DİL BİLİMİ

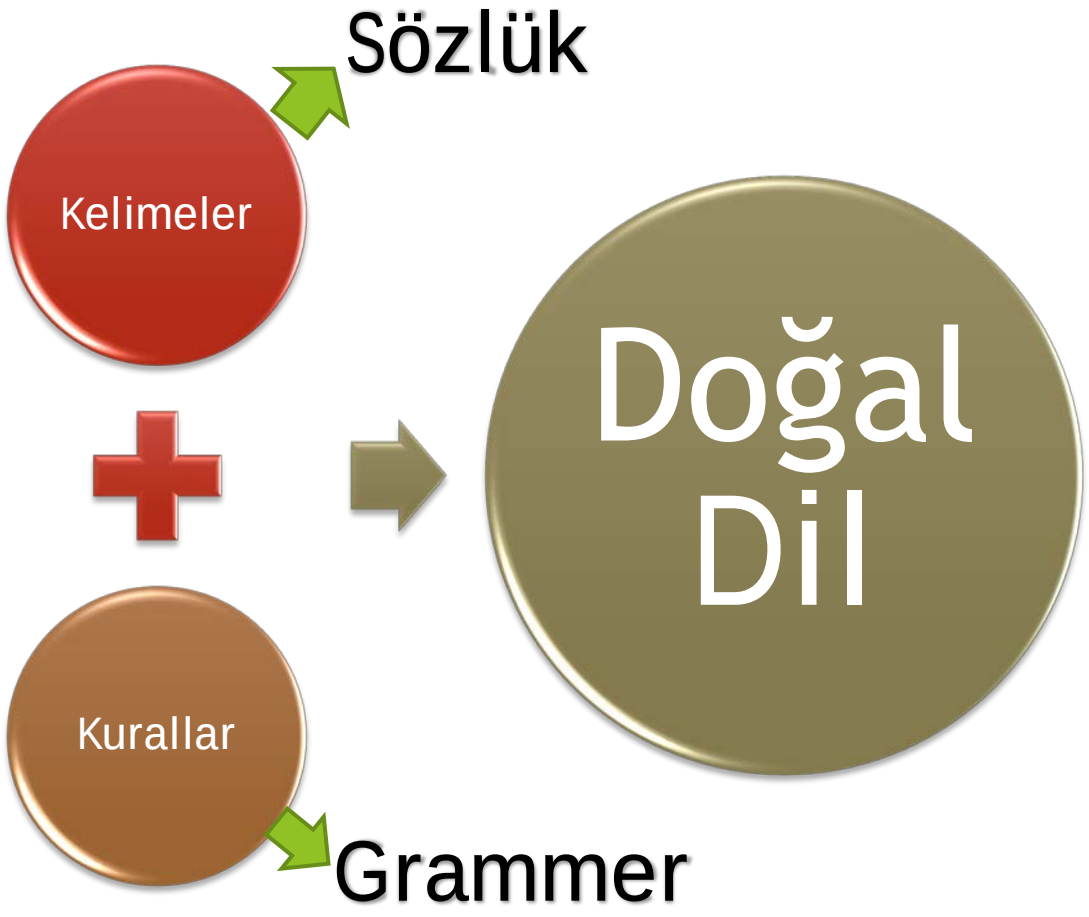
Doç. Dr. PINAR CİHAN

İçindekiler

1-Doğal Dil

- Doğal dil;
  - İnsanların **birbirleri ile** iletişim kurmasını sağlayan **bir** araçtır.
- **Birleşmiş Milletler verilerine göre dünyada 4000'den fazla dil konuşulmaktadır.**
- **Türkçe sondan eklemeli (**bitişken**) bir dildir.**
- **Bir bilgisayarın insanın konuşmasını anlayabilmesi ve insanın anlayabileceği dilde konuşarak insanlarla etkileşimde bulunabilmesi için bilgisayarın dilin tüm özelliklerini bilmesi gerekir.**
  - **Yani dilin özelliklerinin bilgisayara öğretilmesi gerekir.**

1-Doğal Dil



1-Doğal Dil


Dil Bilimi
Ses Bilimi (Phonology)
Bürün Bilgisi
Biçim Bilimi (Morphology)
Söz Dizim (Syntax)
Anlam Bilim (Semantics)
Kullanım Bilgisi (Pragmatics)
Söylem Bilgisi
Edim Bilimi

1-Doğal Dil

- Her bir bileşenin bir giriş ve bir çıkışı var.
- Bir bileşenin çıkışı diğerinin girişi olabilir.
- Birbirlerinin sonuçlarından yararlanabilirler.

