

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

**YENİ BİR SÖZDİZİMSEL İŞARETLEME YÖNTEMİNİN KULLANIMIYLA
TÜRKÇE’NİN İSTATİSTİKSEL AYRIŞTIRMA BAŞARIMININ ARTIRILMASI**

YÜKSEK LİSANS TEZİ

Umut SULUBACAK

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

MAYIS 2015

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

**YENİ BİR SÖZDİZİMSEL İŞARETLEME YÖNTEMİNİN KULLANIMIYLA
TÜRKÇE’NİN İSTATİSTİKSEL AYRIŞTIRMA BAŞARIMININ ARTIRILMASI**

YÜKSEK LİSANS TEZİ

**Umut SULUBACAK
(504151542)**

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Yrd. Doç. Dr. Gülşen ERYİĞİT

MAYIS 2015

İTÜ, Fen Bilimleri Enstitüsü'nün 504151542 numaralı Yüksek Lisans Öğrencisi **Umut SULUBACAK**, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı **“YENİ BİR SÖZDİZİMSEL İŞARETLEME YÖNTEMİNİN KULLANIMIYLA TÜRKÇE’NİN İSTATİSTİKSEL AYRIŞTIRMA BAŞARIMININ ARTIRILMASI”** başlıklı tezini aşağıdaki imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı : **Yrd. Doç. Dr. Gülşen ERYİĞİT**
İstanbul Teknik Üniversitesi

Jüri Üyeleri : **Yrd. Doç. Dr. Ahmet Cüneyd TANTUĞ**
İstanbul Teknik Üniversitesi

Yrd. Doç. Dr. Arzucan ÖZGÜR
Boğaziçi Üniversitesi

.....

Teslim Tarihi : **5 Mayıs 2015**
Savunma Tarihi : **26 Mayıs 2015**

ÖNSÖZ

Tez çalışmamın her aşamasında benimle özverili bir şekilde ilgilenen ve hiçbir konuda desteğini eksik etmeyen hocam Gülşen Eryiğit'e ve çalışmamın gelişim süreci boyunca değerli görüşlerini benimle paylaşarak bana destek olan çalışma arkadaşım Tuğba Pamay'a içten teşekkürlerimi sunarım.

Araştırmalarımıza iki buçuk yıllık bir süreçte destek vererek bu çalışmanın çıkarılabilmesine olanak sağlayan Türkiye Bilimsel ve Teknolojik Araştırmalar Kurumu (TÜBİTAK) ilgililerine ve Avrupa Bilim ve Teknoloji İşbirliği (COST) komitesine de teşekkürü borç bilirim.

Ayrıca, eğitim hayatıma her daim koşulsuz destek olarak bu noktaya gelmeme olanak tanıyan, yoğun çalışmalarım süresince bana karşı sonsuz sabır gösteren sevgili aileme her şey için çok teşekkür ederim.

Mayıs 2015

Umut SULUBACAK
Bilgisayar Mühendisi

İÇİNDEKİLER

Sayfa

ÖNSÖZ	v
İÇİNDEKİLER	vii
KISALTMALAR.....	ix
ÇİZELGE LİSTESİ.....	xi
ŞEKİL LİSTESİ.....	xiii
ÖZET	xv
SUMMARY	xvii
1. GİRİŞ	1
1.1 Tezin Amacı.....	2
1.2 Türkçe ve Bağlılık Formalizmi.....	3
2. SORUNLAR VE ÖNERİLEN ÇÖZÜMLER	7
2.1 Anlamsal Tutarsızlık.....	10
2.2 Hiyerarşi ve Çakışma.....	11
2.3 Belirsizlik	13
2.4 İsteğe Bağlılık.....	14
2.5 Düzen Aşırılığı	15
3. YENİ DERLEMLER	17
3.1 Ön Bilgiler	17
3.2 Çok İyeli İşaretleme	18
3.3 Derlem İstatistikleri	20
4. DENEYLER	23
4.1 Ön Bilgiler	23
4.2 İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi	25
4.3 Yeniden Eşleme Deneyleri	27
4.4 İTÜ Ağaç Yapılı Web Derlemi	29
5. SONUÇ	33
KAYNAKLAR.....	35
EKLER	39
EK A: Bağlılık Geri Dönüşüm Akış Diyagramları	41
EK B: ITU Treebank Annotation Guide (v2.7).....	43
2.1 Elementary Dependencies	43
2.1.1 PREDICATE.....	43
2.1.2 OBJECT	43
2.1.3 SUBJECT	44
2.2 Modifiers.....	44

2.2.1 DETERMINER	44
2.2.2 POSSESSOR	45
2.2.3 MODIFIER.....	45
2.3 Adpositionals	46
2.3.1 ARGUMENT	46
2.3.2 INTENSIFIER.....	47
2.4 Coordinations	48
2.4.1 CONJUNCTION	48
2.4.2 COORDINATION	49
2.4.3 RELATIVIZER.....	50
2.4.4 APPPOSITION.....	50
2.5 Other Dependencies.....	51
2.5.1 PUNCTUATION	51
2.5.2 VOCATIVE	51
2.5.3 DERIV	52
2.5.4 MWE	53
2.6 Deep Dependencies	53
ÖZGEÇMİŞ	57

KISALTMALAR

BS_{E+}	: Etiketli Baęlanma Skoru
BS_{E-}	: Etiketsiz Baęlanma Skoru
ÇK	: Çekim Kümesi
İOSD	: İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi
İWD	: İTÜ Ağaç Yapılı Web Derlemi
OSD	: ODTÜ-Sabancı Ağaç Yapılı Derlemi

ÇİZELGE LİSTESİ

Sayfa

Çizelge 2.1: İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi işaretleme çerçevesi ile Evrensel Stanford Bağlılıkları arasındaki eşleme.	9
Çizelge 3.1: Temel alınan ve yeni tanıtılan derlemlerdeki cümle, birim ve bağlılık istatistikleri.	20
Çizelge 3.2: İTÜ Ağaç Yapılı Web Derlemi’ni oluşturan cümlelerin alan dağılımı.	21
Çizelge 3.3: Temel alınan ve yeni tanıtılan derlemlerdeki bağlılık türü etiketlerinin dağılımları.....	21
Çizelge 4.1: İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi üzerinde doğrulanan modellerin çapraz doğrulama skorları.	26
Çizelge 4.2: İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi’nde MODIFIER , POSSESSOR ve MWE türlerinin diğer türlere eşlendiği üç alternatif kol ile ilişkili modeller üzerindeki çapraz doğrulama skorları.	29
Çizelge 4.3: İTÜ Ağaç Yapılı Web Derlemi üzerinde doğrulanan modellerin çapraz doğrulama skorları.....	30

ŞEKİL LİSTESİ

	<u>Sayfa</u>
Şekil 1.1 : “ <i>Kırmızı arabadaydı</i> ” cümlesi için bağılılık ağacı.	4
Şekil 1.2 : Şekil 1.1’de verilen cümlelerin İngilizce çevirisi olan “ <i>She was in the red car</i> ” cümlesi için bağılılık ağacı.	4
Şekil 2.1 : Nesneleri ifade eden OBJECT bağılılığının bir cümle nesnesinde (<i>solda</i>) ve bir edat tümlecinde (<i>sağda</i>) kullanımı. İkinci kullanım yeni bağılılık türü ARGUMENT kapsamına dahil edilmiştir.	10
Şekil 2.2 : Koordinasyon yapıları için orijinal (<i>solda</i>) ve önerilen (<i>sağda</i>) işaretleme yöntemleri. İkinci yöntemin gerek ayrıştırma kolaylığı açısından, gerekse anlamsal açıdan daha uygun olduğu savunulmaktadır.	11
Şekil 2.3 : İki benzer deyimsel kullanımın ETOL ve COLLOCATION bağılılık türleri ile gösterilmesi. Önerilen çerçevede tüm benzer deyimsel kullanımlar genel bağılılık türü MWE altında işlenmektedir.	11
Şekil 2.4 : Niteleyici görevi gören ad soylu belirteçlerin hallerine göre farklı X.ADJUNCT etiketleri ile ifade edilmesi. Bu tür kullanımlar genel bağılılık türü MODIFIER kapsamına alınmıştır.	12
Şekil 2.5 : Belirli noktalama işareti türlerinin herhangi bir iyeye bağlanmamasına izin verilmesi (<i>solda</i>). Yeni bağılılık türü PUNCTUATION her türden noktalama işaretini kapsayacak ve standart bir işaretleme yöntemi olacak şekilde önerilen çerçeveye dahil edilmiştir.	14
Şekil 2.6 : Orijinal işaretleme çerçevesinde ihmal edilebilen öğelere bağılılık Cümle (<i>I</i>) kurallı dilde gerek duyulacak, ancak web üzerinde yazımı sıklıkla ihmal edilen belirli öğelere bağılılık yüzünden işaretlemenin imkansızlaştığı iki durumu göstermektedir. Cümle (<i>II</i>)’de gösterilen otomatik normalizasyon sonrasında dahi, Cümle (<i>III</i>)’teki gibi bağlaç ve son nokta görevi gören noktalama işaretleri eklenmediği takdirde kurallı işaretleme mümkün olmamaktadır.	15
Şekil 2.7 : Önerilen işaretleme çerçevesinde öge bağımsızlığı yaklaşımı Şekil 2.6 üzerinde görüldüğünün aksine, aynı cümlelerin işaretlenmesi için önerilen etiketler kullanıldığında diğer öğelere bağımlılık oluşmamaktadır. Bu örnek koordinasyon yapıları ve ana cümle yüklemelerinin bağlayıcı noktalama işaretlerinin varlığında ve yokluğunda, birimlerin sırasıyla Cümle (<i>IV</i>) ve Cümle(<i>V</i>) üzerinde görüldüğü şekilde bağlandığı durumları göstermektedir.	16

Şekil 3.1	: Derin bağılıkların işaretlenmesinin gerekli olduğu durumlardan birini gösteren bir örnek. Özne görevindeki öge ‘Deniz’ cümledeki her iki eylemi de nitelemektedir.	19
Şekil A.1	: Yeni MODIFIER bağılılığının orijinal ODTÜ-Sabancı Ağaç Yapılı Derlemi işaretleme çerçevesindeki muadillerine eşlenme akış diyagramı.	41
Şekil A.2	: Yeni POSSESSOR bağılılığının orijinal ODTÜ-Sabancı Ağaç Yapılı Derlemi işaretleme çerçevesindeki eski POSSESSOR ve CLASSIFIER türlerine eşlenme akış diyagramı.....	41
Şekil A.3	: Yeni MWE bağılılığının ilgili sözdizimsel bağılık türlerine eşlenme akış diyagramı.....	42

YENİ BİR SÖZDİZİMSEL İŞARETLEME YÖNTEMİNİN KULLANIMIYLA TÜRKÇE’NİN İSTATİSTİKSEL AYRIŞTIRMA BAŞARIMININ ARTIRILMASI

ÖZET

Bu çalışmada, mevcut tek ağaç yapılı Türkçe bağıllık derlemi olan ODTÜ-Sabancı Türkçe Ağaç Yapılı Derlemi’nde (OSD) kullanılan bağıllık gramerinin eleştirel bir analizi verilmekte ve ardından Türkçe için yeni ve daha gelişmiş bir bağıllık işaretleme çerçevesi önerilmektedir. Yeni çerçeve minimallik ve elle işaretlemenin kolaylaştırılması üzerinde durmakta, orijinal çerçevenin sahip olduğu 26 bağıllık etiketine karşılık yalnızca 16 bağıllık türü ile ifade gücünden bir şey kaybetmeden daha açık ve anlaşılır olabilmektedir.

İşaretleme çerçevesinin ilk uygulamaları olarak çalışma kapsamında iki yeni ağaç yapılı derlem tanıtılmaktadır: 1) OSD’nin İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi (İOSD) adlı, yüksek başarımla sergileyen biçimbirimsel etiketler ve yeni bağıllık türleriyle işaretlenmiş yeni bir sürümü, ve 2) Web üzerinde kullanıcıların girdiği kural dışı cümlelerden derlenerek sözdizimsel olarak işaretlenmiş ilk Türkçe derlem olan İTÜ Web Ağaç Yapılı Derlemi (İWD). Tanıtılan yeni derlemlerimizin her ikisinde de derin bağıllık işaretlemesi yapılarak kaynaklar gelecek anlamsal işlev etiketleme çalışmalarının da yararlanabileceği şekilde düzenlenmiştir.

Önerilen işaretleme yordamlarının temel alınan işaretleme çerçevesine göre başarımlarını ölçebilmek amacıyla sunulan ağaç yapılı derlemlerde ayrıntılı incelemeler yürütülmüş ve çalışma kapsamında bildirilmiştir. Verilen deney sonuçları başarımlarının yalnızca bağıllık etiket kümesinin küçülmesinden çok daha anlaşılır ve tutarlı olan yeni bağıllık gramerinin öngördüğü işaretleme yöntemiyle ilgili olduğunu ortaya koymaktadır.

Çalışmada tanıtılan çok iyeli bağıllık temsili, altın standart kümesinde işaretlenen iyelerden herhangi birinin uygun olması durumunda bir öngörünün kabul edilmesine dayalı yeni bir değerlendirme metriğinin işletilmesine olanak sağlamaktadır. Ayrıştırma deneylerinin sonuçlarına göre İOSD için tanıtılan en iyi model tek iyeli değerlendirmede %75.1, çok iyeli değerlendirmede ise %75.7 etiketli bağlama skoru olarak OSD için şimdiye dek elde edilmiş en iyi etiketli bağlama skoru olan %65.9’u büyük farkla geçmektedir. Ayrıca İWD’nin çapraz doğrulanması sonucu tek iyeli değerlendirmede %78.7, çok iyeli değerlendirmede ise %80.1 olmak üzere yüksek umut vaat eden etiketli bağlama skorları elde edilmektedir.

IMPROVING STATISTICAL DEPENDENCY PARSING PERFORMANCE IN TURKISH BY USE OF A NEW ANNOTATION SCHEME

SUMMARY

In this work, we present a critical analysis of the dependency grammar that has come to be the de facto standard for Turkish language processing studies. Although widely recognized and used in several Turkish corpora including the well-known METU-Sabancı Treebank (MST), the only major syntactically annotated Turkish corpus to date, the grammar is partly outdated, improvable and extensible. Moreover, the METU-Sabancı Treebank itself is often criticized for its inconsistent annotation and difficulty of parsing.

Many recent studies centered around the syntactic parsing of Turkish have focused on fine-tuning specific aspects of their parsing frameworks and failed to make a pivotal overall progress in their parsing performances. We take a detour from specific case studies that would only yield local performance improvements, and delve into the entire structure of the annotation framework. We investigate the current Turkish annotation conventions in detail, identify any flaws and deficiencies with respect to both manual annotation and automatic parsing, and then propose measures that might be taken to alleviate these issues.

Furthermore, as web data become increasingly available for study and the ability to efficiently parse non-canonical sentences gain importance, we place special emphasis on making dependency annotation as lenient on non-canonical texts as possible. The extent by which the colloquial language employed by social web users differs from well-typed formal language is indeed very large, and it is often not enough to orthographically normalize non-canonical sentences in a pre-processing routine to render them as successfully parsable as edited formal texts. As part of this work, we also attempt to parametrize the differences of the language of the web, and likewise suggest what morphosyntactic reforms would likely improve parsing performances.

In accordance with our findings, we also propose a new, improved dependency annotation framework for Turkish. The proposed framework additionally focuses on minimalism and ease of manual annotation, featuring only 16 dependency types that are decidedly more coherent and intuitive compared to the 26 labels of the original framework. We justify all of our proposed changes in the entailed dependency grammar from the original version by either showing conformity with the design principles we explain or demonstrating overlap with universally recognized conventions that have been long since proved.

As the first implementations of the proposed annotation framework, we introduce two new treebanks: 1) A new version of the METU-Sabancı Treebank keeping the same token structure and morphosyntactic features but reannotated with the new dependency types, for which we propose the name ITU-METU-Sabancı Treebank (IMST) in recognition of the considerable previous effort on the original treebank as well as our contribution, and 2) The ITU Web Treebank (IWT), the first Turkish corpus composed of non-canonical user-input sentences extracted from the web, annotated from the ground up in normalization, morphology and syntax layers. Both of our new corpora are marked for deep dependencies in order to support future semantic role labeling studies.

We do not establish any hierarchy between deep and surface dependencies, and rather employ a basic approach that simply supports multiple heads for a single constituent. Although this notation makes our syntactically annotated sentences incompatible with most syntactic parsers in common use, it is straightforward to remap from the multi-headed raw sentences to single-headed projections whenever necessary, and so it boosts the expressiveness of our syntactic annotation without incurring any loss in applicability. For our parsing tests that we discuss in the later sections, we use two elementary single-head choosing methods as a precursor to smarter head choosing routines that may be developed for future work.

Although constituents are conventionally annotated with a single head in dependency parsing, the practice is not always beneficial as there may be more than one head for a dependent that would make sense given clausal structures within the sentence containing the dependent. In such cases, while automatic parsers may predict a meaningful head for a given dependent, the gold-standard validation set may be annotated with another head that is also meaningful, but would still cause the prediction to be determined as incorrect simply because the two heads do not match. We mean for the newly-introduced multi-headed representation to also help in alleviating false negatives caused by such scenarios, by use of a new evaluation metric that we call *relaxed evaluation* (as opposed to the conventional *strict evaluation*) able to validate predicted dependencies that match any one of the heads designated in the gold-standard.

After our discussions, we present our detailed empirical investigations on the new treebanks in order to demonstrate the impact of our proposed annotation schemes with respect to the original framework. We perform cross-validation on all of our models and cross-check parsing models trained from each combination of training sets and single-head choosing routines with each other where appropriate. We provide the figures resulting from our parsing tests and discuss their significance in detail. Additionally, we conduct a series of targeted remapping tests in order to make sure that certain annotation scheme changes were indeed well-founded and effective. Furthermore, our experiments indicate that the parsing performance increases we attain

are not caused by the reduction of the dependency label set, but rather related to our more coherent annotation framework prescribed by the new grammar.

Our final tests show that our best model for the IMST attains labeled attachment scores of 75.1% for strict evaluation and 75.7% for relaxed evaluation, surpassing the state-of-the-art parsing score of 65.9% by a large margin. Cross-validation of the IWT also yields 79.7% for strict evaluation and 80.1% for relaxed evaluation for the best model. Considering these scores, our new resources reveal up to nearly 12 percentage points improvement on the performance of parsing web data.

1. GİRİŞ

Türkçe sözdizimine gösterilen yaygın ilgiye rağmen, yakın zamandaki birçok incelemenin [1–6] de gösterdiği üzere, Türkçe cümlelerin ayrıştırılmasında uzun zamandır geniş çaplı bir ilerleme görülememiştir. Bu konudaki çalışmaların geneli belirli hesaplama veya dilbilim sorunlarına odaklanıp kullandıkları ayrıştırıcılara çeşitli yönlerden ince ayar yaparak yerel gelişmeler kaydetmekte, ancak önemli bir genel ilerleme sağlayamamaktadır. Bu tür çalışmalar ayrıştırma çerçevelerinin geri kalanına müdahale etmemekte ve bu yolla geçmiş çalışmalarla uygun bir biçimde karşılaştırma yapma imkanı oluşturmaktadır, ancak aynı sebeple, odak noktasında bulunan belirli sorunlar dışında kalan birçok konu yeteri kadar araştırılmamakta ve alanın önemli bir kısmı keşfedilememektedir.

