TÜRKÇE İÇİN GÖZETİMSİZ
SÖZDİZİMSEL BELİRSİZLİK GİDERME
Doktora Tezi

Özkan ASLAN

Eskişehir, 2017

TÜRKÇE İÇİN GÖZETİMSİZ SÖZDİZİMSEL BELİRSİZLİK GİDERME

Özkan ASLAN

DOKTORA TEZİ

Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Doç. Dr. Serkan GÜNAL
(İkinci Danışman: Doç. Dr. Bekir Taner DİNÇER)

Eskişehir
Anadolu Üniversitesi
Fen Bilimleri Enstitüsü
Ağustos, 2017

Bu Tez Çalışması BAP Komisyonunca kabul edilen 1410F415 no.lu proje kapsamında desteklenmiştir.

JÜRİ VE ENSTİTÜ ONAYI

Özkan ASLAN'ın "TÜRKÇE İÇİN GÖZETİMSİZ SÖZDİZİMSEL BELİRSİZLİK GİDERME" başlıklı tezi 15/08/2017 tarihinde aşağıdaki jüri tarafından değerlendirilerek "Anadolu Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği"nin ilgili maddeleri uyarınca, Bilgisayar Mühendisliği Anabilim dalında Doktora tezi olarak kabul edilmiştir.

	Unvanı-Adı Soyadı	**İmza**
Üye (Tez Danışmanı)	**: Doç. Dr. Serkan GÜNAL**	**.....**
Üye	**: Prof. Dr. Ümit Deniz TURAN**	**.....**
Üye	**: Prof. Dr. Rifat EDİZKAN**	**.....**
Üye	**: Yrd. Doç. Dr. Alper BİLGE**	**.....**
Üye	**: Yrd. Doç. Dr. Uğur GÜREL**	**.....**

Prof. Dr. Nedim DEĞİRMENCİ

Enstitü Müdürü

ÖZET

TÜRKÇE İÇİN GÖZETİMSİZ SÖZDİZİMSEL BELİRSİZLİK GİDERME

Özkan ASLAN

Bilgisayar Mühendisliği Anabilim Dalı

Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, Ağustos, 2017

Danışman: Doç. Dr. Serkan GÜNAL

(İkinci Danışman: Doç. Dr. Bekir Taner DİNÇER)

Doğal dillerde bir tümce, her biri farklı yapısal yorumlara karşılık gelen birden çok sözdizim ağacı ile gösterilebilir. Bu durum sözdizimsel belirsizlik olarak adlandırılır. Sözdizimsel belirsizlik giderme, basitçe, tümceden elde edilen sözdizim ağaçlarının bağlama göre en uygun olandan en az uygun olana doğru sıralanmasıdır. Bu tezde, sözdizimsel belirsizlik giderme problemi Türkçe için ele alınmış ve gözetimsiz yönteme dayanan bir çözüm önerilmiştir. Yöntemin gözetimsiz olarak adlandırılmasının nedeni sözdizim ağaçlarının sıralanmasında kullanılan olasılık modellerinin imlenmemiş bir metin koleksiyonundan elde edilmiş olmasıdır.

Tez kapsamında, sözdizimsel belirsizlik giderme işini gerçekleştirmek amacıyla, sözdizimsel çözümleyici, Morfolog adlı biçimbilimsel çözümleyici ve TrLex adlı sözlükçe gibi özgün altyapı ögeleri tasarlanmış ve bunları eşgüdümlü biçimde yöneten TMoST adlı bir dizge oluşturulmuştur. Ayrıca öbek yapı dilbilgisine dayanan yeni bir tümce çözümleme gösterimi önerilmiş ve bu gösterimde biçimbilimsel ve sözdizimsel yapıları birlikte işleyebilmeyi sağlayan ve dizimbirim adı verilen yeni bir kavram tanıtılmıştır. Çalışmada, bazıları özgün olan 24 olasılık modeli kullanılmıştır. Modellerin problem üzerindeki başarımını ölçmeye imkân veren AUT adlı bir ağaç yapılı derlem üretilmiştir.

Alanyazında sözdizimsel belirsizlik giderme için başarım, en uygun ağacın sıralamada bulunduğu konum ile veya birinci sıradaki ağacın en uygun ağaca olan benzerliği ile ölçülmektedir. Tezde iki yeni başarım ölçüsü daha önerilmiş ve bağıntı adı verilen ölçünün daha kararlı olduğu değerlendirilmiştir.

Olasılık modelleri tek başına kullanıldığında en iyi başarım, üçlü biçimbirim dil modeliyle elde edilmiştir. Modeller birleştirildiğinde ulaşılan en iyi bağıntı değeri ise yaklaşık 0,41 olmuştur.

Anahtar Sözcükler: Sözdizimsel belirsizlik giderme, Biçimbilimsel çözümleme, Derlem, Sözlükçe, Öbek yapı dilbilgisi.

ABSTRACT

UNSUPERVISED SYNTACTIC DISAMBIGUATION FOR TURKISH

Özkan ASLAN

Department of Computer Engineering

Anadolu University, Graduate School of Sciences, August, 2017

Supervisor: Doç. Dr. Serkan GÜNAL

(Co-Supervisor: Doç. Dr. Bekir Taner DİNÇER)

In natural languages, a sentence can be represented by more than one syntax tree, each one corresponding to different structural interpretations. This is called syntactic ambiguity. To put it simply, in syntactic disambiguation, the syntactic trees obtained from the sentence are ranked from the most appropriate to the least appropriate based on the context. In this dissertation, the problem of syntactic disambiguation is addressed for Turkish and a solution based on an unsupervised method is proposed. The reason for naming the proposed method as unsupervised is that the probability models used for sorting syntax trees are derived from an unannotated text collection.

Within the scope of the dissertation, in order to realize the syntactic disambiguation process, novel infrastructure items including a syntactic parser, a morphologic analyzer called Morfolog, a lexicon called TrLex are designed and a system named TMoST that manages them in a coordinated manner is constituted. Besides, a new sentence representation based on phrase structure grammar is proposed and a new concept called syntheme, which allows morphological and syntactic structures to work together, is introduced. In the study, 24 probabilistic models, some of which are novel, are used. In order to measure the performance of the models over the problem, a treebank called AUT is constituted as well.

In the literature, the performance for syntactic disambiguation is commonly measured by the position of the best tree in the ranking or by the similarity of the first tree to the best one. In the dissertation, two new performance measures are proposed and it is revealed that the measure called correlation is more stable.

When the probabilistic models are used individiually, the best performance is obtained with the morpheme trigram language model. When the models are combined, the best correlation value is achieved as 0.41 approximately.

Keywords: Syntactic disambiguation, Morphological analysis, Corpus, Lexicon, Phrase structure grammar.

TEŞEKKÜR

Tecrübesi ve yapıcı yaklaşımıyla tezime çok önemli katkılarda bulunan danışmanım Doç. Dr. Serkan GÜNAL'a; doğal dil işleme ile hesaplamalı dilbilim alanlarına ilgi duymamı sağlayan ve tezin kilit noktalarının birçoğu için ilham kaynağım olan ikinci danışmanım Doç. Dr. B. Taner DİNÇER'e; dilbilim alanındaki bilgi ve tecrübelerinden yararlandığım Prof. Dr. Ümit Deniz TURAN'a; yerinde müdahaleleriyle tezin uygulama yönünün güçlenmesini sağlayan Yrd. Doç. Dr. Alper BİLGE'ye; yapıcı eleştirileriyle teze katkıda bulunan Prof. Dr. Rifat EDİZKAN ve Yrd. Doç. Dr. Uğur GÜREL'e saygılarımı ve teşekkürlerimi sunarım.

Tez konusunda tavsiyelerini paylaşan Yrd. Doç. Dr. Ahmet ARSLAN'a, pek çok konuda fikir alışverişi yaptığım H. Volkan AGUN'a ve çalışma ortamını benimle paylaşan Arş. Gör. Alper YARGIÇ'a muhabbet ve teşekkürlerimi sunarım.

Tez çalışmam boyunca kendilerine çok az vakit ayırmama rağmen bana her zaman sabır ve hoşgörüyle destek olan aileme, çalışma azmimi artıran ve beni her konuda yüreklendiren değerli eşim Dr. Ezgi ASLAN'a sevgi, saygı ve şükranlarımı sunarım.

…/…/2017

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Bu tezin bana ait, özgün bir çalışma olduğunu; çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilemeyen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; bu çalışmanın Anadolu Üniversitesi tarafından kullanılan "bilimsel intihal tespit programı"yla tarandığını ve hiçbir şekilde "intihal içermediğini" beyan ederim. Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçlara razı olduğumu bildiririm.

……

(İmza)

Özkan ASLAN

(Adı-Soyadı)

İÇİNDEKİLER

TABLOLAR DİZİNİ

ŞEKİLLER DİZİNİ

ŞEKİLLER DİZİNİ (Devam)

KISALTMALAR DİZİNİ

AUT : Anadolu University Treebank (Anadolu Üniversitesi Ağaç Yapılı Derlemi)

AYD : Ağaç Yapılı Derlem (Treebank)

BBG : Biçimbilimsel Belirsizlik Giderme (Morphological Disambiguation)

BÇ : Biçimbilimsel Çözümleyici (Morphological Analyzer)

BD : Bağımlılık Dilbilgisi (Dependency Grammar)

BSÇ : Biçim-Sözdizimsel Çözümleyici (Morpho-Syntactic Parser)

DDİ : Doğal Dil İşleme (Natural Language Processing)

HD : Hesaplamalı Dilbilim (Computational Linguistics)

LMF : Lexical Markup Framework (Sözlüksel İmleme Çerçevesi)

OBBD : Olasılıksal Bağlam Bağımsız Dilbilgisi (Probabilistic Context-Free Grammar)

OD : ODTÜ Derlemi (METU Corpus)

OSTAD : ODTÜ-Sabancı Türkçe Ağaç Yapılı Derlemi (METU-Sabancı Turkish Treebank)

ÖB : Öbek Belirleme (Chunking)

ÖYD : Öbek Yapı Dilbilgisi (Phrase Structure Grammar)

SBG : Sözdizimsel Belirsizlik Giderme (Syntactic Disambiguation)

SÇ : Sözdizimsel Çözümleyici (Syntactic Parser)

SDM : Sonlu Durum Makinesi (Finite State Machine)

TMoST : Turkish Morpho-Syntax Tool (Türkçe Biçim-Sözdizim Aracı)

1. GİRİŞ

İki tümcenin benzerliği nasıl ölçülebilir? Buna verilecek en basit yanıt şöyle olurdu: ***İki tümce ne kadar çok ortak sözcük içeriyorsa o kadar benzerdir***. Yüzeysel olarak bakıldığında bu yanıt yeterlidir. Ancak yapı ve anlam gibi "gözle görünmeyen" kavramlar açısından bakıldığında daha ayrıntılı bir açıklamaya ihtiyaç vardır. Aşağıda verilen tümceler iki farklı benzerliği örnekler:

[1] *Şapkalı çocuk kırmızı bisiklete bindi*
[2] *Şapkalı çocuk kırmızı bisiklete*
[3] *Çocuk bisiklete bindi*

"Ortak sözcük içerme" ölçütüne göre [1]'e en çok [2] benzemektedir. Ancak yapı ve anlam bakımından [1]'e en yakın tümce [3]'tür. [1] ile [3] arasındaki benzerliği açıklayabilmek için şu varsayım yapılmalıdır: ***Sözcüklerin tümcede farklı görevleri vardır; bu görevler yapı ve anlam bakımından eşit önemde değildir***. Öyleyse [2] bir tümce olarak değerlendirilemez çünkü tümceyi oluşturan en önemli ögeden (eylem/yüklem) yoksundur.

Başka bir soruyla devam etmek gerekirse: Tümceyi oluşturan sözcükleri keyfî biçimde sıralamak mümkün müdür? Yanıt, aşağıda verilen örneklerde açıkça görülmektedir:

[4] *Şapkalı kırmızı bindi bisiklete çocuk*
[5] *Kırmızı bisiklete şapkalı çocuk bindi*

Hem [4] hem de [5]'teki sözcükler [1]'e göre farklı sıralanmış olsa da, [1]'e yapıca daha çok benzeyen tümce [5]'tir. Bunun nedenini açıklayabilmek için yeni bir ögeye gerek vardır. Bu öge, sözcükleri içine alan ve tümceyi inşa eden bir yapı olmalıdır. İlgili yapıya kurucu öbek (constituent) adı verilir. Kurucu öbek, herhangi bir sözcük grubu olmanın ötesinde, tümcenin temel bileşenidir. Örneğin "Mavi kapılı evin küçük

penceresi açıldı" tümcesinden "mavi kapılı ev" gibi bir basit öbek çıkarılabilir ancak bu, tümce için bir kurucu öbek değildir. Kurucu öbek, tümcede yapılan işi anlatıyorsa eylem, işi gerçekleştireni işaret ediyorsa özne, eylemi tamamlayan bir öge ise tümleç ve eylemi niteleyen seçimlik bir öge ise eklentidir. Şimdi bu bilgiler ışığında [1]'i oluşturan öbekler incelenirse:

[6] *[Şapkalı çocuk] [kırmızı bisiklete] [bindi]*

[6]'da görüldüğü gibi örnek tümce üç kurucu öbeğe ayrılabilir. Bunlar da sırayla özne, tümleç ve eylemdir[1]. Bu çözümleme, basit bir yapısal çözümleme olup doğal dil işleme alanyazınında kurucu öbek belirlemeye[2] (constituent chunking) karşılık gelir. Öbek belirleme ile [1] ve [5] arasındaki benzerlik açıklanabilir ve buradan şu varsayıma ulaşılır: ***Tümce içinde yalnızca kurucu öbekler yer değiştirebilir***.

Yapısal çözümleme için kurucu öbek belirleme yeterli olsaydı karmaşık algoritmalara ve birçok çalışmaya gerek kalmazdı fakat tümceler her zaman doğrusal biçimde sıralanmış bir öbekler dizisi değildir. Bu noktada önemli bir kavramla karşılaşılır: özyineleme (recursion). Tümce bağlamında özyineleme, bir kurucu öbeğin, içerisinde başka bir tümce yapısını içermesi şeklinde tanımlanabilir. Bu durum kuramsal olarak sonsuz gerilemeye izin verir: "Tümcenin içindeki kurucu öbeğin içindeki tümcenin içindeki kurucu öbeğin içindeki…" Ancak pratikte özyinelemenin bir sınırı vardır; zihin ve bellek böyle yapıların belli bir derinliğe ulaşmasına izin vermektedir.

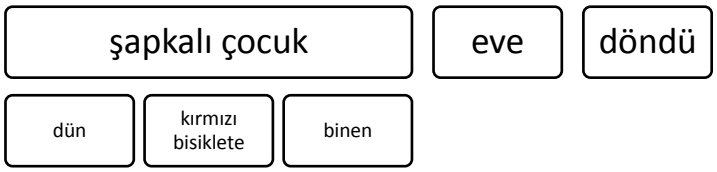
Özyineleme ile birlikte bir tümceyi yapısal olarak çözümlemenin her zaman çok basit olamayacağı açıkça görülür. Aşağıda verilen örnek, özyinelemeli yapılar içermektedir:

[7] *Dün kırmızı bisiklete binen şapkalı çocuk eve döndü*

[1] Yüklem yerine eylem deyimini kullanıyoruz çünkü yüklem, özne hariç diğer kurucu öbekleri içine alan daha büyük bir yapıdır.
[2] Tezde kısaca "öbek belirleme" olarak anılacak.

[8] *[Dün kırmızı bisiklete binen şapkalı çocuk] [eve] [döndü]*

[8], [7]'nin doğrusal gösterimidir. Ancak görüldüğü üzere, bu gösterim ile öznenin içinde yer alan tümcenin öbekleri belirtilememiştir. Özyineleme içeren bir tümce, hiyerarşik gösterimle daha iyi ifade edilebilir. **Şekil 1.1**'de [7] için bir hiyerarşik gösterim örneği verilmiştir.



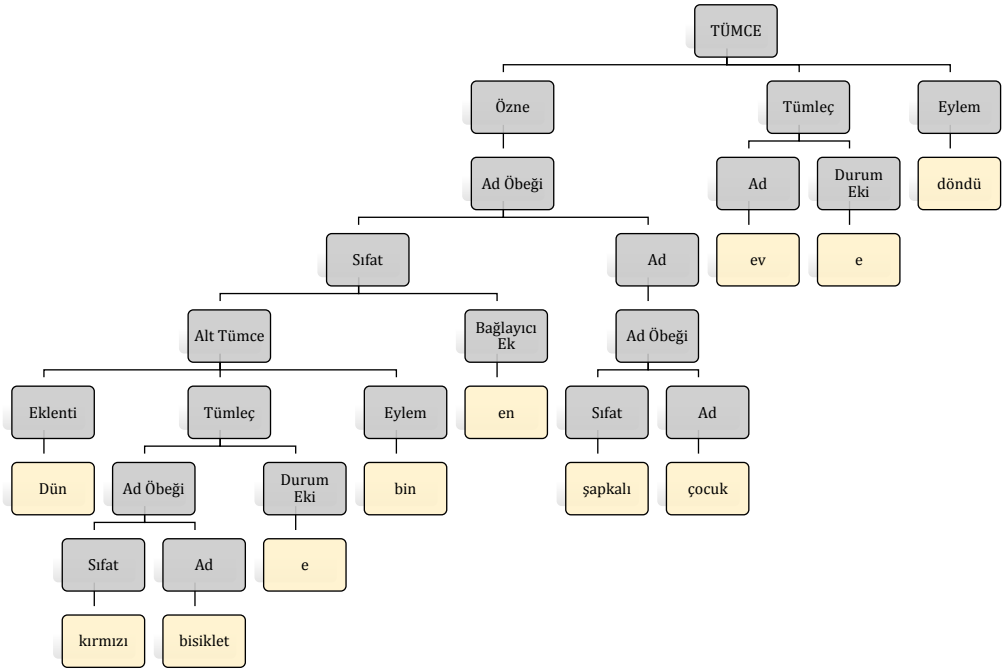
Şekil 1.1. *Hiyerarşik Gösterim*

Hiyerarşik gösterimle öbeklerin farklı düzeylerden oluşabileceği fark edilir. Her bir düzey kendi içinde bir tümcedir. Örneğin "dün kırmızı bisiklete bin" bitimsiz bir tümce olup "en" ekiyle üstteki öbeğe bağlanmıştır. Bu nokta şimdilik akılda tutulup tekrar öbekler arası yer değiştirme meselesine dönülürse: En üstteki düzeyde öbeklerin sıralanmasıyla ilgili kuramsal bir sınırlama yok gibi görünmektedir (bk. **Şekil 1.1**). Örneğin şu tümce devrik yapıda olmasına rağmen kurallıdır: "Döndü eve dün kırmızı bisiklete binen şapkalı çocuk". Ancak öznenin bir alt düzeyindeki tümce için benzer bir yer değiştirme kurallı bir sonuç üretmemektedir: "Binen dün kırmızı bisiklete şapkalı çocuk eve döndü". O hâlde öbek sıralamasına ilişkin varsayım şöyle güncellenmelidir: ***Tümce içinde yalnızca kurucu öbekler yer değiştirebilir. Eğer tümce bir alt tümce ise eylemin yeri sabittir***. Tam da şimdi "en" ekiyle ilgili nokta hatırlanırsa, alt tümcenin eyleminin sabit olmasına yol açan etmenin bu ek olduğu görülür. Buradan da çok önemli bir varsayıma ulaşılır: ***Türkçede bazı ekler tümce yapısını belirleyici niteliktedir***.

Alt tümceleri tümceye yapıştıran bu ekler gibi önemli olan bir diğer grup da durum ekleridir. Durum eklerinin sözdizimsel görevi, bir tümce düzeyinde kurucu öbekler ile eylem arasındaki bağlantıyı göstermektir. [7]'deki "eve" sözcüğü "ev" birimi ile "dön" birimi arasındaki ilişkiyi tanımlamaktadır. Özne ile eylem arasında ise durum

eksiz bir ilişki vardır. Gramercilerin çoğuna göre ekin olmaması da bir bilgidir ve yalın durum olarak adlandırılmaktadır.

Eklerin tümce yapısındaki önemli rolüne değindikten sonra bu tanecikselliğin (granularity) nasıl gösterilebileceği gündeme gelir. Burada hiyerarşik gösterimden daha ayrıntılı bir ifadeye izin veren ağaç yapısı işe yarayacaktır. [7]'nin ağaç yapısındaki gösterimi **Şekil 1.2**'de verilmiştir.

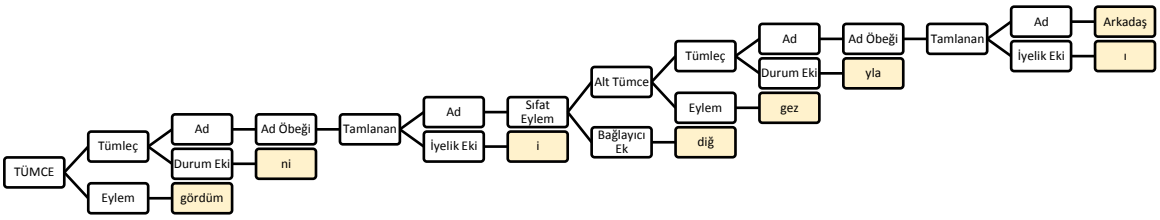


Şekil 1.2. *Tümcenin Ağaç Yapısı Gösterimi*

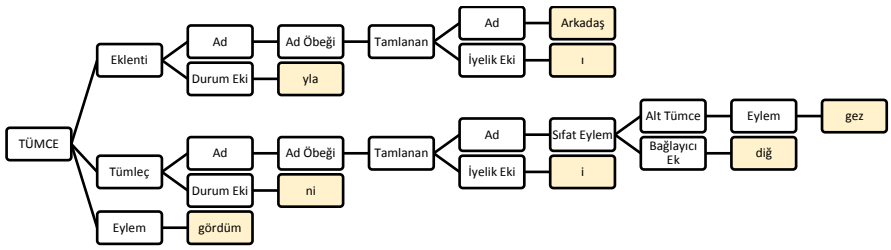
Ağaç yapısı gösterimi sayesinde öbeklerin yalnızca yüzeysel biçimleri değil yapısal ilişkileri de temsil edilebilmektedir. Örnek olarak verilen ağaç gösterimi hiyerarşik ya da doğrusal gösterime kıyasla çok büyük miktarda enformasyon içerir. Ayrıca bu gösterim bir dilbilgisi görüşünü de yansıtır. Tez, ileriki bölümlerde değinilecek olan bu görüşü temel almaktadır.

Bir tümce birden çok biçimsel ve yapısal yorum üretebilir. Dolayısıyla tümce için birden çok ağaç gösterimi söz konusu olabilir. Buna birçok etken neden olmaktadır. Bunlar, bir sözcüğün hangi öbeğe ait olduğunu belirleme, bir sözcüğün nasıl ayrıştırılacağını belirleme ve bir sözcüğün hangi görevde olduğunu belirleme gibi

süreçlerdir. Örneğin “Arkadaşıyla gezdiğini gördüm” tümcesinin ürettiği ağaçlar **Şekil 1.3** ve **Şekil 1.4**’te verilmiştir.