2-Ses Bilgisi (Phonetics-Fonetik)

- Doğal dillerde anlam ayırıcı en küçük öğeye **sesbirim (fonem)** denir.
- Kelimelerdeki vurgulardır.
- Harf; Sesbirimleri göstermek üzere oluşturulan simgeler.
- Bir harfin söyleniş biçimi kelimelere göre farklılık gösterebilir.
- İngilizcede harfler geçtiği kelimeye göre farklı vurgulanabilir.
  - Palm (p→p)
  - Phone (p→f)
- Türkçe'de ise kelimeler yazıldığı gibi okunur.
  - Ancak söylenişi farklı olabilir.
  - Örneğin 'k' harfi bazen 'ke' bazen 'ka' olarak okunur.

2-Ses Bilgisi (Phonetics-Fonetik)

- Bir dilin alfabesi oluşturulurken her bir ses birime bir harf tanımlamak amaçlanır
- Bir dilin alfabesinin tüm sesbirimlerini karşıladığı alfabelere **sesçil alfabe** denir.
  - Böyle dillere ‘Fonetik Dil’ denir.
- Ses bilimi (Phonology); dil içindeki seslerin işlevlerini inceler.
- Doğal Dil İşleme için ses biliminin
  - Giriş bilgisi; alfabe içindeki sesler,
  - Çıkış bilgisi; sıralı fonemlerdir.

3-Bürün Bilgisi

- Bürün; konuşmanın ritim, vurgu ve tonlaması ile ilgilidir.
- Bürün, konuşmacının;
  - Duygu durumu,
  - Konuşmanın söylem, soru veya emir olup olmadığı,
  - Söyleyiş biçimindeki alaycı, iğneleyici, vurgulayıcı vs olup olmadığını belirtir.
- Konuşma sırasında sesin genliği, tonu, vurgusu gibi özellikler duyguları karşıdakine aktarmayı sağlar.
- Yazılı metinler duygu belirtemez.
- Duygunun yazıya aktarılabilmesi için yapılan çalışmalar 2 açıdan ele alınmaktadır;
  - Dil bilgisi
  - Akustik

3-Bürün Bilgisi

- **Dil Bilgisi Açısından Bürün**

Vurgu	• Bir sözcük, hece veya harfin diğerlerine göre daha belirgin söylenmesi
Tonlama	• Konuşma sırasında sesin genlik düzeyindeki değişiklik.
Ezgi	• Konuşma sırasında sesteki değişim örüntüsüne denir. • Ses genliği, 1 (düşük), 2 (orta), 3, 4 biçiminde belirtilir.
Süre	• Kelime içinde bazı ünlü harflerin diğerlerine göre daha uzun okunmasıdır.
Durak	• Sözlü anlatımda vurgulanmak istenen kelimeden sonra konuşmacının bir süre susmasıdır.
Duygu	• Kişinin ses tonunun bulunduğu duygu durumuna yansıması (sakin, üzgün, korkmuş, kızgın)

