeviri modelinin yönlendirmesi ile hedef dildeki bütün tümceler kümesinde aramak yerine, aktarım fonksiyonu tarafından aktarılan sözcük/sözcük öbekleri ile oluşturulabilecek tüm olası tümceler kümesinde aramak”

olarak yorumlanabilir. Bu sayede, Türk dilleri gibi birbirleri ile benzer akraba diller arasında kullanılacak, bilgi tabanlı yöntemler ve istatistik tabanlı yöntemlerin birleşimi olan karma bir çeviri modeli önerilmiştir.

Önerilen bu modele göre denklem (4.1)'in güncellenmiş hali aşağıdaki gibidir:

$$\hat{H} = \arg \max_{H \in \underbrace{F(K)}_{\text{aktarım fonksiyonu}}} \underbrace{P(H)}_{\text{dil modeli}} \quad (4.2)$$

Önerdiğimiz çeviri modeli de iki bileşenden oluşmaktadır. Aktarım fonksiyonu, K tümcesinin karşılığı olabilecek tüm tümceleri üreten bir fonksiyon (transfer function) olarak görev yaparken, dil modeli ise klasik anlamda kullanılarak üretilen karşılıklar arasından hedef dile göre en yüksek olasılık değerine sahip tümcenin seçilmesini sağlar.

Ancak önerilen bilgisayarlı çeviri yöntemi ile istatistiksel çeviri yöntemi arasında vurgulanması gereken önemli bir farklılık bulunmaktadır. İstatistiksel çeviri sisteminde yer alan çeviri modeli, kaynak dildeki tümcenin karşılığı olabilecek aday tümceler kümesini üretirken aynı zamanda bunlar için birer olasılık değeri de atamaktadır. Bu olasılık değeri, dil modeli olasılığı ile birleştirilerek en yüksek olasılıklı çevirinin belirlenmesinde rol oynar. Oysa bizim önerdiğimiz aktarım modelindeki aktarım fonksiyonu, sadece kaynak tümcenin karşılığı olabilecek tümceler üretmektedir, bunlara herhangi bir olasılık değeri atanmamaktadır. Bu tümcelerden en uygun olanı ise dil modeli bileşeni tarafından en yüksek olasılıklı tümcenin seçilmesi ile belirlenir.

4.2 Aktarım Fonksiyonu Modelleri

Çeviri yapılacak dil çiftinin bitişken olması durumunda aktarım fonksiyonu ve dil modeli bileşeni, özelleştirilerek kullanılmalıdır. Sonraki bölümlerde önerilen aktarım fonksiyonu ve dil modeli türlerinin ayrıntıları ele alınmıştır.

4.2.1 Aktarım Modeli 0 – Temel Model

Akraba diller söz konusu olduğunda, diller arasındaki benzerlikleri kullanarak çeviri sürecini basitleştirmenin en kolay yolu, sözcük bazında çalışan doğrudan aktarım

modelinin kullanılmasıdır. Özellikle sözdizimi açısından büyük farklılıklar göstermeyen akraba diller arasında daha uygun olan bu doğrudan aktarım modeli, bitişken diller için, sözcük kökleri ile birlikte biçimbilimsel yapıların da aktarılmasını sağlayacak biçimde değiştirilmiştir. Sonuçta oluşan temel aktarım modelinin matematiksel açıklaması aşağıda verilmiştir.

K , toplam N adet sözcükten oluşan $(k_1 k_2 \dots k_N)$ kaynak dilde bir tümce olsun.

$$K = k_1 k_2 \dots k_N = k_1^N \quad (4.3)$$

Bitişken diller söz konusu olduğunda, her bir sözcüğün hedef dile aktarılması için öncelikle biçimbilimsel çözümlemesinin yapılması, sözcük kökünün ve diğer biçimbilimsel yapıların bulunması gereklidir. Buna göre biçimbilimsel çözümleme aşaması, girişi kaynak dilde yüzeysel biçimdeki sözcük k_i , çıkışı ise bu sözcüğün olası tüm biçimbilimsel çözümlemelerini içeren bir küme olan $C(k_i)$ çözümleme fonksiyonu ile modellenir.

$$C(k_i) = c_{i1}, c_{i2}, \dots, c_{in_i} \quad (4.4)$$

Burada n_i , k_i sözcüğü için üretilen biçimbilimsel çözümlemelerin toplam sayısıdır ve $n_i \geq 1$ şeklinde alttan sınırlıdır². Üretilen her bir biçimbilimsel çözümleme, kök ve bu köke eklenen değişken sayıda³ biçimbilimsel özelliklerden oluşur:

$$c_{ij} = kok_{ij} + b_{ij1} + \dots + b_{ijk} + \dots + b_{ijm_{ij}} \quad (4.5)$$

Biçimbilimsel özellikler b_{ijk} ve sözcük kökleri kok_{ij} aktarılması $A(c_{ij})$ aktarım fonksiyonu ile sağlanır. Bu aktarım fonksiyonu giriş değeri olarak, biçimbilimsel bir çözümleme c_{ij} 'yi almakta, çıkış olarak ise sözcük kökünün ve biçimbilimsel özelliklerin hedef dile aktarılmış halini üretmektedir:

$$A(c_{ij}) = a_{ij1}, \dots, a_{ijk}, \dots, a_{ijn_{ij}} \quad (4.6)$$

² Biçimbilimsel çözümleyicinin başarısız olarak çıktı üretemediği durumların olmadığı varsayılmıştır. Uygulamada, Türkçe için 1 milyon sözcüklük derlemde bir sözcük için ortalama n_i sayısı 1,75'tir.

³ Uygulamada, Türkçede, çözümleme başına düşen (kök dışında) biçimbilimsel özellik sayısı 4,36'dır.

Sözcük köklerinin çevrilmesinde birden-çoğa ilişki olduğu için bir çözümlemeye karşılık birden fazla çeviri oluşabilmektedir. Dolayısı ile A fonksiyonu çokdeğerli (multi-valued) bir fonksiyon olarak işlev görmektedir. Bu koşullarda üretilen sözcük sayısı $n_{ij} \geq 1$ olacaktır⁴. Kaynak tümcedeki k_i sözcüğünün c_j çözümlemesine karşılık olarak üretilen her bir a_{ijk} çıktısı, çözümleme ile benzer yapıya sahiptir:

$$a_{ijk} = kokh_{ijk} + bh_{ijk1} + bh_{ijk2} + \dots + bh_{ijkm_{ijk}} \quad (4.7)$$

Burada $kokh$ hedef dildeki kökü, bh ise hedef dildeki biçimbilimsel özellikleri göstermektedir.

Bütün bu tanımlamalardan sonra, temel amacımız olan transfer fonksiyonun tanımı yapılabilir. Aslında tanımlanması amaçlanan aktarım fonksiyonu F , bir fonksiyon yerine bir bağıntı olarak tanımlanmalıdır. Hedef dildeki tüm tunceler üzerinde tanımlı olan bu bağıntı, yalnızca “kaynak dildeki sözcüklerin hedef dildeki karşılıklarından oluşan bir dizi sözcüğü içeren” bir alt kümedir:

$$\begin{aligned} F(K) = F(k_1^N) &= \bigcup_{c_{1j} \in C(k_1)} A(c_{1j}) \times \bigcup_{c_{2j} \in C(k_2)} A(c_{2j}) \times \dots \times \bigcup_{c_{Nj} \in C(k_N)} A(c_{Nj}) \\ &= \prod_{i=1}^N \bigcup_{c_{ij} \in C(k_i)} A(c_{ij}) \end{aligned} \quad (4.8)$$

Eğer $f_i(K)$, $F(K)$ bağıntısının i . elemanı olarak tanımlanırsa, geliştirilen aktarım modelimizin amacı, olası bütün çeviriler içerisinde en yüksek olasılıklı $\hat{H}_B$ ‘yi bulmak olarak ifade edilebilir. Buradaki alt indis B , oluşan tuncenin sözcüklerinin yüzeysel biçim yerine yapısal biçimde olduğunu belirtmektedir. En yüksek olasılıklı tuncenin bulunması ise, E eğitim derlemi üzerinde eğitilen bir $L(E)$ dil modeli ile sağlanır:

$$\hat{H}_B = \arg \max_{f_i(K) \in F(K)} p(f_i(K) | L, E) \quad (4.9)$$

⁴ Sözcük kökünün karşılığının aktarım sözlüğünde bulunmadığı durumlarda hiçbir çıktı üretmemek yerine, kaynak dildeki sözcük kökü doğrudan hedef dile kopyalanır. Bu sayede $n_{ij} \geq 1$ koşulu her zaman sağlanmış olur.

Aktarım modelinin son aşaması ise hedef dildeki biçimbilimsel üretici tarafından, dönüştürülen sözcük kökleri ve biçimbilimsel yapılardan yüzeysel biçimlerin elde edilmesidir. Bu üretim aşaması ise bir U fonksiyonu ile temsil edilir:

$$\hat{H} = U(\hat{H}_B) = h_1 h_2 \dots h_M \quad (4.10)$$

Modelin son çıktısı olan $\hat{H}$, kaynak dildeki sözcüklerin, hedef dildeki karşılıklarının sıralandığı tümceyi göstermektedir. Aktarım aşamasında birden çoğa bir yöntem izlendiğinden, oluşan çeviri tümcesinin sözcük sayısı $M \geq N$ dir.

4.2.2 Aktarım Modeli I

Temel modelin en önemli olumsuzluğu, sözcük bazında sadece bire-bir ya da birden-çoğa aktarım yapılmasına izin vermesidir. Temel modelin bu kısıtlaması sonucu, kaynak tümcede birden çok sözcükle ifade edilen yapılar hedef dile doğru aktarılamayacaktır. Bunu gidermek amacı ile temel modele çoktan-çoğa (many-to-many) aktarım yapmak üzere bir takım eklemeler yapılarak Aktarım Modeli I elde edilmiştir. Bu eklemelerle öncelikle çoklu sözcük grupları (ÇSG) belirlenmiş, daha sonra bu gruplar uygun şekilde hedef dile aktarılmıştır.

Bitişken diller için çoklu sözcük gruplarının bulunma süreci, İngilizce, Çince gibi yalıtımlı ya da yalıtımlıya yakın dillerdeki kadar basit değildir. Bunlar ve benzeri dillerde basit bir liste kullanılarak çoklu sözcük grupları belirlenebilirken, Türkçe, Fince, Japonca, Macarca gibi bitişken dillerde çoklu sözcük gruplarının bileşenleri çeşitli biçimbilimsel değişikliklere uğrayabilirler [46]. Bu değişiklikler, ÇSG'lerin, basitçe bir listeden bakılarak belirlenmesini engellemektedir. Sonuç olarak, bitişken dillerde ÇSG'lerin bulunması için tümcedeki sözcüklerin kökleri ve diğer biçimbilimsel özellikleri gibi daha ayrıntılı bilgilerle, düzenli ifadeler ya da sonlu durumlu dönüştürücüler gibi daha karmaşık araçlara gerek duyulur.

Temel model tanıtılırken verilen matematiksel altyapıya bağlı kalınarak ÇSG'lerin işlenmesi ile ortaya konulan yeni modelin matematiksel bağıntısı aşağıdaki gibi kurulmuştur.

ÇSG'leri, bir ya da birden fazla sözcüğe ait biçimbilimsel çözümleme kümeleri arasından belirli bir yönteme ya da kural dizisine göre seçilen elemanlardan oluşan sıralı eşleşmeler (ordered pairs) olarak adlandırılabiliriz. Örneğin aşağıda bir K

tümcesinin ardışıl üç sözcüğü (k_i, k_{i+1}, k_{i+2}) için biçimbilimsel çözümleme sonuçları bulunmaktadır:

$$\begin{aligned} C(k_i) &= c_{i,1}, c_{i,2}, \dots, c_{i,x}, \dots, c_{i,n_i} \\ C(k_{i+1}) &= c_{i+1,1}, c_{i+1,2}, \dots, c_{i+1,y}, \dots, c_{i+1,n_{i+1}} \\ C(k_{i+2}) &= c_{i+2,1}, c_{i+2,2}, \dots, c_{i+2,z}, \dots, c_{i+2,n_{i+2}} \end{aligned} \quad (4.11)$$

Varsayalım ki ÇSG bulucu kurallar, bu üç sözcüğün çözümlemeleri içerisindeki $c_{i,x}$, $c_{i+1,y}$ ve $c_{i+2,z}$ çözümlemelerinin bir ÇSG oluşturduğunu belirlesin. Bu durumda bu üç elemanlı sıralı eşleşmeler ($c_{i,x}, c_{i+1,y}, c_{i+2,z}$), $G(K)$ kümesinin bir elemanı olur. Bu durumda $G(K)$ aşağıdaki kümeler üzerinde tanımlı bir bağıntı olmaktadır:

$$G(K) \subseteq \prod_{i=1}^N C(k_i) \quad (4.12)$$

Tümceyi oluşturan bütün sözcüklerin tüm çözümlemelerinin kartezyen çarpımı içerisinde arama yapılarak, kurallara uyan sıralı çiftler $G(K)$ bağıntısının elemanı olarak belirlenir. Ancak bağıntı bu şekli ile matematiksel olarak yanlıştır. Çünkü kartezyen çarpımı ile oluşturulan kümenin elemanlarının hepsi, mutlak olarak N sözcükten oluşmalıdır. Bu ise şu anlama gelmektedir: sadece N sözcükten oluşmuş ÇSG'leri işlenebilir, yukarıda örnekteki ($c_{i,x}, c_{i+1,y}, c_{i+2,z}$) gibi üç elemanlı bir sıralı eşleşme girdisi $G(K)$ kümesinin elemanı olamaz.

Bunu düzeltmek için $G(K)$ bağıntısının üzerinde tanımlı olduğu kümelere etkisiz bir eleman eklemek yeterli olacaktır:

$$G(K) \subseteq \prod_{i=1}^N C(k_i) \cup \{\varepsilon\} \quad (4.13)$$

ÇSG belirleyici kuralların, ε girdisini boş katar olarak yorumlayacağı ve göz ardı edeceği düşünüldüğünde, ($c_{i,x}, c_{i+1,y}, c_{i+2,z}$) girdisi artık ($\varepsilon, \varepsilon, \dots, c_{i,x}, c_{i+1,y}, c_{i+2,z}, \dots, \varepsilon, \varepsilon$) biçimine dönüşerek $G(K)$ kümesine eklenebilir.

Sözcük sözcük aktarma yapılırken, eğer sıradaki sözcüğün herhangi bir çözümlemesi c_{ij} , $G(K)$ içerisindeki ÇSG'lerin bir parçası ise, bu sözcüğün hiçbir çözümlemesi aktarılmaz. Ancak eğer c_{ij} , bu ÇSG'nin son sözcüğü ise, c_{ij} yerine bu ÇSG

birleştirilerek aktarım fonksiyonuna gönderilir. Buna göre yukarıda verilen $(c_{i,x}, c_{i+1,y}, c_{i+2,z})$ örneğinin aktarılma süreci aşağıdaki gibidir:

$$\begin{aligned}
C(k_i) &= c_{i,1}, c_{i,2}, \dots, c_{i,x}, \dots, c_{i,n_i} \\
E(C(k_i)) &= \emptyset, \emptyset, \dots, \emptyset, \dots, \emptyset \\
C(k_{i+1}) &= c_{i+1,1}, c_{i+1,2}, \dots, c_{i+1,y}, \dots, c_{i+1,n_{i+1}} \\
E(C(k_{i+1})) &= \emptyset, \emptyset, \dots, \emptyset, \dots, \emptyset \\
C(k_{i+2}) &= c_{i+2,1}, c_{i+2,2}, \dots, c_{i+2,z}, \dots, c_{i+2,n_{i+2}} \\
E(C(k_{i+2})) &= \emptyset, \emptyset, \dots, B(c_{i,x}, c_{i+1,y}, c_{i+2,z}), \dots, \emptyset
\end{aligned} \tag{4.14}$$

Bu çalışma düzenini sağlayan E fonksiyonunun tanımı aşağıda verilmiştir:

$$E(c_{ij}) = \begin{cases} c_{ij} & \text{eğer } (\forall j: 1 \leq j \leq n_i)(\forall p: 1 \leq p \leq N) c_{ij} \neq x_p \\ \emptyset & \text{eğer } (\exists j: 1 \leq j \leq n_i)(\forall p: 1 \leq p \leq N) c_{ij} = x_p \wedge x_{p+1} \neq \varepsilon \\ B(X) & \text{eğer } (\exists j: 1 \leq j \leq n_i)(\forall p: 1 \leq p \leq N) c_{ij} = x_p \wedge x_{p+1} = \varepsilon \end{cases} \tag{4.15}$$

Denklem (4.15)'te yer alan $x_p, X \in G(K)$ sıralı eşleşmesinin p . elemanıdır. B fonksiyonu ise X sözcük dizisini, geçerli bir biçime getirmek için uygun şekilde birleştirilerek tek bir kök ve uygun biçimbilimsel özellikleri içeren yapıya dönüştüren bir birleştirme fonksiyonudur.

E fonksiyonun üzerinde biraz açıklama yapmak uygun olacaktır. Fonksiyonun (1). alt tanım aralığında, K tümcesinin i . sözcüğü k_i 'nin j . çözümlemesi c_{ij} 'nin aktarılıp aktarılmayacağına karar verilir. Eğer k_i 'ye ait çözümlemelerden hiçbirisi, $G(K)$ 'da belirlenen ÇSG yapılarının herhangi birisinin bileşeni olarak geçmiyorsa, c_{ij} olduğu gibi çıkış olarak üretilir. (2) ile numaralandırılmış alt tanım aralığı ise, eğer k_i sözcüğünün herhangi bir çözümlemesi, $G(K)$ 'daki herhangi bir ÇSG'nin son bileşen ($x_{p+1} \neq \varepsilon$) dışındaki bir bileşeni ise, k_i 'ye ait bütün çözümlemelerin göz ardı edileceğini söylemektedir. Son tanım aralığı (3) bölgesinde ise, k_i 'ye ait bir çözümleme, $G(K)$ bağıntısındaki herhangi bir ÇSG'nin son sözcüğü ise ($x_{p+1} = \varepsilon$), c_{ij} yerine ÇSG'nin tamamı (X) B fonksiyonu tarafından dönüştürülerek üretilir.

ÇSG'lerin aktarılması için bu tanımlamalar yapıldıktan sonra, kaynak tümcenin olası bütün karşılıklarını üreten transfer fonksiyonun denklemi (4.8), aşağıdaki gibi değiştirilir:

$$\begin{aligned}
F(K) &= F(k_1^N) = \bigcup_{c_{1j} \in C(k_1)} A \ E \ c_{1j} \times \bigcup_{c_{2j} \in C(k_2)} A \ E \ c_{2j} \times \dots \times \bigcup_{c_{Nj} \in C(k_N)} A \ E \ c_{1N_j} \\
&= \prod_{i=1}^N \bigcup_{c_{ij} \in C(k_i)} A \ E \ c_{1j}
\end{aligned} \tag{4.16}$$

Dil modelini kullanarak en yüksek olasılıklı tümcenin seçildiği bundan sonraki denklemlerde herhangi bir değişme olmaz.

4.2.3 Aktarım Modeli II

Geliştirilen modellerdeki bir diğer eksiklik de aktarım kurallarının sözcük bazında işlem görmesidir. Akraba diller arasında çeviri yapılsa bile, sözcükler arası ilişkiler her zaman bulunur. Çeviri modelinin başarısının arttırılabilmesi için bu ilişkiler göz önüne alınarak aktarım yapılmalıdır. Örneğin bazı Türk dilleri arasında çeviri yaparken, ortaçların, niteledikleri isimden bazı biçimbilimsel özellikleri alması gerekmektedir. Bu ve buna benzer durumları başarılı çevirebilmek için tümce genelinde işlem gören bir takım aktarım kuralları tanımlanmalıdır. Tümce genelinde çalışan aktarım kuralları, tümcedeki sözcüklerin biçimbilimsel bilgilerini kullanabileceği gibi bazı kurallar da sözcüklerin yüzeysel biçimlerine⁵ gerek duyabilir.

Ancak mevcut aktarım fonksiyonu A , sadece sözcük kökleri ve sözcüğe ait biçimbilimsel yapıların aktarımı sağlamaktadır. Bunu geliştirmek üzere denklem (4.10) aşağıdaki gibi değiştirilmiştir:

$$\hat{H} = S_Y \ U \ S_B \ \hat{H}_B = h_1 h_2 \dots h_M \tag{4.17}$$

Bu denklemde, hedef dilde oluşturulan tümceler üzerinde işlem gören bir S_B fonksiyonu tanımlanmıştır. Bu fonksiyona, giriş olarak yapısal gösterimde sözcüklerden oluşmuş tümceler kümesi gelir. Fonksiyon, her bir tümce üzerinde, sözcükler arasında tanımlanan kurallara uygun olarak aktarımı gerçekleştirilir.

⁵ Özellikle yazımsal (ortagraphic) kurallar bu sınıfa girmektedir. Örneğin Türkçede sözcüklerden ayrı yazılan ama bir önceki sözcüğün son seslisine göre ünlü uyumuna uyan –de bağlacı veya –mi soru eki gibi ekleri doğru şekilde üretebilmek için sözcüklerin yüzeysel biçimlerine gerek duyulur.

Tümcedeki sözcüklerin yüzeysel biçimlerine gerek duyarak yapılan değişiklikler ise S_Y fonksiyonu modellenmiştir.

4.3 Bitişken Diller İçin İDM Oluşturulması

İngilizce, Almanca gibi dillerden farklı olarak, Türkçe için dil modelleri oluşturulurken sözcüklerin yüzeysel biçimlerinin kullanılması, Türkçenin türetken ve çekimli yapısından dolayı seyrek veri sorununa yol açmaktadır. Bu yüzden eğitim verisi olarak sözcüklerin yüzeysel biçimleri yerine, sözcüklerin köklerinin ve diğer bazı biçimbilimsel özelliklerin kullanılması yoluna gidilmiştir. Yüzeysel biçim yerine, sözcüklere ait biçimbilimsel çözümleme sonuçlarının tamamının kullanılarak bir İDM oluşturulması durumunda, gene seyrek veri sorunu oluşmaktadır. Seyrek veri sorununu azaltmak için, biçimbilimsel çözümlemedeki tüm etiketler yerine bunların gruplanarak kullanılması fikri ortaya çıkmıştır [47]. Türkçedeki her sözcük, kök ve bir veya birden fazla çekim grubundan oluşmaktadır. Çekim grupları birbirlerinden DB (derivation boundary) ile ayrılmaktadır [48]:

$$\text{kök} + \text{ÇG}_1^{DB} + \text{ÇG}_2^{DB} + \dots + \text{ÇG}_n^{DB}$$

Burada ÇG_i , sözcük türü ve çekim özelliklerini de içeren ilgili çekim grubunu ifade etmektedir. Örnek olarak aşağıda bir biçimbilimsel çözümleme sonucu verilmiştir:

yararlanmanın :	yarar+Noun+A3sg+Pnon+Nom	ÇG1
	DB +Verb+Acquire+Pos	ÇG2
	DB +Noun+Inf2+A3sg+Pnon+Gen	ÇG3

Bu örnekte, isim türlü **yarar** sözcüğünün sözcük türü, **+lan** yapım eki ile önce eyleme daha sonra da **+ma** mastar eki ile de tekrar isme dönüşmüştür. Bu dönüşme süreci içerisinde oluşan her sözcük türünün de kendisine ait çekim özellikleri bulunabilir. Türetilmiş bir sözcüğün etkin sözcük türü, son ÇG'nin sözcük türü olarak kullanılır (örneğin etkin sözcük türü "isim"dir).

Tablo 4.1'de, 1 M sözcükten oluşan bir derlem üzerinde gözlenen, kök hariç bütün etiketlerin bulunduğu tam çözümlemelerin ve ÇG'lerin sayıları verilmiştir [47]. Bir köke eklenebilecek ek sayısının sınırsız olmasına karşın, derlem üzerinde yapılan sayıma göre 10.531 farklı tam çözümlemeye rastlanmıştır. Tam çözümlemeler ÇG'lere ayrılarak ÇG'ler sayıldığı zaman ise 2.194 farklı ÇG'ye rastlanmıştır. Bu sonuçlar, ise seyrek veri sorunun bir miktar indirgendiği ortaya koymaktadır.

Tablo 4.1 : Derlemde Gözlenen Tam Çözümleme ve Çekim Grubu Sayıları

	Kuramsal Üst Sınır	Gözlenen Adet
Tam Çözümleme	∞	10.531
Çekim Grubu	9.129	2.194

Sadeleştirme açısından yapılan bir başka genelleme de ÇG’lerden oluşan türetilmiş bir sözcüğün sözdizimsel açıdan bağlantısının, son ÇG’den çıkarak sonraki sözcüklerin ÇG’lerinden herhangi bir tanesine bağlanması şeklinde ifade edilir [48].

Bu bölümde, türetme ve çekim özelliklerine sahip bitişken diller için, seyrek veri sorunundan en az etkilenecek şekilde dilin farklı özelliklerini modelleyen farklı İDM türleri önerilmiştir.

4.3.1 İDM Tip-I – Kök

Bu tip dil modeli, sözcük köklerinin n-gram olasılıklarından oluşmuştur. Gerçeklenen ön bir işlemle, eğitim derlemindeki girdilerin, sözcük kökü dışındaki tüm biçimbilimsel özellikleri silinerek Tip-I İDM için yeni bir eğitim derlemi oluşturulmuştur.

Tablo 4.2 : İDM Tip-I İçin Örnek Eğitim Tümcesi

Sözcük Biçimbilimsel Çözümlemesi	İDM Eğitim Verisi
Müzik müzik +Noun+A3sg+Pnon+Nom	müzik
dünyasının dünya +Noun+A3sg+P3sg+Gen	dünya
tanınmış tanı +Verb^DB+Verb+Pass+Pos+Narr+A3sg	tanı
isimleri isim +Noun+A3pl+P3sg+Nom	isim
yeni yeni +Adj	yeni
yılı yıl +Noun+A3sg+Pnon+Acc	yıl
çalışarak çalış +Verb+Pos^DB+Adverb+ByDoingSo	çalış
karşıladı karşıla +Verb+Pos+Past+A3sg	karşıla

4.3.2 İDM Tip-II – Son Sözcük Türü

Sadece sözcük türlerinden oluşan İDM tipidir. Türetilmiş sözcükler için son çekim grubunun sözcük türü kullanılmıştır. Sözcük türü sayısının az olmasından dolayı yüksek dereceli Tip-II İDM’leri oluşturulabilmektedir.

Tablo 4.3 : İDM Tip-II İçin Örnek Eğitim Tümcesi

Sözcük Biçimbilimsel Çözümlemesi	İDM Eğitim Verisi
Müzik müzik+Noun+A3sg+Pnon+Nom	Noun
dünyasının dünya+Noun+A3sg+P3sg+Gen	Noun
tanınmış tanı+Verb^DB+Verb+Pass+Pos+Narr+A3sg	Verb
isimleri isim+Noun+A3pl+P3sg+Nom	Noun
yeni yeni+Adj	Adj
yılı yıl+Noun+A3sg+Pnon+Acc	Noun
çalışarak çalış+Verb+Pos^DB+Adverb+ByDoingSo	Adverb
karşıladı karşıla+Verb+Pos+Past+A3sg	Verb

Bu tip İDM ile sözdizimsel yapı, farklı bir biçimde modellenmiş olur. Örneğin 4. dereceden Tip-II İDM’ye göre aşağıdaki iki tümcenin olasılıkları karşılaştırılırsa, gerçeğe uygun olarak (4.18)’deki tümcenin olasılığının daha fazla olduğu görülür:

$$P(\text{Büyük ev zor temizlenir.}) \approx P(\text{Adj Noun Adverb Verb}) = 1,383 \times 10^{-4} \quad (4.18)$$

$$P(\text{Temizlendi ev zor büyük.}) \approx P(\text{Verb Noun Adverb Adj}) = 6,909 \times 10^{-7} \quad (4.19)$$

4.3.3 İDM Tip-III – Son Çekim Grubu

Sadece son ÇG’ler kullanılarak oluşturulmuş İDM tipidir. Tip-II İDM, tümce sözdizimi ile ilgili bir modelleme yapmasına karşın aynı sözcük türüne sahip çözümlemeleri ayıklama noktasında yetersiz kalmaktadır. Bu amaçla, Tip-II İDM’de son çekim grubunda bulunan sözcük türüne ek olarak son ÇG’deki tüm biçimbilimsel özellikler kullanılarak Tip-III İDM oluşturulmuştur.

Tablo 4.4 : İDM Tip-III İçin Örnek Eğitim Tümcesi

Sözcük Biçimbilimsel Çözümlemesi	İDM Eğitim Verisi
Müzik müzik+Noun+A3sg+Pnon+Nom	Noun+A3sg+Pnon+Nom
dünyasının dünya+Noun+A3sg+P3sg+Gen	Noun+A3sg+P3sg+Gen
tanınmış tanı+Verb^DB+Verb+Pass+Pos+Narr+A3sg	Verb+Pass+Pos+Narr+A3sg
isimleri isim+Noun+A3pl+P3sg+Nom	Noun+A3pl+P3sg+Nom
yeni yeni+Adj	Adj
yılı yıl+Noun+A3sg+Pnon+Acc	Noun+A3sg+P3sg+Nom
çalışarak çalış+Verb+Pos^DB+Adverb+ByDoingSo	Adverb+ByDoingSo
karşıladı karşıla+Verb+Pos+Past+A3sg	Verb+Pos+Past+A3sg

4.3.4 İDM Tip-IV – Kök + Son Sözcük Türü

Tip-IV İDM, çözümlemede yer alan sözcük kökü ile son ÇG'deki sözcük türü ile oluşturulur. Bu anlamda Tip-I ile Tip-II İDM birleşimi olarak yorumlanabilir.

Tablo 4.5 : İDM Tip-IV İçin Örnek Eğitim Tümcesi

Sözcük Biçimbilimsel Çözümlemesi	İDM Eğitim Verisi
Müzik müzik+Noun+A3sg+Pnon+Nom	müzik+Noun
dünyasının dünya+Noun+A3sg+P3sg+Gen	dünya+Noun
tanınmış tanı+Verb^DB+Verb+Pass+Pos+Narr+A3sg	tanı+Verb
isimleri isim+Noun+A3pl+P3sg+Nom	isim+Noun
yeni yeni+Adj	yeni+Adj
yılı yıl+Noun+A3sg+Pnon+Acc	yıl+Noun
çalışarak çalış+Verb+Pos^DB+Adverb+ByDoingSo	çalış+Adverb
karşıladı karşıla+Verb+Pos+Past+A3sg	karşıla+Verb

4.3.5 İDM Tip-V – Kök + Son Çekim Grubu

Tip-V İDM, çözümlemede yer alan sözcük kökü ile son ÇG ile oluşturulur. Bu anlamda Tip-I ile Tip-III İDM birleşimi olarak yorumlanabilir.

