Doğal Dil İşleme (DDİ)
Natural Language Processing (NLP)

Prof.Dr. Banu Diri

Konular

- Doğal Dil İşlemeye Genel Bakış (Course Overview)
- *Dilbiliminin Esasları (Linguistics Essentials)**
- Gramer ve Diller (Grammer and Language)
- Düzenli Diller (Regular Expression)
- Dil Modelleri N-Grams (Language Models)
- Biçimbilimsel Analiz (Morphological Analysis)
- Sözdizimsel Analiz-POS (Syntax Analysis-Part of Speech Tagging)
- Anlam Bilgisi (Semantik)- Söylem (Discource) Bilgisi-Edim (Pragmatic) Bilgisi
- Eşdizimlilik (Collocation)
- *HHM, Viterbi Algoritması**

**** Ek bilgi***

Konular

- Makine Öğrenmesi (Machine Learning)
- Derin Öğrenme (Deep Learning)
- Metin Sınıflandırma (Text Classification)
- Bilgiye Erişim Sistemleri (Information Retrieval)
- Bilgi Çıkarımı (Information Extraction)
- Kelime Anlamları (Word Semantic)
- Kelime Gömmeleri (Word Embeding)
- Duygu Analizi (Sentiment Analizi)
- Soru Cevaplama Sistemleri (Question Answering)
- Görsel için Başlık Üretme (Image Captioning)
- *Machine Translation (Makine Çevirisi)*
- Projeler, Araştırma Ödevi, Seminer

Kaynaklar

- Speech and Language Processing: An Introduction to Natural Language Processing, Coputational Linguistics and Speech Recognition, *D.Jurafsky and J. Martin*
- Foundations of Statistical Natural Language Processing, *C. Manning and H. Schutze*
- Statistical Language Learning, *Eugene Charniak*
- and *INTERNET*

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.

- Doğal Dil İşleme, NLP (Natural Language Processing) olarak bilinen Yapay Zeka ve Dil Biliminin bir alt kategorisidir.

- Türkçe, İngilizce, Almanca, Fransızca gibi doğal dillerin (insana özgü tüm diller) işlenmesi ve kullanılması amacı ile araştırma yapan bilim dalıdır.

Dil bilimi veya **Lengüistik**, insan dilinin ilmi araştırmasıdır. Lisanların gelişmesini, aralarındaki bağları ve dünya üzerinde dağılımını araştırır. Bu araştırmayı yürütene *lengüist* denir.

Hedefi, insanın kendisi ve dünyası hakkında bilgi edinmek, bilgiyi depolamak ve ulaştırmaktır.

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,

- *Yazım yanlışlarının düzeltilmesi (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*
- *Konuşmayı metne çevirme (STT)*
- *Konuşma tanıma ve üretme (TTS)*
- *Metnin içerdiği bilgiyi çıkarma*
- *Bilgiyi çekme*

gibi birçok başlıkla özetlenebilir.

- Bilgisayar teknolojisinin yaygın kullanımı, bu başlıklardan üretilen uzman yazılımların gündelik hayatımızın her alanına girmesini sağlamıştır.

- Örneğin, tüm kelime işlem yazılımları birer imla düzeltme aracı taşır. Bu araçlar aslında yazılan metni çözümleyerek dil kurallarını denetleyen **doğal dil işleme** yazılımlarıdır.

- Konuşma ve komut anlama yazılımları ile insan ve bilgisayar arasındaki klavye, fare gibi veri girişi aygıtları ortadan kalkacaktır.

Karşılaşılan zorluklar nelerdir

- Kuralsız ve anlaşılmaz konuşmalar *(Napıyon len?)*
- Kuralsız ve bozuk yazılar *(kelebkler)*
- Konuşmayı bölme *(iki cümle arasında duraklamadan konuşmak)*
- Metni bölme *(paragraf uzunluğunda cümleler yazmak)*
- Anlam belirsizliklerini giderme *(köprücüler İstanbul'da toplanıyor) (bridge-köprü-briç)*
- Söz dizimsel belirsizlikleri giderme

(Banu armutları ayılara aç oldukları için verdi

Banu armutları ayılara tatlı oldukları için verdi)

Doğal Dil İşleme Nedir ?