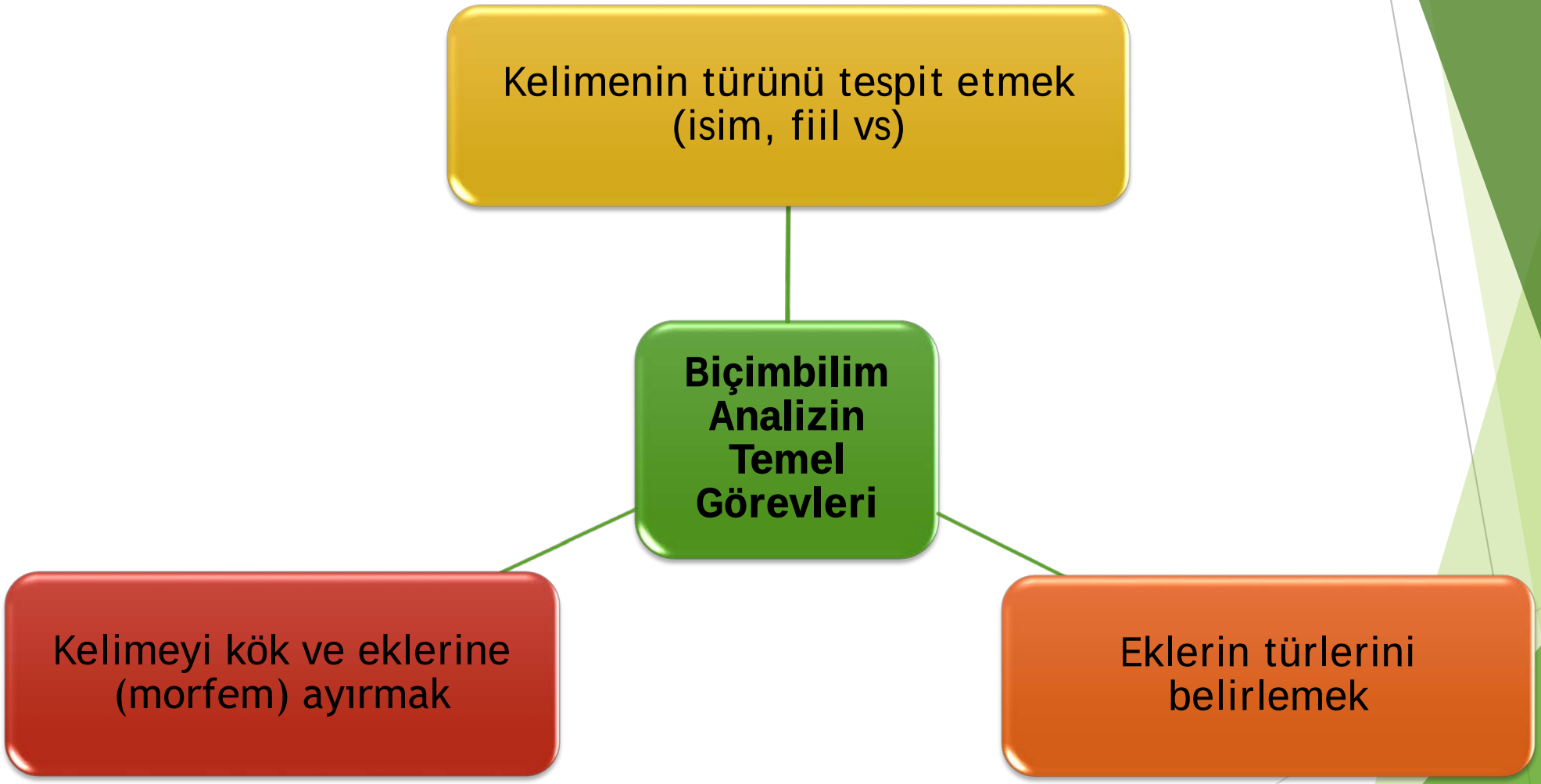
3-Bürün Bilgisi

- **Akustik Açısından Bürün**
  - Konuşma
    - Temel frekans ve konuşmacıya özgü frekans bileşenlerinden oluşur.
    - Mikrofon ile elektrik sinyallerine dönüştürülür.
  - Konuşmacının ses düzeyi duygu durumunu belirlemede etkendir.
  - Ses şiddetinin ölçü birimi desibeldir.

Biçim Birim (Morphology)

- Bir kelime parçalanırken bölünen en küçük parçalar.
- Bir kelimenin yapısının bilgisayarlar tarafından otomatik olarak çözümlenmesine **biçim birimsel analiz** denir.
- Kelime bir kök sözcüğe eklenen biçim birimlerden (morpheme) oluşur.
  - Kelimenin aldığı önekler/son ekler. (Yapım ekleri, çekim ekleri)
  - Kelime tekil mi çoğul mu
  - Kelimenin kökü

Biçim Birim (Morphology)



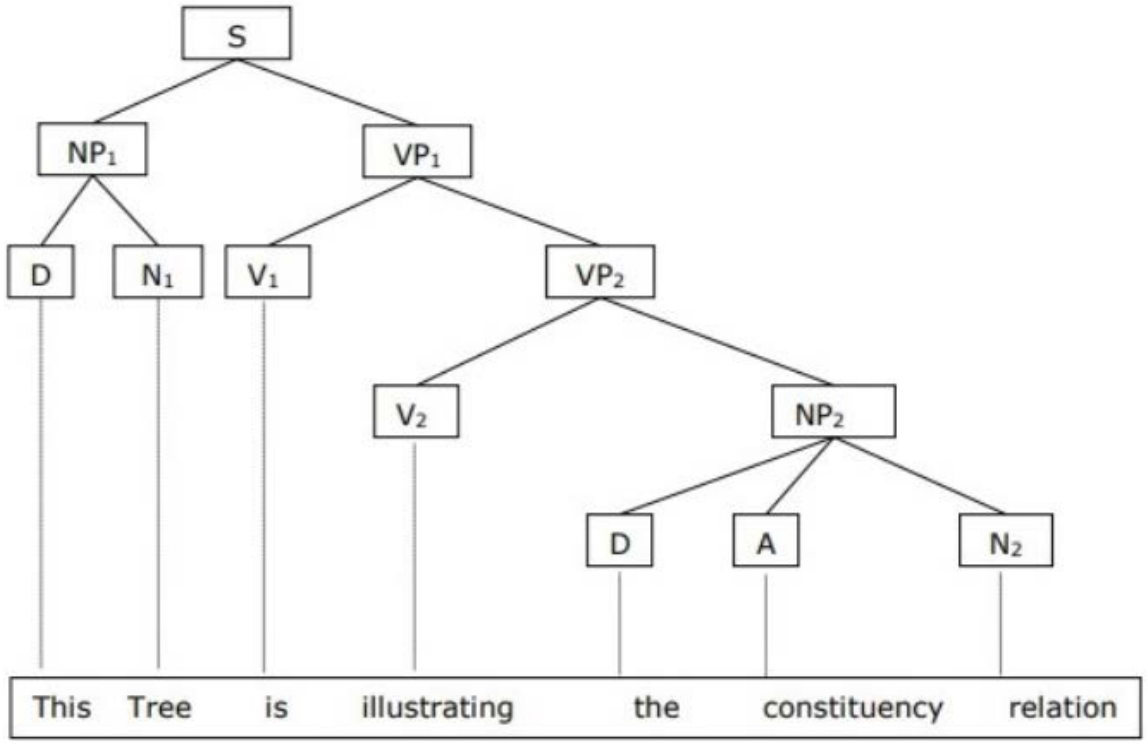
Söz Dizim (Syntax)

