Doğal Dil İşlemede Kullanılan iki temel yaklaşım vardır:

- Sözdizimsel (Syntactic) Analiz
- Anlambilimsel(semantik)

Bir dokümandaki herbir sözcük, metnin en küçük birimi olup POS(part of speech) olarak adlandırılır. Her POS için bir etiket(tag) belirlenir. Örneğin İngilizce için aşağıdaki gibidir.

Etiket	Niteleme	Örnek
DT	Determiner	a, the
FW	Foreign Word	Mea culpa
IN	Proposition-sub conjunction	of, in, by
JJ	adjective	yellow
NNP	Proper noun-singular	
PP	Personel-pronoun	I, you, he
VB	Verb-base form	cut

Word	Lemma	Tag
the	the	+DET
girl	girl	+Noun
kissed	kiss	+Vpast
the	the	+DET
boy	boy	+Noun
on	on	+PREP
the	the	+DET
cheek	cheek	+Noun

Cümledeki her bir sözcüğün gövde ve sözcük türünü gösteren etiketlerle çözümlenmesi

Sözcüklerin Sözdizimsel Fonksiyonları

İsim(noun): Varlıkları tanımlar.

Belirteç(determiner): İsmi belirler(the, a, an)

Sıfat(adjective): İsmi niteler; isimle birlikte ortaya çıkar, tek başına kullanılmaz.(kırmızı top)

Fiil(verb): Bir oluşu, bir durumu kişiye bağlayarak anlatır.

Zarf(adverb): Fiillerin niteliğini belirtir.

> “güzel bir evde oturmak istiyorum” cümlesinde “güzel” sözcüğü sıfattır.
>
> “bu ev uzaktan daha güzel görünüyordu” cümlesinde ise zarftır.

Edat(proposition): Tek başına anlam taşımaz, kendinden önceki sözcükle birlikte kullanıldığında bir anlamı olur(gibi, için, kadar, karşı gibi)

Bağlaç(conjunction): Cümleleri veya aynı görevdeki sözcükleri birbirine bağlar; aradaki anlam ilgisini kurar. (ancak, ama, fakat, ile)

Sözdizim analizinde kullanılan yöntemler: Yukarıdan aşağıya ayrıştırma(top-down parsing) ve aşağıdan yukarıya’dır. Sözdizimsel analiz cümlenin yapısal tanımını oluşturabilmek için morfolojik analizin sonuçlarını kullanır. Amacı, arka arkaya gelen sözcükler yığınını, cümle birimleri olarak ifade eden bir yapıya kavuşturmaktır.

Semantik/Anlamsal Analiz: Sözcüklerin ayrı ayrı veya birbirine bağlanmış olarak oluşturduğu anlama karşılık gelen doğru modelleme yapıları oluşturmaktır. Anlamsal analiz, sözcüksel(lexical) ve bileşik(combination) olmak üzere iki türdür.

Sözcüksel analiz: Üst kavram(hyperonymy), alt kavram(hyponymy), zıt anlam(antonymy), eş anlam(synonymy), çok anlamlılık(polynymy), eş seslilik(homophony), parçanın bütünü(polonymy) ve bütünün parçası(holonymy)

Bileşik analiz: Bütünün anlamı, parçaların anlamından daha fazlası olabilir.

Metin Madenciliğinde **kavram çıkarımı**(concept extraction): Metnin en küçük parçası olup nesnelere veya fikirlere dayanır. Dolayısıyla nesne ve fikirlerin belirlenmesi yoluyla kavram keşfi yapılabilir. Sözdizimsel tanımlamalar yoluyla da kavram çıkarımı yapılabilir. İsim, sıfat, fiil tamlamaları tanımlanarak kavram çıkarılabilir. Aynı kavramın çeşitli görünümleri olabileceğinden bunları standart bir biçime dönüştürmek için normalizasyon süreci uygulanır. Bu işlem, kategori oluşturma ve eğilim yakalama için önemlidir. Örn: “vadesiz hesap”, “bu vadesiz hesaplar”, “benim vadesiz hesaplarımdan biri” gibi ifadeler hep aynı kavramla ilgilidir. Kavram ve varlık çıkarım araçları birlikte kullanılabilir. “Türkiye başbakanının konuşması” bileşik kavramdır. “başbakanın konuşması” şeklinde bir kavrama ve “Türkiye” gibi bir varlığa bölünebilir.

Varlık çıkarımı(entity extraction): Gramere dayalı tam ayrıştırma(parsing) yapar. Gramere dayalı linguistik analiz, örüntü yakalama ve sözlükler yardımıyla varlıkları tanır ve sınıflar.

“Turan Güneş”(kişi) ve “Turan Güneş”(bulvar) arasında ayrım yapabilir.

