

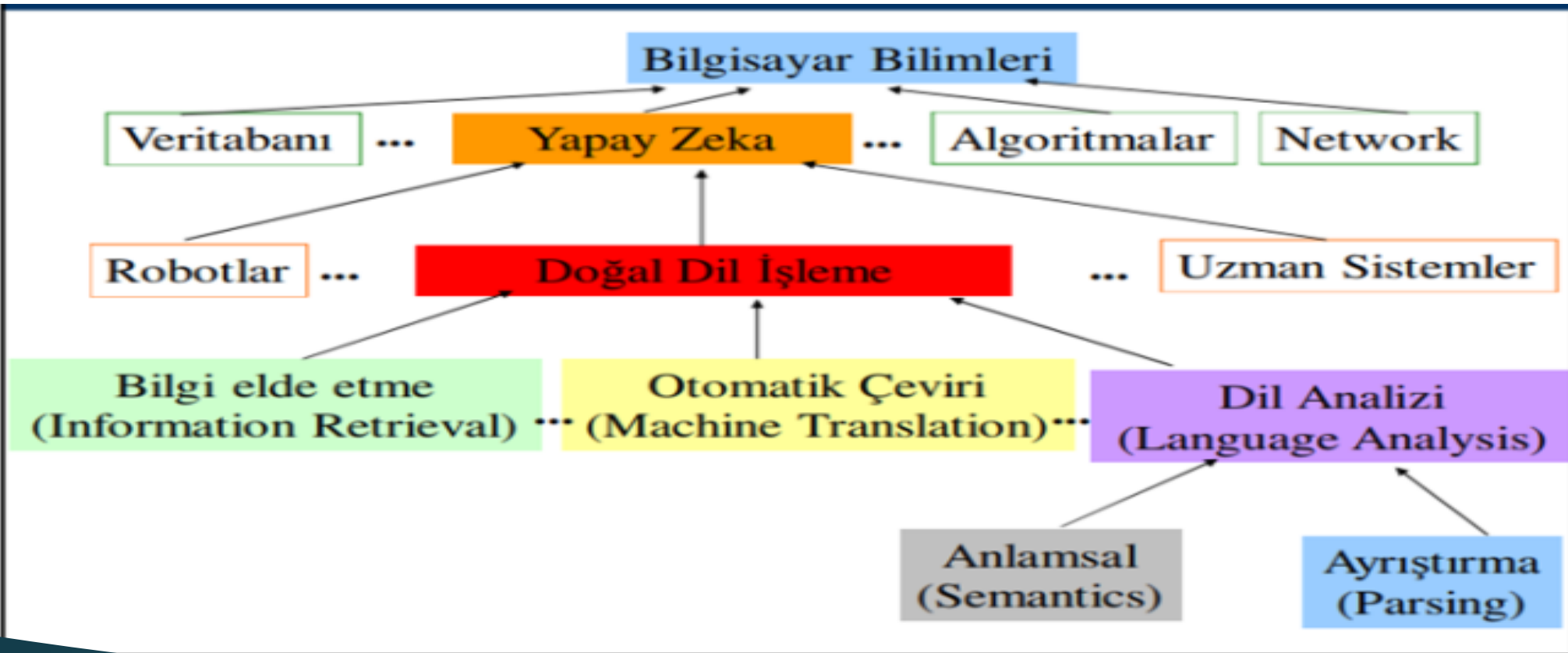


TÜRKÇE İÇİN METİN DÜZELTME

FERHAT KARTAL

DANIŞMAN: DR. FATMA NEDA TOPUZ

- ❑ Doğal dil işleme nedir?
- ❑ Doğal dil işlemenin önemi
- ❑ Doğal dil işleme kullanım alanları
- ❑ Amaç
- ❑ Önceki Çalışmalar
- ❑ Yöntem
 - Aşama1
 - Aşama2
 - Aşama3
- ❑ Örnekler
- ❑ Değerlendirme
- ❑ Çalışma Sonuçları
- ❑ Zemberek ile Karşılaştırma
- ❑ Yorum
- ❑ Öneri



DOĞAL DİL İŞLEMENİN YERİ

Doğal Dil işleme nedir?