Bu şekilde çalışmalara sıklıkla konu olan durumların dışında kalarak sözdizimsel ayrıştırmaya darboğaz teşkil eden birtakım sorunların olduğu görülmektedir. En yüksek başarımlı ayrıştırıcıların istisnasız olarak elle işaretlenmiş derlemlerden denetimli öğrenme yaptığı düşünüldüğünde, sözkonusu sorunların derlemlerle ilgili eksikliklerden kaynaklanması olasıdır. Yıllar boyunca çok değerli bir kaynak teşkil eden ODTÜ-Sabancı Türkçe Ağaç Yapılı Derlemi [7], bu güne kadar tasarlanmış hemen her Türkçe bağıllık ayrıştırıcısı tarafından kullanılmıştır. Buna rağmen derlemin kullandığı bağıllık grameri halen çeşitli açılardan sık sık eleştirilmektedir. Bazı geçmiş çalışmalarda [8, 9] da değinildiği üzere derlemin birtakım kusurları, özellikle de çok sayıda işaretleme tutarsızlığı öne çıkmaktadır. Halihazırda Türkçe için işaretlenmiş, ODTÜ-Sabancı Türkçe Ağaç Yapılı Derlemi'ne muadil veya alternatif olabilecek farklı bir kaynak bulunmamaktadır¹ ve bu durum derlemle ilgili tespit edilebilecek eksikliklerin ortaya çıkmasını daha da

¹Derlemler arasında İTÜ Doğrulama Kümesi [10] de sayılabilir, ancak bu yalnızca 300 cümle içeren küçük bir derlem olup denetimli öğrenme sistemlerinde doğrulama ve test amacıyla oluşturulmuştur ve bu yüzden eğitim için yetersiz kalmaktadır.

engellemektedir.

ODTÜ-Sabancı Ağaç Yapılı Derlemi'nin hem (bağlılık grameri yönünden) teorik, hem de (işaretleme yönünden) pratik açılardan geniş bir gelişme payı olduğu bilinmektedir. Bu düşüncelerin ışığında, belirli ayrıntılara odaklanan incelemelerden uzaklaşarak daha radikal yenilikler düşünmek önemlidir. Bu nitelikte bir çalışma, birden fazla dil üzerine gerçekleştirilen ayrıştırma çalışmalarında ortaya çıkan, Buchholz ve Marsi'nin [11] de değindiği derlem ayrıştırma zorluğu ve Nilsson, Riedel ve Yüret'in [12] de altını çizdiği ayrıştırıcılar arası istikrarsızlık gibi sorunların da hafifletilmesini sağlayabilecektir. Bu yönde kapsamlı bir araştırmaya girmek veya oturmuş bir derlemi yeniden oluşturmaya kalkışmak yorucu ve getirisine göre yüksek maliyetli bir iş gibi görünebilse de, bu tür radikal bir çalışmanın faydalı olma potansiyeli yüksektir.

Bunların yanı sıra, modern dil işleme uygulamalarının ihtiyaçlarına göre düşünüldüğünde, web üzerinde kullanıcıların oluşturduğu metinler ve diğer kural dışı verilerin işlenebilmesine olanak sağlamak, oturmuş bir derlemin bu motivasyonla yeniden işaretlenmesini desteklemektedir. İnternet üzerinde yaygın olarak kullanılan dil basit bir jargona göre standart dilden çok daha fazla farklılaştığından dolayı, web verisi ayrıştırma için çok özgün bir alan teşkil etmektedir. Web verisinin ayrıştırılması, Seddah ve diğ. [13] tarafından da ele alındığı gibi alan adaptasyonuna teorik olarak benzer bir iş olsa dahi, web üzerinde eğilim gösterilen dil kullanımlarını da kapsayabilmek için özel formalizmlere ihtiyaç duymaktadır. ODTÜ-Sabancı Türkçe Ağaç Yapılı Derlemi'nin oluşturulmasında da odak noktalarından biri alan çeşitliliği olmuştur, ancak derlemin tamamen kurallı cümlelerden oluşması derlemi web verisine yönelik eğitilecek bir ayrıştırıcı için yetersiz kılmaktadır.

1.1 Tezin Amacı

Kuralsız veriler düşünülerek hazırlanmış bir bağlılık grameri ve buna uygun bir ayrıştırma altyapısının varlığı ile, alan farklılığı sorununun çözümü olarak web

verisine yönelik farklı ve yeni bir derlemin oluşturulması netlik kazanır. Alan adaptasyonu konusunda daha önceki çalışmalar ile Google Web Treebank [14] ve French Social Media Bank [13] gibi daha yeni web odaklı derlem çalışmalarının da desteklediği gibi, Türkçe için de özel bir web derleminin oluşturulması, web verisinin başarılı bir şekilde işlenebilmesine kurallı ve formal cümlelerden oluşan genel bir derleme göre çok daha fazla katkı sağlayacaktır. Bu sebepten dolayı, yeni önereceğimiz işaretleme çerçevesini kullanarak Türkçe için bir derlem işaretlemek ve yayımlamak bu çalışma için ikincil bir motivasyon oluşturmaktadır.

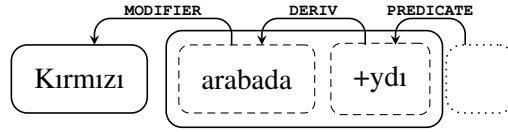
Çalışmada Türkçe için yeni bir bağıllık işaretleme çerçevesi önerilmektedir. Bu çerçevenin öncelikli amaçları, ODTÜ-Sabancı Ağaç Yapılı Derlemi'ndeki bilinen sorunları ekarte etmek, işaretleyici kişilerin üzerindeki yükü hafifleterek, işaretleme hatalarını azaltmak ve kuralsız dili modellemek konusunda daha yüksek başarı sağlamaktır. Çalışmada ODTÜ-Sabancı Ağaç Yapılı Derlemi'ndeki cümlelerin bu çerçeveye göre yeniden işaretlenmesiyle oluşturulan İTÜ-ODTÜ-Sabancı Türkçe Ağaç Yapılı Derlemi, yeni bir sürüm olarak tanıtılmaktadır. Çalışmanın sunduğu bir diğer derlem de İTÜ Ağaç Yapılı Web Derlemi adını verdiğimiz yeni bir web derlemidir. Çalışmanın devamında bu yeni ağaç yapılı derlemler üzerinde deneysel ölçümler yapılmakta ve Türkçe için mevcut en yüksek başarımlı ayrıştırma sistemleri ile karşılaştırmalı sonuçlar sunulmaktadır.

1.2 Türkçe ve Bağıllık Formalizmi

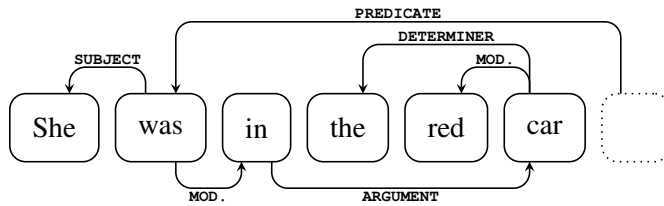
Bağıllık kavramı literatürdeki en eski gramerlerden [15] beri bilinmesine rağmen modern *bağıllık grameri* olarak Tesnière'e [16] atfedilmektedir ve son yıllarda yoğun ilgi görerek hesaplamalı dilbilim alanında yaygın olarak işletilmiştir. Bağıllık grameri, bir cümlenin sözdizimini modellemek için bağıllıkların ne şekilde kullanılacağına yönelik bir kural kümesi belirlemesi açısından, bağıllık olgusunun pratik bir uygulaması niteliğindedir. Olası bağıllık türlerinin tanımlanması, bu bağıllık türlerinin hangi dilbilimsel kavramları modellemekte nasıl kullanılacağına belirlenmesi, cümlelerin hangi öğelerinin nasıl bağıllıklar oluşturacağı konusundaki kısıtların seçilmesi ve cümle içinde hangi bağıllık türlerinin bulunması gerektiği konusundaki

gereksinimlerin ortaya konması bir bağıllık grameri tarafından öngörülür.

Bağıllık formalizminin dışında öge formalizmi ise bir cümleyi kendileri de özyinelemeli olarak alt cümlelere bölünen yan cümlelere bölerek özyinelemeyi cümlelerin öğelerine kadar indirir ve cümlelerin sözdizimini bu şekilde ifade eder. Öge gramerleri bu şekilde temelde hangi tür yan cümlelerin ardışık öğeler tarafından meydana getirilebileceğini modellerler. Türkçe gibi serbest öge dizilimli diller için bu gramerlerin gözetilebileceği tüm öge dizilimlerini düşünmek çok karmaşık ve zorlayıcı gramerlerin tanımlanmasını gerektirir. Türkçe sözdizimini öge gramerleriyle modellemek için başta Çetinoğlu'nun [17] CCG grameri ve Çakıcı'nın [9] LFG grameri olmak üzere bazı çalışmalar yapılmıştır, ancak bu çalışmaların devamı gelmemiş ve araştırma zaman içinde daha sürdürülebilir olarak görülen bağıllık formalizmine kaymıştır [2, 4, 6].



Şekil 1.1: “Kırmızı arabadaydı” cümlesi için bağıllık ağacı.



Şekil 1.2: Şekil 1.1’de verilen cümlelerin İngilizce çevirisi olan “She was in the red car” cümlesi için bağıllık ağacı.

Türkçe’nin sözdizimsel analizine odaklanan çoğu modern çalışmada da olduğu gibi bu çalışmada da bağıllık formalizmi esas alınmaktadır. Bu formalizm, Şekil 1.2 üzerinde de gösterildiği gibi, sözdizimsel bilginin cümlelerin öğeleri arasındaki *bağıllık* adı verilen yönlü ikili bağıllıkların bir kümesi olarak ifade edilmesini gerektirir. Her bağıllık bir asil öge (*iye*) ile onu niteleyen bir ikincil öge (*uydu*) arasında tanımlanır.

Bağlılıklar ayrıca öğeler arasındaki ilişkinin türünü ifade eden, *bağlılık türü* adı verilen birer etikete sahiptir. Bağlılık formalizmi hakkında yakın zamanda yayımlanmış bir çalışma olarak Kübler, McDonald ve Nivre'nin [18] çalışmasına başvurulabilir.

Türkçe biçimbilimsel olarak zengin ve bitişken dillere klasik bir örnek niteliğindedir. Yoğun olarak çekimli olan biçimbilimsel yapısının yanı sıra, Türkçe'de gövdeye geldiğinde sözcük türünü değiştirebilen çok sayıda işlek yapım eki kullanılmaktadır. Bir sözcüğün farklı türetim aşamaları, sözcük çekimi konusunda zayıf olan İngilizce gibi bir dilde ayrı ayrı sözcüklere karşılık gelebileceğinden, Türkçe cümleler genellikle daha az sayıda ve daha çok çekimli sözcükler içerir. Türkçe cümlelerin sözdizimini daha uygun şekilde analiz edebilmek için sözcükler ortografik şekilleriyle işlenmez ve Şekil 1.1'de de bir örneğinin gösterildiği gibi türetim sınırlarından *çekim grubu* (ÇK) adı verilen sözcük altı birimlere bölünürler. Birden fazla ÇK'ya sahip sözcükler Türkçe'de çok yaygındır, hatta dört veya beş taneye kadar ÇK'ya sahip sözcüklere günlük sıradan cümlelerin içinde bile rastlamak mümkündür. Bundan dolayı bağlılık ilişkilerini daha doğru ifade etmek için ayrıştırma birimleri olarak sözcüklerden ziyade ÇK'ları kabul etmek daha uygun olmaktadır. ÇK'ların kullanımı Türkçe'nin ayrıştırılması için bir de facto standart haline gelmiştir ve Hakkani-Tür, Oflazer ve Tür [19], Oflazer [20], Oflazer ve diğ. [7], Eryiğit ve Oflazer [21] ve Eryiğit, Nivre ve Oflazer'in [3] de aralarında bulunduğu çok sayıda etkili çalışmada da bu gösterim yöntemi kabul edilmiştir.

Şekil 1.1 ve 1.2 aynı cümlenin Türkçe ve İngilizce için büyük ölçüde benzer dilbilgisel yapıda olan hallerinin işaretleme yöntemlerini karşılaştırarak bu kavramın belirgin bir örneğini oluşturmaktadır. Türkçe örnekteki 'arabadaydı' sözcüğünün sözcük içi bağlılığı olan **DERIV** ile birbirlerine bağlanan iki ÇK içerdiği gösterilmiştir. Bu bağlılık türüne sahip bağlılıklar ayrıştırıcıların bulmasında herhangi bir zorluk teşkil etmeyecek olduğundan dolayı değerlendirme sırasında hesaba katılmazlar. Örneklerden ayrıca İngilizce cümlelerdeki tanımlıklar için olan **DETERMINER** bağlılığı ve edatlar için kullanılan **ARGUMENT** bağlılığı gibi kolay bağlılıkların cümlenin Türkçe halinde sözdizimsel olarak gösterilmediği ve bunun yerine bu özelliklerin

biçimbirimsel bilgiden çıkarıldığı görülmektedir. Tüm bunlar özünde geri kalan bağılılıkları bulmanın zorluğunu artırmakta ve Türkçe'nin ayrıştırma başarımlarının tatmin edici olmamasına kısmi olarak etki etmektedir.

2. SORUNLAR VE ÖNERİLEN ÇÖZÜMLER

Bir işaretleme çerçevesinin tasarlanması, özünde amaç odaklı bir iş olmaktadır. Tasarlanacak nesnelerin ve tasarım amaçlarının net bir tanımını belirlemiş olmak esastır. Bağlılık gramerlerinin bağlamında bu nesneler bağlılık türlerine karşılık gelmekte, amaçlar ise hangi bağlılık türünün hangi durumlarda kullanılacağıyla ilgili kararlara tesir etmektedir. Bağlılık türleri ideal olarak açıklayıcı, dışlayıcı, anlaşılır ve özlü olmalıdır, ancak sık sık bu özelliklerin birini iyileştirmek diğerlerinden ödün vermeyi gerektirir ve bundan dolayı bir grameri bu noktalardan dengelemek ciddi bir zorluk teşkil etmektedir. Örneğin, daha az sayıda bağlılık türüne sahip olmak hem işaretleme hatalarını hem de ayrıştırma entropisini azaltırken bir yandan da ilgili gramerin anlamsal ifade gücünü azaltır. ODTÜ-Sabancı Ağaç Yapılı Derlemi'nin bahsedilen eksiklikleri göz önünde bulundurulduğunda tutarsızlık ve belirsizliklerin giderilebilmesi adına yeni bağlılık gramerinin açık ve minimal olması önceliklendirilmiştir.

Çalışmaya ODTÜ-Sabancı Ağaç Yapılı Derlemi'nin geniş çaplı bir incelemesiyle başlanmıştır. Bu sırada derlemin büyük bir kısmının üzerinden geçilerek incelenen cümleler üzerinde karşılaşılan tartışmalı noktalar kaydedilmiştir. Bu bağlamda en sık karşılaşılan durumlar, tutarsız veya kararsız şekillerde işaretlenmiş dilbilimsel kavramlar ile, resmi olmayan dilde kullanılması zorunlu olmayan bazı öğelerin varlığına bağımlı standart işaretleme yöntemleri olmuştur. Karşılaşılan bu durumların genel bir sınıflandırması, bunlar üzerindeki duruşumuz ve bu durumlarla karşılaşılan örnek cümleler bu bölümün devamında verilmektedir.

Belirli durumlara karşı çözüm oluşturması amacıyla yapılan yerel değişikliklere karar verme aşamasında literatürdeki diğer önemli ağaç yapılı derlemler ayrıntılı bir şekilde analiz edilmiş ve konuyla ilgili geçmiş çalışmalar incelenmiştir. Bu kapsamda

Çekçe Prague Ağaç Yapılı Bağlılık Derlemi [22], İsveççe Talbanken'05 [23], Macarca Szeged Ağaç Yapılı Derlemi [24] ve Felemenkçe Alpino Ağaç Yapılı Derlemi [25] ile, yaygın olarak tanınan ve kabul gören Stanford Bağlılıkları [26, 27] üzerinde, ilgili dilin Türkçe ile olan benzer dilbilgisel özelliklerinin nasıl modellendiği gözetilerek incelemeler yapılmıştır. Bunların yanı sıra McDonald ve diğerlerinin bağlılık ilişkilerini standartlaştırma üzerine olan çalışması [28], Schneider ve diğerlerinin bağlılık işaretlemesini sadeleştirmek konulu araştırması ve Martin ve diğerlerinin koordinasyon yapılarını işaretleme yöntemlerinin dağılımına yönelik incelemesi gibi, işaretleme yaklaşımlarıyla ilgili çok sayıda çalışma referans olarak alınmıştır. Tüm bu gözden geçirmeler sonucunda yeni işaretleme yöntemlerine karar verebilmek için sağlam bir altyapı oluşturulmuş, geçmiş birikime bakıldığında net bir çözümü belirleyen sorunlar konusunda ise, bahsi geçen çekinceler göz önünde bulundurularak, hem dilbilimsel hem de hesaplamalı olarak kolaylık sağlayabilecek sezgisel değişikliklere gidilmiştir.

Temel alınan işaretleme çerçevesinin arkasındaki fikirler, çerçeveyi daha evrensel bir yapıya uydurmaya yönelik bir çabadan uzak olarak, konu üzerindeki geçmiş çalışmalarımızın devamı niteliğindedir. Buna rağmen Stanford Bağlılıkları'nın giderek artan etkinliğinden dolayı Stanford işaretleme çerçevesi ve bu çerçevenin uygulamalarına özel olarak önem verilmiştir. Stanford Bağlılıkları'nı temel alan De Marneffe ve diğerlerinin [29] diller arası birleşik bir işaretleme çerçevesinin tasarlanması üzerine olan çalışması, yeni işaretleme çerçevemizi evrensel olarak daha iyi tanınan bir çerçeve cinsinden ifade edebilmemize olanak sağlamıştır. Önerilen çerçeve ile Evrensel Stanford Bağlılıkları arasındaki bir eşleştirme Tablo 2.1 üzerinde verilmiştir. Önerilen bağlılık türleri ve bu türlerin kullanımları hakkında ayrıntılı bilgiye Ek B'de verilen işaretleme kitapçığından ulaşılabilir.

Temel kabul edilen çerçeve üzerinde saptanan sorunlar sıradaki bölümlerin altında düzenlenmiştir. Her bir alt bölümde ilgili konu anlatılmakta ve örneklenmekte, ardından konuya ilişkin alınan önlemler gerekçeleriyle birlikte aktarılmaktadır. Önerilen değişiklikler, ilgili bağlılık türlerinin çıkarılması, eklenmesi, farklı türlere

Çizelge 2.1: İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi işaretleme çerçevesi ile Evrensel Stanford Bağlılıkları arasındaki eşleme.