Tablo 4.6 : İDM Tip-V İçin Örnek Eğitim Tümcesi

Sözcük Biçimbilimsel Çözümlemesi	İDM Eğitim Verisi
Müzik müzik+Noun+A3sg+Pnon+Nom	müzik+Noun+A3sg+Pnon+Nom
dünyasının dünya+Noun+A3sg+P3sg+Gen	dünya+Noun+A3sg+P3sg+Gen
tanınmış tanı+Verb^DB+Verb+Pass+Pos+Narr+A3sg	tanı+Verb+Pass+Pos+Narr+A3sg
isimleri isim+Noun+A3pl+P3sg+Nom	isim+Noun+A3pl+P3sg+Nom
yeni yeni+Adj	yeni+Adj
yılı yıl+Noun+A3sg+Pnon+Acc	yıl+Noun+A3sg+P3sg+Nom
çalışarak çalış+Verb+Pos^DB+Adverb+ByDoingSo	çalış+Adverb+ByDoingSo
karşıladı karşıla+Verb+Pos+Past+A3sg	karşıla+Verb+Pos+Past+A3sg

4.3.6 İDM Tip-VI – Tüm Etiketler

Son İDM tipi ise tüm biçimbilimsel özellikleri kapsayan model olarak belirlenmiştir:

Tablo 4.7 : İDM Tip-VI İçin Örnek Eğitim Tümcesi

Sözcük Biçimbilimsel Çözümlemesi	İDM Eğitim Verisi
Müzik müzik+Noun+A3sg+Pnon+Nom	Noun+A3sg+Pnon+Nom
dünyasının dünya+Noun+A3sg+P3sg+Gen	Noun+A3sg+P3sg+Gen
tanınmış tanı+Verb^DB+Verb+Pass+Pos+Narr+A3sg	Verb^DB+Verb+Pass+Pos+Narr+A3sg
isimleri isim+Noun+A3pl+P3sg+Nom	Noun+A3pl+P3sg+Nom
yeni yeni+Adj	Adj
yılı yıl+Noun+A3sg+Pnon+Acc	Noun+A3sg+P3sg+Nom
çalışarak çalış+Verb+Pos^DB+Adverb+ByDoingSo	Verb+Pos^DB+Adverb+ByDoingSo
karşıladı karşıla+Verb+Pos+Past+A3sg	Verb+Pos+Past+A3sg

Sözcüklerin biçimbilimsel çözümlemelerinde yer alan, kök hariç diğer tüm etiketlerden oluşan İDM Tip-VI, biçimbilimsel yapıların en geniş kapsamlı modellendiği İDM tipidir.

4.3.7 Farklı Dil Modeli Tiplerinin Entropi Değerleri

Geliştirilen dil modeli tiplerinin değerlendirilmesi amacı ile yaklaşık 5000 sözcükten oluşan ve ilgili tipin modellediği bilgileri (kök, son sözcük türü gibi) içeren sınama kümeleri hazırlanmıştır. Bu sınama kümesi üzerinde hesaplanan entropi değerleri Tablo 4.8’de verilmiştir.

Bu tablo incelendiğinde, genel olarak $n \geq 3$ için entropi değerinin azalmadığı (hatta bazı durumlarda arttığı) görülmektedir. Bu ise 3. dereceden İDM kullanmanın yeterli olacağı, 4. ve 5. dereceden İDM’leri kullanmanın getirdiği yeni bir bilgi olmadığı sonucunu doğurmaktadır. Bu noktada, sınama verisinin, dilin tamamını temsil ettiği varsayılmıştır.

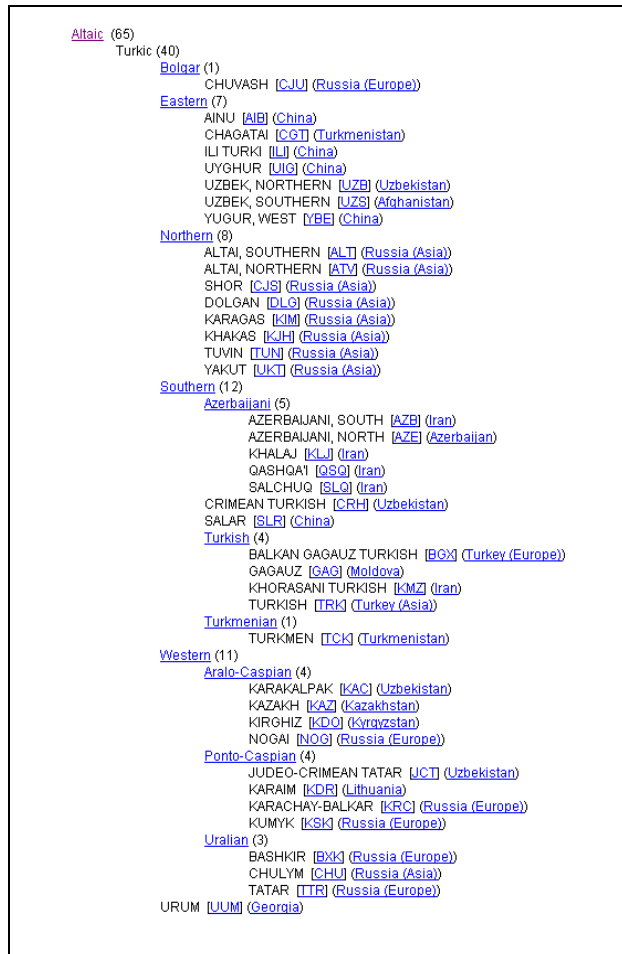
Tablo 4.8 : Farklı Tipte ve Derecede İDM’lerin Entropi Değerleri

İDM Tipi	Entropi (bit)				
	n=1	n=2	n=3	n=4	n=5
Tip-I - Kök	10,32	8,93	8,79	8,79	8,80
Tip-II - Son Sözcük Türü	2,43	2,30	2,27	2,26	2,26
Tip-III - Son Çekim Grubu	6,27	5,77	5,69	5,69	5,69
Tip-IV - Kök + Son Sözcük Türü	10,88	9,50	9,38	9,39	9,39
Tip-V - Kök + Son Çekim Grubu	12,72	11,48	11,41	11,41	11,41
Tip-VI - Tüm Çözümleme (kök hariç)	6,64	6,15	6,08	6,08	6,08

5. TÜRK DİL AİLESİ

Türk dil ailesi, çoğunlukla Orta Asya coğrafyasına yayılmış ve toplamda yaklaşık 180 milyon insanın kullandığı, aynı temelleri paylaşan ve birçok benzer özelliği bulunan dillerden oluşmaktadır. Türkiye Türkçesi ile diğer Türk dilleri arasında benzerliğin yüksek olması, bağlantılı diller olan Türk dillerinde sözcük köklerinin çoğunlukla aynı kalmasından kaynaklanmaktadır [49].

Türk dil ailesindeki Türk dillerinin lehçe mi yoksa ayrı birer dil mi olduğu dil bilimciler arasında halen bir tartışma konusudur. Bu konuda bizim kabul ettiğimiz görüş, karşılıklı anlaşılabilirlik (mutual intelligibility) ilkesine dayanan ve Türk dillerini ayrı birer dil olarak niteleyen görüştür [50]. Türk dil ailesinin sınıflandırılması şu şekildedir [51]:



Şekil 5.1 : Türk Dil Ailesinin Sınıflandırması



Şekil 5.2 : Türk Dillerinin Konuşulduğu Başlıca Coğrafyaların Haritası

Türk dilleri çoğunlukla eski SSCB'den ayrılarak bağımsızlığını ilan eden ülkelerde konuşulmaktadır. Türk dillerinin konuşulduğu başlıca ülkelere ilişkin coğrafi bölgenin haritası Şekil 5.2'de verilmiştir.

Tablo 5.1 : Türk Dilleri İle İlgili Bazı Bilgiler

Dil	Dil Kodu	Kullanıldığı Bölge (Başlıca)	Kullanan Kişi Sayısı (Yaklaşık)
Türkçe	TRK	Türkiye	72 milyon
Azerice	AZJ	Azerbaycan	24,3 milyon
Azerice	AZB	İran	7 milyon
Türkmençe	TCK	Türkmenistan	6,4 milyon
Kazakça	KAZ	Kazakistan	8 milyon
Kırgızca	KDO	Kırgızistan	2,6 milyon
Uygurca	UIG	Çin (Doğu Türkistan bölgesi)	7,6 milyon
Özbekçe	UZB	Özbekistan	18,5 milyon
Özbekçe	UZS	Afganistan	1.4 milyon
Çuvaşça	CJU	Rusya	2 milyon

Günümüz Türkiye Türkçesine en çok benzeyen diller, aynı sınıf içerisinde yer alan Azerice, Türkmence ve Gagauz (Gök Oğuz) Türkçesidir. Özbekçenin, günümüz Türkçesine benzerlik derecesi ortadır. Özbekçe ile Türkmence, Uygurca, Kazakça ve Kırgızca benzerlik derecesi de ortadır. Kırgızca ve Kazakça, coğrafi olarak Türkiye'ye uzak oldukları için Türkçeye Özbekçe ve Türkmenceden daha uzaktır. Kıpçak grubundaki diller arasında Türkçeye en yakın dil Kırım-Tatar Türkçesidir. Yakutça ve Çuvaşçayı ise günümüz Türkçesinden tamamen farklılaşmıştır [49].

5.1 Türk Dilleri Arasındaki Benzerlikler

Türk dilleri anlam ve biçim açısından incelendiğinde birçok benzerlik ve aynılık olduğu görülür. Benzerlikler, sözcük dağarcıklarındaki ortak sözcükler açısından incelendiğinde baskın olarak adılar, sıfatlar, ilgeçler, belirteçler, zamanla ilgili sözcükler, organ isimleri, doğa, bitki ve hayvan isimlerinde ortak kullanımların olduğu görülmektedir. Ayrıca sözdizimsel açıdan bütün Türk dilleri özne-nesne-yüklem (SOV) sırasını kullanmaktadır. Ortak sözcüklerin dışında Türk dillerinin benzerlikleri, biçimbilimsel açıdan incelendiğinde ise ek türlerinin ve ekleniş biçimlerinin çoğu kez aynı olduğu görülmektedir. Örneğin Tablo 5.2'de bazı Türk dilleri için isim durum ekleri verilmiştir [49]. Hemen hemen bütün Türk dillerinde özellikle adlara eklenen çekim eklerinin türleri ve sıraları aynıdır. Büyük ünlü uyumu, küçük ünlü uyumu, ünsüz benzeşmesi gibi bazı sesbilgisel olaylar hepsinde görülmese de bazı Türk dillerinde ortak olarak bulunur. Tüm bu benzerliklere örnek olarak aşağıda farklı Türk dillerinde ortak olarak kullanılan iki ata sözü verilmiştir.

Türkçe	: Ağır	kazan	geç	kaynar.
Azerice	: Ağır	qazan	geç	qaynayar.
Türkmence	: Agyr	gazan	giç	gaýnar.
Özbekçe	: Çuqur	därya	tinç	aqar.
Kırgızca	: Oor	kazan	keç	kaynayt.
Kazakça	: Awur	qazan	keş	qaynaydı.

Türkçe	: Dağ	dağa	kavuşmaz,	insan	insana	kavuşur.
Azerice	: Dağ	dağa	govuşmaz,	insan	insana	govuşur.
Türkmence	: Dag	daga	duşmaz,	adama	adama	duşar.
Özbekçe	: Tâğ	tâğ	bilän	qavuşmas,	adam	adam bilän qavuşar.
Kırgızca	: Too	tooğo	koşulbayt,	adam	menen	adam koşulat.
Kazakça	: Taw	tawğa	qosılmas,	adam	adamğa	qosıldı.

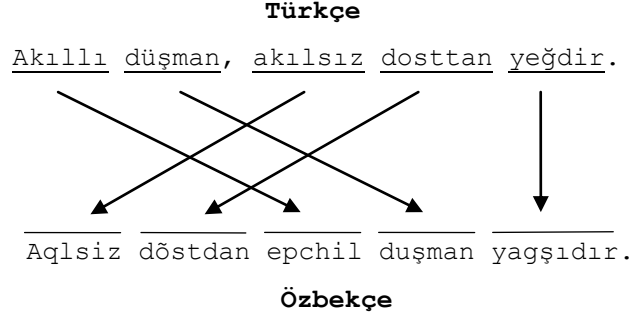
Görüldüğü gibi sözcük sıraları çoğunlukla aynıdır. Yalnızca bazı sözcükler bazı Türk dillerinde iki ya da daha fazla sözcükle ifade edilmektedir.

Tablo 5.2 : Bazı Türk Dilleri İçin İsim Durum Ekleri

Ad Durum Ekleri	Türkçe	Azerice	Türkmençe	Özbekçe	Kırgızca	Kazakça	Uygurca
Belirtme Durumu	+ı (+i,+u,+ü) +yı (+yi,+yu,+yü)	+ı (+i,+u,+ü) +nı (+ni,+nu,+nü)	+y (+i) +ny (+ni)	+ni	+nı (+ni,+nu,+nü) +dı (+di,+du,+dü) +tı (+ti,+tu,+tü) +n	+nı (+ni) +dı (+di) +tı (+ti) +n	+ni
Yönelme Durumu	+a (+e) +ya (+ye)	+a (+ə) +ya (+yə)	+a (+e, +ä)	+gä +kä +qa	+ga (+ge,+go,+gö) +ka (+ke,+ko,+kó) +na (+ne,+no,+nö) +a (+e, +o,+ö)	+ğa (+ge) +qa (+ke) +a (+e)	+ga (+ge) +ka (+ke)
Kalma Durumu	+da (+de) +ta (+te)	+da (+də)	+da (+de)	+dä	+da (+de,+do,+dö) +ta (+te,+to,+tö)	+da (+de) +ta (+te) +nda (+nde)	+da (+de) +ta (+te) +nda (+nde)
Çıkma Durumu	+dan (+den) +tan (+ten)	+dan (+dən)	+dan (+den)	+dän	+dan(+den,+don,+dön) +tan (+ten,+ton,+tön) +nan(+nen,+non,+nön)	+dan (+den) +tan (+ten) +nan (+nen)	+din +tin
Tamlayan Durumu	+ın (+in,+un,+ün) +nın (+nin,+nun,+nün)	+ın (+in,+un,+ün) +nın (+nin,+nun,+nün)	+yň (+iň,+uň,+üň) +nyň (+niň)	+niň	+nın (+nin,+nun,+nün) +dın (+din,+dun,+dün) +tın (+tin,+tun,+tün)	+nıň (+niň) +dıň (+diň) +tıň (+tiň)	+niň

5.2 Türk Dilleri Arasındaki Farklılıklar

Tümceler arasında sözcük sıraları açısından çoğunlukla benzerlik söz konusu olsa da bazı durumlarda tümce içindeki sözcüklerin yerleri de değişebilir. Örnek olarak çoğu Türk dilinde bulunan aşağıdaki atasözü verilmiştir:



Şekil 5.3 : Türkçe-Özbekçe Tümcelerde Sözcük Sıraları Farklılığı Örneği

Ancak Türk dillerinde, sözcük öbeklerinin tümce içerisinde yer değiştirebilmesi özelliği bulunduğundan, Özbekçe tümce Türkçeye çevrilirken sözcük sıralarında bir değişikliğe gidilmese bile anlamı koruyan Türkçe tümce üretilebilir:

Akılsız dosttan akıllı düşman yeğdir.

Yazım dili açısından Türk dilleri tarihte farklı abece sistemlerini kullanmışlardır. Özellikle SSCB'nin zorlaması ile bir bu bölgelerde kullanılan Türk dilleri Kiril abecesi ile yazılmaya başlanmıştır. Bunların dışında Arap ve Latin abecesini kullanan diller birbirlerine yakınlaşırken, bu süreçte Kiril abecesini kullanan diller bilinçli olarak farklılaştırılmıştır. Günümüzde Çin'de konuşulan ve Arap abecesini kullanan Uygurca dışındaki bütün Türk dilleri Latin abecesine geçiş yapmışlardır.

Türk dilleri arasında gözlenebilen diğer farklılıklar özellikle eylem çekimlerinde ortaya çıkan farklı zaman kullanımları, dillere özel kipler ve özne-yüklem uyumlarındaki farklılıklardır. Örneğin Türkçedeki geniş zaman kalıbı Türkmencede gelecek zaman anlamını taşır. Ayrıca Türkmencede, Türkçede bulunmayan **+makçı/+mekçi** ekleri ile kurulan ve “bir eylemi yapmayı düşünmek/yapmaya niyetlenmek” anlamında bir eylem kipi bulunur.

5.3 Türk Dilleri Hakkında Özet Bilgiler

Bu bölümde başlıca Türk dillerinin temel özellikleri ve abeceleri ile ilgili bilgiler verilecektir. Abecelerin gösterildiği tablolarda, koyu ve eğik olarak belirtilen harfler, Türkçede olmayan ya da farklı sesler için kullanılan harfleri göstermektedir.

5.3.1 Azerice

Azeri Türkçesinin yeni abecesinde 32 harf vardır. Türkçeden farklı olan bu üç harf (ə, x ve q) Tablo 5.3’de kalın olarak gösterilmiştir.

Tablo 5.3 : Azericenin Abecesi

	Azeri Abecesi		Türkçe Karşılığı
1	A	a	a
2	B	b	b
3	C	c	c
4	Ç	ç	ç
5	D	d	d
6	E	e	e
7	Ə	ə	ince e
8	F	f	f
9	G	g	g
10	Ğ	ğ	ğ
11	H	h	he
12	X	x	ha
13	I	ı	ı
14	İ	i	i
15	J	j	j
16	K	k	ke
17	Q	q	ka
18	L	l	l
19	M	m	m
20	N	n	n
21	O	o	o
22	Ö	ö	ö
23	P	p	p
24	R	r	r
25	S	s	s
26	Ş	ş	ş
27	T	t	t
28	U	u	u
29	Ü	ü	ü
30	V	v	v
31	Y	y	y
32	Z	z	z

Azericede, Türkçedeki gibi büyük ve küçük ünlü uyumu bulunur. Buna karşın Türkçedeki ünsüz benzeşmesi ise gözlenmez.

Azerice isim çekimi, Türkçeye çok benzemektedir ancak arada küçük farklılıklar da vardır.

Tekillik / Çoğulluk

Çoğulluk ekleri **+lar** ve **+lər** Türkçedeki ile aynı şekilde kullanılır.

Belirtme Durumu

Örneğin belirtme durumu eki Türkçedeki gibi **+ı**, **+i**, **+u** ve **+ü** eklerinden oluşmaktadır. Ancak Azericede ekten önce sesli bulunursa araya **y** harfi yerine **n** harfi gelir:

dəpdəri (defteri)	xalğı (halkı)	kulağımızı (qulağımızı)
arabanı (arabayı)	sürücünü (sürücüyü)	ölçünü (ölçüyü)

Yönelme Durumu

Türkçedeki gibi **+a** ve **+ə** ekleri ile kurulur.

dəpdərə (deftere)	xalga (halka)	anaya (anaya)
-------------------	---------------	---------------

Kalma Durumu

Türkçedekine benzer olarak **+da**, **+də** ekleri ile kurulur. Türkçedeki kullanımdan tek farklı bu eklerin ünsüz benzeşmesine uymamasıdır (**+ta**, **+te** halleri yoktur):

başda (başta)	evdə (evde)	kitapda (kitapta)
---------------	-------------	-------------------

Çıkma Durumu

Çıkma durumu, **+dan** ve **+dən** ekleri ile kurulur. Bu ekler de ünsüz benzeşmesine uymazlar:

atdan (attan)	evdən (evden)	kitaptan (kitaptan)
---------------	---------------	---------------------

Tamlayan Durumu

Tamlayan durumu ekleri de Türkçe ile tamamen aynıdır:

kitabın (kitabın)	xalgın (halkın)	ölünün (ölünün)
-------------------	-----------------	-----------------

5.3.2 Türkmençe

1920'lere kadar Latin abecesini kullanırken, Rusya'nın egemenliği altına girdikten sonra Kiril abecesini kullanmaya başlanmıştır. 1990'lardan sonra bağımsızlığını elde eden Türkmenistan'ın Kiril abecesinden, Latin abecesine dönüşü sancılı olmuştur. Standart bir abece ortaya çıkana kadar çok farklı Latin harfleri kullanılmıştır. Ancak Türkmenistan Milli Meclisi, 12.04.1993 tarihinde "Yeni Türkmen Abecesi"ni kabul ederek, abece konusundaki kargaşalara son vermiştir [49]. Günümüzde kullanımda olan Türkmen abecesi Tablo 5.4'te verilmiştir. Türkmençenin biçimbilimsel yapısı Bölüm 7.1.1'de ayrıntılı olarak ele alınacağından burada anlatılmamıştır.

Tablo 5.4 : Türkmençenin Abecesi

	Türkmen Abecesi		Türkçe Karşılığı
1	A	a	A
2	B	b	B
3	J	j	c
4	Ç	ç	Ç
5	D	d	d
6	E	e	e
7	Ä	ä	ince e
8	F	f	f
9	G	g	g
10	H	h	h
11	Y	y	ı
12	I	i	i
13	Ž	ž	j
14	K	k	k
15	L	l	l
16	M	m	m
17	N	n	n
18	Ñ	ñ	ince (nazal) n
19	O	o	o
20	Ö	ö	ö
21	P	p	p
22	R	r	r
23	S	s	s
24	Ş	ş	Ş
25	T	t	t
26	U	u	u
27	Ü	ü	ü
28	W	w	v
29	Ý	ý	y
30	Z	z	z

5.3.3 Kazakça

Kazakça, Doğu Türkistan'da, Moğolistan'da, Rusya'da ve Orta Asya'nın bazı bölümlerinde konuşulur; Abecesinde toplam 32 harf bulunur:

Tablo 5.5 : Kazakçanın Abecesi

	Kazak Abecesi		Türkçe Karşılığı
1	A	a	a
2	B	b	b
3	J	j	c
4	D	d	d
5	E	e	e
6	Ä	ä	ince e
7	F	f	f
8	G	g	g
9	Ğ	ğ	ğ
10	H	h	he
11	X	x	ha
12	I	ı	ı
13	İ	i	i
14	K	k	ke
15	Q	q	ka
16	L	l	l
17	M	m	m
18	N	n	n
19	Ñ	ñ	ince (nazal) n
20	O	o	o
21	Ö	ö	ö
22	P	p	p
23	R	r	r
24	S	s	s
25	Ş	ş	ş
26	T	t	t
27	U	u	u
28	Ü	ü	ü
29	V	v	v
30	Y	y	y
31	Z	z	z
32	W	w	v

Kazak Türkçesinde büyük ünlü uyumu kuralı gözlenmektedir. Küçük ünlü uyumu da kural olarak bulunmasına karşın bazı sözcükler küçük ünlü uyumuna uymaz. Kazak Türkçesinde ünsüz benzeşmesi kuralı da bulunur:

kitaptı (kitaptı) qısta (kışta) şašta (saçta)

Kazakçanın isim çekim ekleri Türkçeninkinden biraz farklıdır.

Tekillik / Çoğulluk

- Ünlüler ve **r, w, y** ünsüzlerinden sonra **+lar** ve **+ler** eki gelir:

balalar (çocukar) qoylar (koyunlar) tawlar (dağlar)

- Yumuşak ünsüzlerden sonra **+dar** ve **+der** eki gelir:

qazdar (kazlar) qoyındar (koyunlar) öleñder (ölenler)

- Sert ünsüzlerden sonra **+tar** ve **+ter** eki gelir:

ağaştar (ağaçlar) esikter (kapılar) tister (dişler)

Belirtme Durumu

- Ünlülerden sonra **+nı, +ni** eki gelir:

almanı (elmayı) evi (üyini) atanı (atayı)

- Yumuşak ünsüzlerden sonra **+dı, +di** eki gelir:

qalamdı (kalemi) küldi (gülü) qarındı (karnı)

- Sert ünsüzlerden sonra **+tı, +ti** eki gelir:

uşaqtı (uçacağı) esikti (kapıyı) balıqtı (balığı)

Yönelme Durumu

- Ünlülerden ve yumuşak ünsüzlerden sonra **+ğa, +ge** eki gelir:

köşege (caddeye) üyge (eve) olarğa (onlara)

- Sert ünsüzlerden sonra **+qa, +ke** eki gelir:

qonaqqa (konuğa) iske (işe) aşqa (aşa)

- Tekil 1. ve 2. kişi iyelik ekinden sonra **+a, +e** gelir:

qalama (şehrime) dosına (dostuna) üyine (evine)

- Tekil ve çoğul 2. iyelik eklerinden sonra **+na, +ne** gelir:

awlına (köyüne) qolına (eline) köline (gölüne)

Kalma Durumu

- Ünlülerden ve yumuşak ünsüzlerden sonra **+da, +de** eki gelir:

qazda (kazda) qumda (kumda) elde (ülkede)

- Sert ünsüzlerden sonra **+ta, +te** eki gelir:

ata (atta)

tiste (dişte)

köşte (göçte)

Çıkma Durumu

- Ünlülerden ve yumuşak ünsüzlerden sonra **+dan, +den** eki gelir:

joldan (yoldan)

üyden (evden)

sizden (sizden)

- Sert ünsüzlerden sonra **+tan, +ten** eki gelir:

qıştan (kıştan)

küşten (güçten)

ayaqtan (ayaktan)

- **m, n, ñ** ve 3. iyelik eklerinden sonra **+nan, +nen** eki gelir:

adamnan (adamdan)

üvine (evinden)

innen (inden)

Tamlayan Durumu

- Ünlülerden ve m, n, ñ ünsüzlerinden sonra **+nıñ, +niñ** eki gelir:

balanıñ (çocuğun)

şeşeniñ (annenin)

künniñ (güneşin)

- Yumuşak ünsüzlerden sonra **+dıñ, +diñ** eki gelir:

olardıñ (onların)

joldıñ (yolun)

köldiñ (gölün)

- Sert ünsüzlerden sonra **+tıñ, +tiñ** eki gelir:

qustıñ (kuşun)

jürektiñ (yüreğin)

küştiñ (gücün)

5.3.4 Kırgızca

Kırgız abecesinde toplam 30 harf vardır. Kırgızcanın yeni abecesi Tablo 5.6’da verilmiştir:

Tablo 5.6 : Kırgızcanın Abecesi

Kırgız Abecesi		Türkçe Karşılığı	
1	A	a	a
2	B	b	b
3	C	c	c
4	Ç	ç	ç
5	D	d	d
6	E	e	ince e
7	É	é	e
8	F	f	f
9	G	g	g
10	Ğ	ğ	ğ
11	H	h	h
13	I	ı	ı
14	İ	i	i
16	K	k	k
18	L	l	l
19	M	m	m
20	N	n	n
21	Ñ	ñ	ince (nazal) n
22	O	o	o
23	Ö	ö	ö
24	P	p	p
25	R	r	r
26	S	s	s
27	Ş	ş	ş
28	T	t	t
29	U	u	u
30	Ü	ü	ü
31	V	v	v
31	Y	y	y
32	Z	z	z

Türk dil ailesindeki dillerde genellikle iki ünlü yan yana gelmezken Kırgızcada bu durum çok sık görülür. Bu durum, ünlünün uzun okunması gerektiği gösterir. Aslında diğer Türk dillerinde de okunuşu uzun olan ünlülerin bulunmaktadır. Ancak bu dillerde, yazı dilinde bu durum belirtilmemesine rağmen Kırgızcada ünlü harfin iki defa yazılması ile belirtilir.

Kırgızcada büyük ünlü uyumu vardır. Ayrıca Türkçedekine çok benzer küçük ünlü uyumu da bulunur. Ünsüz benzeşmesi de Kırgızca'nın Türkçeye benzeyen özelliklerindendir.

Tekillik / Çoğulluk

- Ünlülerden sonra **+lar, +ler, +lor** ve **+lör** ekleri gelir:

talaalar (bozkırlar) köçölör (köşeler) ündüülör (ünlüler)

- Yumuşak ünsüzlerden sonra **+dar, +der, +dor** ve **+dör** ekleri gelir:

kağazdar (kağıtlar) közdör (gözler) kalemder (kalemler)

- Sert ünsüzlerden sonra **+tar, +ter, +tor** ve **+tör** ekleri gelir:

tooktor (tavuklar) baştar (başlar) cigitter (yiğitler)

Belirtme Durumu

- Ünlülerden sonra **+nı, +ni, +nu** ve **+nü** ekleri gelir:

baanı (kıymetini) toonu (dağı) cazuunu (yazıyı)

- Yumuşak ünsüzlerden sonra **+dı, +di, +du** ve **+dü** ekleri gelir:

kagazdı (kağıdı) üydü (evi) kazdı (kazı)

- Sert ünsüzlerden sonra **+tı, +ti, +tu** ve **+tü** ekleri gelir:

kuştı (kuşu) küçtü (gücü) attı (atı)

Yönelme Durumu

- Ünlülerden ve yumuşak ünsüzlerden sonra **+ga, +ge, +go** ve **+gö** ekleri gelir:

toogo (dağa) küzgö (güze) elge (halka)

- Sert ünsüzlerden sonra **+ka, +ke, +ko** ve **+kö** ekleri gelir:

mektepke (okula) taşka (taşa) tüpko (dibe)

- Tekil 1. ve 2. kişi iyelik ekinden sonra **+a, +e, +o** ve **+ö** gelir:

balama (çocuğuöa) elime (halkıma) közümö (gözüme)

- Tekil 3. kişi iyelik eklerinden sonra **+na, +ne, +no** ve **+nö** gelir:

atasına (atasına) toosuna (dağına) ağasına (ağabeyine)

Kalma Durumu

- Ünlülerden ve yumuşak ünsüzlerden sonra **+da, +de, +do** ve **+dö** eki gelir:

caanda (yağmurda) köldö (gölde) coldo (yolda)

- Sert ünsüzlerden sonra **+ta, +te, +to** ve **+tö** eki gelir:

kitepte (kitapta) atta (atta) tookto (tavukta)

Çıkma Durumu

- Ünlülerden ve yumuşak ünsüzlerden sonra **+dan, +den, +don** ve **+dön** eki gelir:

toodon (dağdan) közdön (gözden) oozdon (ağızdan)

- Sert ünsüzlerden sonra **+tan, +ten, +ton** ve **+tön** eki gelir:

kitepten (kitaptan) tüptön (dipten) tookton (tavuktan)

- 3. tekil iyelik eklerinden sonra **+nan, +nen, +non** ve **+nön** eki gelir:

colunan (yolundan) apasınan (anasından) közünön (gözünden)

Tamlayan Durumu

- Ünlülerden ünsüzlerinden sonra **+nın, +nin, +nun** ve **+nün** eki gelir:

kalaanın (şehrin) talaanın (annenin) toonun (dağın)

- Yumuşak ünsüzlerden sonra **+dın, +dın, +dun** ve **+dün** eki gelir:

ağamdın (ağabeyimin) bizdın (bizim) közümdün (gözümün)

- Sert ünsüzlerden sonra **+tın, +tın, +tun** ve **+tün** eki gelir:

curttun (yurdun) kiteptin (kitabın) çaçtın (saçın)

- **+ki** ekinden önce gelen **n** ünsüzü düşer:

eceniki (ablanınki) üydükü (evinki) sizdiki (sizinki)

5.3.5 Uygurca

Uygurca, Doğu Türkistan’da ve Orta Asya’nın değişik bölgelerinde kullanılır. Doğu Türkistan’daki Uygular Arap abecesini kullanırken diğerleri Kiril abecesini kullanmaktadır. Tablo 5.7’de Uygur abecesindeki Arap harflerinin Latin harfleriyle yazılmış bir sürümü verilmiştir.

