ESKİ TÜRKÇE METİNLERİN GÜNÜMÜZ TÜRKÇESİNE SADELEŞTİRİLMESİ

SIMPLIFICATION OF OLD TURKISH TEXTS INTO MODERN TURKISH

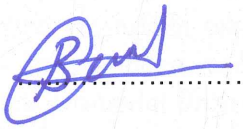
EROL ÖZKAN

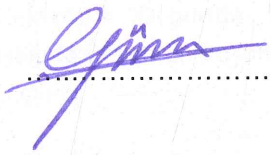
DR. ÖĞR. ÜYESİ GÖNENÇ ERCAN
Tez Danışmanı

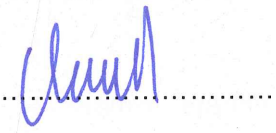
Hacettepe Üniversitesi
Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin
Bilgisayar Mühendisliği Anabilim Dalı için Öngördüğü
YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2018

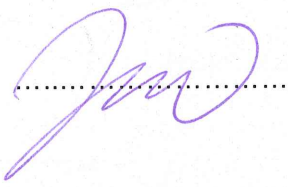
EROL ÖZKAN' ın hazırladığı **"Eski Türkçe Metinlerin Günümüz Türkçesine Sadeleştirilmesi"** adlı bu çalışma aşağıdaki jüri tarafından **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI'** nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Dr. Öğr. Üyesi Burcu Can
Başkan ..... 

Dr. Öğr. Üyesi Gönenç Ercan
Danışman ..... 

Dr. Öğr. Üyesi Umut Özge
Üye ..... 

Dr. Öğr. Üyesi Ufuk Çelikcan
Üye ..... 

Dr. Öğr. Üyesi Adnan Özsoy
Üye ..... 

Bu tez Hacettepe Üniversitesi Fen Bilimleri Enstitüsü tarafından **YÜKSEK LİSANS TEZİ** olarak onaylanmıştır.

Prof. Dr. Menemşe GÜMÜŞDERELİOĞLU
Fen Bilimleri Enstitüsü Müdürü

YAYINLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin / raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma ama iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan **" Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"** kapsamında tezim aşağıda belirtilen koşullar haricinde YÖK Ulusal Tez Merkezi / H. Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- o Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir. (1)
- o Enstitü / Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren .... Ay ertelenmiştir. (2)
- o Tezimle ilgili gizlilik kararı verilmiştir. (3)

06/08/2018

Erol Özkan

"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"

(1) Madde 6. 1. Lisansüstü tezle ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir

(2) Madde 6. 2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. Şahıslara veya kurumlara haksız kazanç imkanı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü ve fakülte yönetim kurulunun gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.

(3) Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, tezin yapıldığı kurum tarafından verilir*. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, ilgili kurum ve kuruluşun önerisi ile enstitü veya fakültenin uygun görüşü üzerine üniversite yönetim kurulu tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.
Madde 7. 2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir.

* Tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu tarafından karar verilir.

ETİK

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

01/08/2018

EROL ÖZKAN

ÖZET

Eski Türkçe Metinlerin Günümüz Türkçesine Sadeleştirilmesi

Erol ÖZKAN

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Danışman: Dr. Öğr. Üyesi Gönenç Ercan

Ağustos 2018, 70 sayfa

Dilde zaman içerisinde meydana gelen değişim eski zamanlarda yazılmış metinlerin günümüzde kullanılmayan birçok kelime içermesine yol açmaktadır. Bu durum okurların eski metinleri anlamasını zorlaştırmaktadır. Metin sadeleştirme görevinin amacı, metin içeriğini ve anlamını koruyarak metnin okunabilirliğini ve anlaşılabilirliğini arttırmaktır.

Bu tezde, Cumhuriyet dönemi Türkçesi ile yazılmış metinlerin karmaşıklığının metin sadeleştirme yöntemleriyle azaltılması hedeflenmiştir. Metin sadeleştirme görevi eski Türkçeden güncel Türkçeye dil içi çeviri problemi olarak görülmüştür. Sadeleştirme işlemi için; kural tabanlı ve istatistiksel sadeleştirme modelleri oluşturulmuştur. Kural tabanlı model için otomatik olarak oluşturulmuş sözlük kullanılmış, istatistiksel model için ise Nutuk kullanılarak oluşturulan paralel veri kümesi kullanılmıştır. İki model karşılaştırıldığı gibi tek bir hibrid sistemde birleştirilmiştir.

Sonuçlar makine çevirisi sistemlerinin başarı ölçümünde kullanılan BLEU metriği kullanılarak değerlendirilmiştir. Çalışma ile eski metinlerin karmaşıklığı azaltılarak, bu metinlerin hedef kitlesi arttırılmaktadır.

Anahtar Kelimeler: Metin sadeleştirme, makine öğrenmesi, istatistiksel makine çevirisi, açımlama.

ABSTRACT

Sımplification of Old Turkish Texts into Modern Turkish

Erol ÖZKAN

Master of Science, Computer Engineering Department

Supervisor: Asst. Prof. Dr. Gönenç ERCAN

August 2018, 70 pages

Changes in a language over time causes the text written in old times contain a lot of words that are not used at the present time. This makes it difficult for readers to understand old texts. The goal of text simplification task is to increase the readability and understandability of the text while preserving its content and meaning.

In this thesis, it is aimed to reduce the complexity of the texts written in republican period Turkish with text simplification methods. Text simplification task is considered as an Intralingual translation problem. For simplification process; rule-based and statistical simplification models are developed. For rule-based model, an automatically built bilingual dictionary is used. For statistical model, a parallel corpus that is built from Nutuk is used. These systems are compared, as well as they are combined in a single hybrid system.

The results are measured using BLEU metric that is used in the evaluation of machine translation systems. With this work, the complexity of old texts is reduced and the target audience of these texts is increased.

Keywords: Text simplification, machine learning, statistical machine translation, paraphrasing.

TEŞEKKÜR

Tez çalışmamın her aşamasında ilgi ve desteğini esirgemeyen, değerli katkılarıyla yol gösteren, her zaman çalışmaya teşvik eden, yönlendirme ve bilgilendirmeleriyle çalışmamı bilimsel temeller ışığında şekillendiren sayın hocam Dr. Öğr. Üyesi Gönenç Ercan'a sonsuz teşekkürlerimi sunarım.

Ayrıca, tüm eğitim ve kariyer hayatım boyunca benden maddi ve manevi desteklerini esirgemeyen, desteklerini hep arkamda hissettiğim, her zaman yanımda olan sevgili aileme teşekkürlerimi bir borç bilirim.

İÇİNDEKİLER

Sayfa

ÇİZELGELER

Sayfa

ŞEKİLLER

Sayfa

KISALTMALAR

TDK	Türk Dil Kurumu
BLEU	BiLingual Evaluation Understudy
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
GNMT	Google’s Neural Machine Translation
MLE	Maximum Likelihood Estimation
IBM	International Business Machines
EM	Expectation Maximization
BP	Brevity Penalty
Ort.	Ortalama
Var.	Varyans
S.S.	Standart Sapma

1. GİRİŞ

Dile zaman içerisinde yeni kelimeler girerken bazı kelimelerin kullanımı azalmakta bazı kelimeler ise tarihin tozlu sayfaları arasında tamamen kaybolmaktadır [1]. Eski zamanlarda yazılmış metinler günümüzde kullanılmayan birçok kelime ve kelime grubu içermektedir. Bu kelimelerin anlamı günümüzde birçok okur tarafından bilinmediği için metinler kolaylıkla anlaşılamamaktadır. Bu sebeple eski edebi eserler dil uzmanları ya da yetkin kişiler tarafından incelenmekte, eserler ya tamamen sadeleştirilerek yeni bir metin olarak sunulmakta ya da metindeki kelimelere atıf yapılarak kelimelerin anlamları sayfa sonlarına eklenmektedir [2], [3]. Fakat bu ve benzeri yöntemleri uygulamak hala büyük ölçüde manuel bir süreç olup insan gücü gerektirmektedir. Ayrıca, bu yöntemlerde işlemi gerçekleştirecek olan kişinin dil alanında önemli bir bilgi birikiminin de bulunması gerekmektedir.

1.1. Türk Dilindeki Hızlı Değişim Süreci

Yüzyıllar boyu büyük toprak parçalarını egemenliği altında bulunduran Osmanlı İmparatorluğu farklı kültürlerin etkisi altında kalmıştır. Dil, Arapça ve Farsça gibi birçok dilden etkilenmiştir [4]. Dilde kullanılan kelime dağarcığındaki Arapça ve Farsça kelimelerin sayısı oldukça fazladır. Bu durum yazılı ve sözlü dil kullanımında zorluklara sebep olmuştur.

Bu amaçla, Cumhuriyet döneminde, Mustafa Kemal Atatürk önderliğinde dilde sadeleştirme çalışmaları başlatılmıştır [4]. 1928 yılında Arapça alfabeden Latin alfabesine geçilmiş ve dili yabancı kelimelerden arındırmak amacıyla Türk Dil Kurumu (TDK) kurulmuştur.

Yapılan reformlarla, yeni neslin artık okuma-yazma öğrenmek için Arapça harfleri öğrenme ve Arapça ve Farsça kelimeleri bilme zorunluluğu ortadan kalkmıştır. Dil daha güncel, pratik, kesin ve anlaşılması daha kolay hale gelmiştir. Okur-yazarlık oranı yapılan sadeleştirme çalışmaları sonrasında hızlı bir artış göstermiştir [5].

Fakat kısa bir zaman dilimi içerisinde dilde önemli bir değişim meydana gelmiştir. Günümüz okurları Cumhuriyet dönemi ve öncesinde yazılan eserleri kolaylıkla anlayamamaktadır. Bazı eserlerin günümüz Türkçesine sadeleştirilmesi yapılsa da, birçok eser günümüz okurlarının anlayabileceği durumda değildir.

1.2. Amaç, Kapsam ve Yöntem

Çalışmada Cumhuriyet dönemi Türkçesi ile yazılmış ve günümüzde kullanılmayan birçok kelime içeren metinlerin sadeleştirilerek metinlerde geçen eski kelime ve kelime gruplarının güncel Türkçedeki karşılıkları ile değiştirilmesi hedeflenmiştir. Yazılma tarihi eski olan metinlerin günümüz okurları tarafından daha kolay anlaşılabilmesi sağlanarak bu metinlerin hedef kitlesi arttırılmaktadır. Aynı zamanda, dil işleme araçlarında eski Türkçe kelimelerin sözlükte bulunmaması probleminden doğan başarı düşme sorunu da azaltılmaktadır.

Literatürde sadeleştirmenin farklı kapsamlarda yapıldığı birçok çalışma bulunmaktadır. Bu çalışmalar, metinlerin sadeleştirilmesi için uzun ve anlaşılması güç olan cümlelerin parçalanması, gereksiz parçaların silinmesi, parçaların sırasının değiştirilmesi ve açımlama gibi farklı sadeleştirme işlemlerinin bir ya da birkaçını içermektedir [6].

Sadeleştirilen eserlerde eski ve anlaşılması zor olan kelimelerin güncel karşılıkları ile değiştirilmesi sağlanırken, çoğu durumda cümlelerin yapısında değişiklik gerçekleştirilmemektedir. Bu durum sadeleştirmelerde yazar tarafından oluşturulan cümle yapısının korunmaya çalışılmasından kaynaklanmaktadır. Tez çalışmasında eski metinlerin günümüz Türkçesine sadeleştirilmesi dil içi çeviri problemi olarak görülmüş [7], cümlenin yapısından doğan karmaşıklığın azaltılması ise kapsam dışı bırakılmıştır.

Eski metinlerin güncel Türkçeye çevrilmesini hedef alan ilk veri kümesi Nutuk [8] kullanılarak oluşturulmuştur. Türkçeye çevrilmiş Nutuk [9] güncel Türkçede bulunmayan birçok ifade içermektedir. Sadeleştirilmiş Nutuk'ta [10] bu ifadelerin günümüzde kullanılan karşılıklarıyla değiştirilmesi sağlanmıştır. Türkçe Nutuk ve sadeleştirilmiş Nutuk'tan örnek bir cümle çifti Çizelge 1.1 üzerinde verilmiştir.

Çizelge 1.1. Nutuk ve sadeleştirilmiş Nutuk'tan örnek bir cümle

Nutuk	"**Vaziyet ve manzarai umumiye:** Osmanlı Devletinin **dahil** bulunduğu grup, **Harbi Umumîde mağlûp olmuş**, Osmanlı ordusu her tarafta zedelenmiş, **şeraiti** ağır, bir **mütarekename** imzalanmış."
Sadeleştirilmiş Nutuk	"**Genel durum ve görünüm:** Osmanlı Devletinin **içinde** bulunduğu grup, **Genel Savaşta yenilmiş**, Osmanlı ordusu her tarafta zedelenmiş, **şartları** ağır bir **ateşkes anlaşması** imzalanmış."

Kelime tabanlı olarak sadeleştirme işlemini gerçekleştirmek üzere kural tabanlı sadeleştirme modeli oluşturulmuştur. Modelde; sadeleştirmeler oluşturulan Osmanlıca-Türkçe sözlük üzerinden gerçekleştirilmektedir. İlk olarak eski kelimeler tespit edilmekte, daha sonra tespit edilen kelimelerin günümüz Türkçesindeki karşılıkları ile değiştirilmesi sağlanmaktadır.

Kural tabanlı modelde gerçekleştirilen sadeleştirmeler kelime bazında yeterli olsa da, tümce yapıları ve sözdizimsel olarak karmaşık olan cümleler için yeterli olmamaktadır. Bu amaçla; sadeleştirmelerin cümle tabanlı olarak hizalanmış paralel veri kümesi üzerinden öğrenildiği istatistiksel sadeleştirme modeli oluşturulmuştur. Modelde ilk olarak veri kümesi üzerindeki cümleler kelime tabanlı olarak eşleştirilmektedir. Daha sonra Moses aracı [11] kullanılarak istatistiksel makine çeviri modeli eğitilmiştir. Aynı zamanda, farklı kelime eşleştirme araçlarının, farklı dil modeli veri kümelerinin ve faktörlü makine çevirisi yönteminin sistem başarısı üzerindeki etkisi test edilmiştir.

İki model birbirleri ile karşılaştırıldığı gibi, tek bir hibrid sistemde birleştirilmiştir. Sonuçlar istatistiksel makine çevirisi sistemlerinin başarı ölçümünde kullanılan BLEU metriği [12] kullanılarak değerlendirilmiştir. Nutuk'un sadeleştirilmiş sürümü aynı zamanda referans metin olarak kabul edilmiş ve geliştirilen sistemin değerlendirilmesinde kullanılmıştır.

Tez çalışmasının bilimsel katkıları;

- Eski Türkçe metinlerin güncel Türkçeye çevrilmesini hedef alan ilk veri kümesinin oluşturulması,
- Sadeleştirme işleminin sözlük üzerinden gerçekleştirildiği kural tabanlı sadeleştirme modelinin oluşturulması ve bu modelde Türkçenin biçimbilimsel yapısından kaynaklanan sorunların giderilmesi,
- Sadeleştirme işlemi için istatistiksel makine çevirisi yönteminin kullanılması, değerlendirilmesi ve kural tabanlı sadeleştirme modeli ile karşılaştırılması,
- İki yöntemin hibrid bir sistemde birleştirilmesi.

1.3. Tezin Yapısı

İkinci bölümde metin sadeleştirme problemi ve bu problem için geliştirilmiş yöntemler açıklanmıştır. Ayrıca, günümüze kadar geliştirilen önemli makine çevirisi yöntemleri incelenmiş ve karşılaştırılmıştır.

Üçüncü bölümde, Nutuk ile ilgili genel bilgiler sunulmuş, metin sadeleştirme veri kümeleri ve bu veri kümelerini oluşturma yöntemleri hakkında bilgiler verilmiş, kullanılan Nutuk veri kümesini oluşturma işlemleri açıklanmış ve oluşturulan veri kümesi üzerinde istatistiksel analizler gerçekleştirilmiştir.

Dördüncü bölümde geliştirilen kural tabanlı sadeleştirme modeli ve bu modelde kullanılan işlem serisi ayrıntılı olarak incelenmiştir.

Beşinci bölümde uygulanan istatistiksel sadeleştirme modeli ve bu modelde kullanılan işlem serisi ayrıntılı olarak açıklanmıştır.

Altıncı bölümde yapılan deney sonuçlarına yer verilmiş, yedinci bölümde bu sonuçlar değerlendirilmiştir.

Sekizinci bölümde ise sonuç bölümüne ve gelecekte çalışmaya nasıl devam edileceğine yer verilmiştir.

2. LİTERATÜR ÖZETİ

Bu bölümde, ilk olarak, metin sadeleştirme problemi ve bu problem için geliştirilmiş yöntemler açıklanmıştır. Daha sonra, günümüze kadar geliştirilen önemli makine çevirisi yöntemleri incelenmiş ve karşılaştırılmıştır.

2.1. Metin Sadeleştirme

Metin sadeleştirme görevinin amacı, metin içeriğini ve anlamını koruyarak metnin okunabilirliğini ve anlaşılabilirliğini arttırmaktır. Metin sadeleştirme, karmaşık belgelere, teknik belgelere, eski belgelere; bu metinlerin karmaşıklığını azaltmak ve bu belgeleri daha geniş bir insan kitlesi tarafından erişilebilir kılmak amacıyla uygulanmaktadır. Bu insan kitleleri çocukları, yaşlıları, okuryazarlık seviyesi düşük olan kişileri, ana dili farklı olan kişileri, disleksi, afazi veya sağırlık gibi rahatsızlıklardan muzdarip olan kişileri kapsamaktadır [13].

Metin sadeleştirme görevinde insan grupları hedeflenebildiği gibi bilgisayar yazılımı ve algoritmalar da hedeflenebilmektedir. Cümlenin karmaşık yapısının ve kelime dağarcığının sorun teşkil ettiği metin özetleme ve makine çevirisi gibi dil işleme uygulamalarında sistem başarısını iyileştirmek amacıyla kullanılabilmektedir [14], [15].

Metin sadeleştirme problemi üzerine son yıllarda birçok önemli çalışma yapılmıştır. Bu çalışmalarda; sadeleştirme problemi için sözcüksel, sözdizimsel, istatistiksel makine çevirisi ve hibrid olmak üzere farklı yöntemler kullanılmıştır [6]. Bu yöntemler cümlelerin parçalanması, gereksiz parçaların silinmesi, parçaların sıralanması ve açımlama gibi işlemlerin bir ya da birkaçından oluşmaktadır.

2.1.1. Sözcüksel Sadeleştirme

Sözcüksel sadeleştirmenin amacı, karmaşık kelimelerin daha basit karşılıkları ile değiştirilmesidir. Bu yöntemde metin üzerinde sözdizimsel ya da biçimbilimsel olarak sadeleştirme işlemi gerçekleştirilmemektedir.

Sözcüksel sadeleştirme sistemlerinde, ilk olarak karmaşık kelimeler belirlenmektedir. Daha sonra, bu kelimelerin yerine kullanılabilecek aday kelime listesi oluşturulmaktadır. En son olarak ise, aday kelimelerden en yüksek skora sahip olan kelimenin karmaşık kelime ile değiştirilmesi gerçekleştirilmektedir. Aday

kelimeleri oluşturmak ve değerlendirmek için word2vec [16] modelleri, kelime frekansları ya da WordNet [17] gibi farklı kaynaklar kullanılabilmektedir [18]–[20].

Glavas ve Stajner [18] sözcüksel olarak anlaşılması zor olan kelimelerin metnin anlaşılırlığını düşürdüğünü belirtmiş, metni daha basit ve anlaşılabilir hale getirmek üzere bu kelimeleri değiştirmek için eğittikleri word2vec modelini kullanmışlardır. Çalışmaları yalnızca tekil kelimelerin değiştirilmesi üzerine odaklanmıştır.

Ligozat ve arkadaşları [19] sözcüksel sadeleştirme görevi için dil modeli istatistiklerini kullanmışlardır. Farklı kaynaklar ile oluşturdukları dil modellerinin sistem başarısı üzerine etkisini karşılaştırmış ve sadeleştirilmiş Wikipedia [21] üzerinden eğittikleri dil modeliyle en yüksek başarı skorunu elde etmişlerdir.

Thomas ve Anderson [20] metinleri daha küçük ve daha basit bir kelime dağarcığı ile ifade edebilmek amacıyla WordNet kullanmışlardır. Bu amaçla metin üzerindeki karmaşık isim ve fiillerin değiştirilmesine odaklanmışlardır. Sistemlerinde, ayrıca, kelime türü işaretleme ve anlam belirsizliği giderme işlemleri uygulanmaktadır.

Sözcüksel sadeleştirmede dilin daha basit kelimelerle ifade edilmesi sağlanmaktadır. Tez çalışmasında; eski kelimelerin, güncel karşılıklarıyla sadeleştirilmesi hedeflenmektedir. Dil modeli istatistiklerinin sadeleştirme işleminde kullanılması sağlanmıştır. WordNet ya da word2vec kullanarak başarılı sonuçlar alabilmek için ise; Türkçe WordNet'in Osmanlıca kelimeleri içermesi ve word2vec modelini oluşturacak yeterli miktarda eski ifade içeren derlemin oluşturulması gerekmektedir.

2.1.2. Sözdizimsel Sadeleştirme

Karmaşık cümleler çocuklar gibi daha az yetenekli okuyucular tarafından daha zor anlaşılmaktadır. Sözdizimsel sadeleştirmenin amacı, cümlenin içerik ve anlamını koruyarak, sözdizimsel ve biçimbilimsel karmaşıklığını azaltmaktır.

Sözdizimsel sadeleştirme sistemlerinde, okumayı ve kavramalarını kolaylaştırmak üzere cümleler üzerine bir dizi sözdizimsel ve biçimbilimsel sadeleştirme işlemi gerçekleştirilmektedir. Bu sadeleştirmeler; cümlelerin parçalanmasını, gereksiz parçaların silinmesini ve bu parçaların sıralanmasını içerebilmektedir. Parçalama işleminde uzun ve anlaşılması zor olan cümleler daha kısa cümlelere çevrilmekte,

silme işleminde cümlelerin önemsiz kısımları silinmekte, sıralama işleminde ise cümledeki parçaların sırası değiştirilmektedir.

Sözdizimsel sadeleştirme çalışmalarında, sadeleştirme kurallarının manuel olarak oluşturulduğu ya da paralel bir veri kümesi üzerinden otomatik olarak öğrenildiği çalışmalar bulunmaktadır.

Siddharthan [22] sözdizimsel sadeleştirme işlemini; analiz, çevrim ve yeniden oluşturma olmak üzere 3 adımda gerçekleştirmektedir. Analiz adımında metin sistemde kullanılmak üzere hazırlanmakta, çevrim adımında manuel olarak oluşturulmuş sözdizimsel sadeleştirme kuralları uygulanmakta ve en son olarak, yeniden oluşturma adımında ise metin yeniden oluşturulmaktadır. Aynı zamanda Siddharthan'ın sisteminde; kelime türü işaretleme, isim tamlaması belirleme, artgönderimsel ifade ayrımı, cümle ve sözlük sıralama işlemleri uygulanmaktadır.

Torunoglu-Selamet ve arkadaşları [23] Türkçe için sözdizimsel sadeleştirme çalışması yapmış, çalışmalarında cümleleri içerdikleri sözdizimsel ve biçimbilimsel yapılara göre farklı kategorilere ayırmış ve bu kategorilere giren cümleleri sadeleştirmek için sözdizimsel ve biçimbilimsel kurallar oluşturmuşlardır.

Eski bir eserin günümüz Türkçesine aktarılmasında cümle yapısına mümkün olduğunca sadık kalınmaya çalışılmaktadır. Yazarlar tarafından oluşturulmuş anlatım özelliklerinin değiştirilmemesi için metinlerde sözdizimsel değişimler gerçekleştirilmemektedir. Tez çalışmasında, cümlenin yapısından doğan karmaşıklığın azaltılması kapsam dışı bırakılmıştır.

2.1.3. Makine Çevirisi Yöntemleri İle Sadeleştirme

Son yıllarda metin sadeleştirme görevi için makine çevirisi yöntemlerinin kullanıldığı çalışmalar yapılmıştır. Bu çalışmalarda sadeleştirmeler orijinal ve sadeleştirilmiş metinler üzerinden makine öğrenmesi yöntemleri kullanılarak öğrenilmektedir.

Zhu ve arkadaşları [24] çalışmalarında PWKP metin sadeleştirme veri kümesini oluşturmuş, bu veri kümesi ile ağaç tabanlı istatistiksel makine çevirisi modeli eğitmişlerdir. Çalışmaları cümle parçalama, silme, sıralama ve öbek değiştirme sadeleştirme işlemlerini kapsamaktadır.