İOSD	ESB	Notlar
SUBJECT	<i>nsubj</i>	—
SUBJECT	<i>csubj</i>	—
SUBJECT	<i>nsubjpass</i>	—
SUBJECT	<i>csubjpass</i>	—
OBJECT	<i>doj</i>	—
MODIFIER	<i>iobj</i>	—
OBJECT	<i>ccomp</i>	—
MODIFIER	<i>xcomp</i>	—
MODIFIER	<i>advmod</i>	—
MODIFIER	<i>advcl</i>	—
ARGUMENT	<i>neg</i>	Eylemler eklerle olumsuz yapılır. Ek eylemler 'değil' olumsuzluk edatının uydusu olur.
INTENSIFIER	—	Vurgulama ve katkı anlamı veren edatlar nitelenen ögeye bu bağlılıkla bağlanır.
DETERMINER, MODIFIER	<i>det</i>	İşaret sıfatları için DETERMINER , soru sıfatları için MODIFIER bağlılığı.
MODIFIER	<i>amod</i>	—
APPOSITION	<i>appos</i>	—
MODIFIER	<i>nummod</i>	—
MODIFIER	<i>relcl</i>	—
MODIFIER	<i>nfincl</i>	—
PREDICATE	<i>root</i>	—
PUNCTUATION	<i>punct</i>	Noktalama işareti olan öğeler cümle yüklemi yerine cümle kökünün uydusu olur.
—	<i>aux</i>	—
—	<i>auxpass</i>	—
—	<i>cop</i>	—
—	<i>expl</i>	—
RELATIVIZER	<i>mark</i>	Ana cümle yüklemi, yan cümle yüklemine bu bağlılıkla bağlanır.
VOCATIVE	<i>discourse</i>	Söylem unsurları cümle köküne bu bağlılıkla bağlanır. Emoticonlar noktalama sayılır.
VOCATIVE	<i>vocative</i>	—
—	<i>dep</i>	—
MODIFIER	<i>nmod</i>	—
—	<i>ncmod</i>	—
ARGUMENT	<i>case</i>	Edatlar yan cümlelerinin iyesi sayılır ve uyduları bunlara ARGUMENT etiketi ile bağlanır.
POSSESSOR, MWE	<i>compound</i>	Ad tamlamaları için POSSESSOR , sıfat tamlamaları için MODIFIER , sayısal ifadelr için MWE .
MWE	<i>name</i>	—
MWE	<i>mwe</i>	—
—	<i>goeswith</i>	—
—	<i>foreign</i>	—
—	<i>reparandum</i>	—
COORDINATION	<i>conj</i>	Bağlılık yayının yönü ters alınır.
CONJUNCTION	<i>cc</i>	Sol koordinasyon ögesinin yokluğunda ara bağlaçlar cümle köküne bağlanır.
COORDINATION	<i>parataxis</i>	Parataktik ilişkiler birer koordinasyon yapısı teşkil ediyormuş gibi bağlanır.
—	<i>list</i>	—
—	<i>remnant</i>	—
—	<i>dislocated</i>	—
DERIV	—	Bir sözcüğün her ÇK'sı sıradaki ÇK'ye bu bağlılıkla bağlanır.

bölünmesi, diğer türlerle birlikte gruplanması veya işaretleme yöntemlerinin değiştirilmesi şeklinde çeşitlilik göstermektedir. Bu bölümde irdelenen konular savımızı kuvvetlendirmek için yaptığımız deneyler için temel oluşturmaktadır. İlgili deneyler Bölüm 4 içerisinde anlatılmakta ve tartışılmaktadır.

2.1 Anlamsal Tutarsızlık

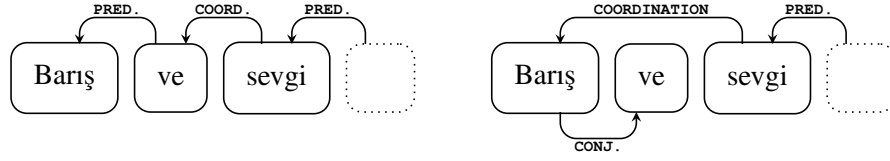
Orijinal çerçevede birtakım bağıllık türleri anlamsal çağrışımlarına uyumsuz olabilecek şekilde kullanılmaktadır. Bu durumlar özellikle sık kullanılan bağıllık türlerinin çok sık geçmeyen ikincil kullanımlarında görülmektedir. Bunların kendilerine özgü bağıllık türleriyle ifade edilmesi veya farklı bir geniş türe dahil edilmesinin olumsuz etkileri de olabilecek olsa bile, özellikle işaretleme aşamasında sebep olunacak zorluk düşünüldüğünde anlamsal tutarsızlığın olumsuz etkisi daha baskın olmaktadır. Bu türden sorunlar çok yaygın olmamakla birlikte yine de dikkat çekecek kadar sık görülmektedir.



Şekil 2.1: Nesneleri ifade eden **OBJECT** bağıllığının bir cümle nesnesinde (*solda*) ve bir edat tümlecinde (*sağda*) kullanımı. İkinci kullanım yeni bağıllık türü **ARGUMENT** kapsamına dahil edilmiştir.

Bu konuya bir örnek edat tümleçlerinin bağlanması nesneleri ifade eden **OBJECT** bağıllık türünün kullanılmasıdır. Edat tümleçlerinin uyduları zaman zaman tümleç nesnesi olarak adlandırılabilir, ancak bu uydular aslen edat görevindeki iyeyi niteleyen argümanlardır ve buna göre cümle ve ilgi yan cümlelerinin nesneleriyle herhangi bir bağlantıları yoktur. Orijinal çerçevede bu öğelerin nesne olarak kabul edilmeleri belirsiz bir durum yaratmakta, ayrıca bu ifade yöntemi ayrıştırıcıların öğrenme aşamasında kararsız kalmasına yol açmakta ve gerçek nesnelerin bulunmasını dahi zorlaştırmaktadır. Diğer tüm yan cümle argümanlarıyla

birlikte bu ögeler de **ARGUMENT** türü kapsamına dahil edilmiştir.



Şekil 2.2: Koordinasyon yapıları için orijinal (*solda*) ve önerilen (*sağda*) işaretleme yöntemleri. İkinci yöntemin gerek ayrıştırma kolaylığı açısından, gerekse anlamsal açıdan daha uygun olduğu savunulmaktadır.

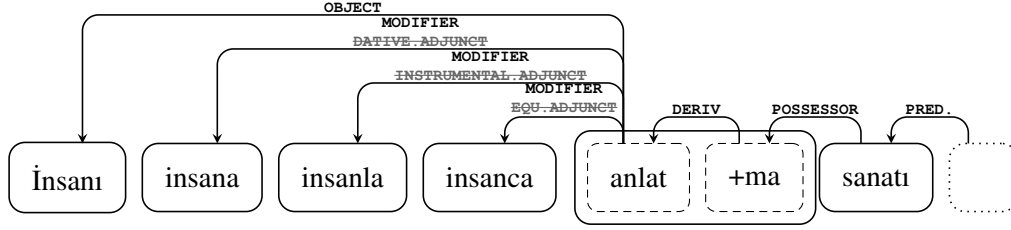
Diğer bir durum ise koordinasyon yapılarında görülmektedir. Orijinal çerçevede bu yapılarda koordinasyon içerisinde olan ögelerin birbirine bağlanması yerine koordinasyon sağlayan bağlacın ikinci ögeye **COORDINATION** etiketi ile bağlanması sözkonusudur. Bağlılığa anlamsal olarak bakıldığında bu durum, bağlacın kendisinin ögelerden biriyle bir koordinasyon yapısı oluşturduğunun çıkarımına yol açmaktadır. Bu açıdan, Ambati, Reddy ve Kilgarrieff'in [30] çalışmasında da değinildiği gibi, koordinasyon içindeki ögelerin birbirine bağlandığı bir gösterim yöntemi daha uygun olmakta ve önerilen yöntemde de **COORDINATION** etiketi bu doğrultuda kullanılmaktadır.

2.2 Hiyerarşi ve Çakışma



Şekil 2.3: İki benzer deyimsel kullanımın **ETOL** ve **COLLOCATION** bağlılık türleri ile gösterilmesi. Önerilen çerçevede tüm benzer deyimsel kullanımlar genel bağlılık türü **MWE** altında işlenmektedir.

Temel alınan işaretleme çerçevesinde bazı bağlılık türleri diğer birtakım bağlılık türlerinin kapsamı ile çakışmaktadır. Gramerin kendisi bağlılık türleri arasında herhangi bir hiyerarşi oturtmadığından işaretlemede farklı ayrıntı düzeyleri işletilememekte ve nihayetinde bu kavramın da olumsuz bir etkisi olmaktadır.



Şekil 2.4: Niteleyici görevi gören ad soylu belirteçlerin hallerine göre farklı **X.ADJUNCT** etiketleri ile ifade edilmesi. Bu tür kullanımlar genel bağıllık türü **MODIFIER** kapsamına alınmıştır.

Bu durum öncelikli olarak işaretleyicileri etkileyerek belirli durumlarda hangi bağıllık türünün kullanılacağını keyfi bir tercihe bırakmakta, daha sonra da ayrıştırıcıya yansiyarak yüksek entropiye yol açabilmektedir. Diğer yandan, önemli bir etki de ayrıştırıcı çıktılarının değerlendirilme aşamasında ortaya çıkmaktadır. Bağıllık işaretlemelerinin kaba ve ince türlere ayrılarak kullanılamaması eş olmayan bağıllık türlerinin birbiriyle her zaman uyumsuz görülmesine yol açmakta, özünde mantıklı olan bazı bağıllık atamalarının hatalı kabul edilmesine neden olmaktadır. Buna verilebilecek örneklerden biri, normalde **COLLOCATION** türünün tamamen kapsadığı, etmek ve olmak gibi yardımcı eylemlerle türetilen çok sözcüklü ifadelerle atanan **ETOL** etiketidir. Bu tür bir kapsam çakışması karşısında, altın standart işaretlemede **COLLOCATION** etiketi atanan bir bağıllığa ayrıştırıcı tarafından **ETOL** etiketi atanması durumunda, atama geçerli olmasına rağmen değerlendirilme sonucunda hatalı olarak görülür. Bu tür çakışmalar önerilen işaretleme çerçevesinde tümüyle kaldırılmıştır.

Bazı bağıllık türleri ise diğer bazı bağıllık türleriyle kullanım itibarı ile fazlaca benzeşmekte ve karışıklığa sebep olmaktadır. Bu duruma verilebilecek en belirgin örnek, niteleyicilere ayrılmış olan genel **MODIFIER** türü karşısında, yalnızca ad soylu niteleyicilere hallerine göre atanması öngörülen **X.ADJUNCT** türleridir. Örneğin, herhangi bir alt sınıfa ait olmayan belirteçlerin geneli aynı çerçevede **MODIFIER** etiketi ile işaretlenirken, bulunma halindeki sıradan bir ad soylu belirteç **LOCATIVE.ADJUNCT** etiketini kullanmaktadır. Dahası, bulunma halindeki birtakım ad soylu belirteçler deyimse anlama geldiklerinde, bunlar için anlamsal kaygılardan

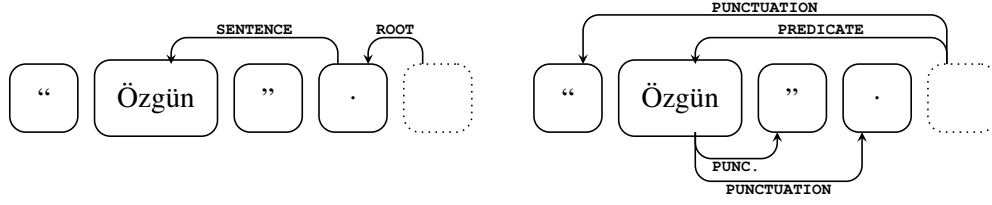
dolayı yine de **MODIFIER** etiketi tercih edilebilmektedir ve bu tercihin ne zaman yapılması gerektiği de belirlenmemiştir. Bu karmaşıklıklara bir yanıt olarak, önerilen çerçevede yalnızca **MODIFIER** etiketi tutulmuş ve zaten biçimbilimsel bilgiden de ayırdına varılabilen tüm **X.ADJUNCT** etiketleri kullanımdan çıkartılmıştır.

Bunların yanında, çok sözcüklü ifadeler karşılık gelen genel anlamsal bağıllık türü **MWE** ile hemen hemen tüm sözdizimsel bağıllık türleri arasında doğal bir çakışma söz konusudur. Buna rağmen, Eryiğit, İlbaş ve Can [2] gibi çalışmalarda da yararının gösterildiği üzere, ayrıştırma sırasında kararsızlığa sebep olacağı halde ifade gücünü zayıflatmamak açısından **MWE** türünün kullanımı yeni çerçevede devam ettirilmektedir.

2.3 Belirsizlik

Bazı öğelerin bağlanacakları iye kesin olarak belirli olsa dahi, zaman zaman arada olması gereken bağıllık türü çok belirgin olmayabilir (ya da bunun tam tersi gerçekleşebilir). Bu tür durumlar keyfi işaretlemelere yol açmakta veya işaretleyiciler tarafından öğrenilmesi oldukça zor olan ayrıntılı işaretleme yordamlarının oluşturulmasını zorunlu kılmakta, bu nedenle eğitim kümesinin tutarlılığına dolaylı olarak zarar vermektedir. Nadiren hiçbir etiketle doğru şekilde ifade edilemeyecek bağıllık ilişkileri olabilse de, bu durum çoğunlukla birden fazla etiketle eşit derecede doğru ifade edilebilecek bağıllıklarda görülmekte, bu durumlarda aradaki ayrım belirsiz olabilmektedir. Yeni çerçevede, sık görülen durumları kapsayacak yeni bağıllık türlerinin kullanılması veya belirsizlik içeren işaretleme için kullanımlara bir netlik getirilmesi sağlanmıştır.

Bu kavramın bir örneği hiçbir bağıllık türü ile tam olarak ifade edilemeyen ve yerine göre **MODIFIER** veya **OBJECT** türleri altında işlenen tümleş argümanlarında görülebilir. Yeni işaretleme çerçevesinde bu tür öğeler için **ARGUMENT** türü getirilmiş, bir iyede birden fazla bulunabilen veya hiç bulunmayabilen niteleyici bağıllıklarından farklı olarak, iyenin kesin ve tam olarak bir tane uydu beklediği edat



Şekil 2.5: Belirli noktalama işareti türlerinin herhangi bir iyeye bağlanmamasına izin verilmesi (solda). Yeni bağıllık türü **PUNCTUATION** her türden noktalama işaretini kapsayacak ve standart bir işaretleme yöntemi olacak şekilde önerilen çerçeveye dahil edilmiştir.

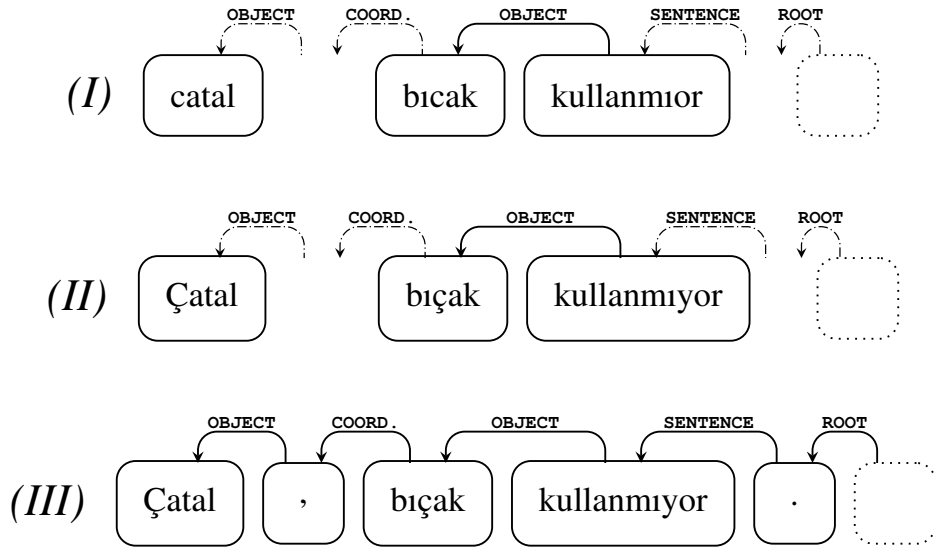
tümleci gibi yapılarda bu bağıllığın kullanılması öngörülmüştür.

2.4 İsteğe Bağlılık

Bağıllık gramerleri gerekirci biçimsel gramerler olduğundan isteğe bağlı herhangi bir durum barındırmamalıdır. Orijinal çerçevede noktalama işareti olan öğeler için birer bağıllık tanımlamak zorunlu olmadığından, bu durum açık şekilde ihlal edilmiştir. Yalnızca belirli noktalama işareti türlerinin, bağlaç görevi gören işaretlerin ve cümle sonu noktasının kullanımıyla kesin olarak ilişkilendirilmiş bağıllık türleri mevcuttur ve diğer noktalama işaretlerinin bir iyeye sahip olmadan devam etmelerine izin verilmektedir. Söz edilen grupların dışında kalan noktalama işaretleri cümle sözdiziminin dışında kabul edilir ve keyfi birer öğeye NOTCONNECTED bağıllığı ile bağlanırlar. Bu durum temel olarak bağıllık gramerinin cümlelerin tüm öğeleri için bir bağıllığı zorunlu tutmadığı anlamına gelmekte ve bu açıdan çoğunluk bağıllık ayrıştırıcısıyla ters düşerek ayrıştırma sonuçlarının değerlendirmesi konusunda zorluğa neden olmaktadır. Bunun yanında, NOTCONNECTED bağıllığı ayrıştırma sırasında olağan bir bağıllık türü olarak görüldüğünden öğrenme başarımları da buradan dolaylı olarak etkilenmektedir. Bu konuya bir çözüm olarak yeni çerçevede noktalama işaretleri için **PUNCTUATION** etiketi getirilmiştir ve bu yolla tüm noktalama işaretlerinin bağıllığı standartlaştırılarak gramerden isteğe bağlı işaretleme kaldırılmıştır.

2.5 Düzen Aşırılığı

Orijinal çerçevedeki belirli bağıllık şemaları birtakım öğelerin cümle içinde değişmez şekilde belirli konumlarda bulunmasına bağlıdır ve bu öğeler yerlerinde bulunmadığında işaretlemelerin doğru şekilde yapılması mümkün olmamaktadır. Bu tür varsayımlara aykırı düşen durumlarda ilgili bağlantıların alternatif olarak nasıl gösterileceği konusunda belirsiz durumlar ortaya çıkmış olur ve buradan işaretleme tutarsızlıkları doğar. Bu şekildeki aykırı durumlar günlük veya kuralsız dilde nadir değildir. Özellikle web jargonunda kısa yazım amacıyla cümlelerin bazı öğeleri ihmal edildiğinde, hatta bazı kurallı cümlelerde dahi yaygın olmayan ifadeler ya da deyimssel veya eski kullanımların varlığında bu tür durumlar gözlenebilmektedir. Bu sebeple bu konuya değinilmesi de önemlidir.



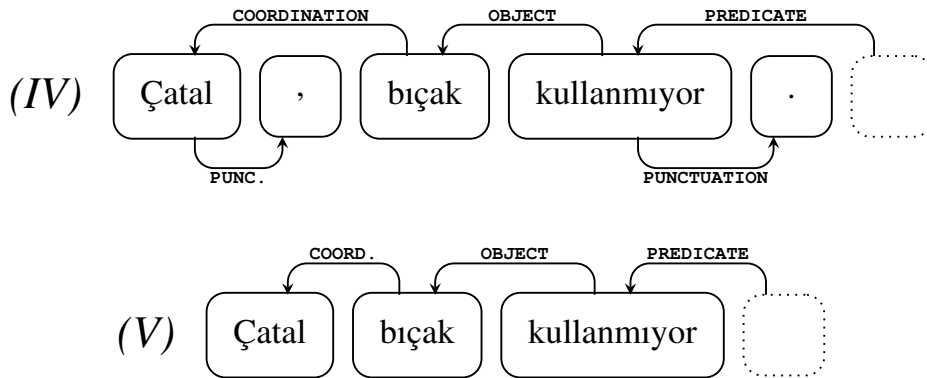
Şekil 2.6: Orijinal işaretleme çerçevesinde ihmal edilebilen öğelere bağıllık

Cümle (I) kurallı dilde gerek duyulacak, ancak web üzerinde yazımı sıklıkla ihmal edilen belirli öğelere bağıllık yüzünden işaretlemenin imkansızlaştığı iki durumu göstermektedir. Cümle (II)'de gösterilen otomatik normalizasyon sonrasında dahi, Cümle (III)'teki gibi bağlaç ve son nokta görevi gören noktalama işaretleri eklenmediği takdirde kurallı işaretleme mümkün olmamaktadır.

Düzen aşırılığı durumu, Şekil 2.6 üzerinde de örneklendiği üzere, ana yüklemnin işaretlenmesinde ihtiyaç duyulan cümle sonu noktalama işaretlerinde en belirgindir. Cümle kökü, orijinal çerçevede cümlemin ana yüklemnin son noktalama işaretine

SENTENCE bağılılığı ile, bu işaretin ise köke **ROOT** bağılılığı ile bağlanması yoluyla belirtilmektedir. Bu şema, kuralsız dilde sıklıkla yapıldığı gibi cümle sonu noktasının görmezden gelinmesini tolere edemez. Bunu çözmek için yüklemlerin yeni **PREDICATE** bağılılığı ile doğrudan kök düğümüne bağlandığı yeni bir işaretleme biçimi getirilerek Şekil 2.7 üzerinde görüldüğü gibi son noktalama işaretinin kullanılmayabileceği durumlara tolerans sağlanmıştır.