Tablo 5.7 : Uygurcanın Abecesi

	Uygur Abecesi		Türkçe Karşılığı
1	A	a	A
2	B	b	B
3	C	c	c
4	Ç	ç	Ç
5	D	d	d
6	É	é	e
7	E	e	ince e
8	F	f	f
9	G	g	g
10	Ğ	ğ	ğ
11	H	h	he
12	H	h	ha
14	I	i	i
15	J	j	j
16	K	k	ke
17	K	k	ka
18	L	l	l
19	M	m	m
20	N	n	n
21	Ñ	ñ	ince (nazal) n
22	O	o	o
23	Ö	ö	ö
24	P	p	p
25	R	r	r
26	S	s	s
27	Ş	ş	ş
28	T	t	t
29	U	u	u
30	Ü	ü	ü
31	V	v	v
32	Y	y	y
33	Z	z	z

Uygurcada büyük ünlü uyumu kısmen gözlense de küçük ünlü uyumu gözlenmez.

Tekillik / Çoğulluk

Uygurcanın çoğul ekleri Türkçedeki **+lar** ve **+ler** ekleridir.

Belirtme Durumu

Belirtme durumu eki büyük ünlü uyumuna uymayan **+ni** ekidir.

adāmni (adamı) işlerni (işlrei) kitapni (kitabı)

Yönelme Durumu

Yönelme durumu eki **+ga** ve **+ge** ekleridir. Ünsüz benzeşmesine uyan bu ek sert ünsüzlerden sonra **+ka** ve **+ke** biçiminde kullanılır.

adāmge (adama) işlerge (işlere) etrafka (etrafka)

Kalma Durumu

Belirtme durumu eki **+da** ve **+de**'dir. Ünsüz benzeşmesine uyarak bu ek sert ünsüzlerden sonra **+ta** ve **+te** biçimlerine dönüşür.

kutuda (kutuda) işte (işte) közlerimizde (gözlerimizde)

Çıkma Durumu

Uygurcadaki çıkma durumu eki **+din**'dir. Bu ek ilginç bir davranış göstererek ünlü uyumuna uymaz ama ünsüz benzeşmesine uymaktadır.

sizdin (sizden) yoldin (yoldan) çeçäktin (çiçekten)

Tamlayan Durumu

Belirtme durumu eki **+niñ**'dir.

Ankara'niñ (Ankara'nın) işniñ (işin) adāmniñ (adamın)

5.3.6 Özbekçe

Özbek dilinde konuşma dilinde gözlenen büyük ünlü uyumu kuralına yazı dilinde rastlanmaz. Bunun nedeni, konuşma dilinde mevcut olan **ı, ö** ve **ü** seslerine karşılık yazı dilinde bir harf kullanılmamasıdır.

Tablo 5.8 : Özbekçenin Abecesi

Özbek Abecesi			Türkçe Karşılığı
1	О (Ў)	о (ў)	a
2	B	b	b
3	Ҳ	ҳ	c
4	Ç	ç	ç
5	D	d	d
6	E	e	e
7	Ў	ў	ince e
8	F	f	f
9	G	g	g
10	Ğ	ğ	ğ
11	H	h	he
12	Х	х	ha
13	И	и	ı, i
14	J	j	j
15	K	k	ke
16	Қ	қ	ka
17	L	l	l
18	M	m	m
19	N	n	n
20	Њ	њ	ince (nazal) n
21	Ў	ӯ	o, ö
22	P	p	p
23	R	r	r
24	S	s	s
25	Ş	ş	ş
26	T	t	t
27	У	u	u, ü
28	V	v	v
29	Y	y	y
30	Z	z	z

Özbekçede büyük ve küçük ünlü uyumu ile birlikte ünsüz benzeşmesi kuralı da yoktur.

Tekillik / Çoğulluk

Büyük ünlü uyumu kuralı olmadığı için Özbekçede çoğul eki taktır : **+lär**

Belirtme Durumu

Belirtme durumu eki yalnızca **+ni**'dir.

uyni (evi) yölni (yolu) kitâbni (kitabı)

Yönelme Durumu

Yönelme durumu eki **+ga**'dır. Ancak **-k** harfî ile biten sözcüklerde **+kâ**, **-q** ve **-ğ** harfi ile biten sözcüklerde ise **+qa** olarak kullanılır.

uygä (eve) körekkä (küreğe) baqqa (bağa)

Kalma Durumu

Kalma durumu eki yalnızca **+dä**'dir.

uydä (evde) bizdä (bizde) kitâbdä (kitapta)

Çıkma Durumu

Çıkma durumu eki yalnızca **+dän**'dir.

uydän (evden) bizdän (bizden) kitâbdän (kitaptan)

Tamlayan Durumu

Tamlayan durumu eki yalnızca **+nin**'dir.

uyniñ (evin) bizniñ (bizim) kitâbniñ (kitabın)

6. TÜRK DİLLERİ ARASINDA BİLGİSAYARLI ÇEVİRİ

Türk dilleri arasında çeviri yaparken ortaya çıkan en büyük sorun, Türkçe dışındaki diğer Türk dilleri için doğal dil işleme çalışmalarının yok denecek kadar az olmasıdır. Bir çok Türk dili için biçimbilimsel çözümleme, sözdizim ayrıştırıcısı gibi araçlar ya da elektronik ortama aktarılmış sözlükler veya işlenmiş metinler bulmak olası değildir. Türkçe için bile genel amaçlı kullanılabilecek yüksek başarımlı bir sözdizimsel çözümleme aracı bulunmamaktadır.

Bütün bu olumsuzluklara karşın, özellikle Türk dillerinin sözdiziminin benzer olması (Azerice, Türkmence gibi bazı diller için neredeyse aynı olması), Bölüm 3.1’de anlatılan aktarım yöntemleri arasında en temel yöntem olan doğrudan aktarım yönteminin bile başarılı sonuçlar üretebileceğini düşündürmektedir.

Doğrudan aktarım yöntemi hariç diğer tüm bilgi tabanlı aktarım yöntemlerinde, sözdizim çözümlemesi, anlamsal çözümleme gibi üst düzey bilgiler gerekmektedir. Ancak Türk dilleri söz konusu olduğunda bu bilgileri üretecek araçlar dahi bulunmamaktadır.

İstatistiksel ve örnek tabanlı çalışan yöntemleri kullanabilmek için ise birbirlerinin karşılığı olan hizalanmış eğitim verilerine büyük miktarlarda gerek duyulur. Bu şekilde hazırlanmış paralel eğitim verilerinin bulunmaması, ayrıca bu tür bir eğitim kümesinin hazırlanmasının yüksek maliyet gerektirmesinden dolayı bu yöntemlerin uygulanabilirliği düşüktür.

Sözcük sıralarını değiştirmeden, sadece sözcükleri birebir çevirerek gerçekleştirilecek doğrudan aktarım yöntemi, gerek sözdizim çözümlemesi gibi daha üst seviyede bilgilere, gerekse de istatistiksel yöntemlerin kullandığı gibi büyük miktarlarda paralel eğitim verilerine ihtiyaç duymadığı için en uygun yöntem olarak görülmektedir. Ayrıca doğrudan aktarım yöntemi, sadece sonlu durum yöntemleri kullanılarak gerçekleştirilebilmektedir. Bu sayede sistematik ve hızlı çalışan bir aktarım sistemi gerçekleştirilebilir.

Ancak Türk dilleri gibi bitişken diller söz konusu olduğunda doğrudan aktarım yöntemi Tablo 6.1'deki gibi geliştirilerek kullanılmalıdır:

Tablo 6.1 : Geliştirilmiş Doğrudan Aktarım Yöntemi Aşamaları

-
1. Kaynak Dil Biçimbilimsel Çözümlemesi
 2. Biçimbilimsel Yapıların Hedef Dile Aktarılması
 3. Sözcük Kökünün Hedef Dile Aktarılması
 4. Hedef Dilde Biçimbilimsel Üretici ile Sözcüğün Yüzeysel Biçiminin Üretilmesi
-

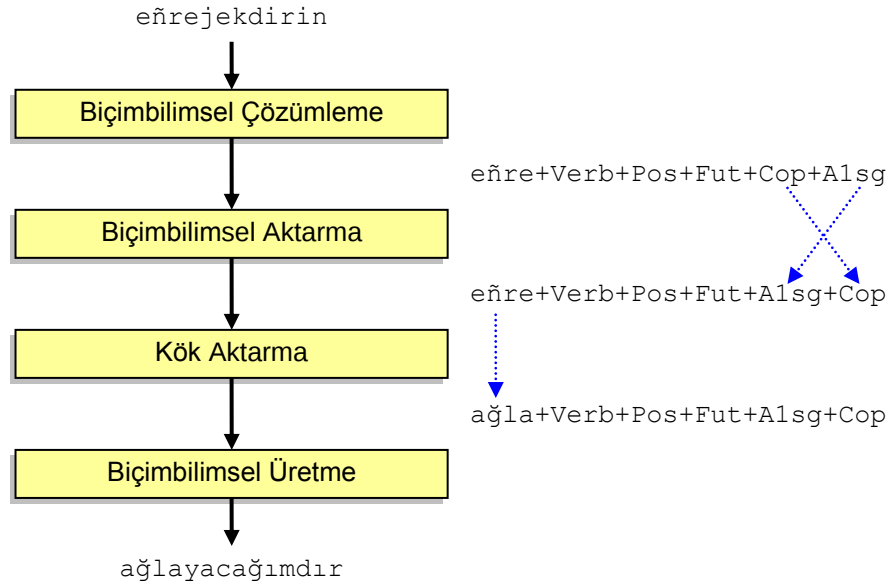
Bitişken yapısından dolayı, kaynak tümcedeki bir sözcüğün karşılığını sözlükte olduğu gibi arayıp bulmak mümkün değildir. Bu noktada, bir ön işlem olarak biçimbilimsel çözümlemenin yapılması gereklidir. Biçimbilimsel çözümleme sonucunda sözcük kökü ve diğer biçimbilimsel yapılar ortaya çıkar. Artık doğrudan aktarım, kaynak dildeki bu yapıların (sözcük kökü ve biçimbilimsel yapılar) hedef dile iki aşamalı olarak aktarımı biçiminde algılanmalıdır.

Her ne kadar Türk dillerinin sözdizimsel ve biçimbilimsel yapıları birbirlerine yakın olsa da, bu diller arasında biçimbilimsel farklılıklar da azımsanmayacak boyuttadır. Biçimbilimsel farklılıkları gidermek üzere bir takım biçimbilimsel dönüşüm kuralları gerçekleştirilmeli ve bu kuralların işletilmesi sonucunda kaynak dildeki biçimbilimsel yapılar, hedef dil için geçerli biçimbilimsel yapılar haline gelmelidir.

İkinci aşama olarak, kaynak dilde çözümlenen sözcük kökünün karşılığı aktarım sözlüğünden bulunmalı ve hedef dildeki karşılığı veya karşılıkları ile değiştirilmelidir.

Son adımda ise elde edilen biçimbilimsel yapı, hedef dilin biçimbilimsel üreticisi tarafından yüzeysel biçime çevrilir.

Anlatılan geliştirilmiş doğrudan aktarım yöntemine göre Türkmençe bir sözcüğün Türkçe karşılığının oluşturulma süreci Şekil 6.1'de gösterilmiştir.



Şekil 6.1 : Örnek Türkmençe Sözcüğün Türkçe Karşılığının Oluşturulması

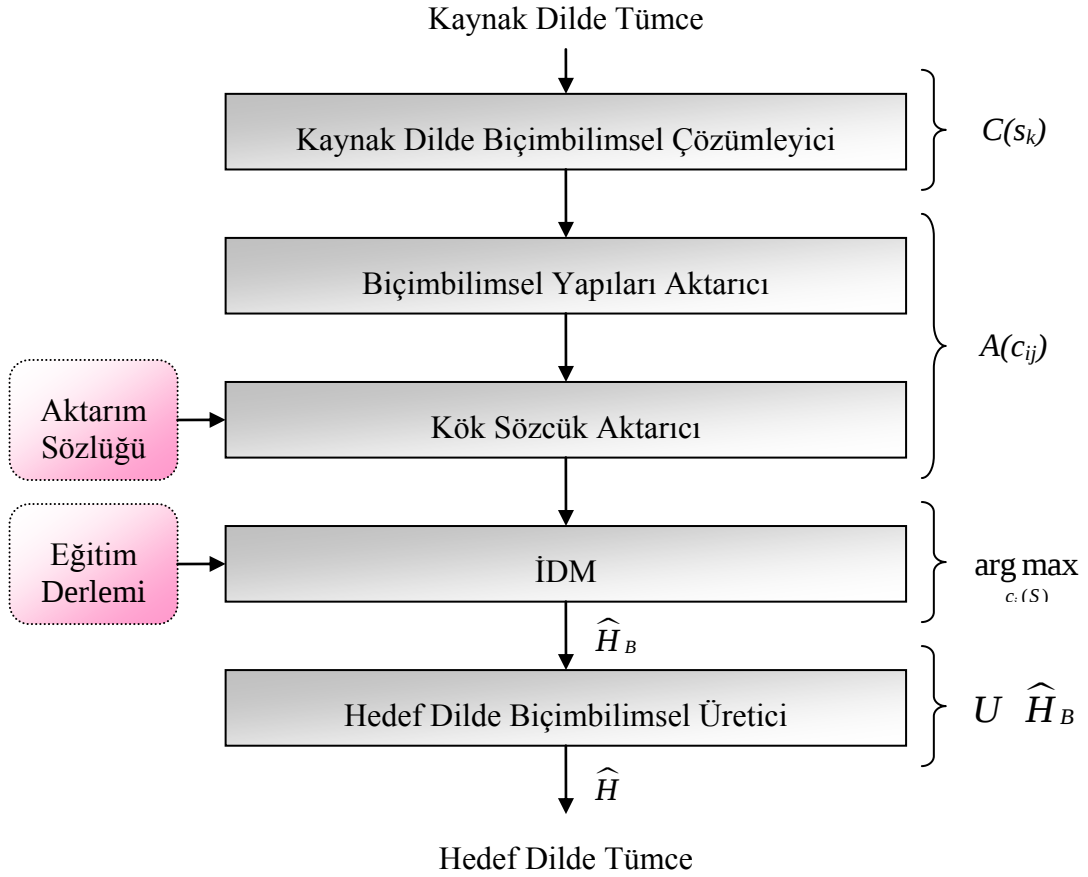
Geliştirilmiş doğrudan aktarım yönteminde dahi biçimbilimsel çözümleme ve sözcük kökü aktarımında belirsizlikler ortaya çıkacaktır. Bu belirsizliklerin çözülmesi, doğrudan aktarım yöntemini kullanan sistemlerde karmaşık kurallarla sağlanır (bkz. Bölüm 3.1.1). Bunun yerine, Bölüm 4’te önerilen ve doğrudan aktarım yaklaşımını istatistiksel yöntemlerle birleştiren modeller kullanılarak Türk dilleri arasında BÇ sistemleri gerçekleştirilebilir. Bu amaçla, Model 0 (temel model) üzerine gerçekleştirilmiş kuramsal bir çeviri sisteminin bileşenleri ve veri akışı Şekil 6.2’de verilmiştir. Bu şekilde kesikli çizgili dikdörtgenler veri kaynaklarını, düz çizgili dikdörtgenler ise süreçleri göstermektedir.

6.1 Kaynak Dilde Biçimbilimsel Çözümleme

Biçimbilimsel çözümleme, Türk dilleri gibi bitişken diller için doğal dil işleme alanında yapılacak her işlemde gerek duyulan bir aşamadır. Sözcüğün yüzeysel biçimlerinin sayısının çok fazla⁶ olduğu diller için biçimbilimsel çözümleme olmadan aktarım sözlükleri hazırlamak ya da aktarım kuralları geliştirmek olanaksızdır. Dolayısı ile kaynak dil olarak hangi Türk dili seçilirse seçilsin, bu dil ile ilgili biçimbilimsel çözümleyicinin de hazırlanması gerekmektedir. Türkçe için

⁶ Türkçede bir tek sözcük kökünden üretilebilecek yüzeysel biçimlerin sayısı bir milyondan fazladır

geliştirilmiş geniş kapsamlı ve yüksek başarımlı bir biçimbilimsel çözümleyici halihazırda bulunmaktadır [13]. Diğer Türk dilleri için bu tür bir çözümleyicinin varlığı (Kırım Tatarcası hariç) bilinmemektedir.



Şekil 6.2 : Temel Modeli Gerçekleyen Örnek Bir Çeviri Sistemi

Kaynak dil olarak seçilen dil, Türkçenin dışında bir Türk dili ise bu dil için bir biçimbilimsel geliştiricinin gerçekleşmesi ön koşuldur. Böyle bir gerçeklemede dikkat edilecek en önemli nokta, geliştirilecek çözümleyicinin var olan Türkçe biçimbilimsel çözümleyici ile benzer mantıkla çalışacak ve benzer biçimbilimsel etiketler üretecek şekilde tasarlanmasıdır. Bu sayede aktarım kuralları (yani A fonksiyonu) daha basit hale getirilebilir.

6.1.1 Kaynak Dilde Biçimbilimsel Belirsizliğin Giderilmesi

Türkçe gibi karışık bir biçimbilimsel yapıya sahip dillerde biçimbilimsel çözümleme sonuçları çoğu zaman birden fazladır. Bu ise **biçimbilimsel belirsizliğin** (morphological ambiguity) ortaya çıkmasına neden olmaktadır. Eğer kaynak Türk

dili için biçimbilimsel belirsizliği gidermek üzere bir araç varsa, bu araç kullanılarak istatistiksel sürecin karmaşıklığı azaltılabilir.

Türkçe için biçimbilimsel belirsizliklerin giderilmesi ile ilgili birçok çalışma olmasına karşın [52-54] bu çalışmaların sonucunda yüksek başarımlı bir araç henüz genel kullanıma sunulamamıştır. Diğer Türk dilleri için ise bu konuda bir çalışmaya rastlanamamıştır. Biçimbilimsel belirsizlik gidericilerin tasarlanması için kullanılan bir çok yöntem gözetimli (supervised) çalıştığı için elle işlenmiş çok miktarda eğitim verisine gerek duyulmaktadır. Türkçe dışındaki diğer Türk dilleri için henüz böyle bir eğitim kümesi olmamasından dolayı bilinen yöntemler kullanılarak bu diller için bir biçimbilimsel belirsizlik giderici tasarlanması yoluna gidilememektedir.

6.2 Sözcük Köklerinin Kaynak Dilden Hedef Dile Aktarımı

Kural tabanlı bütün bilgisayarlı çeviri sistemlerinde olduğu gibi öngörülen temel model için de bir aktarım sözlüğü gereklidir. Aktarım sözlüğünde kaynak dildeki sözcük bazında arama yapılabilmesi ve buna karşılık gelen hedef dildeki tüm sözcükler üretilebilmelidir. Bu noktada altı çizilmesi gereken konu, sözcük kökü aktarımı sırasında birden fazla karşılık üretilebileceğidir. Yani bu bileşenin ürettiği sonuçlar içerisinde bir belirsizlik vardır. Söz konusu bu belirsizlik **sözcüksel belirsizlik** (lexical ambiguity) olarak adlandırılmaktadır.

bar → var, bütün

Belirsizliği bir kademe azaltmak amacı ile sadece sözcük köküne bakarak arama yapmak yerine sözcük kökü ve sözcük türüne göre arama yapılabilir. Böylelikle yazımı aynı, ancak sözcük türleri farklı olan sözcük kökleri, daha az bir belirsizlikle aktarılabilir:

bar (sıfat) → bütün (sıfat)
bar (eylem) → var (eylem)

Aktarım sözlüğü tasarlanırken sözcük türlerine göre sınıflandırma yapılabilirse bu sayede belirsizliğin derecesi düşürülebilir.

6.3 Biçimbilimsel Yapıların Kaynak Dilden Hedef Dile Aktarımı

Kaynak ve hedef dil arasındaki biçimbilimsel farklılıkları gidermek üzere bir dizi dönüşümün yürütüldüğü aktarım bileşenidir. Bu dönüşüm, kaynak dildeki

biçimbilimsel çözümleme sonucunda üretilen biçimbilimsel özelliklerin, hedef dil biçimbilimsel üreticinin beklediği şekile dönüştürme işlemi olarak da nitelendirilebilir. Bu aşamada sözcük köküne dokunulmadan sadece biçimbilimsel etiketler üzerinde dönüştürme ve sıra değiştirme işlemleri yapılır. Bu kurallar, iki dil arasındaki biçimbilimsel farklılıklar incelenerek geliştirilir. Diğer bir yöntem ise birbirlerinin çevirisi olan, kaynak ve hedef dildeki biçimbilimsel yapıları içeren bir eğitim kümesi üzerinde, bilgisayar öğrenmesi yöntemlerinden birisinin eğitilerek kuralların otomatik olarak üretilmesidir.

6.4 İDM Bileşeni

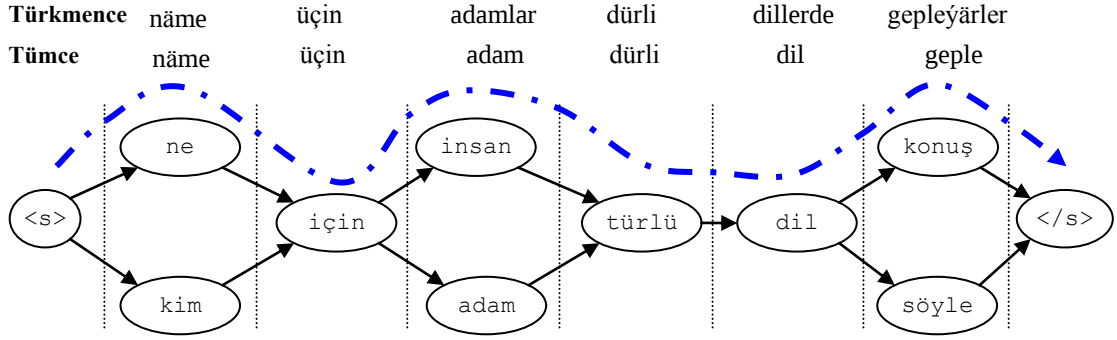
Önerilen aktarım modellerinde yer alan bütün bileşenler içerisinde kaynak dilde biçimbilimsel çözümleyici ve kök aktarımı bileşenlerinin çıktıları belirsizlik içermektedir. Eğer bölüm 6.1.1’de anlatılan kaynak dilde biçimbilimsel belirsizlik giderici kullanılabilir durumda ise biçimbilimsel belirsizlik elenir ve sadece sözcüksel belirsizlik kalır.

Gerek biçimbilimsel belirsizlik, gerekse de sözcüksel belirsizliğin giderilmesini amaçlayan İDM bileşeni, istatistiksel yöntemlerle en olası sözcük dizisini (yani tümceyi) belirler. Bu amaçla Bölüm 2.3’de anlatılan İDM’ler kullanılmaktadır. Ancak gene Türk dillerinin türetme ve çekim özelliklerinden dolayı, İDM’ler salt biçimde kullanılamaz. Sözcüklerin sadece yüzeysel biçimlerini içeren bir İDM’de seyrek veri sorunu ortaya çıkmaktadır. Bunu önlemek için Bölüm 4.3’te anlatıldığı gibi farklı tiplerde İDM’lerin kullanılması yoluna gidilebilir.

Örneğin, sözcüklerin yüzeysel biçimleri yerine sadece sözcük kökleri üzerine kurulmuş bir İDM kullanılması durumunda, hedef dilde ortaya çıkan sözcüksel belirsizliğin giderilmesi sağlanabilir.

Aktarım sistemindeki İDM bileşenine girdi olarak, kaynak dildeki tümcenin bütün sözcüklerinin aday çevirileri gelir. Bileşenin çıktısı olarak ise tüm kombinasyonlar içerisinde seçilen İDM’ye göre en yüksek olasılığa sahip tümce üretilir.

Olası tüm kombinasyonların tamamının olasılıklarının hesaplaması yerine, aday sözcüklerden bir Hidden Markov Modeli (HMM) oluşturularak üzerinde Viterbi [55] algoritmasının çalıştırılmasıyla en yüksek olasılıklı sözcük dizisi elde edilebilir.



Şekil 6.3 : Örnek Bir Tümcenin HMM İle Çözümleme Süreci

Yukarıda Türkmençe bir tümce Türkçeye çevrilirken oluşturulan bir HMM örneği verilmiştir⁷. Şekil 6.3'teki özel simgeler <s> ile </s> sırasıyla tümce başını ve sonunu işaret eden simgelerdir. HMM'deki **gözlem olasılıkları** (observation probabilities) 1 seçilerek sadece durum geçiş olasılıklarının kullanılması sağlanmıştır [56]. Şekilde, durum geçişlerini gösteren oklara iliştilmiş olarak, sözcüklerin seçilen İDM'ye göre olasılıkları bulunmaktadır. Örneğin **"ne"** ile **"için"** durumları arasındaki ok, $P("için"|"ne")$ olasılığını, <s> ile **"ne"** arasındaki ok ise $P("ne" | <s>)$ olasılığını (tümcenin başında **"ne"** sözcük kökünün olma olasılığını) göstermektedir. Viterbi algoritması ile de bu HMM üzerinde en olası yol (path) bulunur. Bu yol üzerindeki sözcükler seçilerek oluşturulan tümce hedef dilde olasılığı en yüksek çeviridir.

Tablo 6.2'de, Şekil 6.3'te kurulan HMM üzerinden hesaplanan en olası 3 aday tümce gösterilmiştir. Farklı derecelerde kök dil modelleri kullanılarak aday tümcelerin olasılıkları hesaplanmıştır. Kalın harflerle yazılan tümce, doğru çeviriyi göstermektedir [57].

Türkçede **dil** ismi ile birlikte **söylemek** yerine daha çok **konusmak** eylemi kullanılmaktadır. Buna uygun olarak da tek başına sözcük sıklıklarına bakıldığında ($n=1$) **söylemek** eylemi **konusmak** eyleminden daha çok geçmesine karşın İDM derecesi arttıkça **konusmak** eylemini içeren tümcelerin olasılığının yükseldiği görülmektedir.

⁷ Örnek HMM, 1. ve 2. dereceden İDM için geçerlidir. Daha yüksek İDM için geçmiş sözcük sayıları birden fazla olduğu için gösterilen modelin yükseltilmesi gereklidir.

Tablo 6.2 : İDM İle En Olası Tümcenin Bulunması

İDM Derecesi	En Olası 3 Tümce	Sıra	Log. Olasılık
n=1	ne için insanlar türlü dillerde söylüyorlar	1	-17.2978
	ne için insanlar türlü dillerde konuşuyorlar	2	-17.5196
	ne için adamlar türlü dillerde söylüyorlar	3	-17.7816
n=2	ne için insanlar türlü dillerde konuşuyorlar	1	-18.1625
	ne için adamlar türlü dillerde konuşuyorlar	2	-18.3105
	kim için insanlar türlü dillerde konuşuyorlar	3	-18.6553
n=3	ne için insanlar türlü dillerde konuşuyorlar	1	-18.2265
	kim için insanlar türlü dillerde konuşuyorlar	2	-18.6196
	ne için adamlar türlü dillerde konuşuyorlar	3	-18.6294

6.5 Hedef Dilde Biçimbilimsel Üretici

İDM bileşeninin çıktısı, yüzeysel biçimdeki sözcükler yerine yapısal biçimdeki sözcüklerden oluşan bir tümcedir. Bu tümcede yer alan tüm sözcükler, hedef dile ilişkin bir biçimbilimsel üreticiden geçirilerek yüzeysel biçimler oluşturulmalı ve sistemin son çıktısı olan tümce üretilmelidir. Bu amaçla hedef dile ilişkin bir biçimbilimsel üreticiye gerek duyulmaktadır. Türkçe için geliştirilen biçimbilimsel çözümleyici, SDD olarak tasarlandığı için ters yönde çalıştırıldığında biçimbilimsel üretici olarak iş görmektedir [13]. Üstelik Türkçe için bu ters çalıştırma durumunda herhangi bir belirsizlik oluşmamaktadır. Yani yapısal biçimde bir sözcüğe karşılık, o sözcüğe ait sadece bir yüzeysel sözcük üretilmektedir ⁸. Türkçe dışındaki diğer Türk dilleri için bilinen bir biçimbilimsel üretici yoktur.

⁸ Çok ender bazı durumlarda birden fazla sonuç üretilmektedir:

aç+Verb+Neg+Imp+A2sg → (1) açma, (2) açmasana

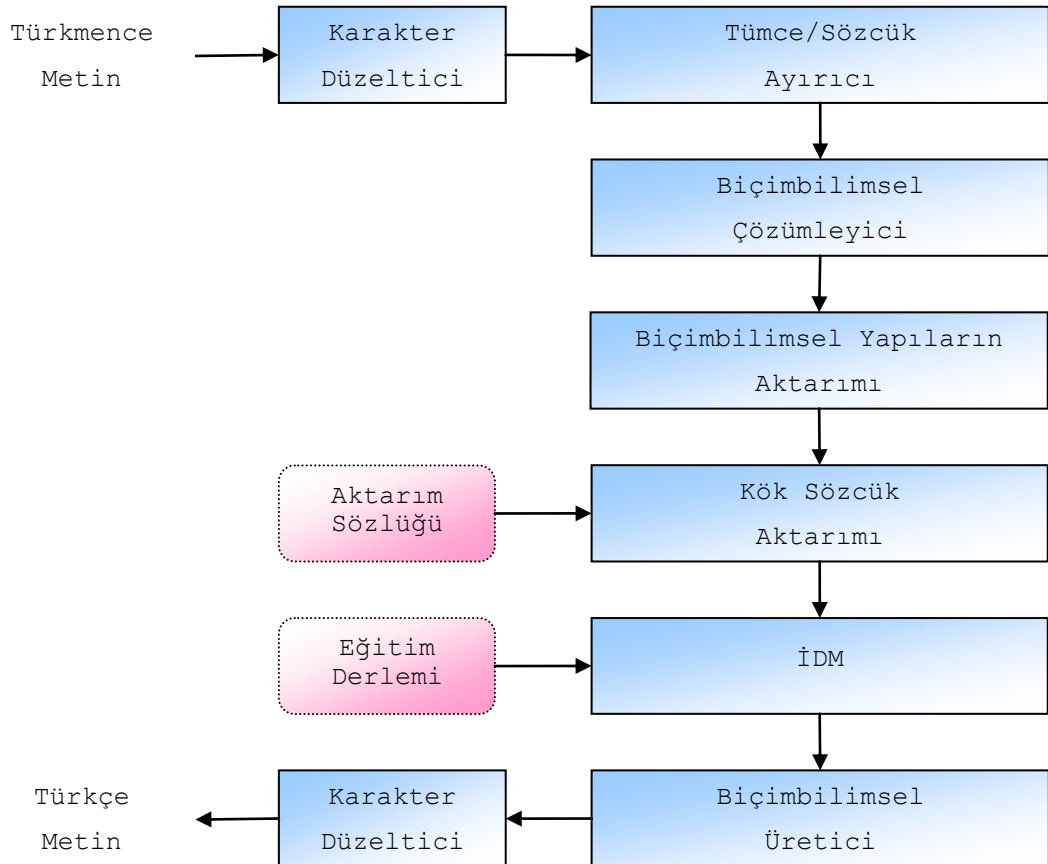
7. TÜRKMENCEDEN TÜRKÇEYE BİLGİSAYARLI ÇEVİRİ SİSTEMİ

Bir önceki bölümde ayrıntıları açıklanan çeviri modellerinin bir uygulaması olarak Türkmenceden Türkçeye bir çeviri sistemi tasarlanmış ve gerçekleştirilmiştir. Çeviri sistemi ilk olarak Model 0'a (temel model) uygun olarak gerçekleştirilmiş daha sonra ise bu temel model üzerine Model I ve Model II'nin getirdiği iyileştirmeler eklenmiştir.