DDİ, ana işlevi bir doğal 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 pisikoloji gibi değişik alanlarda geliştirilmiş kuram, yöntem ve teknolojiler bütünüdür.

Niçin D

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 çı
  - Özg
    çıka
- Otomat
  - Kitabın bir yoğunlaşmak
- Soru cevaplam
- Bilgi elde etme
- Text ve diyalog üretmek

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

Doğal dil alanındaki temel araştırmalar

- Doğal dillerin işlev ve yapısının daha iyi anlaşılması
- Bilgisayar ve insanlar arasında arabirim olarak doğal dili kullanmak ve aradaki iletişimi kolaylaştırmak
- Bilgisayar yardımıyla bir dilden diğerine çeviri yapmak

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

Bilim ve iş alanındaki geçerli dil İngilizce

Türkçe'deki çalışmalar yetersiz kalmaktadır

Doğal?

- Doğal Dil ?
  - İnsanlar tarafından konuşulan diller, İngilizce, Japonca, Türkçe, vs., buna karşılık yapay diller, C++, Java, vs.
  - 3000 ile 4000 arasında değişik dil var
  - UNESCO tarafından 6 tanesi resmi dil olarak kabul edilmiştir (Çince, İngilizce, İspanyolca, Rusça, Fransızca ve Arapça)
  - Türk dili ve lehçeleri *(5. sırada yer alır)*
  - Çok dillilik ve iletişim güçlüğü yapay dillerin doğmasına neden olmuştur
  - Yapay dillerin en tanınmışı 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 ?

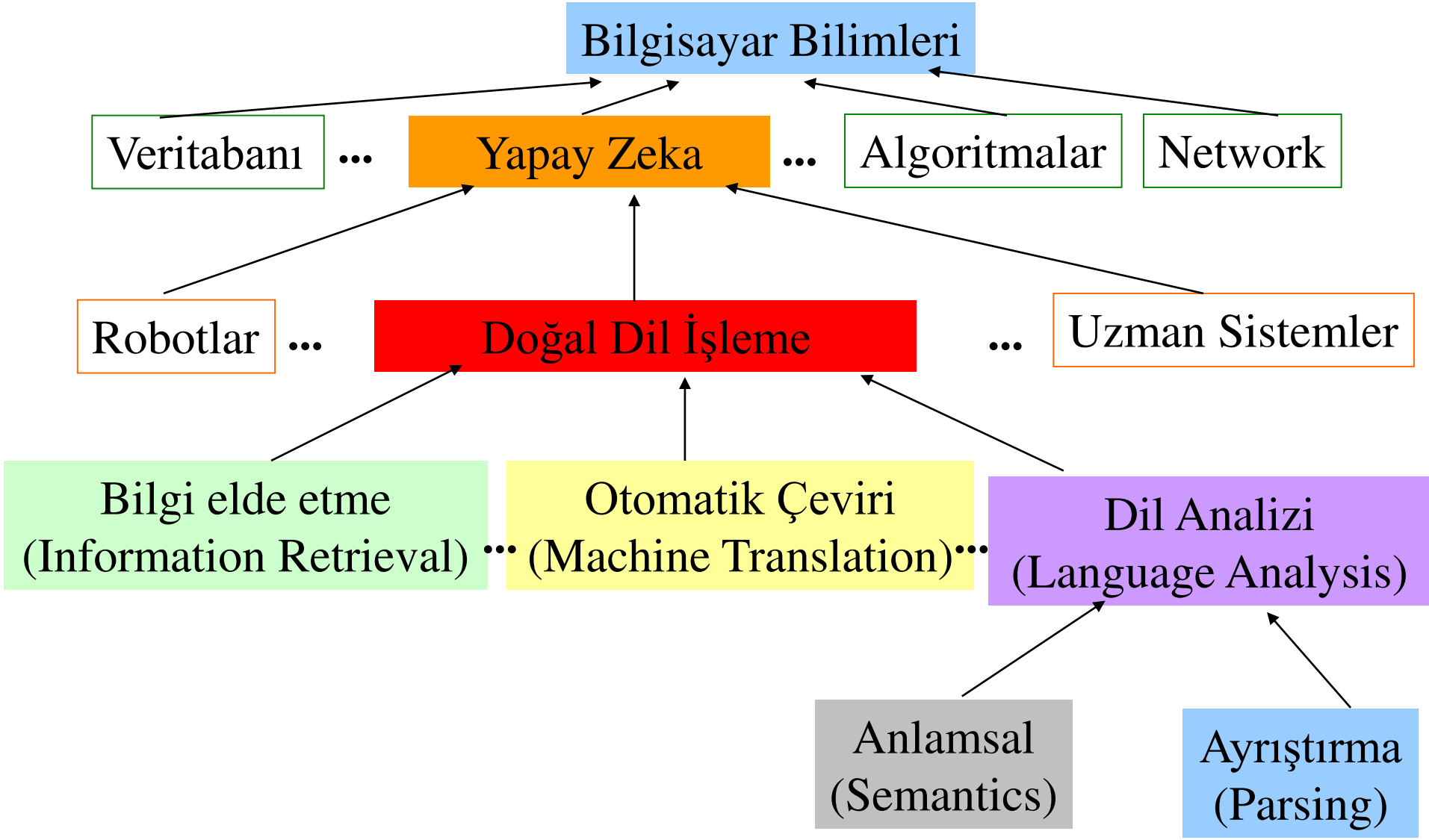
- kJfmmfj mmmvvv nnnffn333
- Uj iheale eleee mnster vensi credur
- Baboi oi cestnitze
- Coovoel2^ ekk; ldsllk lkdf vnnjfj?
- Fgmflmllk mlfm kfre xnnn!