Diğer bir yaygın örnek ise koordinasyon yapıları ve ilgi yan cümlelerinde, sırasıyla **COORDINATION** ve **RELATIVIZER** bağılıklarıyla ilgili olarak görülmekte ve yine Şekil 2.6 üzerinde gösterilmektedir. Bu şemaların her ikisi de koordinasyon içinde bulunan öğelerin (ilgi yan cümleciklerinde cümle yüklemlerinin) arasında bir ara bağlacın olmasını şart koşar ve bu öğeyi seri bir bağılılık zinciri içerisinde koordinasyon içindeki öğeler arasında bir köprü olarak kullanır. Bağlaçlar ve bağlaç görevi gören noktalama işaretleri de cümle içinde sık sık ihmal edildiğinden, bağlaçların ayrı işlendiği ve koordinasyon yapısı içindeki öğelerin birbirine bağlandığı bir işaretleme yöntemi getirilerek Şekil 2.7 üzerinde gösterildiği gibi bağlaç görevindeki öğenin yokluğunda da bu yapıların işaretlenmesine olanak sağlanmıştır.



Şekil 2.7: Önerilen işaretleme çerçevesinde öğe bağımsızlığı yaklaşımı

Şekil 2.6 üzerinde görüldüğünün aksine, aynı cümlelerin işaretlenmesi için önerilen etiketler kullanıldığında diğer öğelere bağımlılık oluşmamaktadır. Bu örnek koordinasyon yapıları ve ana cümle yüklemlerinin bağlayıcı noktalama işaretlerinin varlığında ve yokluğunda, birimlerin sırasıyla Cümle (IV) ve Cümle(V) üzerinde görüldüğü şekilde bağlandığı durumları göstermektedir.

3. YENİ DERLEMLER

3.1 Ön Bilgiler

Önerilen yeni bağıllık gramerinin yeni kaynakları kullanacak gelecek çalışmalara etkisi hakkında önceden bir fikir sahibi olmak için, yeni işaretleme çerçevesine bağlı kalınarak baştan iki derlem işaretlenmiştir. Öncelikle orijinal ve önerilen iki işaretleme yönteminin uygun şekilde karşılaştırılabilmesi için ODTÜ-Sabancı Ağaç Yapılı Derlemi yeni bağıllıklarla en baştan işaretlenmiştir. Geçmişte bu derlem üzerine sarf edilen önemli çaba ve bu çalışmada yapılan katkı göz önünde bulundurularak bu derleme İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi adı verilmiştir. Devamında, web üzerinde serbestçe bulunan çeşitli bağlamlardaki verilerin bir araya getirilmesiyle yeni bir derlem oluşturulmuş, ardından Pamay ve diğerlerinin [31] çalışmasında anlatıldığı şekilde yeni çerçeveye göre işaretlenmiştir. Bu derleme İTÜ Ağaç Yapılı Web Derlemi adı verilmiştir. Bu bölümde bu kaynaklar¹ hakkında ayrıntılara yer verilmektedir.

Yeni derlemlerin her ikisinin de işaretlenmesi İTÜ İşaretleme Aracı'nın [32] güncellenmiş sürümü üzerinde yürütülmüştür. Bu derlemlerin işaretlenme sürecinde dilbilimsel açıdan yüksek yeterliğe sahip üç işaretleyici çalışmıştır. İşaretleyiciler derlemlerin kendilerinin işaretlenmesine başlamadan önce iki haftalık bir eğitim sürecinden geçirilmiştir. İlk olarak İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi için biçimbilimsel çözümlemeleri önceden atanmış altın standart birimler derlemin önceki sürümünden temin edilmiş, geriye kalan bağıllık işaretlemeleri ise iki aylık bir süreç içerisinde tamamlanmıştır. Bunun devamında İTÜ Ağaç Yapılı Web Derlemi için işaretleyiciler sürece web kaynaklarından çekilmiş standart dışı ham cümlelerden başlamak durumunda kalmışlardır. Cümleler elle birimlerine ayrılarak normalize edilmiş, ardından her bir birimin biçimbilimsel çözümlemeleri atanmış ve ancak

¹Bu çalışmada tanıtılan her iki ağaç yapılı derlem de <http://tools.nlp.itu.edu.tr> adresinden araştırma amaçlı olarak edinilebilmektedir.

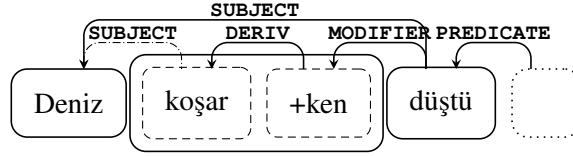
bu aşamadan sonra birimler arasındaki bağılıklar işaretlenmeye başlamıştır. Bu süreçlerden dolayı derlemin geliştirilmesi dört aylık bir süreçte gerçekleşmiştir. İşaretleme süreçleri sırasında ve sonrasında her iki derlemin de cümleleri incelenmiş, cümlelerdeki hatalı ve tutarsız işaretlemeler saptanmış ve ardından iki hafta süren bir düzeltme aşamasının sonunda kaynakların son halleri oluşmuştur.

Sözdizimsel katmanın altyapısını oluşturan biçimbilimsel katman için, önerilen yeni bağılıklar türleriyle daha uyumlu olması beklenen, Eryiğit'in [33] çalışmasında anlatılan güncellenmiş biçimbilimsel etiket kümesi ve ilgili biçimbilimsel çözümleyici kullanılmıştır. İTÜ Ağaç Yapılı Web Derlemi'ni oluşturan cümleler Eryiğit ve Torunoğlu-Selamet'in [34] çalışmasında ortaya konulan yönteme bağlı kalınarak elle normalize edilmiştir. İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi halihazırda kurallı cümlelerden derlendiği için bu normalizasyon süreci yalnızca İTÜ Ağaç Yapılı Web Derlemi'nin ham cümleleri için işletilmiştir. İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi, ODTÜ-Sabancı Ağaç Yapılı Derlemi'nin biçimbilimsel çözümlerinin güncellenmiş biçimbilimsel etiket kümesine otomatik olarak eşlenmiş halinin üzerine kurulurken, İTÜ Ağaç Yapılı Web Derlemi'ne ait birimlerin biçimbilimsel çözümleri baştan çıkartılmış ve her birim için olası çözümler arasından elle seçim yapılmıştır.

3.2 Çok İyeli İşaretleme

İşaretleme yöntemimizle ilgili değinilecek önemli bir konu da yeni derlemlerde *derin* (veya *kısıtsız*) bağılıkların da işaretlenmiş olmasıdır. Derin bağılıklar birimlerin olağan *yüzey* bağılıklarının yanında diğer mantıksal iyelerle, genellikle de farklı bağılıklar türleri ile oluşturdıkları ikincil bağılıklardır. Derin bağılıkların işaretlenmesi, her ögenin tek bir iyesinin olması kısıtını ihlal ederek derlemi önışleme olmaksızın çoğu sözdizimsel ayrıştırıcıyla uyumsuz hale getirmekle birlikte, farklı bir iyeyle olan bağlantının bir bağılılıkla açık bir şekilde *işaretlenmek* yerine mantıksal olarak *ima edilmesi* durumunda oluşan ifade eksikliğini kapatmaktadır. Örneğin, ana cümlenin yüklemine özne olan bir birim, yan cümle içerisindeki

bir eylemin de mantıksal öznesi olabileceksen, bu birim ancak tek bir yüzey bağıllığı kurabileceğinden yalnızca daha önemli görülen ana cümle yüklemiyle ilişkilendirilecektir. Şekil 3.1, bu durumdaki bir cümleyi göstererek, Türkçe’de bu tür yan cümlelerde ana cümle öznesiyle aynı varlığa işaret eden ayrı bir birim hiçbir zaman belirtilmediğinden dolayı derin bağıllıkların kullanımının anlamlı olacağı çok yaygın bir durumu örneklemektedir.



Şekil 3.1: Derin bağıllıkların işaretlenmesinin gerekli olduğu durumlardan birini gösteren bir örnek. Özne görevindeki öge ‘Deniz’ cümledeki her iki eylemi de nitelemektedir.

Derin bağıllıkların işaretlenmesi genellikle anlamsal ayrıştırıcıların eylem argümanlarının anlamsal rollerini çıkarması sırasında kullanılabileceği tüyolar oluşturduğundan dolayı tercih edilmektedir. Buna göre, yeni derlemlerde derin bağıllıkların getirilmesinin altındaki en büyük motivasyon, derlemleri olası anlamsal rol çıkarımı çalışmaları için kullanılabilir hale getirmek olmuştur. Bununla birlikte sözdizimsel ayrıştırma sırasında tek iye kısıtından dolayı meydana gelen yanlış negatif çıktıların azaltılması da ayrı bir motivasyon teşkil etmiştir. Böyle alışlagelmiş bir kısıtın olmasına rağmen, karmaşık anlamsal bağlantılardan dolayı bir ögenin ilişkilendirilebileceği eşit derecede anlamlı birden fazla bağıllık olan durumlar olabilir. Bu durumlarda altın standart doğrulama kümesinde geçerli bir iyeye olan bir bağıllık işaretlenmişken, sözdizimsel ayrıştırıcılar yine geçerli olan fakat farklı bir iyeye bağıllık kurabilir. Nadir olmayan bu senaryolarda ayrıştırıcı çıktısı mantıksal olarak doğru olsa da değerlendirmede yanlış kabul edilmek zorunda kalmaktadır. Yeni derlemlerde yüzey bağıllıkları ve derin bağıllıklar arasında bir hiyerarşi kurulmamış, yalnızca her bir öge için birden fazla iyenin desteklendiği bir işaretleme yöntemi getirilerek tek bir ögenin ilişkilendirilebileceği birden fazla bağıllığı doğru kabul

edebilecek ayrı bir doğruluk metriğinin kullanılabilmesi amaçlanmıştır.

3.3 Derlem İstatistikleri

Tanıtilan derlemlerin etkisini göstermek ve mevcut en yüksek başarımlı sistemlerle makul karşılaştırmalar yapabilmek için, temel alınan ODTÜ-Sabancı Ağaç Yapılı Derlemi'nin yanı sıra yeni derlemler üzerinde çok sayıda deneysel ölçüm yapılmıştır. Öncelikle tüm derlemler üzerinde belirli değerlerin karşılaştırmalı istatistikleri verilmiş, ardından ayrı bir bölümde çeşitli sözdizimsel ayrıştırma modellerinin değerlendirdiği deneysel sonuçlar raporlanmıştır.

Tablo 3.1 derlemlerdeki cümle, birim ve bağıllık sayılarını yan yana vererek önemli çıkarımların yapılmasına olanak tanımaktadır. Tablo 3.2 İTÜ Ağaç Yapılı Web Derlemi'ni oluşturan cümlelerin çekildikleri web sitesi kategorilerine göre dağılımını vermektedir. Tablo 3.3 tüm derlemlerdeki bağıllıkların bağıllık türlerine göre dağılımlarını göstermektedir. Yalnızca bir tire ile işaretlenmiş hücreler ilgili bağıllık türünün o derlemde kullanılan bağıllık gramerinde kullanılmadığına işaret eder.

Çizelge 3.1: Temel alınan ve yeni tanıtilan derlemlerdeki cümle, birim ve bağıllık istatistikleri.

	ODTÜ-SABANCI A. Y. DERLEMI	İTÜ-ODTÜ-SABANCI A. Y. DERLEMI	İTÜ AĞAÇ YAPILI WEB DERLEMI
# Cümle	5635	5635	5009
# Sözcük	56424	56424	43199
# Birim (ÇK)	67403	63089	47245
# Tek İyeli Birim	67403 (%100,0)	60688 (%96,2)	45357 (%96,0)
# Çok İyeli Birim	—	2401 (%3,8)	1888 (%4,0)
# Bağıllık (DERIV hariç)	56424	59425	46136
# Bağıllık (DERIV dahil)	67403	66090	50181
# İzdüşümsel Bağıllık	66145 (%98,1)	64663 (%97,8)	49521 (%98,7)
# İzdüşümsel Olmayan Bağıllık	1258 (%1,9)	1427 (%2,2)	660 (%1,3)

Çizelge 3.2: İTÜ Ağaç Yapılı Web Derlemi’ni oluşturan cümlelerin alan dağılımı.

KATEGORI	# CÜMLE
<i>Haber Yorumları</i>	1049
<i>Kişisel Blog Yorumları</i>	1005
<i>Müşteri Ürün Yorumları</i>	1019
<i>Sosyal Medya Gönderileri</i>	981
<i>Tartışma Forumu Gönderileri</i>	956

Çizelge 3.3: Temel alınan ve yeni tanıtılan derlemlerdeki bağılılık türü etiketlerinin dağılımları.

	ODTÜ-SABANCI A. Y. DERLEMI	İTÜ-ODTÜ-SABANCI A. Y. DERLEMI	İTÜ AĞAÇ YAPILI WEB DERLEMI
ABLATIVE.ADJUNCT	523 (%0,8)	—	—
APPPOSITION	202 (%0,3)	91 (%0,1)	16 (%0,0)
ARGUMENT	—	1805 (%2,7)	1615 (%3,2)
CONJUNCTION	—	1360 (%2,1)	963 (%1,9)
CLASSIFIER	2050 (%3,0)	—	—
COLLOCATION	73 (%0,1)	—	—
COORDINATION	2476 (%3,7)	3078 (%4,7)	2896 (%5,8)
DATIVE.ADJUNCT	1361 (%2,0)	—	—
DERIV	10979 (%16,3)	6665 (%10,1)	4045 (%8,1)
DETERMINER	1952 (%2,9)	2180 (%3,3)	1930 (%3,8)
EQU.ADJUNCT	16 (%0,0)	—	—
ETOL	10 (%0,0)	—	—
FOCUS.PARTICLE	23 (%0,0)	—	—
INSTRUMENTAL.ADJUNCT	271 (%0,4)	—	—
INTENSIFIER	903 (%1,3)	1070 (%1,6)	866 (%1,7)
LOCATIVE.ADJUNCT	1142 (%1,7)	—	—
MODIFIER	11690 (%17,3)	15516 (%23,5)	12225 (%24,4)
MWE	2432 (%3,6)	3552 (%5,4)	2860 (%5,7)
NEGATIVE.PARTICLE	160 (%0,2)	—	—
OBJECT	8338 (%12,4)	5094 (%7,7)	3272 (%6,5)
POSSESSOR	1516 (%2,2)	4070 (%6,2)	2357 (%4,7)
PREDICATE	—	5741 (%8,7)	5187 (%10,3)
PUNCTUATION	—	10375 (%15,7)	6210 (%12,4)
QUESTION.PARTICLE	289 (%0,4)	—	—
RELATIVIZER	85 (%0,1)	129 (%0,2)	98 (%0,2)
ROOT	5644 (%8,4)	—	—
S.MODIFIER	597 (%0,9)	—	—
SENTENCE	7261 (%10,8)	—	—
SUBJECT	4481 (%6,6)	5174 (%7,8)	4333 (%8,6)
VOCATIVE	241 (%0,4)	190 (%0,3)	1308 (%2,6)
(BAĞLANMAYAN BİRİMLER)	2688 (%4,0)	—	—

Tablo 3.3 üzerinde görüldüğü üzere İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi ve İTÜ Ağaç Yapılı Web Derlemi'nden elde edilen figürlerin yüzeysel bir karşılaştırması web alanında kullanılan günlük konuşma dilinin resmi bir dile sahip olan düzenlenmiş metinlerden belirgin derecede farklı olduğuna işaret etmektedir. Bu farklılıkların arasında en önemlisi muhtemelen noktalama işareti kullanımındır. İTÜ Ağaç Yapılı Web Derlemi İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi'nden %3.3 daha az noktalama işareti içermektedir. Diğer bir belirgin farklılık ise ünlemlerin kullanımındadır (**VOCATIVE** bağıllığı). İTÜ Ağaç Yapılı Web Derlemi'nde %2.6 olarak görülen ünlem yoğunluğu İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi'nde ölçülen %0.3'e göre çok daha fazladır. Diğer bir farklılık ise **DERIV** bağıllığı yoğunluğunun her iki yeni derlemde de yeni birimlerin biçimbilimsel çözümlemelerinden dolayı hatırı sayılır derecede düşmüş olmasıdır. İTÜ Ağaç Yapılı Web Derlemi'nin İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi'ne göre daha da düşük bir ÇK yoğunluğu göstermesinden ise web jargonunun daha kısa ve az türetimli sözcükleri tercih ettiği çıkarılabilir.

4. DENEYLER

Bu bölüm yeni derlemlerimiz ve ODTÜ-Sabancı Ağaç Yapılı Derlemi'nin temel sürümü üzerinde gerçekleştirilen karşılaştırmalı istatistiksel analizleri vermektedir. Bölümde testlerin sonucunda alınan figürler sunulmakta ve her test grubu için ayrı birer bölümde testlerin önemi yorumlanmaktadır. Analizlerin amacı kısmen derlemler üzerindeki ayrıştırma başarımlarını karşılaştırmak, kısmen de bağlılık işaretleme yöntemlerinde orijinal işaretleme çerçevesinin üzerinden gidilen büyük değişikliklerin etkilerini yerel olarak ölçmektir. Yeni derlemler üzerindeki başarımların daha geniş görülebilmesi için alternatif ayrıştırma modelleri eğitilmiş ve bu modellerin başarımları çeşitli metriklere göre ölçülerek sunulmuştur.

Bölüm 4.1 öğrenme, ayrıştırma ve değerlendirme sistemlerine ilişkin ön bilgileri içermektedir. Bölüm 4.2 İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi üzerinde gerçekleştirilen genel doğruluk testlerine ayrılmıştır. Bölüm 4.3'te İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi'nde yürütülen yerel testlerin sonuçları yorumlanmakta ve üç büyük işaretleme yöntemi değişikliğinin etkileri bu yolla gösterilmektedir. Son olarak, Bölüm 4.4'te İTÜ Ağaç Yapılı Web Derlemi üzerindeki doğruluk skorları sunulmakta ve derlemi meydana getiren cümlelerin alan özgünlüğü hakkında incelemelere yer vermektedir.

4.1 Ön Bilgiler

Tüm testlerimizde Eryiğit ve diğerlerinin baz alınan çalışmasında [2] kullanılan MaltParser [35] yapılandırması aynı şekliyle hem öğrenme hem ayrıştırma için kullanılmış ve sonuçların bu çalışmayla uygun biçimde karşılaştırılabilmesine olanak tanınmıştır. Yine atıfta bulunulan çalışmayla uyumluluk amacıyla tüm eğitim kümelerinden izdüşümsel olmayan cümleler çıkarılmıştır. Bu uygulamanın

önemli bir başarıml artışı sağladığı aynı zamanda Eryiğit ve diğlerlerinin [3] çalışması ile diğler birçok geçmiş çalışmada gösterilmiştir. Morfosentaktik (hem biçimbilimsel hem de sözdizimsel nitelikteki) **DERIV** bağılılıklarının ayrıştırma sırasında tespiti basit olduğundan, bu türden bağılılıklar doğruluk ölçümlerinde göz önünde bulundurulmamıştır. Literatürde başarıml ölçümlerinde noktalama işaretlerinin işlenme durumuna ise farklı yaklaşımlar mevcuttur: (1) noktalama işaretlerinin ölçümlerde değlerlendirilmemesi (örn. [11]), (2) noktalama işaretlerinin de değlendirmeye katılması (örn. [12]). Önerilen işaretleme çerçevesinde noktalama işaretlerine de anlamlı bir bağılılık türü atanması üzerine bunların da ölçümlerde değlendirilmesi önem kazandığından bu çalışmada ikinci yöntem kabul edilmiştir ve noktalama işaretleri de değlendirmeye katılmaktadır.