4

- ▶ NLP yani Doğal Dil İşleme, doğal dillerin kurallı yapısının çözümlenerek anlaşılması veya yeniden üretilmesi amacını taşır.
- ▶ Bu çözümlemenin insana getireceği kolaylıklar, *yazılı dokümanların otomatik çevrilmesi, soru-cevap makineleri, otomatik konuşma ve komut anlama, konuşma sentezi, konuşma üretme, otomatik metin özetleme, bilgi sağlama* gibi birçok başlıkla özetlenebilir.
- ▶ Bilgisayar teknolojisinin yaygın kullanımı, bu başlıklardan üretilen uzman yazılımların gündelik hayatımızın her alanına girmesini sağlamıştır.
[wikipedia]

Doğal Dil İşlemenin Önemi

5

- Gelecekte, konuşma sentezleyiciler ve konuşma anlama alanındaki gelişmeler ve makine-insan iletişiminin gelişmesi, insanın makineden beklentilerini yükseltecektir.
- Geleceğin en önemli sektörlerinden biri olan yapay zekâ ile insanın iletişim kuracağı tek araç dildir.
[wikipedia]



Doğal Dil İşleme Kullanım Alanları

- ▶ Makine Çevirisi
- ▶ Ses Tanıma
- ▶ Yazım Denetimi
- ▶ Metin Özetleme
- ▶ Soru Cevaplama



Doğal Dil İşleme Kullanım Alanları

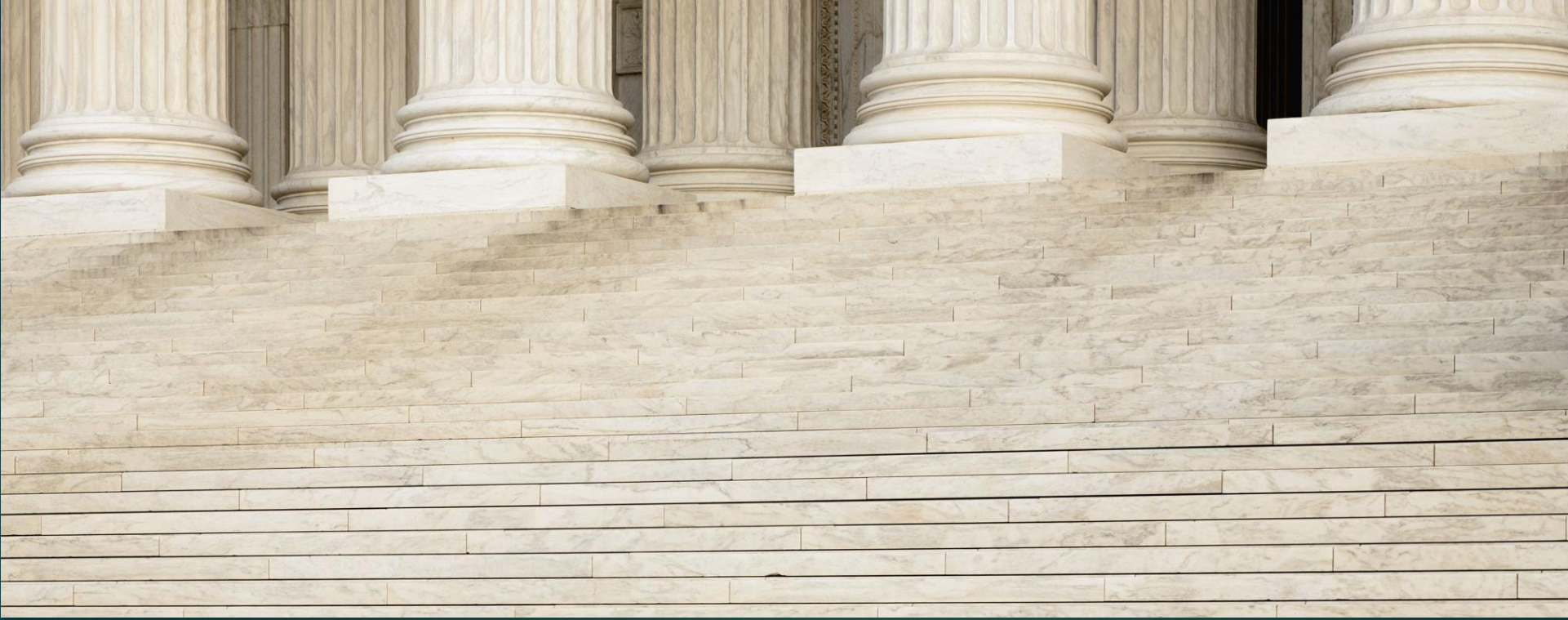
- Makine Çevirisi
- Ses Tanıma
- **Yazım Denetimi**
- Metin Özetleme
- Soru Cevaplama