Şekil 1.3. "Arkadaşıyla gezdiğini gördüm" Tümcesi İçin Birinci Ağaç Gösterimi



Şekil 1.4. "Arkadaşıyla gezdiğini gördüm" Tümcesi İçin İkinci Ağaç Gösterimi

Şekil 1.3 ve **Şekil 1.4**’te görülen ve sözdizim ağacı olarak adlandırılan ağaçlar aynı tümcenin iki ayrı yapısal yorumudur. Öyleyse örnek tümce için bir belirsizlik mevcuttur. Bu belirsizlik tümce yapılarına ilişkin bir belirsizlik olduğu için, tümce yapılarını inceleyen sözdizim alanına atfen, sözdizimsel belirsizlik olarak adlandırılır. Örnekteki ağaçlardan biri, çoğu zaman, bağlama göre ya da anlamca daha olasıdır. İşte bu ağacı belirleme işine sözdizimsel belirsizlik giderme adı verilmektedir.

1.1. Tezin Konusu

Bu tezde ele alınan temel problem, verilen bir tümceden sözdizimsel çözümleme yoluyla üretilen sözdizim ağaçlarının puanlanması ve bunun bir sonucu olarak ağaçların sıralanması işidir. Eğer tümce için yalnızca bir sözdizim ağacı üretilebiliyorsa kesinlik söz konusudur. Ancak çoğu zaman bir tümce için dilbilgisi kurallarına uyan birçok sözdizim ağacı elde etmek mümkündür. Bu durumda ağaçlardan hangisinin ya

da hangilerinin daha olası olduğunu belirlemek önemli bir uğraştır. Bir tümcenin yapısını ortaya koyan geçerli sözdizim ağacı belirlendiğinde tümceyi oluşturan birimler tanımlanmış olur. Bu noktada tümce ile ilgili biçimbilimsel ve sözdizimsel bütün bilgiler elde edilmiş demektir. Bu bilgiler anlamsal ilişkilerin çıkartılmasında ve farklı çözümlemelerde kullanılabilir.

Tümceler, içinde yer aldıkları bağlamla birlikte değerlendirilir ve anlam kazanır. Başka bir ifadeyle, bağlam verildiğinde tümce için yalnızca bir sözdizimsel yapı geçerlidir. Bu çalışmada bağlam bilgisi kullanılmamış olup tümceden elde edilen bütün olası sözdizimsel yapılar işlenmiştir. Özetle doğal dil konuşucularının sıradan bir yeteneği olan bağlamından bağımsız bir tümceyi anlamlandırma işine sonuç odaklı bir yaklaşım sağlanmaya çalışılmıştır.

Doğal dil işlemede farklı zorluk derecelerine sahip olan birçok problem bulunmaktadır. Türkçe için daha çok gövdeleme, sözcük türü belirleme, biçimbilimsel çözümleme ve biçimbilimsel belirsizlik giderme konularında çalışmalar yapılmış olup sözdizimsel çözümleme üzerine çok az sayıda çalışma bulunmaktadır. Doğrudan sözdizimsel belirsizlik gidermeyi konu edinen bir çalışma ise bilindiği kadarıyla yoktur. İlgili konunun seçilmesinin en önemli nedeni bu eksikliktir. Özellikle tümce yapısı çözümlemesinden beslenen üst seviyedeki çalışmalar için kapsamlı bir sözdizimsel çözümleyiciye ihtiyaç vardır. Tezde bu konunun seçilme nedenlerinden bir diğeri de yazarın Türkçenin ekleri ile ilgili çalışmasından (Aslan, 2008) faydalanmak ve biçimbilimsel süreçlerle sözdizimsel süreçleri bir araya getirmek olmuştur.

Çalışma sırasında birçok bulgu elde edilmiş, bazılarının öngörülerle uyuştuğu belirlenmiştir. Örneğin tümce çözümlemesinde eylem merkezli yaklaşımı kullanmanın pratikliği deneyimlenmiştir. Bu, aynı zamanda Türkçe için hem dilbilim çerçevesinde hem de doğal dil işleme çerçevesinde eylemleri konu alan çalışmaların eksikliğini ortaya çıkarmıştır. Bir başka uyuşma, biçimbirimleri temel alma ile ilgilidir. Mevcut çalışmalarda hesaplama birimi ya da çözümleme odağı olarak sözcükler ve çekimleme durakları seçilmektedir. Tezde biçimbirimlerin dil modelleri için en uygun hesaplama birimi olduğu iddia edilmektedir. Edinilen bir başka deneyim de imlenmiş derlem sorunu hakkındadır. Türkçe doğal dil işleme çalışmalarında bilgi-yoğun derlemlerin

eksikliği ortadadır. Bu çalışmada bu eksikliği azaltacak bir ürün sunulamamış ancak problemi çözmek için gözetimsiz yöntemden yararlanılmıştır. Bu yöntem, belirsizlik içeren metinlerden, dolaylı gözetimsiz bir şekilde bilgi çıkarımına dayanmaktadır. Deneylerde elde edilen sonuçlar, modellerin hiçbirinde imlenmiş (altın) veri kullanılmadığı hesaba katıldığında, oldukça tatmin edicidir.

1.2. Tezin Kapsamı

Bu tez çerçevesinde üretilen dizge, yalnızca Türkiye Türkçesinde yazılmış metinleri kabul eder. Dizgenin güncel sürümü noktalama işaretlerini dikkate almamakta, ancak sonraki çalışmalarda bunların işlenmesi planlanmaktadır. Dizge yalnızca kurallı tümceleri işleyebilir. Bunun yanında her bir tümce bir eylem ile bitmeli, özel ad ve birleşik eylem içermemelidir. Bu durum, dizgenin tümce tanıma oranını belirleyen en önemli kısıtlılıktır. Tezin sonrasındaki çalışmalarda dizgenin tanıma oranı artırılmaya çalışılacaktır. Çözümlemelerde tümce ötesi ya da tümceler arası etkileşim içeren herhangi bir bilgi kullanılmamıştır.

1.3. Tezin Düzeni

Tez dokuz bölümden oluşmaktadır. İkinci bölüm olan *Kuramsal Temeller*'de kısaca dilbilimin tanımı ve konusu işlenmiş, ardından Türkçenin temel özelliklerine değinilmiş ve son olarak doğal dil işlemenin tanımı verilerek temel konu başlıkları listelenmiştir. Üçüncü bölüm olan *Sözdizimsel Çözümleyici*'de tezin başlıca ürünü olan TMoST adlı dizgenin ana bileşenleri tanıtılmış, sözdizimsel çözümleme işinin temel adımları açıklanmış ve örneklenmiştir. Dördüncü bölüm olan *Biçimbilimsel Çözümleyici*'de TMoST'un bir bileşeni olarak tasarlanan biçimbilimsel çözümleyici alt dizgesi tanıtılmıştır. Ayrıca bununla ilgili önceki çalışmalar özetlenmiş, önerilen dizgenin ayrıntılı özellikleri tanıtılmış ve dizgenin mevcut bir başka dizge ile karşılaştırılmasından elde edilen bulgular sunulmuştur. Beşinci bölüm olan *Sözlükçe*'de biçimbilimsel çözümleyicinin bir bileşeni olan sözlükçe ayrıntılı olarak açıklanmıştır.

Bu amaçla önceki çalışmalar listelenmiş ve sözlükçenin hazırlanma aşamaları sıralanmıştır. Burada yer alan *Bulgular* bölümünde ise sözlükçeden elde edilen istatistiksel veriler paylaşılmıştır. Altıncı bölüm olan *Veri*'de tezde gerçekleştirilen deneylerde veri olarak kullanılan metin koleksiyonu ve ağaç yapılı derlem tanıtılmıştır. Yedinci bölüm olan *Model*'de mevcut çalışmalarda kullanılan ve tezde önerilen modeller açıklanmıştır. Sekizinci bölüm olan *Sözdizimsel Belirsizlik Giderme*'de tezin ana konusu olan sözdizimsel belirsizlik giderme ile ilgili ilk olarak önceki çalışmalar özetlenmiş, ardından önerilen yöntem açıklanmış ve son olarak deney sonuçları sunulmuştur. Dokuzuncu ve son bölüm olan *Tartışma*'da ise tezde elde edilen bütün sonuçlar bir arada yorumlanmıştır.

2. KURAMSAL TEMELLER

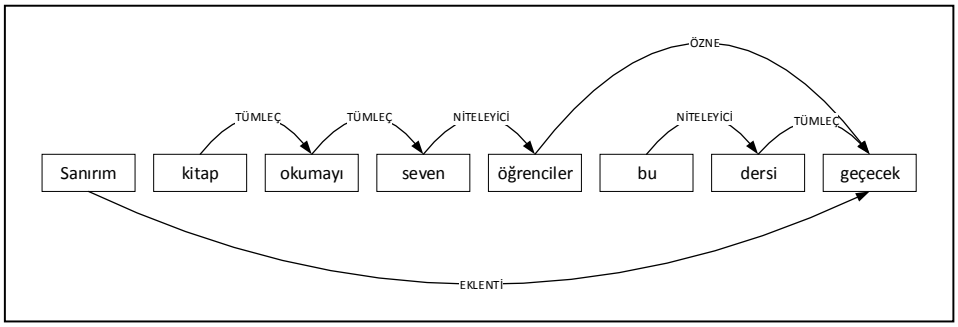
Bu bölümde tezde ele alınan tümce çözümlemesi konusunun kuramsal çerçevesini oluşturan dilbilim alanı ve uygulama dili olarak seçilen Türkçe hakkında temel bilgiler verilecek, buna ek olarak, tez konusunun mühendislik boyutu olan doğal dil işleme kısaca tanıtılacaktır.

2.1. Dilbilim

Dilbilim (linguistics), dili konuşanlar arasındaki iletişimi sağlayan dil yetisini ve doğal dilleri inceleyen bir bilim dalıdır. Dilbilim yalnızca olguları gözlemler ve betimlemeler yapar. Bu anlamda matematik gibi nesnesi, alanı ve yöntemi olan kesin bir bilimdir. "Dilbilimin ve dilbilimcilerin görevi, bir dilsel topluluğa ait bireylerin zihinlerindeki ortak özelliği geniş ve ayrıntılı biçimde tanıtmaya çalışmaktır (Kıran ve Kıran, 2013, s. 47)".

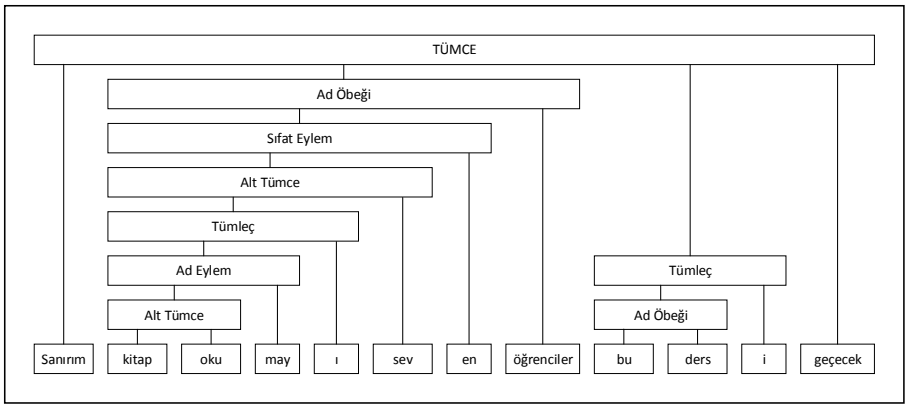
Dilbilgisi (grammar), "Bir dilin sesten söz dizimine kadar bütün birimlerini, bu birimlerin yapı ve anlam özelliklerini araştıran bilim dalıdır (Karaağaç, 2013, s. 289)." Dilbilgisi dilin seslerini, sözcüklerin yapı, anlam ve kökenlerini, tümce kuruluşlarını ve bunlarla ilişkili kuralları inceler ve dört türe ayrılır: kuralcı (normative), betimleyici (descriptive), tarihsel (historical) ve karşılaştırmalı (comparative). Bu sınıflandırmaya göre kuralcı dilbilgisi yazı dilini inceler ve kurallara bağlar; betimleyici dilbilgisi dilin belli bir zamandaki durumunu anlatır; tarihsel dilbilgisi dilin geçirdiği evrimi inceler ve karşılaştırmalı dilbilgisi de ortak bir ata dilden ayrışıp yayılan dilleri karşılaştırır. Karşılaştırmalı dilbilgisi bir ana dilin lehçelerini konu edindiğinde dilbilim, dünyadaki bütün dilleri ele aldığında ise genel dilbilim olarak adlandırılır. Bu bakış açısından bakıldığında, dilbilim, dilbilgisinin alt alanlarından birinin özel bir durumudur. Ancak dilbilimcilere göre dilbilim özellikle kuralcı dilbilgisi ile bağdaşmaz; daha çok betimleyici dilbilgisi ile yakınlık gösterir. Bu nedenle dilbilimciler dilbilimi ayrı bir bilim dalı olarak değerlendirir (Ediskun, 2005, s. 65).

Kuramsal dilbilimde birçok farklı dilbilgisi çerçevesi geliştirilmiştir. Tez kapsamında, bunlar arasından bağımlılık dilbilgisi (BD, dependency grammar) ve öbek yapı dilbilgisi (ÖYD, phrase structure grammar) ele alınacaktır. BD, Tesnière tarafından önerilmiş ve bağımlılık ilişkilerine dayanan bir kuramdır. BD ile bir tümcenin yapısı *baş* ve *bağımlı* adı verilen ögeler arasındaki bağımlılıklara dayanarak modellenir. *Baş*, bir öbeği temsil eden temel ögedir. *Bağımlı* ise öbeğin oluşması için gerekli olan ya da öbeği tamamlayan unsurlardır. BD'de tümce çözümlemesi, bağımlılık ilişkilerini temsil eden oklarla gösterilir. **Şekil 2.1**'de bir BD yapı çözümlemesi görülmektedir.



Şekil 2.1. *Bağımlılık Dilbilgisi İle Tümce Çözümlemesi*

İlk kez Chomsky (1957) tarafından tanımlanan ÖYD ise, öbek yapı kurallarından oluşur. Bu kurallar öbeklerin oluşum şekillerini belirler. Her bir kural bir öbeğin hangi sözlüksel kategoriler bir araya geldiğinde ortaya çıkacağını ifade etmektedir. Örneğin "[Ad Öbeği] → [Sıfat] [Ad]" kuralı, yan yana gelmiş bir sıfat ve bir adın, ad öbeği oluşturacağını bildirir. Kural yapısında, soldaki öge, oluşan birimi, sağdaki ögeler ise bileşenleri göstermektedir. ÖYD'de özyinelemeli yapıları ifade etmek de mümkündür. Örneğin şu kural ile özyinelemeli yapılar tanımlanabilir: "[Sıfat Öbeği] → [Sıfat] (Sıfat Öbeği)". **Şekil 2.2**'de bir ÖYD yapı çözümlemesi görülmektedir.



Şekil 2.2. *Öbek Yapı Dilbilgisi İle Tümce Çözümlemesi*

İki kuramsal çerçeve karşılaştırıldığında şu noktalar öne çıkar:

- BD'de düğümler sözcükler iken, ÖYD'de düğümler sözcükler (Türkçe için biçimbirimler) ya da bunların birleşmesiyle oluşan öbek etiketleri olabilir.
- BD'de ilişkilerin yönünü oklar, türünü ise çizgilerin üzerindeki etiketler belirler; ÖYD'de çizgiler etiketsizdir ve ilişkiler doğrudan değil ancak aynı yapı bileşeni olma kimliğiyle tanımlanabilir.
- BD yapıları ÖYD yapılarına göre daha az düğüm içerir ve bu açıdan ekonomiktir. Buna karşılık ÖYD yapıları ise daha çok bilgi içermektedir.
- Bir ÖYD yapısı eğer uygun etiketlere sahipse BD yapısına dönüştürülebilir, ancak tersi mümkün değildir.

Bu tezde önerilen sözdizimsel çözümleyici daha çok ÖYD tipinde ağaç yapıları üretmektedir. Ancak özne, tümleç, eklenti gibi tümce içi ilişkileri de sunabildiği için BD'nin işlevselliğini de yansıtır.

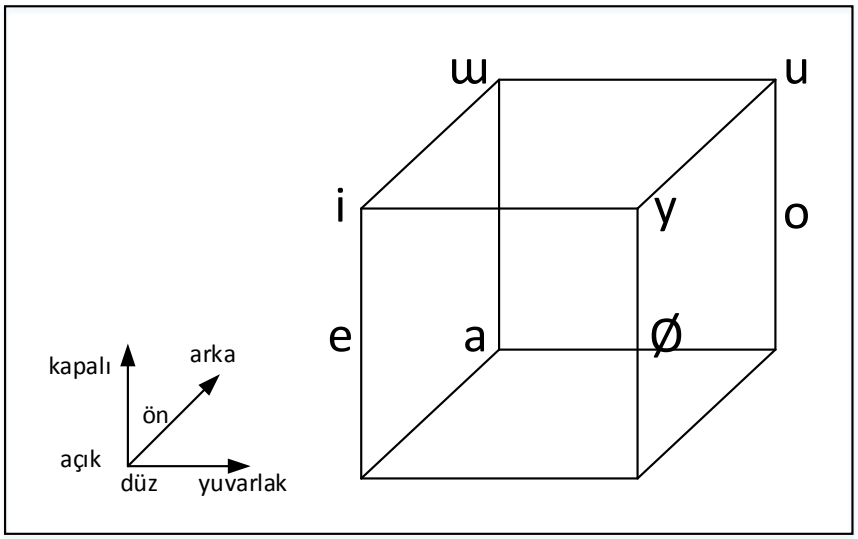
2.2. Türkçe

Altay dil ailesinin bir üyesi olan Türkiye Türkçesi[3] yalnızca Türkiye'de, çoğunluğunun ana dili olmak üzere, 70 milyondan fazla insan tarafından konuşulmaktadır. Ayrıca yabancı dil olarak Türkçe öğretimine ilgi son yıllarda gittikçe

[3] Tez boyunca "Türkçe" olarak anılacak.

artmıştır. Bu artışa özellikle Afrika ve Asya ülkeleri ile kurulan ticari ve kültürel ilişkilerin gelişmesi neden olmuştur. Bu açıdan değerlendirildiğinde Türkçe, önemi ve etkisi günden güne artan bir dil hâline gelmektedir.

Sesbilimsel bakışla incelendiğinde, Türkçede 8 patlamalı (p, b, t, d, c, ɟ, k, g), 7 sürtünmeli (f, v, s, z, š, ž, h), 2 patlamalı-sürtünmeli (č, ǰ), 2 genizsil (m, n), 3 akıcı (l, ɫ, r) ve 1 kayıcı (j) olmak üzere 23 ünsüz sesbirim bulunmaktadır. Ünlü sesbirimler ise ön-arka, düz-yuvarlak ve kapalı-açık boyutları üzerindeki konumlarıyla tarif edilmek üzere 8 adettir (Özsoy, 2004, s. 35). Ünlü sesbirimlerin ayrıntılı sınıflandırması **Şekil 2.3**'te incelenebilir.



Şekil 2.3. *Ünlü Sesbirimlerin Sınıflandırılması*

Yaygın kanıya göre Türkçe sesletim açısından oldukça düzenlidir. Bu düzenliliği ispatlamak için de "yazıldığı gibi okunan" bir dil olduğu iddiası öne sürülmektedir. Ancak sözü edilen özelliğe zıt olarak pek çok örnek gösterilebilir (Altun, 2010). Türkçedeki biçimsel düzenliliği hem yabancı dillerden geçen ögeler hem de iç dinamikleri olumsuz yönde etkilemiştir. Türkçedeki ses uyumsuzluklarıyla ilgili kapsamlı bir çalışma Clements ve Sezer (1982) tarafından yapılmıştır. Ses değişimi, eklerin altbiçimciklerinin çeşitlenmesine neden olur. Bu çeşitlenmede genellikle

belirleyici olan kural, tabanın son ünlüsünün ince veya kalın ve geniş veya dar oluşunu temel alır (Deny, 2000).

Biçimbilimsel açıdan oldukça zengin bir ek dizgesine sahip, eklemeli (agglutinative) bir dil olan Türkçede bir tek kökten çok sayıda başsözcük (lemma) ve sözcük biçimi (word form) üretilebilmektedir. Hankamer (1989) 20.000 ad kökü ve 10.000 eylem kökünden oluşan bir sözlükçe kullanıldığında Türkçede yaklaşık 200 milyar sözcük biçimi oluşabileceğini iddia etmiştir.

Ön ek ve ara ek bulunmayan Türkçede sözcük yapımı (word formation) ve çekimleme daima son ekler aracılığıyla gerçekleştirilir. Ancak özellikle yabancı dillerden geçen kavramları karşılayan sözcüklerde gözlenen öz-, iç-, dış-, ilk-, ön- gibi bazı sözcükler ön ek algısı oluşturmaktadır. Bunlar bağımsız olarak da anlamlı olan sözcükler olduğu için ön ek olarak yorumlanmaz ve bu tip yapıların birleşik sözcükler olduğu kabul edilir (Şahin, 2006). Türkçenin ekleri genellikle türetim ve çekim ekleri olarak iki grupta incelenir. Ancak az sayıda da olsa genellikle çekim eki görevinde olan bazı ekler yeni sözcük yapımında kullanılmışlardır. Benzer şekilde türetim eki olarak nitelenen bazı ekler de çekim ekleri kadar işlek olabilmektedir.

Sözcük türü sınıflaması açısından, ad ve eylem olmak üzere iki temel sözcük tabanı vardır (Grønbech, 2011). Türetim ekleri bu iki taban arasındaki geçişlerden hareketle genellikle addan ad, addan eylem, eylemden ad ve eylemden eylem türeten ekler şeklinde dört sınıfta incelenir. Türetim eklerinin eklendikleri taban üzerinde ürettikleri anlam değişmesine işlev dersek, bazı türetim ekleri tek işlevli bazıları çok işlevlidir. Bu konuda bir envanter çalışması Gedizli (2012) tarafından yapılmıştır.

Sözcükler öbekler halinde hareket ederler ve öbek yapı kuralları (Phrase Structure Rules) başlığı altında toplanan bir takım kurallara tabidirler. Bir öbek (phrase) dilbilgisel olarak birbirine bağlanmış tümce içinde birlikte hareket edebilen sözcük topluluğudur. Öbek aynı zamanda tümce içinde özne, yüklem, nesne gibi bir işlev üstlenen ve en az bir ya da daha fazla sözcük içeren ögedir. Her öbekte bir sözcük öbek kurucudur. Anlam ve yapı olarak öbekteki en önemli ya da merkezi rolü olan sözcüktür. Buna öbekte baş (head) denir. Öbekte başın yeri öbeğin başında ya da sonundadır. Türkçe eylem sonlu ve baş sonlu bir dildir (Alagözlü, 2016).

