

# Dođal Dil İřleme (Natural Language Processing -NLP)

**Dr. Fatih KALEMKUŐ**

*Kafkas Üniversitesi*

# Dil Nedir ?

“Sözcük ve cümle birimleri aracılığıyla, düşünceyi konuşmayla ilişkilendiren çok seviyeli bir sistemdir”

*N.Chomsky*

İnsanlar arasında bir iletişim aracıdır.

Dilin bilgisayar ortamında modeli oluşturulursa iletişim için önemli bir araç elde edilmiş olur.

# Dođal Dil İřleme - NLP

- Dođal Dil İřleme, NLP (Natural Language Processing) olarak bilinen Yapay Zeka ve Dil Biliminin bir alt kategorisidir.
- Trke, İngilizce, Almanca, Fransızca gibi dođal dillerin (insana zg tm diller) iřlenmesi ve kullanılması amacı ile arařtırma yapan bilim dalıdır.

**Dil bilimi** veya **Lengistik**, insan dilinin ilmi arařtırmasıdır. Lisanların geliřmesini, aralarındaki bađları ve dnya zerinde dađılımını arařtırır. Bu arařtırmayı yrtene *lengist* denir.

Hedefi, insanın kendisi ve dnyası hakkında bilgi edinmek, bilgiyi depolamak ve ulařtırmaktır.

# Uzman Sistemler ve Doğal Dil İşleme

**NLP-Doğal Dil İşleme**, doğal dillerin kurallı yapısının çözümlenerek anlaşılması veya yeniden üretilmesi amacını taşır.

Bu çözümlemenin insana getireceği kolaylıklar,

- *Sözcük işlemci (word processing)*
- *Yazılı dokümanların bir dilden diğer bir dile yarı otomatik olarak çevrilmesi*
- *Soru-cevap makineleri (bir veri tabanına SQL ile değilde, bir doğal dil ile sorgu yöneltme ve sistemin bunu çözümleyerek bir SQL sorgusuna çevirdikten sonra sonuçları kullanıcıya vermesi)*
- *Bilgisayar yardımıyla dil öğretmek,*
- *Çok ve tek dilli sözlüklere erişmek*
- *Doğal dilde cümle ve metin üretmek*
- *Metin özetleme*
- *Otomatik konuşma ve komut anlama*
- *Konuşma sentezi*
- *Konuşma tanıma ve üretme*
- *Bilgi sağlama*

gibi birçok başlıkla özetlenebilir.

# Uzman Sistemler ve Dođal Dil İşleme

- Bilgisayar teknolojisinin yaygın kullanımı, bu başlıklardan üretilen uzman yazılımların gündelik hayatımızın her alanına girmesini sağlamıştır.
- Örneđin, tüm kelime işlem yazılımları birer imla düzeltme aracı taşır. Bu araçlar aslında yazılan metni çözümleyerek dil kurallarını denetleyen **dođal dil işleme** yazılımlarıdır.
- Konuşma ve komut anlama yazılımları ile insan ve bilgisayar arasındaki klavye, fare gibi veri girişı aygıtları ortadan kalkacaktır.

# Dođal Dil İřleme (NLP) Nedir?

DDİ, ana işlevi bir dođal dili çözümleme, anlama, yorumlama ve üretme olan bilgisayar sistemlerinin tasarımını ve gerçekleştirilmesini konu alan bir mühendislik dalıdır.

Sabit algoritmalar içermediğinden ve belirsizliklere sahip olduğundan bir NP problemidir.

Yapay zeka, biçimsel diller kuramı, kuramsal dilbilim, bilgisayar destekli dilbilim ve bilişsel psikoloji gibi deđişik alanlarda geliştirilmiş kuram, yöntem ve teknolojiler bütünüdür.