AMAÇ

- Hatalı kelime içeren ve anlamsal bütünlüğü olmayan cümlelerin anlamlı bir cümle yapısına dönüştürülmesidir.





ÖNCEKİ ÇALIŞMALAR

- Yılmaz Ince vd. (2017). Spell Checking and Error Correcting Application for Turkish
- Polat vd. (2019). Otomatik Konuşma Tanıma Sistemlerinde Kullanılan Gerçek Metin Verisinde Biçimbilimsel-Sözdizimsel Hataların Tespiti ve Düzeltmesi
- Çolakoğlu vd. (2019). Normalizing Non-canonical Turkish Texts Using Machine Translation Approaches. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics .

- Bölücü vd. (2019). Context Based Automatic Spelling Correction for Turkish.
- Aydoğan vd. (2020). Spelling Correction with the Dictionary Method for the Turkish Language Using Word Embeddings
- Koksal vd. (2020). #Turki\$hTweets: A Benchmark Dataset for Turkish Text Correction.

- Keleş vd. (2021). Metin Benzerliği Algoritmaları ile Veri Tekilleştirme: Oteller Veri Tabanında Bir Uygulama.
- Arslan vd. (2021). Detecting and correcting automatic speech recognition errors with a new model.
- Uz vd. (2023). Towards Automatic Grammatical Error Type Classification for Turkish

YÖNTEM:
anlamsız cümle -> analiz -> anlamlı cümle

Yeni **bit arba** gördüm



Yeni bir araba gördüm



Aşama 1



Aşama 1: Her kelimeye yakın olan kelimelerin bulunması

15

güzel

bit

arba

gördüm

güzel

bit

araba

gördüm

gazel

fit

ara

görgü

güncel

kit

arma

görüp

gürel

bir

aba

görür

tüzel

bil

arpa

gördü

Kelime benzerlik algoritmaları



- Mesafe Düzenleme Tabanlı Benzerlik Algoritmaları,
 - Levenshtein,
 - Hamming ,
 - Jaro

- **Token Tabanlı Benzerlik Algoritmaları,**
 - **Jacard,**
 - **Sorensendice,**
 - **Tversky,**
 - **Overlap**

- **Diziliş Tabanlı Benzerlik Algoritmaları,**
 - **Ratcliff-Obershelp,**
 - **Longest-Common Substring**

İki kelime karşılaştırılarak benzerlik oranı belirlenir

20

%87.5

gökyüzü



günyüzü

Algoritma	Toplam Kelime	İsabet	Başarı Oran
Levenshtein	50	40	%80
Hamming	50	29	%58
Jaro	50	36	%72
Jacard	50	11	%22
Sorensendice	50	12	%24
Tversky	50	13	%26
Overlap	50	5	%10
Ratcliff-Obershelp	50	38	%76
Longest-Common Substring	50	20	%40

**Neden
Levensthein?**

**PROJEMİZ İÇİN EN
BAŞARILI
ALGORITMA**

Aşama 2: Kelime dizilimleri



N-gram nedir?