Xu ve arkadaşları [13] metin sadeleştirme görevi için sözdizimsel istatistiksel makine çevirisi modeli eğitmişlerdir. Veri kümesi olarak PPDB [25] veri tabanında bulunan açımlamaları kullanmış, bu açımlamalardan basit olanları belirlemek amacıyla SARI ve FKBLEU olmak üzere iki farklı metrik geliştirmişlerdir.

Yine Xu ve arkadaşları [26] Shakespeare tarafından yazılmış metinlerin yazım stilini güncel İngilizceye çevirmeye çalışmışlardır. Bunun için iki taraflı olarak stil transferi yapabilen bir açımlama modelini istatistiksel makine çevirisi yöntemi kullanarak oluşturmuşlardır. Geliştirdikleri model tarafından gerçekleştirilen bir stil transferi örneği Çizelge 2.1 üzerinde verilmiştir.

Çizelge 2.1. Shakespeare yazım stilinden güncel İngilizceye stil transferi örneği

Orjinal	i ' ll bite you by the ear for that joke .
Xu ve ark. [26]	i will bite thee by the ear for that jest .

Örnekte Shakespeare yazım stilindeki metin güncel İngilizceye çevrilmektedir. Transfer işleminde bazı kelimelerin değiştirildiği, bazılarının ise değiştirilmediği görülmektedir.

Jhamtani ve arkadaşları [27] güncel İngilizceyi Shakespeare'in yazım stiline çevirebilmek amacıyla kodlayıcı ve çözümleyiciden oluşan bir nöral çeviri sistemi kullanmışlardır. Ayrıca, sistem başarılarını arttırmak amacıyla, harici bir sözlüğü ve farklı kaynaklar üzerinden eğittikleri word2vec modellerini sistemlerinde kullanmışlardır.

Benzer bir çalışmada, Wang ve arkadaşları [28] metin sadeleştirme görevini normal İngilizceden ve sadeleştirilmiş İngilizceye bir makine çevirisi problemi olarak görmüşlerdir. Bu problemi çözmek amacıyla RNN kodlayıcı ve çözümleyiciden [29] oluşan nöral bir çeviri modeli önermişlerdir. Ayrıca çalışmalarında, makine çevirisi ve metin sadeleştirme problemi arasındaki farkları incelemiş ve metin sadeleştirme problemi için çeviride ortak kelimelerin kullanılabileceğine değinmişlerdir.

2.2. Makine Çevirisi

Makine çevirisi, bir dilden başka bir dile çeviri işlemidir [30]. Bilgisayar kullanılarak yapılan makine çevirisi çalışmaları yirminci yüzyılda başlamış ve günümüze kadar birçok farklı yöntem geliştirilmiştir [31].

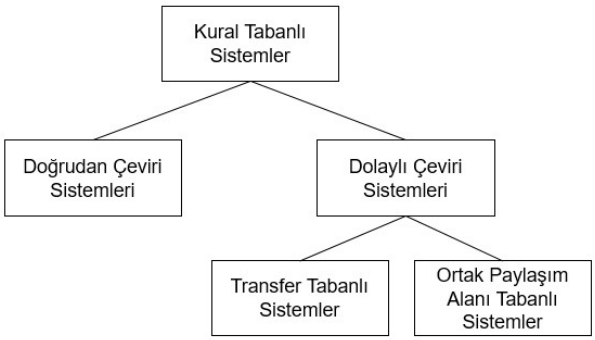
Makine çevirisi çalışmalarında ilk olarak kural tabanlı sistemler kullanılmıştır. Daha sonra bu sistemler yerlerini çevirilerin paralel veri kümeleri üzerinden otomatik olarak öğrenildiği istatistiksel yöntemlere bırakmıştır. Ayrıca, günümüze kadar bağlam tabanlı, örnek tabanlı, nöral ve hibrid olmak üzere farklı yöntemler de geliştirilmiştir.

Makine çevirisinin zengin literatüründeki çalışmalar, çevirinin nasıl modellendiğine bağlı olarak; kural tabanlı sistemler, istatistiksel sistemler, nöral sistemler ve hibrid sistemler olmak üzere farklı başlıklar altında incelenmiştir.

2.3. Kural Tabanlı Sistemler

Kural tabanlı sistemlerde, kaynak ve hedef dil arasındaki dönüşüm dil bilgisine dayalı kurallar yardımı ile yapılmaktadır. Sözdizimsel, biçimbilimsel ve anlamsal özelliklere göre tanımlanan bu kurallar, dil uzmanlarının saatler süren çalışmaları sonucunda oluşturulmaktadır.

Kural tabanlı sistemler, Şekil 2.1'de gösterildiği gibi doğrudan çeviri sistemleri ve dolaylı çeviri sistemleri olmak üzere 2 grup altında incelenebilmektedir.



Şekil 2.1. Kural tabanlı sistemler

2.3.1. Doğrudan Çeviri Sistemleri

Doğrudan çeviri sistemlerinde, kaynak dil yüzeysel biçimdeki kelimeler üzerinden hedef dile çevrilmektedir. Bu sistemler ilk geliştirilen makine çevirisi sistemleri olmakla birlikte, bu sistemlerde çok az miktarda sözdizimsel ve biçimbilimsel analiz işlemleri uygulanmaktadır [30].

2.3.2. Dolaylı Çeviri Sistemleri

Doğrudan çeviri sistemlerinde çeviriler kelimelerin yalnızca yüzeysel biçimleri üzerinden gerçekleştirilmektedir. Bunun sonucu olarak sözdizimsel özellikler ve dil bilgisi kuralları çevirilerde kullanılamamaktadır. Dolaylı çeviri sistemlerinde bu sorunun çözülmesine odaklanılmıştır. Bu sistemlerinde çeviri dilin soyut bir şekilde ifade edildiği ara bir biçim üzerinden gerçekleştirilmektedir.

Transfer tabanlı sistemlerde çeviri üç adımda gerçekleştirilmektedir. İlk olarak, kaynak dil soyut bir metin biçimine dönüştürülmektedir. Daha sonra, bu soyut biçim üzerinden hedef dil biçimine dönüşüm sağlanmaktadır. En son olarak ise, hedef dil metni oluşturulmaktadır.

Ortak paylaşım alanı tabanlı sistemlerde, dilin çevrildiği soyut biçim üzerinden birçok dile çevrim sağlanabilmektedir [30]. Günümüze kadar tüm diller arasında çevrim sağlayabilecek soyut bir biçim bulunamamıştır [30]. KANT sisteminde [32] Fransızca, Almanca ve Japonca arasında çevrim sağlanabilmektedir.

Kural tabanlı sistemler, basit çevirilerin yeterli olduğu yakın dil çiftlerinin çevirisinde oldukça başarılı sonuçlar vermektedir. Bu dil çiftlerine Çekçe-Slovakça, Çekçe-Lehçe, İspanyolca-Katalanca gibi örnekler verilebilmektedir.

Yakın dil çiftleri arasında çevirinin yapıldığı ilk çalışma transfer tabanlı bir çeviri sistemi olan RUSLAN'dır [33]. RUSLAN sisteminde Çekçe belgelerin Rusçaya çevrilmesi sağlanmıştır. Bu amaçla sistemde; Çekçe biçimbilimsel ve sözdizimsel çözümleyici, aktarım, Rusça biçimbilimsel ve sözdizimsel üretici olmak üzere farklı bileşenler kullanılmıştır. Çalışma sonucunda yapılan değerlendirmelerde sistemin ürettiği çevirilerin yaklaşık %40'ının doğru olduğu, %40'ında küçük, %20'sinde ise önemli düzeltmeler gerektiği belirlenmiştir.

ČESĬLKO [34] sisteminde yakın iki dil olan Çekçe ve Slovakça arasında transfer tabanlı çeviri işlemi gerçekleştirilmektedir. Sistem kelime tabanlı olmakla birlikte, Çekçe biçimbilimsel çözümleme, biçimbilimsel belirsizliğin giderilmesi, aktarım ve Slovakça biçimbilimsel üretim işlemleri içermektedir.

Tantuğ ve arkadaşları [35] Türkmen dili ve Türkçenin benzer özelliklerinden yararlanarak, Türkmen dilinden Türkçeye doğrudan çeviri sistemi geliştirmişlerdir.

Sistemleri sözcüksel aktarım, biçimbilimsel aktarım ve istatistiksel anlam ayrımı olmak üzere farklı işlemlerden oluşmaktadır. Çalışmalarında sistemin başarısını ölçmek için BLEU metriğini kullanmışlardır. Farklı istatistiksel dil modelleri kullanarak deneyler yapmış ve en yüksek yaklaşık 33 BLEU skoru elde etmişlerdir.

Benzer şekilde Altıntaş & Çiçekli [36], Türkçe ve Kırım Tatar dilleri arasında makine çeviri sistemi geliştirmişlerdir. Sistemde, ilk olarak, Türkçe kelimeler üzerinde biçimbilimsel analiz gerçekleştirilmekte; daha sonra kök kelime, gramer ve bağlam bağımlı yapıların çevirisi yapılmakta; en son olarak ise, biçimbilimsel sentez yapılarak Kırım Tatar dilindeki metin oluşturulmaktadır.

2.4. İstatistiksel Sistemler

İstatistiksel sistemlerde çeviriler, kaynak ve hedef dildeki metinlerin cümle tabanlı olarak hizalandığı paralel veri kümeleri üzerinden öğrenilmektedir. Bu durum istatistiksel çeviri sistemlerinin, paralel veri kümesi bulunan dil çiftleri için kısa bir sürede uygulanabilir olmasını sağlamaktadır.

İstatistiksel sistemlerde, çevirinin kalitesi büyük oranda kullanılan paralel veri kümesinin kalitesine ve büyüklüğüne bağımlıdır. İnternet teknolojilerinin gelişmesi, büyük veri kümelerinin oluşturulmasına olanak sağlamıştır. Bu veri kümelerinin oluşturulmasında kitap çevirileri, Wikipedia [37] metinleri gibi farklı kaynaklar kullanılabilmektedir.

Europarl [38] veri kümesinin oluşturulmasında Avrupa Parlamentosu'nun web sitesinde yayınlanan metinler kullanılmıştır. Web üzerinden metin çıkarımı yapılmış, doküman tabanlı bir eşleştirme işlemi gerçekleştirilerek paralel metinler oluşturulmuştur. Daha sonra, bu metinler cümlelere bölünmüş ve cümle tabanlı başka bir eşleştirme işlemi daha gerçekleştirilmiştir.

Güncel istatistiksel çeviri sistemlerinde, kelime ve kelime gruplarının birlikte çevirisi gerçekleştirilmektedir [11]. Bu sistemlerde çeviri modeli genellikle kelime çeviri modeli, kelime grubu çeviri modeli, sıralama modeli gibi farklı bileşenlerden oluşmaktadır. Ayrıca, istatistiksel sistemlerde, çeviri farklı işlemler uygulanarak gerçekleşmektedir. İlk olarak, kaynak cümle daha küçük parçalara bölünmektedir. Daha sonra, her parçanın hedef dile çevrilmesi sağlanmaktadır. En son olarak ise,

çevrilen parçalar yeniden düzenlenmektedir. Ayrıca, doğru kelime seçimi ve dilde akıcılığı sağlamak için dil modeli kullanılmaktadır.

Faktörlü çeviri sistemlerinde; kelimelerin sözdizimsel ve biçimbilimsel özelliklerin gerçekleştirilen çevirilerde kullanılabilmesi sağlanabilmektedir. Bu yöntemlerde, her kelime, yüzeysel biçimleri yerine kelime kökü, kelime türü işareti ve biçimbilimsel ekleri gibi birçok faktörden oluşan bir vektör ile ifade edilmektedir [39]. Bu faktörler, çeviride sözdizimsel ve biçimbilimsel özelliklerin kullanılmasına izin vermektedir.

Sözdizimsel çeviri sistemlerinde; kaynak dilden hedef dile çevirim sözdizimsel kurallar üzerinden gerçekleştirilmektedir. Bu sayede yeniden sıralama gibi cümlenin yapısı ile ilgili çeviri sorunlarının çözülmesi sağlanmaktadır [40].

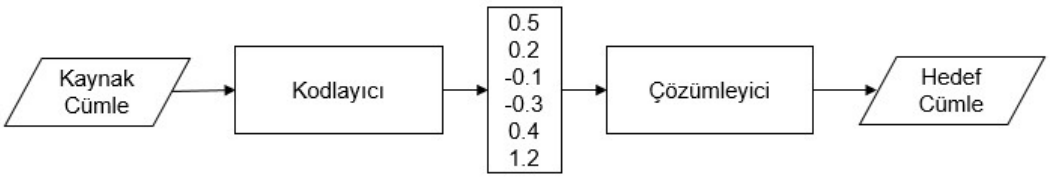
Oflazer [41] İngilizce-Türkçe dil çifti için istatistiksel bir çeviri modeli geliştirmiştir. Türkçedeki biçimbilimsel olarak zengin yapısının çeviride problem yarattığına değinmiş; bunu çözmek için Türkçe kelimeleri kök ve eklerden oluşacak şekilde ifade etmiştir. Sistem başarısını BLEU metriği ile değerlendirmiş ve 19.77 BLEU skorundan 26.87 BLEU skoruna artış sağlamıştır.

El-Kahlout [42] İngilizceden Türkçeye istatistiksel makine çeviri sistemi geliştirmiştir. Çeviri işlemi için; Moses aracı ve faktörlü makine çevirisi yöntemini kullanmış ve farklı sözdizimsel ve biçimbilimsel faktörlerin sistem başarısı üzerine etkisini test etmiştir. Sistem başarısını BLEU metriği ile değerlendirmiş ve yapılan çeviri neticesinde çeviri işlemi için 7.83 BLEU skoru kazanım sağlamışlardır.

Koehn ve Hoang [39], çalışmalarında faktörlü çeviri yöntemini tanıtmışlardır. Yaptıkları deneylerde İngilizce-Almanca dil çifti için 18.04'den 18.22 BLEU skoruna, İngilizce-İspanyolca dil çifti için 23.41'den 24.25 BLEU skoruna, İngilizce-Çekçe dil çifti için ise 25.82'den 27.62 BLEU skoruna artış sağlamışlardır.

2.5. Nöral Sistemler

Nöral sistemlerde çeviriler yapay sinir ağı modelleri kullanılarak öğrenilmektedir. Bu yöntemde; cümle daha küçük parçalara bölünmemekte, cümlenin çevirisi bir bütün olarak gerçekleştirilmektedir. Nöral sistemlerde kullanılan örnek bir çeviri mimarisi Şekil 2.2 üzerinde verilmiştir.



Şekil 2.2. Nöral sistemlerde kullanılan çeviri mimarisi

Kodlayıcı-çözümleyici yapay sinir ağı modellerinden oluşan bu mimari ilk olarak Kalchbrenner ve Blunsom [43] tarafından kullanılmıştır. Bu çeviri mimarisini kullanan sistemlerde çeviri iki aşamada gerçekleştirilmektedir. İlk olarak kaynak cümle, bir kodlayıcı yardımıyla, bağlam vektörüne dönüştürülmektedir. Daha sonra, çözümleyici, bu bağlam vektörünü kullanarak hedef dildeki cümleyi oluşturmaktadır. Kalchbrenner ve Blunsom sistemlerinde kodlayıcı olarak CNN; çözümleyici olarak ise RNN sinir ağı modeli kullanmışlardır. Günümüze kadar farklı yapay sinir ağı modellerinin kullandığı çalışmalar da gerçekleştirilmiştir [44], [45].

Nöral sistemlerde çevirinin bir bağlam vektörü üzerinden gerçekleştirilmesi cümle anlamının da çevirilerde kullanılabilmesine ve farklı uzunluktaki çevirilerin gerçekleştirilebilmesine olanak sağlamaktadır. Fakat farklı uzunluktaki tüm cümlelerin sabit bir bağlam vektörünü üzerinden gerçekleştirilmesi uzun cümlelerde çeviri kalitesinin düşmesine neden olmaktadır [46]. Bu sorunu çözmek için dikkat mekanizmasına sahip nöral sistemler geliştirilmiştir [45].

Son yıllarda nöral sistemler araştırmacılar tarafından büyük bir ilgi görmüştür. Google tarafından sunulan çeviri hizmetinde yine bir nöral çeviri sistemi olan GNMT [47] kullanılmaktadır. Fakat Nöral sistemlerde başarılı sonuçlar alabilmek için, kural tabanlı ve istatistiksel yöntemlere göre, daha büyük bir veri kümesi ihtiyacı bulunmaktadır. Tez çalışmasında, cümle tabanlı olarak hizalanmış yeterli veri kümesi oluşturulamadığı için nöral bir sistem kullanılmamıştır.

2.6. Hibrid Sistemler

Hibrid sistemler, farklı yöntemleri kullanan makine çevirisi sistemlerinin beraber kullanılarak, bu sistemlerin güçlü yönlerinin birleştirildiği sistemlerdir. Literatürde kural tabanlı ve istatistiksel sistemlerin farklı şekillerde birleştirildiği çalışmalar bulunmaktadır.

Costa-jussà ve arkadaşları [48], çalışmalarında hibrid çeviri sistemlerini; kural tabanlı sistemler tarafından yönetilen ve istatistiksel sistemler tarafından yönetilen sistemler olmak üzere iki başlık altında incelemişlerdir. Daha iyi sonucu veren çeviri yönteminin temel, diğer yöntemin ise yardımcı çeviri yöntemi olarak seçilmesini önermişlerdir. Ayrıca, hibrid yaklaşımların konuşma tanıma, bilgi çıkarımı, makine çevirisi gibi birçok alanda faydalı olduğunu da belirtmişlerdir.

Sanchez-Cartagena ve arkadaşları [49], istatistiksel sistemlerinde kullanılacak olan paralel veri kümesi bulunmasının zorluğuna değinmiş ve kural tabanlı çeviri sistemini ve istatistiksel çeviri sistemini birleştirmişlerdir. Çalışmalarında kullandıkları test veri kümesi üzerinde kural tabanlı çeviri sistemini çalıştırmış ve değiştirilmesi belirlenen ifadeleri istatistiksel sistemin kelime grubu çeviri tablosuna eklemişlerdir. Çalışmalarını paralel veri kümesinin az olarak kullanıldığı test senaryoları ile değerlendirmiş ve kullandıkları hibrid yöntem ile başarı skorunda artış gözlemlemişlerdir.

Tan ve arkadaşları [50], istatistiksel sistemlere pasif ve yaygın yöntemlerle iki-dilli sözlüğün eklenmesinin sistem başarısı üzerine etkilerini incelemişlerdir. Kullandıkları pasif yöntemde, istatistiksel sistemin eğitim veri kümesine iki-dilli sözlüğü eklemişlerdir. Yaygın yöntemde ise; istatistiksel sistemi, çözümleme aşamasında sözlük üzerindeki çeviri olasılıklarını kullanacak şekilde düzenlemişlerdir. Sözlük kayıtlarının farklı sayılarda istatistiksel sistemde kullanarak deneyler gerçekleştirmişlerdir. Yaptıkları deneyler sonucunda pasif ve yaygın yöntemin sistem başarısını arttırdığını gözlemlemişlerdir.

Tez çalışmasında, eski ve güncel Türkçe için büyük miktarlarda paralel veri kümesi bulunmasının zorluğundan dolayı hibrid bir sadeleştirme modeli geliştirilmiştir. Kural tabanlı ve istatistiksel modeller farklı yöntemler kullanılarak hibrid bir sistemde birleştirilmiştir.

3. VERİ KÜMESİ

Bu bölümde, ilk olarak Nutuk ile ilgili genel bilgiler sunulmuştur. Daha sonra, sadeleştirme veri kümeleri ve bu veri kümelerini oluşturma yöntemleri hakkında bilgiler verilmiş, kullanılan Nutuk veri kümesini oluşturma işlemleri açıklanmış ve oluşturulan veri kümesi üzerinde istatistiksel analizler gerçekleştirilmiştir.

3.1. Nutuk

Atatürk tarafından yazılan Nutuk 15–20 Ekim 1927 tarihleri arasında CHP kongresinde sunulmuştur. Sunum yaklaşık altı gün sürmüştür. Nutuk, sunumdan kısa bir süre sonra kitap haline getirilmiş ve 1927'de iki cilt halinde Osmanlıca olarak yayımlanmıştır [51].

Yayımlanmasının ardından, Nutuk, farklı dillere çevrilmiş; yabancı elçiliklere ve kütüphanelere dağıtılmış; birçok tarihçi, bilim adamı, sosyolog, edebiyatçı ve filozof tarafından incelenmiştir.

Nutukta 1919-1927 yılları arasında geçen Kurtuluş Savaşı, cumhuriyetin ilanı ve inkılap olayları anlatılmaktadır. Türkiye'nin Bağımsızlık mücadelesi ve Türkiye Cumhuriyeti'nin kurulması hakkında bilgi edinmek için başvurulan en önemli kaynakların başında gelmektedir.

Nutuk'un günümüze kadar yalnızca Türkçeye çevrilerek ya da sadeleştirilerek yayımlanmış birçok farklı sürümü bulunmaktadır [52]. Çalışmada Nutuk'un sadeleştirilmeden yayımlanan 1938 baskısı [9] eski Türkçe olarak, Bedi Yazıcı tarafından sadeleştirilerek 1995 yılından yayımlanan baskısı [10] güncel Türkçe olarak kabul edilmiştir.

3.1.1. Nutuk'un Üslup Özellikleri

Nutuk'un dili o devirde kullanılan Arapça ve Farsça kelimeler, bu dillerden alınan kurallarla oluşturulan tamlamalar ve devlet dilinde yazılmış olmasından dolayı günümüze kullanılmayan birçok kelime içermektedir.

Börekçi [53], Nutuk'ta kullanılan kelimelerin çoğunlukla Osmanlıcadan geldiğini ve bu kelimelerin çoğunun Arapça kökenli olduğunu belirtmiştir. Nutuk'tan seçtiği on sayfa üzerinde istatistiksel bir analiz gerçekleştirerek bunu doğrulamıştır. Yaptığı

analiz sonucunda Nutuk'ta geçen kelimelerin %58'inin Arapça, %30'unun Türkçe, %6'sının ise Farsça olduğunu tespit etmiştir.

Benzer bir çalışmada, Karamanlıoğlu [54] Atatürk'ün Gençliğe Hitabesi üzerinde istatistiksel bir analiz yapmıştır. Gençliğe Hitabe üzerinde geçen kelimelerin %57,3'ünün Arapça, %35,9'unun Türkçe, %6,8'inin ise Farsça olduğunu tespit etmiştir.

3.2. Metin Sadeleştirme Problemi İçin Geliştirilen Paralel Veri Kümeleri

İngilizcede için gerçekleştirilen metin sadeleştirme çalışmalarında PWKP [24] ve Newsela [55] gibi birçok farklı paralel veri kümesi kullanılmıştır. İngilizce için oluşturulmuş paralel veri kümeleri bulunmasına karşın, Türkçe için oluşturulmuş ve eski metinlerin güncelleştirilmesine odaklanan paralel bir veri kümesi bulunmamaktadır.

PWKP veri kümesinin oluşturulması için normal Wikipedia ve sadeleştirilmiş Wikipedia [21] kullanılmıştır. Normal Wikipedia'da bulunan cümleler karmaşık, sadeleştirilmiş Wikipedia'da bulunan cümleler ise basit olarak değerlendirilmiştir. İlk olarak, web üzerinden metin çıkarımı yapılmış; daha sonra, metinlerin cümle tabanlı olarak eşleştirilmesi sağlanmıştır. Oluşturulan veri kümesi toplam 108,016 paralel cümle çifti içermektedir.

Xu ve arkadaşları [55] PWKP veri kümesindeki eşleştirmelerde hatalar bulunduğuna ve eşleştirilmiş cümlelerin normal metinden daha sade olmadığına değinmişlerdir. Daha kaliteli bir sadeleştirme veri kümesi yaratmak amacıyla, editörler tarafından 4 farklı oranda sadeleştirilen gazete yazıları ile Newsela veri kümesini oluşturmuşlardır. Oluşturdukları veri kümesi yaklaşık 60.000 paralel cümle çifti içermektedir.

3.3. Cümle Tabanlı Eşleştirme

Cümle tabanlı eşleştirme işleminde kaynak ve hedef dildeki metinlerdeki cümlelerin hizalanması sağlanmaktadır. Cümle tabanlı eşleştirme işlemi; büyük veri kümelerinde hızlı ve en az hata ile gerçekleştirilebilmelidir. Eşleştirme işlemi için uzunluk tabanlı, kelime tabanlı ya da hibrid yöntemler kullanılabilmektedir [56].