Türkçenin baş sonlu bir dil oluşu öbek yapılarının sola doğru dallanmasına yol açmaktadır. "… sentaks açısından yönetilen ögenin yöneten ögeden önce gelmesi

şeklindeki sözlüksel unsurların yönetilen-yöneten normu hâkim durumdadır. Bu normun, 'Altay dilleri sentaksının temel kuralı' olduğu söylenir … (Johanson, 2014, s. 42)".

Türkçe aynı zamanda serbest sözcük sıralamalı (free word-order) bir dildir. Bu özellik tümce içindeki bütün sözcüklerin hiçbir kısıt olmadan yer değiştirebileceğini ima etmez. Yer değiştirebilen birimler yalnızca kurucu öbeklerdir. Kurucu öbekler özne, tümleç ve eklenti olup bağlı bulundukları eylemle birlikte bir eylem öbeği kurarlar. Bir tümcede birden çok eylem öbeği bulunabilir. Dolayısıyla serbest sıralama ancak her bir eylem öbeği içindeki kurucu öbekler arasında gerçekleşebilir. Taylan'a (1984) göre, Türkçede bu sıralama genellikle özne-tümleç-eylem biçiminde gözlenir ancak farklı sıralamalar, konu, odak, arka plan vb. söylem işlevlerini kodlamak amacıyla kullanılabilmektedir.

2.3. Doğal Dil İşleme

Bu alt bölüme başlarken öncelikle bir adlandırma sorunundan bahsetmek yerinde olacaktır. Alanyazında doğal dil işleme (DDİ, natural language processing) ifadesi ile hesaplamalı[4] dilbilim (HD, computational linguistics) ifadesi zaman zaman birbirinin yerine kullanılmaktadır. Bu durum, her iki terimin açık bir tanımının olmamasından kaynaklanabileceği gibi, DDİ'nin İngilizce karşılığı olan NLP'nin[5] olumsuz imajı da HD'yi kullanma eğilimini artırabilmektedir. Bilinçli olmayan bu tercihlerin dışında aslında iki terim için kabaca da olsa bir ayrım yapılabilir.

HD kuramsal dilbilimin hedefleriyle daha uyumlu iken, DDİ daha uygulamaya dönük ve yararcı, dolayısıyla mühendislik yaklaşımına daha uygundur. HD merkeze aldığı dilbilimin temel problemlerine yanıtlar aramakta iken, DDİ dilbilimi bir araç olarak değerlendirmekte ve çalışmalarının merkezinde "daha verimli" yolları araştırmak yer almaktadır.

[4] "Computational" sözcüğüne karşılık olarak "berimsel" önerilmektedir (Bozşahin ve Zeyrek, 2000). Bu tezde "hesaplamalı" terimi tercih edilmiştir.
[5] Sahte bilim olarak nitelendirilen "Neuro-linguistic programming"in kısaltması da NLP'dir.

Bu çerçevede değerlendirildiğinde bu çalışmanın DDİ konulu bir tez olduğunu söylemek yerinde olacaktır fakat tezin dayandığı altyapının çoğunlukla nitel kaygılarla hazırlanması ve yazarın beslendiği kaynaklar itibarıyla bakıldığında bir HD çalışması olmaktan da çok uzak değildir. Özetle, tezde, dilbilimin temel araştırma sorularına bir yanıt aranmamakla birlikte HD çalışmalarından da yararlanılmış; DDİ'de tanımlı problem (sözdizimsel belirsizlik giderme) için, belirlenen kısıtlar altında, olabildiğince verimli ve sade çözümler önerilmeye çalışılmıştır.

DDİ en genel tanımıyla, insan dillerinin matematiksel olarak modellenmesini, bilgisayar ortamında temsil edilebilmesini ve böylece dil ögelerinin makineler tarafından anlaşılabilmesini sağlayacak amaçlar ve etkinlikler bütünüdür. DDİ bilgisayar bilimleri, yapay zekâ, dilbilim, matematik, istatistik, psikoloji, felsefe vb. pek çok bilim dalı, çalışma alanı ve disiplinin birleşmesiyle vücut bulur. DDİ'nin temel problemleri, ses sinyalleri ile yazıyı uygun şekilde eşleştirmek, sözcüklerin sınırlarını belirlemek ve iç yapıları ile onların işlevleri arasındaki ilişkileri kavramak, tümcelerin yapılarını çözümlemek ve sözcük, öbek veya tümce düzeyindeki belirsizlikleri gidermek, ses, biçim, yapı ve anlam arasındaki bağlantıyı kurmak ve alana özgü (domain-specific) ya da geniş çaplı anlam belirsizliğini gidermektir. Bu amaçları gerçekleştirmek için kuramsal ve uygulamalı dilbilimin birçok alanıyla disiplinler arası çalışmalar yapılmaktadır.

DDİ'nin temel konu başlıkları şunlardır: doğal dil anlama, makine çevirisi, soru yanıtlama, doğal dil üretme, sözcük anlam belirsizliği giderme, varlık ismi tanıma, söylem çözümlemesi, duygu çözümlemesi, sözdizimsel çözümleme, öbek belirleme, biçimbilimsel çözümleme, sözcük türü etiketleme, tümce sonu belirleme, sözcük ayrıştırma, optik karakter tanıma, konuşma tanıma, metinden konuşma sentezleme…

İstatistiksel dil modellemesi, sözdizimsel çözümleme ve makine çevirisi gibi pek çok DDİ konusu bakımından Türkçe eşsiz problemler sunmaktadır (Oflazer, 2014). Birçok çalışmada (Emekligil vd. 2016, Logacev vd. 2014, Sak vd. 2010, Parlak ve Saraçlar 2009, Arısoy ve Arslan 2005) Türkçenin "meydan okuyucu" bir dil olduğu vurgulanır. Sözü edilen zorluk Türkçenin karakteristik özelliklerinin bir sonucu olup

bunun yanı sıra, kaynak eksikliği ve Türkçe DDİ çalışmalarının görece geç başlamış olması da önemli birer etkendir.

3. SÖZDİZİMSEL ÇÖZÜMLEYİCİ

Bu bölümde tezin ana ürünü olan yazılım kütüphanesi ve onun en önemli parçası olan sözdizimsel çözümleyici tanıtılacaktır. Bu çözümleyici, tez kapsamında belirlenen dilbilgisi kurallarını işleterek, verilen Türkçe tümceler için yapısal çözümlemeler üreten bir dizgedir ve özgün bir çalışmanın sonucudur.

Sözdizimsel çözümleyici (SÇ, syntactic parser), en genel tanımıyla, verilen bir sözcük dizisinden dilbilgisi örüntülerini çıkarmaya yarayan bir dizgedir. Çıktı olarak sözcük dizisinin yapısal karşılıkları olan yazılı ya da çizge gösterimlerini üretir. SÇ'yi tanıyıcıdan (recognizer) ayırmak gerekir: tanıyıcı girdinin dilbilgisel olup olmadığını belirler, ancak yapısal gösterimleri üretemez. Ayrıca SÇ girdi üzerinde yalnızca yapısal çözümleme gerçekleştirir; ögelerin anlamlarını dikkate almaz (Sanders ve Sanders, 1989, s. 14).

Özellikle çekim eklerinin sözdizimde önemli rol oynadığı diller için SÇ, çekimsel biçimbilimden de etkilenir. Bu bağlamda, Türkçe için geliştirilen bir SÇ'de biçimbilimsel tanımlamaların, kuralların ve bağlantıların bulunması gerektiği dikkate alınmalıdır. Öyleyse, SÇ'nin biçimbilimsel çözümleyici ve onun bileşenlerini de içermesi uygun olacaktır. Aslen böyle bir dizgenin adı tam olarak biçim-sözdizimsel çözümleyicidir (BSÇ, morpho-syntactic parser). Ancak Türkçede sözdizimsel ilişkilerin doğal olarak kısmi biçimbilimsel dinamikleri de içereceği hatırlanırsa kısaca SÇ terimi kullanılabilir.

3.1. Önceki Çalışmalar

Türkçe sözdizimsel çözümleme çalışmalarının ilk örneklerinden biri olan ve Güngördü (1993) tarafından gerçekleştirilen çalışma, sözlüksel-işlevsel dilbilgisi (lexical functional grammar) esasına dayanmaktadır. Kısıt tabanlı bir dilbilgisi çerçevesi olan sözlüksel-işlevsel dilbilgisi, dil ögelerini bileşen yapısı ve işlevsel yapı şeklinde iki koşut düzeyde ele alır. Söz konusu çalışmada serbest sözcük dizilimi problemiyle başa çıkan ve hem yalın hem de karmaşık tümce yapılarını içine alan dilbilgisi kuralları geliştirilmiştir. Ayrıca biçimbilimsel yapıların çözümlenmesi için iki

düzeyli biçimbilimsel çözümleyiciden yararlanılan çalışmada, ele alınan tümcelerin % 82'sinin çözümlenebildiği belirtilmiştir.

Eryiğit vd.'nin (2008) çalışması bağımlılık ayrıştırma paradigmasını kullanarak Türkçe tümcelerin çözümlenmesi üzerine gerçekleştirilen inceleme ve deneyleri sunmaktadır. Çalışmada *IG* adı verilen yapıların kullanıldığı modellerin sözcük tabanlı modellere kıyasla daha başarılı olduğu rapor edilmektedir. *IG* "inflectional group"un kısaltması olup iki türetim durağı arasındaki biçimbilimsel özellikler kümesi anlamına gelmektedir (Oflazer, 2014). Sözcük ile biçimbirim arasında konumlanan yapay bir birimdir.

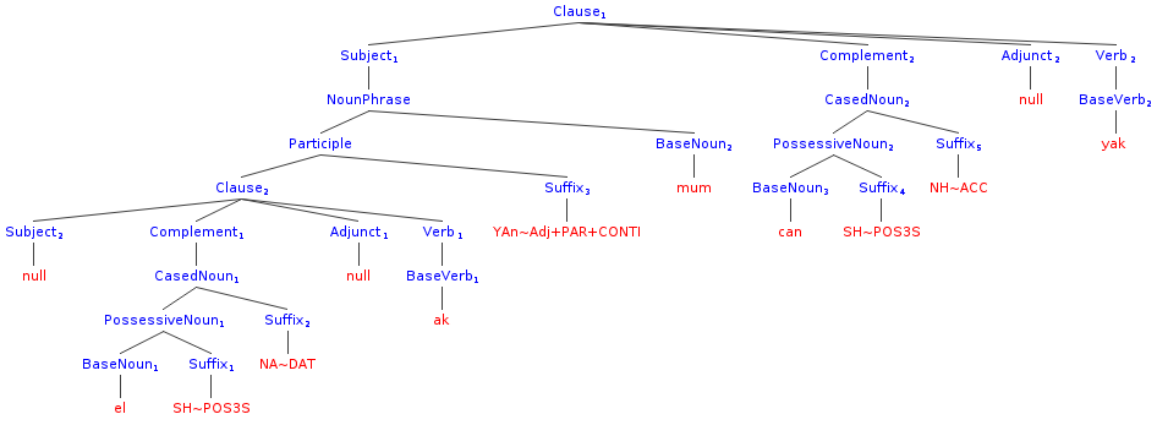
Eryiğit (2014) tarafından tanıtılan ardışık dizge (pipeline), ham metinden sözdizimsel düzeye kadar birçok aşamadan oluşmaktadır. Bu aşamalar şunlardır: simgeleştirme, Türkçe karakter düzeltme, ünlüleştirme, yazım düzeltme, normalleştirme, dil tanıma, biçimbilimsel çözümleme, biçimbilimsel belirsizlik giderme, varlık adı tanıma, bağlılık ayrıştırma. Dizgede sözdizimsel çözümleme bağımlılık ayrıştırma aşamasında gerçekleştirilmekte olup burada girdi, biçimbilimsel belirsizliği giderilmiş bir tümce ve özellikleri iken; çıktı, girdi üzerinde bağımlılık etiketlerinin ve bağlantı numaralarının eklenmiş biçimidir.

Bildiğimiz kadarıyla Türkçe için öbek yapı dilbilgisi temelli ve geliştirilmeye açık bir sözdizimsel çözümleyici bulunmamaktadır. Bu alandaki eğilim özellikle bağımlılık dilbilgisi yönündedir. Hizmete sunulmuş en kapsamlı çözümleyici[6] Eryiğit'in (2014) ardışık dizgesinin bir parçası olup bu çözümleyiciye ancak Web API yardımıyla erişilebilmektedir.

3.2. Önerilen Dizge

Önerilen SÇ, temel olarak öbek yapı dilbilgisinden faydalanıyor olsa da özne, tümleç, eklenti vb. rolleri tanımlamaya izin vermesi itibarıyla bağımlılık dilbilgisinin sunduğu ilişkiselliğe sahiptir. Bu anlamda Türkçe için uygun bir gösterim olduğunu düşündüğümüz bir ağaç yapısı öneriyoruz. Bu yapı **Şekil 3.1**'de örneklenmiştir.

[6] http://tools.nlp.itu.edu.tr/DependencyParsing; erişim: 28.06.2017



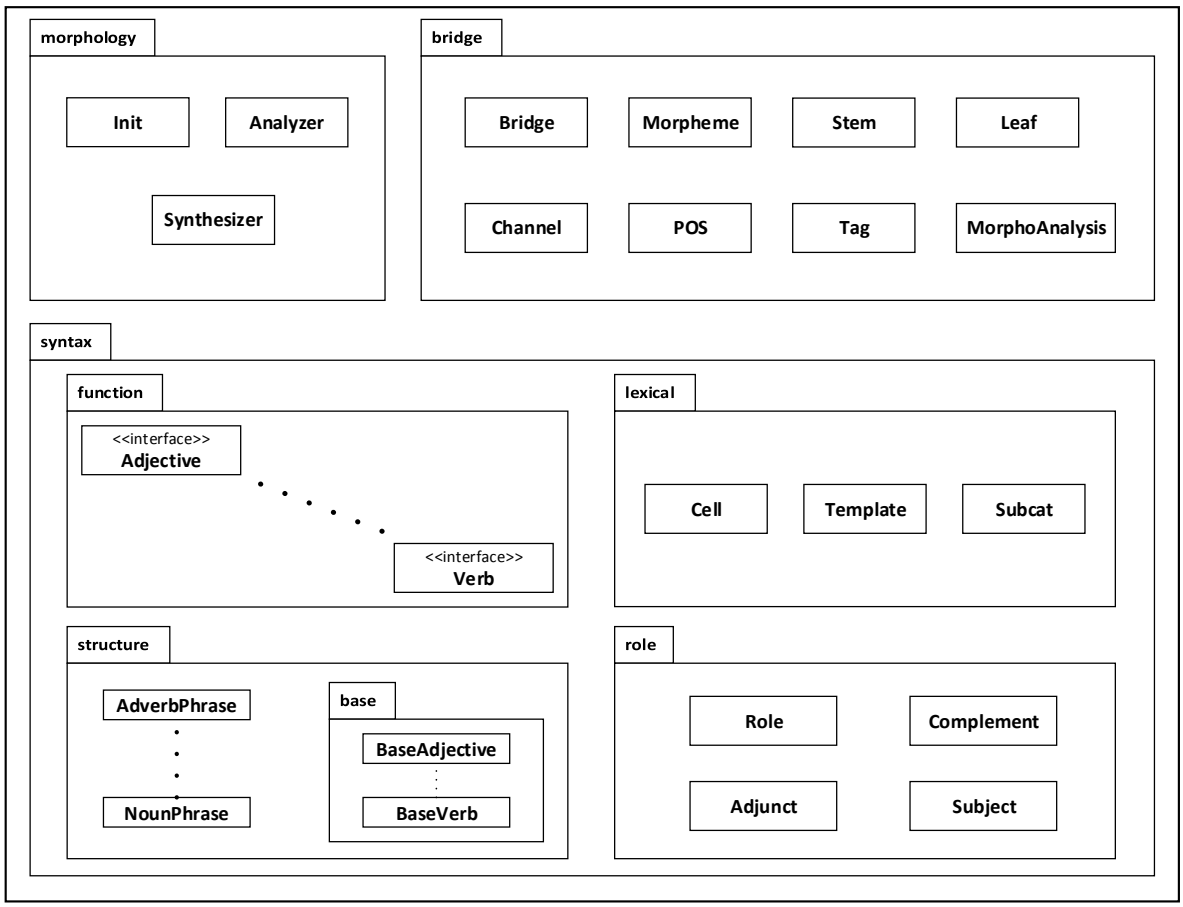
Şekil 3.1. *Tezde Önerilen Sözdizim Ağaç Yapısı*

Şekil 3.1'de verilen sözdizim ağacı "Eline akan mum canını yaktı" tümcesinin yapısal bir yorumudur. Ağaçta <*null*> değeri almamış yapraklar biçimbirimlere karşılık gelir. Biçimbirimler yüzeybiçim formu (örnek: kitap, lar, ı) yerine sözlükbiçim formunda (örnek: kitaP, lAr, SH) gösterilmiştir. Ek tipindeki biçimbirimler sözlükbiçim formunun yanı sıra biçimbilimsel etiketleri de içermektedir. Ekler için gösterim formatı şöyledir: [sözlükbiçim]~[biçimbilimsel etiket(ler)]. Örneğin, ağaçta yer alan *NH~ACC* eki belirtme durum ekini simgelemektedir. *NH* ekin sözlükbiçimi iken, *ACC* accusative anlamına gelen etikettir.

Bu yorumda "eline akan mum" öbeği özne, "canını" sözcüğü tümleç ve "yak" sözcüğü eylemdir. "Eline akan mum" öbeği, içerisinde bir tümcecik (Clause) barındırmaktadır: "eline ak". Alt tümcecikler de kendi içinde özne, tümleç, eklenti, eylem gibi rollere sahiptir. Her bir rol ağaçta sabit biçimde gösterilmekte, eğer boş ise <*null*> ifadesi ile bu durum belirtilmektedir. **Şekil 3.1** için *Clause1* düzeyinde özne rolü dolu iken, *Clause2* düzeyinde özne rolü boştur.

Böyle bir gösterim dolaylı olarak bağımlılık ilişkilerini çıkarmaya izin verebilir. Örneğin **Şekil 3.1**'de "yak" eylemi ile "can" ad tabanı arasındaki eylem-nesne ilişkisi ağaç yapısı tarafından dolaylı olarak işaret edilmektedir. "can" sözcüğünden yukarıya doğru ilerlendiği takdirde *Complement* rol tanımına ulaşılacaktır. Bu da *Clause* üzerinden *Verb* ile bağlantılıdır.

Bu tezde tanıtılan SÇ, Java dilinde oluşturulmuş bir kod kütüphanesinin en önemli parçası olarak tasarlanmıştır. TMoST (Turkish Morpho-Syntax Tool) adı verilen bu kütüphanenin en önemli üç paketi ve içerdiği ögeler özet olarak **Şekil 3.2**'de sunulmuştur.



Şekil 3.2. *TMoST Kütüphanesi Paketleri*

TMoST'u oluşturan paketlerden morphology[7], biçimbilimsel çözümleme işini gerçekleştirir. Bu paketin önemli sınıfları Init, Analyzer ve Synthesizer'dır. Init başlangıç işlemlerini yapar, Analyzer, verilen bir sözcüğün biçimbilimsel çözümlemelerini listeler ve Synthesizer, derin yapıdaki biçimbirim dizisinden yüzeybiçim olan sözcüğü üretir.

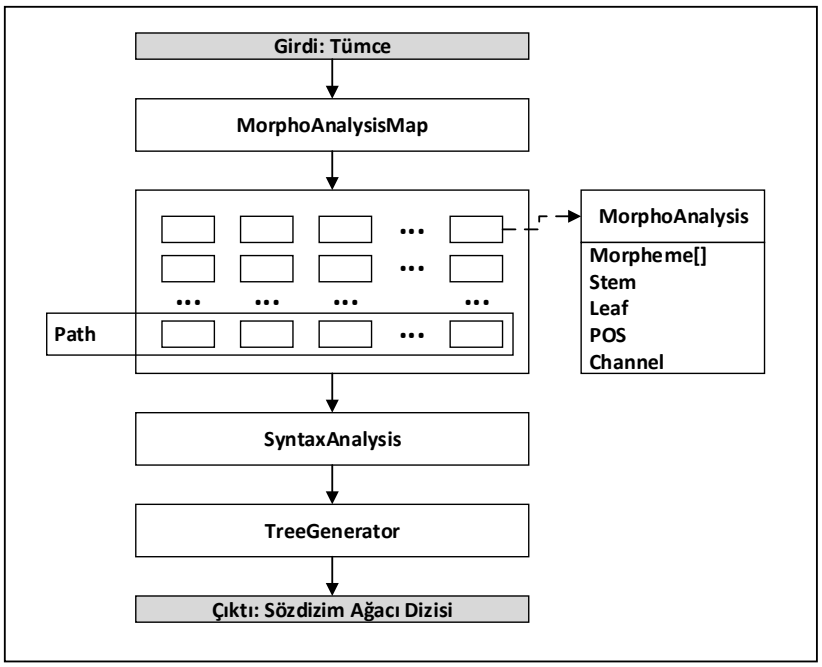
[7] Bu paketin alt birimleri Biçimbilimsel Çözümleme bölümünde açıklanmıştır.

Bridge adlı paket biçimbilim ile sözdizim arasında bir köprü oluşturmak için tasarlanmıştır. Paketle aynı adı taşıyan *Bridge* sınıfı *morphology* paketiyle olan iletişimi kurar ve diğer sınıflardan türetilecek nesneleri oluşturur. Bu sınıf *morphology* paketinde gerçekleştirilen işlemlerin dış dünyadan soyutlanmasını sağlar. *Morpheme* sınıfı biçimbirime karşılık gelir. Bir sözcük biçimbilimsel olarak çözümlendiğinde elde edilen biçimbirimler bu sınıfın nesneleri olarak işlem görür. *Stem* gövdeyi temsil eder ve bir ya da daha çok biçimbirimden oluşabilir. *Leaf* sözcüğe ait çözümlemeden gövde çıkarıldığında geriye kalan biçimbirim ya da biçimbirimlerdir. *Channel* bir çözümleme için mümkün olabilecek temel biçimbilimsel bilgi yuvalarından oluşur. Örneğin çatı, kutup, zaman, iyelik, durum bilgileri bunlardan bazılarıdır. *POS* sözcük türünü simgeler. *POS*, *MorphoAnalysis* ve *Stem* sınıfları için bir özelliktir. *Tag*, biçimbirimlere atanan kısa etiketler için kullanılır. *MorphoAnalysis*, nesneleri bir sözcük için üretilen farklı biçimbilimsel çözümlemeleri yönetmek için kullanılır ve *Bridge* ile birlikte bu paketin en önemli sınıfıdır.

Syntax paketi dört alt paketten oluşur: *function*, *structure*, *lexical* ve *role*. *Function* alt paketi dilbilgisel işlevlere karşılık gelen arayüzleri içerir. *Structure* paketinde alt paket olan *base*'in içinde sıfat, eylem vb. taban yapıları, dışında ise bütün sözdizimsel yapılar bulunur. *Lexical* paketi sözlüksel bir bilgi olan yanulamlama ile ilgili sınıfları kapsar. *Role* paketi ise *structure* paketinde yer alan *Clause* sınıfının bileşenleri olan kurucu üyelerin tümce içi rollerini yönetebilmek için tasarlanmıştır.