23

- ▶ Aynı cümle içerisinde kelimelerin birbirlerini ardınca gelme durumu incelenir.
 - ▶ Bigram: İki kelimenin birbiri ardınca gelmesini dikkate alır.
 - ▶ Trigram: Üç kelimenin birbiri ardınca gelmesini dikkate alır.

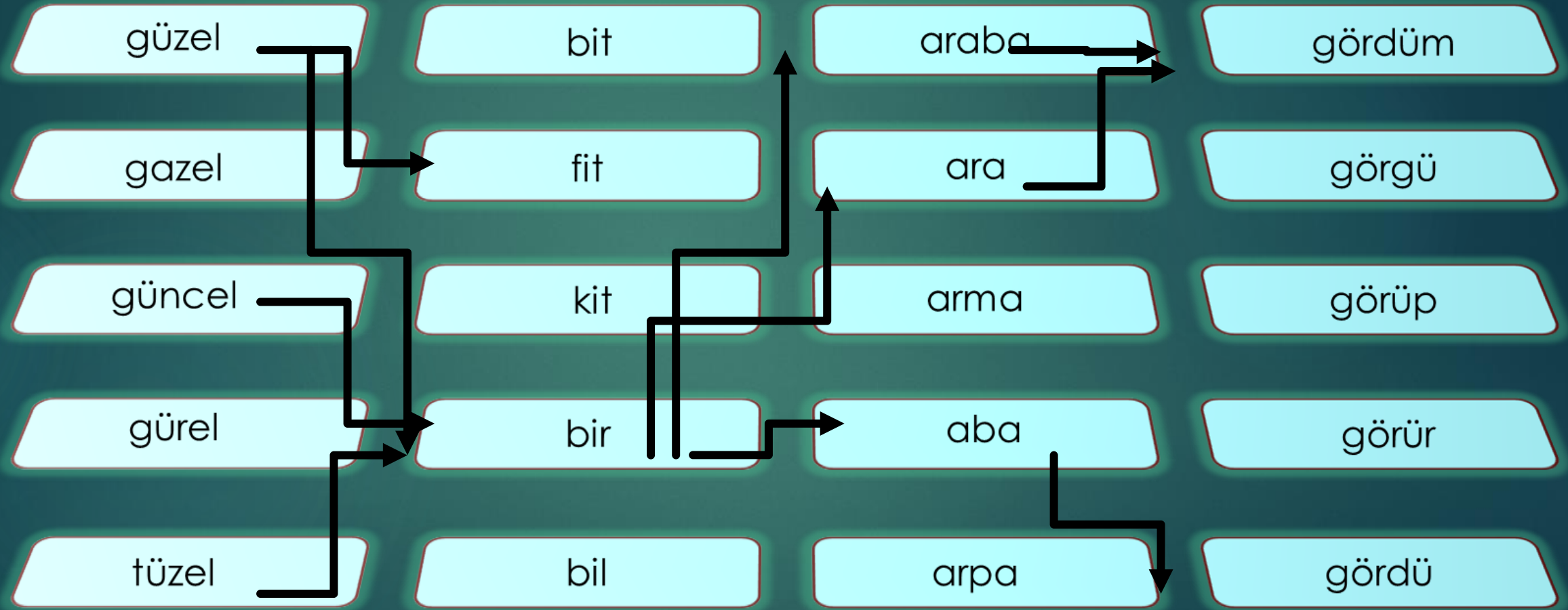
Metin	“Okuldan sonra sinemaya gitti. Eve gelmedi.”
Unigramlar	“okuladan”, “sonra”, “sinemaya”, “gitti”, “eve”, “gelmedi”
Bigram	“okuldan sonra”, “sonra sinemaya”, “sinemaya gitti”, “gitti eve”, “eve gelmedi”
Trigramlar	“okuldan sonra sinemaya”, “sonra sinemaya gitti”, “sinemaya gitti eve”, “gitti eve gelmedi”
N-gramlar (n=4)	“okuldan sonra sinemaya gitti”, “sonra sinemaya gitti eve”, “sinemaya gitti eve gelmedi”