!!!

- 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!

DDİ'nin bilgisayar bilimindeki yeri neresidir ?



Analizin dilbilimsel seviyesi

- Konuşma Dili
- 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"

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)

…

- Yüzeysel ayrıştırma:

  “the dog chased the bear” *(köpek ayıyı kovaladı)*

  “the dog” “chased the bear”

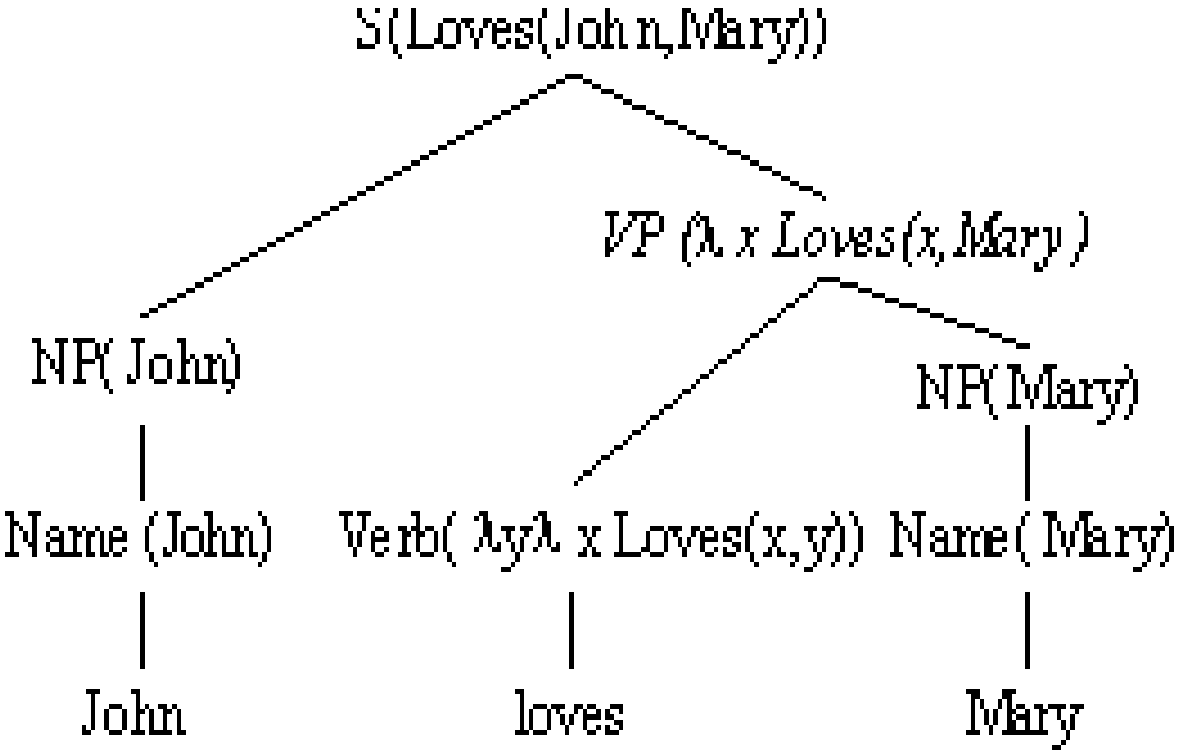
  özne - yüklem ile ilgili olan

  Temel yapının belirlenmesi

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

…

- Tam ayrıştırma: John loves Mary



…

- Zamir Çözümleme (anaphora resolution)