- Kelimelerin cümle oluşturmak için bir araya getirilme biçimi.
- Kelimelerin belirli sıralarda birleşmesine **söz dizimi** denir.
- Türkçe'de cümleler;
  - Özne, nesne ve yüklem.
- Bilgisayarın bir kelime grubunu anlamsal olarak çözümlemesi için önce cümle olup olmadığını belirlemelidir.
- Bunun için de cümlenin öğelerini tanımlaması gerekir.
- Girdi olarak cümleleri alır, çıktı olarak cümlelerin POS (**Part of Speech)** taglerini verir.

Not: POS Tagging (Kelime Türü Etiketleme): Metinde yer alan kelimelerin türlerinin (isim, fiil, sıfat, bağlaç vb.) etiketlenmesidir.

Söz Dizim (Syntax)

- Kelimelerin cümledeki anlamlarına ayrıştırılmasına '**Part of Speech Tagging**' denir.
- Kelime grupları, cümlenin öğeleri vs için tanımlı etiketler vardır.
  - Özne (Ö), Yüklem (Y), Nesne (N), İsim (İ), Sıfat (S), İsim grubu (İG) vs gibi
- Bu işlem sonucunda cümleler bir **ayrıştırma ağacı (parse tree)** de temsil edilebilir.



Söz Dizim (Syntax)

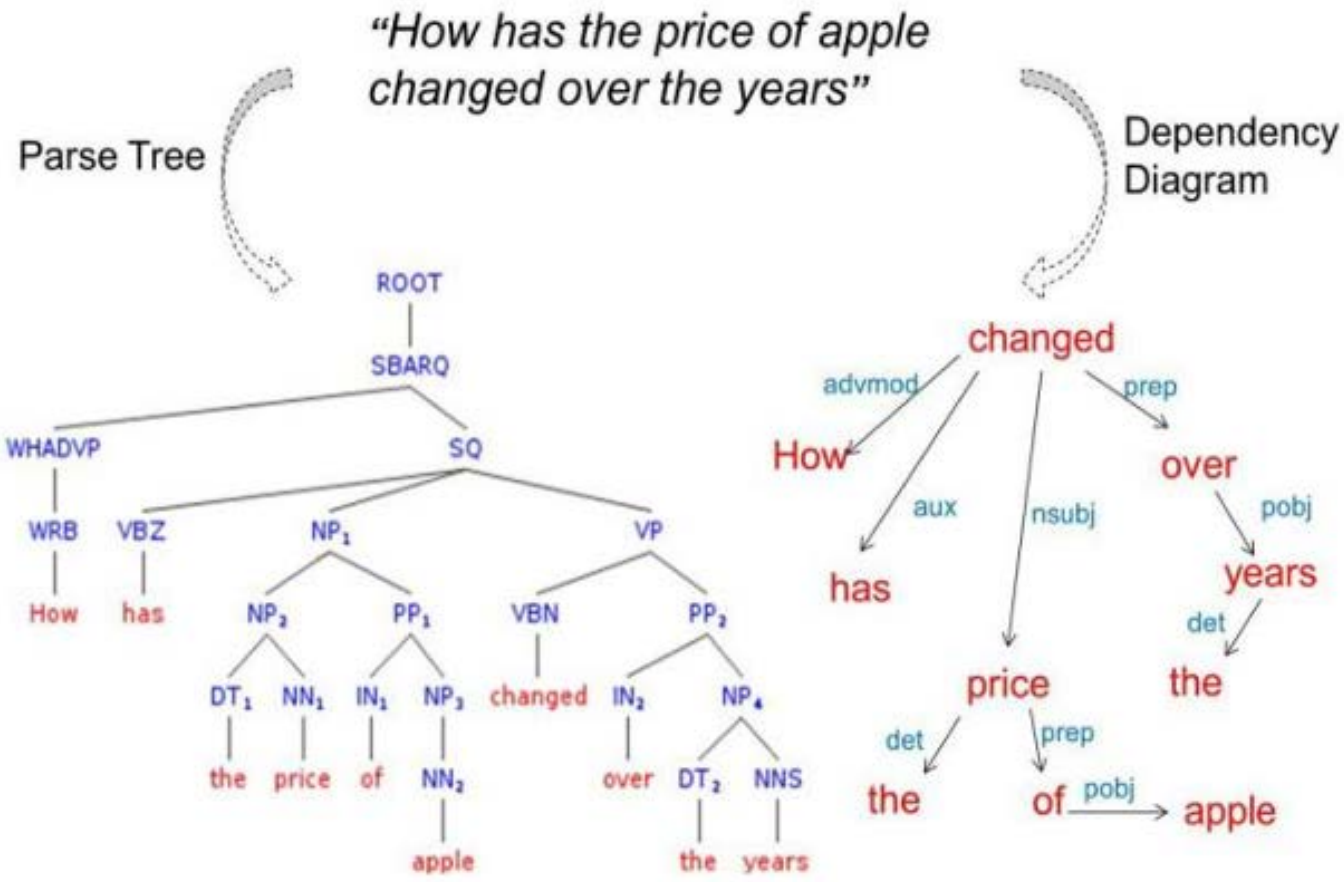
- Özne yüklem vs gibi türlerin bir araya gelme şekli ve sırası dilin kurallarına bağlıdır.
- Bu nedenle ayrıştırma ağaçları dil modellerine bağlıdır.