Seçilen yöntem: Bigram

- ▶ Sebebi, daha fazla veri elde etmek

Kelime dizilimi

26



bigram

27

Güzel fit

Güzel bir

Güncel bir

bir ara

bir araba

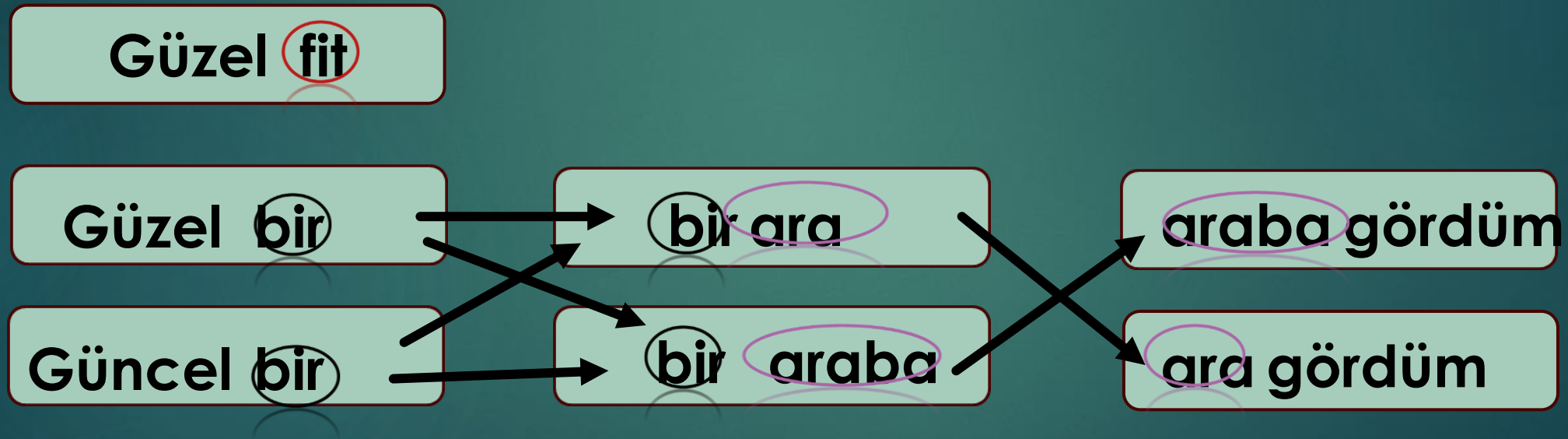
araba gördüm

ara gördüm

Zincirleme

28

- Bigram'ın son kelimesiyle bir sonraki bigram'ın ilk kelimesi aynıysa bağlanır.



Son dizilim

29

- Kelimeler bağlandıktan sonra elde edilen cümleler

Güzel bir araba gördüm

Güzel bir ara gördüm

Güncel bir araba gördüm

Güncel bir ara gördüm

Ařama 3:

İSTATİSTİK



En uygun cümlelerin belirlenmesi

31

- ▶ Tek bir cümle önerisi varsa istatistiğe gerek yok.
- ▶ Fakat birden fazla cümle olduğunda en uygun olanı belirlenmeli.

- ▶ Her cümlenin yüklemi baz alınarak veri tabanına göre istatistik çıkarılması
- ▶ Cümle içerisindeki her kelimenin cümlenin yüklemiyle veri tabanında kaç kere geçtiğinin sayılarak elde edilen skorun kaydedilmesi
- ▶ Tüm kelimelerin skorlarının toplamının cümlenin skoru olarak yazılması

Neden yüklem?