Uzunluk tabanlı yöntemlerde; kısa cümlelerin kısa cümlelerle, uzun cümlelerin ise uzun cümlelerle eşleşeceği varsayılmaktadır. Tüm eşleştirmeler için cümle uzunluğuna dayalı bir olasılık hesaplanmakta ve bu olasılıklardan en yüksek ihtimale sahip olan eşleştirmeler dinamik programlama yöntemi kullanılarak aranmaktadır.

Kelime tabanlı yöntemlerde eşleştirmeler kelimeler üzerinden gerçekleştirilmektedir. Kelimelerin eşleştirme işleminde kullanılması daha az hata içeren eşleştirmelerin üretilebilmesine olanak sağlamıştır. Fakat bu yöntem uzunluk tabanlı yöntemlere göre daha fazla işlem gücü ve zaman gerektirmektedir [56].

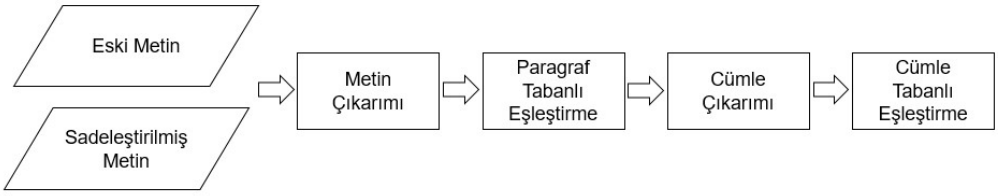
Ayrıca uzunluk tabanlı ve kelime tabanlı yöntemlerin birlikte kullanıldığı hibrid yöntemler de mevcuttur [56]. Bu yöntemlerde, büyük metinler uzunluk tabanlı yöntemler kullanılarak daha küçük parçalara bölünmektedir. Daha sonra uygulaması pahalı olan kelime tabanlı yöntemler bölünen parçalar üzerinde uygulanmaktadır.

Nutuk'un paralel veri kümesine dönüştürülmesi için LF Aligner [57] aracı kullanılmıştır. LF Aligner aracının altyapısında Hunalign [58] cümle tabanlı eşleştirme algoritması kullanılmaktadır. Hunalign'da uzunluk ve kelime tabanlı cümle eşleştirme yöntemleri birlikte kullanılmaktadır. Kaynak ve hedef cümleler için benzerlik skorları kelimeler ve cümle uzunluğu üzerinden birlikte hesaplanmaktadır. Uzunluk tabanlı yöntemde; kaynak cümlenin karakter sayısının, hedef cümlenin karakter sayısına oranı kullanılmaktadır. Çıkartılan cümle benzerlik skorları üzerinden eşleştirme işlemi gerçekleştirilmektedir. Ayrıca, araçta, kelime tabanlı ve uzunluk tabanlı yöntemlerin birbirlerine göre ağırlıkları Macarca ve İngilizce dil çifti üzerinden öğrenilmiştir. Fakat öğrenilen bu ağırlıkların diğer dil çiftleri için de kullanılabileceği belirtilmiştir [58].

Hunalign'de ortak kelime içermeyen farklı metinlerin eşleştirilme başarısını arttırmak için harici iki-dilli sözlükler kullanılabilmektedir. Sözlük kullanılması durumda, kaynak metin üzerinde çeviri işlemi gerçekleştirilmekte; çevrilen metin kullanılarak eşleştirme işlemi yapılmaktadır. Nutuk veri kümesindeki eski ve güncel Türkçede ortak kelimeler bulunmasından dolayı, detaylı bir sözlük oluşturma ihtiyacı duyulmamıştır.

3.4. Nutuk Paralel Veri Kümenin Oluşturma Yöntemi

Paralel veri kümesini oluşturmak için uygulanan işlem serisi Şekil 3.1 üzerinde gösterilmiştir.



Şekil 3.1. Paralel veri kümesini oluşturmak için uygulanan işlem serisi

İlk olarak, dosyalar üzerinden metin çıkarımı yapılmaktadır. PDF dosyaları metin dosyası biçimine dönüştürülmekte ve sonlarında kısa çizgi bulunan kelimelerin normalleştirilmesi sağlanmaktadır. Aynı zamanda sayfa numaraları gibi istenmeyen bazı ifadelerin silinmesi yine bu aşamada gerçekleştirilmektedir. Metin çıkarımı gerçekleştirilen bir örnek Şekil 3.2'de verilmiştir.

1335 senesi Mayısının 19 uncu günü Samsuna çıktım. Vaziyet ve manzarai umumiye:

Osmanlı Devletinin dahil bulunduğu grup, Harbi Umumîde mağlûp ol- ←
muş, Osmanlı ordusu her tarafta zedelenmiş, şeraiti ağır, bir mütarekename im- ←
zalanmış. Büyük Harbin uzun seneleri zarfında, millet yorgun ve fakir bir hal- ←
de. Millet ve memleketi Harbi Umumîye sevkedenler, kendi hayatları endişesi- ←
ne düşerek, memleketten firar etmişler. Saltanat ve hilâfet mevkiini işgal eden
Vahdettin, müstereddi, şahsını ve yalnız tahtını temin edebileceğini tahayyül et- ←
tiği denî tedbirler araştırmakta. Damat Ferit Paşanın riyasetindeki kabine; âciz,
haysiyetsiz, cebîn, yalnız padişahın iradesine tâbi ve onunla beraber şahısları- ←
nı vikaye edebilecek herhangi bir vaziyete razı. **PDF**

⇩

1335 senesi Mayısının 19 uncu günü Samsuna çıktım. Vaziyet ve manzarai umumiye:
Osmanlı Devletinin dahil bulunduğu grup, Harbi Umumîde mağlûp olmuş, Osmanlı ordusu her tarafta zedelenmiş, şeraiti ağır, bir mütarekename imzalanmış. Büyük Harbin uzun seneleri zarfında, millet yorgun ve fakir bir halde. Millet ve memleketi Harbi Umumîye sevkedenler, kendi hayatları endişesine düşerek, memleketten firar etmişler. Saltanat ve hilâfet mevkiini işgal eden Vahdettin, müstereddi, şahsını ve yalnız tahtını temin edebileceğini tahayyül ettiği denî tedbirler araştırmakta. Damat Ferit Paşanın riyasetindeki kabine; âciz, haysiyetsiz, cebîn, yalnız padişahın iradesine tâbi ve onunla beraber şahıslarını vikaye edebilecek herhangi bir vaziyete razı. **TXT**

Şekil 3.2. Metin çıkarımı gerçekleştirilen bir örnek

Metin çıkarımı işleminin sonucunda elde edilen metinler, LF Aligner aracı kullanılarak, paragraf tabanlı olarak eşleştirilmektedir. Daha sonra, eşleştirilen her bir paragraf Zemberek aracının [15] cümle sonu belirleme özelliği kullanılarak cümlelere bölünmekte ve yine LF Aligner aracı ile bu kez cümle tabanlı eşleştirme işlemi gerçekleştirilmektedir. Zemberek aracında cümle sonlarının belirlenmesi için manuel olarak tanımlanmış kurallar ve çok katmanlı algılayıcı [59] ile eğitilen bir sınıflandırıcı model kullanılmaktadır [60]. Çizelge 3.1 üzerine oluşturulan veri kümesinden birkaç örnek cümle verilmiştir. "K" ile kaynak metin, "R" ile sadeleştirilmiş Nutuk'tan elde edilen referans metin ifade edilmektedir.

Çizelge 3.1. Oluşturulan veri kümesindeki örnek cümle çiftleri

Cümle Çiftleri
K: "**1335** senesi Mayısının 19 uncu günü Samsuna çıktım." **R:** "**1919** senesi Mayısının 19 uncu günü Samsuna çıktım."
K: "**Vaziyet ve manzarai umumiye**: Osmanlı Devletinin **dahil** bulunduğu grup, **Harbi Umumîde mağlûp olmuş**, Osmanlı ordusu her tarafta zedelenmiş, **şeraiti** ağır, bir **mütarekename** imzalanmış." **R:** "**Genel durum ve görünüm**: Osmanlı Devletinin **içinde** bulunduğu grup, **Genel Savaşta yenilmiş**, Osmanlı ordusu her tarafta zedelenmiş, **şartları** ağır bir **ateşkes anlaşması** imzalanmış."
K: "Büyük **Harbin** uzun seneleri **zarfında**, millet yorgun ve fakir bir halde." **R:** "Büyük **Savaşın** uzun seneleri **içinde**, millet yorgun ve fakir bir halde."
K: "Millet ve memleketi **Harbi Umumîye** sevkedenler, kendi **hayatları endişesine** düşerek, memleketten **firar etmişler.**" **R:** "Millet ve memleketi **Genel Savaşa** sevkedenler, kendi **hayatlarının kaygısına** düşerek, memleketten **kaçmışlar.**"
K: "**Saltanat** ve **hilâfet mevkiini işgal eden** Vahdettin, **müstereddi**, **şahsını** ve yalnız tahtını **temin edebileceğini tahayyül ettiği denî tedbirler** araştırmakta." **R:** "**Padişah** ve **halife olan** Vahdettin, **soysuz**, **kendini** ve yalnız tahtını **koruyabileceğini hayal ettiği alçakça önlemler** araştırmakta."
K: "Damat Ferit Paşanın **riyasetindeki** kabine; âciz, haysiyetsiz, **cebîn**, yalnız padişahın **iradesine tâbi** ve onunla beraber şahıslarını **vikaye edebilecek** herhangi bir **vaziyete** razı." **R:** "Damat Ferit Paşanın **başkanlığındaki** kabine; âciz, haysiyetsiz, **korkak**, yalnız padişahın **iradesi altında** ve onunla beraber şahıslarını **esirgeyebilecek** herhangi bir **duruma** razı."
K: "Ordunun elinden **esliha** ve cephanesi alınmış ve alınmakta..." **R:** "Ordunun elinden **silâhları** ve cephanesi alınmış ve alınmakta..."
K: "İtilâf Devletleri, **mütareke ahkâmına riayete lüzum** görmüyorlar." **R:** "İtilâf Devletleri, **ateşkes hükümlerine uymayı gerekli** görmüyorlar."
K: "Birer **vesile** ile, İtilâf donanmaları ve askerleri İstanbulda." **R:** "Birer **bahane** ile, İtilâf donanmaları ve askerleri İstanbulda."
K: "Adana **vilâyeti**, Fransızlar; Urfa, Maraş, **Ayıntap**, İngilizler tarafından işgal edilmiş." **R:** "Adana **ili**, Fransızlar; Urfa, Maraş, **Gaziantep**, İngilizler tarafından işgal edilmiş."

Oluşturulan veri kümesi toplam 15.835 cümle çifti içermekte ve eski kelime ve kelime gruplarının güncel Türkçedeki karşılıkları ile değiştirildiği birçok örnek barındırmaktadır.

3.5. İstatistiksel Analiz

Oluşturulan veri kümesi üzerinde çeşitli istatistiksel analiz işlemleri gerçekleştirilmiştir. Gerçekleştirilen bu analizler sonucunda elde edilen veriler Çizelge 3.2 üzerinde verilmiştir.

Çizelge 3.2. Veri kümesi üzerinde gerçekleştirilen istatistiksel analiz sonuçları

	Kaynak Metin	**Sadeleştirilmiş Metin**
Toplam Cümle Sayısı	15.835	15.835
Toplam Kelime Sayısı	273.604	264.759
Toplam Farklı Kelime Sayısı	37.989	36.467
Cümle Başına Ort. Kelime Sayısı	17,27	16,71
Kelime Başına Ort. Karakter Sayısı	7,19	7,47

Cümlelerin çok sayıda kelimeden oluşması, metnin zorluğuna dikkat çekmektedir. Kaynak ve sadeleştirmiş metinde cümle uzunlukları temel olarak benzer boyutlarda kalmıştır. Bu sadeleştirmelerde cümle yapısının büyük oranda korunmasının bir sonucudur. Elle kontrol edilen 100 cümlede eşleştirmelerin %90 oranında doğru olduğu görülmüştür. Hataların ise kaynak ve hedef metindeki farklılıklardan ya da bazı örneklerde paragrafların cümlelere doğru olarak bölünememesinden kaynaklandığı gözlemlenmiştir.

4. KURAL TABANLI SADELEŞTİRME MODELİ

Bu bölümde geliştirilen kural tabanlı sadeleştirme modeli incelenmiştir. Geliştirilen modelde sadeleştirmeler oluşturulan sözlük üzerinden gerçekleştirilmektedir.

4.1. Modelde Kullanılan Bileşenler

Kural tabanlı modelde kelimeler üzerinde biçimbilimsel analiz, sözcüksel aktarım ve biçimbilimsel sentez işlemleri gerçekleştirilerek aday kelimeler oluşturulmaktadır. Daha sonra bu kelimelerin dil modeli sorgulama işleminde değerlendirilmesi gerçekleştirilmektedir.

Modelde biçimbilimsel analiz ve biçimbilimsel sentez işlemleri için TRMorph [61]; sözcüksel aktarım için Osmanlıca-Türkçe sözlük; dil modeli sorgulama işlemi için ise dil modeli kullanılmaktadır. Bu bölümde modelde kullanılan bileşenler incelenmiştir.

4.1.1. TRMorph

TRMorph [61], Türkçe için geliştirilmiş bir biçimbilimsel çözümleyicidir. Araç ile kelimelerin yüzey biçimleri ve biçimbilimsel özellikleri arasındaki ilişkiler belirlenebilmekte ve kelimeler; kelime kökü, türü ve kipi gibi bileşenlerine ayrılabilmektedir. Aynı zamanda kelime kökü ve biçimbilimsel özellikler ile yüzeysel biçimdeki kelimeler de sentezlenebilmektedir. TRMorph aracında biçimbilimsel çözümleme ve sentez işlemleri, araçta tanımlanan sonlu durum kuralları kullanılarak gerçekleştirilmektedir. Bu kurallar Foma [62] aracı yardımıyla, lexc ve xfst [63] sonlu durum kural tanımlama dillerinde oluşturulmuştur.

TRMorph'ta biçimbilimsel analizin gerçekleştirilmesinde bir sözlük kullanılmaktadır [61]. Bu sözlük, büyük web kaynaklarının işlenmesiyle oluşturulmuştur. Sözlükte bulunan hataların giderilmesi; Zemberek aracının yazım denetleyicisi ve elle yapılan incelemeler ile gerçekleştirilmiştir. TRMorph'ta aynı zamanda kelimenin sözlükte bulunmaması durumunda tüm olası biçimbilimsel olasılıkların üretildiği bilinmeyen kelime tahmin edici [61] de bulunmaktadır. Kural tabanlı modelde, eski Türkçe kelimelerin büyük bölümünün sözlük üzerinde bulunmamasından dolayı, bu kelimeler için, bilinmeyen kelime tahminci kullanılmıştır. Bu kelimelerin analizi sonucunda TRMorph tarafından üretilen tüm olasılıklar ele alınmaktadır.

TRMorph'ta, sözlük üzerinde bulunan bir kelime için birden fazla olasılık üretilebilmektedir. Bu durum Türkçedeki eklerin muğlak yapısından kaynaklanmaktadır [64]. Çöltekin [61] TRMorph'un benzer özellikler sunan diğer araçlardan daha fazla muğlak biçimbilimsel olasılık sunduğunu belirtmiştir. TRMorph aracı tarafından birden fazla biçimbilimsel olasılığı bulunduğu belirlenen birkaç örnek Çizelge 4.1'de verilmiştir.

Çizelge 4.1. TRMorph aracında üretilen muğlak biçimbilimsel analiz örnekleri

Kelime	**Biçimbilimsel Olasılık**
yüz	yüz<N>
	yüz<Num>
	yüz<V><imp><2s>
buna	bu<Prn:dem><dat>
	buna<V><imp><2s>
	bun<N><dat>
evleri	ev<N><p3p> (ev-leri)
	ev<N><pl><acc> (ev-ler-i)
doktorlar	doctor<N><pl>
	doctor<N><0><V><cpl:pres><3p>

Örneklerdeki kelimelerin doğru biçimbilimsel olasılığı kullanıldıkları bağlama göre farklılık göstermektedir. Kural tabanlı modelde TRMorph aracı tarafından belirlenen tüm olasılıkların ele alınması sağlanmaktadır. TRMorph aracında üretilen kelime türü işareti ve biçimbilimsel ek olasılıklarının listesi EK 1'de verilmiştir.

4.1.2. Sözlük

Kural tabanlı sistemlerin en önemli bileşeni kullanılan iki dilli sözlüktür. Sözlük, bu sistemlerde kelimenin hedef dile aktarılması için kullanılmaktadır. Kural tabanlı sistemlerin başarısı, büyük oranda kullanılan sözlüğün kalitesine ve büyüklüğüne bağlıdır.

Geliştirilen kural tabanlı sadeleştirme modelinde sadeleştirmeler sözlük üzerinden gerçekleştirilmektedir. Bu amaçla Osmanlıca-Türkçe [65], [66] ve Öztürkçe [67] sözlükler işlenerek Osmanlıca-Türkçe sözlük oluşturulmuştur. Oluşturulan sözlükte toplam 255,747 kayıt bulunmaktadır. Oluşturulan sözlükten birkaç kayıt örneği Çizelge 4.2'de verilmiştir.

Çizelge 4.2. Oluşturulan sözlükten kayıt örnekleri

Kaynak	**Hedef**
kumandan	komutan
kumandanlık	komutanlık
kumanya	gemi zahiresi
kumanya	gemi kileri

Sözlükte, bir kelimenin birden fazla güncel Türkçe karşılığı bulunabilmektedir. Ayrıca, sözlüğe yeni kayıtlar da eklenebilmektedir. Bu etmen geliştirilen kural tabanlı sadeleştirme modelini oldukça genişletilebilir ve sürdürülebilir kılmaktadır. Kural tabanlı modelde, yalnızca işlenen kaynaklardan sözlüğe ekleme yapılmış; Nutuk veri kümesine özel bir işlem gerçekleştirilmemiştir.

4.1.3. Dil Modeli

Dil modelleri; makine çevirisi, konuşma tanıma ve yazım düzeltme gibi dil işleme uygulamalarında cümlelerin ya da kelime gruplarının birlikte bulunma olasılıklarının hesaplanmasında kullanılmaktadır.

Çalışmada cümle ve kelime gruplarının bir arada bulunma olasılığının hesaplanması için KenLM [68] aracı kullanılmıştır. KenLM aracında kelime gruplarının bir arada bulunma olasılıkları N-Gram ve maksimum olasılık tahmini (MLE) yöntemleri kullanılarak hesaplanmaktadır. MLE yönteminde ifadelerin metin üzerindeki frekansları belirlenmekte, daha sonra bu frekanslar normalize edilerek değerlerin 0 ve 1 arasında ölçeklendirilmesi sağlanmaktadır. Bu olasılığın hesaplanmasında kullanılan yöntem Denklem 1 üzerinde verilmiştir.

$$p(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_{w'} C(w_{n-1}w')} \quad (1)$$

Denklemde $p(w_n|w_{n-1})$ ifadesi w_{n-1} kelimesinden sonra w_n kelimesinin gelme olasılığını, $C(w_{n-1}w_n)$ ifadesi $w_{n-1}w_n$ ifadesinin metin üzerindeki sayısını, $\sum_{w'} C(w_{n-1}w')$ ifadesi ise metin üzerinde w_{n-1} ile başlayan ifade sayısını temsil etmektedir. Örneğin; “kuvayi millîye” ifadesinin bir arada bulunma olasılığının hesaplaması için, metin üzerinde geçen “kuvayi millîye” ifade sayısının “kuvayi” kelimesi ile başlayan ifade sayısına oranı kullanılmaktadır.

Denklem 1'i farklı N değerlerini kapsayacak şekilde genelleştirmek mümkündür. Genelleştirilmiş bu yöntem Denklem 2'deki gibidir. Denklemde N değeri w_n kelimesinin öncesindeki kelime sayısını ifade etmektedir.

$$p(w_n|w_{\mathrm{n-N+1}}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}w_N)}{\sum_{w'} C(w_{n-N+1}^{n-1}w')} \quad (2)$$

Cümledeki kelimelerin bir arada bulunma olasılığının hesaplanmasında, cümle üzerindeki bir kelimenin veri kümesinde bulunmaması; tüm cümlenin olasılığının 0 olarak hesaplanmasına neden olmaktadır. Bu sorunu çözmek amacıyla Laplace, Good Turing ve Kneser-Ney gibi farklı yöntemler geliştirilmiştir. Bu yöntemlerde backoff ya da interpolasyon gibi farklı metotlara başvurulabilmektedir. Backoff metodunda üst mertebe bir N-Gram yerine daha alt mertebe bir N-Gram kullanılmakta, interpolasyon yönteminde ise farklı mertebeden N-Gramların birleştirilmesi sağlanmaktadır [69].

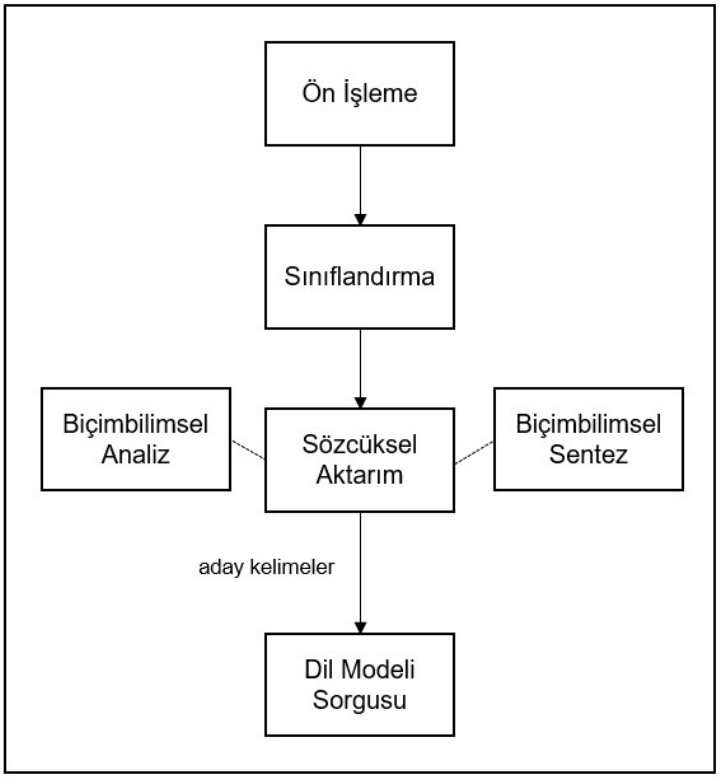
KenLM aracında Kneser-Ney düzgünleştirme yöntemi kullanılmaktadır. Bu yöntemde, interpolasyon metodu kullanılarak, farklı mertebeden hesaplanan N-Gram olasılıklarının birleştirilmesi sağlanmaktadır [70]. Üst mertebe N-Gramların 0 yada 0'a yakın olduğu durumlarda alt mertebe dil modeli sonuçlarına, normal olduğu durumlarda ise üst mertebeden gelen dil modeli sonuçlarına daha fazla ağırlık verilmektedir. Kneser-Ney düzgünleştirmesinde kullanılan yöntem Denklem 3'deki gibidir.

$$Pabs(w_i|w_{\mathrm{i-1}}) = \frac{max(C(w_{i-1}w_i - \delta, 0))}{\sum_{w'} C(w_{i-1}w')} \alpha Pabs(\, w_i\,) \quad (3)$$

Denklemde 2-Gram için uygulanan Kneser-Ney yöntemi verilmiştir. Denklemde; δ indirgeme [69], α ise interpolasyon işleminde kullanılan normalleştirme parametresidir. İndirgeme parametresi ile frekans değerlerinden küçük bir değer çıkarılmakta ve bunun sonucu olarak veri kümesi üzerinde hiç bulunmayan ifadeler için olasılık uzayı oluşturulmaktadır. α değeri kullanılarak ise tüm mertebeden gelen N-Gram ağırlıklarının 1 olması sağlanmaktadır.

4.2. Modelde Uygulanan İşlem Serisi

Şekil 4.1 üzerinde kural tabanlı sadeleştirme modelinde uygulanan işlem serisi verilmiştir.



Şekil 4.1. Kural tabanlı sadeleştirme modelinde uygulanan işlem serisi

Model ilk olarak ön işleme adımı ile başlamaktadır. Bu adımda kaynak cümle sistemin geri kalanında kullanılmak üzere hazırlanmaktadır. Daha sonra, kelimenin değiştirilip değiştirilmemesi gerektiğine göre sınıflandırılması sağlanmaktadır. Sınıflandırma işlemi sonucunda değiştirilmesi gerektiği belirlenen her bir kelime üzerinde sözcüksel aktarım işlemi gerçekleştirilerek sadeleştirme adayları oluşturulmaktadır. Sözcüksel aktarım işleminde doğrudan olarak sözlük üzerinden adaylar oluşturulabildiği gibi, yalnızca kaynak ya da kaynak ve hedef kelimeler üzerinde biçimbilimsel analiz ve biçimbilimsel sentez işlemleri gerçekleştirilerek de sadeleştirme adayları oluşturulabilmektedir. Aday kelimeler, dil modeli sorgusu ile değerlendirilmekte ve en yüksek skora sahip olan kelimenin seçilmesi sağlanmaktadır.