Önceki çalışmalardan devralınan ayrıştırma çerçevesi birden fazla iyesi işaretlenmiş uyduların öğrenilmesini desteklemediğinden dolayı bu çalışmada öğrenme sürecinden önce her uydu için tek bir iyei filtreleyen önışleme rutinleri çalıştırılmıştır. İye seçimi için öncelikle En Yakın İye ve En Uzak İye adını verdiğimiz iki temel yöntem işletilmiştir. En Yakın İye seçimi en soldaki ileri (cümle sırasına göre uydudan sonra gelen) iyei, veya ileri iyelerin olmaması durumunda en sağdaki geri (uydudan önce gelen) iyei tercih etmektedir. Buna karşılık En Uzak İye seçimi en sağdaki ileri iyei, veya ileri iyelerin olmaması durumunda en soldaki geri iyei seçmektedir. Öğrenme için optimal iyei seçen daha karmaşık yöntemlerin tasarlanması da mümkündür, ancak bu tür bir tasarım bu çalışmanın kapsamı dışında kalmaktadır.

Ölçümlerde kullanılan metrikler alışlagelmiş ÇK-tabanlı etiketli ve etiketsiz bağlanma skorlarıdır. Etiketsiz bağlanma skorunda (BS_{E-}) yalnız iye birimin doğru bulunması durumunda bir tahmin doğru kabul edilmekte iken, etiketli bağlanma skoru (BS_{E+}) için aynı zamanda uydu ile iye arasındaki bağılılık türünün de doğru bulunması gerekmektedir. Bu ikisinin arasında yüksek etiketli bağlanma skorlarının elde edilmesi daha zor ve dolayısıyla daha değerli olduğundan, BS_{E+} skorları başarıml karşılaştırmasında birincil metrik olarak kullanılmaktadır. Farklı modeller arasında etiketsiz bağlanma skorlarındaki olası farklılıkların da görülebilmesi için bu skorlar

da tutulmaktadır. Çapraz doğrulama deneylerinde test kümelerindeki standart hatalar ayrıca verilmektedir. Gerekli görülen yerlerde istatistiksel anlamlılık ölçümü için McNemar'ın eşli t-testi kullanılmıştır.

Klasik doğruluk skoru ölçümünde (*katı değerlendirme*) her zaman her birimden tek bir altın standart bağlılık çıkmaktadır ve bu bağlılığın bulunan bağlılıkla eşleşip eşleşmediği kontrol edilir. Bu kısıt bağlılık ayrıştırması için kolay ve etkili bir değerlendirme modeli teşkil ederken, Bölüm 3.2 altında da irdelendiği üzere bir yandan yanlış negatiflere yol açabilmektedir. Bu tür yanlış negatiflerden kaynaklanan başarımların kayıplarını hafifletmek adına bu çalışmada yeni derlemlerdeki çok iyeli işaretlemelerden yararlanılmakta ve her birimden çıkan tüm altın standart iyelerin değerlendirildiği bir ölçüm metriği kullanılmaktadır. Bu yöntemde (*gevşetilmiş değerlendirme*) bulunan bağlılığın aynı birimden çıkan altın standart bağlılıklardan herhangi biri ile eşleşmesi durumunda bağlılık doğru kabul edilmektedir. Gevşetilmiş değerlendirmenin de katı değerlendirme gibi etiketli ve etiketsiz biçimleri bulunmaktadır. Bu bölümde her bir iye seçme yönteminin çıktısı üzerinde katı değerlendirmenin yanı sıra gevşetilmiş değerlendirme ile elde edilen doğruluk skorları da verilmektedir.

4.2 İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi

Tablo 4.1 İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi üzerindeki ayrıştırma başarımlarını göstermektedir. Sonuçlar derlemin iki temel tek iyeli hali üzerinde on katlı çapraz doğrulama yapılması ile çıkarılmıştır. İye seçim işlemi eğitim ve doğrulama kümelerine bölünen ham derlem üzerinde yürütülmüş ve bunun sonucunda ham derlemden gelen çok iyeli veri kümelerinin yanında her iki iye seçim yöntemi için birer tek iyeli veri kümesi elde edilmiştir. Bu paralel veri kümelerinin cümleleri hizalanmış olduğundan, bir grup veri kümesi kullanılarak eğitilen modellerin başarımlarını diğer bir grup veri kümesi üzerinde doğrulanması da mümkün olmuştur. Tüm veri kümelerinin kendi içlerinde ve birbirleriyle çapraz doğrulanması sonucunda elde edilen etiketli ve etiketsiz bağlanma skorları Tablo 4.1 üzerinde verilmektedir.

Çizelge 4.1: İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi üzerinde doğrulanan modellerin çapraz doğrulama skorları.

		Doğrulama Kümesi		
		İTÜ-ODTÜ-SAB. D. (Tek İye: En Yakın)	İTÜ-ODTÜ-SAB. D. (Tek İye: En Uzak)	İTÜ-ODTÜ-SAB. D. (Gevşetilmiş Değerlendirme)
Eğitim Kümesi	İTÜ-ODTÜ-SAB. D. (Tek İye: En Yakın)	$BS_{E+}: \%75.1 \pm \%0.21$	$BS_{E+}: \%74.3 \pm \%0.19$	$BS_{E+}: \%75.7 \pm \%0.21$
		$BS_{E-}: \%84.1 \pm \%0.26$	$BS_{E-}: \%83.1 \pm \%0.25$	$BS_{E-}: \%84.4 \pm \%0.24$
	İTÜ-ODTÜ-SAB. D. (Tek İye: En Uzak)	$BS_{E+}: \%74.0 \pm \%0.23$	$BS_{E+}: \%74.4 \pm \%0.23$	$BS_{E+}: \%75.2 \pm \%0.25$
		$BS_{E-}: \%82.9 \pm \%0.23$	$BS_{E-}: \%83.2 \pm \%0.25$	$BS_{E-}: \%83.8 \pm \%0.25$

Modeller arası karşılaştırma sonucu elde edilen bağlanma skorları aynı model üzerinde eğitim ve doğrulama yapıldığı durumlardan genel olarak daha düşük çıkmıştır, ancak skorlar arasında ciddi bir farkın olmaması da ilgi çekici bir durumdur. Buna ek olarak, çok iyeli birimlerin oldukça düşük bir oranda olmasına rağmen (bkz. Tablo 3.1), gevşetilmiş değerlendirmenin anlamlı bir iyileşmeye yol açtığı görülmektedir. Buradan katı değerlendirmede yanlış negatiflerin kayda değer bir başarımla düşüşüne sebebiyet verebileceği anlaşılabılır, ancak diğer yandan, bu durum tek iye seçme yöntemlerinin fazlaca basit yapısından kaynaklanıyor da olabileceğinden daha akıllı bir iye seçme yöntemi ile katı değerlendirme ve gevşetilmiş değerlendirme arasındaki başarımların kapanması da mümkün olabilir. Her iki şekilde de, En Yakın İye kümeleriyle eğitilen modellerin hem En Yakın İye kümelerinde doğrulandığı durumda katı değerlendirmede, hem de çok iyeli ham kümelerde doğrulandığı durumda gevşetilmiş değerlendirmede en yüksek skorları verdiği görülmektedir ve buradan En Yakın İye seçiminin En Uzak İye varyantından daha başarılı olduğu sonucuna varılmıştır.

Yeni İOSD orijinal OSD'nin yeniden işaretlenmiş hali olduğundan dolayı bu derlemler paralel derlemler sayılabilir ve bu sayede iki derlem üzerinde elde edilen skorların karşılaştırılması anlamlı olmuştur. İki derlem arasında Tablo 3.1 üzerinde görülen bağıllık sayısı farklarının İOSD'deki çok iyeli işaretlemeden kaynaklandığı ve bunun dışında birim sayılarının eşit olduğu göz önünde tutulmalıdır. Bu duruma konu olan ekstra bağıllıklar başarımların ölçümleri için (gevşetilmiş değerlendirme dışında) test kümelerinden çıkarıldığından dolayı bu durum ayrıştırma sonuçlarının

karşılaştırılmasının önünde bir engel teşkil etmemektedir.

Orijinal ODTÜ-Sabancı Ağaç Yapılı Derlemi üzerinde çalışan öncü sistemdeki [2] en yüksek etiketli bağlanma skoru olan %65.9¹ karşısında yeni çerçevede elde edilen en yüksek katı değerlendirme skoru %75.1 ve gevşetilmiş değerlendirme skoru %75.7, etiketli doğrulama skorlarında neredeyse 10 yüzdelik puanlık bir artış olduğunu ortaya koymaktadır. Bu artış, çoğu geçmiş çalışmaya kıyasla çok önemli bir artıştır. Etiketsiz bağlanma skorlarında da benzer bir düzende artış olduğu, yeni skorların geçmiş öncü sistemin skoru olan %76.0'a kıyasla katı değerlendirmede %84.1'e ve gevşetilmiş değerlendirmede %84.4'e kadar yükseldiği görülmektedir.

4.3 Yeniden Eşleme Deneyleri

Orijinal ODTÜ-Sabancı Ağaç Yapılı Derlemi'nin bağıllık grameri üzerindeki değişikliklerin bir uzantısı olarak, Bölüm 2.2 altında da bahsedildiği üzere, benzer nitelikli bağıllık türleri tek etiket altında toplanmıştır. Bu tür gruplamalara ancak bağıllık türleri arasındaki farklılıkların sözdizimsel olmaktan ziyade biçimbilimsel olduğu durumlarda karar verilmiştir. Veri güdümlü bağıllık ayrıştırıcıları neredeyse değişmez olarak birimlerin biçimbilimsel özelliklerinden yararlandıklarından, sözdizimsel bağıllıklar arasındaki bu tür biçimbilimsel nitelikli ayrımlar gereksiz görünmektedir. Ayrıca bu şekildeki bağıllıkların ortak etiketler altında toplandığı bu tür çerçevelerde, uydu ve iye birimlerin biçimbilimsel özellikleri referans alınarak bir sonişleme rutini ile aradaki ortak bağıllık etiketinin ilgili ince etiketlere yeniden eşlenmesi çok kolay olmaktadır. Bu değişiklik anlamsal veya işlemsel temelleri olmayan tek değişiklik sınıfını teşkil ettiğinden görece keyfi kalsa da, temelde

¹Orijinal çalışmadaki bu skor, zamanının ölçüm standartları gereği noktalama işaretleri değerlendirilmeksizin hesaplanmıştır, ancak kendi ölçüm sistemimizle orijinal OSD üzerinde aynı skorları tekrar elde ettiğimizde noktalama işaretlerinin değerlendirilmeye katıldığı ve katılmadığı modellerin arasındaki başarımların istatistiksel olarak anlamsız olduğu tespit edilmiştir. Yeni İOSD üzerinde yapılan ölçümlerde noktalama işaretleri hesaba katılmadığında ise etiketli bağlanma skoru katı ve gevşetilmiş değerlendirme için sırasıyla %75.1 ve %75.7'den %69.8 ve %70.5'e düşmekte ve bu durum ayrıştırma başarımlarındaki iyileşmede yeni getirilen basit noktalama işaretleme yönteminin ancak kısmi bir katkısının olduğunu göstermektedir.

gözetilen minimalizm prensibi ile gerekçelendirilmektedir.

Bu tür gruplamaların iki büyük örneği **MODIFIER** ve **POSSESSOR** etiketleri için yapılmıştır. Bu bağılıklardan ilki ad soylu tümleçleri kapsayan **X.ADJUNCT** bağılıkları ve bunun dışında kalan her niteleyiciyi kapsayan **MODIFIER** bağılılığından derlenmiş, ikinci bağılık ise ad tamlamalarında tümleyen iniyelik halinde olduğu durumlarda kullanılan **POSSESSOR** bağılılığı ile tümleyen yalın halde olduğunda kullanılan **CLASSIFIER** bağılılığını kapsamaktadır. Bu gruplamaların bağımsız etkilerini görmek için İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi'nin her ortak etiket için alternatif kolları oluşturulmuş ve her birinde ilgili etiket Şekil A.1 ve A.2 üzerinde gösterildiği şekilde orijinal gramerin gerektirdiği etiketlere sonişleme rutinleri ile eşlenmiştir. Devamında bu eşlenmiş veriler (önceden gözlenen en yüksek başarılı yöntem olduğu üzere) En Yakın İye seçimiyle tek iyeli hale indirgenmiş ve bunlar üzerinde eğitilen modeller değerlendirilmiştir. Sonuçta elde edilen figürler Tablo 4.2 üzerinde verilmektedir.

Öncelikle **MODIFIER** ve **POSSESSOR** etiketleri üzerinde gerçekleştirilen eşleme deneyleri bu türler için alt türlerin kullanımının ayrıştırma başarımı üzerinde gözle görülür herhangi bir etkisinin olmadığını açık olarak göstermektedir. İlgili alternatif modellerin her ikisinde de etiketli (75.1 ± 0.1) ve etiketsiz (75.7 ± 0.1) bağlanma skorlarının standart model ile arasındaki farkın istatistiksel olarak anlamsız olduğu görülmüştür. Bu bulgular bu tür kategorilerde olabildiğince minimal bir işaretleme yöntemi kullanmanın güvenli olacağını göstermekte ve önceki alt türleri yeni işaretleme çerçevesinde tek bir ortak etiket altında gruplama inisiyatifi desteklemektedir. Bu yolla aynı zamanda başarımlar veya ifade gücünde bir düşüşe yol açmadan işaretleme süreci daha açık ve sezgisel hale getirilmektedir.

Yapılan diğer bir eşleme ise çok sözcüklü deyimsele ifadeleri tanımlamakta kullanılan **MWE** bağılılığı ile ilgilidir. Bu bağılık anlamsal yapısından dolayı ayrıştırma sürecinde doğru olarak atanması en zor olan bağılık türleri arasındadır. Çok sözcüklü ifadelerin sözdizimsel katmanda işaretlenmesi ifade gücü ve ayrıştırma

kolaylığı arasında daha ödünleşimli bir durumu teşkil etmektedir. Bu sebepten dolayı **MWE** türü de Şekil A.3 üzerinde gösterildiği gibi benzer şekilde bir soniştirme rutini ile safi sözdizimsel türlerle açılmış ve böylece derlemin üçüncü bir alternatif kolu oluşturulmuştur. Bu eşleme **MWE** türlerini tamamen ortadan kaldırarak gramerin çok sözcüklü ifadeleri göstermesi için bir yol bırakmamış ve ifade gücünde net bir azalmaya yol açmıştır. Bunun yanında, eşlenmiş veri kümesinde eğitilen modellerle yapılan çapraz doğrulama, Tablo 4.2 üzerinde gösterildiği gibi etiketli bağlanma skorlarında %2.7'ye yakın bir artış göstermiş olduğundan, bu değerlendirmenin yapılmış olması önemlidir. Bu eşlemenin ayrıştırıcının genel başarımı üzerindeki belirgin etkisine rağmen işaretleme çerçevesinden **MWE** türü kaldırılmamış, çok sözcüklü ifadelerin ifade edilebilmesi ayrıştırma başarımındaki yükselişe tercih edilmiştir.

Çizelge 4.2: İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi'nde **MODIFIER**, **POSSESSOR** ve **MWE** türlerinin diğer türlere eşlendiği üç alternatif kol ile ilişkili modeller üzerindeki çapraz doğrulama skorları.

		Doğrulama Kümesi	
		(Tek İye: En Yakın)	(Gevşetilmiş Değerlendirme)
Eğitim Kümesi	İOSD	$BS_{E+}: \%75.1 \pm \%0.21$	$BS_{E+}: \%75.7 \pm \%0.21$
	(Tek İye: En Yakın)	$BS_{E-}: \%84.1 \pm \%0.26$	$BS_{E-}: \%84.4 \pm \%0.24$
	İOSD - MODIFIER Eşlemesi	$BS_{E+}: \%75.0 \pm \%0.22$	$BS_{E+}: \%75.6 \pm \%0.21$
	(Tek İye: En Yakın)	$BS_{E-}: \%84.0 \pm \%0.25$	$BS_{E-}: \%84.4 \pm \%0.23$
	İOSD - POSSESSOR Eşlemesi	$BS_{E+}: \%75.2 \pm \%0.20$	$BS_{E+}: \%75.8 \pm \%0.20$
	(Tek İye: En Yakın)	$BS_{E-}: \%84.2 \pm \%0.25$	$BS_{E-}: \%84.4 \pm \%0.23$
	İOSD - MWE Eşlemesi	$BS_{E+}: \%77.8 \pm \%0.17$	$BS_{E+}: \%78.4 \pm \%0.16$
	(Tek İye: En Yakın)	$BS_{E-}: \%84.2 \pm \%0.25$	$BS_{E-}: \%84.5 \pm \%0.23$

4.4 İTÜ Ağaç Yapılı Web Derlemi

İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi'nde olduğu gibi İTÜ Ağaç Yapılı Web Derlemi üzerindeki çapraz doğrulama skorları da Tablo 4.3 üzerinde verilmiştir. Tek iyeli modellerin birbirleriyle karşılaştırılması ve gevşetilmiş değerlendirme sonuçları İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi'ne benzer bir skor dağılımı göstermektedir. İTÜ Ağaç Yapılı Web Derlemi'nde de En Yakın İye seçim

yöntemi katı değerlendirmede %79.7 ve gevşetilmiş değerlendirmede %80.1'e varan etiketli bağlanma skorlarıyla daha iyi başarımlar göstermiştir. Elde edilen bu skorları İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi'yle karşılaştırmak çok anlamlı olmayacak olsa da, İTÜ Ağaç Yapılı Web Derlemi'nde ölçülen temel başarımların İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi'ndekileri açık arayla geçtiği görülmektedir. Bu sonuçların ışığında bu yeni derlemin kendi içinde oldukça tutarlı bir yapıda olduğu sonucu çıkarılabilir.

Çizelge 4.3: İTÜ Ağaç Yapılı Web Derlemi üzerinde doğrulanan modellerin çapraz doğrulama skorları.

		Doğrulama Kümesi		
		İTÜ A. Y. WEB D. (Tek İye: En Yakın)	İTÜ A. Y. WEB D. (Tek İye: En Uzak)	İTÜ A. Y. WEB D. (Gevşetilmiş Değerlendirme)
Eğitim Kümesi	İTÜ-ODTÜ-SAB. D. (Tek İye: En Yakın)	BS_{E+} : %68.2 BS_{E-} : %79.6	BS_{E+} : %67.7 BS_{E-} : %79.0	BS_{E+} : %68.7 BS_{E-} : %80.1
	İTÜ-ODTÜ-SAB. D. (Tek İye: En Uzak)	BS_{E+} : %66.9 BS_{E-} : %78.4	BS_{E+} : %67.0 BS_{E-} : %78.4	BS_{E+} : %67.7 BS_{E-} : %79.2
	İTÜ A. Y. WEB D. (Tek İye: En Yakın)	BS_{E+} : %79.7 $\pm$ %0.22 BS_{E-} : %87.5 $\pm$ %0.14	BS_{E+} : %78.8 $\pm$ %0.23 BS_{E-} : %86.5 $\pm$ %0.14	BS_{E+} : %80.1 $\pm$ %0.23 BS_{E-} : %87.9 $\pm$ %0.15
	İTÜ A. Y. WEB D. (Tek İye: En Uzak)	BS_{E+} : %78.8 $\pm$ %0.23 BS_{E-} : %86.7 $\pm$ %0.18	BS_{E+} : %79.7 $\pm$ %0.22 BS_{E-} : %86.6 $\pm$ %0.15	BS_{E+} : %79.5 $\pm$ %0.23 BS_{E-} : %87.5 $\pm$ %0.17

İki yeni derlem arasındaki alan farklılığı hakkında bir fikir sahibi olmak adına İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi üzerinde eğitilen modellerle İTÜ Ağaç Yapılı Web Derlemi'nin ayrıştırılması da denenmiştir. Beklendiği üzere, her iki derlemin de aynı işaretleme çerçevesine göre etiketlenmiş olmasına rağmen, İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi'nin kurallı cümleleri üzerinde eğitilen modellerin İTÜ Ağaç Yapılı Web Derlemi'nin yoğun ölçüde kural dışı olan cümlelerini ayrıştırmakta, her ne kadar bu kural dışı cümleler elle dikkatli bir şekilde normalize edilmiş olsalar da, daha düşük başarımlar gösterdiği görülmüştür. İTÜ Ağaç Yapılı Web Derlemi'nin İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi üzerinde eğitilen modeller kullanılarak ayrıştırılması sonucunda, her iki derlemin de kendi içinde çapraz doğrulanmasıyla elde edilen skarlardan ciddi derecede daha düşük skorlar

alınmıştır. Bu durumlar, bu iki derlemin oldukça farklı alanlardaki cümleleri kapsadığı ve birbirlerinin muadili olamayacağına güçlü bir kanıt teşkil etmektedir.