33

- Cümle analiz edilirken tüm sorular yükleme sorulur

özne

dolaylı
tümleç

zarf
tümleci

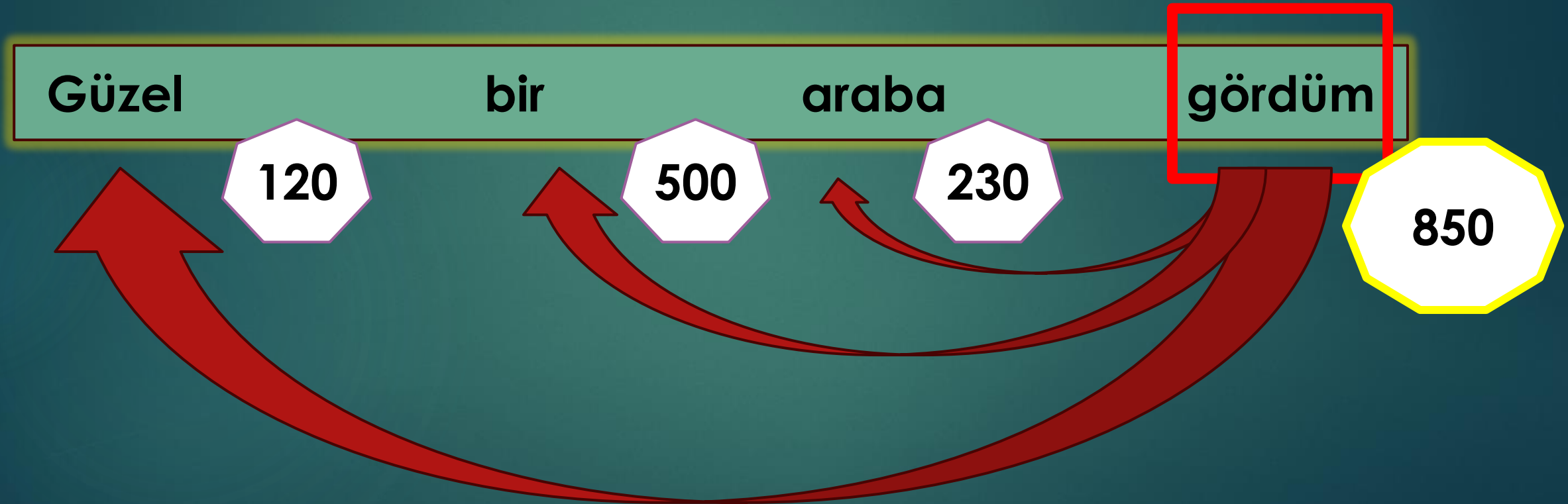
...

yüklem

kim?

nereye?

ne
zaman?



Güzel bir araba gördüm

850

Güzel bir ara gördüm

650

Güncel bir araba gördüm

750

Güncel bir ara gördüm

450

- En çok skora sahip olan cümle en uygun cümle olarak kabul edilir.

Güzel bir araba gördüm

850

Örnekler

Örnek1: Bir hatalı kelime içeren cümleler

 TEXT CORRECTION

yazın hav güneşli olur

mesaj girin

öneriler

yazın hava güneşli olur -->10

yaların hal güneşli olur -->6

temizle

düzeltil

TEXT CORRECTION

mavi bir arba yolda gidiyor

mesaj girin

öneriler

mavi bir araba yolda gidiyor -->187
mavi bir arma yolda gidiyor -->182
mavi bir ara yolda gidiyor -->215
mavi bir arka yolda gidiyor -->184