4.2.1. Ön İşleme

Bu işlemde, kaynak cümle sistemin geri kalanında kullanılmak üzere hazırlanmaktadır. Veri kümesinden istenmeyen karakterler temizlenmekte, tüm metin küçük harfe dönüştürülmekte ve cümlelerin kelimelere bölünmesi gerçekleştirilmektedir. Bu işlem için Zemberek aracı kullanılmıştır. Örnek bir cümle üzerinde gerçekleştirilen ön işleme adımı Çizelge 4.3'te verilmiştir.

Çizelge 4.3. Ön işleme adımı gerçekleştirilen örnek bir cümle

Kaynak Cümle	"1335 senesi **M**ayısının 19 uncu günü **S**amsuna çıktı**m.**"
İşlenen Cümle	"1335 senesi **m**ayısının 19 uncu günü **s**amsuna çıktı**m .**"

4.2.2. Sınıflandırma

Sınıflandırma işleminde, kelimenin eski olup olmadığına göre sınıflandırılması gerçekleştirilmektedir. Bu işlem için kelimelerin eski ve güncel Türkçedeki frekans değişim oranları hesaplanmıştır. Değişimin hesaplanmasında kullanılan yöntem Denklem 4'teki gibidir.

$$Yüzdesel\ Değişim\ (\%) = \frac{f_{sadeleştirilmiş}(\text{kelime}) - f_{kaynak}(\text{kelime})}{f_{kaynak}(\text{kelime})}\ x\ 100 \qquad (4)$$

Denklem sonucunda elde edilen değer, kelimenin eski ve günümüz Türkçesinde hangi oranda değişim gösterdiğini yüzdesel olarak ifade etmektedir. Kelime frekansları için Nutuk veri kümesinin eğitim için ayrılan kısmı kullanılmıştır. Birkaç örnek cümle üzerinde gerçekleştirilen sınıflandırma örneği Çizelge 4.4 üzerinde verilmiştir. Koyu renkle ifade edilen kelimelerin değiştirilmesi gerektiği belirlenmiştir.

Çizelge 4.4. Örnek cümleler üzerinde gerçekleştirilen sınıflandırma işlemi

birinci **fırka kumandanı** mustafa asım **vesika** , **163** .
şam **ahalisinin** emiri **değilsin** .
düzce ve **havalisinde vaziyetin şayanı emniyet** olduğunu benden daha iyi **istihbar ediyor** .
ermenilerin **fazla mutalebatına** hak vermeksizin **hudutlarda** bazı **tashihatın icrasına razı** oluruz .

Farklı eşik değerleri kullanıldığında sınıflandırma işleminin sonucu değişiklik göstermektedir. Çalışmada, farklı eşik değerlerinin etkisi Nutuk veri kümesinin test

için ayrılan kısmı kullanılarak değerlendirilmiştir. Test veri kümesindeki referans metinlerde değiştirilen kelimeler doğru örnek olarak kabul edilmektedir.

4.2.3. Biçimbilimsel Analiz

Biçimbilimsel açıdan zengin dillerde, yapım ve çekim ekleri kullanılarak yeni kelimeler üretilebilmektedir. Geliştirilen kural tabanlı sadeleştirme modelinde birçok kelime, yapım ve çekim ekleri nedeniyle, sözlük üzerinde doğrudan bulunamamaktadır. Bu nedenle, bu işlemde, kelimeler üzerinde biçimbilimsel analiz işlemi gerçekleştirilerek, kelimeler yüzeysel biçiminden kelime kökü, kelime türü işareti ve biçimbilimsel eklerine ayrıştırılmaktadır. Gerçekleştirilen birkaç biçimbilimsel analiz örneği Çizelge 4.5'de verilmiştir.

Çizelge 4.5. Biçimbilimsel analiz örnekleri

Kaynak Kelime	**Kelime Kökü**	**Kelime Türü İşareti**	**Biçimbilimsel Ek Olasılıkları**
kumandanı	kumanda	<N>	<p2s><acc>, <p2s><acc><0><V>, <p2s><acc><0><V><cpl:pres>, <p2s><acc><0><V><cpl:pres><3s>
	kumandan	<N>	<acc>, <acc><0><V>, <acc><0><V><cpl:pres>, <acc><0><V><cpl:pres><3s>, <p3s>, <p3s><0><V>, <p3s><0><V><cpl:pres>, <p3s><0><V><cpl:pres><3s>
ahalisinin	ahali	<N>	<p3s><gen>, <p3s><gen><0><V>, <p3s><gen><0><V><cpl:pres>, <p3s><gen><0><V><cpl:pres><3s>
	ahâli	<N>	<p3s><gen>, <p3s><gen><0><V>, <p3s><gen><0><V><cpl:pres>, <p3s><gen><0><V><cpl:pres><3s>
vaziyeti	vaziyet	<N>	<acc>, <acc><0><V>, <acc><0><V><cpl:pres>, <acc><0><V><cpl:pres><3s>, <p3s>, <p3s><0><V>, <p3s><0><V><cpl:pres>, <p3s><0><V><cpl:pres><3s>
tashihatın	tashihat	<N>	<p2s><0><V><cpl:pres><3s>, <gen>, <p2s><0><V><cpl:pres>, <gen><0><V><cpl:pres><3s>, <gen><0><V>, <p2s>, <gen><0><V><cpl:pres>, <p2s><0><V>

"kumandanı" kelimesi için "kumanda" ve "kumandan" olmak üzere iki farklı kelime kökü belirlenmiştir. TRMorph aracında bir kelime için birden fazla biçimbilimsel olasılık belirlenebilmektedir. Kelimelerin belirlenen kelime türü işaretleri ve biçimbilimsel ek olasılıkları, kelimelerin sağında verilmiştir.

4.2.4. Biçimbilimsel Sentez

Biçimbilimsel sentez, biçimbilimsel analiz işleminin tersidir. Bu işlemde, kelimenin kökü, kelime türü işareti ve biçimbilimsel ekleri kullanılarak yüzeysel biçimdeki kelime oluşturulmaktadır. Gerçekleştirilen birkaç biçimbilimsel sentez örneği Çizelge 4.6 üzerinde verilmiştir.

Çizelge 4.6. Biçimbilimsel sentez örnekleri

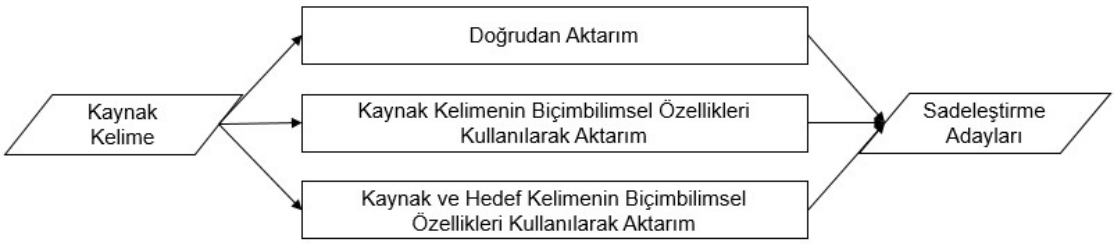
Kaynak Kelime	Hedef Kelime Kökü	Kelime Türü İşareti	Biçimbilimsel Ek Olasılıkları	Belirlenen Kelime
kumandanı	komutan	<N>	<acc>	komutanı
			<acc><0><V>	
			<acc><0><V><cpl:pres>	
			<acc><0><V><cpl:pres><3s>	
			<p3s>	
			<p3s><0><V>	
			<p3s><0><V><cpl:pres>	
			<p3s><0><V><cpl:pres><3s>	
ahalisinin	halk	<N>	<p3s><gen>	halkının
			<p3s><gen><0><V>	
			<p3s><gen><0><V><cpl:pres>	
			<p3s><gen><0><V><cpl:pres><3s>	
vaziyeti	durum	<N>	<acc>	durumu
			<acc><0><V>	
			<acc><0><V><cpl:pres>	
			<acc><0><V><cpl:pres><3s>	
			<p3s>	
			<p3s><0><V>	
			<p3s><0><V><cpl:pres>	
			<p3s><0><V><cpl:pres><3s>	
tashihatın	düzeltme	<N>	<V><vn:inf><N><pl><gen>	düzeltmelerin
			<V><vn:inf><N><pl><gen><0><V>	
			<V><vn:inf><N><pl><gen><0><V><cpl:pres>	
			<V><vn:inf><N><pl><gen><0><V><cpl:pres><3s>	

Çizelgedeki ilk örnekte; “komutan” kelime kökü, “<N>” kelime türü işareti ve “<acc>” biçimbilimsel eki kullanılarak, yüzeysel biçimdeki “komutanı” kelimesinin belirlenebildiği görülmektedir.

4.2.5. Sözcüksel Aktarım

Sözcüksel aktarım işleminde, kaynak kelimenin oluşturulan sözlük üzerinden hedef kelimeye çevrilmesi sağlanmaktadır. Çevrim, yalnızca sözlük üzerinden yapılabildiği gibi, biçimbilimsel analiz ve sentez işlemleri gerçekleştirilerek de yapılabilmektedir.

Farklı biçimbilimsel analiz ve sentez işlemlerinin bileştirilmesine göre farklı sözcüksel aktarım yöntemleri uygulanabilmektedir. Sözcüksel aktarımda kullanılan bu yöntemler Şekil 4.2'de verilmiştir.



Şekil 4.2. Sözcüksel aktarım yöntemleri

Sadeleştirmeler biçimbilimsel analiz ve sentez işlemleri kullanılmadan doğrudan aktarım yöntemiyle oluşturulabildiği gibi, kelimeler üzerinde biçimbilimsel analiz ve sentez işlemleri gerçekleştirilerek de oluşturulabilmektedir. En son olarak tüm yöntemler tarafından üretilen sadeleştirme adaylarının birleştirilmesi sağlanmaktadır.

Doğrudan Aktarım

Doğrudan aktarım yönteminde kaynak ve sözlük üzerinden bulunan kelimeler üzerinde biçimbilimsel analiz ve sentez işlemleri gerçekleştirilmemekte; sadeleştirme adayları yalnızca oluşturulan sözlük kullanılarak üretilmektedir. Birkaç doğrudan aktarım örneği Çizelge 4.7 üzerinde verilmiştir.

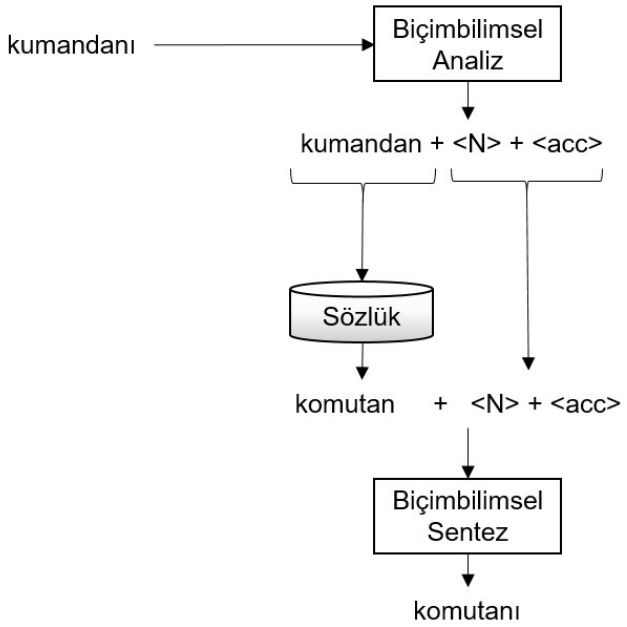
Çizelge 4.7. Doğrudan aktarım örnekleri

Kaynak	**Hedef**
kumandan	komutan
kumanda	komuta
ahali	umum, yaşayanlar, ahali, nas, halk, insan topluluğu
vaziyet	duruş, durum, konum, hal
tashihat	tashihler, düzeltmeler

Sözlük üzerinde kaynak bir kelime için birden fazla hedef kelime bulunabilmektedir. Bu durumda tüm kelimeler aday sadeleştirme olarak değerlendirilmektedir.

Kaynak Kelimenin Biçimbilimsel Özellikleri Kullanılarak Aktarım

Bu yöntemde yalnızca kaynak kelime üzerinde biçimbilimsel analiz işlemi gerçekleştirilerek aday kelime listesi oluşturulmaktadır. Şekil 4.3 üzerinde yalnızca kaynak kelime üzerinde analiz işlemi gerçekleştirilerek yapılan bir aktarım örneği verilmiştir.



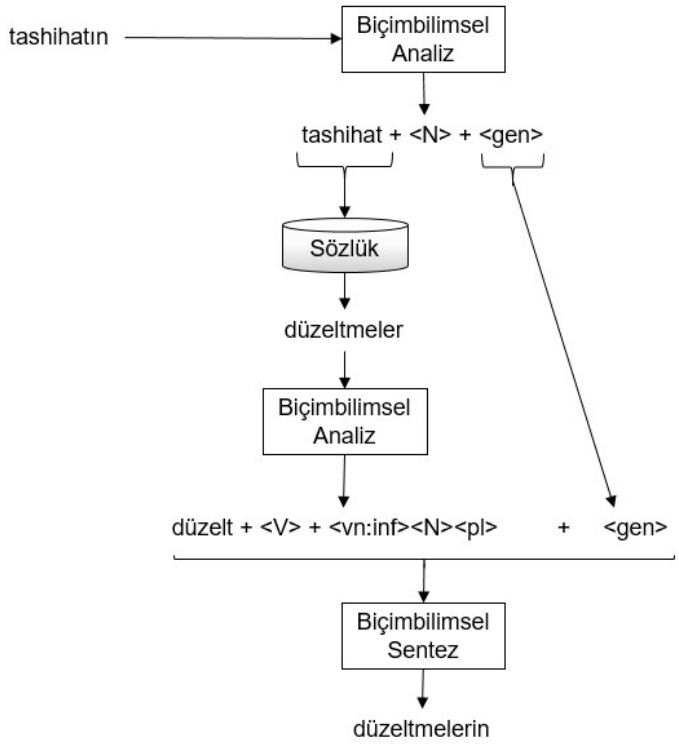
Şekil 4.3. Kaynak kelimenin biçimbilimsel özellikleri kullanılarak yapılan bir aktarım örneği

İlk olarak “kumandanı” kaynak kelimesi üzerinde biçimbilimsel analiz işlemi gerçekleştirilerek kelime; kelime kökü, kelime türü işareti ve biçimbilimsel eklerine ayrılmıştır. Daha sonra kelimenin kökü kullanılarak sözlük araması işlemi gerçekleştirilmiş, “komutan” kelimesi tespit edilmiş ve tespit edilen bu kelimeye kaynak kelimenin biçimbilimsel özellikleri eklenerek sadeleştirilmiş “komutanı” kelimesi oluşturulmuştur.

Kaynak ve Hedef Kelimenin Biçimbilimsel Özellikleri Kullanılarak Aktarım

Bu yöntemde kaynak ve hedef kelimeler üzerinde biçimbilimsel analiz işlemi gerçekleştirilmektedir. Bu sayede yalnızca doğrudan aktarım ve kaynak kelimenin biçimbilimsel özelliklerinin kullanıldığı aktarım yöntemlerinde oluşturulamayan sadeleştirmelerin oluşturulması sağlanmıştır. Şekil 4.4’de hem kaynak hem de

hedef kelime üzerinde biçimbilimsel analiz işleminin gerçekleştirildiği bir örnek verilmiştir.



Şekil 4.4. Kaynak ve hedef kelimenin biçimbilimsel özellikleri kullanılarak yapılan bir aktarım örneği

İlk olarak “düzeltmeler” kaynak kelimesi üzerinde biçimbilimsel analiz işlemi gerçekleştirilerek kelime; kelime kökü, kelime türü işareti ve biçimbilimsel eklerine ayrılmıştır. Daha sonra kelimenin kökü kullanılarak sözlük araması işlemi gerçekleştirilmiş, “düzeltmeler” kelimesi tespit edilmiş ve bu kelime üzerinde yeniden biçimbilimsel analiz işlemi gerçekleştirilmiştir. En son olarak, hedef kelime kökü, hedef kelime biçimbilimsel özellikleri ve kaynak kelime biçimbilimsel ekleri birleştirilerek sadeleştirilmiş “düzeltmelerin” kelimesi oluşturulmuştur.

4.2.6. Dil Modeli Sorgusu

Biçimbilimsel analiz, sözcüksel aktarım, biçimbilimsel sentez işlemleri sonucunda birden fazla sadeleştirme adayı belirlenebilmektedir. Aynı zamanda, dilin yapısından doğan belirsizlikten dolayı, biçimbilimsel analiz aşamalarında yanlış

kelime kökleri bulunabilmekte ve bunun sonucunda yanlış sadeleştirme adayları oluşturulabilmektedir.

Dil modeli sorgulama işleminde, aday kelimelerin değerlendirilmesi ve en yüksek başarı skoruna sahip olan kelimenin seçilmesi sağlanmaktadır. Gerçekleştirilen örnek bir dil modeli sorgusu örneği Çizelge 4.8 üzerinde verilmiştir. Cümleler için 10 tabanında logaritma işlemi uygulanarak elde edilen N-Gram sonuçları kullanılmaktadır. Kural tabanlı modelde, dil modelinin eğitimi ve sorgulaması için KenLM aracı [68] kullanılmıştır.

Çizelge 4.8. Dil modeli değerlendirme örneği

Değerlendirilen Cümle	**Değerlendirme Sonucu**
[1] "muhterem efendiler , bu vaziyeti hep beraber mütalea etmeye medar olacak kadar malûmat arzettiğimi ümit ederim ."	-56.20
[2] "muhterem efendiler , bu **hali** hep beraber mütalea etmeye medar olacak kadar malûmat arzettiğimi ümit ederim .."	-55.86
[3] "muhterem efendiler , bu **konumu** hep beraber mütalea etmeye medar olacak kadar malûmat arzettiğimi ümit ederim ."	-54.97
[4] "muhterem efendiler , bu **durumu** hep beraber mütalea etmeye medar olacak kadar malûmat arzettiğimi ümit ederim ."	-52.11*

Örnekte "vaziyeti" kelimesi için "hali", "konumu" ve "durumu" olmak üzere 3 farklı sadeleştirme adayı oluşturulmuştur. Adaylar kaynak cümledeki orjinal kelimeyle değiştirilerek yeni cümleler oluşturulmuş; bu cümleler üzerinde dil modeli sorgusu gerçekleştirilerek en yüksek başarı skoruna sahip olan cümlenin seçilmesi sağlanmıştır. Değerlendirme sonucunda 4 numaralı cümledeki "durumu" kelimesi en başarılı kelime olarak belirlenmiştir.

4.3. Sonuç

Bu bölümde geliştirilen kural tabanlı sadeleştirme modeli incelenmiştir. Geliştirilen modelde sadeleştirmeler oluşturulan sözlük üzerinden gerçekleştirilmektedir. Yapılan sadeleştirmeler kelime bazında yeterli olsa da, tümce yapıları ve sözdizimsel olarak karmaşık olan cümleler için yeterli olmamaktadır. Bu amaçla; cümle tabanlı olarak hizalanmış paralel bir veri kümesi üzerinden sadeleştirmelerin öğrenildiği istatistiksel sadeleştirme modeli oluşturmuştur.

5. İSTATİSTİKSEL SADELEŞTİRME MODELİ

Bu bölümde uygulanan istatistiksel sadeleştirme modeli incelenmiştir. Geliştirilen modelde sadeleştirmeler kullanılan paralel veri kümesi üzerinden otomatik olarak öğrenilmektedir. Modelde ilk olarak veri kümesindeki cümlelerde kelime tabanlı eşleştirme işlemi gerçekleştirilmektedir. Daha sonra Moses [11] aracı kullanılarak istatistiksel makine çeviri modeli eğitilmiştir. Aynı zamanda, farklı kelime eşleştirme araçlarının, farklı dil modeli veri kümelerinin ve faktörlü makine çevirisi yönteminin sistem başarısı üzerinde etkisi test edilmiştir.

5.1. İstatistiksel Sistem

İstatistiksel sistemde; bir dilin diğer bir dile çevrilmesi amacıyla kullanılan istatistiksel makine çevirisi yöntemi kullanılmaktadır. Kural tabanlı modelden farklı olarak, sadeleştirmeler cümle tabanlı olarak hizalanmış paralel veri kümesi üzerinden öğrenilmektedir. Bunun sonucu olarak; sadeleştirmelerin kalitesi kullanılan paralel veri kümesinin kalitesine ve büyüklüğüne bağımlıdır.

Sistemde; ilk olarak, kaynak cümle daha küçük parçalara bölünmektedir. Daha sonra, her parçanın hedef dile çevrilmesi sağlanmaktadır. En son olarak ise, çevrilen parçalar yeniden düzenlenmektedir. Ayrıca, doğru kelime seçimi ve dilde akıcılığı sağlamak için dil modeli kullanılmaktadır.

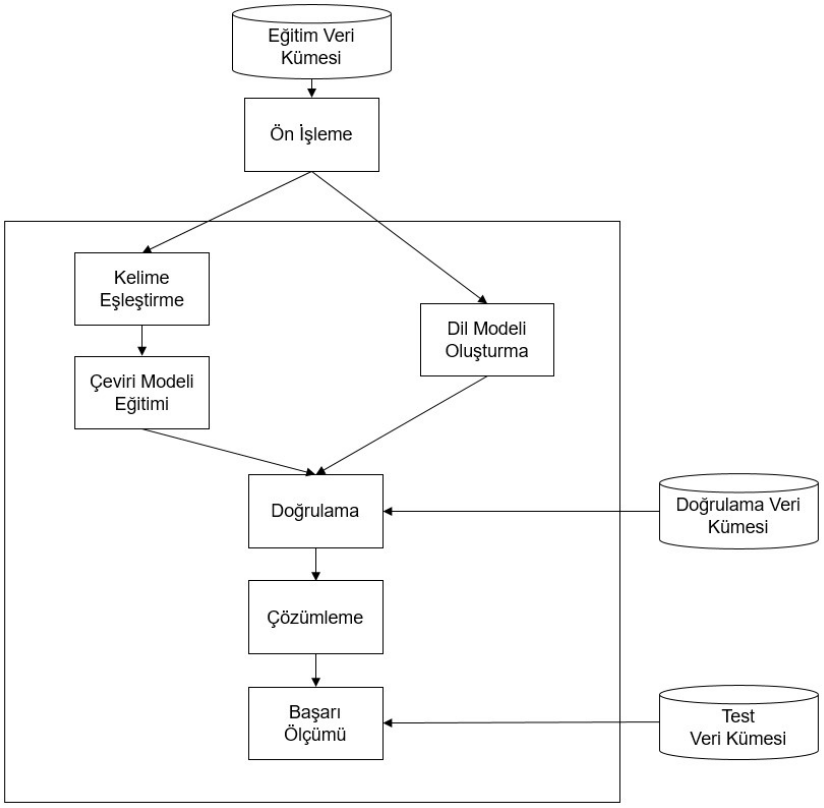
Ayrıca, modelde faktörlü makine çevirisi yöntemi kullanılarak sözdizimsel ve biçimbilimsel özelliklerinde sadeleştirme işleminde kullanılması sağlanmaktadır. Her kelime, yüzeysel biçimdeki basit bir kelime yerine kelime kökü, kelime türü işareti ve biçimbilimsel eklerden oluşan bir vektör şeklinde ifade edilmektedir.

5.2. Uygulanan İstatistiksel Sadeleştirme Modeli

İstatistiksel sadeleştirme modelinde Bayes modeli kullanılmaktadır. Kaynak k cümlesinin, s sadeleştirilmiş cümlesine çevrilme ihtimali Denklem 5'te ifade edilmektedir. Denklemde sadeleştirmenin yeterliliği $P(k|s)$ ve akıcılığı $P(s)$ birbirinden ayrılmaktadır. Sadeleştirmelerin yeterliliği için kelime eşleştirmeleri, akıcılığı için dil modeli kullanılmaktadır.

$$arg \max_{s \in s^*} P(s|k) = arg \max_{s \in s^*} P(k|s)\ P(s) \tag{5}$$

Oluşturulan paralel veri kümesi üzerindeki cümle çiftleri ile Moses [11] aracı kullanılarak istatistiksel makine çevirisi modeli eğitilmiştir. Kullanılan model cümle tabanlı olmakla birlikte ön işleme, kelime eşleştirme, çeviri modeli eğitimi, dil modeli oluşturma, doğrulama, çözümleme ve başarı ölçümü olmak üzere farklı işlemlerden oluşmaktadır. Uygulanan işlem serisi Şekil 5.1 üzerinde gösterilmiştir.