- **Morfolojik analiz kelime düzeyinde, söz dizimsel analiz cümle düzeyinde yapılır.**

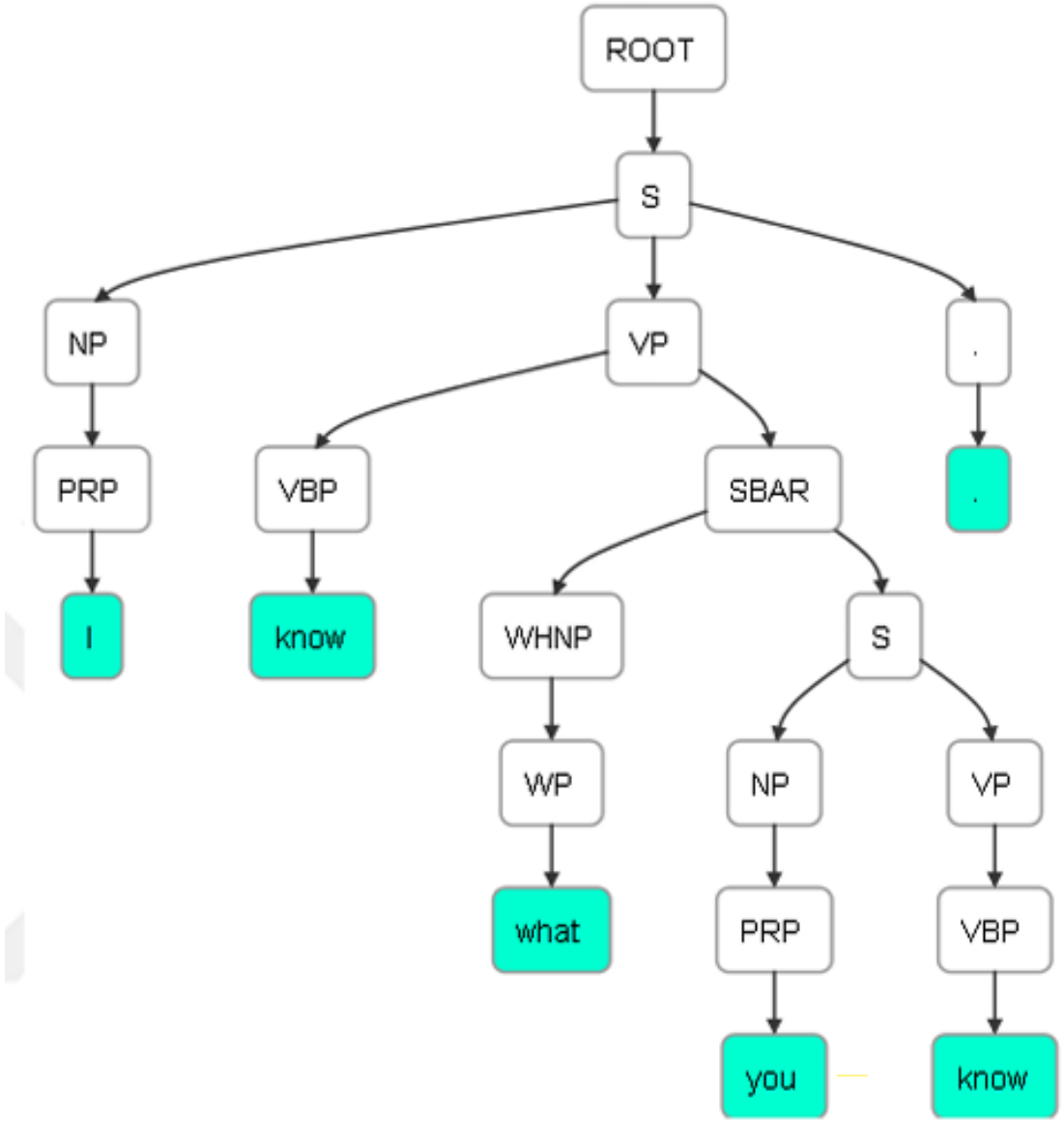
Söz Dizim (Syntax)