# Niçin Doğal Dil İşleme ?

Tür, cinsiyet, sahiplik(yazar)

- Büyük miktarlarda veri
  - Internet
  - Intranet
- Çok fazla sayıdaki dokümanların işlenmesi

DDİ'de uzmanlık gerektirir

- Dokümanların kategorilerine göre sınıflandırılması
- Dokümanlarda arama ve indeksleme
- Otomatik çeviri
- Konuşma anlama
  - Telefon konuşmalarını anlama
- Bilgi çıkarılması
  - Özetleme
  - Çıkartma
- Otomatik yazma
  - Kitabın özetini yazdırma
  - Yorum yazdırma
- Soru cevaplama
  - Sorulara cevap yazdırma
- Bilgi elde etme
- Text ve diyalog üretmek

DDİ ile bir soru yöneltildiğinde sistem bunu çözümler ve SQL sorgusuna dönüştürüp işler sonra kullanıcıya cevap döndürür

# Dođal Dil Alanındaki Temel Arařtırmalar

- Dođal dillerin iřlev ve yapısının daha iyi anlaşılması
- Bilgisayar ve insanlar arasında arabirim olarak dođal dili kullanmak ve aradaki iletiřimi kolaylařtırmak
- Bilgisayar yardımıyla bir dilden diđerine çeviri yapmak

Japonya, Almanya, İngiltere, ABD, Hollanda gibi ülkelerde bu alanda yazılımlar geliştirilmiř

Bilim ve iř alanındaki geçerli dil İngilizce

Türkçe'deki çalışmalar yetersiz kalmaktadır

# Dođal?

- Dođal Dil ?
  - İnsanlar tarafından konuşulan diller, İngilizce, Japonca, Türkçe, vs., buna karşılık yapay diller, C++, Java, vs.
  - 3000 ile 4000 arasında deđişik dil var
  - UNESCO tarafından 6 tanesi resmi dil olarak kabul edilmiştir (Çince, İngilizce, İspanyolca, Rusça, Fransızca ve Arapça)
  - Türk dili ve lehçeleri
  - Çok dillilik ve iletişim güçlüğü yapay dillerin doğmasına neden olmuştur
  - Yapay dillerin en tanınmışısı Polonyalı *L.L. Zamenkov*'un ortaya attığı *Esperanto*'dur
  - Bilim ve iş dünyasının dili İngilizce
  - Türkiye Cumhuriyetleri'nde Türkiye Türkçesi önemli bir yer tutmaktadır