Şekil 5.1. İstatistiksel sadeleştirme modeli

Modelde, ilk olarak, eğitim veri kümesi üzerinde ön işleme adımı gerçekleştirilerek paralel cümle çiftleri modelin diğer işlemlerinde kullanılmak üzere uygun hale getirilmektedir. Daha sonra, cümleler kelime tabanlı olarak eşleştirilmekte ve kelimelerin çeviri olasılıkları hesaplanmaktadır. Çeviri modeli eğitimi işleminde çeviri tablosu, kelime grubu çeviri tablosu ve sıralama modeli tablosu; dil modeli oluşturma işleminde modelde kullanılacak dil modeli oluşturulmaktadır. Doğrulama işleminde modelde kullanılacak parametrelerin düzenlenmesi sağlanmakta, çözümleme işleminde test veri kümesi üzerinde sadeleştirme işlemi gerçekleştirilmektedir. Tüm işlemler tamamlandıktan sonra ise sistemin başarısı ölçülmektedir.

5.2.1. Ön İşleme

Bu işlemde, kaynak cümle sistemin geri kalanında kullanılmak üzere hazırlanmaktadır. Veri kümesinden istenmeyen karakterler temizlenmekte, metin küçük harfe dönüştürülmekte ve cümleler kelimelere bölünmektedir.

5.2.2. Kelime Eşleştirme

Kelime eşleştirme işleminde; cümle çiftlerindeki kelimeler arasındaki $p(s|k)$ çeviri olasılıkları hesaplanmaktadır. Hesaplanan çeviri olasılıkları çeviri modeli eğitimi işleminde kullanılmaktadır. Bu yüzden, kelime eşleştirme işleminde sağlanan doğruluğun yüksek olması doğrudan olarak veri kümesi üzerinden öğrenilen çevirilerin doğru olmasını sağlamaktadır. Kelime eşleştirme problemi Şekil 5.2 üzerinde ifade edilmektedir.

Şekil 5.2. Kelime eşleştirme işlemi

Kelime eşleştirme işlemi için; Brown ve arkadaşları [17] çalışmalarında IBM kelime eşleştirme modellerini önermişlerdir. Tez çalışmasında, IBM 1-2-3 ve 4'ün kelime eşleştirme işleminde kullanıldığı deneyler gerçekleştirilmiştir. IBM Modellerinde kayıp veri içeren problemlerde kullanılan Expectation Maximization (EM) algoritması [71] kullanılmaktadır. Bu algoritmada kullanılan yöntem Denklem 6 üzerinde verilmiştir.

$$p(s|k,\theta) = \sum_{a} p(s,a|k,\theta) \rightarrow max_{\theta} \tag{6}$$

Denklemde s sadeleştirilmiş metni, k kaynak metni, a ise gizli değişken olarak kelime eşleştirmelerini ifade etmektedir. Veri kümesi üzerinden gözlemlenebilir s ve k değişkenleri kullanılarak, denklemin başarısını en yüksek duruma getiren kelime eşleştirmeleri bulunmaya çalışılmaktadır. Algoritma tahmin (expection) ve iyileştirme (maximization) olmak üzere iki adımdan oluşmaktadır [71]. Tahmin

adımında farklı eşleştirme olasılıkları tahmin edilmekte, iyileştirme adımında eşleştirme parametrelerinin güncellenmesi sağlanmaktadır. Günümüze kadar IBM modellerin geliştirilerek sunulduğu ya da farklı yöntemlerle kelime eşleştirme işleminin gerçekleştirildiği farklı çalışmalar da yapılmıştır [72], [73].

IBM Modelleri

IBM Model 1'de eşleştirme işlemi için yalnızca tekil kelimelerin olasılıkları kullanılmaktadır. Cümle çiftlerinde gerçekleştirilen tüm eşleşmelerin eşit bir olasılığa sahip olduğu varsayılmaktadır. Bu durum yanlış kelime eşleştirmelerinin üretilebilmesine neden olabilmektedir. Denklem 7 üzerinde IBM Model 1'de kullanılan yöntem verilmiştir.

$$p(s,a|k) = p(J|k) \prod_{j=i}^{J} p(a_j)\, p(s_j|a_j,k) \tag{7}$$

$p(J|k)$ ifadesi ile ilk olarak sadeleştirmiş cümlenin uzunluğu belirlenmektedir. Daha sonra her bir j kelimesi için; $p(a_j)$ kelime eşleştirme ve $p(s_j|a_j,k)$ kelime eşleştirme olasılığı verildiğinde s_j sadeleştirilmiş kelimesinin üretilme olasılığı kullanılarak çevrim gerçekleştirilmektedir.

Model 1 tarafından gerçekleştirilen bazı eşleştirmelerde sorun olması; kelime konumlarının eşleştirme işleminde kullanılmamasından kaynaklanmaktadır. IBM Model 2'de kelimelerin cümle içerisindeki konumları ve uzunluk bilgisi de eşleştirme işleminde kullanılmaktadır. IBM model 2'de kullanılan yöntem Denklem 8'de verilmiştir.

$$p(s,a|k) = p(J|k) \prod_{j=i}^{J} p(a_j|j,I,J)\, p(s_j|a_j,k) \tag{8}$$

Denklemde, IBM Model 1'den farklı olarak; j kelimesinin sadeleştirilmiş cümle içerisindeki konumu, I kaynak cümle uzunluğu ve J sadeleştirmiş cümle uzunluğu kelime eşleştirme olasılığının hesaplanmasında kullanılmaktadır. Yine IBM model

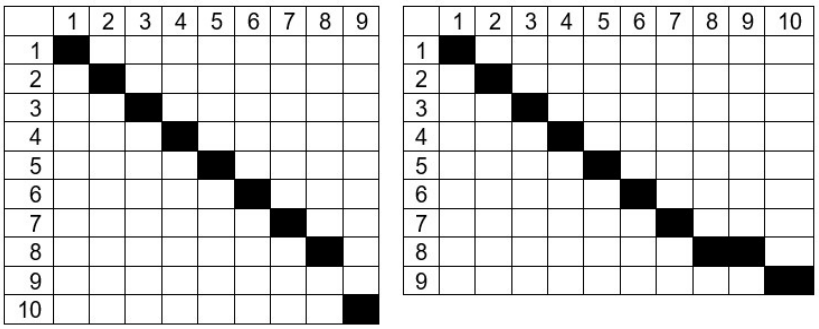
1'deki gibi; $p(a_j|j,I,J)$ ve $p(s_j|a_j,k)$ kullanılarak sadeleştirilmiş s_j kelimesi belirlenmektedir.

IBM Model 3'te, farklı sayıdaki kelime gruplarının eşleştirilmesine odaklanılmıştır. Modelde kaynak kelimenin kaç adet hedef kelimeyle eşleştirileceğini belirten φ parametresi bulunmaktadır [74]. Bu sayede, farklı sayılardaki kelimelerin eşleştirilebilmesi, cümleden kelime silinebilmesi ya da cümleye yeni kelimeler eklenebilmesi sağlanabilmektedir.

IBM Model 4'te kelime sınıflarının eşleştirme işleminde kullanılması sağlanmıştır. Eşleştirme işleminde daha önceden eşleştirmiş kelimeler ve bu kelimeleri çevreleyen kelimelerin sınıfları da kullanılabilmektedir. Bu sayede sıralama değişiklerinin eşleştirmelerde kapsamlı olarak kullanılabilmesi mümkün olmuştur.

Kelime Eşleştirmelerinin Birleştirilmesi

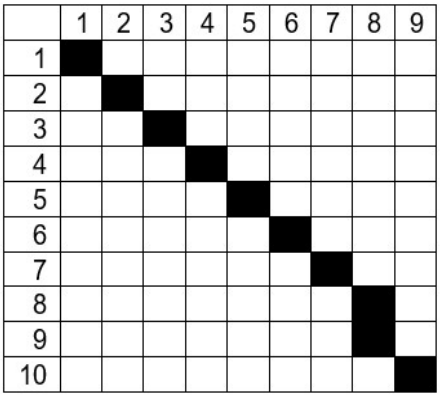
Şekil 5.2'de verilen örnek cümle üzerinde gerçekleştirilen kelime eşleştirme işlemi sonuçları Şekil 5.3 üzerinde verilmiştir. Çizelgelerde satır ve sütunlardaki sayılar kelimelerin cümle içerisindeki konumlarını ifade etmektedir.



Şekil 5.3. Kaynak-Hedef ve Hedef-Kaynak kelime eşleştirme sonuçları

Kelime eşleştirme işlemi için genellikle kaynak cümledeki kelimelerin hedef cümledeki kelimelerle eşleşme olasılıkları ve hedef cümledeki kelimelerin kaynak cümlelerdeki kelimelerle eşleşme olasılıkları ayrı ayrı hesaplanmakta, daha sonra bu eşleştirmelerin birleştirilmesi sağlanmaktadır. Birleştirme işlemi için iki eşleştirmenin kesişimi ya da birleşimi kullanılabilmektedir. İstatistiksel sadeleştirme modelinde; Moses aracında sunulan "grow-diag-final" yöntemi kullanılmıştır. Bu yöntemde ilk olarak tüm kelimeler için kesişim işlemi uygulanmaktadır.

Eşleştirilemeyen kelimelerde ise; bu kelimelerin yalnızca bir birim uzağındaki eşleştirmeler ile birleşim işlemi uygulanmaktadır. Şekil 5.5'te birleştirilmiş kelime eşleştirme sonucu verilmiştir.



Şekil 5.4. Birleştirilmiş kelime eşleştirme sonucu

İstatistiksel modelde, farklı yöntemleri kullanan farklı kelime eşleştirme araçlarının sistem başarısı üzerine etkisi test edilmiştir.

5.2.3. Dil Modeli Oluşturma

Dil modeli, makine çevirisi uygulamalarında üretilen çıktının akıcılığını arttırmak için kullanılmaktadır. Ayrıca birden fazla çeviri olasılığı bulunan kelimeler için en yüksek başarı skoruna sahip olan hedef kelimenin seçilmesi sağlamaktadır. Bu yüzden dil modelinin oluşturulmasında kullanılan dil modeli veri kümesinin kalitesi ve büyüklüğü uygulanan sadeleştirme modeli için büyük önem taşımaktadır. İstatistiksel sadeleştirme modelinde, Nutuk veri kümesinin eğitim için ayrılan kısmı dil modeli veri kümesi olarak kullanılmıştır. Ayrıca farklı bir dil modeli veri kümesinin sistem başarısı üzerine etkisi test edilmiştir.

5.2.4. Çeviri Modeli Eğitimi

Çeviri modeli eğitim işleminde, modelde kullanılacak olan kelime çeviri tablosu, kelime grubu çeviri tablosu ve sıralama modeli tablosu oluşturulmaktadır. Kelime çeviri tablosunda tekil kelimelerin birbirleri arasındaki çeviri olasılıkları, kelime grubu çeviri tablosunda kelime ve kelime gruplarının birbirleri arasındaki çeviri olasılıkları, sıralama modeli tablosunda ise kelimelerin çevrilirken yer değiştirme olasılıkları bulunmaktadır.

5.2.5. Çözümleme

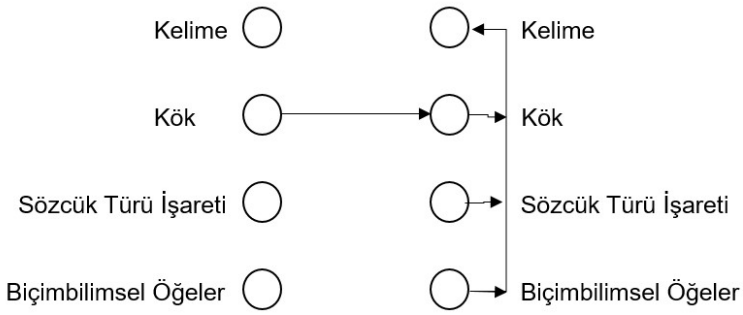
Verilen bir cümlenin hedef tarafta birden fazla çeviri ihtimali mevcuttur. Bu işlemde verilen bir cümle kelime ya da kelime gruplarına bölünmekte, daha sonra bu kelime ve kelime gruplarının çevirisi yapılarak aday çeviri cümleleri oluşturulmaktadır. En son olarak aday cümlelerden en yüksek skora sahip olan cümle çeviri olarak seçilmektedir. Çözümleme yapılırken cümleden kelimeler silinebilmekte, kaynak tarafındaki tek bir kelime hedef tarafta birden fazla kelime ile ifade edilebilmekte ya da çeviride kelimelerin sırası değiştirilebilmektedir.

5.3. Faktörlü Makine Çevirisi

Oluşturulan istatistiksel modelde çeviriler kelimelerin yüzeysel biçimleri kullanılarak gerçekleştirilmektedir. Bu durum, yapım ve çekim almış kelimelerin türedikleri kelimelerden tamamen farklı kelimeler olarak değerlendirilmesine neden olmaktadır. Çevirinin yalnızca kelimelerin yüzeysel biçimleri kullanılarak yapılması İngilizce gibi biçimbilimsel olarak zengin olmayan dillerde önemli sorunlara yol açmasa da; Türkçe, Almanca ya da Fince gibi biçimbilimsel olarak zengin dillerin çevirisinde çeviri kalitesinin düşmesine neden olmaktadır. Modelde, bu sorunun çözülmesi amacıyla Koehn [39] tarafından geliştirilen faktörlü makine çevirisi yöntemi kullanılmıştır.

Faktörlü makine çevirisi yönteminde, sözdizimsel ve biçimbilimsel özellikler cümle tabanlı sistemler ile bütünleştirilmektedir. Bu sistemlerde, her kelime, yüzeysel biçimdeki basit bir kelime yerine birçok faktörden oluşan bir vektör şekilde ifade edilmektedir. Çevirilerde kelimelerin yüzeysel biçimleri yerine; kelimelerin sözdizimsel, biçimbilimsel ya da anlamsal özellikleri de kullanılabilmektedir.

Moses aracında sunulan faktörlü makine çevirisi yönteminde; kelime eşleştirme işlemi kelimelerin yüzeysel biçimleri yerine kelime kökleri kullanılarak yapılabilmekte ya da kelime sıralaması için kelimelerin sözdizimsel özellikleri kullanılabilmektedir. Aynı zamanda, kelimelerin kelime türü işareti ve biçimbilimsel ekleri hedef kelimelerin oluşturulma aşamasında kullanılabilmektedir. Koehn [39] tarafından gerçekleştirilen çalışmada sunulan örnek bir faktörlü çeviri modeli Şekil 5.5 üzerinde verilmiştir.



Şekil 5.5. Örnek bir faktörlü çeviri modeli

Modelde, kelime köklerinin çevrilmesi için ayrı bir model, kelime türü işareti ve biçimbilimsel özelliklerinin çevrilmesi için ayrı bir model oluşturulmaktadır. Daha sonra, bu iki model bir üretim modeli kullanılarak birleştirilmekte ve hedef dildeki kelimeler oluşturulmaktadır. Bu model kullanılarak, örnek bir kelimenin Almancadan İngilizceye çevrilmesi Çizelge 5.1 verilmiştir.

Çizelge 5.1. Faktörlü model kullanılarak gerçekleştirilen bir çeviri örneği

Kaynak Faktör	hauser|haus|NN|plural-nominative-neutral
Çeviri 1: Köklerin çevrilmesi	haus → house, home, building, shell
Çeviri 2: Biçimbilimsel özelliklerin çevrilmesi	NN|plural-nominative-neutral → NN|plural, NN|singular
Üretim: Hedef dildeki kelimelerin üretilmesi	house|NN|plural → houses house|NN|singular → house home|NN|plural → homes

5.3.1. Faktörlü Makine Çevirisinde Uygulanan İşlemler

İstatistiksel modelde uygulanan faktörlü çeviri yönteminde ilk olarak paralel veri kümesi üzerinde veri hazırlama işlemi gerçekleştirilmektedir. Bu aşamada; kelimeler üzerinde biçimbilimsel analiz işlemi gerçekleştirilmekte ve kelimelere; kelime kökü, kelime türü işareti ve biçimbilimsel ekleri gibi özellikler eklenmektedir. Gerçekleştirilen örnek bir veri hazırlama işlemi örneği Çizelge 5.2 üzerinde verilmiştir. Veri hazırlama işlemi için tez çalışmasının diğer bölümlerinde kullanılan TRMorph aracı kullanılmıştır. Ayrıca, eski Türkçe ve günümüz Türkçesi için aynı biçimbilimsel kurallar kullanılmıştır.

Çizelge 5.2. Cümle üzerinde gerçekleştirilen veri hazırlama işlemi

Cümle	“vaziyet ve manzarai umumiye : osmanlı devletinin dahil bulunduğu grup , harbi umumîde mağlûp olmuş , osmanlı ordusu her tarafta zedelenmiş , şeraiti ağır , bir mütarekename imzalanmış .”
	“vaziyet|vaziyet|N|N ve|ve|Cnj:coo|Cnj:coo manzarai|UNK|UNK|UNK umumiye|umumiye|N|N :|:|Punc|Punc osmanlı|UNK|UNK|UNK devletinin|devlet|N|N.p2s.gen dahil|dahil|N|N bulunduğu|bulun|V|V.vn:past.N.p3s grup|grup|N|N ,|,|Punc|Punc harbi|harbi|N|N umumîde|umumî|N|N.loc mağlûp|mağlûp|Adj|Adj olmuş|ol|V|V.evid ,|,|Punc|Punc osmanlı|UNK|UNK|UNK ordusu|ordu|N|N.p3s her|her|Det:def|Det:def tarafta|taraf|N|N.loc zedelenmiş|zedele|V|V.pass.evid ,|,|Punc|Punc şeraiti|şerait|N|N.p3s ağır|ağır|V|V ,|,|Punc|Punc bir|bir|Num|Num mütarekename|UNK|UNK|UNK imzalanmış|imza|N|N.lan.V.evid .|.|Punc|Punc”
Hedef Cümle	“genel durum ve görünüm : osmanlı devletinin içinde bulunduğu grup , genel savaşta yenilmiş , osmanlı ordusu her tarafta zedelenmiş , şartları ağır bir ateşkes anlaşması imzalanmış .”
	“genel|genel|Adj|Adj durum|durum|N|N ve|ve|Cnj:coo|Cnj:coo görünüm|görünüm|N|N :|:|Punc|Punc osmanlı|UNK|UNK|UNK devletinin|devlet|N|N.p2s.gen içinde|iç|N|N.p2s.loc bulunduğu|bulun|V|V.vn:past.N.p3s grup|grup|N|N ,|,|Punc|Punc genel|genel|Adj|Adj savaşta|savaş|N|N.loc yenilmiş|ye|V|V.pass.evid ,|,|Punc|Punc osmanlı|UNK|UNK|UNK ordusu|ordu|N|N.p3s her|her|Det:def|Det:def tarafta|taraf|N|N.loc zedelenmiş|zedele|V|V.pass.evid ,|,|Punc|Punc şartları|şart|N|N.p3p ağır|ağır|V|V bir|bir|Num|Num ateşkes|ateşkes|N|N anlaşması|anlaşma|N|N.p3s imzalanmış|imza|N|N.lan.V.evid .|.|Punc|Punc”

Gerçekleştirilen veri hazırlama işleminde, örneğin; “imzalanmış” kelimesi “imzalanmış|imza|N|N.lan.V.evid” ifadesine dönüştürülmüştür. “imza” kelime kökünü, “N” kelime türü işaretini, “N.lan.V.evid” ise kelimeye eklenen biçimbilimsel ekleri ifade etmektedir. Biçimbilimsel analiz sonucunda biçimbilimsel özellikleri belirlenemeyen kelimeler için tüm faktörler “UNK” olarak ifade edilmektedir.

Veri hazırlama işlemi gerçekleştirildikten sonra; çeviri ve üretim olmak üzere, iki aşamada, çeviri işlemi gerçekleştirilmektedir. Çeviri aşamasında kaynak faktörlerin hedef faktörlere çevrilmesi sağlanmaktadır. Çeviri işleminde kelime kökü, kelime, kelime türü işareti ya da biçimbilimsel ekler kullanılabilmektedir. Ayrıca bir ya da birden fazla faktörün çevrimi birlikte gerçekleştirilebilmektedir. Üretim aşamasında ise çevrilen faktörler kullanılarak hedef dilindeki kelimeler oluşturulmaktadır. Çalışmada çeviri ve üretim aşamalarında farklı faktörlerin kullanıldığı deneyler gerçekleştirilmiştir.

6. DENEYSEL ÇALIŞMALAR

Sadeleştirme sonuçlarının değerlendirilmesinde makine çevirisi sistemlerinin başarı ölçümünde kullanılan BLEU metriği kullanılmıştır. Ayrıca, deneylerde çapraz doğrulama yöntemi uygulanmıştır.

Bu bölümde ilk olarak BLEU metriği ve uygulanan çapraz doğrulama yöntemi açıklanmıştır. Daha sonra kural tabanlı sadeleştirme modeli, istatistiksel sadeleştirme modeli ve hibrid sadeleştirme modeli deney sonuçlarına yer verilmiştir.

6.1. BLEU Metriği

Metin sadeleştirme çalışmalarında başarıyı ölçmek için birçok farklı yöntem geliştirmiştir. Başarı ölçümündeki temel ölçüt insanlar tarafından yapılan sadeleştirmeler ve makineler tarafından yapılan sadeleştirmeler arasında mümkün olduğunca yüksek bir benzerlik ortaya çıkarmaktır.

Bu amaçla farklı metrikler geliştirilmiştir. Zhu ve arkadaşları [12] 2002 yılında makine çevirisi sonuçlarını ölçmek için ucuz ve dilden bağımsız olarak çalışabilen BLEU metriğini geliştirmişlerdir. İlerleyen yıllarda bu metrik makine çevirisi sonuçlarının değerlendirilmesinde standart haline gelmiştir.

Metrikte, sistem çıktısı aday cümleler ve insanlar tarafından yapılan sadeleştirmeler karşılaştırılmaktadır. Metrik N-Gram tabanlı olmakla birlikte, ucuz ve dilden bağımsızdır. Standart BLEU metriği Denklem 9'daki gibidir.

$$BLEU = BP \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{9}$$

Metriğin çıktısı 0 ve 1 arasında değişmektedir. Çalışmada sonuçlar 100 ile çarpılıp sunulmaktadır. BP indirgeme parametresini, N kullanılan N-Gram mertebesini, w_n düzenlenmiş hassasiyet skoru [12] olarak adlandırılan p_n'in ağırlığını ifade etmektedir. Çalışmada 4-Gram eşleştirmeye kadar hesaplama yapılmakla birlikte, w_n değeri $\frac{1}{N}$ olarak kullanılmaktadır. p_n'in formülü Denklem 10 üzerinde verilmiştir.

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(ngram)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram' \in C'} Count(ngram')} \tag{10}$$

Denklemde $Count_{clip}(ngram)$ aday sadeleştirme ve referans metin arasındaki ortak N-Gram sayısını, $Count(ngram')$ ise aday sadeleştirmedeki toplam N-Gram sayısını ifade etmektedir. Aday sadeleştirmenin kısa ve referans metinle benzer olması durumunda, p_n yanıltıcı bir yükseklik sunmaktadır. Bu yüzden BP parametresi kullanılarak farklı uzunluktaki cümlelerin belirli bir oranda indirgenmesi sağlanmaktadır. BP ceza parametresinin hesaplanmasında kullanılan yöntem Denklem 11'deki gibidir.

$$BP = \begin{cases} 1, & if\ c > r \\ e^{(1-\frac{r}{c})} & if\ c \leq r \end{cases} \tag{11}$$

Denklemde r referans metnin uzunluğunu, c ise aday metnin uzunluğunu ifade etmektedir.

6.2. Çapraz Doğrulama

Makine öğrenmesi uygulamalarında eğitim, doğrulama ve test veri kümelerinin büyüklükleri sonuçların tutarlı olarak elde edilebilmesi için büyük önem arz etmektedir. Test veri kümesi boyutunun küçük olduğu durumda, farklı test örnekleri ile farklı sonuçlar elde edilebilmektedir. Aynı şekilde, eğitim ve doğrulama veri kümeleri boyutunun küçük olduğu durumda, farklı eğitim örnekleriyle farklı parametreler öğrenilebilmektedir.