5. SONUÇ

Bu çalışmada yaygın olarak tanınan ODTÜ-Sabancı Türkçe Ağaç Yapılı Derlemi'nin kullandığı bağıllık gramerinin üzerine kurulan yeni sözdizimsel işaretleme çerçevesi anlatılmıştır. Yeni işaretleme çerçevesi, temel alınan gramerin 24 bağıllık etiketine karşılık yalnızca 16 etiket içermekte, bununla birlikte daha açık ve anlaşılır olan, tutarlılığı daha yüksek ve birbirleriyle daha az çakışan bağıllık türleri kullanmaktadır. Bu işaretleme çerçevesi altında işaretleme zorluğunun ciddi derecede azaldığı ve bu kolaylığa karşılık yeni gramerin eskisine kıyasla ifade gücünün azalmadığı görülmektedir.

Çalışmanın devamında yeni işaretleme çerçevesinin birer uygulaması olarak, ODTÜ-Sabancı Ağaç Yapılı Derlemi'nin yeniden oluşturulmuş bir sürümü olarak İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi, ardından web kullanıcılarının oluşturduğu kural dışı yazılardan derlenen İTÜ Ağaç Yapılı Web Derlemi tanıtılmıştır. İki derlemde de ayrıca derin bağıllıklar işaretlenerek gelecekte bu kaynakları kullanacak anlamsal rol çıkarımı çalışmalarına imkan tanınmıştır. Bu işaretleme ayrıca öğelerin tüm işaretli iyelerinin değerlendirildiği, yanlış negatif çıktılarından daha az etkilenen, daha adil bir değerlendirme metriğinin işletilmesine olanak sağlamıştır.

Çalışmanın devamında işaretleme çerçevesinin anlatılan teorik avantajlarını somut olarak göstermeyi amaçlayan çok sayıda deneysel sonuç verilmiş, özellikle eski kaynaklarla karşılaştırılmak üzere, tanıtılan yeni kaynaklar üzerindeki sözdizimsel ayrıştırma başarımları sunulmuştur. Verilen en iyi ayrıştırma modeli için İTÜ-ODTÜ-Sabancı Ağaç Yapılı Derlemi'nde etiketli bağlanma skorları katı değerlendirmede %75.1 ve gevşetilmiş değerlendirmede %75.7 olarak hesaplanmış, bu skorların orijinal ODTÜ-Sabancı Ağaç Yapılı Derlemi üzerinde %65.9 olarak bildirilen en yüksek etiketli bağlama skorunu ciddi derecede geçtiği görülmüştür.

İTÜ Ağaç Yapılı Web Derlemi'nin karşılaştırılacağı bir temel derlem olmamasına rağmen, çapraz doğrulama deneylerimiz sonucunda katı değerlendirme için %78.7 ve gevşetilmiş değerlendirme için %80.1 olarak bulunan etiketli bağlanma skorları çok olumlu görünmektedir. Sonuçlardan aynı zamanda çalışmanın kural dışı web verisinin ayrıştırılması konusuna önemli bir katkısının olduğunu göstermektedir. Web alanı için yapılmış olan adaptasyon çalışması ile bağlanma skorları her bir değerlendirme metriği için neredeyse 12 yüzdelik puanlık bir artış göstermiştir. Yeni işaretleme çerçevesi ve alan adaptasyonu üzerine yapılan çalışmaların ortak etkisinin bundan daha da büyük olduğu öngörülmekle beraber, İTÜ Ağaç Yapılı Web Derlemi'nin geçmiş bağlılık şemasına göre işaretlenmiş bir sürümü bulunmadığından bunun net bir ölçümü yapılamamaktadır.

Çalışmanın çıktıları düşünüldüğünde, denetimli öğrenme sistemlerinin eğitim kümeleri üzerinde doğrudan çalışarak kayda değer bir ilerleme kaydedildiğini söylemek mümkündür. Üzerinde çalışılan verinin kalitesini artırmak açık uçlu bir uğraş olsa da, bildirilen sonuçlar, başta teorik olarak öngörüldüğü gibi, bu uğraşın ayrıştırma başarımları üzerinde potansiyel olarak ciddi etkilerinin olabileceğini doğrulamıştır. Bu çalışmanın böylece Türkçe veri kümeleri üzerine yapılacak benzer çalışmalar için motivasyon sağlaması umulmaktadır.

KAYNAKLAR

- [1] **Sulubacak, U. ve Eryiğit, G.**, 2013. Representation of Morphosyntactic Units and Coordination Structures in the Turkish Dependency Treebank, Proceedings of the 4th Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL), Association for Computational Linguistics, Seattle, Washington, USA, s.129–134, <http://www.aclweb.org/anthology/W13-4915>.
- [2] **Eryiğit, G., Ilbay, T. ve Can, O.A.**, 2011. Multiword Expressions in Statistical Dependency Parsing, Proceedings of the 2nd Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL), Association for Computational Linguistics, Dublin, Ireland.
- [3] **Eryiğit, G., Nivre, J. ve Oflazer, K.**, 2008. Dependency Parsing of Turkish, *Computational Linguistics*, **34(3)**, 357–389.
- [4] **Çetinoğlu, O. ve Kuhn, J.**, 2013. Towards Joint Morphological Analysis and Dependency Parsing of Turkish, Proceedings of the 2nd International Conference on Dependency Linguistics (DepLing), Charles University in Prague, Matfyzpress, Prague, Czech Republic, Prague, Czech Republic, s.23–32, <http://www.aclweb.org/anthology/W13-3704>.
- [5] **Özlem Çetinoğlu**, 2014. Turkish Treebank as a Gold Standard for Morphological Disambiguation and Its Influence on Parsing, Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), European Language Resources Association (ELRA), Reykjavík, Iceland.
- [6] **Durgar El-Kahlout, I., Akın, A.A. ve Yılmaz, E.**, 2014. Initial Explorations in Two-phase Turkish Dependency Parsing by Incorporating Constituents, Proceedings of the 1st Joint Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL) and Syntactic Analysis of Non-Canonical Languages (SANCL), Dublin City University, Dublin, Ireland, s.82–89, <http://www.aclweb.org/anthology/W14-6108>.
- [7] **Oflazer, K., Say, B., Hakkani-Tür, D.Z. ve Tür, G.**, 2003. Building a Turkish Treebank, *Treebanks*, Springer, s.261–277.
- [8] **Eryiğit, G.**, 2006. Dependency Parsing of Turkish, *Doktora Tezi*, Istanbul Technical University.
- [9] **Çakıcı, R.**, 2008. Wide-Coverage Parsing for Turkish, *Doktora Tezi*, The University of Edinburgh.

- [10] **Eryiğit, G.**, 2007. ITU Validation Set for METU-Sabancı Turkish Treebank.
- [11] **Buchholz, S. ve Marsi, E.**, 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing, Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, s.149–164.
- [12] **Nilsson, J., Riedel, S. ve Yüret, D.**, 2007. The CoNLL 2007 Shared Task on Dependency Parsing, Proceedings of the CoNLL Shared Task Session of the Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, s.915–932.
- [13] **Seddah, D., Sagot, B., Candito, M., Mouilleron, V., Combet, V. ve diğerleri**, 2012. The French Social Media Bank: a Treebank of Noisy User Generated Content, Proceedings of the 24th International Conference on Computational Linguistics (COLING).
- [14] **Bies, A., Mott, J., Warner, C. ve Kulick, S.**, 2012. English Web Treebank, *Linguistic Data Consortium*.
- [15] **Percival, W.K.**, 1990. Reflections on the History of Dependency Notions in Linguistics, *Historiographia Linguistica*, **17(1-2)**, 29–47.
- [16] **Tesnière, L.**, 1959. *Éléments de Syntaxe Structurale*, Éditions Klincksieck.
- [17] **Çetinoğlu, O.**, 2009. A Large Scale LFG Grammar for Turkish, *Doktora Tezi*, Sabancı University.
- [18] **Kübler, S., McDonald, R. ve Nivre, J.**, 2009. Dependency Parsing, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers.
- [19] **Hakkani-Tür, D.Z., Oflazer, K. ve Tür, G.**, 2002. Statistical Morphological Disambiguation for Agglutinative Languages, *Computers and the Humanities*, **36(4)**, 381–410.
- [20] **Oflazer, K.**, 2003. Dependency Parsing with an Extended Finite-State Approach, *Computational Linguistics*, **29(4)**, 515–544.
- [21] **Eryiğit, G. ve Oflazer, K.**, 2006. Statistical Dependency Parsing of Turkish, Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Trento, s.89–96.
- [22] **Böhmová, A., Hajič, J., Hajičová, E. ve Hladká, B.**, 2003. The Prague Dependency Treebank, *Treebanks*, Springer, s.103–127.
- [23] **Nivre, J., Nilsson, J. ve Hall, J.**, 2006. Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation, Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), s.1392–1395.

- [24] **Csendes, D., Csirik, J., Gyimóthy, T. ve Kocsor, A.**, 2005. The Szeged Treebank, Text, Speech and Dialogue, Springer, s.123–131.
- [25] **Van der Beek, L., Bouma, G., Malouf, R. ve Van Noord, G.**, 2002. The Alpino Dependency Treebank, *Language and Computers*, **45**(1), 8–22.
- [26] **De Marneffe, M.C. ve Manning, C.D.**, 2008. The Stanford Typed Dependencies Representation, Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation (COLING), Association for Computational Linguistics, s.1–8.
- [27] **De Marneffe, M.C., Connor, M., Silveira, N., Bowman, S.R., Dozat, T. ve Manning, C.D.**, 2013. More Constructions, More Genres: Extending Stanford Dependencies, Proceedings of the 2nd International Conference on Dependency Linguistics (DepLing), Charles University in Prague, Matfyzpress, Prague, Czech Republic, Prague, Czech Republic, s.187–196, <http://www.aclweb.org/anthology/W13-3721>.
- [28] **McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N. ve Lee, J.**, 2013. Universal Dependency Annotation for Multilingual Parsing, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, Sofia, Bulgaria, s.92–97, <http://www.aclweb.org/anthology/P13-2017>.
- [29] **De Marneffe, M.C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J. ve Manning, C.D.**, 2014. Universal Stanford Dependencies: a Cross-Linguistic Typology, Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), European Language Resources Association (ELRA), Reykjavík, Iceland.
- [30] **Ambati, B.R., Reddy, S. ve Kilgarrieff, A.**, 2012. Word Sketches for Turkish, Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), s.2945–2950.
- [31] **Pamay, T., Sulubacak, U., Torunoğlu-Selamet, D. ve Eryiğit, G.**, 2015. The Annotation Process of the ITU Web Treebank, Proceedings of the 9th Linguistic Annotation Workshop (LAW), Denver, CO, USA.
- [32] **Eryiğit, G.**, 2007. ITU Treebank Annotation Tool, Proceedings of the ACL Workshop on Linguistic Annotation (LAW), Prague.
- [33] **Eryiğit, G.**, 2015. Morphological Processing of Turkish: From Theory to Practice, (Under review).
- [34] **Eryiğit, G. ve Torunoğlu-Selamet, D.**, 2015. Social Media Text Normalization for Turkish, (Under review).

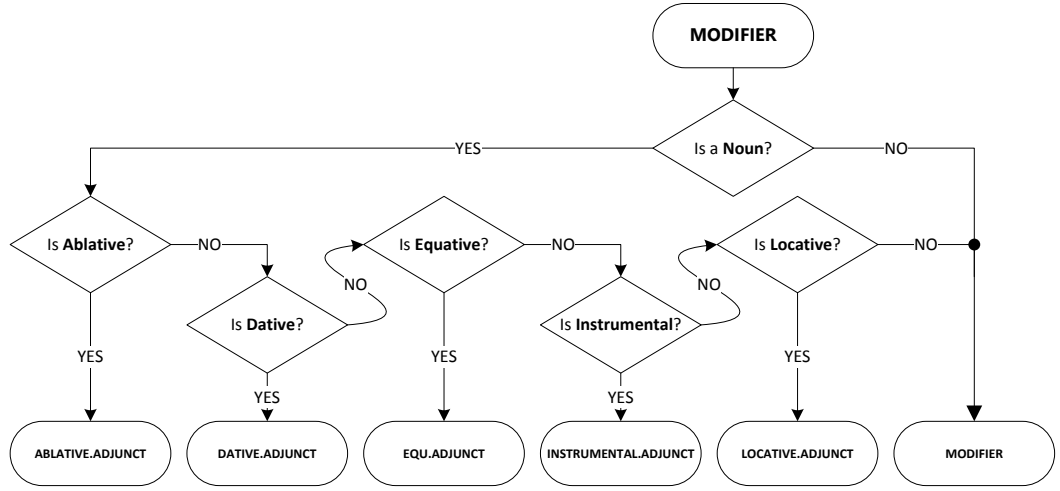
- [35] **Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryiğit, G., Kübler, S., Marinov, S. ve Marsi, E.**, 2007. MaltParser: A Language-Independent System for Data-Driven Dependency Parsing, *Natural Language Engineering*, **13**, 95–135.

EKLER

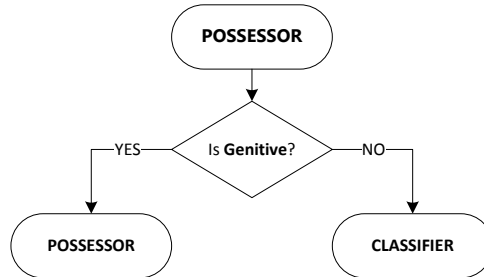
EK A: Baęlılık Geri Dönüşüm Akış Diyagramları

EK B: Baęlılık İşaretleme Kitapçığı (*İngilizce*)

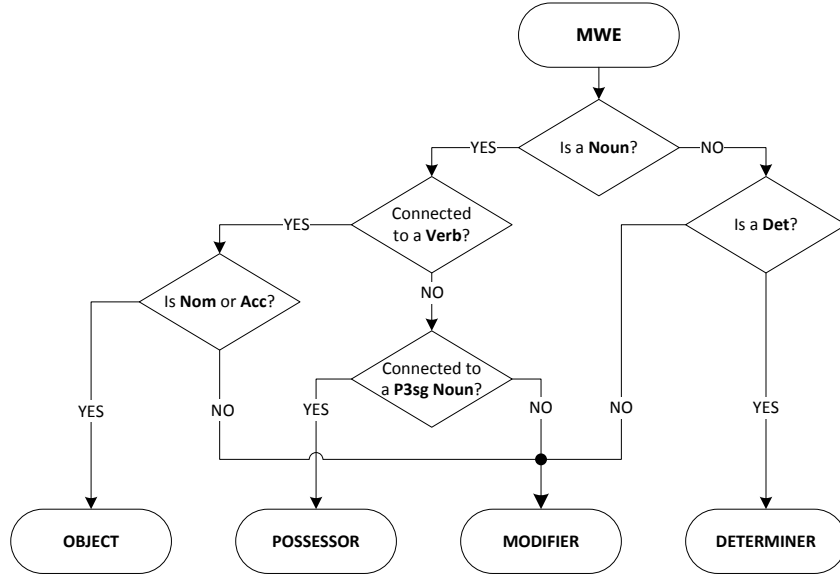
EK A: Bağlılık Geri Dönüşüm Akış Diyagramları



Şekil A.1: Yeni **MODIFIER** bağlılığının orijinal ODTÜ-Sabancı Ağaç Yapılı Derlemi işaretleme çerçevesindeki muadillerine eşlenme akış diyagramı.



Şekil A.2: Yeni **POSSESSOR** bağlılığının orijinal ODTÜ-Sabancı Ağaç Yapılı Derlemi işaretleme çerçevesindeki eski **POSSESSOR** ve **CLASSIFIER** türlerine eşlenme akış diyagramı.



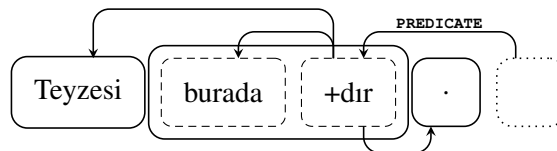
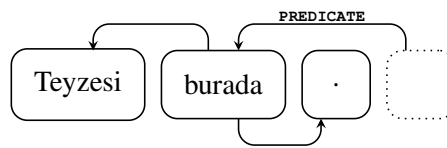
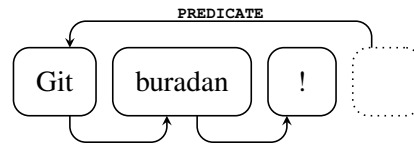
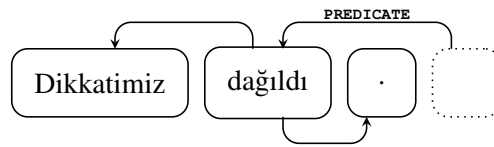
Şekil A.3: Yeni MWE bağlılığının ilgili sözdizimsel bağlılık türlerine eşlenme akış diyagramı.

EK B: ITU Treebank Annotation Guide (v2.7)

2.1 Elementary Dependencies

2.1.1 PREDICATE

Usage: Predicates are connected to the root node of the sentence with the **PREDICATE** relation. This dependency type is used for both inflected verbs and copulative nominals.

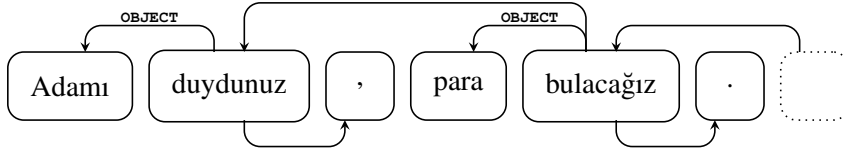


2.1.2 OBJECT

Usage: Nominative and accusative dependents that function as the direct object of a verb are connected to the verbal head with the **OBJECT** relation. If the object and the

verbal head combine to denote an idiomatic meaning, the **MWE** relation should be used instead of **OBJECT**.

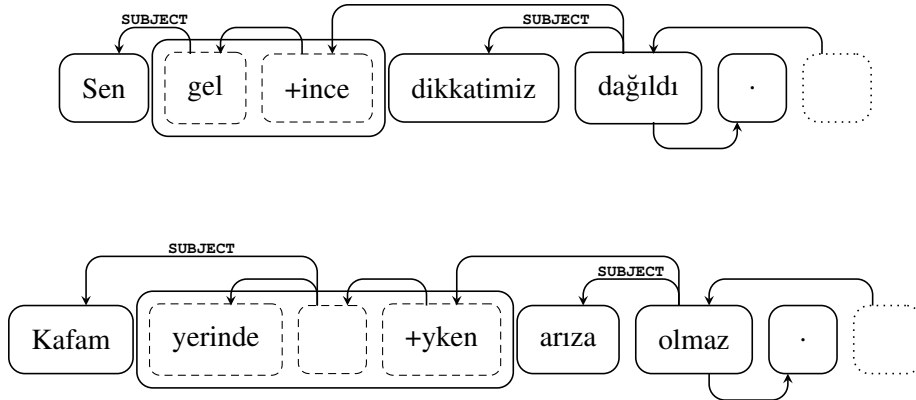
Might be mistaken for: MWE



2.1.3 SUBJECT

Usage: Nominative dependents that function as the direct subject of a predicate are connected to the predicative head with the **SUBJECT** relation. Subjects of subordinate clause predicates derived to non-predicative stems via affixation also call for this dependency type. If the subject and the predicative head combine to denote an idiomatic meaning, the **MWE** relation should be used instead of **SUBJECT**.