# Niçin Doğal Dil İşleme?

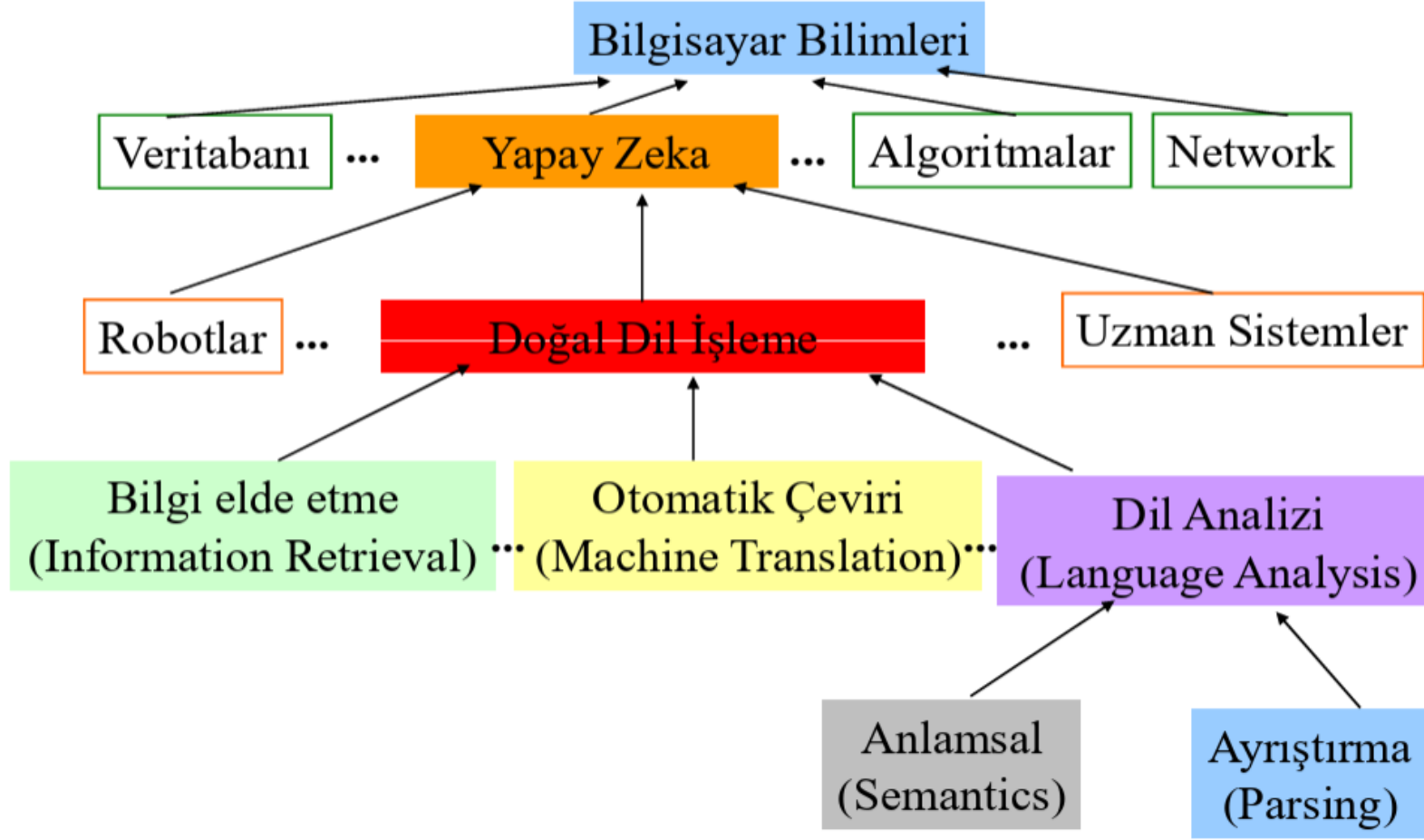
- kJfmmfj mmmvvv nnnffn333
- Uj iheale eleee mnster vensi credur
- Baboi oi cestnitze
- Coovoel2^ ekk; ldsllk lkdf vnnjfj?
- Fgmflmlk mlfm kfre **xnnn!**

# Niçin Doğal Dil İşleme?

- Bilgisayarlar doğal dilde yazılmış bir dokümanı bizim bir önceki slaytı gördüğümüz gibi görür !
- İnsanların bir dili anlaması zor değildir
  - Sağduyuya sahip
  - Mantıklı düşünebilme kapasitesi (reasoning capacity)
  - Deneyim
- Bilgisayarlar ise
  - Sağduyuya sahip değil
  - Mantıklı düşünemez

**Biz onlara öğretmediğimiz sürece!**

# NLP'nin bilgisayar bilimindeki yeri neresidir ?



# Analizin Dilbilimsel Seviyesi

- Konuşma
- Yazım Dili
  - Sesbilim (phonology): sesler / harfler / telaffuz
  - Biçimbilim (morphology): kelimenin yapısı
  - Sözdizim (syntax): cümlenin anlamını oluşturan birimlerin hiyerarşik bir yapıda ifade edilmesi
  - Anlamsal (semantic): cümlenin anlamı
- Seviyeler arasındaki etkileşim

# Biçimbilim-Morphology

## **Örnek: çocukları**

Çocuk +İsim+ Çoğul+ 3.tekil kişi iyelik

(Sevgi'nin çocukları Ayşe ve Mehmet geldiler.)

çocuk+İsim+ Çoğul+-i hali

(Yeni gelen çocukları gördünüz mü?)