Makine öğrenmesi uygulamalarındaki bu problemi en aza indirmek amacıyla çapraz doğrulama yöntemi kullanılmaktadır. Bu yöntemde veri kümesi belirli sayıdaki eşit parçalara bölünmektedir. Daha sonra bölünen parçalar farklı deneylerde test edilmekte ve bu test sonuçlarının ortalaması alınmaktadır. Kohavi [75] çalışmasında K parametresinin 10 olarak seçilmesinin daha iyi sonuç verdiğini belirtmiştir. Çalışmada paralel veri kümesi eşit büyüklükteki 10 parçaya bölünmüş ve bölünen parçalardan her biri farklı deneylerde test edilmiştir.

6.3. Kural Tabanlı Sadeleştirme Modeli

Geliştirilen kural tabanlı sadeleştirme modelinde ilk olarak eski kelimeler tespit edilmekte, daha sonra bu kelimelerin günümüz Türkçesine sadeleştirilmesi gerçekleştirilmektedir. Modelde, eski kelimelerin tespit edilmesi için sınıflandırma

işlemi uygulanmaktadır. Bu bölümde ilk olarak sınıflandırma işleminin; daha sonra, tüm modelin değerlendirme sonuçlarına yer verilmiştir.

6.3.1. Sınıflandırma İşlemi

Sınıflandırma işlemi farklı eşik değerleri kullanılarak değerlendirilmiştir. Nutuk veri kümesinin referanslarında değiştirilen kelimeler doğru örnek olarak kabul edilmiştir. Çizelge 6.1 üzerinde farklı eşik değerleri kullanılarak elde edilen duyarlılık, hassasiyet ve F1 skoru değerleri verilmiştir.

Çizelge 6.1. Farklı eşik değerleri için yapılan değerlendirme sonuçları

Kelime Değişim Yüzdesi (%)	Duyarlılık	Hassasiyet	F1 skoru
%10	0.8558	0.9203	0.8863
%20	0.8728	0.9103	0.8912
%30	0.8849	0.8985	0.8917
%40	0.9013	0.8827	0.8919*
%50	0.9119	0.8671	0.8889
%60	0.9156	0.8573	0.8855
%70	0.9193	0.8425	0.8792
%80	0.9226	0.8136	0.8647

Sınıflandırma aşamasında en iyi sonuç; eşik değerinin %40 olduğu deneyde elde edilmiştir. Bu yüzden gerçekleştirilen diğer deneylerde eşik değeri %40 olarak alınmıştır.

6.3.2. Kural Tabanlı Sadeleştirme Modeli Sonuçları

Kural tabanlı sadeleştirme modelinde farklı biçimbilimsel analiz ve sentez işlemlerinin kullanılmasına göre sözcüksel aktarım farklı kapsamlarda gerçekleştirilebilmektedir. Çizelge 6.2'de kural tabanlı sadeleştirme modelinin sonuçlarına yer verilmiştir.

Çizelge 6.2. Kural tabanlı sadeleştirme modeli sonuçları

Aktarım Türü	Var.	S.S.	Ort.
Başlangıç	0.36	0.60	26.07
Doğrudan aktarım	0.66	0.44	29.29
Kaynak kelimenin biçimbilimsel özellikleri kullanılarak aktarım	0.50	0.71	33.51
Kaynak ve hedef kelimenin biçimbilimsel özellikleri kullanılarak aktarım	0.57	0.75	34.36*

Çizelgenin ilk satırında Nutuk [9] ve sadeleştirilmiş Nutuk [10] arasında BLEU metriği kullanılarak ölçülen benzerlik skoru yer verilmiştir. Diğer satırlarda kullanılan farklı sözcüksel aktarım yöntemleriyle elde edilen sonuçlar yer almaktadır. Modelde en iyi sonuç sadeleştirme işleminde kaynak ve hedef kelimenin biçimbilimsel özelliklerinin kullanıldığı deneyde elde edilmiştir. Kural tabanlı model tarafından gerçekleştirilen sadeleştirmelerin sonucunda 26.07 BLEU skorundan 34.36 BLEU skoruna artış sağlanmıştır. Kural tabanlı modelin değerlendirilmesi sonucunda elde edilen diğer sonuçlar Çizelge 6.3'de verilmiştir.

Çizelge 6.3. Kural tabanlı sadeleştirme modelinde elde edilen diğer sonuçlar

Aktarım Türü	**Değiştirilmesi belirlenen kelime sayısı**	**Sözlükte Bulunan Kelime Sayısı**	**Değiştirilen Kelime Sayısı**	**Doğru Olarak Değiştirilen Kelime Sayısı**
Doğrudan aktarım	140,349	49,069	34,079	12,479
Kaynak kelimenin biçimbilimsel özellikleri kullanılarak aktarım	140,349	49,069	57,764	23,187
Kaynak ve hedef kelimenin biçimbilimsel özellikleri kullanılarak aktarım	140,349	49,069	65,326	25,854*

Modelde değiştirilen kelime sayısı ve doğru olarak değiştirilen kelime sayısı farklı biçimbilimsel analiz ve sentez işlemlerinin kullanılmasına göre değişim göstermektedir. En iyi sonuç; kaynak ve hedef kelimenin biçimbilimsel özelliklerinin kullanıldığı deneyde elde edilmiştir. Ayrıca sınıflandırma işlemi sonucunda toplam 273.604 kelimeden 140,349 kelimenin değiştirilmesi gerektiğine karar verilmiştir. Bu kelimelerden 49,069'i direk olarak sözlük üzerinden bulunabilmektedir. Dil modeli sorgusu işlemi tarafından bu kelimelerden bazılarının değiştirilmemesi gerektiği belirlenmiştir. Sadeleştirmelerde biçimbilimsel özelliklerin kullanılmasıyla, değiştirilen kelime sayısının sözlükte bulunan kelime sayısını geçtiği görülmüştür. Bu durum biçimbilimsel analiz işlemi gerçekleştirilmeden sözlük üzerinden direk olarak çoğu kelimenin bulunamamasından kaynaklanmaktadır.

6.4. İstatistiksel Sadeleştirme Modeli

İstatistiksel modelde; veri kümesi büyüklüğünün, farklı kelime eşleştirme araçlarının, farklı dil modeli veri kümelerinin ve faktörlü makine çevirisi yönteminin sistem başarısı üzerinde etkisi değerlendirilmiştir. Yapılan deneylerde çözümleme

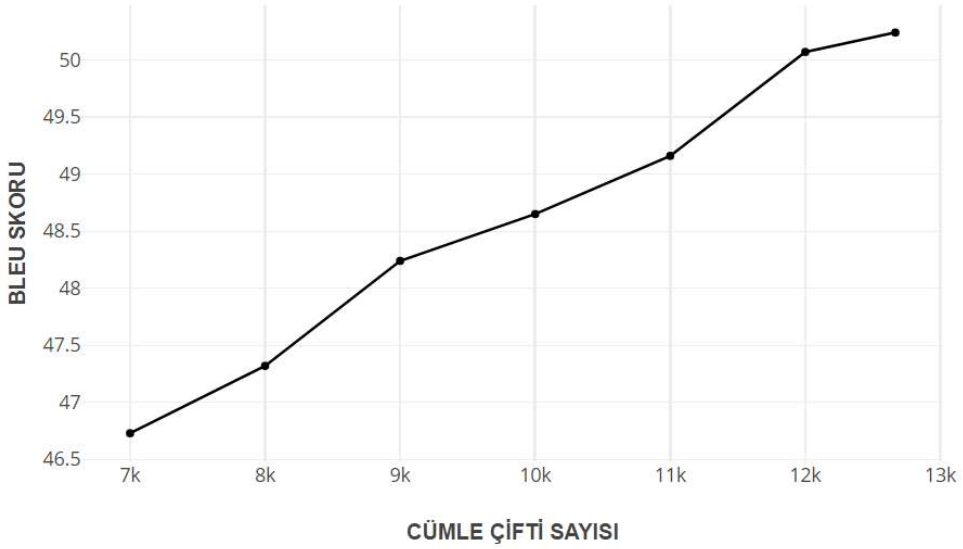
adımında üretilecek cümle sayısı 100 olarak belirlenmiştir. Ayrıca, her deneyde eğitim için ayrılan parçalardan biri doğrulama veri kümesi olarak kullanılmıştır.

6.4.1. Doğrulama İşlemi

İstatistiksel modelde; doğrulama veri kümesi kullanılarak, çeviri modelinde kullanılacak olan parametreler başarıyı yükseltecek şekilde düzenlenmektedir. Kelime çeviri tablosu katkısı, kelime grubu çeviri tablosu katkısı, dil modeli katkısı, kelime sırası değişim olasılığı, yeni kelime ekleme veya silme olasılığı gibi parametreler eğitim veri kümesinden ayrı bir veri kümesi üzerinden öğrenilmektedir.

6.4.2. Veri Kümesinin Büyüklüğünün Etkisi

Veri kümesi büyüklüğünün sonuçlara etkisini değerlendirmek üzere; doğrulama ve test veri kümelerinin boyutu sabit olarak tutulmuş, eğitim veri kümesinin boyutu ise dereceli olarak arttırılmıştır. Şekil 6.1 üzerinde farklı büyüklükteki eğitim veri kümelerinin kullanıldığı testlerin sonuçlarına yer verilmiştir.



Şekil 6.1. Farklı eğitim veri kümesi büyüklüğünün sonuçlara etkisi

Veri kümesi büyüklüğünün arttırılmasının sonuçları önemli ölçüde arttırdığı görülmüştür. Kullanılan en büyük eğitim veri kümesi 12.668 cümleden oluşmaktadır.

6.4.3. Kelime Araçlarının Karşılaştırılması

Farklı kelime eşleştirme araçlarının sistem başarısı üzerine etkisinin olup olmadığı bağımlı örneklem t-testi ile test edilmiştir. Kelime eşleştirme araçları olarak Giza++

[74], Berkeley Aligner [72] ve Fast Align [73] araçları karşılaştırılmıştır. Çizelge 6.4 üzerinde bu araçlar kullanılarak elde edilen ortalama BLEU skorları gösterilmiştir.

Çizelge 6.4. Farklı kelime eşleştirme araçları için t-testi sonuçları

Araç Çifti	n	Ortalama	Var.	S.S.	df	t	P (iki-uçlu)
Berk. Aligner	10	52.24*	0.68	0.83	9	23.15	2.48×10^{-9}
Giza++	10	51.19	0.78	0.88			
Berk. Aligner	10	52.24	0.68	0.83	9	21.33	5.13×10^{-9}
Fast Align	10	51.00	0.71	0.84			
Giza++	10	51.19	0.78	0.88	9	2.75	0.02
Fast Align	10	51.00	0.71	0.84			

Testlerde α değeri 0.05 olarak alınmıştır. Ayrıca, iki-uçlu t-kritik değeri tüm karşılaştırmalarda yaklaşık olarak 2.26'dır. Kelime eşleştirme araçlarından Berkeley Aligner ile en yüksek başarı skoru elde edilmiştir. Berkeley aracı ve diğer kelime eşleştirme araçlarının başarı skoru arasında önemli ölçüde bir fark bulunmakta, diğer araçlar arasında ise önemli derecede bir fark bulunmamaktadır.

Giza++ ve Fast Align araçlarında kaynak-hedef ve hedef-kaynak eşleştirmeleri için eğitilen asimetrik modellerin daha sonradan kesişimi alınmaktadır. Berkeley Aligner aracında ise bu iki model veri olabilirliğini arttıracak şekilde birlikte eğitilmektedir. Bu durum kullanılan veri kümesi üzerinde başarılı sonuçlar vermektedir.

6.4.4. Farklı Dil Modellerinin Karşılaştırılması

Farklı kaynaklardan kullanılarak oluşturulan dil modellerinin sistem başarısı üzerine etkisi değerlendirilmiştir. İlk deneyde Nutuk kullanılarak eğitilen dil modeli kullanılmış; ikinci deneyde 412.028 cümleden oluşan ve gazete yazıları[1] üzerinden eğitilen dil modeli kullanılmıştır. Ayrıca iki modelinin birlikte kullanıldığı üçüncü bir deney daha gerçekleştirilmiştir. Çizelge 6.5'te gerçekleştirilen deneylerin sonuçlarına yer verilmiştir.

Çizelge 6.5. Farklı dil modellerinin sonuçlara etkisi

	Var.	S.S.	Ort.
Nutuk Dil Modeli	0.68	0.83	52.24
Gazete Yazıları Dil Modeli	0.57	0.75	45.69
İki Dil Modelinin Birlikte Kullanılması	0.76	0.87	52.28*

[1] http://gurmezin.com/modelinizi-nasil-egitirsiniz/

Gazete yazıları kullanılarak eğitilen dil modelinde başarının düştüğü gözlemlenmiştir. Bu durum veri kümesi ve kullanılan dil modelindeki metinlerin farklılık gösterip, sadeleştirmelerin doğru olarak gerçekleştirilememesinden kaynaklanmaktadır. Ayrıca, İki dil modelinin birlikte kullanılması durumunda deney sonuçlarının yalnızca 0.04 değerinde değişiklik gösterdiği tespit edilmiştir. Moses aracında iki dil modelinin ağırlıkları doğrulama işleminde düzenlenmektedir. Bu durum yüksek sonuç alınan dil modelinin çevirilerdeki ağırlığının yüksek olmasına yol açmaktadır.

6.4.5. Faktörlü Sadeleştirme Modeli

İstatistiksel sadeleştirme modelinde, faktörlü makine çevirisi yönteminin sistem başarısı üzerindeki etkisi değerlendirilmiştir. Çeviri ve üretim aşamaları için farklı faktörler kullanılarak deneyler gerçekleştirilmiştir. Bu deneylerin sonuçları Çizelge 6.6'da verilmiştir.

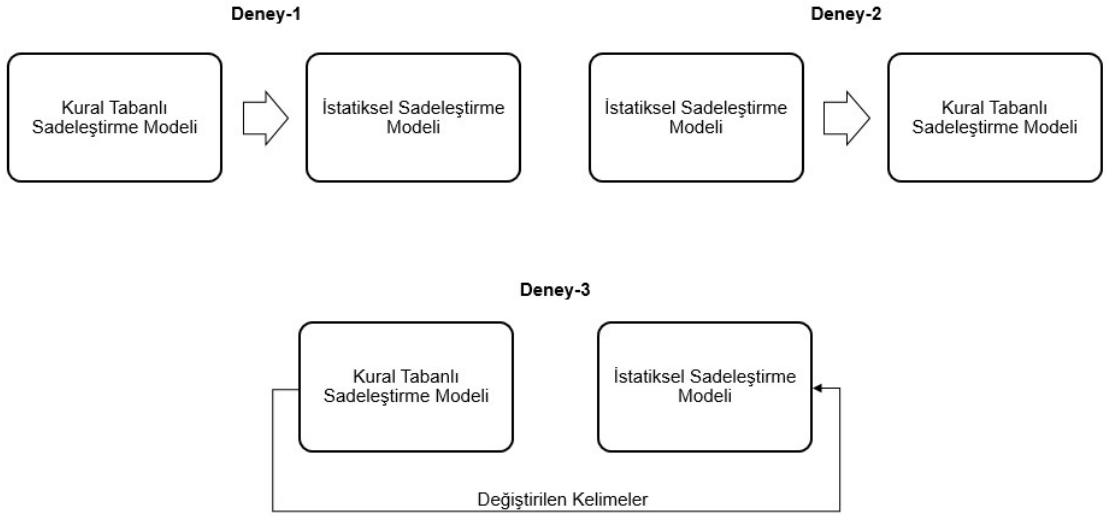
Çizelge 6.6. Faktörlü makine çevirisi sonuçları

Kelime Eşleştirme	Çeviri	Üretim	Var.	S.S.	Ort.
kelime → kelime	kelime → kelime	-	0.68	0.83	52.24
kök → kök	kelime → kelime	-	0.72	0.85	53.12
kök → kök	kelime → kelime + kelime türü işareti	kelime → kelime türü işareti	0.73	0.85	53.11
kök → kök	kelime → kelime + kelime türü işareti (LM)	kelime → kelime türü işareti	0.64	0.80	53.11
kök → kök	kelime → kelime + biçimbilimsel ekler	kelime → biçimbilimsel ekler	0.63	0.79	53.13*

Faktörlü makine çevirisi yönteminde en iyi sonuç kelime eşleştirme işleminin kelimenin köküne göre, çeviri ve üretim işlemlerinin ise biçimbilimsel ekler üzerinden gerçekleştirildiği deneyde elde edilmiştir. Fakat deneylerde elde edilen BLEU skorunda çok fazla bir değişim gözlemlenmemiştir. İstatistiksel modelde biçimbilimsel analiz işlemi kelime tabanlı olarak gerçekleştirilmekte, her kelime için yalnızca tek bir biçimbilimsel özellik belirlenebilmektedir. Aynı zamanda kaynak ve hedef faktörlerin oluşturulmasında yalnızca TRMorph aracı kullanılmakta, her iki taraf için de kelime türü işareti ve biçimbilimsel ek biçimleri aynı kalmaktadır. Bu durum faktörlü çeviri yöntemi kullanılarak çok fazla bir kazanç elde edilememesine neden olmaktadır.

6.5. Hibrid Sadeleştirme Modeli

Kural tabanlı sadeleştirme modeli ve istatistiksel sadeleştirme modeli farklı yöntemler kullanılarak hibrid bir işlem serisinde birleştirilmiştir. Gerçekleştirilen ilk deneyde; ilk olarak kural tabanlı sadeleştirme modeli, daha sonra istatistiksel sadeleştirme modeli uygulanmıştır. İkinci deneyde, bu işlemlerin uygulama sırası değiştirilmiştir. Üçüncü deneyde; kural tabanlı sadeleştirme modeli tarafından gerçekleştirilen sadeleştirmeler, pasif olarak, istatistiksel sadeleştirme modelinin eğitim veri kümesine eklenmiştir. Şekil 6.2 üzerinde gerçekleştirilen deneyler görsel bir şekilde ifade edilmiştir.



Şekil 6.2. Hibrid sadeleştirme modeli için gerçekleştirilen deneyler

Çizelge 6.7'de hibrid yöntemler kullanılarak elde edilen deney sonuçları verilmiştir.

Çizelge 6.7. Hibrid sadeleştirme modeli sonuçları

	Var.	**S.S.**	**Ort.**
Deney-1 (Kural tabanlı – İstatistiksel)	0.80	0.89	52.95
Deney-2 (İstatistiksel – Kural tabanlı)	0.75	0.87	51.09
Deney-3 (Sadeleştirmelerin eklenmesi - 1 defa)	0.86	0.93	54.09*
Deney-3 (Sadeleştirmelerin eklenmesi - 2 defa)	0.82	0.90	54.03
Deney-3 (Sadeleştirmelerin eklenmesi - 3 defa)	0.93	0.96	53.98

En iyi sonuç; kural tabanlı model tarafından gerçekleştirilen sadeleştirmelerin, pasif olarak, istatistiksel modele 1 defa eklendiği deneyde elde edilmiştir.

6.5.1. Eğitim Veri Kümesinde Değişmeyen Kelimelerin Kullanılması

Eski ve güncel Türkçedeki kelimelerin yalnızca bir bölümü değişim göstermektedir. Bu yüzden, değişmeyen bu kelimelerin model tarafından da değiştirilmemesi istenmektedir. Hibrid modelde, eğitim veri kümesindeki değişmeyen kelimeler tespit edilmiş ve bu kelimeler tekrar veri kümesine eklenmiştir. Ayrıca 4-Grama kadar değişmeyen tüm N-Gramların veri kümesine eklendiği ayrı bir deney daha gerçekleştirilmiştir. Deneylerin sonuçları Çizelge 6.8'de verilmiştir.

Çizelge 6.8. Eğitim veri kümesinde değişmeyen kelimelerin kullanılması

	Var.	**S.S.**	**Ort.**
Değişmeyen kelimelerin eklenmesi	0.94	0.97	54.40
Değişmeyen N-Gramların eklenmesi	0.88	0.94	54.03

Veri kümesindeki değişmeyen kelimelerin veri kümesine 1 defa eklenmesiyle başarının arttığı, tüm N-Gramların eklenmesiyle ise başarının azaldığı gözlemlenmiştir.

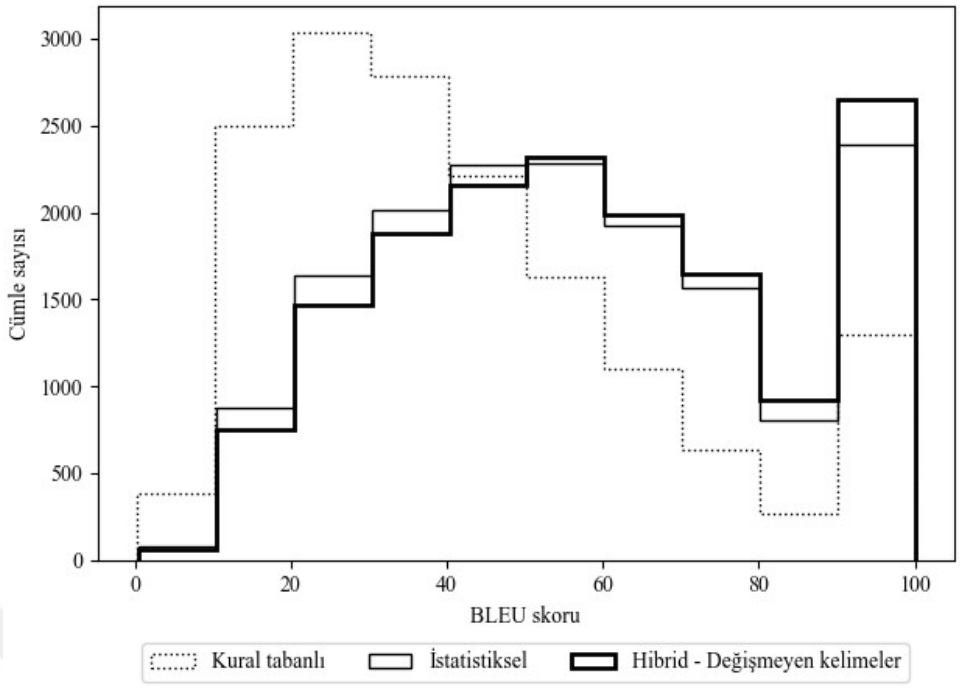
6.6. Sonuç

Gerçekleştirilen tüm deney sonuçları özet olarak Çizelge 6.9 üzerinde verilmiştir.

Çizelge 6.9. Özet olarak deney sonuçları

	Var.	**S.S.**	**Ort.**
Başlangıç	0.36	0.60	26.07
Kural Tabanlı Sadeleştirme Modeli	0.57	0.75	34.36
İstatistiksel Sadeleştirme Modeli	0.68	0.83	52.24
İstatistiksel Sadeleştirme Modeli (Faktörlü)	0.63	0.79	53.13
Hibrid Sadeleştirme Modeli	0.86	0.93	54.09
Hibrid Sadeleştirme Modeli - Değişmeyen kelimeler	0.94	0.97	54.40
Hibrid Sadeleştirme Modeli - Değişmeyen kelimeler (Faktörlü)	0.62	0.79	55.17*

Hibrid sadeleştirme modelinin; değişmeyen kelimelerle zenginleştirilip, istatistiksel model için faktörlü yöntemin kullanılmasıyla çalışmadaki en iyi sonuç elde edilmiştir. Ayrıca kural tabanlı, istatistiksel ve hibrid sadeleştirme modelleri tarafından sadeleştirilmiş cümlelerin referans metinlerle karşılaştırılması ile elde edilen BLEU skorları 10 farklı kümede gruplanmış ve Şekil 6.3 üzerinde histogram olarak sunulmuştur. Histogramı oluşturmak amacıyla; çapraz doğrulama yönteminde gerçekleştirilen deneylerde kullanılan test veri kümelerinin birleştirilmesi sağlanmıştır.



Şekil 6.3. Kural tabanlı, istatistiksel ve hibrid sadeleştirme modelleri ile elde edilen histogram sonuçları

Grafikte yatay eksen BLEU skoru küme aralığını, dikey eksen ise bu kümede bulunan cümle sayısını göstermektedir. Grafikte; hibrid sadeleştirme modeli ile; düşük BLEU skoru elde edilen daha az, yüksek BLEU skoru elde edilen daha fazla cümle üretilebildiği görülmektedir.

7. DEĞERLENDİRME

Bu bölümde yapılan deneylerde elde edilen sonuçlar değerlendirilmiştir. Ayrıca, kural tabanlı ile istatistiksel sadeleştirme modellerinin avantajları ve dezavantajları ele alınmıştır.