temizle

düzeltil

TEXT CORRECTION

toplantı bugün öğle satinde olur

mesaj girin


öneriler

toplantı bugün öğle saatinde olur

temizle

düzeltil

Örnek2: İki hatalı kelime içeren cümleler

 TEXT CORRECTION

yazın hav güneşli olur


mesaj girin

öneriler

yazın hava güneşli olur -->10
yalın hal güneşli olur -->6

temizle

düzelt

 TEXT CORRECTION

mavi bir arba yoda gidiyor

mesaj girin

öneriler

mavi bir araba yolda gidiyor -->187

mavi bir arma yolda gidiyor -->182

mavi bir ara yolda gidiyor -->215

mavi bir arka yolda gidiyor -->184

temizle

düzelt

TEXT CORRECTION

toplani bugün öğle satinde olu|

mesaj girin


öneriler

toplanır bugün öğle saatinde olur

temizle

düzeltil

Örnek3: Üç hatalı kelime içeren cümleler

 TEXT CORRECTION

yazın hav güneşli oluş


mesaj girin

öneriler

yazın hava güneşli oluş -->79
yalın hal güneşli oluş -->56

temizle

düzelt

 TEXT CORRECTION

mavi bir arba yoda gidiyor

mesaj girin

öneriler

mavi bir araba yolda gidiyor -->187

mavi bir arma yolda gidiyor -->182

mavi bir ara yolda gidiyor -->215

mavi bir arka yolda gidiyor -->184

temizle

düzeltil

TEXT CORRECTION

toplani bugün öğle satinde oluk

mesaj girin

öneriler

toplanır bugün öğle saatinde oluk

temizle

düzeltil

SONUÇ	Hatalı Kelime Sayısı	Test Sayısı	Doğru Sonuç Sayısı	Başarı Oranı
	1	10	7	%70
	2	10	3	%30
	3	10	1	%10

Test sonuçları



Değerlendirme

Uygulama Sonucu

52

- Test sonuçları incelendiğinde, tek kelime hatası içeren cümlelerde başarılı sonuçlar elde edilirken cümle içerisindeki hatalı kelime sayısındaki artış başarı oranını düşürmektedir. Bu durum artan karmaşıklık ve cümledeki gerçek anlam belirsizleşmesi ile açıklanabilir.

ZEMBEREK İLE KARŞILAŞTIRMA

Zemberek nedir?

54

- ▶ Zemberek, Türkçe dili için Doğal dil işleme alanında çalışmak isteyen herkes için kullanımı kolay ve açık bir kütüphanedir.



TEXT CORRECTION



güzel bit araba gördüm

mesaj girin

öneriler

```
güzel bir araba gördüm %58 [7, 341, 4]
güzeli bir araba gördüm %58 [7, 341, 4]
güncel bir araba gördüm %57 [0, 341, 4]
güncel bir arama gördüm %58 [0, 341, 8]
güncel bir araca gördüm %65 [0, 341, 50]
güncel bir arada gördüm %65 [0, 341, 50]
güzel bit araba gördüm
```