3.2.1. Çözümleme adımları

Şekil 3.3, TMoST'ta bir girdi tümcesinden sözdizim ağaçları üretme adımlarını özetlemektedir.

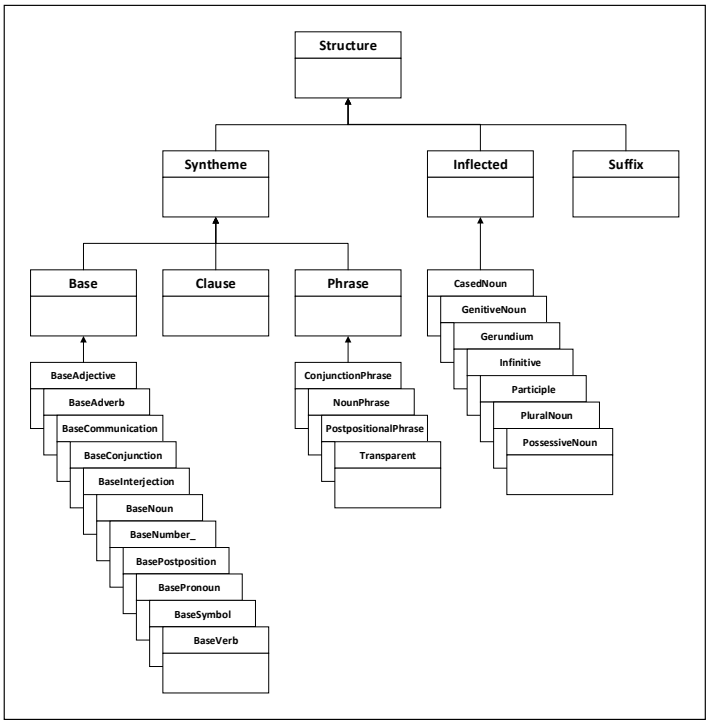


Şekil 3.3. *Girdi Tümceden Sözdizim Ağacı Üretme Adımları*

Şekil 3.3'e göre, ilk olarak, girdi olan tümce *MorphoAnalysisMap* sınıfından türetilen nesneye verilir. Bu nesne tümceyi simgelerine ayırıp her bir simge için biçimbilimsel çözümleme işlemi gerçekleştirerek *MorphoAnalysis* nesneleri üretir. *MorphoAnalysis* sınıfı biçimbilimsel çözümlemeden elde edilen biçimbirim dizileri *Morpheme[]* nesneler dizisinde, gövdeyi oluşturan biçimbirimleri *Stem* nesnesinde, gövdeden arta kalan biçimbirimleri *Leaf* nesnesinde, sözcük türü bilgisini *POS* nesnesinde ve biçimbilimsel özelliklerin tamamını *Channel* nesnesinde saklar. Bir simge için bir ya da daha çok *MorphoAnalysis* nesnesi üretilebilir. Tümce bir simge dizisi olduğu için bu dizi boyunca *MorphoAnalysis* nesneleri farklı yollarla birbirini izleyebilir. Bu yollardan her birine *Path* (patika) adı verilmektedir. *MorphoAnalysisMap* işte bu patikaları üretir. Patikaların sayısı her bir simge için üretilen *MorphoAnalysis* nesnelerinin Kartezyen çarpımıyla hesaplanabilir. *SyntaxAnalysis*, patikaları tek tek *TreeGenerator* nesnesine gönderir ve patika üzerinde dilbilgisi kuralları işletilerek sözdizim ağaçları elde edilir. Bir patika için bir ya da daha çok sözdizim ağacı üretilebileceği gibi, patika kuralları sağlamadığı için hiçbir ağaç da üretilemeyebilir.

3.2.2. Dilbilgisi kuralları

Dilbilgisi kuralları çözümleyici içinde tanımlanmış yapı nesneleri ile temsil edilmektedir. Bunlar bağlam bağımsız dilbilgisi kuralları kadar yalın olmayıp, nesne tabanlı programlamanın verdiği imkânlar ölçüsünde çeşitli bilgi ve çerçevelerle donatılmış görece karmaşık kurallardır. Burada bir kural bir yapı nesnesinin yaratılması ile gerçeklenir. Nesnenin yaratılamaması kuralın işletilemediği anlamına gelmektedir. **Şekil 3.4**'te yapı nesnelerinin türetildiği sınıflar ve aralarındaki hiyerarşi görülmektedir.



Şekil 3.4. *Yapı Sınıfları Hiyerarşisi*

Şekil 3.4'te verilen hiyerarşide görüldüğü gibi bütün sınıflar *Structure*'dan türemektedir. *Structure* üç alt sınıfa ayrılır: *Syntheme*, *Inflected* ve *Suffix*. *Suffix* yapısı basitçe ekleri temsil eder. Bunlar yalnızca çekim ekleri olabilir. Türetim ekleri *Base* tipindeki yapıların içinde gömülü durumdadır. *Inflected* yapısı ise çekimlenmiş yapıları temsil etmektedir. *Inflected* tipindeki yapılar başka bir *Inflected* tipindeki yapının içinde yer alabilir. Hangi yapıların hangi yapıları içerebileceğini dilbilgisi kuralları

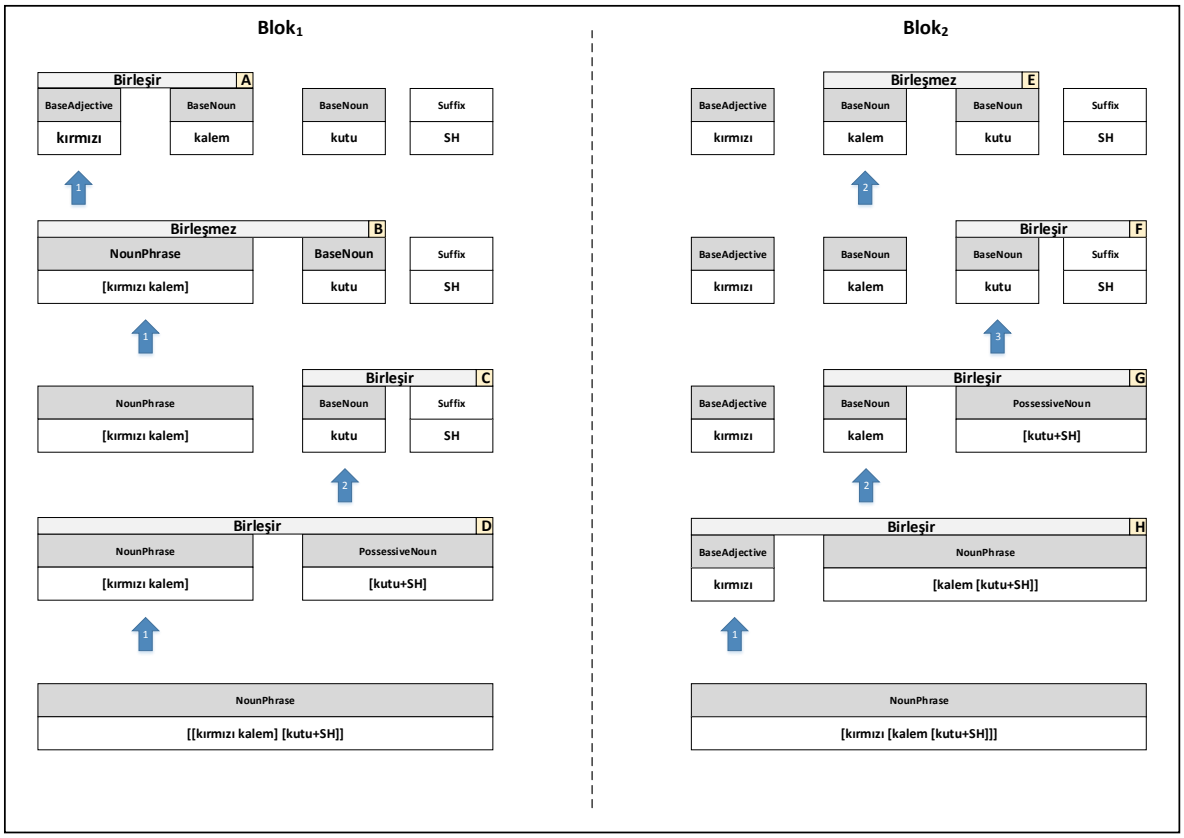
belirlemektedir. *Syntheme* sözdizimin temel birimi olan yapıları temsil eden soyut bir tanımlama olup TMoST'un çalışma ilkesini en kısa biçimde açıklayan ve tez açısından büyük bir öneme sahip olan yapıdır. *Syntheme* için Türkçe bir karşılık öneriyoruz: dizimbirim. Dizimbirim kavramı sayesinde bir sözcük tabanıyla (Base) bir öbeği (Phrase) denk biçimde işleyebiliriz. Örneğin, *Inflected* tipinde bir yapı kısaca şöyle tanımlanabilir: *Syntheme+Suffix*. *Inflected* tipinde bir yapı olarak *CasedNoun* yapısını ele alalım. *Syntheme* sayesinde *kitaP* (BaseNoun) ile *ders kitabı* (NounPhrase) yapıları eşbiçimde işlenerek sözgelimi yönelme durum eki ile birleşebilir: *kitaba* ile *ders kitabına*.

Kurallar belli bir anda yalnızca bir *MorphoAnalysis* patikası üzerinde çalışmaktadır. Bu patika bir ya da daha çok *MorphoAnalysis* nesnesinden oluşur. Her bir *MorphoAnalysis* nesnesi *Stem* ve *Leaf* içerik nesnelerine sahiptir. *Stem* nesnesi kullanılarak *Base* yapı nesnesi ve *Leaf* nesnesi kullanılarak da bir ya da daha çok *Suffix* yapı nesnesi üretilir. Böylece *MorphoAnalysis* patikası yapı nesnelerinden oluşan bir diziye dönüşür.

Yapı nesneleri dizisi dilbilgisi kuralları tarafından işlenmeden önce birkaç önişlemden geçer. Bu önişlemler sırayla, ikilemelere özgü ekler denetimi, *Clause* yapı nesnesi (eylem öbeği) oluşturma, eylem ile eylem çekim eklerini birleştirme, çoğul ekini sol tarafındaki ad birimiyle birleştirme şeklindedir. Önişlemler yapılmadığı takdirde çok sayıda ağaç oluşturma yolu ortaya çıkmakta ve çözümleme zamanı artmaktadır.

Dilbilgisi kuralları, yapı nesneleri dizisi üzerinde yalnızca ikili birleşimlere izin vermektedir. Her bir nesne bağımlı olma (Dependent adlı arayüz yardımıyla) ve olmama şeklinde tanımlanabilmektedir. Buna göre yalnızca *Suffix* yapı nesneleri bağımlıdır. Nesneler ikili biçimde incelenirken bazı kurallar bağımlı olmayan yapı ile bağımlı olan yapıyı birleştirme üzerinde karar verir. Buna örnek olarak şu birleşim verilebilir: *okul (BaseNoun) + NDA~LOC (Suffix)*. Bazı kurallar ise bağımlı olmayan yapı ile yine bağımlı olmayan yapıyı birleştirme üzerinde bir karar verir. Buna örnek olarak şu birleşim verilebilir: *kırmızı (BaseAdjective) + balıK (BaseNoun)*. Burada birinci tür kurallar çekimlemeye ilişkin dilbilgisi kuralları iken, ikinci tür kurallar türetim ve diğer

birleşimleri tanımlayan kurallardır. **Şekil 3.5**'te dilbilgisi kurallarının işleyişi örneklenmiştir:



Şekil 3.5. *Dilbilgisi Kurallarının İşleyişi*

Şekil 3.5'te "kırmızı kalem kutusu" tamlamasının kurallar tarafından nasıl işlendiği ve blok mekanizmasının çalışma biçimi örneklenmiştir. TMoST mevcut durumu itibarıyla bitimli olmayan (yüklemsiz) tümceleri tanımamaktadır; bu örnekte ise basitçe incelenebilmesi için bir tamlamanın çözümlenme aşamaları gösterilmiştir. Öncelikle belirtmek gerekir ki **Şekil 3.5**'te sıralanan işlemlerin tümü **Şekil 3.3**'te verilen *TreeGenerator* sınıfı içinde gerçekleşmektedir. *TreeGenerator* belli bir anda tek bir patikayı işleyebilir. Hatırlanacağı üzere, bir patika, tümceden elde edilebilecek biçimbilimsel çözümleme dizilerinden her biridir.

Yapı nesneleri ikişer ikişer incelenmektedir. Dilbilgisi kuralları işaretçinin (numaralı ok işaretleri) gösterdiği yapı ile bir sağdaki yapının birleşip

birleşmeyeceğine ilişkin bir karar verir. Etkin patika için başlangıçta yalnızca bir blok vardır. İki bağımlı olmayan yapı için "birleşir" kararı verildiğinde yeni bir blok oluşur ve bu blokta ilgili yapıların birleşmediği senaryo gerçeklenir. $Blok_2$ A kararının verilmesiyle oluşmuş ve işaretçi 2. konumdan başlamıştır. Burada belirtilmesi gereken bir başka husus da blokların eşzamanlı olarak çalıştırılabileceğidir. Ancak TMoST'un güncel sürümünde paralel programlama uygulanmamıştır.

Yapılar için ikişerli birleşme-birleşmeme kararları vermeye dayalı olan bu dizge kolaylıkla özyinelemeli biçimde de tasarlanabilirdi. Bundan özel olarak kaçınılmıştır. Çünkü özyinelemeli algoritma daha kısa kodlanmaya elverişli olmasına rağmen anlaşılması ve güncellenmesi daha zor olabilmektedir. Bunun yanı sıra mevcut dizgenin, ileri sürümlerde eklenebilecek paralel programlamaya daha iyi uyum sağlayacağı öngörülmektedir.

4. BİÇİMBİLİMSEL ÇÖZÜMLEYİCİ

Bu bölümde TMoST'un bir alt dizgesi olan ve tez çerçevesinde önerilen, özgün biçimbilimsel çözümleyici tanıtılacaktır.

Biçimbilimsel çözümleme basitçe, sözcükleri oluşturan biçimbirimlerin elde edilmesi işi olarak tanımlanabilir. Biçimbirimleri belirlemek; sözcük türü belirleme, gövdeleme, sözcük anlam belirsizliği giderme, sözdizim çözümlemesi gibi birçok süreç için önemlidir. Biçimbilimsel çözümleme sonucunda biçimbilimsel belirsizlik adı verilen önemli bir problemle karşılaşılır. Biçimbilimsel belirsizlik, kısaca, bir sözcüğün farklı biçimbilimsel yorumlara sahip olması şeklinde tanımlanabilir. Bu durum Türkçede çok yaygındır. Biçimbilimsel çözümleyicinin (BÇ, morphological analyzer) birincil görevi, sözü edilen belirsizliği gidermek değil, biçimbirimleri doğru ve tam olarak belirlemek ve olası bütün çözümlemeleri sunmaktır.

Biçimbilimsel çözümleme eklemeli dillerde ek dizgesinin modellenmesi ve kök sözlüğünün oluşturulması gibi ayrıntılı çalışmalar gerektirmektedir. Türkçe ek dizgesinde istisnalar az olmasına rağmen ek ardışıklıklarının kurallandırılması kapsamlı bir çalışma konusudur. Morfotaktik olarak adlandırılan bu ardışıklıklar üç önemli noktaya etki eder:

1. kökleri, dolayısıyla sözlüğün büyüklüğünü belirler.
2. gövdeye dâhil olmayacak çekim eklerini ve gövdede yer alması gereken türetim eklerini tanımlar.
3. biçimbilimsel çözümleme sonucundaki belirsizliği etkiler.

Türkçede bir sözcük, bağlamından bağımsız olarak düşünüldüğünde genellikle birden çok olası biçimbirim dizisi üretir. Bu dizilerden biri, sözcüğün içinde yer aldığı bağlamda geçerlidir. Biçimbilimsel belirsizlik burada sözcük başına düşen çözümleme sayısı olarak yorumlanabilir. Tanımlanan ek sayısı arttıkça ve ek modeli karmaşık hâle geldikçe belirsizliğin artması beklenir.

Bu bölümde kural tabanlı ve üretici sesbilim modeline dayanan bir Türkçe BÇ tanıtılmıştır. Bu çözümleyici bir önceki bölümde tanıtılan TMoST kütüphanesinin bir

parçasıdır. TMoST bünyesine girmeden önce çözümleyiciye Morfolog adı verilmiştir ve sonraki bölümlerde bu adla anılacaktır.

4.1. Önceki Çalışmalar

Biçimbilimsel çözümleme konusunda kural tabanlı ilk genel (dilden bağımsız) yaklaşım, Koskenniemi (1983) tarafından önerilen iki seviyeli biçimbilim modelidir. Bu model üretici sesbilimin mirasını ileriye götürerek, özellikle karmaşık biçimbilimsel yapıdaki diller için etkili bir çözüm sunmuştur. Üretici sesbilim ise Chomsky ve Halle (1968) tarafından geliştirilmiştir. Üretici sesbilimde dildeki ses etkileşimleri, sözlük seviyesindeki sembolleri yüzey seviyedeki sembollere dönüştüren yeniden yazım kurallarıyla açıklanır. Üretim terimi bu dönüşümü ifade etmek amacıyla kullanılmaktadır. Üretim süreci her zaman sözlük seviyesinden yüzey seviyeye doğru tek yönlü olarak gerçekleşir. Yeniden yazım kuralları sözlükbiçimleri aşama aşama yüzeybiçimlere çeviren sıralı bir kural listesidir. Her bir kural, girdi üzerinde değişiklik yapar ve ardından gelen kural da girdinin değişmiş hâli üzerinde çalışır. Yeniden yazım kuralları, girdi üzerinde seri biçimde iş görür ve aynı anda başka kuralların erişimine izin vermeyen ara seviyeler oluşturur. İki seviyeli modelde ise kurallar paralel olarak işlenir; böylece kuralların sıralaması çözümlemeyi etkilemez.

İki seviyeli modelin gerçekleştirimi için PC-KIMMO adlı bir program geliştirilmiştir. PC-KIMMO tabanlı ilk Türkçe biçimbilimsel çözümleyici ise Oflazer (1994) tarafından hazırlanmıştır. Bu çözümleyici 22 adet iki seviyeli kural ve 23.000 girdilik bir sözlükten (lexicon) oluşur.

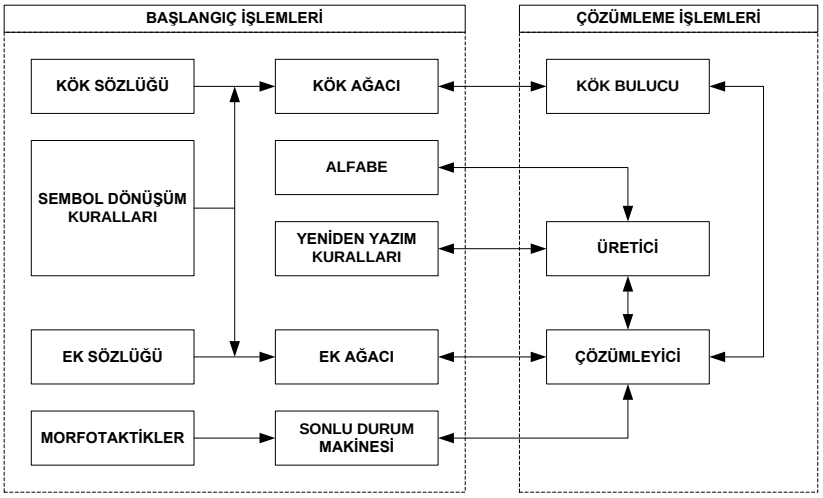
Kural tabanlı bir başka çözümleyici, Akın ve Akın (2007) tarafından geliştirilen Zemberek adlı sistemdir. Zemberek, biçimbilimsel çözümlemenin yanı sıra yazım denetimi, sözcük türetimi, sözcük önerme ve ASCII karakter Türkçeleştirme gibi işleri de kapsayan açık kaynaklı bir kütüphanedir. Özellikle Türk dilleri için tasarlanmış olan Zemberek, OpenOffice yazılım eklentisi şeklinde ve Pardus işletim sisteminde yazım denetimi amaçlı olarak kullanılmıştır.

İki seviyeli modele dayanan diğer bir çözümleyici TRmorph'tur (Çöltekin, 2010). Stuttgart finite state transducer tools (SFST) aracıyla geliştirilen bu çözümleyici General Public License (GPL) ile kullanıma ve geliştirime açık olarak sunulmuştur. Kök sözlüğü Zemberek yazım denetleyicisinden uyarlanmış ve 37.101 kök içermektedir. Kökler için bağlaç, edat, fiil, isim, özel isim, sıfat, ünlem, zamir ve zarf şeklinde dokuz sözcük türü tanımlanmıştır.

Cebiroğlu ve Adalı (2002) tarafından yapılan çalışmada Türkçe ek sistemini modelleyen bir sonlu durum makinesi ile sözcükleri sondan başa doğru çözümleyen (sağdan ek atarak) ve sözlüksüz köke ulaşmayı sağlayan bir sistem ortaya konmuştur. Bu yaklaşımla gerçekte biçimbilimsel çözümleme yapılmakta ve bunun yan ürünü olarak gövde elde edilmektedir. Biçimbilimsel çözümleme ile dolaylı olarak ilişkili olan başka bir çalışma ise gövdeleme algoritmaları yazmak için özel olarak tasarlanan Snowball diliyle (Porter, 2001) gerçeklenmiş Türkçe gövdeleyicidir (Kapusuz, 2006). Bu gövdeleyicide Cebiroğlu ve Adalı'nın (2002) çalışmasında olduğu gibi sağdan ek atarak sözlüksüz gövdeleme tekniği kullanılmıştır.

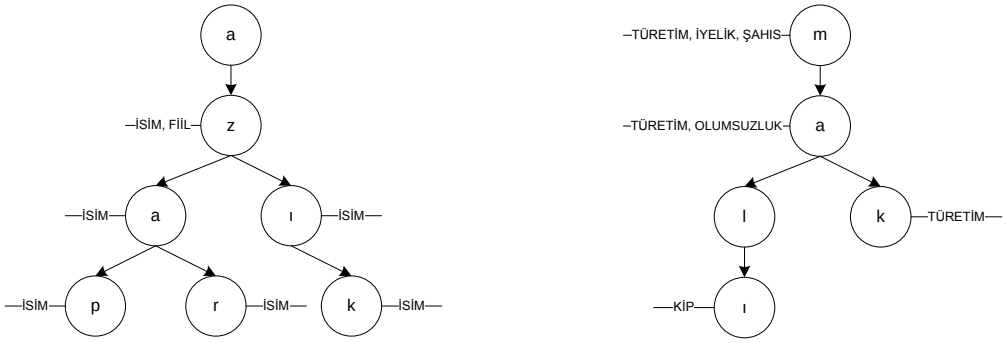
4.2. Önerilen Dizge

Şekil 4.1'de Morfolog'un genel işleyişi verilmiştir:



Şekil 4.1. *Önerilen Biçimbilimsel Çözümleyicinin Genel İşleyişi*

Şekil 4.1'de görüldüğü gibi bu alt dizge, başlangıç işlemleri ve çözümleme işlemleri şeklinde iki temel kısımda incelenebilir. Başlangıç işlemlerinde kök sözlüğü kök ağacını, ek sözlüğü de ek ağacını oluşturmak için kullanılır. Burada çözümleme sırasında harf birliklerinin daha hızlı bir şekilde kök ve eklerle eşleştirilebilmesi amacıyla ağaç veri modelinden faydalanılır. Ağaç veri modelinde kök ve ekleri oluşturan semboller bir ağaç üzerine yerleştirilir ve kök veya ekler başlangıçtan itibaren aynı sembol dizisini paylaşıp daha sonra farklılaşabilir. Bu şekilde oluşacak ortak sembol dizileri bir kere belirtileceği için bellek açısından, kök/ek sayısına değil, aranan sembol dizisinin uzunluğuna bağlı olduğu için de hesaplama zamanı açısından büyük bir avantaj sağlamaktadır. **Şekil 4.2**'de bu çözümleyiciye ilişkin ağaç veri yapısı görülmektedir.