"The dog entered my room. It scared me"

"Köpek odama girdi ve beni korkuttu"

- Edat ekleme (preposition attachment)

"I saw the man in the park with a telescope"

Anlamsal-Semantic

Doğal dili anlamak ! Ama nasıl?

- Kelimelerdeki belirsizlikler

  "plant" = industrial plant

  "plant" = living organism

- Anlamsal analizin önemli mi?
  - 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

...

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.

...

- 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.

…

Indekslemenin 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

…

- 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

…

- 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

…

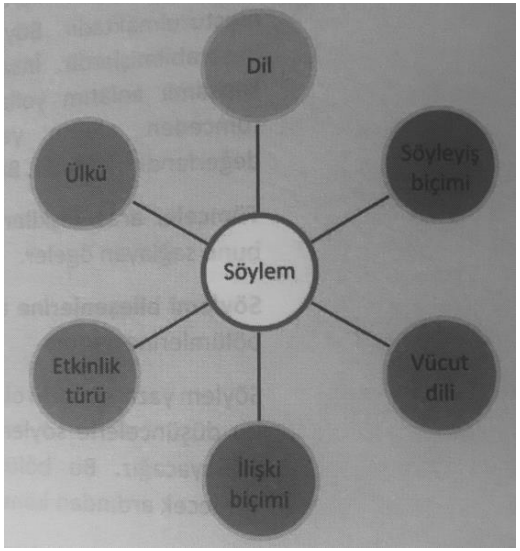
- Text bir dilden diğerine nasıl çevrilir ?
- Önceden yapılmış olan çeviriler sisteme öğretilir
- → Paralel bir külliyata ihtiyaç vardır
- Fransızca-İngilizce, Çince-İngilizce
- Makul çeviriler
- Çince-Hintçe – günümüzde uygun bir külliyat yoktur!

Söylem Bilimi-Discource

Sözcükler→Tümceler→Paragrafar→Dokümanlar

Birden fazla tümceden oluşan yazılı veya sözlü söylemleri inceler

- Tümceler arası ilişkiler çıkarılır
- Söylemi, başlık-giriş-gelişme-sonuç kısımlarına ayırma
- Bir söylemin etkili olması şartları



- Yazılı, sözel, elektronik söylem

15.03.2019 Garipçe köyü

«Mayo ve bikini ile denize girmek yasaktır»
Anlam 1: Bu köyde mayo ve bikini ile denize girilmesi yasaktır.
Çıplak girilmelidir

Anlam 2: Bu köyde denize elbise ile girilir
Anlam 3: Bu köyde plaj kıyafeti ile denize girilmesi yasaktır.
Burası plaj değildir.

Edim/Kullanım Bilimi-Pragmatic

- Sözdizimi ve Anlam bilimi tümce bazında çalışır
- Söylem bilimi ise birden çok tümce üzerinde çalışır
- Tümcelerin tek tek anlamları ile ilgilenmek yerine metni anlar ve yorumlar
- Sözcük ve tümceleri kullanıldıkları bağlam içerisinde değerlendirir

Bir yolcu ile yolda giden bir kişi arasında aşağıdaki konuşma geçmiş olsun…

- Metro istasyonu nerede, biliyor musunuz?

- Evet biliyorum *(der ve yürümeye devam eder)*

Her iki tarafından sorunun cevabı konusunda beklentisi farklıdır

Ev sahibi : Çocuğunuz var mı?
Kiracı : 10 yaşında bir oğlum var
Ev sahibi : Allah bağışlasın
Kiracı : Bir de küçük köpeğim var
Ev sahibi : Bu kötü

Eğer konuşmanın bağlamının ev sahibi-kiracı arasında olduğu bilinmez ise adamın köpeğinin olmasına kötü denmesi anlaşılmaz