Gerçeklenen uygulamada, aktarım fonksiyonunda yer alan bütün bileşenler SDD biçiminde tasarlanmıştır.

7.1 Aktarım Modeli 0 Gerçeklemesi

Aktarım Modeli 0'ı temel olarak tasarlanan Türkmenceden Türkçeye çeviri sisteminin bileşen şeması Şekil 7.1'de verilmiştir:



Şekil 7.1 : Aktarım Modeli 0 Temelinde Oluşturulan Sistemin Bileşenleri

7.1.1 Türkmence Biçimbilimsel Çözümleyicinin Geliştirilmesi

İki-düzeyle biçimbilimsel çözümleme ilkeleri esas alınarak Xerox sonlu durumlu araçlarıyla Türkmence için bir biçimbilimsel çözümleyici tasarlanmıştır [58]. Bu biçimbilimsel çözümleyicinin tasarım aşamaları aşağıdaki bölümlerde verilmiştir.

7.1.1.1 Türkmen Dilinin Biçimbilimsel Özellikleri

Türkmence dilinin biçimbilimsel yapısı Türkçe ile benzerlik göstermektedir. Özellikle isim çekimlerinde eklerin türleri ve geliş sıraları Türkçeye çok benzerdir. Bu benzerliklerden dolayı Türkçe için gerçekleştirilmiş olan biçimbilimsel çözümleyici [13] temel olarak alınmıştır.

Her ne kadar Türkmence, Türkçeye en çok benzeyen dillerden birisi olsa da, iki dil arasında harfler, ses olayları, sözcük çekimleri ve anlamsal açıdan pek çok farklılıklar bulunmaktadır [59-64]. Türkmencenin Türkçeye benzerliği ilk bakışta yararlı görülse de, bazı açılardan zararlı olmaktadır. Örneğin Türkçe ile birebir aynı olan bazı sözcükler ya da ekler, Türkmencede farklı anlamlara gelmektedir. Türkçe bilen birisi, Türkmence öğrenirken bu tür yanılgılara kolayca düşebilmektedir.

Ses Olayları

Türkçeye en yakın dillerden biri olsa da Türkmencede yazı dili ile konuşma dili arasında büyük farklılıklar bulunmaktadır. Aslında Türkçede de bütün sözcükler, yazıldığı gibi okunmaz ama Türkmencede bu durum istisna olmaktan çıkmış ve çok sık karşılaşılan bir durum olmuştur. Kısıtlı sayıdaki Türkmence dilbilgisi kaynaklarının çoğunda, yazı dili ile konuşma dili arasındaki ayrım net olarak verilmemiştir. Bu nedenle bazı kuralların sadece konuşma dili için geçerli olduğunu ortaya çıkarmak oldukça zahmetli olmuştur.

Türkmencede sözcüklerin okunuşları ile yazılışları arasında Türkçenin tersine çok büyük farklılıklar bulunmaktadır. Bütün seslilerin kısa ve uzun okunuşları bulunmaktadır. Ancak yazı dilinde herhangi bir seslinin kısa mı uzun mu okunacağına ilişkin bir işaret yoktur. Aşağıda bu konu ile ilgili örnekler verilmiştir (uzun okunan sesliler, “:” işaretiyle belirtilmiştir):

Uzun Okunuş

at (a:t) ad, isim

ot (o:t) ateş

daş(da:ş) taş

Kısa Okunuş

at (at) at

ot (ot) ot

daş (daş) dış

Türkmencede büyük ünlü uyumu vardır. Sözcüklerin bazıları küçük ünlü uyumuna uyarken bazıları da uymaz. Türkçede geniş-yuvarlak seslilerden (**o, ö**) sonra dar-yuvarlak (**u, ü**) ya da geniş-düz (**a, e**) seslileri gelir. Türkmencede ise **o, ö** seslilerinden sonra dar-düz sesliler (**y, i**) gelir. Bu nedenle Türkmencede bazı sözcükler küçük ünlü uyumuna uymaz.

Türkmencede de Türkçede olduğu gibi sessiz yumuşaması vardır. Sözcük sonundaki **p, ç, t, k** sessizleri, sesli ile başlayan bir ek aldıklarında **b, c, d, g** harflerine dönüşürler. Sessiz benzeşmesi de kısmen görülür.

Sesli düşmesi kuralı ise Türkmencede daha kurallıdır. Bir seslinin düşmesi için:

1. iki heceli sözcük olmalı
2. ilk V kısa, hece açık olmalı (V, CV)
3. ikinci hece kapalı olmalı (CVC)

Ancak ne yazık ki 2. maddede söylenin V'nin yani seslinin kısa olması, yazı dilinde belirtilmemektedir.

Okunuşlarla ilgili bir çok kurala, biçimbilimsel çözümleyicinin geliştirilmesi ile ilgisi olmadığı için burada yer verilmeyecektir.

Tekillik / Çoğulluk

Çoğulluk ekleri **+lar** ve **+ler** Türkçedeki ile aynı şekilde kullanılır.

Belirtme Durumu

Belirtme durumu eki Türkçedekinden farklı olarak sadece **+y** ve **+i** eklerinden oluşmaktadır. Ancak Türkçede ekten önce sesli harf bulunursa araya **n** harfi yerine **y** harfi gelir:

kitab**y** (kitabı)

gol**y** (kolu)

baş**y** (başı)

göz**i** (gözü)

gül**i** (gülü)

Yönelme Durumu

Türkçedeki gibi **+a** ve **+e** ekleri ile kurulur.

depdere (deftere) göze (göze) bilbile (bülbüle)

Sesli ile biten isimlerde, yaklaşma durumu eki (+*a*, +*e*, +*ü*) farklılık göstermektedir.

- i) –*a*, –*o* ile biten sözcüklere yaklaşma durumu eki eklenmez, yaklaşma durumu sadece sözcüğün sonunda seslinin uzun okunuşu ile belirtilir.

ata (ata) baba

ata (ata:) babaya

- ii) –*i*, –*e*, –*ü* ile biten sözcüklere yaklaşma durumu eki geldiğinde, sözcüğün son seslisi –*ü* olarak değişir.

Berdi ⇒ *Berdä*

Berdi ⇒ *Berdi'ye*

- iii) –*y* seslisi ile biten sözcüklere yaklaşma durumu eki geldiğinde, sözcüğün son seslisi –*a* seslisi olarak değişir.

Mary ⇒ *Mara*

Marı ⇒ *Marı'ya*

Kalma Durumu

Türkçedekine benzer olarak +*da*, +*de* ekleri ile kurulur. Türkçedeki kullanımdan tek farklı bu eklerin ünsüz benzeşmesine uymamasıdır (+*ta*, +*te* halleri yoktur):

guşda (kuşta)

altda (altta)

kitapda (kitapta)

Kalma durumundan sonra –*ki* eki gelirse, kalma durumu ekindeki sesli uzar:

depderdaki (defterdeki)

bizdaki (bizdeki)

süýtdaki (sütteki)

kitapdaki (kitapdaki)

adamdaki (adamdaki)

ondaki (ondaki)

Çıkma Durumu

Çıkma durumu, +*dan* ve +*den* ekleri ile kurulur. Bu ekler de ünsüz benzeşmesine uymazlar:

öýden (evden)

altdan (alttan)

kitapdan (kitaptan)

Tamlayan Durumu

Tamlayan durumu ekleri +*yñ*, +*iñ*, +*uñ* ve +*üñ* ekleridir:

golun (kolun)

burnynyñ (burnunun)

diliñ (dilin)

İsim çekimi ve eylem çekimi açısından incelendiğinde, Türkmençe, isim çekimi yönünden Türkçeye çok benzerken, eylem çekiminde ciddi farklılıklar vardır. Özellikle çatı kurulumu Türkmençede Türkçeye oranla çok daha karmaşıktır. Türkçede çatı kurulumu genelde aşağıdaki gibidir:

görmek	(Yalın)
görüşmek	(İşteş)
görüştürmek	(İşteş-Ettirgen)
görüştürülmek	(İşteş-Ettirgen-Edilgen)

Sadece bazı durumlarda ettirgenlik birkaç defa tekrarlanabilir.

Türkmençede ise çatı eklerinin geliş sırası çok daha karışıktır:

Tablo 7.1 : Türkmençede Çatı Eklerinin Sıralanışı

İki Çatı Ekli Eylemler	Üç Çatı Ekli Eylemler	Dört Çatı Ekli Eylemler
Ettirgen + Dönüşlü	Dönüşlü + Ettirgen + Edilgen	Dönüşlü + İşteş + Ettirgen + Edilgen
Dönüşlü + Edilgen	İşteş + Ettirgen + Edilgen	Ettirgen + Dönüşlü + Ettirgen + Edilgen
İşteş + Edilgen	Ettirgen + Edilgen + İşteş	
Ettirgen + Edilgen	Dönüşlü + Ettirgen + İşteş	
Dönüşlü + İşteş	Dönüşlü + İşteş + Ettirgen	
Edilgen + İşteş	Edilgen + İşteş + Ettirgen	
Ettirgen + İşteş	Ettirgen + İşteş + Ettirgen	
İşteş + İşteş	Dönüşlü + Ettirgen +	
Dönüşlü + Ettirgen	Ettirgen	
İşteş + Ettirgen		
Ettirgen + Ettirgen		

Ayrıca Türkmençede bazı eylem kipleri çekilmezler. Örneğin gelecek zamanı belirten **+jek / +jak** ekinden sonra kişi çekim eki gelmez. Örnek kullanımı aşağıdaki gibidir:

Men	geljek	(geleceğim)
Sen	geljek	(geleceksin)
O	geljek	([o] gelecek)

Ayrıca gereklilik kipi **+malı / +meli** de benzer şekilde kişi eki almaz.

Ancak bu kiplere kesinlik anlamı katan **+dyr** eki geldiğinde, bu **+dyr** ekinden sonra kişi çekim ekleri gelebilir.

Men	geljekdirin.	(geleceğimdir)
-----	--------------	----------------

Türkmencede, Türkçede bulunmayan bazı kipler de vardır. Örnek olarak bir iş için hazırlık yapıldığını ya da o işin yapılmasının düşünüldüğünü gösteren **+mekçi / +makçy** eki bulunmaktadır. Bu ek de istisna olarak çekim eki almayan kipler grubundandır.

Belirsiz geçmiş zaman eki olarak kullanılan **+mış / +miş** eki Türkmencede ilk zaman olamaz. Bunun yerine **+ypdy / +ipdi / +updu / +üpdü** ekleri gelmektedir. Ancak ikinci zaman olarak **+mış / +miş** eki gelebilmektedir.

Türkçede geniş zaman olarak kullanılan **+ar / +er** eki, Türkmencede gelecek zaman anlamını taşımaktadır.

Gene Türkçedekine benzer şekilde geniş zamanın 3. tekil şahsının olumsuzu farklıdır. Ancak daha büyük bir farklılık olarak bazı kiplerde olumsuzluk eki olarak **+ma / +me** gelmemekte bunun yerine eylemden sonra **däl** (“değil”) getirilmektedir. Örnek:

Men gelcek däl

Kimi durumlarda **däl** eylemi de çekime uğramaktadır.

Biçimbilimsel Çözümleyicinin Gerçeklenmesi

Türkmençe için biçimbilimsel çözümleyici geliştirirken iki düzeyli biçimbilimsel çözümleme yöntemi benimsenmiş ve *XEROX*’un sonlu durumlu araçlarından yararlanılmıştır [23]. Öncelikle kökler ve eklerle ilgili durum geçişleri yani morfolojik kurallar tasarlanmış ve *LEXC* aracılığıyla gerçekleştirilmiş, daha sonra iki-düzeyle kurallar *TWOLC* ile oluşturulmuştur. Ek olarak, bazı geçersiz durumların elenmesi için *XFST* ortamında kurallar yazılmış ve elde edilen bu üç SDD birleştirilerek tek bir SDD elde edilmiştir. Oluşan bu çözümleyici, ters yönde çalıştırıldığı zaman üretici olarak da çalışabilmektedir.

7.1.1.2 İki Düzeyli Kurallar

Türkmencedeki çeşitli ses olaylarını ve değişimlerini gerçeklemek için bir dizi iki-düzeyle kural tanımlanmış ve *TWOLC* derleyicisi yardımı ile bu kuralları gerçekleyen bir SDD oluşturulmuştur.

İki düzeyli kuralları tanımlamadan önce, bu kuralların üzerinde işlem göreceği abecenin tanımlanması gerekmektedir. Bu abece güncel Türkmen harfleri ile sadece ara aşamalarda kullanılan ve yazıda görünmeyen bazı ek karakterler içermektedir.

Her ne kadar *TWOLC* derleyicisi UTF-8 karakter kümesini destekleyerek standart olmayan ASCII karakterlerinin kullanımına izin verse de, bu tür bir kullanımda hata ayıklama ve komut satırından sınamaların yapılması olanaksız olmaktadır. Bu nedenle standart ASCII tablosunda olmayan karakterler için bir ASCII karakteri, Tablo 7.2'deki gibi seçilmiş ve kurallarda bu şekilde gösterilmiştir.

Tablo 7.2 : ASCII Olmayan Karakterler Yerine Kullanılan Karşılıklar

ASCII dışı karakterler	ü	ö	Ç	ñ	ş	ý	ž	ä
Seçilen ASCII karşılık	<i>U</i>	<i>O</i>	<i>C</i>	<i>N</i>	<i>S</i>	<i>Y</i>	<i>Z</i>	<i>E</i>

Ayrıca kuralların üzerinde işlem göreceği bir takım kümeler de tanımlanmıştır:

! Sessiz harfler

CONS = b C d f g h j Z k l m n N p r s S t w Y z;

! Sesli harfler

VOWEL = a E e y i o O u U;

! Kalın sesliler

BACKV = a y u o;

! İnce sesliler

FRONTV = e E i O U;

! Kalın-yuvarlak sesliler

BKROV = o u;

! Kalın-düz sesliler

BKUNROV = a y;

! İnce-yuvarlak sesliler

FRROV = O U;

! İnce-düz sesliler

FRUNROV = e i E;

! Bazı durumlarda silinen harfler (kaynaştırma gibi)

X = s M n %\$;

! Sesli düşmesine uğrayabilecek harfler (burun->burnu gibi)
VS = y i O o U u;

Sesli düşmesi, sessiz yumuşaması gibi ses olaylarını gerçekleyebilmek için çeşitli Türkmençe kaynaklarından çıkartılan bir dizi kural hazırlanmıştır [59, 62, 64].

1. A:a => [:BACKV] [:CONS]* (%':%') (%+:0) [CONS: | :CONS | :0]* _
2. A:e => [:FRONTV] [:CONS]* (%':%') (%+:0) [CONS: | :CONS | :0]* _

Bu iki kural ile eklerde **A** ile belirtilen sesli harfin, ünlü uyumuna göre **a** ya da **e** harfine dönüştürülmesi sağlanmıştır.

Yapısal biçim : galam+Noun+A3pl+Pnon+Nom
Ara biçim : galam+lAr
Yüzeysel biçim : galam0lar
galamlar (kalemler)

Yapısal biçim : depder+Noun+A3pl+Pnon+Nom
Ara biçim : depder+lAr
Yüzeysel biçim : depder0ler
depderler (defterler)

3. V:a => [:BACKV] [CONS]* (%+:0) [CONS: | :CONS | :0]* _
4. V:E => [:FRONTV] [CONS]* (%+:0) [CONS: | :CONS | :0]* _

Bu iki kural ile eklerde **V** ile belirtilen sesli harfin, ünlü uyumuna göre **a** ya da **E (ä)** harfine dönüştürülmesi sağlanmıştır.

Yapısal biçim : bar+Verb+Pos+Prog2+Anon
Ara biçim : bar+YVn
Yüzeysel biçim : bar0Yan
barýan (varmakta olan)

Yapısal biçim : geple+Verb+Pos+Prog2+Anon
Ara biçim : geple+YVn
Yüzeysel biçim : geple0YEn
gepleýän (konuşmakta olan)

5. I:y => [:BACKV] [CONS]* (%':%') (%+:0) [CONS: | :CONS | :0]* _
6. I:i => [:FRONTV] [CONS]* (%':%') (%+:0) [CONS: | :CONS | :0]* _

Bu iki kural ile eklerde **I** ile belirtilen sesli harfin, ünlü uyumuna göre **y** ya da **i** harfine dönüştürülmesi sağlanmıştır.

Yapısal biçim : gol+Noun+A3sg+Plpl+Nom
 Ara biçim : gol+HmIz
 Yüzeysel biçim : gol0umyz
 golumyz (kolumuz)

Yapısal biçim : Col+Noun+A3sg+Plpl+Nom
 Ara biçim : Col+HmIz
 Yüzeysel biçim : Col0Umiz
 çölümüz (çölümüz)

7. H:i => [:FRUNROV] [CONS]* (%)'(:%) (%)+:0 [CONS: | :CONS | :0]* _
 8. H:y => [:BKUNROV] [CONS]* (%)'(:%) (%)+:0 [CONS: | :CONS | :0]* _
 9. H:u => [:BKROV] [CONS]* (%)'(:%) (%)+:0 [CONS: | :CONS | :0]* _
 10. H:U => [:FRROV] [CONS]* (%)'(:%) (%)+:0 [CONS: | :CONS | :0]* _

Bu kurallar ile eklerde **H** ile belirtilen sesli harfin, kendinden önce gelen seslinin incelik-kalınlık ve yuvarlaklık-düzlük özelliklerine göre **i, y, u** ya da **U** harflerinden birine dönüştürülmesi sağlanmıştır.

Yapısal biçim : dogan+Noun+A3sg+Plpl+Nom
 Ara biçim : dogan+HmIz
 Yüzeysel biçim : dogan0ymyz
 doganymyz (kardeşimiz)

Yapısal biçim : depder+Noun+A3sg+Plpl+Nom
 Ara biçim : depder +HmIz
 Yüzeysel biçim : depder0imiz
 depderimiz (defterimiz)

Yapısal biçim : suw+Noun+A3sg+Plpl+Nom
 Ara biçim : suw+HmIz
 Yüzeysel biçim : suw0umyz
 suwumyz (suyumuz)

Yapısal biçim : sOz+Noun+A3sg+Plpl+Nom
 Ara biçim : sOz+HmIz
 Yüzeysel biçim : sOz0Umiz
 sözümüz (sözümüz)

11. H:0 <=> [:VOWEL] %+:0 _

Eğer **H** ile başlayan bir ek, sesli ile biten bir gövdeye eklenirse, bu **H** sesi silinir.

Yapısal biçim : kEdi+Noun+A3sg+Plsg+Nom
Ara biçim : kEdi+Hm
Yüzeysel biçim : kEdi00m
kädim (kabağım)

12. T:a <=> [:BACKV] [CONS: | :CONS | :0]+ [%+:0] _
13. T:e <=> [:FRONTV] [CONS: | :CONS | :0]+ [%+:0] _
14. Cx:Cy <=> _%+:0 [T:] where Cx in (i e A y)
Cy in (E E E a) matched

Türkmencede yönelme durumu (dative case), sıra dışı bir durumdur. Eğer son hecesinde kalın sesli içeren ve sessiz ile biten bir gövdeye eklenirse yönelme durumu eki **a**, son hecesinde ince sesli içeren ve sessiz ile biten bir gövdeye eklenirse yönelme durumu eki **e** olur. Eğer eklendiği gövde sesli ile bitiyorsa, yönelme hali eki, gövdenin son seslisini (**i⇒E**, **e⇒E**, **A⇒E**, **y⇒a** şeklinde) değiştirir.

Yapısal biçim : Berdi+Noun+A3sg+Pnon+Dat
Ara biçim : Berdi+T
Yüzeysel biçim : BerdE00
Berdä (Berdi'ye)

Yapısal biçim : Mary+Noun+A3sg+Pnon+Dat
Ara biçim : Mary+T
Yüzeysel biçim : Mara00
Mara (Marı'ya)

Yapısal biçim : ata+Noun+A3sg+Pnon+Dat
Ara biçim : ata+T
Yüzeysel biçim : ata00
ata (babaya)

Son örnekten de görüldüğü gibi Türkmencede, sonu **a** ile biten gövdelerin yönelme durumu ile yalın durumu yazılış olarak aynıdır, farklılık okunuşta belli edilir.

15. e:E <=> _ %+:0 [H:] [m: | N: | p:] (I: z:); _ %+:0 [H:] b e r;

Bu kural, **e** harfi ile biten bir gövdeye bazı ekler geldiğinde, gövdenin sonundaki **e** harfinin yumuşayarak **E** sesine dönüşmesini sağlar.

Yapısal biçim : iSle+Verb+Pos+Past+A3sg
Ara biçim : iSle+Hpdi
Yüzeysel biçim : iSlE00pdi
işläpdi (istemişti)

16. Cx:Cy <=> [:CONS] %+:0 _ (CONS VOWEL); where Cx in (s n S Y)
Cy in (0 0 s 0) matched;

Gövdenin sonunda sessiz varsa, gövdeye gelecek ekin başındaki **s**, **n**, **Y** harfleri silinir, **S** ise silinmeyerek **s**'ye dönüştürülür (kaynaştırma harfi olarak kullanılmayan ve silinmemesi gereken **s** anlamında kullanılmaktadır).

17. A:E => %+:0 m _ [k:] %+:0 [T:]; %+:0 d _ %+:0 k [I:];

İki özel durumda eklerdeki **A** harfi **E** olarak çözümlenir. Bunlardan bir tanesi, eylem kökünün **+mAk** mastar eki ile isim haline dönüşerek yönelme durumda kullanıldığı durumdur. İkincisi ise isim çekiminde bulunma hali eki **+dA**'dan sonra **+kI** ekinin geldiği durumdur. Her iki durumda da **A** harfi **E** olarak çözümlenir.

Yapısal biçim : ber+Verb+Pos^DB+Noun+Infl+A3sg+Pnon+Dat
Ara biçim : ber+mAk+T
Yüzeysel biçim : ber0mEg0e
bermäge (varmaya)

Yapısal biçim : galam+Noun+A3sg+Pnon+Dat
Ara biçim : galam+dA+kI
Yüzeysel biçim : galam0dE0ki
galamdäki (kalemdeki)

18. Cx:Cy => _ %+:0 (X:0) [:VOWEL]; where Cx in (p t C k)
Cy in (b d j g) matched;

Ünsüz yumuşamasını gerçekleyen bu kural, sert sessizler **p**, **t**, **C** ve **k** ile biten gövdelere sesli ile başlayan bir ek gelirse, bu sert sessizlerin yumuşamasını sağlar.

Yapısal biçim : kitap+Noun+A3sg+Pnon+Dat
Ara biçim : kitap+T
Yüzeysel biçim : kitab0a
kitaba (kitaba)

19. VS:0 <=> %\$:0 _ [CONS %+:0 (X:0) [A: | H: | I: | T:]];

20. H:0 <=> %\$:0 _ [CONS %+:0 (X:0) [A: | H: | I: | T:]];

Bazı sözcüklerin köklerindeki bazı harfler, belirli ekler geldiğinde düşerler. Bunun belirli bir kuralı olmadığı için, düşmesi olası bu harflerden önce alt sözlüklerde \$ işareti konularak bu durum belirtilir. Yüzeysel biçime dönüştürülürken \$ işareti silindiği için yüzeysel biçimde \$ işareti görünmemektedir.

7.1.1.3 Morfotaktik Kurallar

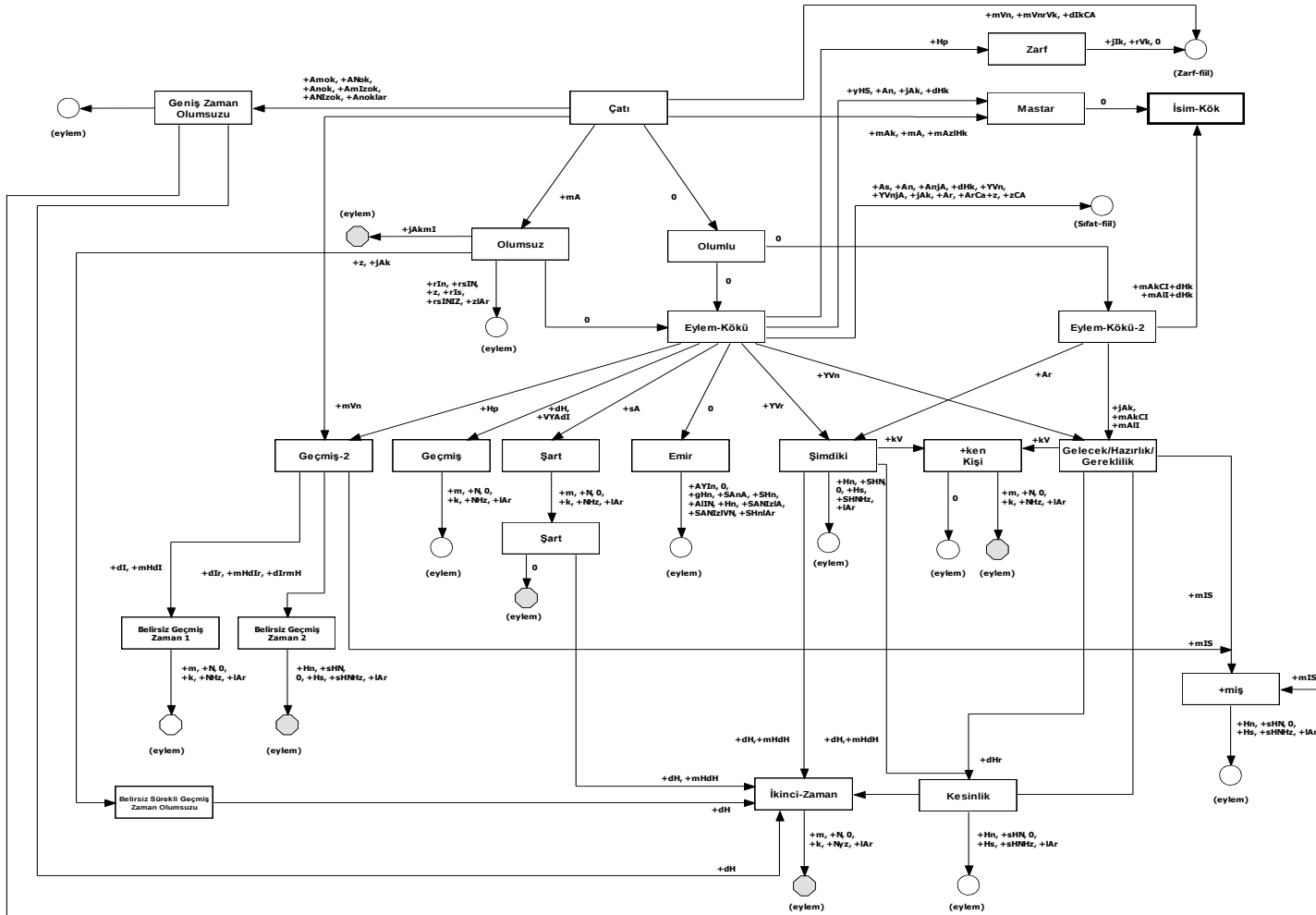
Morfotaktik kurallar, geçerli bir sözcük oluşturmak için eklerin geliş sıralarını inceleyen ve modelleyen kurallardır [24]. Morfotaktik kurallar da, tıpkı iki-düzeyle kurallar gibi SDD'ler ile gerçekleşmiştir. Oluşturulan sonlu durum makineleri Şekil 7.2, Şekil 7.3 ve Şekil 7.4'te verilmiştir. Bu şekillerde kutular durumları, oklar ise durum geçişlerini göstermektedir. Durum geçişleri sadece okların üzerindeki eklerden bir tanesi geldiğinde yapılmaktadır. Durum geçişlerindeki “0” ifadesi, herhangi bir ek gelmeden yapılabilen boş geçişi belirtmektedir. Şekildeki yuvarlaklar, geçerli bir sözcük olduğu zaman ulaşılan bitiş durumlarıdır. Oluşan sözcüğün etkin sözcük sınıfı, bitiş durumlarının yanında parantez içerisinde belirtilmiştir. Tasarımı yapılan bu sonlu durumlu makinelerin oluşturulması için XEROX LEXC aracı kullanılmıştır. Buna uygun olarak, durumlar için alt sözlükler tanımlanmıştır. Her alt sözlük, iki sütundan oluşmaktadır. İlk sütunda birbirinden : işareti ile ayrılmış biçimde sözcük kökünün ya da ekin yüzeysel ve yapısal biçimleri bulunur. Eğer her iki biçimin yazılışı da aynı ise araya : işareti konulmadan bir defada yazılır (*geple:geple* şeklinde yazılmadan sadece *geple* yazılır). İkinci sütunda ise bir sonraki durum yani alt sözlük belirtilir. Geliştirilen sonlu durumlu dönüştürülerden alınan bir kesit aşağıda verilmiştir:

```
LEXICON VERBS
diy          VERB-POST;
bil          VERB-POST;
geple        VERB-POST;
...

LEXICON VERB-POST
+Verb : 0      VERB-ROOT;

LEXICON VERB-ROOT
0      : 0      VERB-PASSIVE;
+Recip : +nHS    VERB-RECIP;
+Recip : +S      VERB-RECIP;
...

LEXICON VERB-PASSIVE
0 : 0
^DB+Verb+Hastily : +Hber    VERB-MAIN;
^DB+Verb+Recip   : +nHS     VERB-RECIP;
```

Şekil 7.4 : Eylem Soylu Sözcükler İçin Morfotaktik Kuralların SDMSi (Devamı)

7.1.2 Kök Sözcük Aktarım Kuralları

Biçimbilimsel çözümlemesi yapılmış Türkmençe sözcük köklerinin Türkçeye aktarılmasını sağlayan kurallar, SDD’ler ile gerçekleştirilmiştir. Örnek bir aktarım kuralı aşağıda verilmiştir:

```
"tatlı" <- "Yakymly"
```

Bölüm 6.2’de belirtildiği gibi, bu aktarım kurallarında sözcük türlerinin kullanılması, sözcüksel belirsizliği azaltmaktadır. Yazılan kurallar bu ilke çerçevesinde oluşturulmuş ve kuralların sağ bağlamları sözcük türleri ile kısıtlandırılmıştır:

```
"gri" <- "boz" \/_ "+Adj" .o.  
"sil" <- "boz" \/_ "+Verb"
```

Bu sayede sistemin rastladığı bütün “**boz**” köklerini, “**gri**” ve “**sil**” kökleri ile değiştirmesinin önüne geçilerek, aktarılacak sözcüğün sıfat ya da eylem olma durumuna göre sadece uygun karşılıkların dönüştürülmesi sağlanmıştır. Kök aktarım bileşenin örnek girdisi ve çıktısı aşağıdaki şekilde verilmiştir:



Şekil 7.5 : Kök Aktarım Bileşeni

7.1.2.1 Birden Fazla Sözcükten Oluşan Karşılıklar

Dillerin doğası gereği, Türkmencede bir tek sözcükle ifade edilen bazı kavramlar Türkçede bir tek sözcük ile ifade edilememekte, ancak iki ya da daha fazla sözcükten oluşan ÇSG’ler ile ifade edilebilmektedir. Bu durumda kök değiştirmek yerine daha akıllı bir yöntemle başvurulması gereklidir. Bu tür durumlara örnek olarak aşağıdaki sözlük girdileri gösterilebilir:

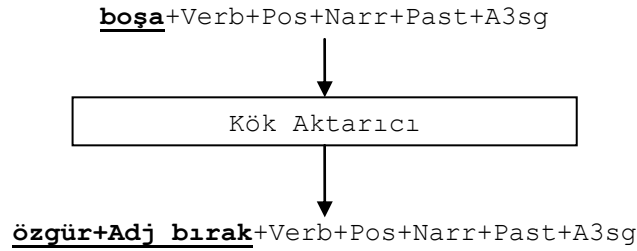
Türkmençe	Türkçe
boşatmak	özgür bırakmak
dillenmek	dile gelmek
entegem	uzun süre

Hedef dil karşılığı ÇSG olan sözlük girdileri için standart kök aktarım kuralları yerine gelişmiş kuralların oluşturulması gereklidir. Önemli olan bir diğer nokta da, bu değiştirme sürecinde, ÇSG'nin son sözcüğü hariç bütün sözcüklerin yapısal biçimde olması zorunluluğudur. Bu, hedef dilde üretilecek tüm sözcüklerin biçimbilimsel özelliklerinin de bulunması zorunluluğu anlamına gelmektedir çünkü sistemin diğer bileşenleri yürütülürken, sözcüklerin yapısal biçimlerine gerek duymaktadır.