Doğal dildeki metinlerin dilbilimsel analizlerinin daha iyi takip edilebilmesi amacıyla görselleştirme araçları geliştirilmektedir. Bunun için bir çözümleyiciye (biçimbirimsel, sözdizimsel veya anlambilimsel) ve onun analizlerini yorumlayarak bir diyagrama dökecek olan görselleştirme programına ihtiyaç duyulmaktadır. Bu tür görselleştirme uygulamalarının en bilinenlerine örnek olarak bağımlılık diyagramı (dependency diagram) ve ayrıştırma ağacı (parse tree) verilebilir.

Doğal dil işlemede söz konusu dil İngilizce olduğunda birden fazla sözdizimsel çözümleyiciye ve görselleştirme programlarına rastlamak mümkündür. Çok bilinen çözümleyiciler genellikle üniversitelerin araştırma gruplarınca geliştirilmiştir. Bunlara örnek olarak Standford Parser, Berkeley Parser, TurboParser (Carnegie Mellon) verilebilir. Bu çözümleyicilerin ürettikleri ayrıştırma ağaçları ve gerçekleştirdiği bağımlılık ayrıştırmaları çeşitli yazılımlar veya yazılım kütüphaneleri ile görselleştirilebilmektedir (Şekil 1.1).



Şekil 1. 1 Bir cümle için ayrıştırma ağacı ve bağımlılık diyagramı [1]



Şekil 1. 7 Bir ayrıştırma ağacı için NlpViz'in ürettiği diyagram

Anlam Bilim (Semantics)

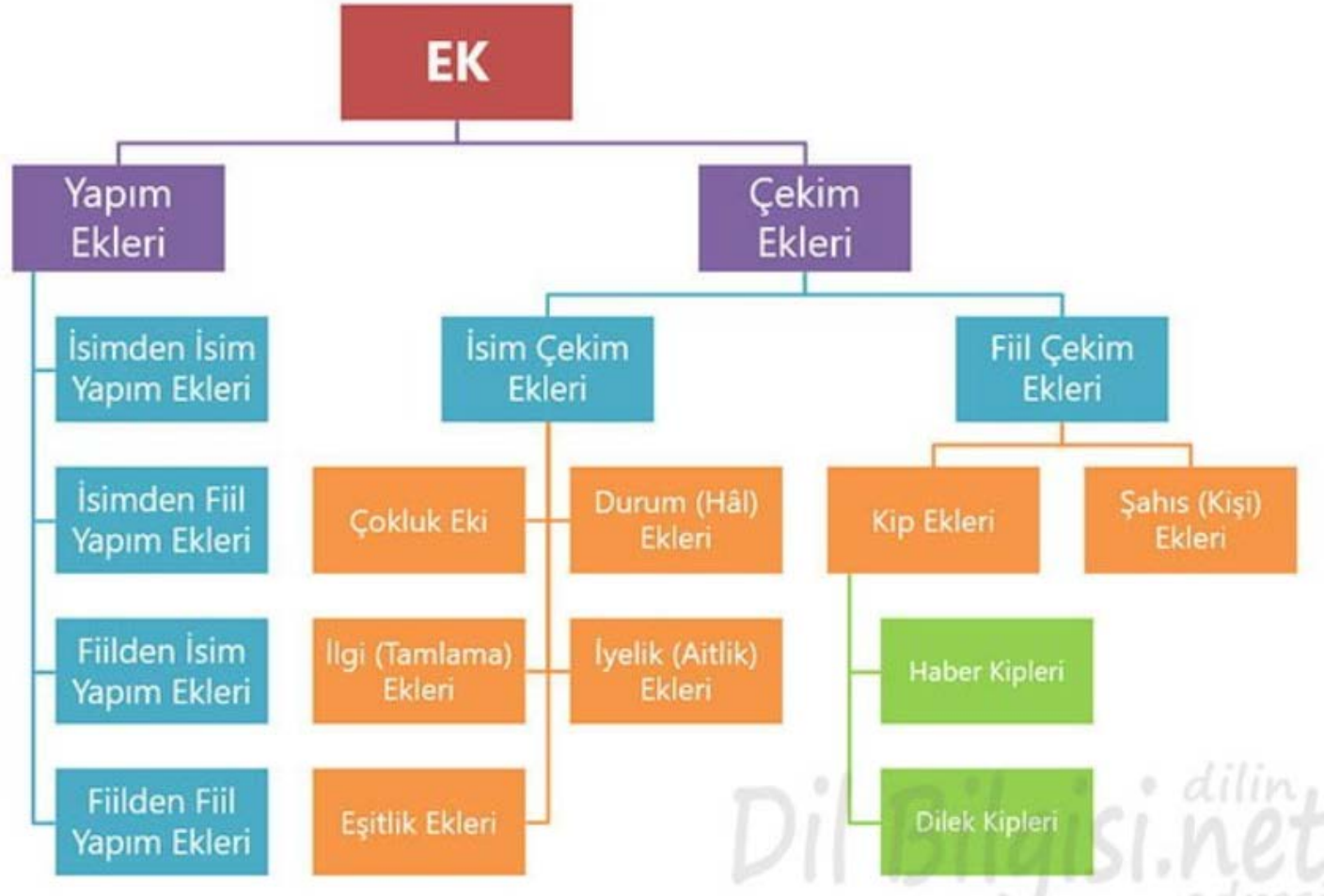
- Semantic, cümlenin anlamı ile ilgilidir.
- İnsan zihni, bir cümle ile karşılaştığında;
  - Bir kelime / kelime öbeği / nesne / senaryoyu çevreleyen tüm içeriği filtreler,
  - İlgili parçaları çıkarır,
  - Geçmiş deneyimlerle karşılaştırır
  - ve bunları elindeki içerik anlayışını derinleştirmek için kullanır.
- Doğal Dil İşleme ile,
  - Söz dizimi oluşturan morfolojik öğeler sınıflandırılır,
  - Bu aşamada kelimeler ek ve cümle hiyerarşisi bakımından sınıflandırılır.
  - Böylece birbirleri ile ilişkileri kurulabilir.
  - Bu ilişkiler sayesinde cümleden anlam çıkarılabilir.