Might be mistaken for: MWE, OBJECT

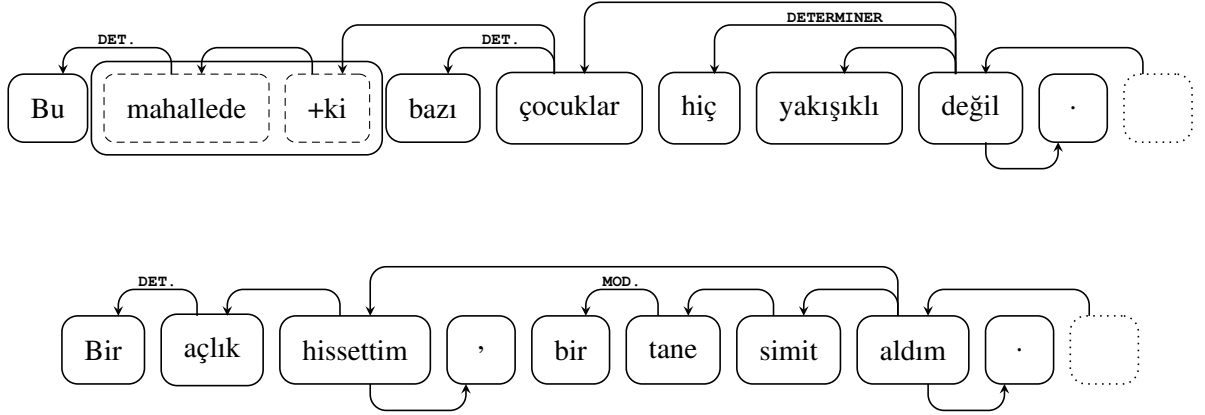


2.2 Modifiers

2.2.1 DETERMINER

Usage: The demonstrative adjectives *bu*, *şu*, *o* and indefinite quantifiers such as *bazı*, *bir*, *biraz*, *birçok*, *birkaç*, *birtakım*, *çok*, *en*, *her*, *hiç* and *hiçbir* are connected to the heads they modify with the **DETERMINER** relation.

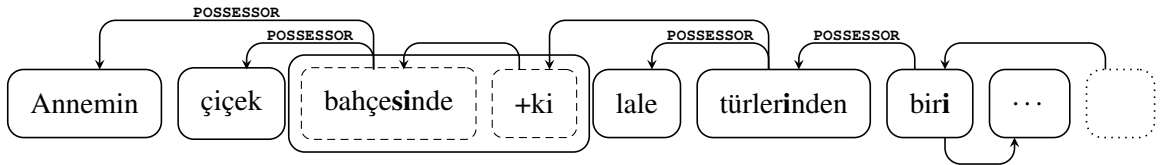
Might be mistaken for: MODIFIER



2.2.2 POSSESSOR

Usage: The **POSSESSOR** relation is used to mark phrases where a nominal dependent is in a possession or classification relation with a nominal head. In other words, noun phrases in the form of “A B’si”, “A’nın B’si” or “A’dan biri”, where the head has third person singular possession, are marked by the **POSSESSOR** relation.

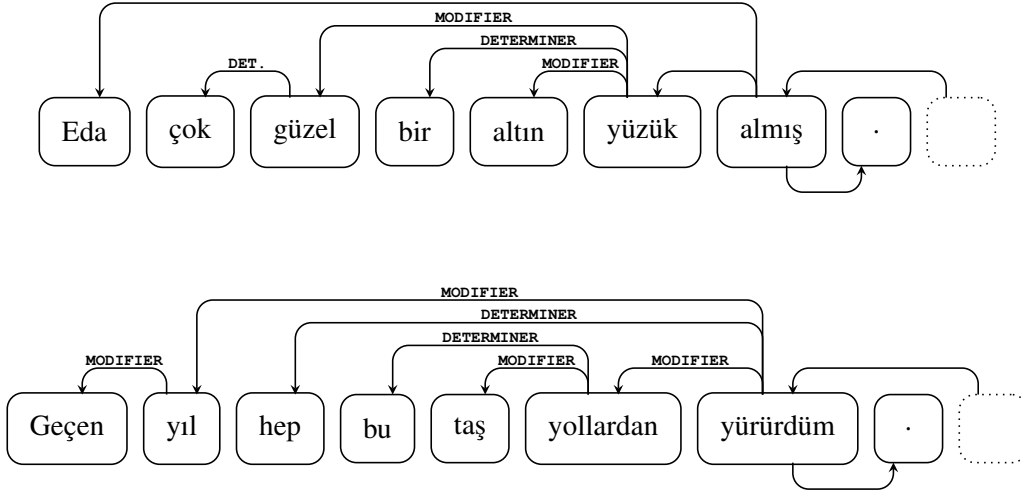
Might be mistaken for: MODIFIER, MWE



2.2.3 MODIFIER

Usage: All indirect objects and modifiers comprising adjectives and adverbs not examined under a specific category are connected to their heads with the relation **MODIFIER**. Although most double-noun phrases fall under the **POSSESSOR** dependency type, certain cases, primarily when a nominal dependent modifies a nominal head without possession, may still call for the **MODIFIER** relation. Many examples regarding **MODIFIER** usages are given as counter-examples under other dependency types.

Might be mistaken for: DETERMINER, MWE, POSSESSOR

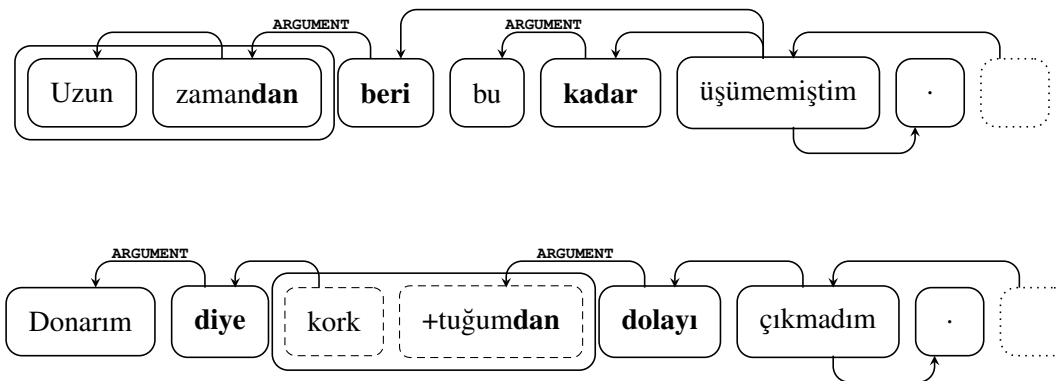


2.3 Adpositionals

2.3.1 ARGUMENT

Usage 1: Dependents of adpositional phrases, where the head is a postposition such as *başka, bağlı, beri, dair, doğru, değin, dek, denli, diye, dolayı, evvel, gayrı, gibi, göre, için, ile, ilişkin, kadar, karşı, karşın, olarak, önce, öte, ötürü, rağmen, sonra* or *üzere*, are connected to the head with the **ARGUMENT** relation. Although such dependents are sometimes called adpositional objects, the **OBJECT** relation is never used in such phrases.

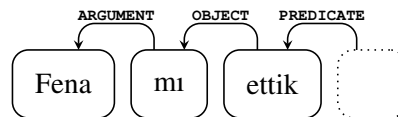
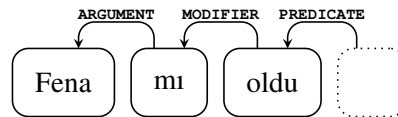
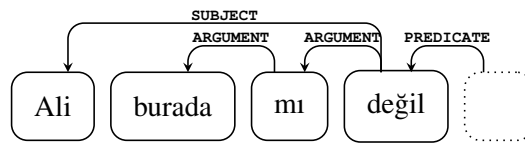
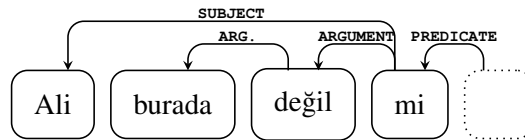
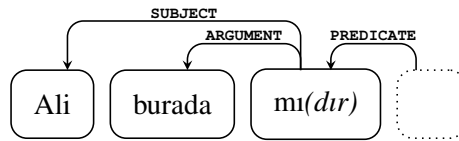
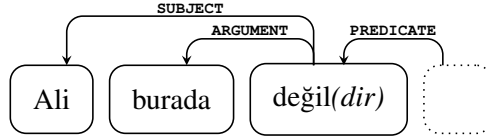
Might be mistaken for: MODIFIER, OBJECT



Usage 2: When the negative particle *değil*, the interrogative particle *mi*, or any copulative form of either of these particles modifies a constituent, the constituent is connected to the particle with the **ARGUMENT** relation. In such phrases, the particle becomes the head of the phrase and undertakes the syntactic function of the modified constituent. These particles typically modify predicates, so they themselves are often

connected to the root node with the **PREDICATE** relation.

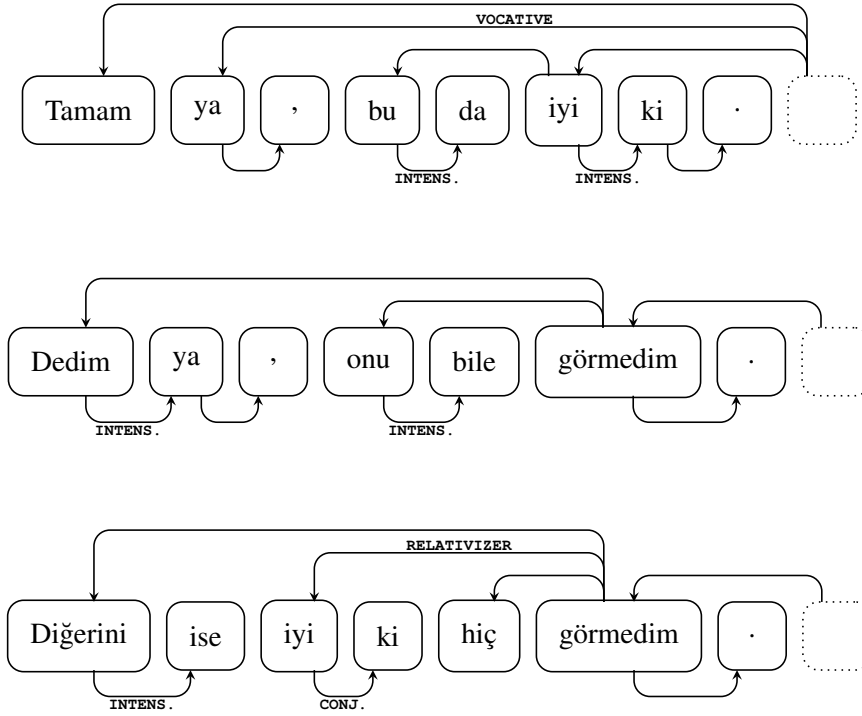
Might be mistaken for: **MODIFIER, OBJECT, PREDICATE, SUBJECT**



2.3.2 INTENSIFIER

Usage: The particles *de* and *daha* indicating addition, as well as particles such as *bile*, *dahi*, *de*, *ise*, *ki* and *ya* indicating emphasis, are connected to the words they modify with the **INTENSIFIER** relation.

Might be mistaken for: CONJUNCTION, RELATIVIZER, VOCATIVE



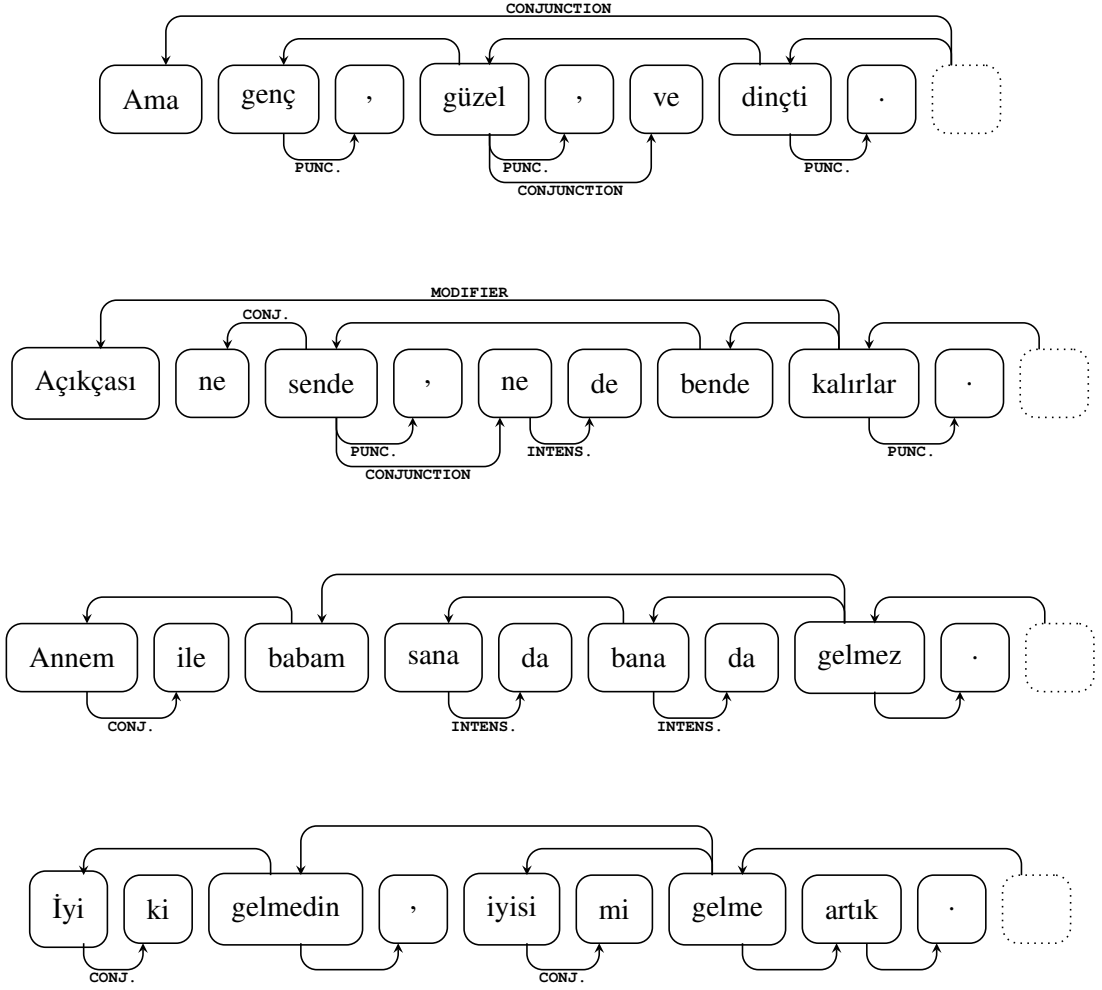
2.4 Coordinations

2.4.1 CONJUNCTION

Usage: This dependency relation marks conjunctions facilitating coordination structures and relative clauses. Single conjunctions occurring between the constituents in coordination, such as *ama*, *ancak*, *çünkü*, *demek ki*, *fakat*, *halbuki*, *hatta*, *hele*, *hem*, *ile*, *lakin*, *madem(ki)*, *ne var ki*, *ne yazık ki*, *nitekim*, *oysa(ki)*, *öyle ki*, *öyleyse*, *üstelik*, *ve*, *veya*, *veyahut*, *ya da*, *yahut*, *yalnız*, *yeter ki*, *yoksa* and *zira*, are connected to the constituent occurring before it, with the **CONJUNCTION** relation. Both parts of correlative conjunctions such as *gerek...gerek(se)*, *ha...ha*, *hem...hem (de)*, *ister...ister(se)*, *kah...kah*, *ne...ne (de)* and *ya...ya (da)*, are connected to the constituent they enclose, likewise with the **CONJUNCTION** relation. Conjunctive punctuation such as commas and semicolons do not fall under this dependency type, and instead call for the **PUNCTUATION** relation.

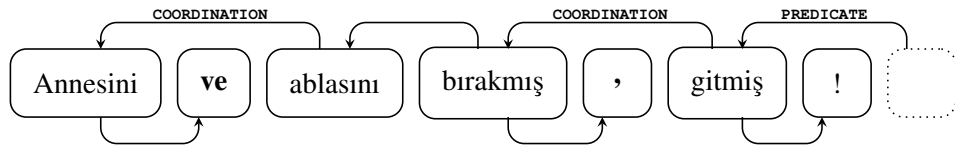
When sentences start with a conjunction and appear to constitute the latter clause of a bipartite sentence, the initial conjunction is connected to the root node with the **CONJUNCTION** relation. Disjunctions such as *açıkçası*, *kısacası*, *ne yazık ki* and *zaten* that begin a sentence do not fall under this scope, and should rather be connected to the predicate with the **MODIFIER** relation.

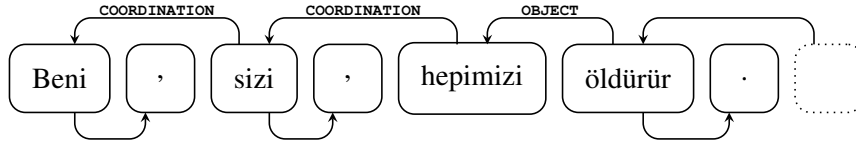
Might be mistaken for: INTENSIFIER, RELATIVIZER, MODIFIER



2.4.2 COORDINATION

Usage: Constituents with similar syntactic functions logically associated with the same head are considered to form a coordination structure. In coordination structures, the rightmost constituent is the head of the structure and is connected to an external head representing the whole structure, while every other constituent is connected to the next constituent after it with the **COORDINATION** relation. It should be noted that only constituents with **PREDICATE**, **SUBJECT**, **OBJECT**, **POSSESSOR** and **MODIFIER** relations with the external head are allowed to form coordination structures.

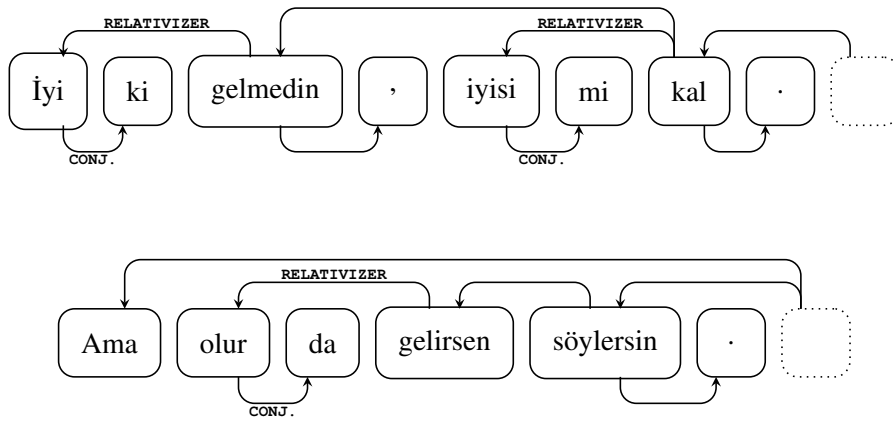




2.4.3 RELATIVIZER

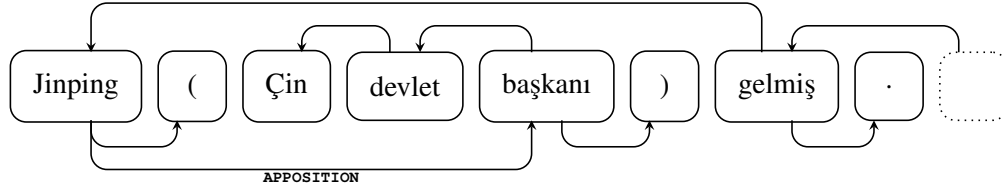
Usage: This dependency type is used for relative clauses, which are treated as coordination structures between subordinate clauses where one clause describes the other, often by the help of conjunctions such as *ki*, *de* and *mi*. In relativizer structures, the constituents in coordination are considered to be the heads of the clauses, and they are connected in a chain as with regular coordination structures, but with the **RELATIVIZER** relation rather than **COORDINATION**.