çocuk+İsim+ Çoğul+ 3. çoğul kişi iyelik

(Ayşe ile Mehmet'in çocukları Gökhan ile Sevgi'dir.)

çocuk+İsim+ Tekil+ 3. çoğul kişi iyelik

(Ayşe'nin çocukları Gökhan ile Sevgi'dir).

# Sözdizim-Syntax

*“the dog ate my homework”* - Who did what?

1. Part of speech tagging (POS etiketleri)  
belirlenmesi

Dog = noun ; ate = verb ; homework = noun

2. Identify collocations

mother in law, hot dog

Birleşik isimler (kitap kurdu)

# Sözdizim-Syntax

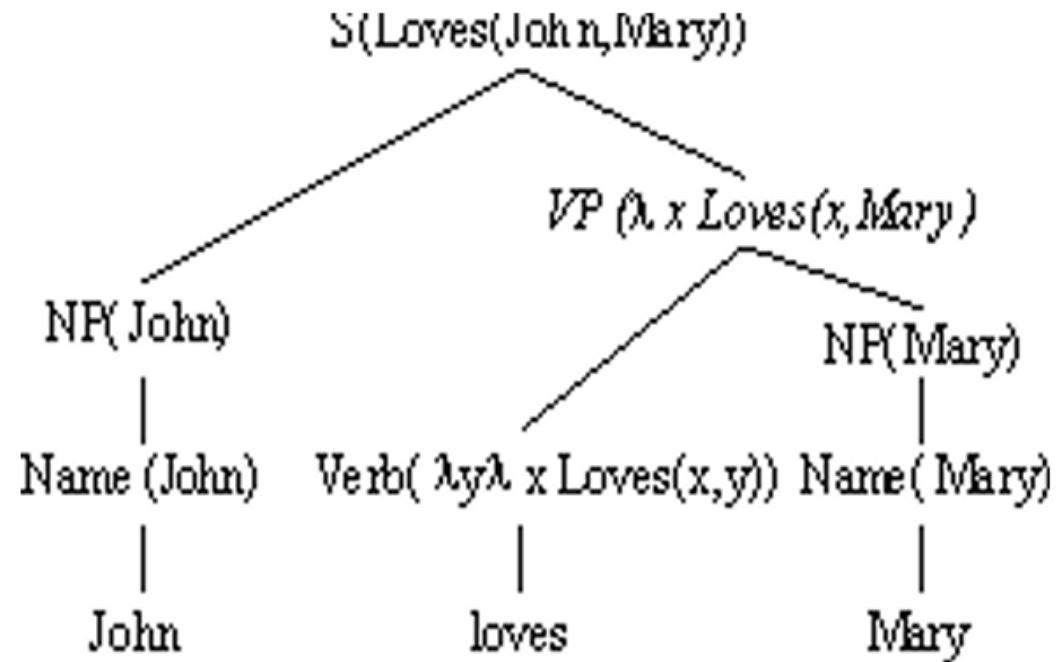
- Yüzeysel ayrıştırma:  
“the dog chased the bear”  
“the dog” “chased the bear”  
özne - yüklem ile ilgili olan

Temel yapının belirlenmesi

NP-[the dog] VP-[chased the bear]

# Sözdizim-Syntax

- Tam ayrıştırma: John loves Mary



# Sözdizim-Syntax

- Anaphora Ayırıştırma (anaphora resolution):

*“The dog entered my room. It scared me”*

*“Köpek odama girdi ve beni ısırdı”*

- Edat ekleme (preposition attachment)

*“I saw the man in the park with a telescope”*

# Anlamsal-Semantics

- Doğal dili anlamak ! Ama nasıl?
- “*plant*” = *industrial plant*
- “*plant*” = *living organism*
- Kelimelerdeki belirsizlikler
- Anlamsal analizin önemi ?
  - Machine Translation: hatalı çeviri
  - Information Retrieval: hatalı bilgi
  - Anaphora Resolution: hatalı referans