(a) *Kök Ağacı* **(b)** *Ek Ağacı*

Şekil 4.2. *Ağaç Veri Yapısı*

Şekil 4.2'de görüldüğü gibi, kök ve eklerin sözlükte sahip oldukları sözlüksel biçimler ağaç yapısına aktarılırken yüzey biçimlere dönüştürülür. Sözü edilen dönüşümü gerçekleştirmek için sembol dönüşüm kuralları kullanılır. Bu aşamada bir kök/ek için yüzeyde gözlenebilecek tüm şekiller üretilir. **Tablo 4.1**'de sözlük seviyesindeki kökleri yüzeybiçimlere dönüştüren sembol dönüşüm kuralları verilmektedir.

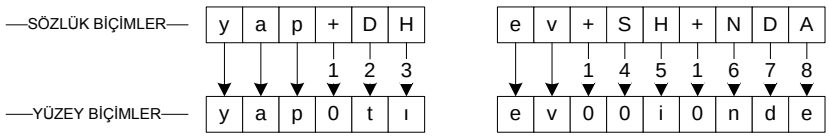
Tablo 4.1. *Kök Sembol Dönüşüm Kuralları*

	Kural	Örnek		Kural	Örnek		Kural	Örnek
1	_T:td	giT: git, gid	7	_Y:y0	suY: suy, su	13	öü_E:eü	döşE: döşe, döşü
2	_P:pb	kaP: kap, kab	8	_%:0	cami%: cami	14	*_E:ei	dE: de, di
3	_K:kğ	göK: gök, göğ	9	_&:0	denizatı&: denizatı	15	_0:0-	hak0: hak, hakk
4	_Ğ:ğg	biyoloĞ: biyoloğ, biyolog	10	aı_E:aı	adE: ada, adı	16	_$:|>	ak$ıl: akıl, akl
5	_Ç:çc	ağaÇ: ağaç, ağac	11	ei_E:ei	beklE: bekle, bekli	17	_/:0	bak/: bak
6	_G:gk	renG: reng, renk	12	ou_E:au	boyE: boya, boyu			

Kök sözlüğü, çözümleyicinin dilbilgisel doğruluğunu önemli derecede etkileyen bir bileşendir. Sözlüğün kapsamını ise morfotaktikler belirler. Örneğin morfotaktikler türetim eklerinin tamamını içeriyorsa sözlük daralır. Başka bir deyişle morfotaktikler karmaşıklaştıkça sözlük sadeleşir. Morfolog'da tanımlanan morfotaktikler çekim ekleri ve işlek türetim eklerini içermektedir. Güncel sürüm için, ek sözlüğünde 110 ek bulunmaktadır. Kök sözlük boyutu ise (özel adlar hariç) 20.000 dolaylarındadır.

Türkçede türetim eklerinin tamamını içeren bir ek modeli ortaya koymak oldukça zordur. Bunun başlıca nedeni eklerin aynı işlekliğe sahip olmamasıdır. Örneğin addan ad türeten "lık" eki hemen hemen her tür ad köküne getirilebilir iken, addan ad türeten "man" eki şiş-man, koca-man gibi birkaç sözcükte görülür. Dizgenin bu tip sözcükleri mutlaka işlemesi isteniyorsa istisnai durumlar morfotaktiklerle ifade edilmemeli, uygun çözümleme sözlükte durağan biçimde tanımlanmalıdır.

Yeniden yazım kuralları, kök ve eklerin sözlük biçimlerini yüzey biçimlere çevirir. Bu kurallarla, modellenen dilin sesbilimsel özellikleri ve biçimbilim bağlamında sözcük içi kök-ek ve ek-ek etkileşimleri dikkate alınır. Yeniden yazım kuralları dizgede üretici bileşenini besler; üretici bileşeni de verilen biçimbirim dizilerini sentezleyerek yüzey biçimleri üretir. **Şekil 4.3**'te yeniden yazım kuralları ile yüzey biçimlerin üretimi örneklenmiştir:



Şekil 4.3. *Yeniden Yazım Kuralları ve Üretim*

Şekil 4.3'te verilen birinci örnekte, kural-1 ile (+:0) sözlük seviyesinde biçimbirimleri ayıran + sembolü yüzeyde 0 sembolüne dönüşür; başka bir deyişle kaybolur. Kural-2 ile (D:t S_) D sembolü kendinden önce bir sert ünsüz bulunması şartıyla t sembolüne dönüşür. Kural-3 ile (H:ı |KZ|_) H sembolü kalın ve düz bir ünlüden sonra ı sembolüne dönüşür. İkinci örnekte kural-4 ile (S:0 C_) S sembolü bir ünsüzden (consonant) sonra geliyorsa 0 sembolüne dönüşür. Kural-5 ile (H:i |İZ|_) H sembolü ince ve düz bir ünlüden sonra i sembolüne dönüşür. Kural-6 ile (N:n (SH)0_) N sembolü SH (iyelik) ekinden sonra geliyorsa n sembolüne dönüşür. Bu kural ile bir istisna tanımlanmaktadır. Kural-7 ile (D:d) eğer kural-2'nin şartı sağlanmamışsa D -> d dönüşümü gerçekleşir. Kural-7 yeniden yazım kuralları listesinde kural-2'den sonra gelmelidir. Aksi takdirde herhangi bir şart belirtilmemiş olan kural-7 her zaman çalışır ve her zaman D -> d dönüşümü meydana gelir. Kural-8 ile (A:e İ_) A sembolü kendinden önceki son ünlü ince bir ünlü ise e sembolüne dönüşür.

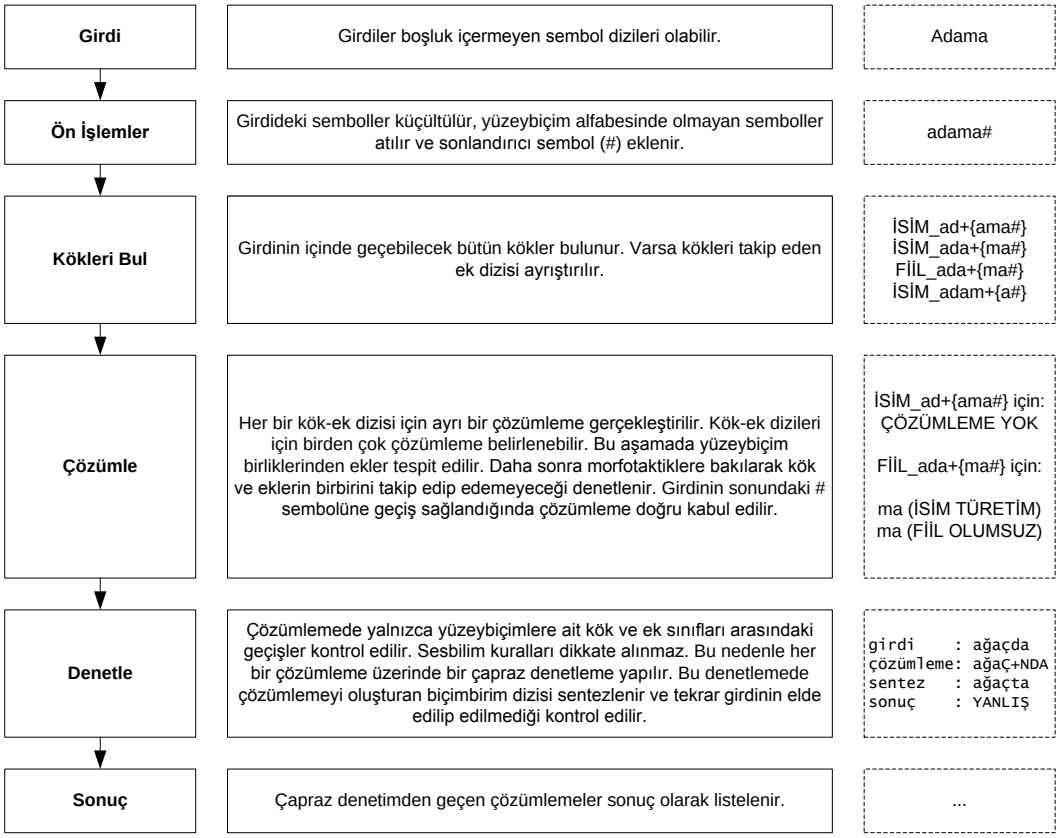
Alfabe, yeniden yazım kurallarında, kök sözlüğünde ve ek sözlüğünde geçen sözlüksel sembollerin ve yüzeyde gözlenen tüm sembollerin özelliklerinin tanımlandığı bölümdür. **Tablo 4.2**'de, alfabede yer alan semboller için tanımlanmış özellikler görülmektedir.

Tablo 4.2. *Sembol Özellikleri*

Sembol	Açıklama
C-V	C: CONSONANT (ÜNSÜZ); V: VOCAL (ÜNLÜ)
S-Y	S: SERT; Y: YUMUŞAK
K-İ	K: KALIN; İ: İNCE
G-D	G: GENİŞ; D: DAR
Z-R	Z: DÜZ; R: YUVARLAK
0-1	0: YÜZEYDE SIFIR OLABİLİR; 1: YÜZEYDE SIFIR OLMAZ
2-3	2: HARF; 3: İŞARET
4-5	4: SABİT SEMBOL; 5: DEĞİŞKEN SEMBOL

A-B	A: A GRUBU FİİLLER (BAK-AR, YAP-AR, YET-ER); B: B GRUBU FİİLLER (BİL-İR, GEL-İR, OL-UR)

Biçimbilimsel çözümlemede, verilen bir sözcük üzerinde sırasıyla ön işlemler, kök bulma, çözümleme, denetleme ve sonuç döndürme gibi süreçler takip edilir. Bu süreçler ve yapılan işlemler **Şekil 4.4**'te verilmiştir.

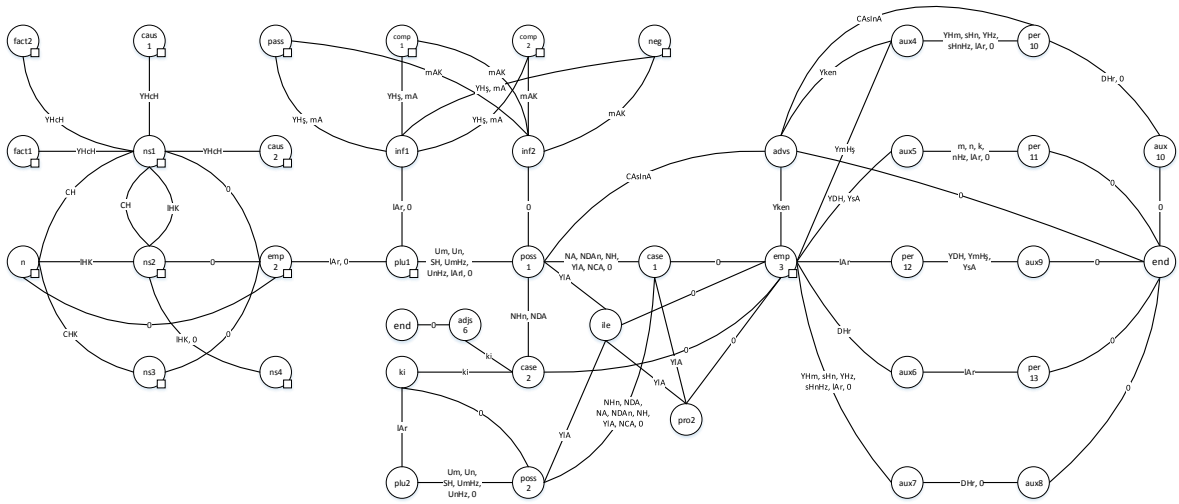


Şekil 4.4. *Biçimbilimsel Çözümleme Algoritması*

Çözümleyici, üretici sesbilim kurallarını denetim sürecinde kullanır. Dolayısıyla üretici sesbilim kurallarına çözümleme sırasında değil, çözümlemeden sonra başvurulmaktadır. Bu sayede, elde edilen biçimbirimler birleştirilerek tekrar girdi sözcüğü üretip üretmedikleri test edilmektedir.

4.2.1. Morfotaktikler

Morfotaktikler Türkçe ek sistemini modelleyen bir şemadır. Sözlükçede tanımlanan sözcük türlerine uygun türetim ve çekimlemeler bu şemada gösterilir. **Şekil 4.5**'te Türkçe ad morfotaktikleri görülmektedir:



Şekil 4.5. *Türkçe Ad Morfotaktikleri*

Şekil 4.5'te örnek olarak verilen Türkçe ad morfotaktikleri *n* (ad kökü) durumuyla başlayıp *end* (son) durumuyla biten ad türetim ve çekimlemelerini göstermektedir. Kök sözlükte *n* durumu ile etiketlenmiş tüm kökler bu ek modelinde işletilebilir. Morfotaktik şemasında daireler durumları, oklar ise bir durumdan diğerine geçiş yapılabilen biçimbirimleri (ekler) simgeler.

Çözümlemede, X durumundan Y durumuna Z ekiyle geçilip geçilmediği denetlenir. Böyle bir denetime, tek bir sözcüğün çözümlenmesinde bile onlarca kez başvurulabilir. Bu nedenle durumlar arasındaki tüm geçişler başlangıç işlemlerinde (bk. **Şekil 4.1**) gerçekleştirilen bir derleme ile sonlu durum makinesi şeklinde ifade edilir. Böylece her defasında morfotaktikler üzerindeki patikalar dolaşılarak arama yapılmasına gerek kalmaz. Sonlu durum makinesini temsil eden veri modelinde bir denetimin hesaplama maliyeti O(1)'dir.

Bir ad kökü bir ek alarak sonlanabilir. Örnek: kitap-lar. **Şekil 4.5**'te *n* durumu (kitap) ile *plu1* durumu (lar) arasında ve *plu1* durumu (lar) ile *end* durumu arasında doğrudan bir bağlantı olmadığı görülmektedir. Modelde doğrudan gösterilmeyen bu bağlantılar "sıfır geçişi" ile sağlanır. Modelde görülen "0" biçimleri ile durumlar birbirlerine sayısız bağlantı kurar. Bu sayede model daha sade bir şekilde sunulabilir.

Morfotaktikler üzerinde daha çok, en olası durumlar betimlenmeye çalışılmıştır. Meydana gelebilecek bazı hatalar sentezleme sırasında filtrelenmektedir. Buna örnek olarak **Şekil 4.5**'te görülen lAr/plu1 ve lArI/pos1 eklerinin art arda gelmesi durumu verilebilir. Morfotaktiklerin bu şekliyle, örneğin "kalemlerleri" gibi bir sözcük doğru şekilde çözümlenir. Ancak sentezleyicide tanımlanan istisnalar yardımıyla, ilgili sözcük hatalı olarak işaretlenecektir. Buna benzer şekilde, eylem kimliği ile ifade edilen eylem ekleşmeleri de sentezleyici tarafından düzenlenir. Eylem kimliği kullanmak yerine ek alırken farklı davranış gösteren eylemler morfotaktik üzerinde her birine birer durum tanımlanacak şekilde ifade edilebilirdi ancak eylemler ekleşme davranışları bakımından 22 sınıfta toplanmaktadır. Bu sınıfları ve her birinin kendi ekleşmelerini şemada tanımlamak morfotaktikleri karmaşık hâle getirecektir. Bu sınıflardan 11 adeti bütün eylemlerin % 95'ini oluşturmaktadır. Çatı özellikleri ve geniş zaman ekine göre başlıca eylem sınıfları **Tablo 4.3**'te verilmiştir.

Tablo 4.3. Çatı Özellikleri ve Geniş Zaman Ekine Göre Başlıca Eylem Sınıfları

İşteş (Uş)	Edilgen	Geniş Zaman	Ettirgen	Dönüşlü (Un)	Yüzde
0	1	1	1	0	20,4
1	1	0	1	1	17,5
0	1	1	0	0	12,7
1	0	1	0	1	12,4
0	1	0	1	0	10,2
0	0	1	0	1	7,3
0	0	1	0	0	6,1
1	1	0	1	0	4,2
0	0	1	1	0	1,9
1	0	0	1	1	1,4
0	1	0	1	1	1,0

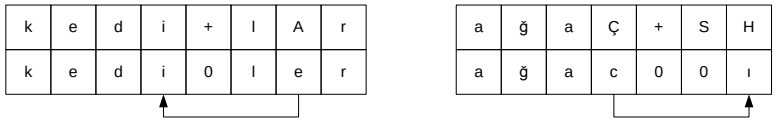
4.2.2. Çözümleme işlemleri

Çözümleyicinin girdi olarak aldığı veriler sözcüklerin yüzeybiçimleridir. Yüzeybiçim, sözlük seviyesinde tanımlanan sözlükbirimler ve eklerin, bazı sesbilimsel ve biçimbilimsel kuralların belirlediği şekilde bir araya gelerek oluşturduğu biçim dizisidir. Biçimbilimsel çözümlemenin amacı yüzeybiçimi, başka bir deyişle sözcüğü oluşturan biçimbirimleri belirlemektir. Bir sözlükbirim ile bir ek dizisinin birleşimi sonucunda (sentezleme) yalnızca bir çeşit yüzeybiçim ortaya çıkabilir. Ancak bir yüzeybiçimin birden çok biçimbirim dizisinden oluşacak şekilde çözümlenmesi mümkündür. Bunun sebebi her bir sözlükbirim veya ekin bir ya da daha çok sembol dizisini tanımlayabilmesidir. Bu sembol dizilerine de biçimlik adını veriyoruz. Örneğin *lAr* eki tanımında *A* üstbiçimi bulunduğundan bu ekin *lar* ve *ler* şeklinde iki biçimliği vardır. Verilen bir yüzeybiçimde *lar* veya *ler* sembol dizileri geçmesi halinde çözümleyici bunların *lAr* ekine ait olması olasılığını dikkate alacaktır. Benzer şekilde bir sözlükbirimin de birden çok biçimliği olabilir. Örneğin, *ağaÇ* sözlükbirimi *Ç* üstbiçimi nedeniyle *ağaç* ve *ağac* şeklinde iki biçimliğe sahiptir. Morfolog'da, sözlükte yer alan sözlükbirimler ve morfotaktiklerde tanımlanmış olan eklerin sahip olduğu biçimlikler indeksleme zamanında (başlangıç işlemlerinde) elde edilmektedir. Sözlükbirimlerin biçimliklerini elde etmek amacıyla bir kural listesi tanımlanmıştır (bk. **Tablo 4.1**). Eklerin biçimlikleri ise sayıca az olmaları ve çok daha karmaşık kurallar gerektirmeleri nedeniyle elle belirlenmiştir.

Morfolog, çözümleme zamanında en çok iki bileşeni kullanır: çözümleyici ve üretici. Çözümleyici bileşeni bir sözcüğü (yüzeybiçim) oluşturabilecek biçimbirim dizilerini belirler. Söz konusu biçimbirim dizileri listelenmeden önce bir araya getirilip sentezlenmeli ve girdiyle harfi harfine aynı sonucu üretip üretmedikleri denetlenmelidir. Bu denetim sırasında üretici bileşenine başvurulur. Bu şekilde, çözümleme, sesbilimsel kurallarla uyumlu hâle getirilir. Sentezleme denetimi gerçekleşmezse, örneğin, "Öğrencilar okule gideyor" tümcesindeki sözcükler doğru kabul edilecektir. Çünkü çözümleme sırasında sembol dizilerini karşılayan sözlükbirim ve ekler birbirinden bağımsız olarak belirlenmekte, ardından bu sözlükbirim ve eklerin

morfotaktiklerde temsil ettiği durumlar arasında bir geçiş olup olmadığına bakılmaktadır. Üretici hataları çözümleme sonuçlanmadan önce belirler. Bunu yaparken, biçimbirimleri, yeniden yazım kurallarını kullanarak onlardan oluşabilecek tek yüzeybiçime dönüştürür.

Olası çözümlemelerin biriktirilerek sentezleyici tarafından denetlenmesi yerine, bir başka yol olarak, çözümleyici, sözcüğün sembolleri üzerinde ilerledikçe elde edilen biçimbirimler anlık olarak sentezlenebilir. Ancak bu yöntem yeniden yazım kurallarının sadece önceki içeriğe bağlı çalışması halinde işe yarar. Oysa Türkçede bazı ses olayları önceye değil sonraya, diğer bir deyişle gerideki değil ilerideki bağlama bağlı gerçekleşir. Örnek vermek gerekirse, büyük ünlü uyumu geriye dönük bir ses olayı iken, ünsüz yumuşaması ileriye dönük bir ses olayıdır. **Şekil 4.6** bu durumu göstermektedir.



Şekil 4.6. *Sembol Dönüşümünde Etkileşim Yönü*

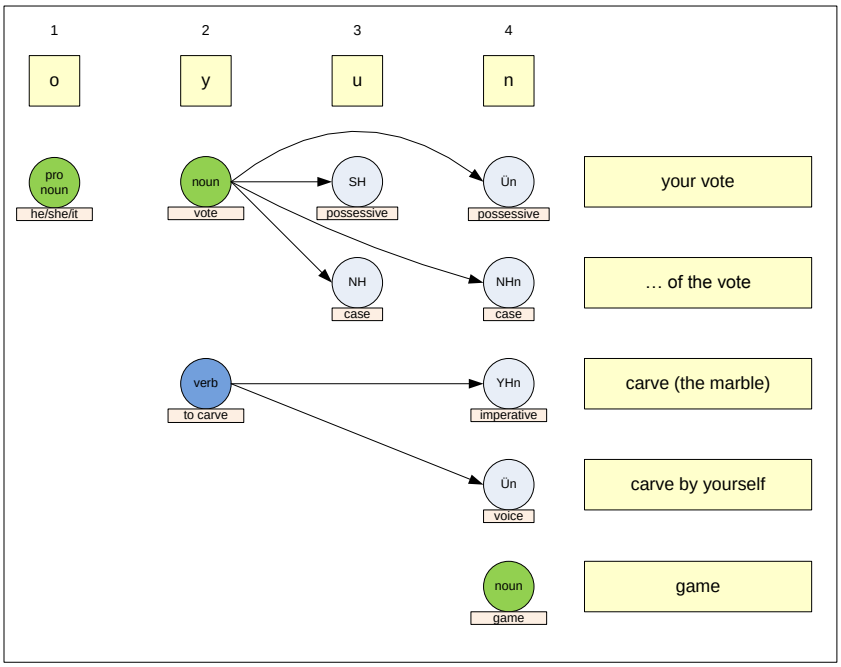
Şekil 4.6'daki okların yönü sözlüksel sembolü yüzeysel sembole dönüştürmek için kullanılan başvurunun ne tarafa yapıldığını gösterir. *A* sembolü kendinden önce en son bir ince ünlü bulunması hâlinde *e* sembolüne, *Ç* sembolü de kendinden sonra bir ünlü bulunması hâlinde *c* sembolüne dönüşür. Burada *+* ve *S* gibi bazı semboller ünlülük, incelik-kalınlık gibi özellikler bakımından yansızdır. Bu nedenle bu tür semboller sonraki veya önceki sembollerin özelliklerini aynen geçiren "saydam" sembollerdir.