Dilbilgisi (İngilizce)

- İngilizce bükümlü (çekimli) bir dildir.
- Sözcükler ön ve son ek alabilir.
  - Ön ek sayısı; 1, son ek sayısı 1 veya 2 olabilir.
- Ön ekler yapım ekidir.
- Son ekler yapım ve çekim eki olabilir.
- Çekim eki yapım ekinden sonra eklenebilir.
- Çekim ekleri bir kelimeyi çoğul yapmak veya bir fiilin kipini değiştirmek için kullanılır.
- Yapım ekleri ön veya son ek olarak eklenir ve kelimenin anlamı değiştirmek için kullanılır.
  - Derivation→derivational
  - Cool→coolness

Dilbilgisi (Türkçe)

- Türkçe sondan eklemeli bir dildir.
- Kelimelerin cümle içindeki görevlerini belirtmek veya onlardan yeni sözcükler türetmek için kullanılan hecelere **ek** denir.
- Türkçe'de 2 temel ek grubu var



Dilbilgisi

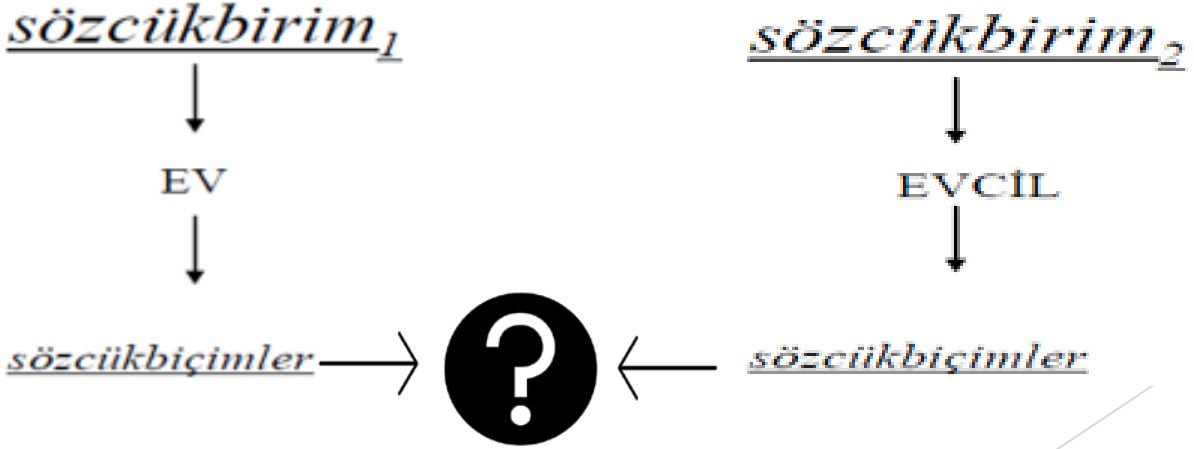
- Bir kelimeye eklendiğinde, farklı bir anlam türetilmesini sağlayan eklere **yapım ekleri (derivation)** denir.
- Bir kelimenin anlamını değiştirmeden farklı görevlerde kullanılmasını sağlayan eklere **çekim ekleri (inflection)** denir.
- Yapım ekleri her zaman çekim eklerinden önce gelir.
- Bir kelimenin kök ve eklerine ayrılması işlemine **biçim bilim çözümleme** denir.