# Niçin Anlamsal Analiz ?

- The sea is home to million of plants and animals
- English → French [commercial MT system]
- Le mer est a la maison de billion des usines (fabrika) et des animaux
- French → English

# Niçin Anlamsal Analiz ?

- Kelimenin anlamını nasıl öğreniriz ?
- Sözlük kullanarak:

plant, works, industrial plant -- (buildings for carrying on industrial labor; "they built a large plant to manufacture automobiles")

plant, flora, plant life -- (a living organism lacking the power of locomotion)

They are producing about 1,000 automobiles in the new plant

The sea flora consists in 1,000 different plant species

The plant was close to the farm of animals.

Word Sense Disambiguation (Kelime Anlamını Berraklaştırma)

# Niçin Anlamsal Analiz ?

- Etiketlenmiş örneklerden öğrenme:
  - İçerisinde “plant” geçen 100 örneğin elle etiketlendiğini varsayalım
  - Öğrenme algoritmalarıyla sistemi eğitelim (machine learning alg.)
  - Sistemin duyarlılığını kontrol edelim

İngilizce çalışmalardaki başarı 60%-70%-(80%)

# Bilgiyi Elde Etme-**Information Retrieval**

- Genel model:
  - Çok fazla sayıda doküman
  - Sorgu
- Görev: Verilen sorgu ile ilgili dokümanları bulma  
Nasıl? İndeks yarat, bir kitabın indeksi gibi
- Sonra ...
  - Vektörel modeller (vectorial models)
  - Boolean modeller
- Örnek: Google, Yahoo, Altavista, vs.

# Bilgiyi Elde Etme-**Information Retrieval**

- **İndekslemenin anlamı**
- (=living organism) anlamını taşıyan “plant” kelimesi aranırken içerisinde (=industrial plant) anlamına gelen “plant” kelimesinin geçtiği dokümanların gelmemesi
- Fakat “flora” veya ilgili bir başka kelimenin yer aldığı dokümanların arama sonucunda getirilmesi
- **Index parsed relations**

# Bilgi Çıkarımı- **Information Extraction**

- “There was a group of about 8-9 people close to the entrance on Highway 75”
- Who? “8-9 people”
- Where? “highway 75”
  
- İstenilen bilgiyi çıkarma
- Yeni kalıplar (patern) bulmak
  - Saklı bilgi, vs.
- US-Gov./mil. Milyonlarca dolar harcamaktadır IE araştırmalarına

# Bilgi Çıkarımı- **Information Extraction**

- Özel bir bilgininde getirilmesi istenebilir
- Soru Cevaplama (question answering)  
“What is the height of mount Everest?”  
11,000 feet  
Current state-of-the-art 40-50%

Belirlenmiş özel bir alanda soru cevap yapmak

# Bilgi Çıkarımı- **Information Extraction**

- Karşı dilde bilgiyi bulma!
- **Cross Language Information Retrieval**
- “What is the minimum age requirement for car rental in Italy?”
- İtalyanca text’lerde de arama yapabilmek için cümle İtalyancaya çevrilir. “eta minima per noleggio macchine”

# Makine Çevirisi-**Machine Translations**

- Text to Text Machine Translations
- Speech to Speech Machine Translations
- Bu tip çalışmalar yaygın olan dil çiftleri için yapılmıştır

İngilizce-Fransızca, İngilizce-Çince

# Makine Çevirisi-**Machine Translations**

- Text bir dilden diğere nasıl çevrilir ?
- Önceden yapılmış olan çeviriler sisteme öğretilir
- → **Paralel bir külliyata ihtiyaç vardır**
- Fransızca-İngilizce, Çince-İngilizce
- Makul çeviriler
- Çince-Hintçe – günümüzde uygun bir külliyat yoktur!

# Sonuç



*Dr. Fatih KALEMKUŞ*

# Sorular



*Dr. Fatih KALEMKUŞ*

# TEŐEKKÜRLER

*Dr. Fatih KALEMKUŐ*