Çözümleyici bileşeni (bk. **Şekil 4.1**) girdi olarak bir sözcük alır ve sonuç olarak biçimbirim dizileri döndürür. Herhangi bir biçimbirim dizisi döndürmemesi hâlinde, sözcük Türkçe açısından yanlış kabul edilir. Girdi sözcük işlenmeden önce, alfabede yer almayan semboller silinmiş ve bütün semboller küçültülmüş olmalıdır.

Çözümleyici, sözcük üzerinde sembol sembol ilerler ve bu semboller biriktirilir. İlk önce, biriken semboller sözlükçede aranarak kökler (sözlükbirim) bulunur. Bir kök bulunduğunda o anki sembol üzerinde bir düğüm oluşur. Semboller üzerinde

ilerlemeye devam ettikçe bu düğümün konumu ile o anki konum arasında kalan sembol birliğine ait bir ek olup olmadığı ve ekler belirlendikçe de eklerin konumlarıyla o anki konum arasındaki sembol birliklerine ait bir ek olup olmadığı araştırılır. Özetle, semboller üzerindeki her adımda, daha önce belirlenmiş düğümler üzerinde gezilir. Her bir düğümün bulunduğu nokta ile o an üzerinde bulunulan sembol arasında kalan birlik, kök veya ek listesi üzerinde aranır. Sembol birliğine ait bir sözlüksel tanımlama bulunursa yeni bir düğüm oluşturulur.

Söz gelimi girdi "evdesin" sözcüğü ve etkin sembol *d* olsun. Bu noktada iken *ev* kökü bulunmuş olmalıdır yani *v* sembolü görüldüğü anda *ev* sözlükbirimini işaret eden bir düğüm oluşturulmuştur. Eğer bu sözlükbirimi tanımlayan bilgilerden yararlanılmazsa *d* ve sonraki sembolleri olabilecek bütün ek biçimlikleri dizisinde aramak gerekir. Morfolog'da morfotaktiklerde bulunan 80 ek için 424 farklı biçimlik tanımlanmıştır. Fakat *ev* sözlükbiriminin bir ad olduğu dikkate alınırsa bakılması gereken biçimlik sayısı 186'ya düşer. *ev* ile *de* düğümleri belirlendikten sonra bu liste 61 biçimlikten oluşur. Bu şekilde, semboller üzerinde ilerledikçe morfotaktikler üzerindeki patika seçenekleri, dolayısıyla bakılacak biçimlikler azalacaktır. Yöntem çözümlemeyi kayda değer şekilde hızlandırmaktadır çünkü sembol birliklerinin biçimlikler listesinde aranması bir kelime için onlarca kez yapılan bir işlemdir. Düğümlerin çözümleme sırasında oluşumuna ilişkin bir örnek **Şekil 4.7**'de görülmektedir.



Şekil 4.7. *Düğümlerin Çözümleme Sırasında Oluşumu*

Çözümleyici, semboller üzerinde ilerlerken önceki düğümleri gezmektedir. Eğer o anki sembol konumu ile güncel düğümün konumu arasında kalan birlik, biçimlik listesinde yoksa diğer düğüme geçilmektedir. Örneğin **Şekil 4.7**'deki 4. sembol olan *n* üzerindeyken *SH* ve *NH* düğümlerinden uygun bir geçiş sağlanamamıştır. Bazı durumlarda geçiş yapılamayan düğüm, sonraki semboller tarafından hiçbir zaman kullanılmaz. Bu tür düğümlerin tespit edilip o anda silinmesi gerekir. Biçimlik listesinde aranan sembol birliğini içeren başka bir birlik yoksa güncel düğümden geçilecek bir nokta kalmamış demektir. Bu bilgi biçimlik sorgusundan alınır. Sembol dizileri önceden kendisini içeren başka bir dizi var veya yok şeklinde indekslenmiştir. Örneğin, *insan* sözcüğü üzerinde *s* sembolüne kadar *in* (eylem) düğümü mevcutken (çünkü insin, inse gibi noktalara gitme olasılığı vardır) *a* sembolüne gelindiğinde bu düğüm geçersiz hale gelir ve silinir. Düğüm silme işlemi yapılmasaydı, "insanlarımızdaki" gibi bir sözcükte defalarca *in* (eylem) düğümünden geçiş olup olmadığına bakılmak zorunda kalınabilirdi.

Üretici bileşeni (bk. **Şekil 4.1**), sözlüksel sembollerden oluşan biçimbirim dizisi üzerinde yeniden yazım kurallarını uygular ve bir yüzeybiçim üretir. Yeniden yazım

kuralları, önceki ve/veya sonraki sembolleri inceleyerek üstbiçimi alabileceği yüzey sembol değerlerinden birine dönüştürürler. Sembol bir üstbiçim değilse herhangi bir dönüşüm gerçekleşmez. Her sembol dönüşümünde sonlu durum makinesi (SDM) güncellenir. Bu güncelleme tüm semboller için ünlü veya ünsüz olma, ünlüler için kalınlık-incelik-düzlük, ünsüzler için sert veya yumuşak olma gibi ses özelliklerine göre yapılır. Sentezleyici bu şekilde bir üstbiçimin dönüşmesi gereken yüzeybiçimi belirlerken üzerinden geçtiği sembollerin ses özelliklerini kullanır. Sonraki semboller henüz ziyaret edilmediği için dönüşümü belirleyecek özellikteki sembol bulunana kadar ileriye bakılır.

4.3. Bulgular

Bu bölümde, önerilen biçimbilimsel çözümleyici (Morfolog) ile Zemberek (Akın ve Akın, 2007) arasında bir karşılaştırma yapılmıştır. Bunun için ODTÜ derlemindeki (Say vd., 2002) metinler kullanılmıştır. **Tablo 4.4**'te ilgili karşılaştırma sonuçları verilmektedir.

Tablo 4.4. *Biçimbilimsel Çözümleyicilerin Karşılaştırılması*

	Zemberek	**Morfolog**
Toplam Simge (Token) Sayısı (a)	236.655	
Çözümlenen Simge Sayısı (b)	202.582	188.982
Toplam Çözümleme Sayısı (c)	493.034	489.593
Toplam Biçimbirim Sayısı (d)	1.647.316	1.763.503
Ortalama Biçimbirim Sayıları Toplamı (e)	607.256	615.330

Tablo 4.4'te verilen toplam birim sayısı (a) ODTÜ derlemini oluşturan tümcelerdeki ",", ";", ".", "?", "!" vb. noktalama işaretleri ve boşlukların ayırdığı birimlerden oluşan tekrarsız sözcük dağarcığı listesindeki toplam sözcük sayısıdır. Çözümlenen birim sayısı (b) ise çözümleyicinin tanıyabildiği birimlerin sayısını ifade etmektedir. Toplam çözümleme sayısı (c) çözümleyicinin bütün sözcükler için bulduğu bütün çözümlemelerin sayısıdır. Buradaki *çözümleme sayısı* ifadesini daha anlaşılır kılmak için **Şekil 4.4**'teki örnek tekrar hatırlatılabilir: "adama" sözcüğü için farklı

bağlamlarda geçerli olabilecek 4 farklı çözümleme söz konusudur. Toplam biçimbirim sayısı (d) çözümleyicinin bütün çözümlemeler için ayrıştırdığı biçimbirimlerin sayısıdır. Ortalama biçimbirim sayıları toplamı ise her bir sözcük için bulunan bütün biçimbirimlerin o sözcük için söz konusu çözümleme sayısına bölünmesiyle elde edilen ortalama biçimbirim sayılarının toplamıdır.

İstatistiksel analizde çözümleme sayısı, biçimbirim sayısı ve ortalama biçimbirim sayısı değişkenleri kullanılmıştır. İstatistiksel analiz sonuçlarına göre değişkenler her iki çözümleyici için ayrı ayrı değerlendirildiğinde hiçbirinin normal dağılmadığı belirlenmiştir (Kolmogorov-Smirnov Test, sig.=0).

Tablo 4.5. Çözümleyicilerin Belirsizlik ve Taneciksellik Açısından Karşılaştırılması

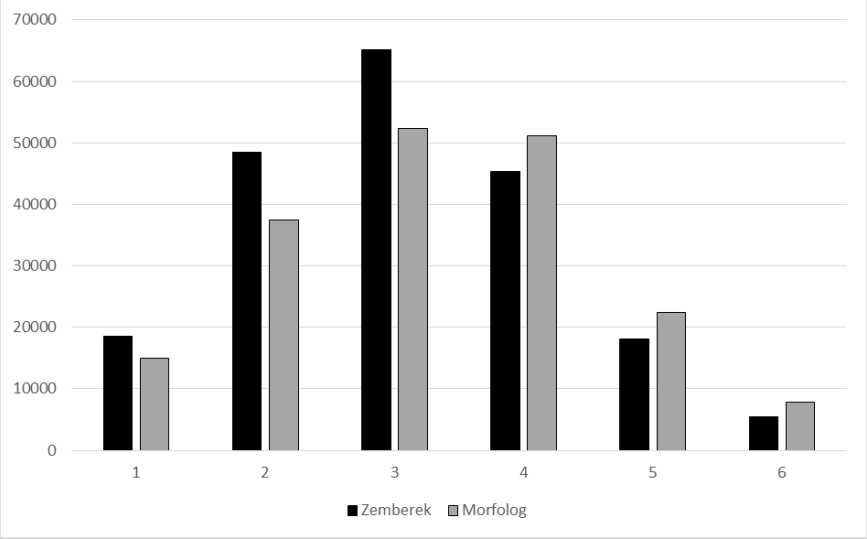
	Formül	**Zemberek**	**Morfolog**
Sözcükleri tanıma oranı	b/a=	86%	80%
Sözcük başına düşen çözümleme sayısı (belirsizlik)	c/b=	2,43	2,59
Sözcük başına düşen biçimbirim sayısı (taneciksellik)	e/b=	3	3,26

Tablo 4.5'te görülebileceği gibi, iki çözümleyici, sözcükleri tanıma oranı açısından karşılaştırıldığında Zemberek %86 ile daha başarılıdır. İyi bir biçimbilimsel çözümleyicinin kurallı olarak yazılmış hemen her sözcüğü tanıyabilmesi gerekir. Bu da sözlük bileşeninin kapsayıcılığı ile ilgilidir.

Sözcük başına düşen çözümleme sayısı biçimbilimsel belirsizlik olarak yorumlanırsa Zemberek daha az belirsizlik üretiyor görünmektedir. Bununla ilişkili olarak çözümleme sayısı değişkeni açısından iki çözümleyici arasında (Mann-Whitney U sig.=0) anlamlı bir farklılık vardır.

Sözcük başına düşen biçimbirim sayısına taneciklilik (granularity) adını verirsek, bu değişken çözümleyicinin bir sözcük için üreteceği ortalama biçimbirim sayısını belirtir ve ne kadar büyükse çözümleyicinin o ölçüde ayrıntılı bir çözümleme yaptığını gösterir. Belirsizlikte olduğu gibi taneciklilikte de sayısal değerlerden yola çıkarak bir çözümleyicinin diğerinden iyi olduğunu söylemek zordur.

İki çözümleyici arasında biçimbirim sayısı değişkeni açısından istatistiksel olarak (Mann-Whitney U sig.=0) anlamlı bir fark bulunmaktadır. Zemberek'te sözcük başına 3 biçimbirim düşerken Morfolog'da bu oran 3,26'dır. Çözümleyiciler arasında sözcük başına düşen biçimbirim sayısı dağılımı **Şekil 4.8**'de görülmektedir. (Şekilde ortalama biçimbirim sayısı değerleri tam sayılara yuvarlanarak frekansları hesaplanmıştır).



Şekil 4.8. *Taneciklilik Dağılımı*

Bu analizde elde edilen sözcük başına düşen biçimbirim oranları Türkçenin orijinal karakteristiği ile bire bir uyumlu olmayabilir. Gerçek veriler için her bir sözcüğün metin içerisinde kendi bağlamı içindeki doğru çözümlemesi seçilmeli, ardından bu çözümlemenin içerdiği biçimbirimler sayılmalı ve son olarak toplam biçimbirim sayısı toplam sözcük sayısına bölünerek sözcük başına düşen ortalama biçimbirim sayısı hesaplanmalıdır. Bu süreçte hem sözcüklerin seçildiği derlem hem de çözümleyicinin yapısı hesaplama sonucunu farklılaştıracaktır. Böyle bir inceleme başka bir çalışmanın konusu olabilir.

4.4. Sonuç

Morfolog'da Türkçe sesbilimsel kuralları modellemek için üretici sesbilim yaklaşımından yararlanılmıştır. Üretici sesbilim kuralları sözlük seviyesindeki

sembolleri yüzey seviyedeki sembollere dönüştüren yeniden yazım kurallarıdır. İki seviyeli modelde, sesbilimsel etkileşimleri denetleyen kurallar çözümleme sırasında devreye girer. Burada ise kurallara çözümlemeden sonraki denetleme aşamasında başvurulur. İki seviyeli model yerine böyle bir sistem oluşturmamızın iki temel nedeni vardır. Birincisi, iki seviyeli modelde üretim, çözümleme ve sesbilimsel kurallar iç içe geçmiş durumdadır. Bu nedenle sonlu durum makinesini oluşturmak ve güncellemek zordur. İkincisi, üretici sesbilim iki seviyeli modele göre daha anlaşılır kurallar oluşturmaya imkân verir. Öte yandan üretici kuralların belli bir düzende sıralanması zorunluluğu vardır.

Sistemin dile bağımlı kısımları sözlükler, kurallar ve morfotaktiklerdir. Bu üç bileşen herhangi bir dil için düzenlenirse çözümleyici o dilin sözcüklerini işleyebilir ancak biçimbilimsel yapı itibarıyla çok sayıda istisna bulunduran diller (ör. bükünlü diller) için kural sayısı artacak ve buna bağlı olarak model karmaşık hâle gelecektir. Dolayısıyla bu çalışmada tanıtılan çözümleyicinin sondan eklemeli diller için uygun olduğu söylenebilir.

Algoritmada en büyük maliyet çözümleme yordamında ortaya çıkmaktadır. Hem ek ağacı arama hem de sonlu durum makinesi sorgulama maliyetleri üstel fonksiyondur. Söz gelimi, 30 harften oluşan bir dizinin çözümlenmesi milyarlarca işlem adımı gerektirir. Ancak Türkçede bu kadar uzun ek dizilerine rastlama olasılığının düşük olması beklenir.

Biçimbilimsel çözümlemede, çözümleyicinin kalitesi beş ölçüte göre incelenebilir: hesaplama performansı, tanıma oranı, belirsizlik, dilbilgisel doğruluk ve taneciksellik. Morfolog'un performansını kök bulma, çözümleme ve denetleme yordamları belirlemektedir. Özellikle çözümleme sonrasında gerçekleştirilen ve üretici sesbilim kurallarını kullanan denetleme yordamı performansı düşürmektedir. Deneyler sırasında Zemberek'in çok daha hızlı çalıştığı gözlenmiştir.

Sözcükleri tanıma oranı açısından bir karşılaştırma yapıldığında da Zemberek (%86) Morfolog'dan (%80) daha iyidir. Belirsizlik açısından, bir çözümleyici, sözcükleri mümkün olduğu kadar çok biçimbirime ayırabilmeli ve belirsizliğin artmasına neden olacak yanlış dizilimleri elemelidir. Bu ölçüte göre değerlendirildiğinde, deneysel

bulgular Zemberek'in Morfolog'dan daha az belirsizlik üreten bir çözümleyici olduğunu söylemektedir ancak daha az belirsizlik üretmek çözümleyicinin dilbilgisel doğruluğunu ve tamlığını garanti etmez çünkü bir çözümleyici düşük belirsizlik ortalamasına sahip olabilir fakat doğru çözümlemelerin birçoğunu seçenekler arasında sunamayabilir. Bu noktada dilbilgisel doğruluk, diğer bir ifadeyle biçimbilimsel olarak ilgili sözcüğün bütün bağlamlardaki olası dizilimlerini önerebilme yeteneğine sahip olması gerekir. Yine de hesaplama zamanı açısından Zemberek daha az karmaşıklık üretiyor denilebilir.

Taneciksellik çözümleyicinin ne kadar ayrıntılı bir çözümleme gerçekleştirdiğinin bir ölçüsü olarak yorumlanabilir. Bu da çözümleyicide morfotaktiklerin ve eklerin nasıl tanımlandığıyla doğrudan ilgilidir. Bulgulara göre Morfolog Zemberek'ten daha ayrıntılı çözümleme gerçekleştirmektedir. Çözümleyici özellikle ileri aşamalara (sözdizimsel çözümleme vb.) girdi oluşturmak amacıyla kullanıldığında daha duyarlı iş görebilecektir.

Çözümleyicinin dilbilgisel doğruluğunu belirlemek için insan tarafından işaretlenmiş bir derleme ihtiyaç duyulur. Bağlamı içinde uygun çözümlemesi kodlanmış olan her bir sözcük çözümleyiciye verilir ve seçenekler arasında ilgili çözümlemenin üretilip üretilmediği ve çözümlemeler için bir sıralama (olasılık) veriliyorsa mümkün olduğu kadar üst sırada olup olmadığı denetlenir.

Kural tabanlı yöntemlerde, bir dile ait dilbilgisi kurallarını tam olarak modellemek mümkün değildir. Şüphesiz, dilin değişkenliği ve birçok etken nedeniyle hiçbir kural tabanlı sistemin eksiksiz olduğu iddia edilemez. Bu çalışmada önerilen Türkçe biçimbilimsel çözümleyici (Morfolog) tanıma oranı, belirsizlik ve hız açısından Zemberek'in gerisinde kalmaktadır. Ancak esas amacı olan dilbilgisel doğruluğu yansıtıcı bir değişken olarak tanecikselliğini daha iyi olduğu gözlenmiştir.

5. SÖZLÜKÇE

Tezin bu bölümünde TrLex adını verdiğimiz Türkçe biçimbilimsel sözlükçe, onu oluşturan bileşenler ve hazırlama süreci açıklanacak ve bazı bulgular sunulacaktır.

DDİ'de kural tabanlı yöntemler yaygın biçimde kullanılır ve bu amaçla yapılandırılmış dilsel veriye ihtiyaç duyulur. Yapılandırılmış dilsel verinin en karakteristik örneği sözlüklerdir. Sözlük daha çok insan kullanıcılara bilgi sağlayan fiziksel bir nesneyi hatırlatırken, sözlükçe bilgisayar programları tarafından kullanılan girdiler listesidir (Hayashi ve Ishida, 2006).

Sözlükçeler türlerine, kapsamlarına ve amaçlarına göre birçok sınıfa ayrılabilir. DDİ'de geçerli olan sözlükçeler hesaplamalı sözlükçelerdir. Bunlar, basitçe, bildiğimiz sözlüklerin manipüle edilebilir ve makinece-okunabilir türleridir (Litkowski, 2005).

Hesaplamalı sözlükçeler bilgileri birçok biçimde sunabilir: XML vb. formatlarda, ağ yapısında, veri tabanı yapısında, veri tablosu biçiminde. Temel olarak sözlükçeler maddeler hâlinde listelenmiş içeriklerinde, ses bilgisi, biçim bilgisi, sözdizimsel bilgi, dilbilgisel bilgi, anlam bilgisi, köken bilgisi, ontoloji vb. birçok başlıkta veri barındırırlar ve bu sayede sözcük anlam belirsizliği giderme, bilgi çıkarımı, soru cevaplama, metin özetleme, konuşma tanıma gibi birçok çalışma alanında kullanılırlar (Litkowski, 2005).

5.1. Önceki Çalışmalar

Hesaplamalı sözlükçelere örnek olarak verilebilecek bazı çalışmalar şöyledir:

• Comlex Syntax (Grishman vd., 1994): 38.000 İngilizce maddebaşından oluşan ve sözdizimsel bilgi (yanulamlama) içeren bir sözlükçedir.

• CLIPS (Ruimy vd., 2002): sesbilimsel, biçimbilimsel ve sözdizimsel olarak kodlanmış 55.000 başsözcükten (lemma) oluşan çok katmanlı İtalyanca bir hesaplamalı sözlükçedir.

• Maltilex (Rosner vd., 1998): Malta dili için yapılmış, biçimbilimsel ve sözdizimsel bilgiler içeren bir sözlükçedir.

• PDT-Vallex (Urešová, 2009): Prague Dependency Treebank'in (PDT) kodlanmasından elde edilen valens (valency) bilgisini içeren ve Çek dili için yapılmış bir sözlükçedir.

• Lefff (Sagot, 2010): Fransızca için sözdizimsel ve biçimbilimsel bilgilerden oluşan geniş kapsamlı ve serbest erişilebilir bir sözlükçedir. Lexical Markup Framework (LMF) ile uyumlu bir dizge olan Alexina framework[8] ile geliştirilmiştir.

• CML (Tadić ve Fulgosi, 2003): Hırvatça için üretilmiş biçimbilimsel bir sözlükçedir. Türetimsel ve çekimsel olmak üzere iki alt sözlükçeden oluşur.

• SKEL morphological lexicon (Petasis vd., 2001): Yunanca için hazırlanmış bir biçimbilimsel sözlükçedir.

• Leffe (Molinero vd., 2009): İspanyolca için geniş kapsamlı biçimbilimsel ve sözdizimsel bir sözlükçedir.

Biçimbilimsel sözlükçe özellikle biçimbilimsel çözümlemede doğrudan kullanılabilir. Bunun yanında gövdeleme ve sözcük türü belirleme gibi süreçlerde de yararlanılabilir. Ayrıca taban-ek çiftlerinin belirlenmesiyle türetim eklerinin sözlüksel çeşitliliği ve üretkenliği incelenebilir. Sözlük çerçevesinde biçim, yapı ve anlam arasındaki ilişkiler için sayısız çözümleme gerçekleştirilebilir. Bu anlamda böyle bir sözlükçe çok zengin bir kaynaktır.