Belirsizlik (Ambiguity)

- Türkçe'de
  - Dilin sondan eklemeli yapısı,
  - Çekim eklerinin çok olması,
  - Bazı eklerin hem yapım hem de çekim eki olarak kullanılabilmesi
  - PROBLEM!
- Gözlükleri
  - Gözlükleri getir. (göz+lük+ler+i (durum eki))
  - Bunlar onun gözlükleri. (göz+lük+ler+i (iyelik eki))
- **Belirsizlik (ambiguity);** Aynı kelime farklı anlam içerecek şekilde çözümlenmesi.
- Cümle seviyesinde analiz gerektirir.

Belirsizlik (Ambiguity)

- Belirsizlik sadece eklerin ayrıştırılmasında değil, kelime türlerini belirlemede de olabilir.
  - Bu gider tablosu çok detaylı hazırlanmış.
  - Ayşe her gün okula gider.
- Kök VS gövde
  - Kök; bir kelimenin parçalanamayan anlamlı, en küçük biçimidir.
  - Gövde; isim veya fiil köklerinden yapım ekleri ile türetilmiş olan yeni kelimedir.
- Sözcükbirim VS sözcükbiçim
  - Bir kelimenin cümlede kullanılan anlamına göre gövdesine **sözcükbirim**, o gövdeden çekim ekleri ile oluşturulan kelimelere ise **sözcükbiçim** denir.



Belirsizlik (Ambiguity)

sözcükbirim$_1$

↓

EV

↓

sözcükbiçimler

ev

evi

evde

evim

evimiz

sözcükbirim$_2$

↓

EVCİL

↓

sözcükbiçimler

evcil

evcilin

evcilsin

evciller

evcilden

Kullanım Bilgisi (Pragmatics)

- Kelimelerin sadece anlamları ve cümledeki yerleri ile değil,
  - Güncel hayatta nasıl kullanıldığının yorumlanması,
  - Metnin bir bütün olarak ele alınmasıdır.
- Doğal Dil İşleme'de
  - Konu modelleme
  - Soru cevaplama
  - Özetleme,
  - Metinlerdeki zamirlerin referanslarını belirleme gibi alanlarda gereksinim duyulmaktadır.

Temel DDİ Kavramları

- Vocabulary-Sözlük;
  - Bir metinde veya konuşmada kullanılan terimler
- Corpus-Derlem;
  - Film incelemesi, sosyal medya gönderileri vb. gibi benzer türde bir metin koleksiyonudur.
- Preprocessing-Önişleme;
  - Metinden, istenmeyen metin, terim ve gürültüyü temizlemek için uygulanan adımlar. Herhangi bir DDİ probleminin çözümü için ilk adımdır.
- Tokenization;
  - Büyük bir metnin küçük parçalara ayrılmasıdır. Her bir küçük parçaya 'token' denir. Ayrılan her küçük parça da bir metindir ve anlamlı bilgi içerir.
- Embeddings-Gömmeler;
  - Her token in, makine öğrenme modeline verilmeden önce bir vektör haline getirilmesi işlemidir. Gömmeler, kelimeler, kelime öbekleri veya karakterler üzerinde oluşturulabilir.

Temel DDİ Kavramları

- N-Grams;
  - Bir metnin, n'er tane kelime veya karakterden oluşan tokenlere ayrıştırılmış halidir.
- Transformers;
  - Paralel hesaplama yapabilen derin öğrenme mimarileridir. Metinlerdeki uzun vadeli bağımlılıkları öğrenmeye yarar.
- Parts of Speech (POS);
  - Kelimelerin cümledeki işlevleridir. İsim fiil gibi
- Parts of Speech Tagging;
  - Kelimelerin isim, fiil gibi etiketlerini belirleme işlemidir.
- Stop Words;
  - Metine anlamsal olarak katkısı olmayan, bağlaç gibi kelimelerin belirlenerek çıkarılması işlemidir.

Temel DDİ Kavramları

- Normalization;
  - Benzer terimleri kanonik bir forma eşleme sürecidir
- Lemmatization;
  - Kelimenin temel formunun belirlenmesi.
- Stemming;
  - Lemmatization gibi kelimelerin temel formunu belirlemeyi amaçlar. Ancak Lemmatization daki gibi kelimeleri POS taglerine ayırmadan bu işlemi yapar.
- DDİ'de Özellik Çıkarımı;
  - Yapılmakta olan işleme yönelik anahtar kelime veya ifadelerin belirlenmesi süreci. Böylece belirli bir kelime veya kelime öbeklerinin sınıflandırılması da sağlanır.

Kaynaklar

- Eşref Adalı, Türkçe Doğal Dil İşleme, Akçağ Yayınları