Sözdizimsel kurallara dayanarak varlıkları türlendirir. Örn: “Turan Güneş’te oturuyorum” ifadesindeki ismin “bulvar” olarak yorumunu yapabilir. Özel isimleri, sayısal varlıkları(banka hesabı, telno), alfanümerik varlıkları(otomobil plakaları, web adresleri), Twitter kullanıcıları, hashtagler vb Özel isimleri de farklı tiplere göre sınıflar: İnsanlar/yerler/ticari/kuruluş vb Varlıklar farklı yazılmış olsalar da bu farklılıklar yakalanabilir. (20:00, saat 20, 20s) Normalizasyon süreci uygulanarak standart bir biçime dönüştürülür. NYSE, New York Stock Exchange, NY Stock Exchange gibi aynı varlığın çeşitli yazım biçimleri standart bir biçime dönüştürülür.

Metin kategorizasyonu(text categorization): İsim, sıfat, fiil tamlamalarının bağlama göre kategorize edilmesi için sözdizim analizi kullanılır. Tekil sözcüklerin kategorizasyonu için de kullanılır. Metnin dilbilimsel temsili, taksonomiyi içeren bir sözlük aracılığı ile kontrol edilir. Metindeki bir sözcük ya da tamlama bir sözlük girişi ie karşılaştığında b giriş için olan kategori metne atanır. Örn: Mobil telefon alanında tipik bir kategorizasyon ekran, kılıf, kap, kamera, pil gibi ürün kategorisine sadece isim olarak bağlananları hesaba katar. Bu durumda

“I love the **screen** on my new Kindle Fire” veya “I ‘ve bought a great new **cover** for my IPAd” gibi cümleler olduğunda bunlar **ürün kategorisine** ait olarak sınıflanır**.**

“I hate it when they **screen** my IPAD at security**” veya “**They are going **to cover** the new Galaxy Tab in next week’s review**”** olduğunda ise ekran ve kap fiil olarak analiz edilir.

Sözlük Yaratma Süreci

Basit anahtar sözcük eşleşmesine değil, kullanılan sözcüklerin anlamlarına dayanır. Sözcüklerin biçimlerini(anlamlarını değil) değiştiren linguistik varyasyonlar doğru biçimde işlenir. Bu, linguistik fenomenleri (morfolojik varyasyonlar: zaman, kişi, cinsiyet, sayı vb göre farklı biçimleri) ve sözdizimsel kuralları içerir.

Kategorizasyon kullanıcının sağladığı bir taksonomi ile birlikte çalışır. Fakat genelde kolayca entegre edilebilecek önceden var olan bir kategori sözlüğü/tesarus yoktur. Bu durumda zaman ve maliyeti düşünerek bir tane yaratmak gerekir.

Duygu Analizleri: Duygu skoru, konu tanımlama, kategori oluşturma ile pozitif ve negatif değerlerden daha fazlası yapılabilir. Gramer analizi kullanılarak yalnız cümle düzeyinde değil, cümle içinde tamlama düzeyinde de görüş analizi yapılabilir. Sözdizimsel analiz farklı tamlamaları(isim, sıfat, fiil gibi) ve bunların bağımlılıklarını tanımlayabildiği için mümkündür. Cümledeki tek bir görüşün çıkarılması ile de sınırlı değildir. Cümlenin içerdiği birçok görüşü algılayabilir. Örn:

“The phone was awesome but it was too expensive and the screen is not big enough” cümlesinden 3 görüş çıkarılabilir. “phone” + ”awesome”, “phone” + “too expensive”, “screen” + “not big”

Duygu analizi aracı ile konu yakalama

Varlık(marka, kişi, ürün, yer vb) – kavram(“küresel ısınma”, “kamu politikaları” veya “finansal kriz”) ve tartışılan konunun özelliklerini/özniteliklerini tam olarak algılar. Olumsuzlama da işlenebilir. “their camera is really not bad at all” gibi. Duygu analizinde rol oynayan karmaşık dil yapıları (negatif veya karşılaştırma) da ele alınır. Bu yapılar görüş farklılıklarını yakalayabilir.

“The phone is much better than my old phone” (+)

“The phone is not much better than my old phone” (-)

Duygu skorunda yoğunluk yakalanabilir. Kişinin konu hakkındaki duygularının yoğunluğu arttıkça, skor yükselir veya azaldıkça tersi olur. Bunu elde etmek için linguistik özellikleri örn sözcüğün semantik gücü veya yoğunluğu belirten sözcüklerin kullanımı (really, very, extreemly gibi) belirlemek gerekir.

“installing software on this machine is really very painfull ” gibi bir cümlenin duygu skoru, “installing software on this machine is really very painfull indeed” ifadesine göre daha düşüktür.

Duygu analizi ve kategorizasyon araçları birlikte de kullanılabilir. Böylelikle konunun özellik ve öznitelikleri bir taksonomiden üretilmiş kategoriye atanabilir.

Adlandırılmış Varlık Tanıma(Named Entity Recognition): Varlık tanımlama, varlık çıkarımı olarak da bilinir. Enformasyon çıkarımının bir alt işlemidir.