5.2. Sözlükçenin Hazırlanması

TrLex oluşturulurken kaynak olarak Türk Dil Kurumu sözlüğü (2005) kullanılmıştır. Hazırlık aşamasında sözlük maddeleri ve bu maddelere ilişkin diğer bilgiler bilgisayar ortamına aktarılarak makinece-okunabilir bir veri elde edilmiştir. Bu veri bir tablo biçiminde yapılandırılmıştır. Hazırlık sürecinden sonra tablodaki her bir girdi incelenerek basit biçimbilimsel ayrıştırma gerçekleştirilmiştir. Bunun sonucunda türetilmiş başsözcükler (lemma) için taban-ek çiftleri elde edilmiş ve bunların dilbilgisel özellikleri kodlanmıştır. Son aşamada ise başsözcüğün ekleşmesinde ortaya

[8] Bk. https://gforge.inria.fr/projects/alexina/; erişim: 16.08.2017

çıkabilecek ses değişimlerini modellemek için özel işaretler kullanılarak sesbilimsel etiketleme yapılmıştır.

TrLex'teki başsözcükler, kökler ve varsa türetim eklerini içerir. Sözcüklerin çekimlenmiş biçimleri bu sözlüğün kapsamında değildir. Almanca, İngilizce vb. diller için sözcüklerin çekimlenmiş biçimlerinin en azından belirli bir kısmının sözlükçede bulunması tavsiye edilse de (Lieber, 1980), eklemeli bir dil olan Türkçede bir sözcük için çekimlenmiş binlerce biçim olabileceğinden sözlüğün başsözcüklerinde çekim ekleri yer almaz. İstisna olarak bazı çekim ekleri sözcük yapımında rol almıştır. Ancak bunların örnekleri sınırlı sayıdadır ve bu ekler üretken değildir.

5.2.1. Veri

Veri tablosunda şu alanlar bulunmaktadır: ID, lemma, base, suffix_morph, suffix, base_POS, lemma_POS, {reciprocal, passive, aorist, causative, reflexive, causative_form}, {transformation, attachment, deletion, accent}, final_lemma, origin, meaning, example_sentence. **Tablo 5.1**'de veri tablosunun içerdiği alanlar ile bu alanların açıklamaları ve örnekleri sunulmuştur.

Tablo 5.1. *Veri Tablosunun Alanları*

Alan	Açıklama	Örnek
ID	Girdileri ayırmak için bir anahtar (Sesteş girdiler farklı ID'lere sahiptir.)	10557
Lemma	Maddenin sözlükte yer alan yazılı formu	çıkmak
Base	Eğer varsa başsözcüğün sonundaki ek çıkarıldığında elde edilen kısım	çık
Suffix_morph	Eğer varsa lemma'nın sonundaki ekin görünen biçimi	
Suffix	Ekin sözlüksel, genelleştirilmiş biçimi	
Base_POS	Tabanın sözcük türü	
Lemma_POS	Başsözcüğün sözcük türü	Verb
Reciprocal	İşteşlik	0
Passive	Edilgenlik	1
Aorist	Eylemin seçtiği geniş zaman eki	0
Causative	Ettirgenlik	1
Reflexive	Dönüşlülük	0
Causative_form	Eğer varsa eylem ettirgenlik ekini alırken ortaya çıkan istisnai durum	çıkar

Transformation	Başsözcüğün sonundaki ses üzerinde meydana gelebilecek ses dönüşümü	
Attachment	Maddedeki ünsüz türemesi vb. olaylar	
Deletion	Ses düşmesi	
Accent	İnceltmeler	
Final_lemma	Ses olaylarını açıklayan sembollerin eklenmesiyle elde edilen son biçim	çık
Origin	Başsözcüğün geldiği dil	Turkish
Meaning	Başsözcüğün tanımı	Birdenbire görünmek
Example_sentence	Başsözcüğü söz konusu anlamı çerçevesinde örnekleyen bir tümce	Neden hiçbir korsan filosu önümüze çıkamadı?

5.2.2. Biçimbilimsel ayrıştırma

Maddelerin biçimbilimsel ayrıştırması yapılırken yalnızca eşzamanlı biçimbilimsel süreçler ve Türkçenin ekleri dikkate alınmış; artzamanlı örneklerde taban veya ek çok bariz ise bir ayrıştırma yapılmış ve artzamanlı yapı ilgili alanda işaretlenmiştir. Örneğin göster- eyleminin eşzamanlı bir inceleme ile biçimbilimsel ayrıştırması mümkün değildir. Artzamanlı bir araştırma yapıldığında ise göster- eyleminin <köster-< kö-z-ter- biçiminde çözümlenebileceği tespit edilmiştir (Günşen 2006: 35). Art zamanlı bu tür incelemeler tez kapsamında değildir.

Sözlükçede, içinde boşluk içermeyen yapılar biçimbilimsel ayrıştırmaya tabi tutulmuştur. İçinde boşluk içermese de birden çok taban bulunduran bitişik yazılmış birleşik sözcükler için bu aşamada bir ayrıştırma yapılmamıştır. Çoklu sözcükler (multiword) basit sözcüklere kıyasla incelenmesi daha zor olan birimlerdir. Ayrıca yer adı, kurum adı, sayı, deyim gibi farklı yapılarda yeni çoklu sözcükler her an türetilebilir. Bu nedenle onları durağan olarak listeleme yerine örüntülerinden yola çıkarak tanıma amaçlı çalışmalar yapılmaktadır (Eryiğit vd., 2015).

Burada gerçekleştirilen biçimbilimsel ayrıştırmadaki temel strateji taban odaklıdır. Daha açık ifadeyle, anlamı tanımlanmış olan sözlük maddesinin içinden, bariz olan ek çıkarıldığında elde edilen tabana göre bir ayrıştırma gerçekleştirilmiştir. Eğer taban kullanımda olan bir sözcük ise ek ayrıştırılmış; taban kullanımda ancak biçimce dönüştüyse güncel biçimi işaretlenmiş; fakat taban kullanımda olan bir sözcük değilse ek ayrıştırılmamıştır.

5.2.3. Sesbilimsel etiketleme

Çalışmanın son aşamasında başsözcüklerin ek aldıklarında dönüşebilecekleri bütün biçimler belli sözlüksel sembollerle formüle edilmiştir. **Tablo 5.2**'de başsözcüğün son harfinde gerçekleşen ses olayını modelleyen semboller ve bunların yüzeybiçimleri verilmiştir. Soldaki yüzeybiçim başsözcük hiçbir ek almazsa sözlüksel sembolün dönüşeceği varsayılan biçimi göstermektedir. Örnek verirsek, "kitaP" başsözcüğündeki P sözlüksel sembolü "kitabı oku" tümcesinde olduğu gibi "b" sesine ya da "bu kitapta" öbeğinde olduğu gibi "p" sesine dönüşebilir; ancak hiçbir ek almazsa da "güzel kitap" öbeğinde olduğu gibi varsayılan biçim olan "p" sesine dönüşecektir. Tablodaki E sembolü diğerlerinden farklı olarak ünlü sesleri modellemektedir ve yalnızca eylemlerde gözlenir.

Tablo 5.2. *Sözlüksel Semboller ve Yüzeybiçimleri*

Sözlüksel Sembol	Yüzeybiçimler		Örnek
P	p	b	kitaP
K	k	ğ	odaK
T	t	d	taT
Ç	ç	c	ağaÇ
G	k	g	renG
Ğ	g	ğ	psikoloĞ
E	e	i	dE

Bu sembollerin yanı sıra, iyelik eki barındıran ögeleri belirtmek için & (ahududu& - ahududuna), ünsüz türemesi için 0 (af0 - affı), istisnai bir durum için % (cami% - camii), sonunda "su" birimi bulunanlar için Y (suY - suyu), ünlü düşmesi için $ (ağ$ız - ağzı) sembolleri kullanılmıştır. Ayrıca incelmeyi göstermek için aksan işareti eklenmiştir. Örnek: hâl - hali.

5.2.4. Sözlüksel imleme çerçevesi

Sözlüksel imleme çerçevesi (Lexical Markup Framework; LMF) DDİ alanında kullanılacak sözlükçeler geliştirmek için tasarlanmış bir standart modeldir (Francopoulo et al., 2006). Bir sözlüksel girdideki bilgi hiyerarşisini tanımlayan "the core package" ve spesifik sözlüksel kaynakları hazırlarken bu bilgilerin nasıl tekrar kullanılacağını açıklayan "extensions of the core package" bileşenlerinden oluşur.

Bu çalışmada LMF etiketleri Türkçeye uyarlanmış ve bazı ek özellikler tanımlanmıştır. Örnek bir sözlüksel girdi aşağıda sunulmuştur:

```
<LexicalEntry>
        <feat att="partOfSpeech" val="noun"/>
        <Lemma>
                <feat att="lexicalForm" val="abajurculuK"/>
        </Lemma>
        <MorphologicalStructure>
                <Base>
                        <feat att="lexicalForm" val="abajurcu"/>
                        <feat att="partOfSpeech" val="noun"/>
                </Base>
                <Suffix>
                        <feat att="lexicalForm" val="lHK"/>
                        <feat att="surfaceForm" val="luk"/>
                </Suffix>
        </MorphologicalStructure>
        <WordForm>
                <feat att="surfaceForm" val="abajurculuk"/>
        </WordForm>
        <WordForm>
                <feat att="surfaceForm" val="abajurculuğ"/>
        </WordForm>
</LexicalEntry>
```

Örnekte görüldüğü gibi bir sözlüksel girdi, başsözcük (lemma), onun sözlüksel biçimi (lexical form), biçimbilimsel yapı (morphological structure) ve onun taban (base) ve eki (suffix), başsözcüğe ait sözcükbiçimlerden (word form) oluşur. Bu gösterimde, standart LMF'den farklı olarak biçimbilimsel yapı bilgisi de yer alır. Öte yandan, veri tablosunda anlam bilgisi bulunurken LMF gösterimine bu bilgi dâhil edilmemiştir.

5.3. Bulgular ve Tartışma

Çalışma sonucunda elde edilen sözlükçeye (veri tablosu) ait bazı temel istatistikler **Tablo 5.3**'te verilmiştir.

Tablo 5.3. *Sözlükçeye İlişkin Temel İstatistikler*

	Frekans	**Yüzde**
Girdi	110,960	-
Başsözcük	83,381	75 (in entry)
Sözlüksel biçim (tekrarsız)	82,627	99 (in lemma)
Tekil sözcük	78,426	71 (in entry)
Tekil sözcük başsözcüğü (tekrarsız)	54,670	70 (in single-word)
Tekil sözcük biçimi (tekrarsız)	54,042	69 (in single-word)

Tablo 5.3'te girdi, veri tablosundaki her bir satıra verilen isimdir. Başsözcük ise belirli bir anlam grubunu temsil eden sözlüksel ögedir. Aynı biçime sahip birden çok başsözcük olabilir. Örneğin bar (a place where alcoholic drinks are sold and drunk) ile bar (a small block of something solid)[9] iki ayrı başsözcüktür ve bu iki başsözcüğün de kendi alt anlamları olabilir. Sözlüksel biçim bu örnekteki "bar" sözcüğüdür. Bu çalışmada özel olarak ele alınan ögeler olan tekil sözcükler, içinde boşluk karakteri bulunmayan başsözcüklerdir. 78,426 adet (%71) tekil sözcük yapısındaki girdi bu çalışma için temel veriyi oluşturmaktadır.

Sözlükte bulunan bütün sözlüksel biçimlerin sayısını bütün başsözcüklerin sayısına oranladığımızda %99 elde ederken, tekil sözcük biçimlerinin sayısını tekil sözcük girdilerine oranladığımızda %69 elde edilmektedir (**Tablo 5.3**). Bunun nedeni kaynak sözlükte anlam çeşitlenmesinin tekil sözcük yapılarında yoğunlaşması veya maddebaşı seçimiyle ilgili bir eğilim olabilir. Bu oranlamayı tersten yaparsak, tekil sözcük girdi sayısını tekil sözcük biçim sayısına bölersek 1.45 oranını buluruz. Bu oran sözlükte tekil sözcük yapısındaki her bir biçim başına ortalama 1.45 adet anlam düştüğünü söyler. Bu sayı Türkçede anlam belirsizliği giderme probleminin zorluğunu yansıtan bir ölçü olarak yorumlanabilir. Bir başka çalışmada (İlgen ve Adalı, 2012) elde edilen 1.61 oranı ile karşılaştırma yapmak, ilgili çalışmada bir sözlükçe örneklemi

[9] http://dictionary.cambridge.org/dictionary/turkish/bar_1; 6/13/2015

kullanıldığı için zordur. Ancak yine de Türkçe için anlam belirsizliği olarak adlandırılabilecek bu oranın bu seviyelerde dolaştığını söyleyebiliriz.

Tekil sözcük başsözcüklerinin yarıdan fazlasının (%56.7) türetilmiş yapıda olduğu gözlemlenmiştir (**Tablo 5.4**). Bu oran birleşik sözcüklerin oranıyla (%2.7) kıyaslanırsa Türkçede sözcük oluşumu için biçimbilimsel süreçlerin çok ağırlıklı biçimde tercih edildiği sayısal olarak ortaya konmuş olur. Ancak sözcük oluşumunda türetim ile birleştirmeyi net olarak ayırmak zordur: birleşik sözcüklerin içinde türetim örnekleri yer alabilir. Örnek olarak "açıkgöz" sözcüğünü incelersek, "açık" ve "göz" şeklinde iki ögeden oluşur. Birinci ögenin içinde bir türetim varken ikinci öge bir köktür. Bu örnekte önce türetimin ardından birleşim işleminin gerçekleştiği açıktır. Ancak öyle sözcükler olabilir ki bu süreçlerin sözcük oluşum sürecinde hangi sırada gerçekleştiğini belirlemek zordur. Örnek vermek gerekirse, "özveri" sözcüğünde bu süreçlerin işletilme sırası belirsizdir. Bu bölümde tanıtılan sözlükçe için sözcük oluşumunda türetim işlemi baskın görünmektedir. Ancak türetim birleşik sözcük yapımının içine de sızmıştır. Bu durum sözcük oluşturma yollarından birleştirme ve türetim arasındaki sınırın belirsizliğine bir örnektir (Booij, 2005; Ralli, 2010).

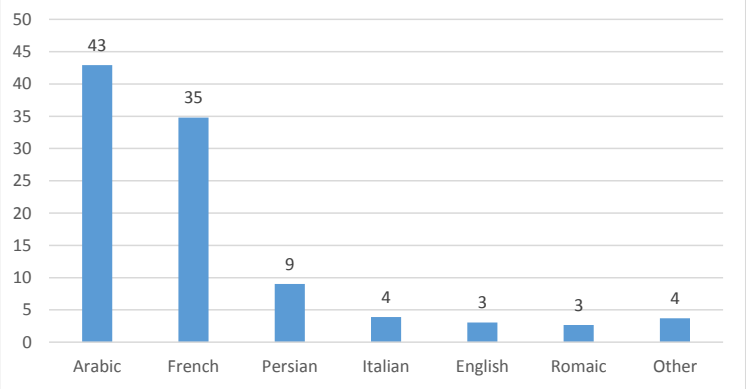
Tablo 5.4'te tekil sözcük başsözcüklerinin sınıflarına ilişkin istatistikler verilmiştir. *Türetilmiş* sınıfı, bir taban ve bir ekten oluşan yapıları (göz-lük), *kök* ek içermeyen tekil ögeleri (göz), *birleşik* bitişik biçimde yazılmış sözcükleri (açıkgöz), *problemli* çözümlenemeyen başsözcükleri (gözene), *iyelik yapısı* sonunda iyelik eki bulunduran bileşik sözcükleri (gözyaşı), *pekiştirme* pekiştirmeleri (masmavi, upuzun) temsil eder.

Tablo 5.4. Tekil Sözcük Başsözcükleri Sınıflarına İlişkin İstatistikler

Tekil Sözcük Başsözcükleri Sınıfları	Frekans	Yüzde
Türetilmiş	31119	56.7
Kök	20762	37.8
Birleşik	1465	2.7
Problemli	861	1.6
İyelik yapısı	557	1.0
Pekiştirme	100	0.2

Sözlükteki başsözcüklerin sözcük türleri ve onlara ait yüzdelik oranları şöyledir: ad %64, sıfat %18, eylem %14, belirteç %3, diğer %1.

Sözlükte yer alan başsözcüklerin kökenleri iki temel grupta toplanırsa, sözlüğün %73'ü Türkçe kökenli sözcüklerden %27'si yabancı dillerden geçen sözcüklerden oluşmaktadır. **Şekil 5.1**'de yabancı dilden geçen sözcüklerin kendi içindeki dağılımı yüzdelik oranlarıyla verilmiştir.

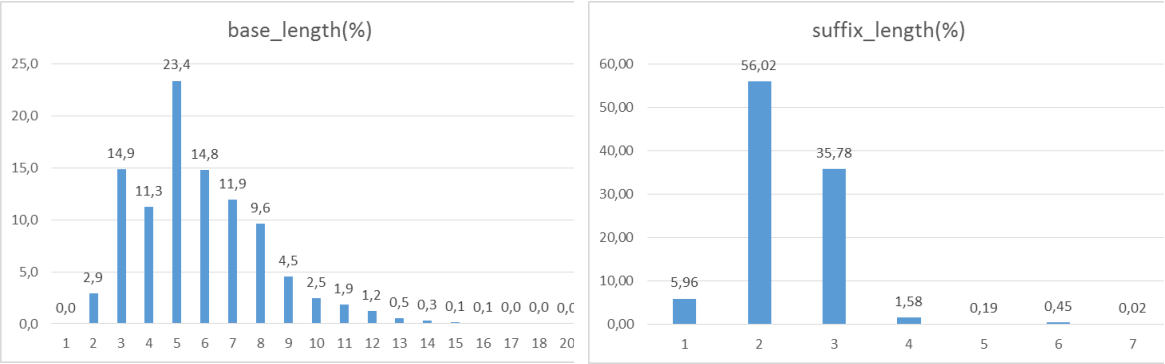


Şekil 5.1. *Alıntı Sözcüklerde Yabancı Dillerin Oranı (Yüzde)*

Şekil 5.1'deki sonuçlara göre en çok alıntı sözcüğün Arapça olması şaşırtıcı değildir. Ancak alıntı sözcüklerin kökenleri Doğu ve Batı dilleri olarak iki gruba ayrılırsa sırasıyla %52 ve %45 oranlarına sahip oldukları görülür ki ikinci oranın bu kadar büyük olmasına sebep olan ağırlıklı olarak Fransızcadır. Ayrıca bu dağılımın alanyazınla (Karaca, 2012) uyumlu olduğu söylenebilir.

5.3.1. Biçimbilimsel çözümleme

Şekil 5.2'de üzerinde biçimbilimsel ayrıştırma gerçekleştirilen, türetilmiş yapıdaki tekil sözcük girdisinden elde edilen istatistikler sunulmuştur.



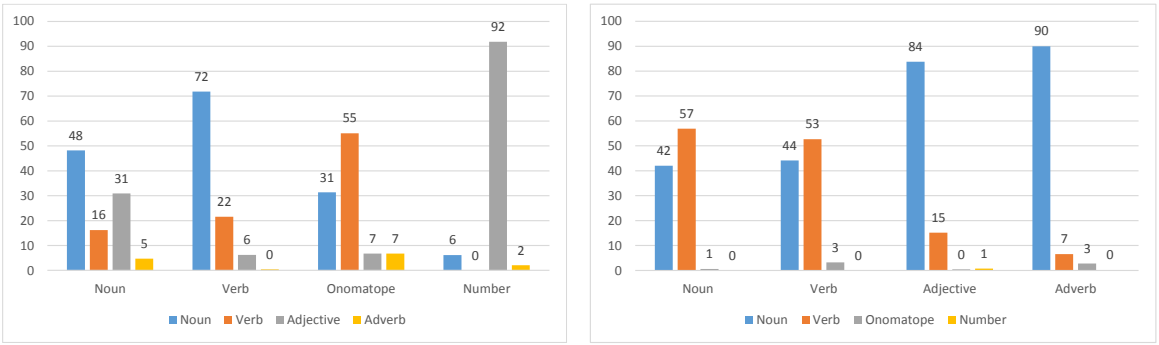
Şekil 5.2. *Tekil Sözcük Girdilerinde Taban ve Ek Uzunluk Dağılımı*

Şekil 5.2'de Base_length değişkeni tabanların harf uzunluklarını vermekte iken, suffix_length değişkeni eklerin harf uzunluklarını gösterir.

Biçimbilimsel çözümlemenin ilk sonucu olarak taban ve ek uzunlukları hesaplandığında (bk. **Şekil 5.2**) en çok tekrar eden (mod) taban uzunluğunun 5 olduğu görülmüştür. Bu sonuçla Türkçe için önerilmiş en hızlı gövdeleme yöntemi olan sözcüğün ilk 5 karakterini gövde olarak alma (Köksal, 1981; Sever ve Tonta, 2006) arasında bir ilişki kurulabilir. Ancak gövdeleme yönteminde tümce seviyesinde çekimlenmiş sözcükler işlenirken sözlükçede yalnızca türetim eki almış sözcükler bulunmaktadır. Bu çalışmada gerçekleştirilen biçimbilimsel çözümlemede türetilmiş sözcükleri oluşturan taban ve son ek çiftleri belirlenmiştir. Tabii olarak, bu çözümleme gövdelemeden farklı bir işlemdir. Ek uzunlukları için de ortalama değer 2.35 olarak hesaplanmıştır. Bu da Güngör'ün (2003) yaptığı çalışmada, corpus'tan elde edilen 2.44 ortalama ek uzunluğu değerine oldukça yakındır. Söz konusu çalışmadaki derlem için ortalama bir sözcüğün 40%'ı ek iken, TrLex'teki bir başsözcüğün 29%'u ekten oluşur.

Biçimbilimsel ayrıştırma sonuçlarına göre, 407 yüzeybiçim, 149 ek biçimi ve 309 ek sayılmıştır. Yüzeybiçim, eklerin başsözcüklerde gözlenen biçimlerini ifade eder. Örneğin "kitap-lık" çözümlemesinde "lık" bir yüzeybiçimdir. Ek biçimi farklı yüzeybiçimlerin bir arada toplanmış şeklidir. Türkçede ünlü uyumu, ünsüz benzeşmesi gibi ses olaylarının yarattığı çeşitlilik bazı joker semboller kullanılarak standartlaştırılır. Örneğin "lık, lik, luk, lük" gibi yüzeybiçimler için "lHK" ek biçimi kullanılır. Ek ise bir ek biçimine sahip, belli sözcük türündeki tabanları belli sözcük türündeki başsözcüklere dönüştüren dil ögesidir. Bu tanım yalnızca türetim ekleri için

değil çekim ekleri için de genellenebilir. **Şekil 5.3**'te sözcük türlerinin dönüşüm oranları verilmiştir.



(a) Taban sınıflarına göre **(b)** Başsözcük sınıflarına göre

Şekil 5.3. *Sözcük Türlerinin Dönüşüm Oranları (Yüzde)*

Şekil 5.3 a'da türetim eklerinin tabanlar için dört sözcük türünü (Noun, Verb, Onomatope, Number) sonuç sözcük türlerine (Noun, Verb, Adjective, Adverb) nasıl bir dağılımla dönüştürdüğü görülmektedir. **Şekil 5.3Şekil 5.3** b'de türetilmiş başsözcüklerin sözcük türlerinin (Noun, Verb, Adjective, Adverb) hangi taban sözcük türlerinden (Noun, Verb, Onomatope, Number) ne oranda geldiği görülmektedir.