Might be mistaken for: **COORDINATION**, **INTENSIFIER**



2.4.4 APPPOSITION

Usage: When a detached clause is placed next to another clause in order to identify it, they are considered to be in apposition. Both clauses are usually noun phrases, and the latter clause, called the appositive, is often given in parentheses. In apposition structures, the head of the appositive is connected to the head of the original clause with the **APPPOSITION** relation. Although apposition structures are syntactically similar to coordination structures, it should be noted that the head of an apposition structure is on the left rather than the right.

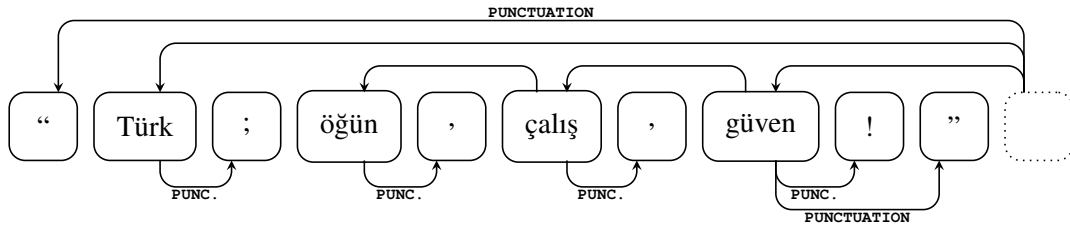


2.5 Other Dependencies

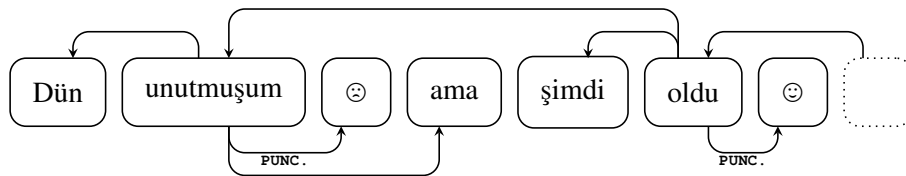
2.5.1 PUNCTUATION

Usage 1: All punctuation, including conjunctive punctuation such as a comma or a semicolon, should be at all times connected to the last non-punctuation constituent occurring before it, with the **PUNCTUATION** relation. Punctuation that begins a sentence should be connected to the sentence's root node instead.

Might be mistaken for: **CONJUNCTION**

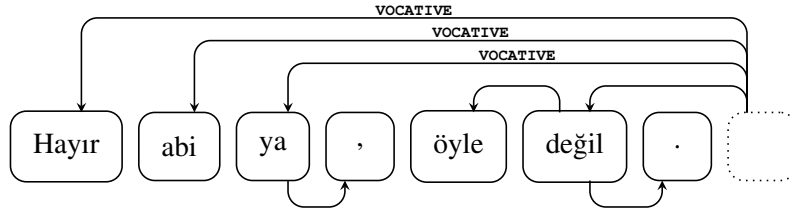


Usage 2: Emoticons such as smileys form dependencies in the same way as regular punctuation.



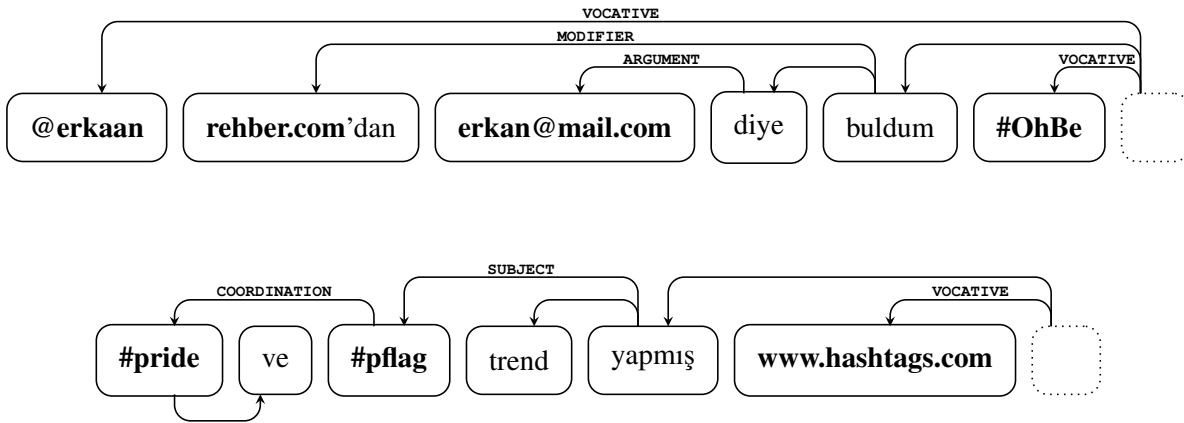
2.5.2 VOCATIVE

Usage 1: This dependency type is used in interjections and exclamations such as *ah*, *of*, *ya*, titular addresses such as *abi*, *canım* and *efendim*, responses such as *evet*, *hayır* and *tamam*, as well as calls or addresses by name, that otherwise do not modify any other constituent within the sentence. Such dependents are connected to the sentence's root node with the **VOCATIVE** relation.



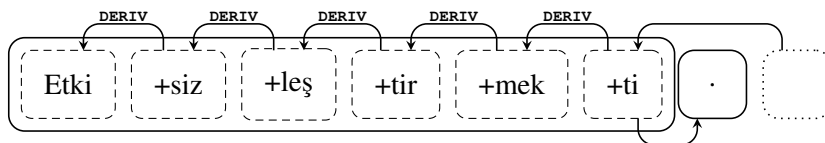
Usage 2: Web-specific constituents such as hashtags, mentions, URLs and e-mail addresses are morphologically regarded as nouns and usually take on regular functions within the sentence, where they form dependencies such as **OBJECT** and **SUBJECT**. However, when they are detached from the other constituents of the sentence, as in a mention in a tweet, they are connected to the sentence's root node with the **VOCATIVE** relation.

Might be mistaken for: A variety of dependencies.



2.5.3 DERIV

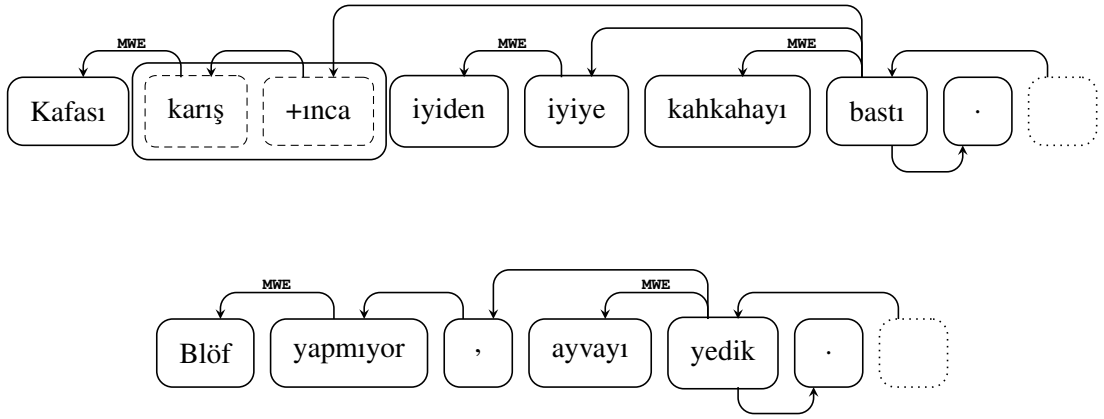
Usage: In words derived via affixation, 'derivational boundaries' are assumed to precede derivational affixes that caused a part-of-speech shift in the stem. These boundaries split the word into so-called 'inflectional groups' (IGs), which are widely accepted to be the proper morphosyntactic parsing units. In constituents comprising multiple IGs, every IG is connected the next with the **DERIV** relation, except for the last IG, which is considered to be the head of the word.



2.5.4 MWE

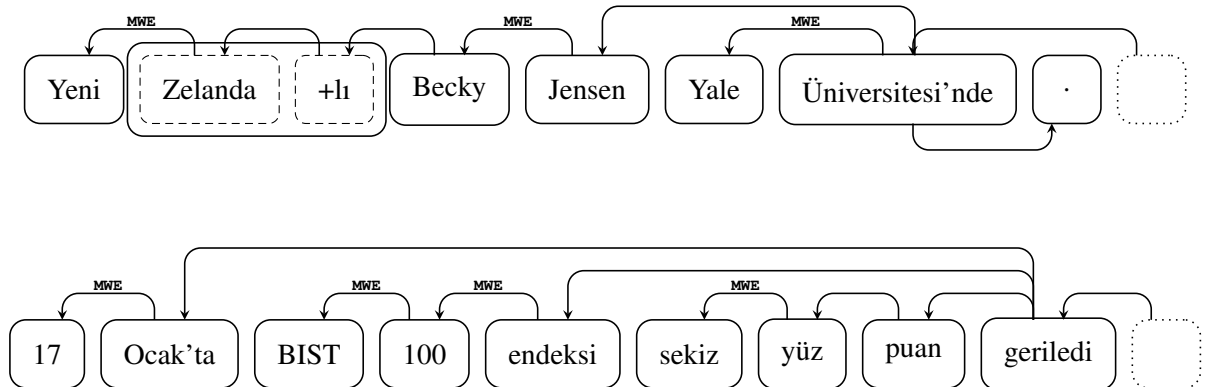
Usage 1: This is a semantic dependency type indicating multi-word expressions. For any pair of constituents that might syntactically form another type of dependency, if they combine to denote an idiomatic meaning, the **MWE** relation overrides the syntactic dependency relation that would otherwise be applied. Each constituent of a multi-word expression should be connected to the next with the **MWE** relation, except for the last one, which is considered to be the head of the phrase.

Might be mistaken for: A variety of dependencies.



Usage 2: Multi-word phrases that would be classified as *named entities*, comprising proper names and titles for persons, organizations, locations and similar entities, as well as the various numerical expressions for numbers, percentages, dates, currencies and measures, also make use of the **MWE** relation.

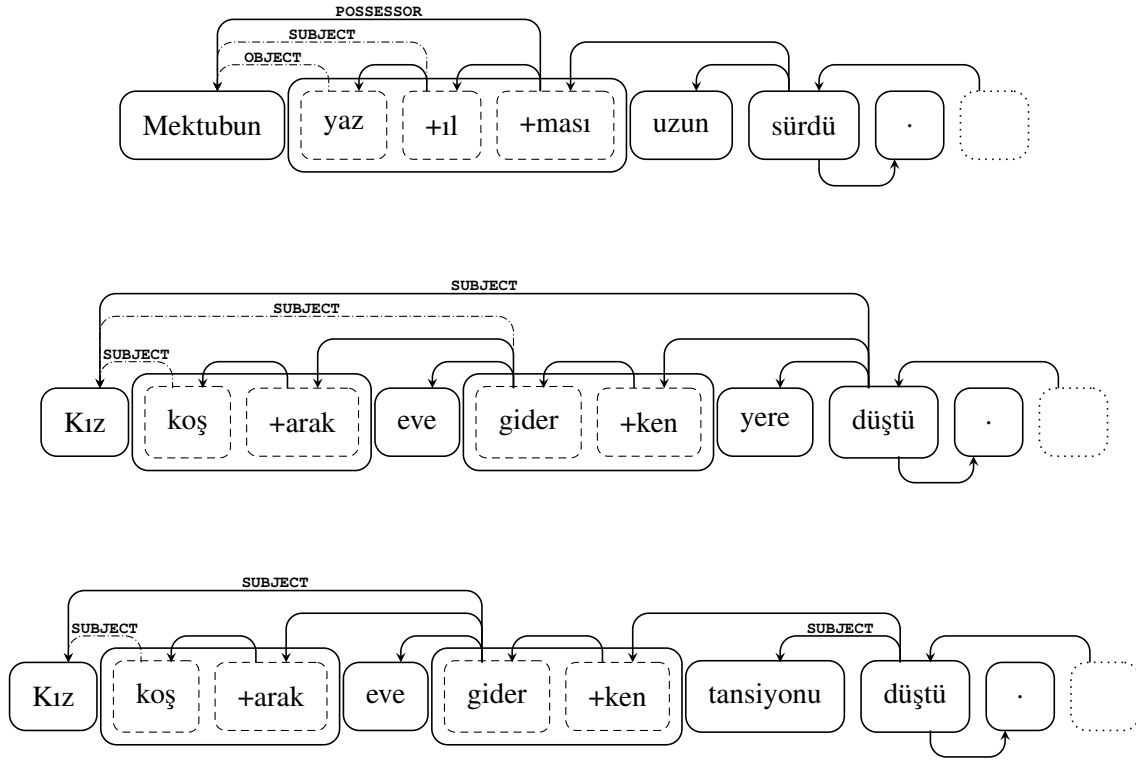
Might be mistaken for: A variety of dependencies.



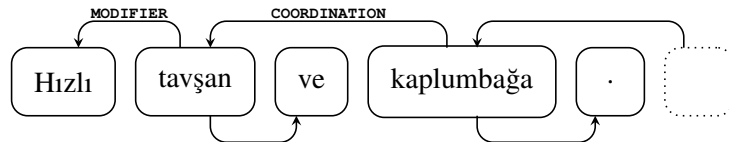
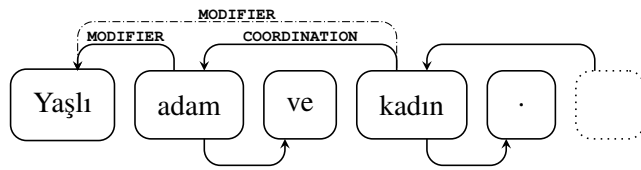
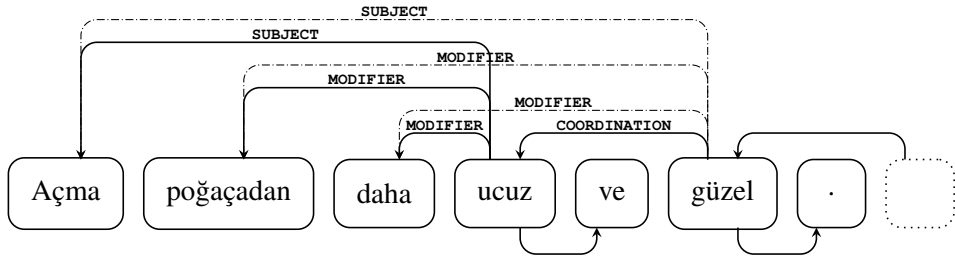
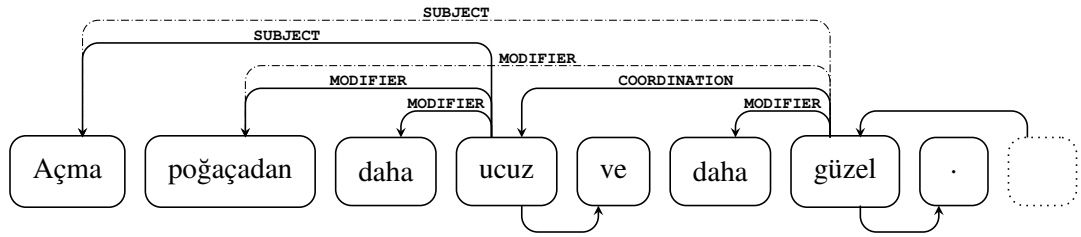
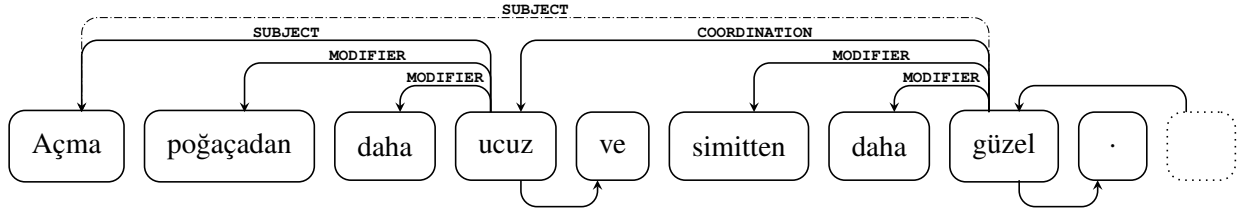
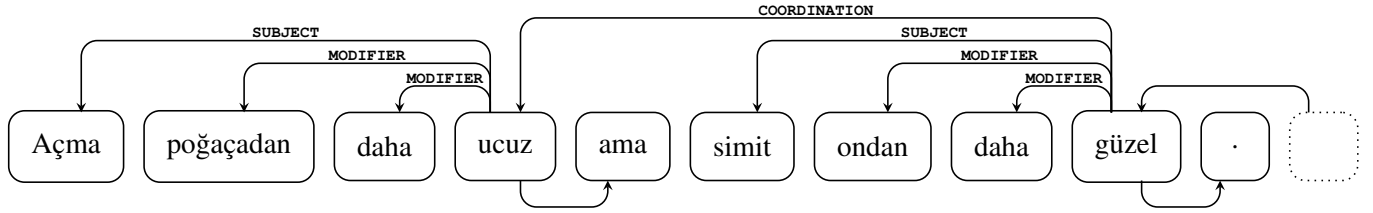
2.6 Deep Dependencies

Definition: *Deep dependencies* are a semantic dependency category that goes beyond the syntactic restriction dictating that each dependent must be connected to a single head, because semantically, a constituent is not necessarily associated with only a single other constituent. When a constituent semantically modifies more constituents than the one it syntactically modifies in its *surface dependency*, that constituent is allowed multiple heads, and additional *deep dependency* arcs may be drawn from it.

Usage 1: All verbs, as well as non-terminal IGs that were verbs at their current stage of derivation, may be investigated for their logical subjects, objects and modifiers, which would be connected to the verb via deep dependencies even if their surface dependencies are with other constituents. In addition to the predicate of the main sentence, such verbal constituents are most often found as the predicates of subordinate clauses, as well as the stems of gerundives.



Usage 2: Constituents in coordination may or may not be sharing their modifiers, and surface dependencies are not sufficient to express this distinction. Therefore, coordination structures are an important focal point of deep dependencies. Shared modifiers should be connected to all of the common heads in coordination, rather than to only one with a surface dependency, while non-shared modifiers should be left as they are.



ÖZGEÇMİŞ



Ad Soyad: Umut Sulubacak

Doğum Yeri ve Tarihi: İstanbul, 19 Nisan 1990

E-Posta: sulubacak@itu.edu.tr

Lisans: İstanbul Teknik Üniversitesi

Y. Lisans: İstanbul Teknik Üniversitesi

Yayın ve Patent Listesi:

TEZDEN TÜRETİLEN YAYINLAR/SUNUMLAR

- **Sulubacak U.** ve Eryiğit G., 2015: A Redefined Turkish Dependency Grammar and Its Implementations: A New Turkish Web Treebank and the Revised Turkish Treebank. *Under review.*
- Pamay T., **Sulubacak U.**, Torunoğlu-Selamet D. ve Eryiğit G., 2015: The Annotation Process of the ITU Web Treebank. *The Ninth Linguistic Annotation Workshop (LAW 2015), co-located with the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, June 5, 2015 Denver, Colorado, USA.
- Eryiğit G., Adalı K., Torunoğlu-Selamet D., **Sulubacak U.** ve Pamay T., 2015: Annotation and Extraction of Multiword Expressions in Turkish Treebanks. *The Eleventh Workshop on Multiword Expressions (MWE 2015), co-located with the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, June 4, 2015 Denver, Colorado, USA.
- **Sulubacak U.** ve Eryiğit G., 2014: ITU Treebank Annotation Manual. *Tez eki.*