Türkçede ÇSG'lerin türetme ve/veya çekim eklerinden etkilenen kısmı sadece ÇSG'nin sonunda yer alan sözcüktür [46]. Bu gerçekten hareketle, kaynak dildeki sözcüğe ait biçimbilimsel özelliklerin, hedef dildeki ÇSG'nin sonundaki sözcüğe ait olduğu, ÇSG'nin başında yer alan diğer sözcüklerin sabit bir yapıya sahip olduğu sonucuna varılabilir. Bu koşullarla, yukarıdaki sözlük girdilerini aktarmak üzere oluşturulması gereken kurallar aşağıda verilmiştir:

```
"dil+Noun+A3sg+Pnon+Dat gel" <- "dillen"  
"özgür+Adj bırak" <- "boSat"  
"uzun+Adj süre+Noun+A3sg+Pnon+Nom"<-"entegem+Adverb"
```

Aşağıdaki şekilde ise kök aktarıcının örnek bir ÇSG'yi aktarması gösterilmiştir. Altı çizili olmayan Türkmence biçimbilimsel yapıların, ÇSG'nin son sözcüğüne eklendiği görülmektedir.



Şekil 7.6 : ÇSG'lerin Aktarılması

7.1.2.2 Sözcüksel Aktarım Kuralları

Uygulamada ortaya çıkan bazı durumlar göstermiştir ki bir takım sözcükler için sadece sözcük kökünü değiştiren basit bir kural yeterli olmamaktadır. Örneğin Türkmencedeki **ulumsy** sözcüğü Türkçedeki **kibirli** sözcüğünün karşılığıdır. Standart kurallar uygulanarak sadece sözcük kökü değiştirildiğinde aşağıdaki dönüştürme işlemi gerçekleşir:

kibirli+Adj ← ulumsy+Adj

İlk bakışta göze çarpan herhangi bir sorun olmamasına karşın, oluşan yapısal biçimdeki sözcük, Türkçe biçimbilimsel üretici tarafından yüzeysel biçime dönüştürüleceği zaman herhangi bir çıktı üretilmemektedir. Bunun altında yatan neden ise, Türkçedeki **kibirli** sözcüğünün aslında türemiş bir sözcük olması ve bu sözcüğün doğru yapısal biçiminin aşağıdaki gibi olmasıdır:

kibir+Noun+A3sg+Pnon+Nom^DB+Adj+With

Ortaya çıkan bu sorunun düzeltilmesi için, Türkmencedeki **ulumsy** sözcüğü için aşağıdaki gibi özel bir kural oluşturulmalıdır:

"kibir+Noun+A3sg+Pnon+Nom^DB+Adj+With"<-"ulumsy+Adj"

Örnekte açıklandığı gibi sözcüğe bağlı özel durumları kotaran kurallar, **sözcüksel kurallar** (lexicalized rules) olarak adlandırılmıştır.

Ancak her iki dilde de ortak olan türetme ekleri ile türetilebilecek sözcükler için ayrı kuralların oluşturulmasına gerek yoktur. Örneğin Türkmencedeki **+lyk** eki ile Türkçedeki **+lık** eki, sıfattan isim yapan aynı göreve sahip iki yapım ekidir. Dolayısı ile Türkmencede bulunan **ulumsylyk** sözcüğünün karşılığı da **kibirlilik** sözcüğüdür. Her iki sözcüğün biçimbilimsel çözümlemesi aşağıda belirtilmiştir:

ulumsy+Adj^DB+Noun+Ness+A3sg+Pnon+Nom

kibir+Noun+A3sg+Pnon+Nom^DB+Adj+With^DB+Noun+Ness+A3sg+Pnon+Nom

Örnekten de görüldüğü gibi, kalın ve altı çizili olarak gösterilmeyen biçimbilimsel yapılar aynıdır. Dolayısı ile bu iki sözcük için ayrı bir sözcüksel aktarım kuralı hazırlanmasına gerek yoktur, yukarıda anlatılan ve **ulumsy** sözcüğünü aktaran sözcüksel aktarım kuralının çalışması yeterli olmaktadır.

7.1.3 Biçimbilimsel Aktarım Kuralları

Türkmence ve Türkçe arasındaki biçimbilimsel farklılıkların giderilerek Türkmence biçimbilimsel çözümleme sonucu üretilen yapıların, kabul edilebilir Türkçe biçimbilimsel yapılara dönüştürülmesini sağlayan kurallardır.

Örneğin Türkmencede bulunan ve emir kipinin 1. tekil ve 1. çoğul kişiler için çekimi, Türkçede istek kipine karşılık gelmektedir:

Türkmence	Türkçe Karşılığı
alaYyn (al+Verb+Pos+Imp+A1sg)	alayım (al+Verb+Pos+Opt+A1sg)
algyn (al+Verb+Pos+Imp+A2sg)	al (al+Verb+Pos+Imp+A2sg)
alsyn (al+Verb+Pos+Imp+A3sg)	alsın (al+Verb+Pos+Imp+A3sg)

Bu deęişiklięi saęlamak üzere ařaęıdaki kural geliřtirilmiřtir:

```
define FromImpToOptA1sgA1pl "+Opt+A1sg" <- "+Imp+A1sg" .o.  
"+Opt+A1pl" <- "+Imp+A1pl";
```

Her iki dil arasındaki biçimbilimsel farklılıklardan bir tanesi de Türkmencede olup da Türkçede olmayan eylem kipleridir. Örneęin Türkmencede “+makçy/+mekçi” eki ile kiřinin, ekin geldięi eylemi yapmayı düřündüęü veya niyetlendięi anlamı kurulur. Bunun Türkçede doğrudan karřılıęı olmadıęı için ÇSG üreten bir kural geliřtirilmiřtir:

```
define mAkCI "^DB+Noun+Infl+A3sg+Pnon+Nom iste+Verb+Pos+Progl+A3sg"  
<- "+Think+Anon";
```

Bu gibi biçimbilimsel farklılıkları gidermek üzere toplam 24 kural kullanılmıřtır.

7.1.4 İstatistiksel Dil Modeli Bileřeni

Aktarım sırasında ortaya çıkan biçimbilimsel ve sözcüksel belirsizliklerin giderilmesi için İDM’leri kullanan bu bileřenin görevi ve iřleyiř tarzı, Bölüm 4.1’de açıklanmıřtır. Bu amaçla, bitiřken diller için Bölüm 4.3’te önerilen farklı türlerde İDM’ler üretilmiřtir. İDM’lerin oluřturulması için yaygın olarak kullanılan iki farklı yardımcı araç bulunmaktadır: CMU-Cambridge Language Modeling Toolkit [65] ve SRILM [66]. Bu çalışmada kullanılan İDM’ler, En Büyük Olabilirlik Kestirimi yöntemi ile SRILM [66] kullanılarak oluřturulmuřtur. Olasılıklar oluřturulurken yumuřatma için Good-Turing [29] yöntemi ile derece-düřürme modelleme [30] yöntemi beraber kullanılmıřtır.

Uygulamada önerilen farklı İDM tiplerinin başarımları ayrı ayrı incelenmiř ve en başarılı sonuç üreten İDM belirlenmeye çalışılmıřtır. Önerilen İDM tiplerinin başarımların sonuçları Bölüm 8.3.4’te verilmiřtir.

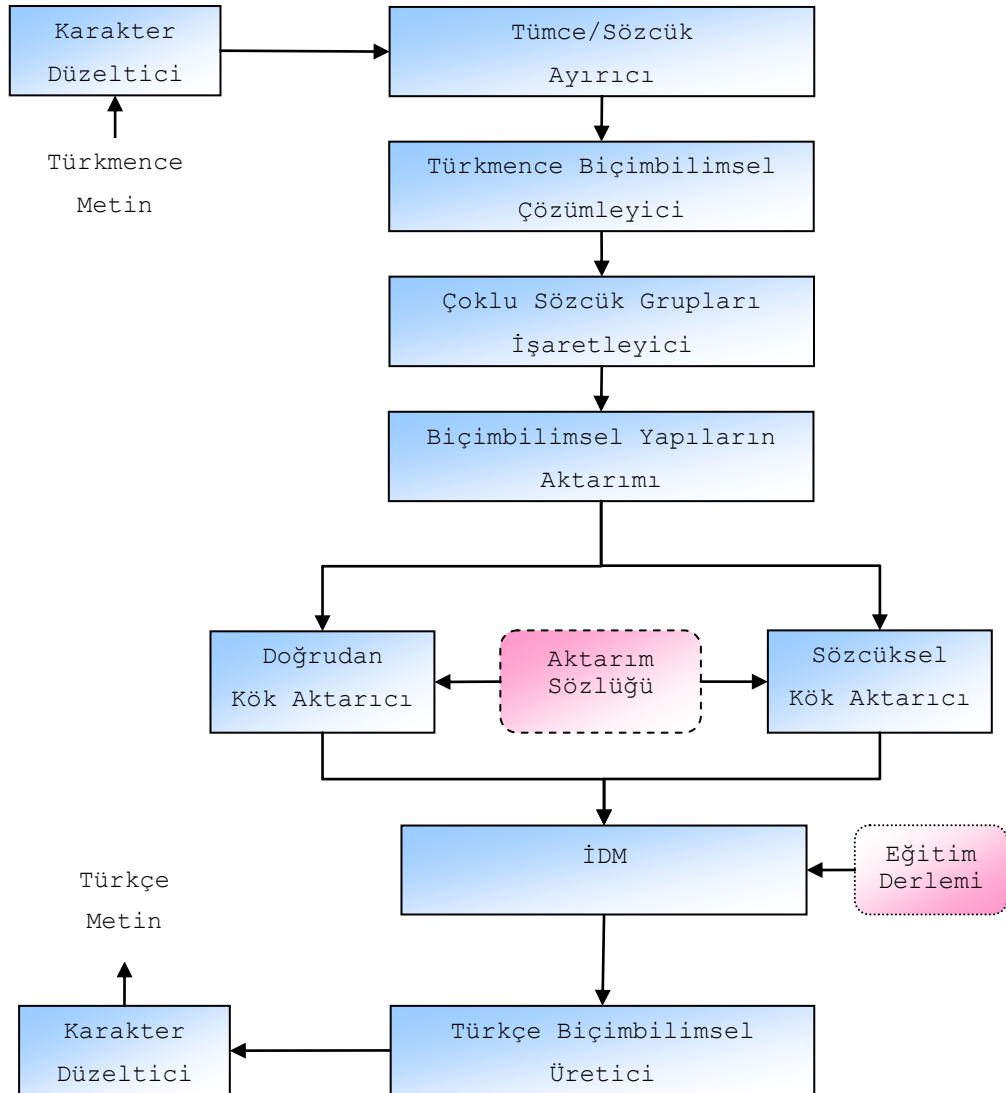
7.1.5 Türkçe Biçimbilimsel Üretici

Türkçe için kullanılan biçimbilimsel çözümleyici [13] SDD yapısında tasarlanmıř olduęu için ters yönde çalıştırıldıęında biçimbilimsel üretici olarak görev görmektedir.

7.2 Aktarım Modeli I Gerçekleşmesi

Model I kullanılarak gerçekleştirilen sistemin genel bileşenleri Şekil 7.7’de gösterilmiştir.

Model 0’ı temel alarak oluşturulan çeviri sisteminde çeviri işlemi, kaynak dildeki tümceyi oluşturan sözcüklerin, birbirlerinden bağımsız olarak aktarılması ile sağlanmaktadır. Ancak, sözcüklerin birbirlerinden bağımsız olarak değerlendirilmesi her koşulda uygun olmamaktadır. Bazı durumlarda, kaynak dildeki ÇSG’lerin bir bütün olarak aktarılması gereklidir. Bu çalışma düzenini sağlamak amacıyla, kaynak tümcedeki ÇSG’leri bularak bunları işaretleyen bir bileşen hazırlanmıştır. Böylelikle, kök aktarım bileşeni, bu yapıların bir bütün olduğunu anlayarak ona uygun aktarım işlemi gerçekleştirmektedir.



Şekil 7.7 : Aktarım Modeli I Temelinde Oluşturulan Sistemin Bileşenleri

Kaynak tümcedeki ÇSG’lerin belirlenmesi için, öncelikle Türkmencede ÇSG’lerin nasıl oluştuklarının incelenmesi gereklidir. Bunun için Türkçe üzerine yapılmış olan bir çalışma baz alınarak Türkmenceye uyarlanmıştır [46]. Bu çalışma uyarınca Türkçede dört tür ÇSG bulunmaktadır:

1. Sabit sözcüklerden oluşan gruplar
2. Kısmen sabit sözcüklerden oluşan gruplar
3. Sabit sözcük içermeyen kurallı gruplar
4. Yer, kişi vb. gibi özel isimleri ifade eden gruplar

Yukarıdaki sınıflandırma Türkmence için de uygun görülmektedir. Bu bağlamda Türkmence için de aynı sınıflandırma öngörülmüştür:

- Sabit sözcüklerden oluşan gruplar

Türkmencede birden fazla sözcükten oluşan gruplar genellikle araya “-“ işareti konularak yazılır:

Türkmence	Türkçe
dady-bidat	yazık
höre-köşe	güzel, tatlı (söz)
az-kem	hayal meyal
gürüm-jürüm	gizli

Ancak bu durumu garantilemek adına, bu şekilde birleşik yazılması gereken sözcüklerin bir listesi hazırlanmış ve bu listeye uygun olarak eksik “-“ işaretleri varsa, bunu metnin ilgili yerlerine ekleyen bir araç geliştirilmiştir.

- Kısmen sabit sözcüklerden oluşan gruplar

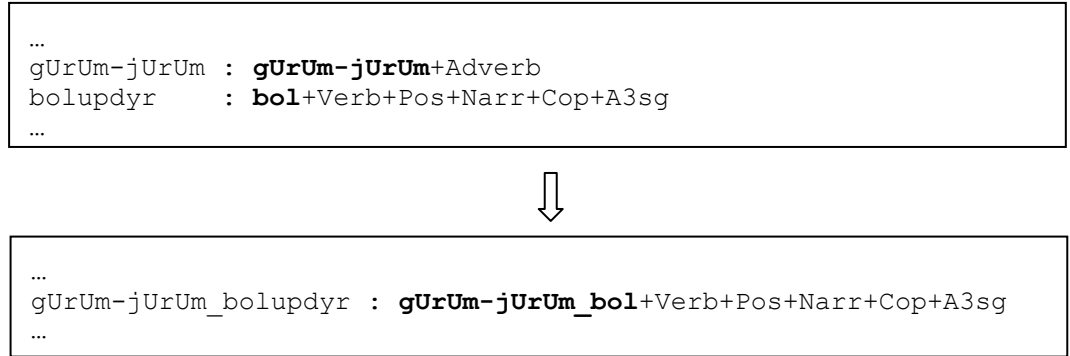
Bu sınıftaki ÇSG’ler, biçimbilimsel açıdan türetme ve çekim olaylarından etkilenirler. Ancak genellikle ÇSG’nin içinde sadece son sözcük türetme ve/veya çekim eki alabilir. Bu da ÇSG’nin başındaki sözcüklerin sabit olduğu anlamına gelmektedir. Aşağıda bu duruma uyan örnekler verilmiştir:

Türkmence	Türkçe
gürüm-jürüm bolmak	kaybolmak
gürüm-jürüm etmek	saklamak
emele gelmek	ortaya çıkmak
ekin meýdanı	tarla

Yukarıda örnekleri verilen ÇSG’lerin sadece son sözcüğüne yapım ve çekim eki getirilebilir:

gürüm-jürüm bolmakçy (kaybolmayı düşünüyor)
gürüm-jürüm bolypdy (kaybolmuştu)
gürüm-jürüm etjek (saklayacak)

Geliştirilen bir araç ile, biçimbilimsel çözümlemesi yapılmış sözcükler işlenerek, farklı yüzeysel biçimlere sahip bu ÇSG’ler aşağıdaki şekilde işaretlenmektedir:



Şekil 7.8 : Çekimlenen ÇSG’lerin Bulunması

Bu şekilde işaretlenen ÇSG kökü aktarılırken, aktarım sözlüğünde **gUrUm-jUrUm_bol** eylemi aranmaktadır. Bu sayede ÇSG türetme ya da çekim eki alsa bile doğru olarak aktarılabilir.

- Sabit sözcük içermeyen kurallı gruplar

Türkmencede bazı ÇSG’ler, sabit sözcük içermeyen, belirli kurallar çerçevesinde oluşturulur. Bu tür ÇSG’lere ilişkin birkaç örnek aşağıda verilmiştir:

Türkmençe	Türkçe	Kurallı Yapı
geçip bildi	geçebildi	eylem_kökü+Hp bil+...
olup bilmez	olamaz	eylem_kökü+Hp bil+...
geljek däl	gelmeyecek	eylem_kökü+jAk däl+...
var eken	varmış	isim_kökü+... eken
öy be öy	evden eve	isim be isim

Bu tür ÇSG’lerde genellikle ekler sabit kalmakta, ama grup içerisinde sözcükler değişebilmektedir.

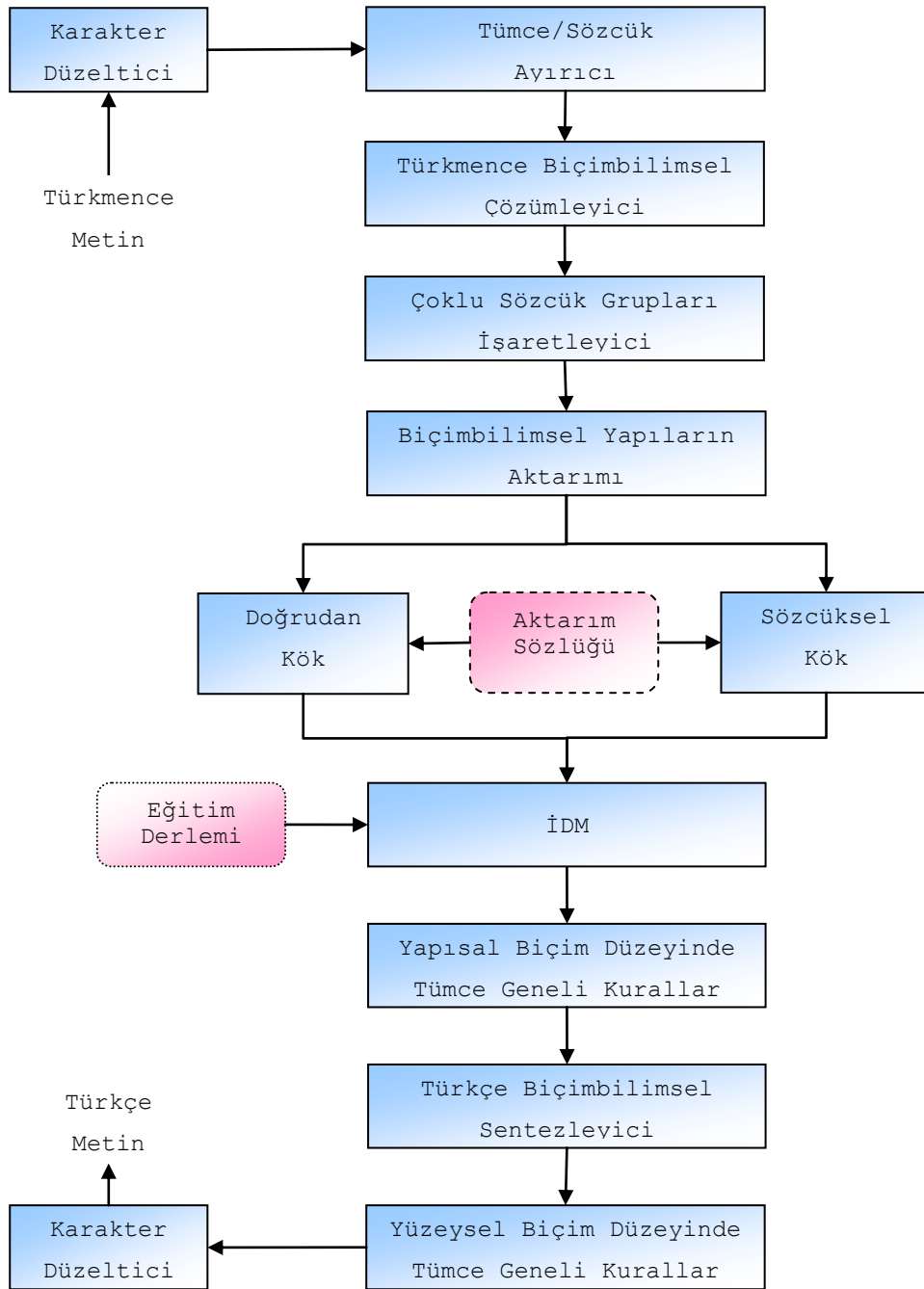
- Yer, kişi vb. gibi özel isimleri ifade eden gruplar

Bu sınıfta yer alan yapılar, Türkçeye olduğu gibi aktarılmaktadır, herhangi bir çeviri yapılmamaktadır.

Türkmençe giriş tümcesindeki ÇSG'leri bulmak üzere hazırlanan bu araçlar birleştirilerek temel çeviri sistemine eklenmiştir.

7.3 Aktarım Modeli II Gerçekleşmesi

Aktarım Modeli I'e ek olarak tümce genelinde hem biçimbilimsel düzeyde hem de yüzeysel düzeyde çalışan kuralların eklenmesi ile gerçekleştirilen Aktarım Modeli II'nin bileşenleri, Şekil 7.9'de gösterilmiştir:



Şekil 7.9 : Aktarım Modeli II Temelinde Oluşturulan Sistemin Bileşenleri

Model 0’da sözcük bazında çeviri yapılmasından doğan bazı sıkıntılar Model I’de ÇSG’lerin işlenmesi ile çözülmüştür. Ancak ÇSG’ler bulunurken sadece birbirlerine yakın sözcükler arasındaki ilişkiler dikkate alınmaktadır. Uygulamada, birbirlerine uzak sözcükler arasındaki bağımlılıklardan doğan çeviri hataları oluşmaktadır. Sorunun çözümü için tümce bazında sözdizimsel çözümlemenin yapılmasına gerek olmasa da, tümce bazında yapılması gereken bazı işlemlerin olduğu sonucuna varılmıştır.

Örneğin Türkmencede bazı kipler (gelecek zaman, zorunluluk ve planlama kipleri) şahıs eki almazlar⁹, ancak Türkçede bu eylemleri yapan şahsın belirtilmesi zorunludur:

Türkmence	Türkçe
Men geljek	geleceğ+im
Sen geljek	gelecek+sin
O geljek	gelecek+Ø

Hedef dildeki tümcenin sözdizimsel çözümlemesinin olması durumunda özneye bakarak eyleme uygun şahıs bilgisini iliştiirmek son derece basit bir işlemdir. Ancak daha önceki bölümlerde de nedenleri açıklandığı üzere, gerçekleştirilecek çeviri sisteminde sözdizimsel çözümlemeden kaçınılmaktadır. Aktarım başarımının yükseltilebilmesi için, sözdizimsel çözümleme yapılmasa dahi, tümce bazında bazı basit işlemlerin gerçekleştirilmesi gereklidir. Örnek olarak yukarıda anlatılan, eksik olan şahıs bilgisinin çıkarılması sorununun kısmen çözülmesi için tümcenin başında bazı adılların varlığı sorgulanabilir:

Men bu gije öýe geljek.
Ben bu gece eve geleceğim.

Sen bu gije öýe geljek.
Sen bu gece eve geleceksin.

Men ve Berdi bu gije öýe geljek.
Ben ve Berdi bu gece eve geleceğiz.

Berdi ve Seyid bu gije öýe geljek.
Berdi ve Seyid bu gece eve gelecekler.

Bu örneklerden de görüldüğü gibi, tümce başında bazı adıl kalıpları aranarak, eylemin eksik olan kişi bilgisi çıkarılabilir. Ancak bu kural, şahıs bilgisinin eksik olduğu tüm eylemler için çalışmayacaktır. Aşağıdaki karşı örnekte, tümce başında

⁹ Bu durum, biçimbilimsel çözümleyici tarafından +**Anon** etiketi ile temsil edilir.

kişi zamiri olmadığı için doğru kişi bilgisi çıkarılamaz. Bu tür durumlarda eyleme ön tanımlı bir davranış olarak 3. tekil kişi etiketi (+A3sg) eklenir.

Bu gije **men** öýe geljek.

Tümce genelinde yapılması gereken işlerden bir diğeri de Türkmencedeki ortaç öbeklerinde kendini göstermektedir. Türkçede ortaçların kendisine gelen iyelik eki, Türkmencede ortacın nitelediği ismin sonuna gelmektedir:

Türkmence	Türkçe
berjek çöregi	vereceği eklemek
geljek yoluňyz	geleceğiniz yol
geljek owadan yoluňyz	geleceğiniz güzel yol
görmedik çölleriň	görmediğin çölleri
gidýan yerimiz	gitmekte olduğumuz yer

Görüldüğü gibi, ortaçlarla kurulan ad öbekleri (noun phrase), sözdizimsel çözümleme kadar kesin olmasa da yaklaşık olarak kestirilmeli ve iyelik etiketinin yeri değiştirilmelidir.

Tümce bazında yürütülmesi gereken bir başka işlem de Türkçede ayrı yazılan ama kendinden önceki sözcüğün son seslisine göre değişen **de / da** bağlacı ile **mi / mı / mü / mu** soru ekleridir. Hedef dile çevrilen sözcükler yüzeysel biçime dönüştürüldükten sonra bazı kuralların tümce genelinde bu harf değişikliklerini yapması gereklidir.

Yukarıda verilen üç örnek gibi, tümce genelinde iş yapan toplam sekiz farklı kural kümesi oluşturulmuştur. Bu kurallar, işlem yaptıkları düzeyler bakımından iki sınıfa ayrılırlar:

- **Yapısal Biçime Gerek Duyan Kurallar**

Bu kuralların işlem görebilmesi için sözcüklerin biçimbilimsel verilerinin de bulunması gereklidir. Bundan dolayı bu kurallar, hedef dil biçimbilimsel üretici bileşeninden önce çalıştırılır.

- **Yüzeysel Biçim Üzerinde İşlem Yapan Kurallar**

Genellikle sözcükler arası sesli uyumları ile ilgili harf değişikliklerini yapan bu kurallar ise hedef dil biçimbilimsel çözümleyiciden sonra üretilen sözcüklerin yüzeysel biçimleri üzerinde işlem yaparlar.

7.4 İki Seviyeli İDM Kullanılması

Farklı tipte İDM'lerin oluşturulmasının nedeni Bölüm 4.3'te anlatıldığı gibi seyrek veri sorunundan en az biçimde etkilenmektir. Bu nedenle tüm biçimbilimsel etiketleri içeren tam çözümleme yerine, bu çözümlemelerin belirli bölümleri kullanılmıştır. Dolayısı ile her İDM, hedef dili, farklı açılardan modellemektedir. Örneğin İDM Tip-I kökleri modellerken İDM Tip-II sözcük türlerini modellemektedir.

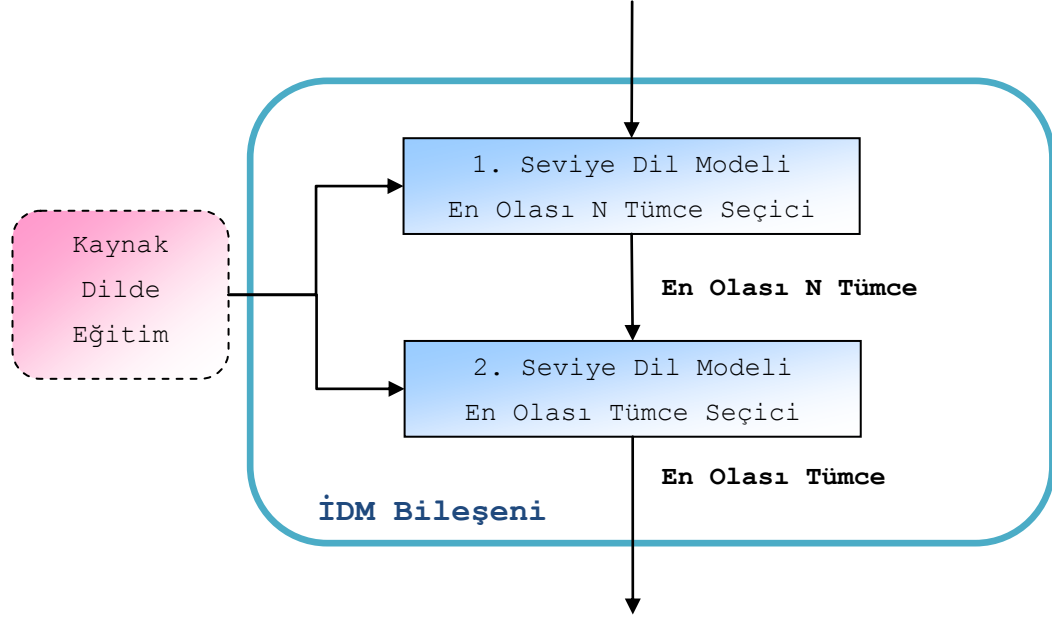
Ancak belirli bir tip İDM seçildiğinde oluşan aday çeviriler, sadece seçilen İDM'nin modellediği açıdan değerlendirilmektedir. Örneğin kökleri modelleyen bir İDM, diğer biçimbilimsel bilgilerin hiçbirini değerlendirmeye katamaz. Bu olumsuzluğu gidermek adına, farklı tipte dil modellerinin beraber kullanılmasının yolları araştırılmıştır. Bu sayede hedef dili farklı açılardan modelleyen farklı tiplerdeki İDM'lerin sonuçları birleştirilerek en olası tümcenin bulunması denenmiştir.