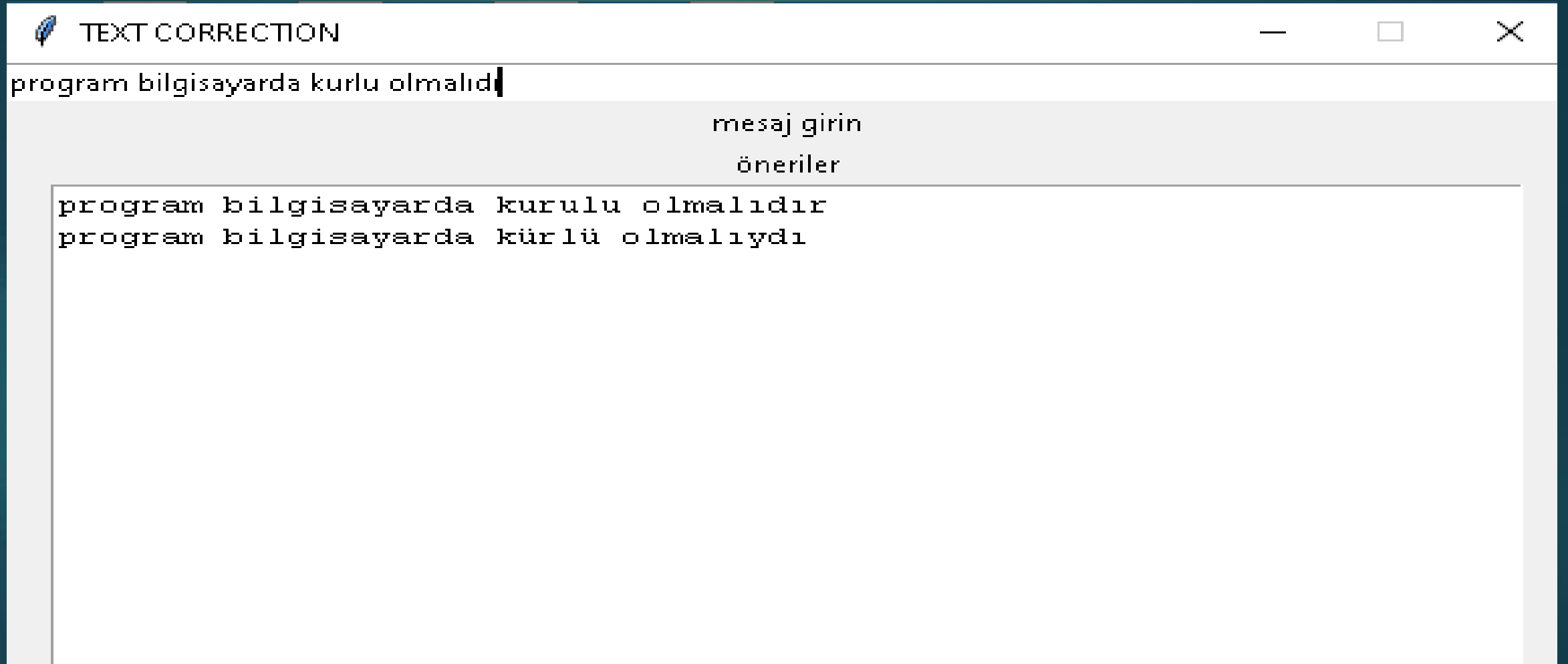



uygulama beta sürümünü yayınladı

mesaj girin

öneriler

```
uygulama beta sürümünü yayınladı %6 [5, 1, 1]  
uygulama beta sürümünde yayınladı %6 [5, 1, 1]  
uygulama beta sürümünü yayınladı
```



Sonuç

- Bir kelimesi hatalı olan 10 cümle üzerinden bir karşılaştırma yapıldığında uygulamanın Zemberek kütüphanesine göre %50 oranında başarılı olduğu gözlemlenmiştir.

Sonuç(2)

59

- ▶ Önceki çalışmalar incelendiğinde tamamının cümledeki yanlış yazılmış kelimeleri esas alarak düzeltmeye çalıştığı görülmektedir.
- ▶ Çalışmamızda farklı olarak, cümle içerisinde doğru yazılmış fakat cümlenin genel anlamına uygun olmayan kelimelerin de incelenerek düzeltilmesi hedeflenmiştir.
- ▶ Doğru olarak verilen bir cümle analiz sonucu yanlış bir cümleye dönüşebilir.

Öneri:


60

- ▶ Bu projede veri tabanları esas alınarak istatistiksel çıkarımlarla sonuca ulaşılmaya çalışılmıştır.
- ▶ Doğal dil çalışmaları, kelimelerin ve cümle içindeki bağlamlarının incelendiği bir alandır. Bu alan, uzun zamanlar ve geniş ekipler gerektiren derin araştırmalarla ilerlemektedir. Bu tür çalışmalarla daha doğru sonuçlar elde edilebileceği öngörülmektedir.

- ▶ Çalışmamızda işlem kapasitesi kısıtlı bir bilgisayar ile bu sonuçlar elde edilmiştir.
- ▶ Bazı etmenler artırılarak daha başarılı sonuçlar elde edilebilir.

Dil yaşıyan bir varlıktır ve binlerce yıldır gelişmektedir. Böyle bir yapının çözümlenmesindeki başarı, katılımcı sayısı ile doğru orantılı olarak artacaktır.





Veri analizi için güçlü makinelerin kullanılması da başarıyı olumlu derecede etkileyecektir.

TEŞEKKÜRLER