7.1. Kural Tabanlı Sadeleştirme Modeli

Kural tabanlı sadeleştirme modelinde kelime tabanlı olarak sadeleştirme işlemi gerçekleştirilmektedir. Çizelge 7.1 üzerinde kural tabanlı sadeleştirme modeli tarafından gerçekleştirilen sadeleştirmeler bulunmaktadır. “K” ile kaynak metin, “S” ile model tarafından gerçekleştirilen sadeleştirme ifade edilmektedir. Cümle çiftlerinin sağında başlangıç ve son durum BLEU skorları yer almaktadır.

Çizelge 7.1. Kural tabanlı sadeleştirme modeli tarafından gerçekleştirilen sadeleştirmeler

Cümle Sadeleştirmeleri	**BLEU Skoru**
K: “1335 senesi mayısının 19 uncu günü samsuna çıktım .” **S:** “1335 senesi mayısının 19 uncu günü samsuna çıktım .”	88 → 88
K: “**vaziyet** ve manzarai umumiye : osmanlı devletinin **dahil** bulunduğu grup , **harbi** umumîde mağlûp olmuş , osmanlı ordusu her tarafta zedelenmiş , **şeraiti** ağır , bir mütarekename imzalanmış .” **S:** “**durum** ve manzarai umumiye : osmanlı devletinin **içinde** bulunduğu grup , **düşman** umumide mağlüp olmuş , osmanlı ordusu her tarafta zedelenmiş , **şartları** ağır , bir mütarekename imzalanmış .”	34 → 53
K: “büyük **harbin** uzun seneleri **zarfında** , millet yorgun ve fakir bir halde .” **S:** “büyük **savaşın** uzun seneleri **içinde** , millet yorgun ve fakir bir halde .”	65 → 100
K: “millet ve memleketi **harbi** umumîye sevkedenler , kendi hayatları **endişesine** düşerek , memleketten **firar** etmişler .” **S:** “millet ve memleketi **düşman** umumiye sevkedenler , kendi hayatları **kaygısına** düşerek , memleketten **kaçmak** etmişler .”	27 → 36
K: “**saltanat** ve hilâfet **mevkiini** işgal eden vahdettin , müstereddi , **şahsını** ve yalnız tahtını **temin** edebileceğini tahayyül ettiği denî **tedbirler** araştırmakta .” **S:** “**kuvvet** ve hilafet **durumunu** işgal eden vahdettin , müstereddi , **kişisini** ve yalnız tahtını **sağlamak** edebileceğini tahayyül ettiği deni **önlemler** araştırmakta .”	15 → 17
K: “damat ferit paşanın **riyasetindeki** kabine ; âciz , haysiyetsiz , cebîn , yalnız padişahın iradesine tâbi ve onunla beraber şahıslarını **vikaye** edebilecek herhangi bir **vaziyete** razı .” **S:** “damat ferit paşanın **başkanlığındaki** kabine ; aciz , haysiyetsiz , cebin , yalnız padişahın iradesine tabi ve onunla beraber şahıslarını **korumak** edebilecek herhangi bir **duruma** razı .”	41 → 50
K: “ordunun elinden esliha ve cephanesi alınmış ve alınmakta ...” **S:** “ordunun elinden esliha ve cephanesi alınmış ve alınmakta ...”	70 → 70
K: “itilâf devletleri , **mütareke** ahkâmına **riayete lüzum** görmüyorlar .” **S:** “itilaf devletleri , **ateşkes** ahkamına **korumaya gerek** görmüyorlar .”	31 → 31
K: “birer **vesile** ile , itilâf donanmaları ve askerleri istanbulda .” **S:** “birer **bahane** ile , itilaf donanmaları ve askerleri istanbulda .”	80 → 80

Geliştirilen kural tabanlı sadeleştirme modelinde birçok kelimenin günümüz Türkçesine doğru olarak sadeleştirilmesi gerçekleştirilebilmektedir. Gerçekleştirilen sadeleştirmelerde cümledeki kelime sıralaması ve cümlenin sözdizimsel yapısı ise değiştirilmemektedir.

Modelde, tekil kelimelerin birden fazla kelime ile sadeleştirilebilmesi doğrudan olarak gerçekleştirilebilmektedir. Tekil bir kelimenin birden fazla kelime ile sadeleştirildiği birkaç örnek Çizelge 7.2 üzerinde verilmiştir.

Çizelge 7.2. Tekil kelimelerin birden fazla kelime ile sadeleştirilmesi

Kaynak Kelime	**Hedef Kelime**
bihakkın	tam olarak
bililtizam	bile bile
binaenaleyh	bunun üzerine
binnetice	sonuç olarak
şiddetle	kesin olarak
tekabül	karşı karşıya
tevfikan	uygun olarak

Geliştirilen modelde; yalnızca tekil kelimelerin sadeleştirilmesinden ve kelime sıralamasının değiştirilmemesinden kaynaklanan sorunlar olduğu gözlemlenmiştir. Bu sadeleştirmelerde harici bir düzeltme işlemi gerekmektedir. Sadeleştirilmesinde sorun tespit edilen birkaç örnek ve bu örneklerin referans sadeleştirmeleri Çizelge 7.3. üzerinde verilmiştir.

Çizelge 7.3. Sadeleştirilmesinde problem tespit edilen birkaç örnek

Kaynak Kelime Grubu	**Sistem Çıktısı**	**Referans Sadeleştirme**
amali millîye	istekleri milliyeyi	ulusal emeller
devleti osmanîye	devleti osmanîye	osmanlı devleti
düveli mütelife	devletleri mütelife	itilâf devletleri
hakimiyeti millîye	egemenliği milliye	ulusal egemenlik
heyeti temsiliye	heyeti temsiliye	temsilci kurul
siyaseti osmanîye	politikası osmaniye	osmanlı politikası

Nutuk veri kümesinde orijinal metinden farklı kelimeler kullanılarak sadeleştirilmiş yüklem yapıları bulunmaktadır. Geliştirilen modelde bu yüklem yapıları kelime tabanlı olarak sadeleştirilebilse bile, metin üzerinde harici bir düzeltme işlemi gerekmektedir. Bu sorunun tespit edildiği sadeleştirme örnekleri Çizelge 7.4 üzerinde verilmiştir.

Çizelge 7.4. Sadeleştirilmesinde problem tespit edilen birkaç örnek

Kaynak Kelime Grubu	Sistem Çıktısı	Referans Sadeleştirme
tebliğ olundu	bildiri olundu	gereği bildirildi
tebliğ ettim	bildiri ettim	bildirdim
intizar eyleriz	beklemek eyleriz	bekleriz
icap etmiştir	gerek etmiştir	gerekli olmuştur

Kural tabanlı sadeleştirme modelinde karşılaşılan bu sorunların giderilmesi ancak; sözdizimsel ve anlamsal özelliklerin sadeleştirmelerde kullanılmasıyla mümkün olabilmektedir.

7.1.1. Biçimbilimsel Analiz ve Sentez İşlemlerinin Etkisi

Türkçenin biçimbilimsel açıdan zengin bir dil olmasından dolayı kelimeler doğrudan olarak sözlük üzerinden bulunamamaktadır. Deneylerde kaynak ve hedef kelimeler üzerinde biçimbilimsel analiz işleminin gerçekleştirilmesinin başarıyı arttırdığı tespit edilmiştir. Biçimbilimsel analiz ve sentez işlemlerinin sonucunda doğru olarak gerçekleştirilen sadeleştirme örnekleri Çizelge 7.5 üzerinde verilmiştir.

Çizelge 7.5. Biçimbilimsel analiz ve sentez sonucunda doğru olarak gerçekleştirilen sadeleştirmeler

Kaynak Kelime Grubu	Sistem Çıktısı	Referans Sadeleştirme
feshini	dağıtmasını	dağıtmasını
hakimiyeti	egemenliği	egemenliği
hukuku	hakları	hakları
icrasına	yapılmasına	yapılmasına
istirahate	dinlenmeye	dinlenmeye
mutalebatına	isteklerine	isteklerine
şeraiti	koşulları	koşulları
tahkikatı	soruşturması	soruşturması
tedabire	önlemlere	önlemlere
vazifesini	görevini	görevini
zarureti	zorunluluğu	zorunluluğu

Ayrıca, yapılan incelemelerde biçimbilimsel analiz ve sentez işlemlerinde dilin yapısından doğan belirsizlikten dolayı, bazı durumlarda, hatalı aday sadeleştirmelerin oluşturulduğu tespit edilmiştir. Biçimbilimsel analiz ve sentez işlemlerinde kullanılan kuralların kalitesinin arttırılmasının bu hataları azaltacağı öngörülmektedir.

7.1.2. Dil Modelinin Etkisi

Modelde sözlükte doğru karşılığı bulunmasına rağmen dil modeli tarafından değiştirilmeyen kelimeler bulunmaktadır. Bu kelimelerde; doğru aday

sadeleştirmeler bulunabilmesine rağmen; bu kelimeler, dil modeli tarafından değiştirilmemiştir. Bu durum sözcüksel aktarım sonucunda belirlenen aday kelimelerin dil modelinde az ya da hiç kullanılmamasından kaynaklanmaktadır. Dil modeli sorgusu sonucunda değiştirilmeyen birkaç örnek Çizelge 7.6 üzerinde verilmiştir.

Çizelge 7.6. Dil modeli tarafından değiştirilmeyen sadeleştirme adayları

Kaynak Kelime	**Sadeleştirme Adayları**
zaferyâb	üstünlük kazanan, zafer kazanan, muzaffer olan
samimiyetine	içtenliğine
tecziyeleri	cezalandırmaları
mukadderatı	mazgısı, mazgıyı, yazgısı, yazgıyı

7.1.3. Kural Tabanlı Sadeleştirme Modelinin Avantajları ve Dezavantajları

Avantajları

- Kural tabanlı sadeleştirme modelinde kalite limiti bulunmamaktadır. Sistem çıktısı analiz edilebilmekte, hatalar düzeltilebilmekte, yeni biçimbilimsel kurallar ve sözlük kayıtları kolaylıkla eklenebilmektedir.
- Yalnızca basit bir sözlük kullanılarak sadeleştirmeler gerçekleştirilebilmektedir. Kural tabanlı model ile BLEU skorunda +8.29'luk bir kazanç sağlanmıştır.
- Çok fazla işlem gücü gerekmemektedir.
- Biçimbilimsel ve sözdizimsel özelliklerin kullanılması istatistiksel yöntemlere göre daha kolaydır.

Dezavantajları

- Kural tabanlı sistemde yeterli bir kalite eşiğine rahatlıkla ulaşılabilirken; ulaşılan kaliteyi iyileştirmek için; sözlük kayıtlarının ve biçimbilimsel kuralların iyileştirilmesi gerekmektedir. Bu işlem; dil alanında uzman insan gücü gerektirmektedir. Aynı zamanda uzun ve pahalı bir süreçtir.

7.2. İstatistiksel Sadeleştirme Modeli

İstatistiksel sadeleştirme modelinde, paralel veri kümesi üzerinden öğrenilen istatistiksel bilgiler kullanılarak sadeleştirme işlemi gerçekleştirilmektedir. Bu model tarafından gerçekleştirilen sadeleştirmeler ve bu sadeleştirmelerin BLEU skorları Çizelge 7.7 üzerinde verilmiştir.

Çizelge 7.7. İstatistiksel sadeleştirme modeli tarafından gerçekleştirilen sadeleştirmeler

Cümle Sadeleştirmeleri	**BLEU Skoru**
K: "**1335** senesi mayısının 19 uncu günü samsuna çıktım ." **S:** "**1919** senesi mayısının 19 uncu günü samsuna çıktım ."	88 → 100
K: "**vaziyet** ve **manzarai umumiye** : osmanlı devletinin **dahil** bulunduğu grup , **harbi umumîde mağlûp olmuş** , osmanlı ordusu her tarafta zedelenmiş , **şeraiti** ağır , bir **mütarekename** imzalanmış ." **S:** "**durum** ve **aldatıcı savaşına** : osmanlı devletinin **içinde** bulunduğu grup , **genel savaşta yenilmiş** , osmanlı ordusu her tarafta zedelenmiş , ağır bir **ateşkes anlaşması** imzalanmış ."	34 → 78
K: "büyük **harbin** uzun seneleri **zarfında** , millet yorgun ve fakir bir **halde** ." **S:** "büyük **savaşın** uzun seneleri **içinde** , millet yorgun ve fakir bir **durumda** ."	65 → 85
K: "millet ve memleketi **harbi umumîye** sevkedenler , kendi **hayatları endişesine** düşerek , memleketten **firar etmişler** ." **S:** "millet ve memleketi **genel savaşa** sevkedenler , kendi **canları kaygısına** düşerek , memleketten **kaçan başvurmuşlar** ."	27 → 64
K: "saltanat ve **hilâfet mevkiini** işgal eden vahdettin , müstereddi , **şahsını** ve yalnız tahtını **temin edebileceğini tahayyül** ettiği **denî tedbirler** araştırmakta ." **S:** "saltanat ve **halifelik durumunu** işgal eden vahdettin , müstereddi **düşünmekle** ve yalnız tahtını **sağlayabileceğini hayal** ettiği **alçakça önlemler** araştırmakta ."	15 → 37
K: "damat ferit paşanın **riyasetindeki** kabine ; âciz , haysiyetsiz , cebîn , yalnız padişahın iradesine **tâbi** ve onunla beraber **şahıslarını vikaye** edebilecek herhangi bir **vaziyete** razı ." **S:** "damat ferit paşanın **başkanlığındaki** kabine ; âciz , haysiyetsiz , cebîn , yalnız padişahın iradesine **bağlı** ve onunla beraber **kişiliklerini korumak** edebilecek herhangi bir **duruma** razı ."	41 → 57
K: "ordunun elinden **esliha** ve cephanesi alınmış ve alınmakta ..." **S:** "ordunun elinden **silâh** ve cephanesi alınmış ve alınmakta ..."	70 → 70
K: "itilâf devletleri , **mütareke ahkâmına riayete lüzum** görmüyorlar ." **S:** "itilâf devletleri , **ateşkes hükümlerine göstermesi gerekli** görmüyorlar ."	31 → 37
K: "birer **vesile ile** , itilâf donanmaları ve askerleri istanbulda ." **S:** "birer **fırsattan yararlanarak** , itilâf donanmaları ve askerleri istanbulda ."	80 → 69

İstatistiksel sadeleştirme modelinde, sadeleştirmeler paralel veri kümesi üzerinden otomatik olarak öğrenilmektedir. Bu yüzden, başarı eğitim veri kümesinin kalitesine ve büyüklüğüne bağlıdır. Yapılan gözlemler sonucunda; veri kümesi üzerinde sık geçen kelimelerin başarıyla sadeleştirilebildiği, az geçen kelimelerin sadeleştirilmesinde ise kalitenin düştüğü gözlemlenmiştir. Veri kümesi üzerinde geçmeyen kelimeler ise sadeleştirilememektedir.

İstatistiksel modelde, kelimeler yerine kelime grupları üzerinden sadeleştirme işlemi gerçekleştirilmektedir. Bu sayede; tekil kelimelerin birden fazla kelimeyle ya da birden fazla kelimenin tek bir kelime ile sadeleştirilebilmesi sağlanabilmektedir. Ayrıca, sadeleştirmelerden kelimeler silinebilmekte, yeni kelimeler eklenebilmekte

ya da kelimelerin sıralaması değiştirilebilmektedir. Bu sadeleştirmelerin gerçekleştiği birkaç örnek Çizelge 7.8 üzerinde verilmiştir.

Çizelge 7.8. Farklı değişim türlerine göre gerçekleştirilen sadeleştirme örnekleri

Kaynak Kelime	**Hedef Kelime**
mütarekename	ateşkes anlaşması
aynen	olduğu gibi
binaenaleyh	bu nedenle
saniyen	ikinci olarak
bilâkaydüşart	kayıtsız şartsız
temin etmek	sağlamak
devam etti	sürdü
girişildiği takdirde	girişilirse
beyanatta bulundu	konuştu
dahil olmuş	girmiş
harbi umumîde	genel savaşta
müdafaai hukuk cemiyeti	hakları savunma derneği
gazetelerde filhakika gördük	gerçekten gazetelerde gördük

İstatistiksel sadeleştirme modelinde; sadeleştirmelerin kelime gruplarına göre gerçekleştirilmesi, kural tabanlı modelde karşılaşılan bazı sorunların giderilmesini sağlamıştır. Kural tabanlı sadeleştirme modeli tarafından yanlış, istatistiksel sadeleştirme modeli tarafından doğru olarak sadeleştirilen birkaç örnek Çizelge 7.9 üzerinde verilmiştir.

Çizelge 7.9. İstatistiksel sadeleştirme modeli tarafından doğru olarak sadeleştirilen örnekler

Kaynak Kelime Grubu	**Sistem Çıktısı**
amali millîye	ulusal emeller
devleti osmanîye	osmanlı devleti
düveli mütelife	itilâf devletleri
hakimiyeti millîye	ulusal egemenlik
heyeti temsiliye	temsilci kurul
siyaseti osmanîye	osmanlı politikası

İstatistiksel modelde; birleşik yüklem yapılarının sadeleştirilmesinin, veri kümesinde sık geçen kelimeler için doğru olarak gerçekleştirilebildiği tespit edilmiştir. Bunun yanı sıra, sadeleştirilmesi doğru olmayan örnekler de bulunmaktadır. Yüklem yapıları üzerinde gerçekleştirilen birkaç sadeleştirme örneği Çizelge 7.10 üzerinde verilmiştir.

Çizelge 7.10. İstatistiksel modelde birleşik yüklem yapılarının sadeleştirilmesi

Kaynak Kelime Grubu	Sistem Çıktısı	Referans Sadeleştirme
tebliğ olundu	gereği bildirildi	gereği bildirildi
intizar eyleriz	bekleriz	bekleriz
tebliğ ettim	bildirdim	bildirdim
icap etmiştir	zorunlu olmuştur	gerekli olmuştur
talik olunur	erteleme olunur	ertelenir
tebliğ ettim	bildirdim	bildirdim
istizan etti	önermiş etti	sordu
haberdar etmemiştim	haber sürmemiştim	bildirmemiştim
tercih ettim	üstün bildirdim	yeğledim

7.2.1. Faktörlü Yöntemin Gerçekleştirilen Sadeleştirmelere Etkisi

Faktörlü yöntemde gerçekleştirilen sadeleştirmelerde kelimelerin sözdizimsel ve biçimbilimsel özellikleri de kullanılabilmektedir. İstatistiksel sadeleştirme modeli tarafından yanlış, faktörlü yöntem kullanılarak doğru olarak sadeleştirilen birkaç örnek Çizelge 7.11 üzerinde verilmiştir.

Çizelge 7.11. Faktörlü yöntemle gerçekleştirilen sadeleştirmeler

Cümle Sadeleştirmeleri	BLEU Skoru
K: “itilâf devletlerini , bilhassa ingilizleri filen mağlûp etmek icap eder .” **S:** “itilâf devletlerini , özellikle ingilizleri edimli olarak **yok etmek** gerekir .” **F:** “itilâf devletlerini , özellikle ingilizleri edimli olarak **yenmek** gerekir .”	23 → 67 → 100
K: “hükûmet daha evvel , istanbuldaki teşkilâtımız rüesassının muvafakat ve muavenetini de temin etmiş ...” **S:** “hükûmet daha önce , istanbuldaki örgütlerimizin rüesassının olur ve yardımını da **söz etmiş . .**” **F:** “hükûmet daha önce , istanbuldaki örgütlerimizin rüesassının olurunu ve yardımını da **sağlamış . .**”	15 → 41 → 72
K: “" 2 mart günü , fırka grubu içtima ettirildi .” **S:** “" 2 mart günü , parti grubu **toplantı verildi .**” **F:** “" 2 mart günü , parti grubu **toplandı .**”	49 → 69 → 100
K: “10 - avrupaca teşkili mutasavver ermenistan hududu hakkında ne düşünüyorsunuz ?” **S:** “10 - avrupaca kurulması düşünülen ermenistan **sınırları ile** ilgili olarak ne düşünüyorsunuz ?” **F:** “10 - avrupaca kurulması düşünülen ermenistan **sınırlarıyla** ilgili olarak ne düşünüyorsunuz ?”	28 → 69 → 100
K: “yolda , rauf beye tesadüf ediyor .” **S:** “yolda , rauf beye **rastladım veriyor** .” **F:** “yolda , rauf beye **rastlıyor** .”	54 → 54 → 84
K: “bunu da yaptıktan sonra ankarayı terkettim .” **S:** “bunu da yaptıktan sonra **ankarayı ayrıldım** .” **F:** “bunu da yaptıktan sonra **ankaradan ayrıldım** .”	54 → 59 → 100
K: “bu hususta ordunun dahi muaveneti talep olunur .” **S:** “bu konuda ordunun de **yardım istemek** .” **F:** “bu konuda ordunun da **yardımı istenir** .”	19 → 36 → 100
K: “gaye , milletin necat ve vatanın halâsıdır .” **S:** “amaç , milletin **esenlik** ve vatanın halâsıdır .” **F:** “amaç , milletin **esenliği** ve vatanın halâsıdır .”	30 → 36 → 75

7.2.2. İstatistiksel Sadeleştirme Modelinin Avantajları ve Dezavantajları

Avantajları

- İstatistiksel sadeleştirme modelinde çeviriler oluşturulan paralel veri kümesi üzerinden öğrenilmektedir. Bu yüzden insan gücü ihtiyacı bulunmamaktadır.
- İstatistiksel sadeleştirme modelinde; kelimelerin birden fazla kelime ile sadeleştirilmesi gerçekleştirilebilmekte, kelimelerin sırası değiştirilebilmekte, yeni kelime eklenebilmekte ve var olan kelimeler silinebilmektedir.
- İstatistiksel sadeleştirme yöntemiyle, yeni veri kümeleri kullanılarak yeni sadeleştirme modelleri kısa bir zaman içerisinde hazırlanabilmektedir.

Dezavantajları

- İstatistiksel sadeleştirme modeli, kural tabanlı modele göre, daha fazla işlem gücü ve zaman gerektirmektedir.
- İstatistiksel sadeleştirme modelinde gerçekleştirilen sadeleştirmelerin kontrolü ve yönetimi oldukça zordur.
- İstatistiksel sadeleştirme modelinde, biçimbilimsel ve sözdizimsel özelliklerin çeviride kullanılması kural tabanlı modele göre oldukça zordur.
- İstatiksel modelde elde edilen başarı direk olarak kullanılan paralel veri kümesinin kalitesine ve büyüklüğüne bağımlıdır. Fakat yeterli kalitede ve büyüklükte paralel veri kümesinin bulunması her dil çifti ya da her alan için mümkün olmamaktadır.

7.3. Hibrid Sadeleştirme Modeli

Kural tabanlı sadeleştirme modeli tarafından üretilen sadeleştirmelerin, pasif olarak, istatistiksel sadeleştirme modeline eklendiği yöntem kullanılarak çalışmadaki en iyi sonuç elde edilmiştir. Ayrıca veri kümesi üzerindeki değişmeyen kelimelerin belirlenip, bu kelimelerin tekrar veri kümesine eklenmesiyle, sonuçların daha da iyileştiği gözlemlenmiştir. Çizelge 7.12 üzerinde; istatiksel sadeleştirme modeli tarafından gerçekleştirilen sadeleştirmeler, kural tabanlı model tarafından belirlenen kelimeler ve hibrid sadeleştirme modeli tarafından gerçekleştirilen sadeleştirmeler verilmiştir. Ayrıca bu sadeleştirmelerin referans metin ile karşılaştırılması sonucu elde edilen BLUE skorları da cümlelerin yanında yer almaktadır.