Benzer sorunların çözümü için literatürde sıklıkla kullanılan bir yöntem, olasılığı en yüksek sıralamanın bulunması için iki farklı seviye kurulmasını öngörmektedir. “En İyi N Aday” (N-Best Lists) adı verilen bu yöntemde, 1. seviyede dar kapsamlı bir bilgi kaynağı ile tüm arama uzayında bir arama gerçekleşir ve çıktı olarak en olası N aday, olasılıklarına göre sıralanmış olarak 2. seviyeye aktarılır. Daha geniş kapsamlı bir bilgi kaynağı kullanan 2. seviye arama yöntem(ler)i, en iyi N adaydan oluşan daraltılmış arama uzayında işlem yaparak bu N adayı tekrar sıralar (re-ranking) ve sonuçta elde edilen en iyi adayı çıktı olarak üretir [67]. En İyi N Aday arama yöntemi kullanılırken, 1. aşama için kullanılan arama yönteminin maliyetinin düşük tutulmasına özen gösterilmesi gerekmektedir.

Çeviri sorununa geri dönersek, oluşturulan aday tümceler içerisinde bütün İDM Tipleri kullanılarak en olası tümceyi seçmenin maliyetinin çok yüksek olduğu görülmüştür.¹⁰ Bu noktada, En İyi N Aday yöntemini kullanabilmek amacıyla iki adet farklı İDM tipinin beraber kullanıldığı 2 seviyeli bir istatistiksel modelleme gerçekleştirilmiştir. 1. seviyede kurulan HMM üzerinde çalışan geliştirilmiş Viterbi [68, 69] algoritması sayesinde olasılığı en yüksek bir tümce yerine N adet tümce üretilir. 2. seviyede ise arama uzayı küçültülen bu tümceler arasında (tercihen hedef dili daha geniş kapsamlı modelleyen) bir diğer İDM tipi ile 2. seviye tümce olasılığı

¹⁰ Standart bir bilgisayarda 255 tümcenin aktarımı yaklaşık 9 dakika sürmektedir.

hesaplanır. Son adımda ise her 1. ve 2. seviyelerde hesaplanan olasılık değerleri birleştirilir ve tümcelerın son olasılık değeri bulunur. Bu sıralamada N tümce arasında en yüksek olasılığa sahip tümce çıktı olarak üretilir.



Şekil 7.10 : İki Seviyeli İDM Uygulaması

8. UYGULAMA VE SONUÇLAR

8.1 Eğitim ve Sınama Verisi

İDM'lerin eğitilmesi için bir eğitim derlemine gereksinim duyulmaktadır. Eğitim verisi olarak yaklaşık 1 milyon sözcükten ve 50 bin tümceden oluşan bir derlem kullanılmıştır. Derlemdeki tümceler, ağırlıklı olarak günlük bir gazetenin belirli bir dönemdeki haber metinlerinden oluşmaktadır. Derlem ile ilgili istatistiksel bazı bilgiler Tablo 8.1'de verilmiştir.

Tablo 8.1 : Eğitim Derlemi İle İstatistikler

Toplam Kelime Adedi	948.403
Toplam Tümce Adedi	50.674
Farklı Sözcük Kökleri Adedi	25.897
1 Defa Geçen Sözcük Kökleri	11.447
2 Defa Geçen Sözcük Kökleri	3505
2'den Fazla Geçen Sözcük Kökleri	10.945

Eğitim derlemi, sözcüklerin yüzeysel biçimlerini ve doğru biçimbilimsel çözümlemesini içermektedir. Derlemdeki sözcükler ilk önce Türkçe biçimbilimsel çözümleyiciden [13] geçirilmiş, ikinci aşamada da doğru biçimbilimsel çözümlemelerinin seçilmesi, kısmen elle kısmen de otomatik yöntemlerle gerçekleştirilmiştir. Bu derlem Türkçe üzerinde yapılan çeşitli biçimbilimsel belirsizlik giderme (morphological disambiguation) araştırmalarında kullanılmaktadır [52, 53, 70]. Eğitim verisinden alınan örnek bir tümce aşağıda verilmiştir:

```
<S> <S>+BStag  
İlköğretim ilköğretim+Noun+A3sg+Pnon+Nom  
önce önce+Adverb  
8 8+Num+Card  
, ,+Punc  
ardından ardından+Adverb  
da da+Conj  
mutlaka mutlaka+Adverb  
11 11+Num+Card  
yıla yıl+Noun+A3sg+Pnon+Dat  
çıkarılmalıdır çık+Verb^DB+Verb+Caus^DB+Verb+Pass+Pos+Neces+Cop+A3sg  
</S> </S>+ESTag
```

Ayrıca Türkmençe ile Türkçe arasındaki farklılıklardan dolayı derlem üzerinde uyumlaştırıcı bazı yapay değişiklikler yapılmıştır:

1. Türkmencede, sözcüğe bitişik yazılan **+yla** aracılık durum eki olmadığı için, Türkçe eğitim derleminde **+Ins** durumunda olan tüm isim soylu sözcükler yalın hale (**+Nom**) getirilmiş ve bu sözcükten hemen sonra **ile+Postp+PCNom** eklenmiştir (arabayla ⇔ araba ile).
2. Soru eki Türkmencede eyleme bitişik yazılmaktadır. Ancak Türkçede ayrı yazılan ve derlemde **+mı**, **+mu**, **+mi**, **+mü** şeklinde geçen farklı soru kökleri bulunmaktadır. Eğitim derleminde bu girdiler silinmiş, bunlardan önceki eylemin sonuna **+Ques** takısı eklenmiştir.

Sisteminin başarısının sınanabilmesi için bir sınama kümesi oluşturulmuştur. Bu sınama kümesi, ağırlıkla hikayelerden alınmış 255 adet Türkmençe tümce ve bu tümcelerin 2 farklı kaynaktan sağlanmış Türkçe karşılıklarını (referans çevirilerini) içermektedir.

Çeviri sırasında biçimbilimsel ve sözcüksel belirsizlikler oluşmaktadır. Bu belirsizliklerin derecesini göstermek amacıyla, sınama kümesi çevrilirken çıkan belirsizlik oranları, Tablo 8.2’de verilmiştir.

Tablo 8.2 : Sınama Kümesinde Ölçülen Belirsizlik Oranları

Türkmencede Sözcük Başına Düşen	1,55
Biçimbilimsel Çözümleme Adedi	
Türkçe Karşılığı 1 Tane Olan Türkmençe Sözcük Oranı	% 44,7
Türkçe Karşılığı 2 Tane Olan Türkmençe Sözcük Oranı	% 34,0
Türkçe Karşılığı > 2 Olan Türkmençe Sözcük Oranı	% 21,3

8.2 Değerlendirme Ölçütleri

Önerilen çeviri modellerinin ve İDM türlerinin başarımlarını ölçmek için *BLEU* ölçütü seçilmiştir. Ancak *BLEU* ölçütünün, Türk dilleri için bazı olumsuz yönleri bulunmaktadır.

Türkçe, tümcedeki öğelerin (belirli bir derecede) yer değiştirebildiği (free constituent order) bir dildir. Burada kullanılan öğe tanımı, sözcük öbekleri anlamındadır. Çoğu Türkçe tümce özne-nesne-yüklem sırasını izlemesine rağmen özellikle konuşma

dilinde farklı sıralar izlenebilir. Birçok dilde sözcüklerin tümce içindeki görevleri tümce içerisindeki yerlerine yani sıralarına göre belirlenirken, Türkçede sözcük öbeklerindeki biçimbilimsel yapılar, tümcedeki yerinden bağımsız olarak öbeğin görevini gösterir. Sözcük öbeklerinin farklı sıralamaları ise farklı bağlamlarda farklı noktaları vurgular [71].

BLEU yönteminde, birden fazla referans çeviri kullanılarak sistemin çıktısı ile bu referanslar arasındaki sözcük dizilerinin (genellikle 4 uzunluğa kadar) eşleşme oranlarına bakılır. Dolayısı ile çeviri sistemi, tümcedeki öğeleri referanslardan farklı şekilde sıralamışsa, söz konusu sistem çıktısı *BLEU* tarafından çok düşük puanlandırılır.

Ancak Türkçede öğelerin sırası değişse bile tümcenin anlamı aynı kalabildiği için sistem çıktısının, referans(lar)la aynı anlamı taşıyan tümceler üretme olasılığı da bulunmaktadır:

Referans	:	Kedi	evdeki	vazoyu	kazayla	kırdı.	
Sistem Çıktısı 1	:	Evdeki	vazoyu	kazayla	kedi	kırdı.	<i>BLEU</i> = 0,55
Sistem Çıktısı 2	:	Kazayla	evdeki	vazoyu	kırdı	kedi.	<i>BLEU</i> = 0,38
Sistem Çıktısı 3	:	Kırdı	vazoyu	kazayla	evdeki	kedi.	<i>BLEU</i> = 0,38

Örnekler incelendiğinde, 3. çıktı, çok daha bozuk bir Türkçe tümce olmasına karşın, 1. sistemin çıktısı ile arasında aynı oranda puan farkı yoktur. Oysa ki 1. sistem çıktısı, referans ile aynı anlamı taşımaktadır ve sadece referanstaki “**kazayla**” sözcüğünde yer alan vurguyu “**kedi**” sözcüğüne aktarmıştır. Örnek sistem çıktısı 2, sistem çıktısı 3’den çok daha doğru bir Türkçe tümce olmasına rağmen çıktı 3 ile aynı puanı almıştır. Bu örneklerden görüldüğü gibi *BLEU* ölçütü, Türkçe sistemlerin başarımını tam anlamı ile ölçememektedir.

Ancak öngörülen aktarım modelleri sözcük sıralarında büyük bir değişikliğe gitmediği için, yukarıda açıklanan sakıncaların etkisi çok daha az olacaktır.

BLEU ölçütünün bitişken diller açısından bir başka olumsuz yönü ise, referans ile aday çeviriler arasında eşleme yaparken sözcüklerin sadece yüzeysel biçimlerinin işlenmesidir. Aktarım sistemi, biçimbilimsel özelliklerden bir tanesinde bile hata yaparsa, oluşan sözcüğün yüzeysel biçimi farklı olmakta ve *BLEU* puanı çok düşük olmaktadır:

Referans	:	Kedi	evdeki	vazoyu	kazayla	kırmış.	
Sistem Çıktısı	:	Kedi	evdeki	vazoyu	kazayla	kırmıştır.	<i>BLEU</i> = 0,66

Örnekte, aktarım sisteminin, biçimbilimsel yapıların aktarımı sırasında seçtiği, ancak çok büyük anlam kaymasına yol açmayan bir farklılığın, *BLEU* puanını nasıl önemli ölçüde düşürdüğü görülmektedir. Bu olumsuz etkiyi azaltmak için, sistemin başarımı ölçülürken, sadece köklerin eşleşmelerinden hesaplanan *BLEU_r* puanı ile normal *BLEU* puanı beraber kullanılmıştır. *BLEU_r* puanını hesaplamak amacıyla, referans çeviriler kümesindeki sözcüklerin biçimbilimsel çözümlemesi yapılmış ve sözcük kökleri elle seçilerek kök referans kümesi de oluşturulmuştur.

Sonuç olarak *BLEU* ölçütü, Türk dil ailesindeki dillerin hedef dil olduğu bilgisayarlı çeviri sistemlerinin başarımlarını değerlendirmesi için en uygun yöntem değildir. Ancak otomatik olması, yaygın olarak kullanılması ve alternatifinin olmaması gibi nedenler göz önüne alındığında uygulamamızın başarımlarını değerlendirmesi *BLEU* yöntemi ile yapılmıştır. Ancak *BLEU* ölçütü, gerçekleşen çeviri sisteminin başka sistemlerle karşılaştırılması amacı yerine daha çok sisteme eklenen bileşenlerin ve yeni kuralların etkilerini saptamak amacı ile kullanılmaktadır [43]. İDM kullanılırken, bağlamı belirlediği için noktalama işaretleri kullanılmış olsa da *BLEU* puanları hesaplanırken noktalama işaretleri gözardı edilmiştir.¹¹

Önerilen modellerin sözcük sıralarında neredeyse hiçbir değişiklik yapmadığı dikkate alındığında, sıradan bağımsız olarak tek sözcük eşleşmelerinin de (1 uzunluklu eşleşmeler) başarımlarını hakkında bir ipucu verebileceği akla gelmektedir. *BLEU-1* puanı olarak gösterilen bu tekli eşleşmeler, sistemin referanstaki sözcükleri sıradan bağımsız olarak üretebilme yeteneğini gösterir. Bu nedenle, sonuçlar verilirken *BLEU-1* puanları da belirtilmiştir. Sözcük sıralarında herhangi bir değişiklik yapılmadığı varsayımı ile sistemin ulaşabileceği başarımların üst sınırı, en yüksek *BLEU-1* puanı olan 100 almaktır. Ancak *BLEU-1* puanı olarak 100 elde edilmesi, genel *BLEU* puanının da yüksek olmasını gerektirmediği unutulmamalıdır. Örneğin referanstaki sözcüklerin dağınık bir şekilde birleştirilmesiyle oluşan bir çeviri çıktısı, tekli eşleşmelerden (*BLEU-1*) 100 puan alsa bile ikili (*BLEU-2*), üçlü (*BLEU-3*) ve dördü (*BLEU-4*) eşleşmelerden puan alamayağı için, çıktının genel *BLEU* puanı çok daha düşük olacaktır.

¹¹ BLEU puanı hesaplaması için kullanılan araç, 2005 yılında düzenlenen istatistiksel çeviri yarışmasının değerlendirme aracıdır: <http://www.statmt.org/wmt05/shared-task/> (2005) Bu aracın çıktılarında, normalde 0-1 aralığında olan BLEU puanı 0-100 arasına çekilerek verilmektedir.

8.3 Sonular

Tasarlanan aktarım fonksiyonu modellerinin başarımlarını incelemek için her bir modelin sınama girdisine karşılık ürettiği çevirilerin *BLEU* ve *BLEUr* puanları hesaplanmıştır. İlk önce temel model puanı hesaplanmış ve bu bir alt sınır olarak kabul edilmiştir. Daha sonra diğer modellerin *BLEU* ve *BLEUr* puanları ile temel sistemin başarımları karşılaştırılmıştır.

8.3.1 Temel Modelin Başarımı

Geliştirilen temel aktarım modeli, 1-gram kök İDM (Tip-I) ile çalıştırılmış ve başarımları ölçülmüştür:

Tablo 8.3 : Aktarım Modeli 0'ın (Temel Model) Başarımı

<i>Aktarım Modeli</i>	<i>BLEU</i>	<i>BLEU-1</i>	<i>BLEUr</i>	<i>BLEUr-1</i>
Model 0	26,57	58,80	35,58	68,40

Bu tabloda *BLEU-1* ve *BLEUr-1* puanları, sırasıyla yüzeysel biçimde ve kök biçiminde referanslarla eşleşen sözcük yüzdelerini (1-gram eşleşmelerini) vermektedir. Buna göre Model 0 (temel model), referanslarda geçen sözcüklerin %58,80'lik bir oranını doğru olarak üretebilmeyi başarmış, üstelik sözcük köklerinin %68,40'ını doğru seçmiştir. Ancak hatalı çeviriler yüzünden 2-gram, 3-gram ve 4-gram eşleşmelerinin sayıları azalmış ve sonuçta daha düşük *BLEU* puanları alınmıştır.

BLEU ile *BLEUr* arasındaki farklar ise, sözcüksel belirsizliğin doğru olarak giderilebildiği ancak biçimbilimsel belirsizliğin giderilmesinde yanlışlık yapıldığı durumlardan kaynaklanmaktadır. Başka bir deyişle bu fark, doğru sözcük kökü seçilmesine rağmen, Türkmence biçimbilimsel belirsizlik doğru giderilemediği veya için yanlış bir yüzeysel biçimin oluşturulduğu örneklerden kaynaklanmaktadır.

Bu başarımlar, sistem üzerinde yapılacak her türlü değişikliğin sonuçlarını ölçmemize yarayan bir temel kıyaslama (baseline) sonucu olacaktır. Yapılan her türlü değişiklik sonucu elde edilen başarımlar, bu temel kıyaslama sonucuyla göreceli olarak değerlendirilmelidir. Bu puanlardan düşük sonuçlar, sistem üzerinde iyileştirme amaçlı yapılan değişikliklerin olumsuz sonuç verdiğini göstermektedir.

8.3.2 Aktarım Modeli I'in Başarımı

Temel modele ek olarak ÇSG'lerin de çevrilmesi sağlanarak sözcük bazında çoktan-çoğa çeviri yapılmasını gerçekleyen Aktarım Modeli I, gene köklerden oluşan Tip-I İDM ile $n=1$ seçilerek çalıştırılmış ve aşağıdaki sonuçlar elde edilmiştir:

Tablo 8.4 : Aktarım Modeli I Başarımı

<i>Aktarım Modeli</i>	<i>BLEU</i>	<i>BLEU-1</i>	<i>BLEUr</i>	<i>BLEUr-1</i>
Model 0	26,57	58,80	35,58	68,40
Model I	28,45	60,90	38,03	70,20

Tablo 8.4'den görüldüğü gibi ÇSG'lerin işlenerek aktarılması, *BLEU* puanlarını yükseltmektedir. Örneğin, temel sistem tarafından doğru aktarılamayan sınama derlemindeki aşağıdaki yapılar, ÇSG'lerin belirlenerek hedef dile aktarılması sayesinde doğru olarak çevrilebilmiştir:

ulanylyp başlaýar	→	kullanılıp başlıyor kullanılmaya başlıyor	(Model-0) (Model-I)
durup bilmeýip	→	durup bilmeyip duramayıp	(Model-0) (Model-I)
gürüm-jürüm bolydyr	→	gizli olmuştur kaybolmuştur	(Model-0) (Model-I)

8.3.3 Aktarım Modeli II'in Başarımı

Tümce genelinde yapısal ve yüzeysel düzeylerde çalışan kuralların Model I'e eklenmesiyle geliştirilen Model II'nin başarımı aşağıda verilmiştir. Diğer aktarım modellerinin başarımları ile adil bir şekilde karşılaştırma yapılabilmesi için aktarım Modeli II için de köklerden oluşan Tip-I İDM $n=1$ seçilerek çalıştırılmıştır.

Tablo 8.5 : Aktarım Modeli II'nin Başarımı

<i>Aktarım Modeli</i>	<i>BLEU</i>	<i>BLEU-1</i>	<i>BLEUr</i>	<i>BLEUr-1</i>
Model 0	26,57	58,80	35,58	68,40
Model I	28,45	60,90	38,03	70,20
Model II	30,53	62,30	37,81	70,60

Tümce genelinde gerçekleşen aktarım kuralları ile Model II'de başarımın arttığı görülmektedir. Aşağıdaki örnekler, tümce seviyesinde çalışan kurallar ile düzeltilen bazı durumları göstermektedir:

günleriň birinde bir owadanja guş bar eken.	(Türkmençe)
günlerin birinde bir güzelce kuş var imiş.	(Model-I)
günlerin birinde bir güzelce kuş varmış.	(Model-II)
olar günüň öňki duran ýerine jemlenişmege başladylar.	(Türkmençe)
onlar günün önceki duran yerine toplaşmaya başladılar.	(Model-I)
onlar günün önceki durduğu yere toplaşmaya başladılar.	(Model-II)
men hem suwuň içinde gezip ýadadym.	(Türkmençe)
ben da suyun içinde bulunup yoruldum.	(Model-I)
ben de suyun içinde bulunup yoruldum.	(Model-II)

8.3.4 Farklı Türde Dil Modellerinin Başarımı

Aktarım fonksiyonunun ürettiği tümcelerden en yüksek olasılığa sahip olanını bulmak için değişik İDM tipleri kullanılmış ve üretilen çevirilerin *BLEU* puanları hesaplanmıştır. İDM derecesinin etkisinin de görülebilmesi amacı ile $n=1,3$ ve 5 seçilerek deneyler tekrarlanmıştır.

Tablo 8.6 : İDM Tip-I Başarımı

Aktarım Modeli	BLEU			BLEU-1		
	$n=1$	$n=3$	$n=5$	$n=1$	$n=3$	$n=5$
Model 0	26,57	27.60	27.88	58,80	59.3	59.4
Model I	28,45	29,47	29,75	60,90	61,40	61,50
Model II	30,53	31,03	31,30	62,30	62,60	62,70
	BLEUr			BLEUr-1		
Model 0	35,58	35,52	35,67	68,40	68,90	69,00
Model I	38,03	37,98	38,14	70,20	70,90	70,90
Model II	37,81	37,77	37,92	70,60	71,30	71,30

Tablo 8.6’da göze çarpan en önemli nokta, beklendiği üzere dil modeli derecesi n ’nin artması ile çeviri başarımının yükselmesidir. Ancak $n=3$ ve $n=5$ arasında çok büyük bir fark bulunmamaktadır. Gerçekten de söz konusu dil modellerinin derecelerine göre entropi değişimlerine bakıldığında (Tablo 4.8), $n>3$ için entropi değerinin çok fazla değişmediği görülmektedir. Uygulama açısından $n=5$ için çalışma süresi çok

yüksek olduğundan önerilen aktarım modeli için $n=3$ seçilmesi uygun görülmektedir¹².

Daha önce de gözlendiği gibi, Tablo 8.6 ile aktarım modelleri başarımları karşılaştırıldığında Model II'nin başarısının diğerlerinden yüksek olduğu görülmektedir.

Yapısal gösterimdeki son sözcük türlerinden oluşan İDM Tip-II'nin başarımlarını değerlendiren sonuçları içeren tablo aşağıda verilmiştir:

Tablo 8.7 : İDM Tip-II Başarımları

Aktarım Modeli	BLEU			BLEU-1		
	$n=1$	$n=3$	$n=5$	$n=1$	$n=3$	$n=5$
Model 0	22,79	22,49	22,63	54,30	54,30	54,50
Model I	24,48	24,15	24,35	56,40	56,30	56,70
Model II	24,95	24,73	24,89	56,40	56,70	56,80
	BLEUr			BLEUr-1		
	$n=1$	$n=3$	$n=5$	$n=1$	$n=3$	$n=5$
Model 0	26,24	26,49	26,57	61,10	61,40	61,50
Model I	28,19	28,30	28,50	62,70	62,90	63,10
Model II	28,10	28,13	28,16	62,80	63,20	63,30

İDM olarak Tip-II kullanıldığında elde edilen sonuçlarda gözlenen başlıca farklılık, dil modeli derecesinin artması ile başarımın fazla değişmemesidir. Bu durum, önerdiğimiz çeviri yönteminde, sözcük türlerinden oluşan bir İDM bileşenin tek başına yeterli olmadığının bir göstergesi olarak yorumlanabilir. Ayrıca İDM Tip-II ile elde edilen puanların tamamı, temel sistemin puanından geride kalmaktadır.

Tablo 8.8'de, biçimbilimsel çözümlemelerin son ÇG'lerinden oluşturulan İDM Tip-III'ün BLEU sonuçları verilmiştir.

¹² Ses tanıma, istatistiksel çeviri gibi bir çok günümüz uygulamalarında genellikle İDM için $n=3$ seçilmektedir.

Tablo 8.8 : İDM Tip-III Başarımı

Aktarım Modeli	BLEU		BLEU-1		BLEU-1	
	n=1	n=3	n=1	n=3	n=1	n=3
Model 0	23,17	23,88	23,99	55,00	55,90	56,00
Model I	24,92	25,63	25,66	57,30	58,20	58,10
Model II	25,25	26,39	26,41	57,30	58,80	58,60
Aktarım Modeli	BLEUr		BLEUr-1		BLEUr-1	
	n=1	n=3	n=1	n=3	n=1	n=3
Model 0	28,00	28,19	28,13	62,20	62,60	62,70
Model I	30,10	30,30	30,24	64,00	34,40	64,50
Model II	29,74	30,05	30,08	64,10	64,70	64,70

Çözümlerlerin son ÇG'leri kullanılarak oluşturulan Tip-III İDM, sözcük kökleri içeren Tip-II İDM'ye göre bir miktar daha fazla biçimbilimsel bilgiyi kullanmaktadır. Bu nedenle Tip-II BLEU puanlarına göre ufak bir artış gözlenmiştir. Ayrıca Tip-II'nin tersine, dil modeli derecesi arttıkça başarımlarda küçük artışlar sağlanabilmektedir. Ancak gene de temel sistemin başarısına çoğu durumda ulaşamamış, ancak Model II kullanıldığında $n=3$ ve 5 için temel sistem başarımına yakınsanmıştır. Tip-I ve Tip-II İDM'nin birleştirilmesi ile oluşan kök ve sözcük türü bilgilerini kullanan Tip-IV İDM'nin başarımlar tablosu aşağıda verilmiştir.

Tablo 8.9 : İDM Tip-IV Başarımı

Aktarım Modeli	BLEU			BLEU-1		
	n=1	n=3	n=5	n=1	n=3	n=5
Model 0	26,31	29,52	29,52	57,90	60,40	60,40
Model I	28,20	31,37	31,37	60,10	62,60	62,60
Model II	29,69	33,34	33,34	61,30	64,00	64,00
Aktarım Modeli	BLEUr			BLEUr-1		
	n=1	n=3	n=5	n=1	n=3	n=5
Model 0	34,30	35,51	36,51	67,20	69,00	69,00
Model I	36,66	38,84	38,84	69,10	70,80	70,80
Model II	36,44	38,62	38,62	69,40	71,20	71,20

Kök bilgisinin eklenmesi ile $n=1$ için temel sistem başarısı yakalanmış, $n=3$ ve $n=5$ dereceleri için ise temel sistem başarımı geçilebilmiştir. Tip-IV İDM ile Tip-I İDM karşılaştırılacak olursa, *BLEUr* puanlarında bir değişiklik olmamasına karşın son sözcük türü gibi biçimbilimsel bir bilginin eklenmesi ile toplamda bir başarı artışının sağlandığı söylenebilir.

Sözcük kökleri ile çözümlemelerin son ÇG'sinden oluşan Tip-V İDM kullanılarak yapılan sınamaların sonuçları Tablo 8.10'da verilmiştir. Sonuçlar incelenmeden önce sezgisel olarak şu sav ileri sürülebilir : “eğer sözcük kökü bilgisine sözcük türü gibi biçimbilimsel bir bilgi eklenerek oluşturulan İDM ile başarı yükseltilebiliyorsa, sözcük köküne daha fazla biçimbilimsel bilgi eklenerek başarı daha da arttırılabilir”. Ancak Tablo 8.10 incelendiğinde sonuçların hiç de beklendiği gibi çıkmadığı görülebilir. Genel *BLEU* puanının düşmesinin temel nedeni, *BLEUr-1* değerlerinin düşmesine bağlıdır. Yani sistem sözcük köklerini aktarırken bu İDM ile daha fazla hata yapmaktadır. Bu sonuç, İDM Tip-V'in seyrek veri sorunundan fazlaca etkilendiğini ve bu yüzden sözcük kökü aktarımındaki başarının düştüğünü göstermektedir.

Tablo 8.10 : İDM Tip-V Başarımı

Aktarım Modeli	BLEU			BLEU-1		
	$n=1$	$n=3$	$n=5$	$n=1$	$n=3$	$n=5$
Model 0	27,80	27,71	27,71	59,40	59,10	59,10
Model I	29,80	29,71	29,71	61,60	61,40	61,40
Model II	31,69	31,51	31,51	62,60	62,70	62,70
	BLEUr			BLEUr-1		
	$n=1$	$n=3$	$n=5$	$n=1$	$n=3$	$n=5$
Model 0	33,29	32,39	32,39	67,10	66,70	66,70
Model I	35,58	34,66	34,66	68,90	68,40	68,40
Model II	35,59	34,46	34,46	69,10	68,80	68,80

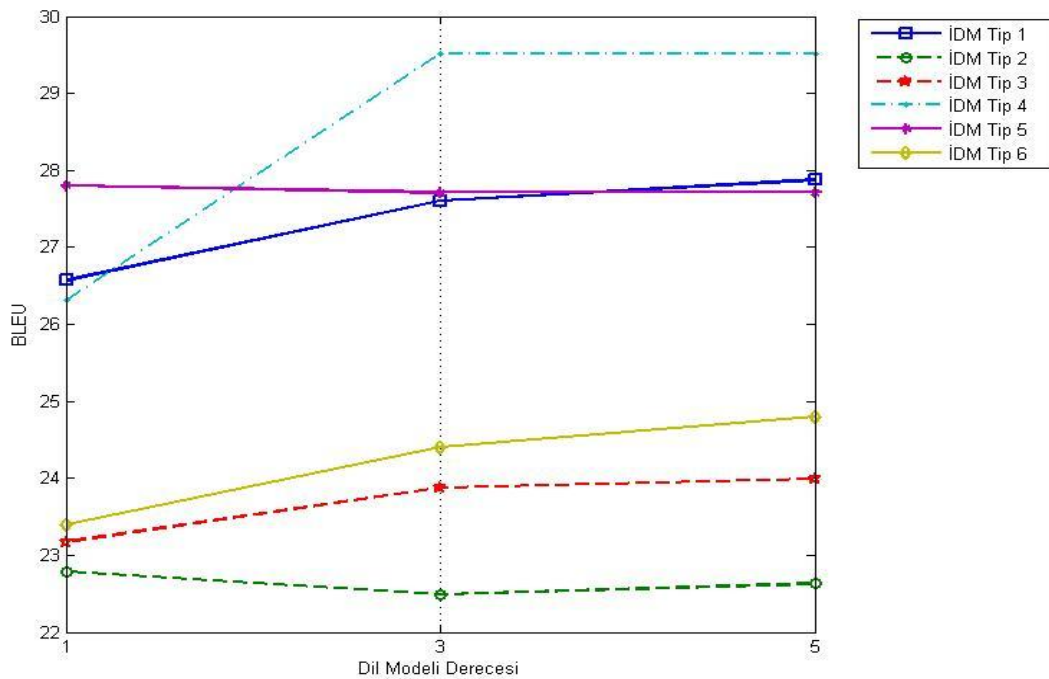
Kök bilgilerini içermeyen, sadece biçimbilimsel çözümlemenin kök harici bileşenlerini içeren İDM Tip-VI ile elde edilen çevirilerin *BLEU* puanları Tablo 8.11'de verilmiştir.