Çizelge 7.12. Hibrid sadeleştirme modeli tarafından gerçekleştirilen sadeleştirmeler

Cümle Sadeleştirmeleri	**K. Tabanlı Model Sadeleştirmeleri**	**BLEU Skoru**
K: "birer **vesile** ile , itilâf donanmaları ve askerleri istanbulda ." **S:** birer **fırsattan yararlanarak** , itilâf donanmaları ve askerleri istanbulda ." **H:** "birer **bahane** ile , itilâf donanmaları ve askerleri istanbulda ."	vesile → bahane	80 → 69 → 100
K: "halbuki biliyorsunuz , mütarekenamede yalnız **şimendiferler** için kontrol mevzuubahstir ." **S:** "oysa biliyorsunuz , ateşkes anlaşmasında yalnız **şimendiferler** için denetim sözkonusudur ." **H:** "oysa biliyorsunuz , ateşkes anlaşmasında **demiryolları** için yalnız denetim sözkonusudur ."	şimendiferler → demiryolları	17 → 55 → 100
K: "her iki kongrede beyanname ve **nizamnamelerle** ilân edilmiş mukarrerattan başka kat'iyyen bir karar mevcut değildir ." **S:** "her iki kongrede bildiri ve **nizamnamelerle** ilân edilmiş mukarrerattan başka kesin bir karar yoktur ." **H:** "her iki kongrede bildiri ve **tüzüklerle** ilân edilmiş kararlardan başka kesin bir karar yoktur ."	nizamnamelerle → tüzüklerle	20 → 41 → 73
K: "**cümlenize** ihtiramatımı takdim ederim efendim ." **S:** "**cümlenize** ihtiramatımı sunarım efendim ." **H:** "**hepinize** saygılarımı sunarım efendim ."	cümlenize → hepinize	29 → 50 → 100
K: "19 - **rüfekam** da aynen benim gibi düşünmekledirler ." **S:** "19 - **rüfekam** da olduğu gibi benim gibi düşünmekledirler ." **H:** "19 - **arkadaşlarım** da tıpkı benim gibi düşünmekledirler ."	rüfekam → arkadaşları	25 → 22 → 78
K: "ehemmiyet , işin **hakikatinde** ve **mahiyetindedir** ." **S:** "önem , işin **hakikatinde** ve **niteliğindedir** ." **H:** "önem , işin **gerçeğinde** ve **niteliğindedir** ."	hakikatinde → gerçeğinde mahiyetindedir → niteliğindedir	27 → 50 → 100

Hibrid sadeleştirme modeliyle kural tabanlı ve istatistiksel modeller tek bir sistemde birleştirilmiştir. Yalnızca istatistiksel model tarafından sadeleştirilemeyen ifadeler, geliştirilen hibrid sistem ile günümüz Türkçesine sadeleştirilebilmektedir.

8. SONUÇLAR

8.1. Sonuçlar

Eski metinlerde geçen fakat günümüzde kullanılmayan kelimeler metinlerin anlaşılmasını zorlaştırmaktadır. Çalışmada eski kelimeleri ve kelime gruplarını günümüzde kullanılan karşılıkları ile değiştirilebilecek bir sistem geliştirilmiştir. Bu amaçla, kural tabanlı sadeleştirme modeli ve istatistiksel sadeleştirme modeli oluşturulmuştur.

Kural tabanlı sadeleştirme modelinde; sadeleştirmeler, oluşturulan sözlük üzerinden gerçekleştirilmektedir. Modelde, ilk olarak eski kelimeler tespit edilmekte, daha sonra belirlenen kelimeler üzerinde biçimbilimsel analiz, sözcüksel aktarım ve biçimbilimsel sentez işlemleri gerçekleştirilerek aday kelime listesi oluşturulmaktadır. En son olarak, aday kelimelerden en yüksek skora sahip olan kelime dil modeli sorgusu ile belirlenmektedir.

İstatistiksel sadeleştirme modelinde; sadeleştirmeler, paralel veri kümesi üzerinden otomatik olarak öğrenilmektedir. Modelde ilk olarak veri kümesindeki cümlelerde kelime tabanlı eşleştirme işlemi gerçekleştirilmektedir. Daha sonra, Moses aracı kullanılarak istatistiksel makine çeviri modeli eğitilmiştir. Aynı zamanda, farklı kelime eşleştirme araçlarının, farklı dil modeli veri kümelerinin ve faktörlü makine çevirisi yönteminin sistem başarısı üzerindeki etkisi test edilmiştir.

Oluşturulan kural tabanlı sadeleştirme modeli ve istatistiksel sadeleştirme modeli birbirleri ile karşılaştırılmış ve hibrid sistemde birleştirilmiştir. En iyi sonuç, kural tabanlı sadeleştirme sistemi tarafından üretilen sadeleştirmelerin ve paralel veri kümesi üzerinde değişmeyen kelimelerin istatistiksel sadeleştirme modelinin eğitim kısmına eklendiği birleştirme yönteminde elde edilmiştir. Sistemlerin başarısının değerlendirilmesinde makine çevirisi sistemlerinin başarı ölçümünde kullanılan BLEU metriği kullanılmıştır.

Çalışma ile eski metinler sadeleştirilerek, okurların bu metinleri daha kolay anlamaları sağlanmaktadır. Ayrıca, dil işleme araçlarında eski Türkçe kelimelerin sözlükte bulunmaması probleminden doğan başarı düşme sorunu da azaltılmaktadır.

8.2. Gelecek Çalışma

Eski Türkçe metinlerin günümüz Türkçesine sadeleştirilmesi için oluşturulan sözlük ve Nutuk paralel veri kümesi kullanılmaktadır. Bu iki kaynak Cumhuriyet dönemi Türkçesinde kullanılan kelimelerin yalnızca bir bölümünü kapsamaktadır. Geliştirilen modellerde, kullanılan kaynaklarda bulunan kelimeler güncel karşılıkları ile sadeleştirilebilmesine rağmen, bu kaynaklarda bulunmayan kelimelerin sadeleştirilmesi mümkün değildir. Gelecekte, kullanılan sözlük ve paralel veri kümesinin büyütülmesi sağlanacaktır. Ayrıca, eğitim kümesinde bulunmayan kelimeleri yakın anlamlıları ile değiştirebilecek bir word2vec modeli üzerine yoğunlaşılacaktır. Bu sayede veri kümesinin sınırlı olmasından kaynaklanan sorunun önüne geçilerek sadeleştirmenin daha geniş bir kapsamda yapılabilmesi sağlanacaktır.

KAYNAKLAR

[1] Akay, R., Dil değişiminin dilsel ve toplumsal nedenleri, *Uluslararası İnsan Bilimleri Dergisi*, **2007**.

[2] Ersoy, M. A., *Safahat - Orijinali ve Günümüz Türkçesi*, (sadeleştirme: Akbaş, A. V.), **2007**.

[3] Adıvar, H. E., *Ateşten Gömlek* (sadeleştirme: Can Yayınları), **2016**.

[4] Bulut, M., Atatürk'ün Türkçeye yönelik özleştirme/sadeleştirme çalışmaları ve bu bağlamda yaşanan dil tartışmaları üzerine bir değerlendirme, *International Periodical For The Languages, Literature and History of Turkish or Turkic*, c. 9/11, sayı Fall, ss. 131–147, **2014**.

[5] Y. Doğaner, "Elifba'dan alfabeye: Yeni Türk harfleri", *Modern Türklük Araştırmaları Dergisi*, c. 2/4, sayı Aralık, ss. 27–44, **2005**.

[6] Shardlow, M., A survey of automated text simplification, *International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Natural Language Processing*, ss. 58–70, **2014**.

[7] Albachten, Ö. B., Intralingual translation as 'modernization' of the language: The Turkish case, *Perspectives: Studies in Translatology*, c. 21, sayı 2, ss. 257–271, **2013**.

[8] Atatürk, M. K., *Nutuk,* (orijinal metin), **1927**.

[9] Atatürk, M. K., *Nutuk,* (çevrim yazı), **1938**.

[10] Atatürk, M. K., *Nutuk,* (sadeleştirme: Yazıcı, B.), **1995**.

[11] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Wade, S., Moran, C., Zens, R., Dyer, C., Ondřej, B., Constantin, A., Herbst, E., Moses: Open source toolkit for statistical machine translation, *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL'07),* sayı June, ss. 177–180, **2007**.

[12] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., BLEU: a method for automatic evaluation of machine translation, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ss. 311–318, **2002**.

[13] Xu, W., Napoles, C., Pavlick, E., Chen, Q., Callison-Burch, C., Optimizing statistical machine translation for text simplification, *Transactions of the Association for Computational Linguistics*, c. 4, ss. 401–415, **2016**.

[14] Angrosh, M., Nomoto, T., Siddharthan, A., Lexico-syntactic text simplification and compression with typed dependencies, *25th International Conference on Computational Linguistics*, ss. 1996–2006**, 2014.**

[15] Chen, H., Huang, H., Chen, H., Tan C.., A simplification-translation-restoration framework for cross-domain smt applications, *Coling-2012*, c. 2, sayı December, ss. 545–560, **2012.**

[16] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., Distributed representations of words and phrases and their compositionality, *NISP*, ss.

3111–3119, **2013.**

[17] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J., Introduction to wordnet: An on-line lexical database, *International Journal of Lexicography*, c. 3, sayı 4, ss. 235–244, **1990**.

[18] Glavas, G., Stajner, S., Simplifying lexical simplification : do we need simplified corpora ?, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 1, ss. 63–68, **2015**.

[19] Ligozat, A.-L., Garcia-Fernandez, A., ANNLOR: a naïve notation-system for lexical outputs ranking, *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012),* ss. 487–492, **2012**.

[20] Thomas, S., Anderson S., Wordnet-based lexical simplification of a document, *Proceedings of KONVENS*, c. 2012, ss. 80–88, **2012**.

[21] Wikipedia, Simple Wikipedia, https://simple.wikipedia.org (Temmuz, **2018**).

[22] Siddharthan, A., Syntactic simplification and text cohesion*, Research on Language and Computation*, c. 4, sayı 1, ss. 77–109, **2006**.

[23] Torunoglu-Selamet, D., Pamay, T., Eryigit, G., Simplification of Turkish sentences, *The First International Conference on Turkic Computational Linguistics*, ss. 55–59, **2016**.

[24] Zhu, Z., Bernhard, D., Gurevych, I., A monolingual tree-based translation model for sentence simplification*, 23rd International Conference on Computational Linguistics*, sayı August, ss. 1353–1361, **2010**.

[25] Ganitkevitch, J., Van Durme, B., Callison-Burch, C., PPDB: The paraphrase database, *Proceedings of NAACL-HLT*, sayı June, ss. 758--764, **2013**.

[26] Xu, W., Ritter, A., Dolan, W. B., Grishman R., Cherry, C., Paraphrasing for style, *Proceedings of COLING 2012*, c. 4, sayı 1234, ss. 2899–2914, **2012**.

[27] Jhamtani, H., Gangal, V., Hovy, E., Nyberg, E., Shakespearizing modern language using copy-enriched sequence-to-sequence models, *arXiv preprint arXiv:1707.01161*, **2017**.

[28] Wang, T., Chen, P., Rochford, J., Qiang, J., Text simplification using neural machine translation, *30th AAAI Conference on Artificial Intelligence*, **2016**.

[29] Cho, K., Merrienboer, B. V., Gulcehre, C., Bahdanau D., Bougares, F., Schwenk, H., Bengio, Y., Learning phrase representations using RNN encoder–decoder for statistical machine translation, *In Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP)*, **2014.**

[30] Okpor, M. D., Machine translation approaches: Issues and challenges, *International Journal of Computer Science Issues*, c. 11, sayı 5, ss. 159–165, **2014**.

[31] Hutchins, W. J., Machine translation:a brief history, *Concise history of the language sciences: from the Sumerians to the cognitivists*, ss. 431–445, **1995**.

[32] Nyberg, E. H., Mitamura, T., The KANT system: fast, accurate, high-quality translation in practical domains, *Coling '92*, sayı July, ss. 1069–1073, **1992**.

[33] Hajic, J., An MT system between closely belated languages", *In Proceedings of the 3rd Conference of The European Chapter of the Association for Computational Linguistics*, Copenhagen, Denmark, ss. 113–117, **1987**.

[34] Hajič, J., Hric, J., Kubon, V., Machine translation of very close languages, *Sixth conference on Applied natural language*, ss. 7–12, **2000**.

[35] Tantuğ, C., Adalı, E., Oflazer, K., Türkmenceden Türkçeye bilgisayarlı metin çevirisi, *İTÜ Dergisi*, c. 7, sayı 4, ss. 83–94, **2008**.

[36] Altintas, K., Çiçekli, İ., A machine translation system between a pair of closely related languages", *Proceedings of the 17th International Symposium on Computer and Information Sciences (ISCIS),* c. October, ss. 192–196, **2002**.

[37] Wikipedia, Wikipedia, https://www.wikipedia.org (Temmuz, **2018**).

[38] Koehn, P., Europarl: A parallel corpus for statistical machine translation, *MT Summit*, c. 11, ss. 79--86, **2005**.

[39] Koehn, P., Hoang, H., Factored translation models, *Computational Linguistics*, sayı June, ss. 868–876, **2007**.

[40] Khalilov, M., Fonollosa, J. A. R., Syntax-based reordering for statistical machine translation, *Computer Speech and Language*, c. 25, sayı 4, ss. 761–788, **2011**.

[41] Oflazer, K., Statistical machine translation into a morphologically complex language, *Computational Linguistics and Intelligent Text Processing, 9th International Conference (CICLing 2008),* Haifa, Israel, c. 4919, ss. 376–387, **2008**.

[42] El-Kahlout, D*., A prototype english-turkish statistical machine translation system*, Doktora Tezi, Sabancı Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, **2009.**

[43] Kalchbrenner, N., Blunsom, P., Recurrent continuous translation models, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, sayı October, ss. 1700–1709, **2013**.

[44] Sutskever, I., Vinyals, O., Le, Q. V., Sequence to sequence learning with neural networks, *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, c. 2, ss. 3104–3112, **2014.**

[45] Bahdanau, D., Cho, K., Bengio, Y., Neural machine translation by jointly learning to align and translate, *arXiv preprint arXiv:1409.0473*, **2014**.

[46] Cho, K., Merrienboer, B. V., Bahdanau, D., Bengio, Y., On the properties of neural machine translation : encoder – decoder approaches, *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8),* ss. 103–111, **2014**.

[47] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu,

X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J., Google's neural machine translation system: bridging the gap between human and machine translation, *arXiv preprint arXiv:1609.08144*, ss. 1–23, **2016**.

[48] Costa-Jussà, M. R., Fonollosa, J. A. R., Latest trends in hybrid machine translation and its applications, *Computer Speech and Language*, c. 32, sayı 1, ss. 3–10, **2015**.

[49] Sánchez-Cartagena, V. M., Sánchez-Martínez, F., Pérez-Ortiz, J. A., Enriching a statistical machine translation system trained on small parallel corpora with rule-based bilingual phrases, *International Conference Recent Advances in Natural Language Processing (RANLP)*, **2011**.

[50] Tan, L., Genabith, J. V., Bond, F., Passive and pervasive use of bilingual dictionary in statistical machine translation", *Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra),* ss. 30–34, **2015**.

[51] Uzun, H., *Atatürk'ün Nutuk'unun içerik analizi*, Doktora Tezi, Hacettepe Üniversitesi Atatürk İlkeleri ve İnkılap Tarihi Enstitüsü, Ankara, **2005**.

[52] Vikipedi, Nutuk hakkında bigiler ve Nutuk'un farklı baskıları, https://tr.wikipedia.org/wiki/Nutuk_(Mustafa_Kemal_Atatürk) (Haziran, **2018**).

[53] Börekçi, M., Atatürk'ün Nutuk'unda söz dizimi ve üslûp özellikleri, *Sosyal Araştırmalar Dergisi,* c. 1(5), ss. 104–125, **2008**.

[54] Karamanlıoğlu, F., Atatürk'ün Gençliğe Hitabesi'nin üslup ve dil özellikleri, *Milli Kültür*, c. Kasım, ss. 97–99, **1963**.

[55] Xu, W., Callison-Burch, C., Napoles, C., Problems in current text simplification research: new data can help", *Transactions of the Association for Computational Linguistics*, c. 3, ss. 283–297, **2015**.

[56] Li, P., Sun, M., Xue, P., Fast-Champollion: A fast and robust sentence alignment algorithm, *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)* , ss. 710–718, **2010**.

[57] Sourceforge, LF Aligner, https://sourceforge.net/projects/aligner, (Temmuz, **2018**).

[58] Varga, D., Halacsy, P., Kornia, A., Nagy, V., Nemeth, L., Tron, V., Parallel corpora for medium density languages, *Proceedings of the International Conference of Recent Advances in Natural Language Processing*, sayı 2003, ss. 590–598, **2004.**

[59] Collins, M., Discriminative training methods for hidden Markov models, *Proceedings of the ACL-02 conference on Empirical methods in natural language processing (EMNLP)*, c. 10, ss. 1–8, **2002**.

[60] Github, Zemberek, https://github.com/ahmetaa/zemberek-nlp, (Temmuz, **2018**).

[61] Coltekin, C., A freely available morphological analyzer for Turkish, *Proceedings of the 7th International Conference on Language Resources and Evaluation,* ss. 820–827, **2010**

[62] Hulden, M., Foma: A finite-state compiler and library, *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, sayı April, ss. 29–32, **2009**.

[63] Beesley, K. R., Karttunen, L., *Finite state morphology*, c. 30, sayı 2. **2003**.

[64] Kutlu, M., Cicekli, I., A hybrid morphological disambiguation system for turkish, *International Joint Conference on Natural Language Processing*, ss. 1230–1236, **2013**.

[65] Devellioğlu, F., *Osmanlıca-Türkçe ansiklopedik lügat*, **2013**.

[66] Anonim, Osmanlıca-Türkçe Lügat, http://www.osmanlimedeniyeti.com/makaleler/sozluk, (Temmuz, **2018**).

[67] Türk Bilim, Öz Türkçe Karşılıklar Kılavuzu, https://media.turuz.com/Dictionary/2011/0113-Eshanlam-_oz_turkce_qarshiliklar_gilavuz_sozluyu(637KB).pdf, (Temmuz, **2018**).

[68] Heafield, K., KenLM : Faster and smaller language model queries, *Proceedings of the Sixth Workshop on Statistical Machine Translation*, sayı 2009, ss. 187–197, **2011**.

[69] Chen, S. F., Goodman, J., An empirical study of smoothing techniques for language modeling, *Proceedings of the 34th annual meeting on Association for Computational Linguistics (ACL)*, c. 13, sayı August, ss. 310–318, **1998**.

[70] Teh, Y. W., A Bayesian Interpretation of Interpolated Kneser-Ney, *Citeseer*, sayı 50000, ss. 1–19, **2006**.

[71] Dempster, A. P., Laird, N. M., Rubin, D. B., Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, Series B, c. 37, ss. 1–22, **1977**.

[72] Liang, P., Taskar, B., Klein, D., Alignment by agreement, *Naacl2006*, sayı June, ss. 104–111, **2006**.

[73] Dyer, C., Chahuneau, V., Smith, N. A., A simple, fast, and effective reparameterization of IBM model 2, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, sayı June, ss. 644–649, **2013**.

[74] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Mercer, R. L., The mathematics of statistical machine translation: Parameter estimation, *Computational linguistics*, c. 19, ss. 263–311, **1993**.

[75] Kohavi, R., A study of cross-validation and bootstrap for accuracy estimation and model selection, *International Joint Conference on Articial Intelligence (IJCAI),* ss. 1–7, **1995**.

EK 1: BİÇİMBİLİMSEL ETİKETLER

Bu ekte TRMorph aracında üretilen kelime türü işaretlerinin listesi Çizelge 1.1 üzerinde verilmiştir.

Çizelge 1.1. TRMorph aracında kullanılan kelime türü işaretleri

Kelime Türü İşareti	Kelime Türü
<Alpha>	Alfanümerik
<Adj>	Sıfat
<Adv>	Zarf
<Cnj>	Bağlaç
<Det>	Belirteç
<Exist>	"var" ve "yok" kelimeleri
<Ij>	Ünlem
<N>	İsim
<Not>	"değil" kelimesi
<Num>	Sayı
<Onom>	Yansıma kelimeler
<Postp>	Edat
<Prn>	Zamir
<Punc>	Noktalama
<Q>	Soru parçacığı
<V>	Fiil

TRMorph aracında isimler için üretilebilen biçimbilimsel eklerin listesi Çizelge 1.2 üzerinde verilmiştir.

Çizelge 1.2. TRMorph aracında kullanılan isim ekleri

Biçimbilimsel Ek	Yüzeysel Biçim
<pl>	-lAr (Çoğul eki)
<abl>	-DAn (Ayrılma durumu eki)
<acc>	-(y)I (Belirme durumu eki)
<dat>	-(y)A (Yönelme durumu eki)
<gen>	-(n)In (İlgi eki)
<ins>	-(y)lA (İsmin vasıta hali)
<loc>	-DA (Bulunma durumu eki)
<p1s>	-(I)m (1. tekil şahıs eki)
<p2s>	-(I)n (2. tekil şahıs eki)
<p3s>	-(s)I (3. tekil şahıs eki)
<p1p>	-(I)mIz (1. çoğul şahıs eki)
<p2p>	-(I)nIz (2. çoğul şahıs eki)
<p3p>	-lArI (3. çoğul şahıs eki)

TRMorph aracında fiiller için üretilebilen şahıs eklerinin listesi Çizelge 1.3 üzerinde verilmiştir.

Çizelge 1.3. TRMorph aracında fiiller için kullanılan şahıs ekleri

Biçimbilimsel Ek	Yüzeysel Biçim
<1s>	-(y)Im
<2s>	-sIn
<3s>	-
<1p>	-(y)Iz
<2p>	-sInIz
<3p>	-lAr

TRMorph aracında fiiller için üretilip birleşik fiil oluşturan biçimbilimsel ek olasılıklarının listesi Çizelge 1.4 üzerinde verilmiştir.

Çizelge 1.4. TRMorph aracında kullanılan fiil ekleri

Biçimbilimsel Ek	Yüzeysel Biçim
<abil><V>	-(y)Abil
<iver><V>	-(y)Iver
<agel><V>	-(y)Agel
<adur><V>	-(y)Adur
<ayaz><V>	-(y)Ayaz
<akal><V>	-(y)Akal
<agor><V>	-(y)gör

TRMorph aracında fiiller için üretilip fiilde kullanılan kipi belirten biçimbilimsel ek olasılıklarının listesi Çizelge 1.5 üzerinde verilmiştir.

Çizelge 1.5. TRMorph aracında kullanılan fiil ekleri

Biçimbilimsel Ek	Yüzeysel Biçim
<evid>	-mIş
<fut>	-(y)AcAk
<obl>	-mAlI
<impf>	-mAktA
<cont>	-(I)yor
<past>	-DI
<cond>	-sA,-(y)A
<opt>	-(y)A
<imp>	-
<aor>	-Ar,-Ir,-z

ÖZGEÇMİŞ

Kimlik Bilgileri

Adı Soyadı:	Erol Özkan
Doğum Yeri:	Ankara, Türkiye
Medeni Hali:	Bekâr
E-posta:	erolozkan@outlook.com
Adresi:	Bilgisayar Mühendisliği, Hacettepe Üniversitesi Beytepe/ANKARA

Eğitim

Lisans:	Başkent Üniversitesi, Bilgisayar Mühendisliği, 3.68/4.00, Bölüm Birincisi
Yüksek Lisans:	Hacettepe Üniversitesi, Bilgisayar Mühendisliği, 3.75/4.00

Yabancı Dil Düzeyi

YDS – 82

İş Deneyimi

Agmlab Bilişim Teknolojileri, Ağustos 2015 – Ocak 2016

Tübitak Bilgem – Siber Güvenlik Enstitüsü, Şubat 2016 - Günümüz

Deneyim Alanları

Metin Madenciliği, Doğal Dil İşleme, Java, Python, Openstack, Kubernetes.

Tezden Üretilmiş Projeler ve Bütçesi

-

Tezden Üretilmiş Yayınlar

Özkan, E., Ercan, G., Eski Türkçe metinlerin modernleştirilmesi, *26. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SIU),* İzmir, **2018**.

Tezden Üretilmiş Tebliğ ve/veya Poster Sunumu ile Katıldığı Toplantılar

-

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
YÜKSEK LİSANS/~~DOKTORA~~ TEZ ÇALIŞMASI ORJİNALLİK RAPORU

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLER ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI BAŞKANLIĞI'NA

Tarih: 06/08/2018

Tez Başlığı / Konusu: ESKİ TÜRKÇE METİNLERİN GÜNÜMÜZ TÜRKÇESİNE SADELEŞTİRİLMESİ

Yukarıda başlığı/konusu gösterilen tez çalışmamın a) Kapak sayfası, b) Giriş, c) Ana bölümler d) Sonuç kısımlarından oluşan toplam .63..... sayfalık kısmına ilişkin, 06/08/2018 tarihinde ~~şahsım~~/tez danışmanım tarafından *Turnitin* adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % 2 'tür.

Uygulanan filtrelemeler:
1- Kaynakça hariç
2- Alıntılar hariç/~~dâhil~~
3- 5 kelimeden daha az örtüşme içeren metin kısımları hariç

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü Tez Çalışması Orjinallik Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve bu Uygulama Esasları'nda belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

Tarih ve İmza

06/08/2018

Adı Soyadı: EROL ÖZKAN
Öğrenci No: N14326457
Anabilim Dalı: BİLGİSAYAR MÜHENDİSLİĞİ
Programı: BİLGİSAYAR MÜHENDİSLİĞİ
Statüsü: ☒ Y.Lisans ☐ Doktora ☐ Bütünleşik Dr.

DANIŞMAN ONAYI

UYGUNDUR.

Dr. Öğr. Üyesi Gönenç Ercan
(Unvan, Ad Soyad, İmza)