Tablo 8.11 : İDM Tip-VI Başarımı

<i>Aktarım Modeli</i>	<i>BLEU</i>			<i>BLEU-1</i>		
	<i>n=1</i>	<i>n=3</i>	<i>n=5</i>	<i>n=1</i>	<i>n=3</i>	<i>n=5</i>
Model 0	23,40	24,40	24,80	55,20	56,40	56,60
Model I	25,15	25,91	26,47	57,40	58,60	58,80
Model II	25,47	27,02	27,46	57,50	59,20	59,30
	<i>BLEUr</i>			<i>BLEUr-1</i>		
	<i>n=1</i>	<i>n=3</i>	<i>n=5</i>	<i>n=1</i>	<i>n=3</i>	<i>n=5</i>
Model 0	29,50	29,85	29,93	62,90	63,30	63,40
Model I	31,58	31,93	32,01	64,70	65,10	65,20
Model II	31,24	31,60	31,74	64,80	65,50	65,50

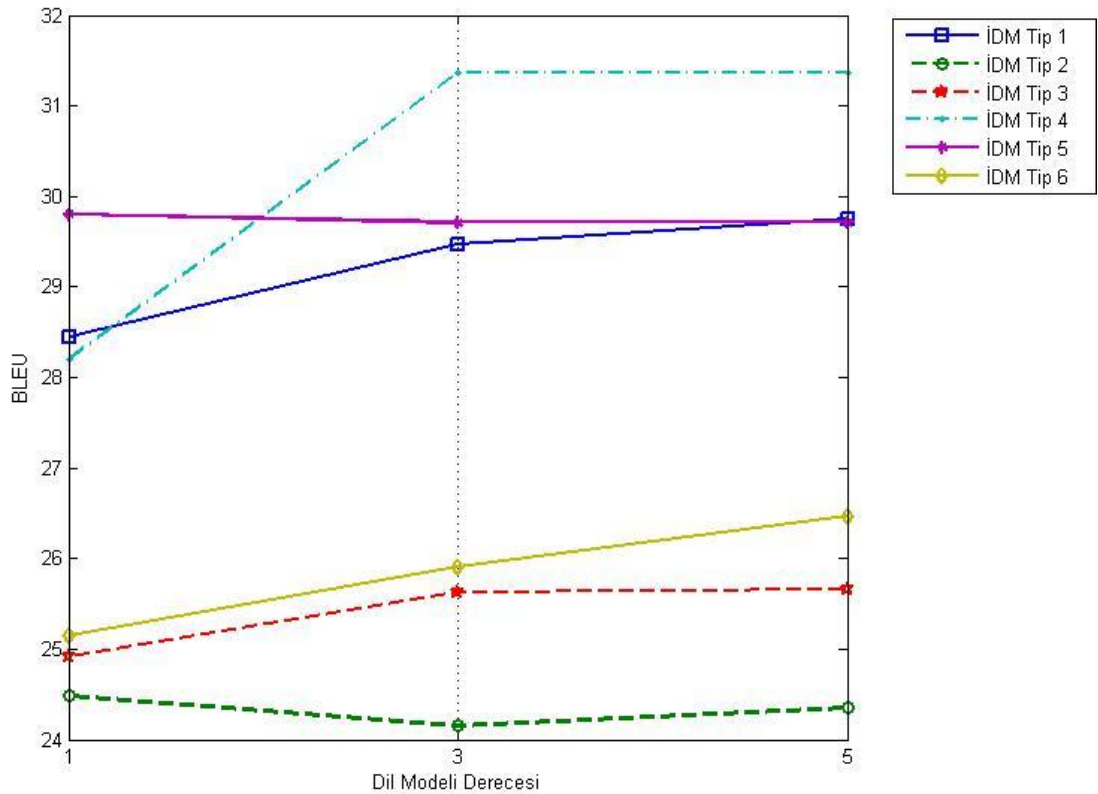
İDM Tip-VI'nın başarımlar tablosu incelendiğinde bütün durumlarda temel sistemin başarımlarının altında bir sonuç elde edildiği görülmektedir. Kuşkusuz bunda en önemli etken, kök bilgisi içermeyen bu İDM ile köklerin aktarımında hatalar yapılmasıdır. Bu durum, *BLEUr-1* puanları incelendiğinde açıkça gözlenmektedir.

Modeller arasında farklı dil modeli tiplerinin kullanılmasının incelenmesi amacı ile de aşağıdaki çizelgeler oluşturulmuştur. Sınanan dil modeli tiplerinin, Model 0 ile birlikte kullanıldığı çevirilerde ölçülen *BLEU* puanları Şekil 8.1'de verilmiştir:

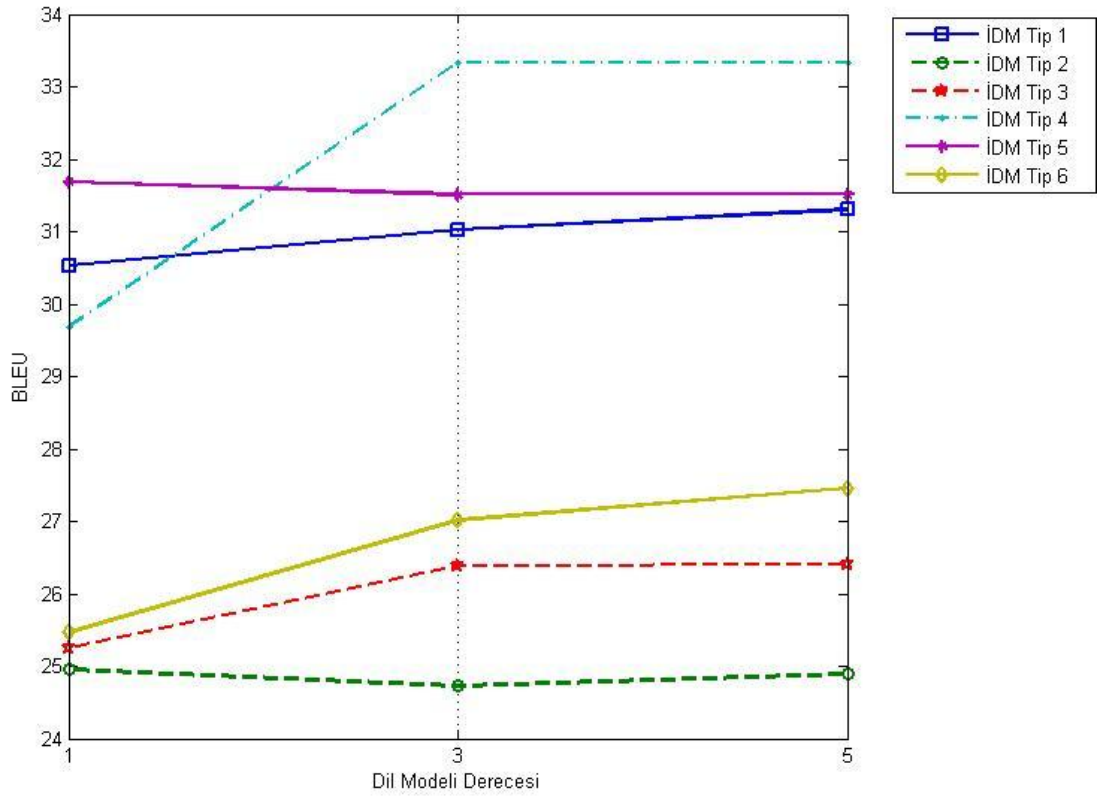
**Şekil 8.1 : Aktarım Modeli 0'nın Farklı İDM Tipleri İle Başarım Grafiği**

Çizelgeden de görülebileceği gibi, en başarısız İDM tipi, sözcük türlerinden oluşan Tip-II olurken en başarılı *BLEU* puanları kök ve son ÇG üzerinde üretilen Tip-IV olmuştur. Gözlenen bir başka sonuç ise kök kullanmayan İDM tipleri ile yapılan deneylerde (Tip-II ve Tip-III) *BLEU* sonuçlarının diğerlerine göre düşük çıkmasıdır. Bunun nedeni, sadece sözcük türünün ya da son ÇG'nin, çeviri sırasında ortaya çıkan belirsizlikleri çözmek için yeterli bilgiye sahip olmaması olarak gösterilebilir. Nitekim bu modellere kök bilgisi eklendiğinde (Tip-IV ve Tip-V) başarımlar gözle görülür bir şekilde yükselmektedir.

Şekil 8.2 ve Şekil 8-3'de, aktarım Modeli I ve aktarım Modeli II için farklı türlerdeki İDM'lerin başarımları çizelge olarak verilmiştir. Her iki çizelge incelendiğinde, Şekil 8.1'deki çizelge ile aynı yapıda oldukları gözlenmektedir. Model I çizelgesi Model 0'a göre, Model II çizelgesi de Model I'e göre yukarı ötelenmiştir. Bu da modellerin başarımları arasındaki farklardan kaynaklanmaktadır. Ancak İDM türlerinin başarımları arasındaki sıralama, bütün çizelgelerde aynı kalmaktadır.



Şekil 8.2 : Aktarım Modeli I'in Farklı İDM Tipleri İle Başarımların Grafiği



Şekil 8-3 : Aktarım Modeli II'nin Farklı İDM Tipleri İle Başarım Grafiği

Şimdiye kadar sunulan veriler ışığında, sonuçlar aşağıdaki gibi özetlenebilir:

1. Hangi dil modeli türü kullanılırsa kullanılsın, modeller arasında aşağıdaki sıralama (bir ya da iki durum dışında) değişmemiştir:

$$BLEU_{Model-II} > BLEU_{Model-I} > BLEU_{Model-0}$$

2. Sonuçlar, dil modeli derecesinin seçilmesi açısından değerlendirildiğinde, 5. dereceden İDM'lerin en başarılı sonuçları ürettikleri görülmüştür. Ancak 3. derece İDM'lerin başarımları, 5. derece İDM'lerden çok düşük olmadığı gibi, çeviri süresini yaklaşık olarak 2/3 oranında azalttığı için, İDM derecelerinin $n=3$ seçilmesinin getireceği başarı kaybı önemsenebilir.
3. Kök bilgisini içermeyen İDM tipleri kullanıldığında, sistemin başarısı temel sistemden geriye düşmektedir. Bunun ana nedeni, kök bilgisini içermeyen İDM tiplerinin sözcük kökü seçiminde çok hata yapmalarıdır. Ancak kök bilgisi eklendiğinde belirgin bir başarı artışı sağlanmaktadır.
4. Kök bilgisinin kullanılmadığı durumlarda, biçimbilimsel bilgi miktarı arttıkça başarımın her durumda arttığı gözlenmiştir:

$$\text{İDM}_{\text{Tip-VI}} > \text{İDM}_{\text{Tip-III}} > \text{İDM}_{\text{Tip-II}}$$

Ancak kök bilgisinin eklenmesi ile seyrek veri sorunu gözlenmeye başlandığından bu sıralama geçerliliğini kaybetmektedir.

5. Diğerlerine göre daha fazla bilgi içeren, sözcük kökü ile son ÇG'deki biçimbilimsel bilgiler kullanılarak oluşturulan Tip-V dil modeli, seyrek veri sorunundan etkilenmiş görülmektedir.
6. *BLEUr* ve *BLEU* puanları arasındaki fark, sözcük kökünün doğru seçilmesine rağmen biçimbilimsel belirsizliğin doğru giderilememesi ve buna bağlı olarak da sözcüğe ait yanlış bir yüzeysel biçimin oluşturulmasından kaynaklanmaktadır.

8.3.5 İki Seviyeli İDM Başarımı

Bölüm 7.4'de anlatılan nedenlerden dolayı iki seviyeli istatistiksel değerlendirme alt yapısı oluşturulmuş ve değişik İDM tipleri ile deneyler yapılmıştır. Deneyler yapılırken, aktarım fonksiyonu olarak Model II, İDM derecesi $n=3$, 1. seviyede üretilen çıktı sayısı $N=20$ olarak seçilmiştir. Tablo 8.12 ve Tablo 8-13'de, bu deneylerin başarımları, *BLEU* ve *BLEUr* puanları olarak verilmiştir. Tablo 8.12 ve Tablo 8-13'de zemini gri olan hücreler, ilgili dil modeli tipinin 1 seviyeli kullanımının *BLEU* ve *BLEUr* puanlarını göstermektedir.

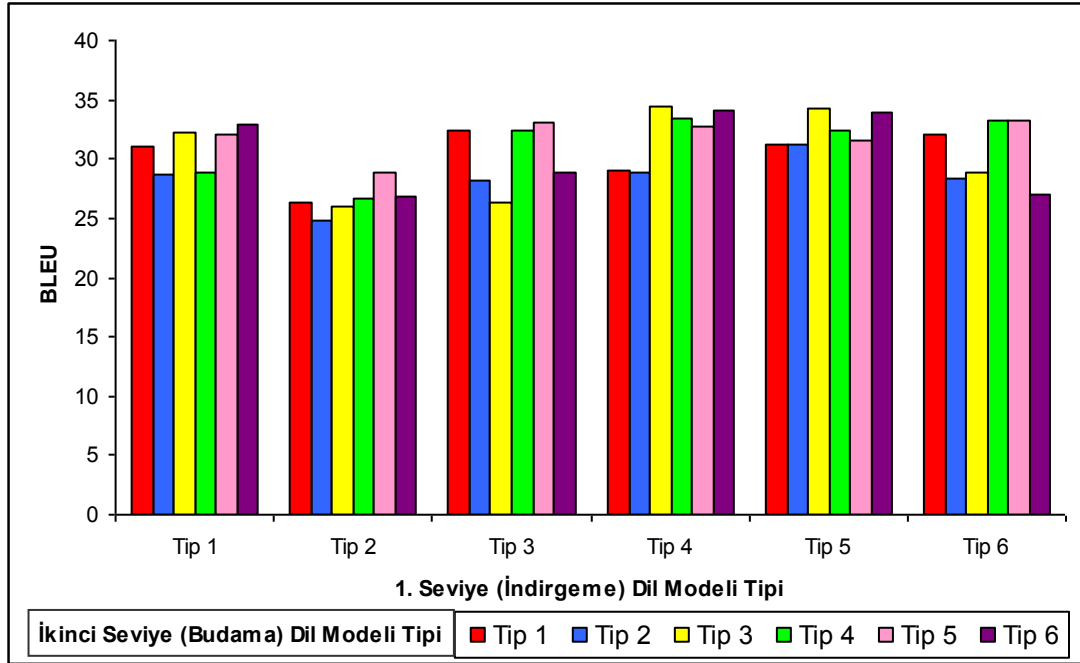
Tablo 8.12 : İki Seviyeli İDM İndirgemesinin *BLEU* Puanları

Model II n=3 N=20		2. Seviye İDM Türü					
		Tip-I	Tip-II	Tip-III	Tip-IV	Tip-V	Tip-VI
1. Seviye İDM Türü	Tip-I	31,03	28,61	32,32	28,82	31,99	32,92
	Tip-II	26,38	24,73	26,02	26,65	28,94	26,85
	Tip-III	32,46	28,18	26,39	32,48	33,12	28,81
	Tip-IV	29,04	28,83	34,36	33,34	32,78	34,03
	Tip-V	31,17	31,20	34,20	32,41	31,51	33,86
	Tip-VI	32,10	28,41	28,87	33,30	33,33	27,02

Tablo 8-13 : İki Seviyeli İDM İndirgemesinin BLEUr Puanları

Model II n=3 N=20		2. Seviye İDM Türü					
		Tip-I	Tip-II	Tip-III	Tip-IV	Tip-V	Tip-VI
1. Seviye İDM Türü	Tip-I	37,77	36,55	37,47	38,01	38,37	38,11
	Tip-II	33,80	28,13	28,43	33,76	32,79	29,02
	Tip-III	37,71	30,87	30,05	37,31	35,77	31,39
	Tip-IV	38,90	37,50	38,38	38,62	38,87	39,31
	Tip-V	37,13	35,05	36,16	37,64	34,46	36,21

Şekil 8-4’de ise BLEU puanlarının çizelgesi verilmiştir. Tablolar ve çizelge incelendiğinde, tek seviyeli İDM kullanımında temel sistem puanlarının altında kalan Tip-II ve Tip-III İDM türleri 1. seviyede, daha geniş kapsamlı modelleme yapan diğer İDM türleri ise 2. seviyede kullanılırsa başarımın arttırılabildiği gözlenmiştir.



Şekil 8-4 : İki Seviyeli İDM Kullanımında Başarım Karşılaştırmaları

Şimdiye dek yapılan tüm deneylerin sonucunda, önerilen çeviri modeli temelinde gerçekleştirilen Türkmence-Türkçe çeviri sisteminin aktarım fonksiyonu bileşeni olarak Model II’nin, İDM bileşeni olarak ise iki seviyeli bir kullanımla Tip-IV ve Tip-III’ün beraber kullanılmasının ($n=3$, $N=20$ parametreleri ile) en yüksek başarımı verdiği görülmüştür.

8.4 Hatalı Durumların İncelenmesi

Gerçeklenen bütün iyileştirmelere karşın, önerilen aktarım yönteminin istatistiksel tabanlı oluşu, eğitim derleminin kısıtlı olması ve sözlük girdilerinin yetersiz olması gibi bazı nedenlerden dolayı hatalı durumlar da oluşmaktadır. Bu bölümde kısaca sık rastlanılan bazı hata türleri üzerinde durulacaktır.

Sistem genelinde kullanılan aktarım sözlüğü, [72] kaynağından yararlanılarak gerçekleştirilmiştir. Uygulamanın gerçekçiliğinin korunması adına, sözlük girdilerine birebir sadık kalınmıştır. Dolayısı ile bu sözlük girdilerindeki eksik anlamlar ve karşılıklar çeviri sırasında bazı sözcüklerin karşılıklarının, referanslarda kullanılan sözcüklerden farklı olmasına neden olmuştur:

Türkmençe Söz.	Sözlükteki Karşılıklar	Ref. Tümcedeki Karşılık
gurna	topla, derle	sağla
bükgildi	çarpıntı, heyecan	tedirgin
mertebe	aşama, derece	seviye

Bu sorunun çözümü, sözlük girdilerinin ayrı bir çalışma ile yeniden düzenlenmesiyle sağlanabilir. İkinci bir çözüm önerisi ise, BalkaNET Projesi [73] kapsamında geliştirilen Türkçe kavramsal sözlükten yararlanarak, aynı anlama sahip sözcüklerden bir karşılıklar kümesinin dinamik olarak oluşturulmasıdır.

Tümce genelinde çalışan kurallar genellikle sözcükler arasındaki ünlü uyumunu düzeltmek amaçlıdır. Örneğin Türkçede isimden ayrı yazılan **de/da** bağlacı ve soru ekleri **mi/mı/mü/mu** bu kullanımlara örnektir. Ancak bazı durumlarda yazımda görülen harf ile konuşma dilinde kullanılan harf farklılık göstermektedir:

Son araba **da** geçti. (1)
Son harf **de** tüketildi. (2)

Görüldüğü gibi **de/da** bağlacından önceki her iki (**araba** ve **harf**) sözcüğün son seslisi (**a**), yazı dilinde kalın olmasına rağmen bu sözcüklerin okunuşları farklı olduğundan takip eden bağlacın seslisi de değişiklik gösterir. İkinci örnekte bulunan **a** harfi aslında Türkçe sesçil abecede bulunan (art a - [α]) ve kalından çok ince olarak seslendirilen bir harftir [49]. Ancak aktarım kurallarının, yazımda değişiklik göstermeyen bu farklılığı sezmesi olası olmadığı için aşağıdaki hatalı çeviri üretilmektedir:

Son harf **da** tüketildi.

8.5 Çeviri Örnekleri

Bu bölümde bazı çeviri örnekleri verilerek, gerçekleştirilen sistemin başarısı ile *BLEU* puanlama sisteminin uyumu incelenecektir. Örneklerde kaynak tümce, sistem çıktısı aday tümce ve *BLEU* puanlamasında kullanılan iki adet referans tümce alt alta verilmiştir.

İlk olarak, üretilen aday çeviri ile, referans tümcelelerin anlamının büyük ölçüde sağlandığı, ancak *BLEU* tarafından düşük puanlandığı durumlara ilişkin örnekler sunulacaktır.

Kaynak : bu mesele bilen adamlar gyzyklyanypdyrlar , emmä ol soraga her taraply jogap berip_bilmändirler .

Aday Çeviri : bu meseleyle insanlar ilgilenmişler , ama o soruya her yönlü cevap verememişler .

Referans 0 : bu mesele ile insanlar ilgilenmişlerdir , ama bu soruya her yönü ile cevap verememişlerdir .

Referans 1 : bu meseleyle insanlar ilgilenmişlerdir , ama bu soruya her yönü ile cevap verememişlerdir .

Kaynak tümcede “_” işareti ile birleştirilen sözcükler, uygulama tarafından bir ÇSG olarak işaretlenerek Türkçeye doğru olarak çevrilebilmiştir. Aday çeviri ve referans karşılaştırıldığında anlamsal açıdan çok büyük bir farklılık olmamasına karşın, bu tümcenin *BLEU* puanının 25,82 çıkması, başarı puanlamasında sözcüklerin yüzeysel biçimlerinin eşleşmelerinin ne kadar önemli olduğunun bir göstergesi olarak yorumlanmalıdır. Benzer şekilde, çeviri ile referans arasında hemen hemen hiçbir anlam farkı içermeyen aşağıdaki örnek için dahi *BLEU* puanı 56,69 olarak hesaplanmıştır:

Kaynak : bu dana adam birden gülmesini goýup aýlamaga başlapdyr.

Aday Çeviri : bu bilgin adam birden gülmesini bırakarak ağlamaya başlamış.

Referans 0 : bu bilgin insan birden gülmesini kesip ağlamaya Başlamış.

Referans 1 : bu bilgin insan birden gülmesini bırakıp ağlamaya başlamış.

Bu nedenle genel sistemin *BLEU* puanı yaklaşık 34 iken, aslında üretilen çevirilerin referanslarla uyumunun çok daha yüksek olduğu yorumu getirilebilir.

İncelenmeye değer bir diğer durum ise, sistem tarafından üretilen bazı aday tümcelerde, referanslarda geçen sözcüklerin eş anlamlılarının tercih edilmesi nedeniyle *BLEU* puanının düşük çıkması durumudur. Sözcük kökünün aktarımındaki seçim farklılıklarından kaynaklanan durumlar, *BLEU* yönteminde çok sayıda referans tümce kullanılarak düzeltilmeye çalışılmaktadır. Ancak bizim değerlendirme yaptığımız tümce kümesinde, her tümce için iki adet referans çeviri bulunmakta ve bazı eş anlamlı sözcükler referanslarda geçmediğinden toplam *BLEU* puanlamasını düşürmektedir. Aşağıda bu duruma uyan bir örnek tümce verilmiştir:

Kaynak : mertebesini bilmeýän adamlary diri hasap etmegin .
Aday Çeviri : derecesini bilmeyen insanlary sađ hesap etme .
Referans 0 : seviyesini bilmeyen adamlary diri sayma .
Referans 1 : mertebesini bilmeyen adamlary diri hesap etme .

Yukarıdaki örnek tümcede, “**mertebe**” Türkmence sözcüğünün karşılığı olarak sistem “**derece**” kökünü seçmiştir. Ancak bu kök, referans tümcelerin ikisinde de kullanılmamıştır. Benzer şekilde aday tümce oluşturulurken sistemin tercih ettiği “**insan**” ve “**sađ**” sözcük kökleri yerine referans tümcelerde farklı sözcükler kullanılmıştır. Bu hali ile 31.62 olarak hesaplanan *BLEU* puanı, aday tümcedeki “**derece**”, “**adam**” ve “**sađ**” kökleri sırası ile “**mertebe**”, “**insan**” ve “**diri**” kökleri ile değiştirilince 75.98 olarak hesaplanmaktadır. Burada çıkartılacak sonuç ise, sistemin çıktılarının referanslarla uyumunun, *BLEU* puanından daha yüksek bir derecede olduğudur.

Çeviriler incelendiğinde görülen sorunlardan bir diğeri de kısa tümcelerde hata yapma oranının artmasıdır. Aşağıda bu duruma uyan bir örnek verilmiştir:

Kaynak : nebsim janym agyrdy .
Aday Çeviri : çok kişim ađrıdı .
Referans 0 : buna çok canım sıkıldı .
Referans 1 : çok canım sıkıldı .

Bu örnekte, Türkmence “**jan**” sözcüğü “**kişi**” anlamı ile aktarılmıştır. Yine benzer bir hata aşağıdaki örnekte de görülebilir:

Kaynak : hem gülki hem gözýaş !
Aday Çeviri : da gülme de gözýaşı !

Referans 0 : hem kahkaha hem gözyaşı !

Referans 1 : hem kahkaha hem gözyaşı !

Bu örnekte ise, Türkçede “**de**” ve “**hem**” bağlaçlarının görevini gören Türkmence “**hem**” sözcüğü aktarılırken “**de**” bağlacı anlamıyla aktarılmıştır. Özellikle tümcenin başındaki sözcükler için yeterli sayıda geçmiş sözcük bulunmadığı bunlara benzer kısa tümcelerde, yukarıdaki gibi hatalı çeviri örnekleri görülebilmektedir.

8.6 Aktarım Süreleri

Bulunan en iyi parametrelerle çalıştırılması durumunda, geliştirilen çeviri sisteminin çıktı üretmesi için geçen toplam süre ve her bir bileşenin işlem süresi, Tablo 8.14’de verilmiştir. Tablodaki sürelerin hesaplanması için, çeviri sistemi, standart bir kişisel bilgisayar ¹³ üzerinde, 255 giriş tümcesi için 10 defa çalıştırılmış ve her çalışma sonunda ölçülen sürelerin ortalaması alınmıştır ¹⁴.

Tablo 8.14 : Aktarım Süreleri

Tümce/Sözcük Ayırıcı	0,67 s
Türkmence Biçimbilimsel Çözümleme	0,19 s
ÇSG İşleyici	8,79 s
Biçimbilimsel Yapıların Aktarımı	0,82 s
Kök Aktarımı	1,61 s
1. Seviye İDM	5,45 s
2. Seviye İDM	22,92 s
Yapısal Biçim Düzeyinde Tümce Geneli Kurallar	3,45 s
Türkçe Biçimbilimsel Üretici	44,98 s
Yüzeysel Biçim Düzeyinde Tümce Geneli Kurallar	0,56 s
Toplam Süre	88,79 s

Tablo incelendiğinde, sistemin saniyede 2,88 tümce çevirdiği hesaplanabilmektedir. Bileşenlerin çalışma süreleri karşılaştırıldığında, Türkçe biçimbilimsel üreticinin en çok süreyi kullandığı görülmektedir. Türkmence biçimbilimsel çözümleyici ise çok

¹³ 1.86 GHz frekanslı çift çekirdekli işlemci, 1 GB ana bellek

¹⁴ İDM dosyalarının yüklenmesi için gereken süre hesaplamalara dahil edilmemiştir.

daha kısa sürede çalışmaktadır. Türkçe için kullanılan üretici araç, genel kullanıma açık geniş kapsamlı bir araçtır. Oysa Türkmence için geliştirilen biçimbilimsel çözümleyicinin sözlük boyutu, sadece sınama kümesindeki sözcükleri işleyebilecek kadar dar kapsamlıdır. İkinci seviye İDM bileşeninin, birinci seviye İDM'den daha uzun sürmesinin nedeni ise birinci seviyede HMM ve geliştirilmiş Viterbi algoritması ile çabucak sonuca gidebilirken, ikinci seviyede üretilen N aday tümce üzerine bir HMM oluşturulamamasıdır. İkinci seviyede oluşan N aday tümcenin her birisi için olasılık ayrı ayrı hesaplanması işlemi de uzun sürmektedir. Sistemin kapsamının genişlemesi durumunda Türkmence biçimbilimsel çözümleme ve kök aktarımı bileşenlerinin çalışma sürelerinin uzayacağı kesindir. Ayrıca tüm bileşenlerin çalışma süreleri, giriş tümcesi ile orantılı olarak artmaktadır.

9. DEĞERLENDİRME VE TARTIŞMA

Bu tez çalışması kapsamında, dilbilgisel açıdan benzerlikler taşıyan akraba diller arasında bilgisayarlı çeviri yöntemleri incelenmiş, özellikle seçilen dil çiftinin bitişken olması durumu göz önünde bulundurularak bir çeviri yöntemi önerilmiştir. Önerilen çeviri yöntemi, hem bilgi tabanlı ve kural temelli hem de istatistiksel bileşenlerden oluşan karma (hibrid) bir modeldir. Çoğu kez, kural tabanlı çeviri sistemlerinin kural havuzu çok karmaşık bir hale gelir ve zamanla yönetilmesi zorlaşır. Diğer yandan istatistiksel sistemler, tasarımcıları bu karmaşadan kurtarmaktadır ancak başarılı bir çeviri için yoğun miktarda çift dilli tümce derlemelerinin hazırlanmasını şart koşmaktadır. Bu tür bir derlemin hazırlanamadığı durumlarda, istatistiksel yöntemler kullanılamamaktadır. Önerdiğimiz sistem, akraba diller arasında, göreceli olarak basit kurallarla, belirli bir belirsizlik düzeyi korunarak çeviri yapılmasını ve belirsizliklerin hedef dilde oluşturulan istatistiksel yapılarla giderilmesini öngörmektedir. Böylelikle akraba dil çiftleri için istatistiksel ve bilgi tabanlı çeviri yöntemlerinin olumlu özellikleri birleştirilerek bir çeviri modeli önerilmiştir. Bu çeviri modeli, özellikle bitişken dil çiftleri arasında çeviri yapacak biçimde tasarlanmıştır.

Önerilen çeviri modeli iki temel bileşenden oluşmaktadır :

- 1) aktarım fonksiyonu ve
- 2) hedef dilde istatistiksel dil modeli.

Aktarım fonksiyonunun tasarımında, özellikle bitişken diller için üç değişik model geliştirilmiştir. Hedef dil için istatistiksel dil modeli oluşturulması konusunda da, bitişken dillerin özellikleri göz önünde bulundurularak altı farklı İDM tipi tanımlanmıştır.

Aktarım fonksiyonu, biçimbilimsel çözümleme destekli doğrudan aktarım olarak düşünülebilir. Kaynak dildeki sözcüklerin biçimbilimsel çözümlemesi yapıldıktan sonra çift dilli sözlük yardımı ile sözcük köklerinin aktarımı sağlanır. Daha sonra çözümlemelerdeki biçimbilimsel yapılar hedef dile aktarılır. Bu fonksiyon içerisinde

iki farklı belirsizlik kaynağı bulunmaktadır. Birincisi biçimbilimsel çözümlemeye, ikincisi ise sözcük kökünün aktarılmasında ortaya çıkmaktadır.

Yöntemin ikinci bileşeni olan istatistiksel dil modeli ile de bu belirsizlikler, istatistiksel yöntemlerle giderilmeye çalışılarak en yüksek olasılığa sahip tümce, aday çeviri olarak üretilir. İstatistiksel çeviride kullanılan İDM ile benzer şekilde kullanılan bu bileşenin, sadece hedef dil üzerinde hazırlanması yeterlidir. Böylelikle her iki dil arasında çok sayıda çift dilli tümcenin hizalanması ile oluşturulacak bir eğitim kümesine gereksinim duyulmamaktadır.

Akraba diller arasında önerdiğimiz bu çeviri modeli, Türk Dilleri arasında çeviri amacıyla irdelenmiş, aktarım modellerinin ve dil modeli tiplerinin başarımlarını ölçmek amacıyla Türkmence-Türkçe dil çifti için bir çeviri sistemi gerçekleştirilmiştir. Gerçeklenen sistemle, önerilen üç aktarım fonksiyonu modeli ve altı farklı İDM tipi kullanılarak sına derlemi üzerinde başarımlar ölçülmüştür.

Önerdiğimiz modeller, kaynak dilde biçimbilimsel belirsizlik giderici bulunmasını zorunlu kılmamaktadır. Önceki çalışmalardan farklı olarak biçimbilimsel belirsizlik hedef dilde giderilmektedir. Bu özellik, biçimbilimsel belirsizlik giderici araçları bulunmayan Türkçe dışındaki diğer Türk Dilleri arasında çeviri yapabilmek için çok önemlidir.

Yapılan deneyler sonucunda, geliştirilen modeller ve gerçekleştirilen sistem, sına derlemindeki Türkmence tümceleri, referans tümcelerdeki anlamları da koruyacak şekilde, Türkçeye aktarabilmiştir. Bu da, birçok bilgisayarlı çeviri sisteminin geliştirilmesinde harcanan emekten çok daha az bir çalışma ile Türk Dilleri arasında çeviri yapabilen sistemlerin tasarlanabileceğini göstermektedir. Her ne kadar tez çalışması kapsamında Türk Dilleri incelenmiş olsa da, geliştirilen modeller, bütün akraba ve bitişken özellikli dil çiftleri için uygulanabilir niteliktedir.

Sonuçlar incelendiğinde, önerilen aktarım fonksiyonlarından Model II'nin, iki seviyeli İDM kullanımı ile (1.Seviye : İDM Tip-IV, 2.Seviye : İDM Tip-III, $N=20$, $n=3$ paramterleriyle) birlikte en yüksek başarıyı sağladığı gözlenmiştir.

Uygulamada, İDM hariç diğer tüm bileşenler SDD'ler ile yapılmış olduğundan, gerçekleştirilen çeviri sistemi ters yönde de rahatlıkla kullanılabilir. Doğal olarak ters yönde kullanımda hedef dil için İDM'nin yeniden oluşturulması gereklidir. Buradan hareketle Türkçeden Türkmenceye çeviri sisteminin gerçekleştirilmesi çok zor

görünmemektedir. Üstelik Türkçe için biçimbilimsel belirsizlik çalışmaları bulunduğundan, sistemin çözmek zorunda kalacağı biçimbilimsel belirsizlik de yoktur, İDM sadece sözcüksel belirsizliğin çözümünde kullanılır.

Gerçeklenen tez çalışmasının bilimsel katkısı, aşağıdaki maddelerle özetlenebilir:

- Akriba ve bitişken dil çiftleri arasında bilgisayarlı çeviri için kural tabanlı ve istatistiksel bileşenlerden oluşan karma bir yöntem önerilmiştir.
- Önerilen modeller çerçevesinde Türk Dilleri arasında bilgisayarlı çeviri konusu incelenmiştir.
- Önerilen modellerin etkinliği ve başarımı, Türkmenceden Türkçeye çeviri yapan bir sistem gerçekleştirilerek sınanmıştır.
- Geliştirilen sistem ile, diğer Türk Dilleri arasında bilgisayarlı çeviri gerçekleştirmek için de kullanılabilecek bir altyapı ortaya konulmuştur.

Çalışma sırasında karşılaşılan çeşitli güçlükler ve sorunlar aşağıda sıralanmıştır:

- İstatistiksel yöntemlerin en büyük gereksinimi, yüklü miktarda eğitim verisidir. İngilizce, Almanca gibi yaygın kullanılan bir çok dil için bu tür eğitim derlemleri kullanıma hazır ve erişilebilir bir durumdayken, Türkçe için kullanılan derlemin büyüklüğü henüz yeterli seviyede değildir. Uygulamada yaklaşık 1M çözümlenmiş sözcük üzerinden çıkartılan İDM, hem sözcük köklerinin hem de karşılaşılan biçimbilimsel çözümleme tiplerinin çeşitliliği yüzünden seyrek veri sorunundan değişik oranlarda etkilenmektedir.
- Türk Dilleri, özellikle uygulamanın geliştirildiği Türkmence dili hakkında doğal dil işleme yaklaşımı dil bilgisi kaynaklarının yetersiz olması, kaynak Türk dilinde biçimbilimsel çözümleyicinin gerçekleştirilmesi ve aktarım kurallarının hazırlanmasında büyük sıkıntılara yol açmıştır.
- Türkçede belirli bir ölçekte olmasına karşın, diğer bir çok Türk Dili üzerinde yeterli sayıda bilgisayarlı dil işleme çalışması ve kullanılabilir araçların olmaması, sistemin geliştirme süresinin uzamasını ve geliştirilen sistemin başarısını düşürmektedir. Örneğin Türkçe dışındaki hemen hemen hiçbir dil için (Kırım Tatarcası hariç) kullanılabilir genel amaçlı bir biçimbilimsel çözümleyiciye erişilememiştir. Benzer şekilde biçimbilimsel belirsizlik

giderilmesi için de bir araç olmadığı için bu biçimbilimsel belirsizlik bir çok aşama boyunca taşınmakta ve potansiyel hatalı karşılıklar üretilmektedir.

- Önerilen modellerin ve İDM tiplerinin başarımlarının karşılaştırılmasında kullanılan BLEU ölçütü, Türk dillerinin yapısına çok uygun değildir. Değişken sözcük/sözcük öbeği sıralaması, zengin yapım ve çekim ekleri dolayısı ile sözcüklerin yüzeysel biçimlerinin farklı oluşabilmesi, BLEU değerlendirme sistemi kapsamında ciddi puan kayıplarına neden olmaktadır.
- BLEU puanlarının hesaplanması için yaygın olarak dört referans tümce kullanılırken, bizim bulabildiğimiz referans çeviri sayısı iki olmuştur. Ayrıca daha sağlıklı bir değerlendirme için sınama derleminin boyutu da arttırılmalıdır.

Çalışma kapsamında eksik görülen ve ileriki çalışmalarda tamamlanması beklenen noktalar şunlardır:

- Gerçeklenen uygulama sisteminde, Türkmence için tasarlanan biçimbilimsel çözümleyici, SDD temelinde tasarlanmış olduğundan ters yönde üretim amacıyla da kullanılabilir. Bu sayede aynı araç, örneğin Türkçeden Türkmenceye çeviri yaparken de kullanılabilir. Ancak bu çalışma kapsamında sadece çözümleme yönünde kullanıldığı için, ters yönde çalıştırılması durumu üzerinde fazla durulmamıştır. Geliştirilen Türkmence biçimbilimsel çözümleyici SDD ters yönde üretici olarak çalıştırılırsa, geçersiz ve fazlalık bazı çıktılar üretmektedir. Bu konuda bir çalışma yapıp Türkmence biçimbilimsel çözümleyici/üretici aracın üretim yönündeki başarımı yükseltilmelidir.
- Geliştirilen Türkmence biçimbilimsel çözümleyicinin güncel sürümü, sadece sınama kümesindeki sözcük kökleri ile yaygın kullanılan bazı sözcük köklerini içermektedir. Genel amaçlı bir sistemin geliştirilmesi için alt sözlükler, tüm sözcük köklerini kapsayacak şekilde güncellenmelidir.
- Benzer şekilde geliştirilen sistemin aktarım sözlüğü, sadece sınama derlemindeki sözcüklerin köklerini içermektedir. Genel amaçlı bir sistemin tasarımı için bu aktarım sözlüğünün kapsamının genişletilmesi gereklidir.

Gerçekleřtirdiđimiz çalıřmayı geliřtirmek üzere sonradan yapılabilircek tamamlayıcı nitelikli çalıřmalar ařađıda verilmiřtir.

- Diđer Türk Dili çiftlerinde çeviri yapılabilmesi için, her Türk Dili için biçimbilimsel çözümleyicinin ve üreticinin gerçekenmesi uygun olacaktır. Geliřtirilecek araçlar, çeviri sistemlerinin geliřtirilmesinde kullanılabilirceđi gibi diđer bir çok DDİ çalıřması için mutlak yapılması gereken bir işlevi yerine getirecektir.
- Farklı türlerde (haber, hikaye, siyasi,...) metinlerden ve bunların referans çevirilerinden oluřan yeni bir sınaama derlemi oluřturarak sistemin, bu genel konulu derlemdeki başarısının ölçülmesi, sistemin genel kullanıma açılması yönünde önemli bir deđerlendirme olanađı sađlayacaktır.

10. KAYNAKLAR

- [1] **Banguoğlu, T.**, 2000. Türkçenin Grameri, Türk Dil Kurumu.
- [2] **Nagao, M.**, 1984. A Framework of a Mechanical Translation Between Japanese and English by Analogy Principle in *Artificial and Human Intelligence*, Eds Elithom and Banerji, North-Holland.
- [3] **Hajič, J.**, 1987. RUSLAN - An MT System Between Closely Related Languages, *Third Conference of the European Chapter of the Association for Computational Linguistics (EACL'87)*, Copenhagen, Denmark.
- [4] **Hajič, J., Hric, J. and Kuboň, V.**, 2000. Machine Translation of Very Close Languages, *Proceedings of the Sixth Conference on Applied Natural Language Processing*
- [5] **Hajič, J., Homola, P. and Kuboň, V.**, 2003. A simple multilingual machine translation system, *MT Summit IX*, New Orleans, USA.
- [6] **Dvořák, B., Homola, P. and Kuboň, V.**, 2006. Exploiting similarity in the MT into a minority language, *LREC-2006: Fifth International Conference on Language Resources and Evaluation. 5th SALTMIL Workshop on Minority Languages: "Strategies for developing machine translation for minority languages"*, Genoa, Italy.
- [7] **Canals-Marote, R., Esteve-Guillén, A., Garrido-Alenda, A., Guardiola--Savall, M. I., Iturraspe-Bellver, A., Montserrat-Buendia, S., Pérez-Antón-Rojas, P., Ortiz-Pina, S., Pastor-Antón, H. and Forcada, M. L.**, 2000 interNOSTRUM: a Spanish-Catalan Machine Translation System, *Machine Translation Review*, **11**, 21-25.
- [8] **M. Corbi-Bellot, A., Forcada, M. L., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Alegria, I., Mayor, A. and Sarasola, K.**, 2006. An open-source shallow-transfer machine translation engine for the Romance languages of Spain, *10th EAMT*

conference "Practical applications of machine translation", Budapest, Hungary.

- [9] **Oller, C. A. i. and Forcada, M. L.**, 2006. Open-source machine translation between small languages : Catalan and Aranese Occitan, *LREC-2006: Fifth International Conference on Language Resources and Evaluation. 5th SALT MIL Workshop on Minority Languages: "Strategies for developing machine translation for minority languages"*, Genoa, Italy.
- [10] **Tomás, J. and Casacuberta, F.**, 2001. Monotone statistical translation using word groups, *MT Summit VIII: Machine Translation in the Information Age*, Santiago de Compostela, Spain.
- [11] **Hamzaoglu, İ.**, 1993. Machine translation from Turkish to other Turkic languages and an implementation for the Azeri languages, *MSc Thesis*, Bogazici University Institute for Graduate Studies in Science and Engineering, İstanbul.
- [12] **Altıntaş, K.**, 2000. Turkish to Crimean Tatar Machine Translation System, *MSc Thesis*, Bilkent University Computer Engineering Department, Ankara.
- [13] **Oflazer, K.**, 1995 Two-level Description of Turkish Morphology, *Literary and Linguistic Computing*, **9**, 137-148.
- [14] **Altıntaş, K. and Çiçekli, İ.**, 2001. A Morphological Analyser for Crimean Tatar, *Proceedings of the 10th Turkish Symposium on Artificial Intelligence and Neural Networks, TAINN*, North Cyprus.
- [15] **Koskenniemi, K.**, 1983. Two-Level Morphology : A General Computational Model for Word Form Recognition and Production, Department of General Linguistics, University of Helsinki.
- [16] **Karttunen, L.**, 1983. KIMMO : A General Morphological Processor, *Texas Linguistic Forum*, Texas, USA, 163-186.
- [17] **Antworth, E. L.**, 1990. PC-KIMMO: A Two-Level Processor for Morphological Analysis, Summer Institute of Linguistics, Dallas, Texas, USA.

- [18] **Sproat, R.**, 1992. Morphology and Computation, MIT Press.
- [19] **Karttunen, L. and Wittenburg, K.**, 1983. A Two-Level Morphological Analysis of English, *Texas Linguistic Forum*, Texas, USA, 217-228.
- [20] **Khan, R.**, 1983. A Two-Level Morphological Analysis of Rumanian, *Texas Linguistic Forum*, Texas, USA, 253-270.
- [21] **Koskenniemi, K.**, 1985. An Application of the Two-Level Model to Finnish, University of Helsinki Department of General Linguistics.
- [22] **Lun, S.**, 1983. A Two-Level Morphological Analysis of French, *Texas Linguistic Forum*, Texas, USA, 271-278.
- [23] **Karttunen, L., Gaal, T. and Kempe, A.**, 1997. Xerox Finite-State Tool, XEROX Research Centre, Europe, Technical Report.
- [24] **Beesley, K. R. and Karttunen, L.**, 2003. Finite State Morphology, CSLI Publications, Stanford.
- [25] **Bahl, L. R., Jelinek, F. and Mercer, R. L.**, 1983 A Maximum Likelihood Approach to Continuous Speech Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **5**, 179-190.
- [26] **Jurafsky, D. and Martin, J. H.**, 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Prentice Hall
- [27] **Jeffreys, H.**, 1948. Theory of Probability, Clarendon Press, Oxford.
- [28] **Witten, I. H. and Bell, T. C.**, 1991 The Zero-Frequency Problem : Estimating the Probabilities of Novel Events in Adaptive Text Compression, *IEEE Transactions on Information Theory*, **37**, 1085-1094.
- [29] **Gale, W. A.**, 1994. Good-Turing Smoothing without Tears, Bell Labs.
- [30] **Katz, S. M.**, 1987 Estimation of Probabilities from Sparse Data for Language Model Component of a Speech Recognizer, *IEEE Transactions on Acoustics, Speech and Signal Processing*, **35:3**, 400-401.
- [31] **Manning, C. D. and Schütze, H.**, 1999. Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge.

- [32] **Goodman, J. T.**, 2001. A Bit of Progress in Language Modeling Extended Version, Microsoft Research, Redmond, USA.
- [33] **Jelinek, F., Mercer, R. L., Bahl, L. and Baker, J. K.**, 1977 Perplexity - A Measure of the Difficulty of Speech Recognition Tasks, *Journal of the Acoustical Society of America*.
- [34] **Chandioux, J.**, 1976 MÉTÉO : un système opérationnel pour la traduction automatique des bulletins météorologiques destinés au grand public., *Meta*, **21**, 127-133.
- [35] **Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L. and Roossin, P. S.**, 1990 A Statistical Approach to Machine Translation, *Computational Linguistics*, **16**, 79-85.
- [36] **Brown, P. F., Pietra, V. J. D., Pietra, S. A. D. and Mercer, R. L.**, 1993 The mathematics of statistical machine translation: parameter estimation, *Computational Linguistics* **19**, 263 - 311.
- [37] **Koehn, P.**, 2003. Noun Phrase Translation, *PhD Thesis*, University of Southern California, Los Angeles.
- [38] **Brown, R. D.**, 1996. Example-Based Machine Translation in the Pangloss System, *The 16th International Conference on Computational Linguistics (COLING-96)*, Copenhagen, Denmark.
- [39] **Guvenir, H. A. and Cicekli, I.**, 1998 Learning Translation Templates from Examples, *Information Systems*, **23**, 353-363.
- [40] **Somers, H.**, 1999. Review Article: Example-based Machine Translation.
- [41] **Papineni, K., Roukos, S., Ward, T. and Zhu, W. J. J.**, 2002. BLEU : A Method for Automatic Evaluation of Machine Translation, *Association of Computational Linguistics, ACL'02*, Philadelphia, PA, USA.
- [42] **NIST Report**, 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics.

- [43] **Callison-Burch, C., Osborne, M. and Koehn, P.**, 2006. Re-evaluating the Role of BLEU in Machine Translation Research, *Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, Trento, Italy.
- [44] **Melamed, I. D., Green, R. and Turian, J. P.**, 2003. Precision and Recall of Machine Translation, *HLT-NAACL 2003*.
- [45] **Banerjee, S. and Lavie, A.**, 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, MI, USA.
- [46] **Oflazer, K., Çetinoğlu, Ö. and Say, B.**, 2004. Integrating Morphology with Multi-word Expression Processing in Turkish, *The ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, Barcelona, Spain.
- [47] **Hakkani-Tür, D. Z., Oflazer, K. and Tür, G.**, 2002 Statistical Morphological Disambiguation for Agglutinative Languages, *Computers and the Humanities*, **36**, 381-410.
- [48] **Oflazer, K.**, 1999. Dependency Parsing with a Extended Finite State Approach, *College Park, Maryland*.
- [49] **Hengirmen, M.**, 2000. Türkçe Dilbilgisi, Engin Yayıncılık, Ankara.
- [50] **Tekin, T.**, 2006 Türk Dilleri Ailesi, www.turkoloji.com.tr.tc.
- [51] <http://www.ethnologue.com>, 2006.
- [52] **Oflazer, K. and Kuruöz, İ.**, 1994. Tagging and Morphological Disambiguation of Turkish Text, *The Fourth ACL Conference on Applied Natural Language Processing*, Stuttgart.
- [53] **Oflazer, K. and Tür, G.**, 1996. Combining Hand-crafted Rules and Unsupervised Learning in Constraint-based Morphological Disambiguation, *Conference on Empirical Methods in Natural Language Processing, EMNLP*, Somerset, New Jersey.

- [54] **Yüret, D. and Türe, F.**, 2006. Learning Morphological Disambiguation Rules for Turkish, *North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL 2006)*, New York City, June, 2006.
- [55] **Fomey, G. D.**, 1973 The Viterbi Algorithm, *IEEE Proceedings*, **61**, 268-278.
- [56] **Tür, G.**, 2000. A Statistical Information Extraction System for Turkish, *PhD Thesis*, Bilkent University The Department of Computer Engineering, Ankara.
- [57] **Tantuğ, A. C., Adalı, E. and Oflazer, K.**, 2006. A Prototype Machine Translation System Between Turkmen and Turkish, *Fifteenth Turkish Symposium on Artificial Intelligence and Neural Networks, TAINN*, Gökova, Muğla, Türkiye.
- [58] **Tantuğ, A. C., Adalı, E. and Oflazer, K.**, 2006. Computer Analysis of the Turkmen Language Morphology, *FinTAL, Lecture Notes in Computer Science*.
- [59] **Arnazarow, S., Borjakow, A., Saruhanow, M., Söyegow, M. and Hojayew, B.**, 2000. Türkmen Dilinin Grammatikasy, Türk Dil Kurumu, Ankara.
- [60] **Kara, M.**, 2000. Türkmençe (Giriş-Gramer-Metinler-Sözlük), Kültür Bakanlığı Yayınları, Ankara.
- [61] **Sarı, B. and Güder, N.**, 1998. Türkmencenin Grameri - I (Fonetika-Ses Bilgisi), Türk Dünyası Gençlerinin Mahtumkulu Yayın Birliği.
- [62] **Sarı, B. and Güder, N.**, 1998. Türkmencenin Grameri - II (Morfologiya – Şekil Bilgisi), Türk Dünyası Gençlerinin Mahtumkulu Yayın Birliği.
- [63] **Sarı, B. and Güder, N.**, 1998. Türkmencenin Grameri - III (Sintaksis), Türk Dünyası Gençlerinin Mahtumkulu Yayın Birliği.
- [64] **Clark, L. V.**, 1998. Turkmen reference grammar, Harrassowitz Verlag, Wiesbaden.
- [65] **Clarkson, P. and Rosenfeld, P. R.**, 1997. Statistical Language Modeling Using CMU-Cambridge Toolkit, *ESCA Eurospeech'97*.

- [66] **Stolcke, A.**, 2002. SRILM - An Extensible Language Modeling Toolkit, *International Conference on Spoken Language Processing*, Denver, Colorado.
- [67] **Chow, Y.-L. and Schwartz, R.**, 1989. The N-Best Algorithm: An Efficient Procedure for Finding Top N Sentence Hypotheses, *Proceedings of a Workshop on Speech and Natural Language*, Philadelphia.
- [68] **Oliveira, L. S., Bortolozzi, R. S. F. and Suen, C. Y.**, 2002 Automatic Recognition of Handwritten Numerical Strings: A Recognition and Verification Strategy, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**, 1438-1554.
- [69] **Oliviera, L. E. S.**, 2003. Automatic Recognition of Handwritten Numerical Strings, *PhD Thesis*, Université du Québec École de Technologie Supérieure, Quebec.
- [70] **Oflazer, K. and Tür, G.**, 1997. Morphological Disambiguation by Voting Constraints, *The Thirty-Fifth Annual Meeting of the ACL and Eighth Conference of the EACL*, Somerset, New Jersey.
- [71] **Erguvanli, E. E.**, 1979. The Function of Word Order in Turkish, *PhD Thesis*, University of California, Los Angeles.
- [72] **Tekin, T., Ölmez, M., Ceylan, E., Ölmez, Z. K. and Eker, S.**, 1995. Türkmençe-Türkçe Sözlük, Simurg Yayınları, İstanbul.
- [73] **Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufis, D., Koeva, S., Totkov, G., Dutoit, D. and Grigoriadou, M.**, 2002. Balkanet: A multilingual Semantic Network for Balkan Languages, *First International WordNet Conference*, Mysore India.

EK A – BİÇİMBİLİMSEL ETİKETLERİN AÇIKLAMALARI

Etiket	Açıklama
+Noun	İsim
+Adj	Sıfat
+Adverb	Belirteç
+Det	Belirteç
+Dup	İkileme
+Interj	Ünlem
+Ques	Soru
+Verb	Eylem
+Postp	İlgeç
+Num	Sayı
+Pron	Zamir
+Punc	İşaretleme
+Card	Asıl Sayı
+Ord	Sırasal Sayı
+Percent	Yüzdesel Sayı
+Range	Aralıksal Sayı
+Real	Gerçel Sayı
+Ratio	Oransal Sayı
+Distrib	Üleştirme Sayı
+Time	Zaman Belirteci
+Inf	Master
+Prop	Özel İsim
+PastPart	Geçmiş Zaman Ortacı
+FutPart	Gelecek Zaman Ortacı
+PresPart	Geniş Zaman Ortacı
+Demons	İşaret Zamiri
+Ques	Soru Zamiri

+Reflex	Dönüşlü Zamir
+Pers	Kişi Zamiri
+Quan	Belgisiz Zamir
+A1sg	1. Tekil Şahıs Uyum Özelliği
+A2sg	2. Tekil Şahıs Uyum Özelliği
+A3sg	3. Tekil Şahıs Uyum Özelliği
+A1pl	1. Çoğul Şahıs Uyum Özelliği
+A2pl	2. Çoğul Şahıs Uyum Özelliği
+A3pl	3. Çoğul Şahıs Uyum Özelliği
+P1sg	1. Tekil Şahıs İyelik Eki
+P2sg	2. Tekil Şahıs İyelik Eki
+P3sg	3. Tekil Şahıs İyelik Eki
+P1pl	1. Çoğul Şahıs İyelik Eki
+P2pl	2. Çoğul Şahıs İyelik Eki
+P3pl	3. Çoğul Şahıs İyelik Eki
+Pnon	Belirsiz İyelik
+Nom	Yalın Durum
+Acc	Belirtme Durumu
+Dat	Yönelme Durumu
+Abl	Çıkma Durumu
+Loc	Kalma Durumu
+Gen	Tamlayan Durumu
+Ins	Aracılık Durumu
+Equ	Eşitlik Durumu
^DB+Verb+Pass	Edilgen Çatı
^DB+Verb+Caus	Ettirgen Çatı
^DB+Verb+Reflex	Yansımalı Çatı
^DB+Verb+Recip	Dönüşlü Çatı
^DB+Verb+Able	Yetkinlik Çatı
^DB+Verb+Repeat	Sürerlik Yardımcı Eylemi (-dur eki)
^DB+Verb+Hastily	Tezlik Yardımcı Eylemi (-ver eki)
^DB+Verb+EverSince	Sürerlik Yardımcı Eylemi (-gel eki)
^DB+Verb+Almost	(-yaz eki)
^DB+Verb+Stay	(-kal eki)
^DB+Verb+Start	Sürerlik Yardımcı Eylemi (-koy eki)

+Pos	Olumlu
+Neg	Olumsuz
+Past	Belirli Geçmiş Zaman
+Narr	Belirsiz Geçmiş Zaman
+Fut	Gelecek Zaman
+Aor	Geniş Zaman
+Pres	Şimdiki Zaman
+Desr	İstek Kipi
+Cond	Şart Kipi
+Neces	Gereklilik Kipi
+Opt	Dilek Kipi
+Imp	Emir Kipi
+Prog1	Şimdiki Zaman (-yor)
+Prog2	Şimdiki Zaman (-makta)
+Cop	Kesinlik / kuşku anlamı (-dır eki)
+AfterDoingSo	Bağlama Ulacı (-ıp eki)
+SinceDoingSo	Zaman Ulacı (-alı eki)
+AsLongAs	Zaman Ulacı (-dıkça eki)
+When	Zaman Ulacı (-ınca eki)
+ByDoingSo	Durum Ulacı (-arak eki)
+While	Durum Ulacı (-ken eki)
+AsIf	Kıyaslama Ulacı (-arcasına eki)
+WithoutHavingDoneSo	Durum Ulacı (-maksızın eki)
+With	İsimden Sıfat Yapan Yapım Eki (-lı eki)
+Without	İsimden Sıfat Yapan Yapım Eki (-sız eki)
+FitFor	İsimden Sıfat Yapan Yapım Eki (-lık eki)
+Agt	Yapım Eki (-çı eki)
+Dim	Küçültme Sıfatı Eki (-cik eki)
+Ness	Sıfattan İsim Yapan Yapım Eki (-lık eki)
+Become	İsimden Eylem Yapan Yapım Eki (-laş eki)
+Acquire	İsimden Eylem Yapan Yapım Eki (-lan eki)
+Zero	Ek gelmeden türetim olduğunu gösterir
+PCAbI	İlgeç, çıkma durumunda bir isimden sonra gelir
+PCAcc	İlgeç, belirtme durumunda bir isimden sonra gelir
+PCDat	İlgeç, yönelme durumunda bir isimden sonra gelir

+PCGen	İlgeç, tamlayan durumunda bir isimden sonra gelir
+PCIns	İlgeç, aracılık durumunda bir isimden sonra gelir
+PCNom	İlgeç, yalın durumda bir isimden sonra gelir

EK B – İNGİLİZCE TERİMLERİN TÜRKÇE KARŞILIKLARI

İngilizce	Türkçe
Ablative Case	Çıkma Durumu
Accusative Case	Belirtme Durumu
Agglutinative	Bitişken
Alignment	Hizalama
Ambiguity	Belirsizlik
Back-Off	Derece Düşürme
Beam-Search	Demetli Arama
Best First Search	İlk En İyiyle Arama
Bound Morpheme	Ek biçimbilim
Chunk	Grup
Chunking	Gruplama
Computer Aided Machine Translation	Bilgisayar Destekli Dil Çevirisi
Corpus	Derlem
Dative Case	Yönelme Durumu
Decoder	Çözücü
Depth-First Search	Derinliğine Arama
Derivation	Türetme
Derivational	Türetimsel
Direct Translation	Doğrudan Aktarım
Disambiguation	Belirsizlik Giderimi
Evaluation	Başarım Değerlendirmesi
Example-Based Machine Translation	Örnek Tabanlı Bilgisayarlı Çeviri
Finite State Machine	Sonlu Durumlu Makine
Finite State Recognizer	Sonlu Durumlu Tanıyıcı
Finite State Transducer	Sonlu Durumlu Dönüştürücü
Formel	Biçimsel
Free Constituent Order Language	Öğelerin Yer Değiştirebildiği Dil
Free Morpheme	Serbest biçimbilim
Genitive Case	Tamlayan Durumu
Inflection	Çekim
Inflectional	Çekimli
Inflectional Group (IG)	Çekim Grubu (ÇG)

Information Extraction	Bilgi Çıkarımı
Information Retrieval	Bilgi Getirimi
Instrumental Case	Aracılık Durumu
Interlingua	Dilden Bağımsız Anlamsal Yapı
Information Acquisition	Bilgi Toplama
Lexical Ambiguity	Sözcüksel Belirsizlik
Lexical Representation	Yapısal Biçim
Lexicon	Alt Sözlük
Locative Case	Kalma Durumu
Machine Translation	Bilgisayarlı Metin Çevirisi
Maximum Likelihood Estimation	En Büyük Olabilirlik Kestirimi
Minimum Edit Distance	En Kısa Değişim Uzaklığı
Morfotactics	Morfotaktik
Morpheme	Biçimbilim
Morphological Analyzer	Biçimbilimsel Çözümleyici
Morphological Generator	Biçimbilimsel Üretici
Morphology	Biçimbilim
Multi-Word Unit	Çoklu Sözcük Grubu
Mutual Intelligibility	Karşılıklı Anlaşılabilirlik
N-Best Lists	En İyi N Aday
Natural Language Processing	Doğal Dil İşleme
Noisy Channel Model	Gürültülü Kanal Modeli
Nominal	İsim Soylu
Noun Phrase	Ad Öbeği
Observation Probability	Gözlem Olasılıkları
Ortography	Yazım, Yazım Dili
Part-Of-Speech (POS)	Sözcük Türü
Phonetic	Sesbilgisel
Phrase	Sözcük Öbeği
Pos Tagger	Sözcük Türü Etiketleyici
Pragmatics	Bağlambilim
Precision	Kesinlik
Recall	Gerigetirim
Regular Expression	Düzenli İfade
Regular Language	Düzenli Dil
Semantics	Anlambilim
Smoothing	Yumuşatma
Statistical Language Model	İstatistiksel Dil Modeli
Sublanguage	Alt Dil
Surface Representation	Yüzeysel biçim
Syntax	Sözdizimi

Translation Model	Çeviri Modeli
Two-Level Morphology	İki-Düzeyli Biçimbilim
Verb Phrase	Eylem Öbeği
Verbal	Eylem Soylu

ÖZGEÇMİŞ

Ahmet Cüneyd TANTUĞ, 1978 yılında Adana’da doğdu. 2000 yılında İstanbul Teknik Üniversitesi Otomatik Kontrol ve Bilgisayar Mühendisliği Bölümü’nden dereceyle mezun oldu. 2002 yılında İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Yüksek Lisans Programını bitirerek Yüksek Mühendis ünvanını aldıktan sonra gene aynı bölümde doktora çalışmalarına başladı.