Başsözcüklerin sözcük türleri açısından dağılımı incelenmiş ve büyük çoğunluğunun (%99) dört temel grupta toplandıkları gözlenmiştir. Bunlar frekans sırasına göre ad, sıfat, eylem ve belirteçtir. Sıfatların eylemlerden de çok olmasını Türkçedeki sözcük türlerinin karakteristik özellikleriyle açıklayabiliriz. Türkçede adlar ve sıfatlar arasında kesin sınırlar bulunmaz: sıfatlar adlar gibi çekimlenebilir. Bu durum hesaplamalı çalışmalarda sıfatları adlar gibi işlemeyi gerektirebilir (Oflazer, 1994).

Eklemeli bir dil olan Türkçede türetim ekleri birbirlerinin peşine eklenerek bir kök üzerinde defalarca sözcük türü değişimi yapabilirler. Burada tabanlar için Ad, Eylem, Yansıma ve Sayı olmak üzere dört taban sınıfı ve sözcüklerin bütünü için de Ad, Eylem, Sıfat ve Belirteç olmak üzere dört temel sözcük türünü esas aldık. Buna göre belli bir taban sınıfından hangi sözcük türlerine nasıl bir dağılım gerçekleştiği (**Şekil 5.3** a) ve belli bir sözcük türüne hangi taban sınıfından nasıl bir dağılımla gelindiği (**Şekil 5.3** b) hesaplanmıştır. Burada dikkat çekici olan sonuçlardan bazıları, adların sıfatlara

dönüşme eğilimi (%31) ve sıfatların da en çok adlardan oluşmasıdır (%84). Bu durumun, daha önce de bahsedilen, Türkçede adlarla sıfatlar arasındaki ayrımı yapmanın zorluğuyla bir ilişkisi olması muhtemeldir.

5.3.1.1. Eylemler

Biçimbilimsel çözümlemede eylemler için 5 özellik kodlanmıştır. Bunlar sırayla reciprocal (0: -Uş çatı eki alamaz, 1: -Uş çatı eki alabilir, 2: ne alabilir ne alamaz), passive (0: -Hl çatı eki alamaz, 1: -Hl çatı eki alabilir, 2: ne alabilir ne alamaz), aorist (0: -Ar geniş zaman eki alabilir, 1: -Ur geniş zaman eki alabilir, 2: -z geniş zaman eki alabilir), causative (0: -Ut ettirgenlik eki alabilir, 1: -DHr ettirgenlik eki alabilir, 2: her ikisini de alamaz), reflexive (0: -Un dönüşlülük eki alamaz, 1: -Un dönüşlülük eki alabilir, 2: ne alabilir ne alamaz). **Tablo 5.5**'te eylem özelliklerinin üç kategoriye göre yüzdelikleri verilmiştir.

Tablo 5.5. *Eylem Özelliklerinin Üç Kategoriye Göre Yüzdeleri*

	0	**1**	**2**
reciprocal	81	18	0
passive	41	59	0
aorist	5	95	0
causative	34	56	10
reflexive	82	17	0

Eylemler üzerinde yapılan incelemeye göre (**Tablo 5.5**) eylemlerin büyük çoğunluğu (%81) işteşlik eki -Uş alamaz. Bunun birkaç nedeni olduğu söylenebilir. Birincisi morfotaktik nedendir: Biçimbilimsel olarak belli çatı eklerinden sonra işteşlik eki gelemez. Örneğin acıt- eyleminin sonuna bu nedenle bir –Uş eki gelmesi mümkün değildir. Çünkü acıt- acı- eyleminin ettirgen biçimidir. Türkçe bu sorunu çözmek için sözdizimi kullanır: "birbirini acıt-". İkincisi sesbilimsel nedendir: sonunda "ş" sesi barındıran bir eylemin sonuna işteşlik eki eklenemez. Örnek: çalış- eyleminin sonuna işteşlik gelirse *çalışış- gibi bir sonuç ortaya çıkar. Türkçe burada da birinci durumdaki gibi sözdizimi kullanarak işin üstesinden gelebilir: "birlikte çalış-". Sonuncu neden

anlambilimsel nedendir: Bazı eylemler için işteşlik anlamsal olarak uygun değildir. Örnek: yut- eylemi karşılıklı veya birlikte yapılması mümkün olmayan bir eylemdir fakat bu son neden belirlenmesi en zor olandır. Bu nedenle çözümlemede öznel bir değerlendirmeye yol açmış olabilir.

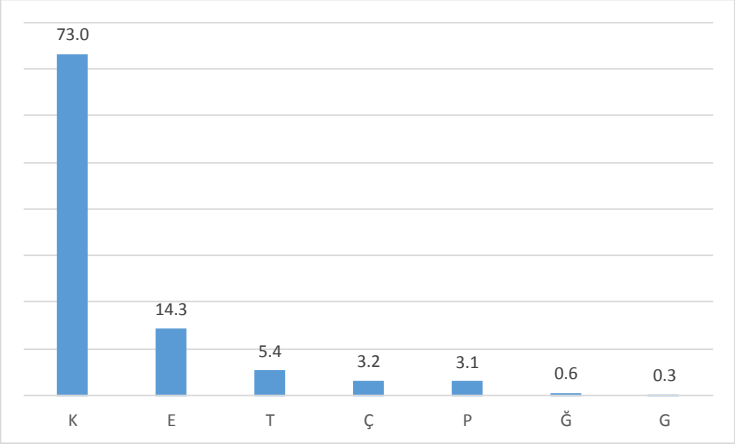
Benzer şekilde dönüşlülük eki –Un eylemlerin büyük çoğunluğuna (%82) eklenemez. Burada da işteşlik ekinde olduğu gibi üç temel neden (biçimbilimsel, sesbilimsel, anlambilimsel) sayılabilir. Edilgenlik ekine gelirsek, –Hl eylemlerin yarısından fazlasına (%59) eklenebilir. Çünkü -Hl eki biçimbilimsel olarak yalnızca hâlihazırda bir edilgenlik eki almış eylemlere gelemez, bunun dışındaki çatıları almış eylemlerin hepsine gelebilir. Bu ekin tek kısıtı sesbilimsel kısıttır: ek, gel-, ol-, ata-, yaşa- vb. sonunda bir ünlü veya "l" sesi olan eylemlere gelemez.–Hl edilgenlik ekini alamayan bu örnekler –Un ekiyle edilgen olurlar: gelin-, olun-, atan-, yaşan- gibi.

Eylemlerin geniş zaman eki alırken büyük çoğunlukla (%95) –Ur ekini tercih ettikleri gözlenmiştir. -Ur eki alan eylemlerin ortalama uzunluğu 7.96 iken –Ar ekini tercih eden eylemlerin ortalama uzunluğu 3.87'dir. Bu istatistiğe göre tek heceli eylemlerin çoğunlukla –Ar eki aldığını söyleyebiliriz.

Ettirgenlik eklerinin durumuna bakıldığında, eylemlerin yarısından çoğunun –DHr ettirgenlik ekini alabilmekte olduğu gözlemlenir. –Ut ettirgenlik ekinin kullanılma nedeni bazen morfotaktik gereği –DHr eki almış eylemlere getirmek (yap-tır-t) bazen ise sesbilimsel nedenlerle ünlü ile biten eylemlere eklemektir. Geriye kalan grup ise (%10) ettirgenlik eki alamayan eylemlerdir. Bunlar çoğunlukla edilgenlik eki almış eylemler ve çok az olarak da olumsuzluk eki almış eylemlerdir. Bu eylemlerin ettirgenlik eki alamamasının nedeni morfotaktiklerdir.

5.3.2. Sesbilimsel etiketleme

Şekil 5.4'te son harfteki ses dönüşümlerinin bu tip dönüşümler içeren başsözcükler içindeki oranları verilmiştir.



Şekil 5.4. *Başsözcük Dönüşümlerinde Son Sembollerin Oranları (Yüzde)*

Başsözcüğün sonuna eklenen sembollerle modellenen ses olaylarından ünsüz türemesi 63 başsözcükte (af0, hak0), iyelik eki barındıran sözcük 557 başsözcükte (bilinçaltı&, kartopu&), su ve ne köklerinin çekimlenmesindeki istisnai durum 6 başsözcükte (akarsuY, neY) gözlenmiştir.

Ses düşmesi 228 başsözcükte gözlenir. Bunların 188'i Arapça (as$ıl, em$ir), 29'u Türkçe (ağ$ız, bey$in) ve 7'si Farsça (müh$ür, şeh$ir) kökenli sözcüklerdir. Ses incelmesi 413 başsözcükte gözlenmiştir. Bunların 209 adeti Arapça (emsâl, hakikât), 155 adeti Fransızca (materyâl, üniversâl), 13 adeti Farsça (hemhâl, tembûl) ve 10 adeti de İngilizce (hôl, vâlf) kökenli sözcüklerdir.

Sesbilmsel etiketleme işleminden elde edilen sonuçlara bakılacak olursa, özellikle ses incelmesi olayı Türkçedeki ünlü uyumunu etkilemesi nedeniyle sesbilimsel olarak incelenmelidir. Ses incelmesi en fazla Arapça kökenli başsözcüklerde görülse de ses incelmesi içeren örneklerin toplam Arapça kökenli sözcük içindeki oranına bakıldığında (%3,24) bu oranın Fransızca kökenli sözcükler için hesaplanan orana (%2,96) çok yakın olduğu görülmüştür (Farsça için %0,96).

Ses dönüşüm incelemesinin sonucuna göre (**Şekil 5.4**), ses dönüşümü içeren başsözcüklerde en çok K dönüşümü gözlenmiştir. Bunu da E ve T dönüşümleri takip etmektedir. E dönüşümünün frekansının çok olmasının nedeni bu dönüşümün e ve a sesleriyle biten eylemlerde gözlenmesidir. T dönüşümü çok nadir olarak eylemlerde gözlenir. Örnekleri giT ve eT eylemleridir.

5.4. Sonuç

Bu bölümde TrLex adını verdiğimiz hesaplamalı biçimbilimsel sözlükçenin oluşturulma süreçleri açıklanmış ve sonuçlardan elde edilen bazı istatistikler sunulmuştur. TrLex esas olarak biçimbilimsel çözümlemede kaynak olarak kullanılmak üzere hazırlanmıştır. Ancak doğal dil işlemenin birçok alanında kullanılma potansiyeli olan zengin bir sözlükçedir. Gövdeleme ve sözcük türü etiketleme gibi birçok alanda veri kaynağı olarak kullanılabilir.

Sözlükçe çalışması sonucunda biçim, yapı, anlam ve diğer bilgilerin yer aldığı daha yoğun bilgi içeren bir veri tablosu ve bu tablodan elde edilen yalnızca biçimbilimsel bilgi ve sözcük türü bilgisi içeren LMF formatında bir XML dosyası elde edilmiştir.

6. VERİ

Bu bölümde Bölüm 8'de tanıtılan deneylerde kullanılan metinlerin özellikleri ve elde edilme süreçleri açıklanacaktır. Tezde iki tür veri kullanılmıştır: Web'den elde edilen metinlerden oluşan ve modeller için kaynak veri niteliğindeki metin koleksiyonu ve tezde üretilen bir ağaç yapılı derlem.

6.1. Metin Koleksiyonu

Tezde tümcelerin sözdizimsel belirsizliklerini gidermek için kullanılacak dil modelleri Türkçe metinlerden gözetimsiz olarak elde edilmiştir. Bu amaçla çok sayıda tümceye ihtiyaç duyulmuştur. Mevcut Türkçe derlemler bu ihtiyacı karşılayamadığı için web üzerinden toplanan metinlerden çok büyük bir koleksiyon oluşturulmuştur. Oluşturulan bu koleksiyon bir derlem değildir. Çünkü metin türlerinin dengelenmesi, temsil edicilik gibi derlem olma ölçütlerini sağlamak amaçlanmamıştır.

Koleksiyon, filtrelenmiş 42.630.365 adet tümceden oluşmaktadır. Bu tümceler arasından 1.394.491 tanesi TMoST tarafından tanınan, geçerli tümcelerdir.

6.2. Ağaç Yapılı Derlem

Yazılı derlemlerin (corpus) hesaplamalı dilbilim çalışmalarında önemli bir yeri vardır. Bir derlem, yalnızca gözlenen verilerden oluşuyorsa, ham derlem (raw corpus); buna ek olarak çeşitli biçimsel, yapısal, anlamsal bilgilerle zenginleştirilmişse, imlenmiş derlem (annotated corpus) olarak adlandırılır. İmlenmiş derlemler iki temel amaç için kullanılmaktadır: öğrenme ve sınama. Öğrenme amaçlı kullanımda derlemlerin içerdiği yazılı dil örnekleri incelenerek kurallar ve/veya istatistiksel dil modelleri üretilmeye çalışılır. Sınama amaçlı kullanımda ise imlenmiş derlem altın standart olarak kabul edilir ve üretilen kuralların ve/veya modellerin doğruluğu bu standarda göre belirlenir. Gözetimli yöntemlerde, imlenmiş derlemler her iki temel amaç doğrultusunda değerlendirilirken, gözetimsiz yöntemlerde yalnızca sınama amaçlı olarak kullanılır.

Ağaç yapılı derlem (AYD; treebank), sözdizimsel yapısı imlenmiş tümcelerden oluşan derlemdir. AYD'ler farklı kuramsal yaklaşımlara göre hazırlanabilmektedir. Bunlar arasında en çok başvurulanlar, BD ve ÖYD'dir.

6.2.1. Önceki çalışmalar

İngilizce için oluşturulan ilk büyük ölçekli AYD, Penn Treebank'tir (Marcus vd., 1993). ÖYD esaslı olan bu derlem, sözcük türü etiketlenmiş yaklaşık 7 milyon sözcük, yapısal olarak ayrıştırılmış 3 milyon sözcük, yüklem-üye yapısı (predicate-argument structure) imlenmiş 2 milyonun üzerinde sözcük ve konuşma metninden elde edilen 1,6 milyon sözcükten oluşur (Marcus vd., 2012, s. 5). Penn Treebank'te kaynak metin olarak Wall Street Journal haberleri ve Brown Corpus kullanılmıştır.

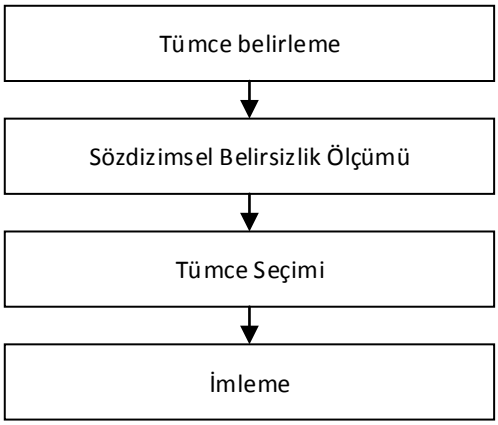
Bildiğimiz kadarıyla Türkçe dili için hazırlanmış tek AYD ODTÜ-Sabancı Türkçe Ağaç Yapılı Derlemidir (OSTAD; Atalay vd., 2003; Oflazer vd., 2003). OSTAD'ın internet üzerinden ulaşılabilen, elden geçirilmiş sürümü[10], BD yaklaşımı esas alınarak çözümlenen 5635 tümceden oluşmaktadır. Aslen ODTÜ Derleminin (OD) bir alt derlemi olan OSTAD, XCES standardına göre imlenmiştir. İmleme işlemi iki aşamadan oluşmuştur: Birinci aşamada biçimbilimsel çözümleyicinin ürettiği geçerli çözümlemeler seçilmiş; ikinci aşamada çözümlemelerin içerdiği çekim öbekleri (inflection groups) bağımlılık dilbilgisi kurallarına göre birbiriyle ilişkilendirilmiştir.

6.2.2. Derlemin hazırlanması

Bu tezde öbek yapı dilbilgisi esasına dayanan bir AYD inşa edilmiştir. Bu derleme AUT (Anadolu University Treebank) adını veriyoruz. AUT, OD'den elde edilmiş tümcelerden oluşmaktadır.

AUT'nin hazırlanma aşamaları **Şekil 6.1**'de verilmiştir.

[10] https://web.itu.edu.tr/gulsenc/METUSABANCI_treebank_v-1.rar; erişim: 03.05.2017



Şekil 6.1. *AUT'nin Hazırlanma Aşamaları*

Tümce belirleme, OD'nin tümcelere ayrılmış durumda olmamasından kaynaklanan bir önişlem olup OpenNLP[11] adlı Java kütüphanesi yardımı ile gerçekleştirilmiştir. Belirlenen tümceler, *sözdizimsel belirsizlik ölçümü* aşamasında, ürettikleri sözdizim ağacı sayısına göre sıralanmıştır. Bir tümceden ne kadar çok sözdizim ağacı üretilebiliyorsa tümcenin o ölçüde sözdizimsel belirsizliğe sahip olduğu varsayılmıştır. *Tümce seçiminde*, sözdizimsel belirsizliğe göre sıralanmış tümceler arasında 10'dan fazla sözdizim ağacı üretenler alınmıştır. Son aşama olan imlemede, seçilen tümceler üzerinde ayrıntılı sözdizimsel ve biçimbilimsel imleme gerçekleştirilmiştir.

6.2.3. Bulgular

AUT'nin inşa edileceği temel tümcelerini seçmeden önce OD üzerinde üç aşamalı bir filtreleme işlemi uygulanmıştır. Birinci aşama simge filtrelemesidir. Simge filtrelemesi ile simgenin içerdiği ve geçerli alfabede yer almayan bütün semboller atılır. Böylece noktalama işaretleri ve rakamlar çıkartılarak simgelerin yalnızca harfleri içermesi sağlanır. İkinci aşama olan tümce filtrelemesi ile simge sayısı 20'den daha çok olan tümceler ya da biçimbilimsel belirsizliği 64'ten daha çok olan tümceler ya da herhangi bir biçimbirim patikası için 5000'den daha çok blok üreten tümceler

[11] https://opennlp.apache.org/; erişim: 04.05.2017

elenmiştir. Yapılan çeşitli denemeler sonucunda belirlenen bu sayılar tümce çözümlemesi açısından çalışma zamanının kabul edilebilir düzeylerde kalmasını sağlamaktadır. Üçüncü aşamada ise elde edilen tümceler içinde TMoST tarafından tanınmayan, dolayısı ile hiç sözdizim ağacı üretilemeyen tümceler ve yalnızca bir adet sözdizim ağacı üretilen tümceler elenmiştir. Sözü edilen filtrelemeler sonucunda elde edilen temel tümcelerine ilişkin genel istatistikler **Tablo 6.1**'de verilmiştir.

Tablo 6.1. *Tümce Filtrelemesi*

	Tümce Sayısı	**Ortalama Simge Sayısı**
OD (Kaynak)	167034	13,98
Tümce Filtrelemesi	100011	7,69
AUT Temel Tümceleri	10000	4,46

Tümce filtrelemesi sonucunda elde edilen tümcelerin ortalama simge sayısının özgün OD tümcelerinin ortalama simge sayısından kayda değer derecede küçük olmasının nedeni, noktalama işaretleri ve rakamları eleyen simge filtrelemesinden çok, tümce filtrelemesindeki 20'den çok simge içeren tümcelerin elenmesinde aranmalıdır. Zira OD içerisinde 20'den fazla simge içeren 33086 adet tümce bulunmaktadır.

AUT temel tümceleri için filtreleme değişkenlerine ilişkin bir betimleyici istatistiksel inceleme **Tablo 6.2**'de verilmiştir.

Tablo 6.2. Temel Tümceler İçin Filtreleme Değişkenlerine İlişkin İstatistikler

	Enküçük	**Enbüyük**	**Ortalama**	**Standart Sapma**
Simge Sayısı	1	12	4,46	1,74
Biçimbilimsel Belirsizlik	1	64	18,91	16,41
Blok Sayısı	2	4878	229,70	491,25
Sözdizimsel Belirsizlik	2	304	6,12	10,65

AUT temel tümcelerinin tümü sözdizimsel imleme için kullanılabilir durumdadır ancak bu ölçüde bir AYD inşa etmek emek-yoğun bir iş olduğu için tez çerçevesinde 10'dan fazla sözdizim ağacı üreten tümcelerin öncelikli olarak işlenmesine karar verilmiş ve bu amaç doğrultusunda 1158 adet tümce incelenmiştir. İlgili istatistikler **Tablo 6.3**'te verilmiştir.

Tablo 6.3. *İmlenen Tümcelere İlişkin İstatistikler*

Geçerli	**510**	**%44**
Geçersiz	451	%39
Deyim içeren	138	%12
Bozuk tümce	59	%5

İşlenen 1158 adet tümcenin tamamı TMoST tarafından çözümlenebilen ve en az 11 sözdizim ağacı üreten tümcelerdir. Bunlardan bazıları (%44) için üretilen sözdizim ağaçlarından biri tümcenin geçerli sözdizimsel çözümlemesi olarak belirlenebilmişken, bazıları (%39) için üretilen tümce ağaçlarının hiçbiri geçerli değildir. Geçersizliğe yol açan etkenler, tümcede özel ad, kısaltma vb. unsurların yer alması, tümcenin bağlaçla başlaması, eylemin yanulam bilgisindeki eksiklik ve dizge tarafından işlenemeyen bazı yapıların bulunmasıdır. Bu noktada, dizge tarafından işlenemeyen yapıları içeren bir tümcenin nasıl çözümlenebildiği sorulabilir. Buna yanıt verirken yanlış alarm kavramından yardım istemek faydalı olacaktır: Bazı durumlarda dizge aslında çözümlenemez olarak bildirmesi gereken bir tümceyi yanlış yorumlayarak çözümlenebilir bir tümce gibi değerlendirebilmektedir. Bu durum, sonuçlar insan gözüyle incelenmeden ortaya çıkamamaktadır.

7. MODEL

Bu bölümde Bölüm 8'de açıklanan deneylerde kullanılan modeller tanıtılacaktır. Bu modeller üç temel grupta toplanabilir. Birinci grup, istatistiksel dil modelleri; ikinci grup, olasılıksal bağlam bağımsız dilbilgisi (OBBD) ve öbek olasılık modeli; üçüncü grup ise ilişkisel modellerdir. İstatistiksel dil modelleri arasında *kanal* ve *rol* adlı modeller; öbek olasılık modeli ve ilişkisel modeller bildiğimiz kadarıyla ilk kez bu tezde önerilmiş, özgün dil modelleridir.

7.1. İstatistiksel Dil Modeli

Dil modeli ifadesi konuşma tanıma (speech recognition) çalışmalarından gelmektedir. Konuşma tanıma için söz konusu olan temel denklem aşağıda görüldüğü gibidir:

$$\widehat{W} = \underset{W}{\operatorname{argmax}} \frac{P(A|W).P(W)}{P(A)} \qquad \textbf{(7.1)}$$

Bu denkleme göre, olasılıkların oranını en büyükleyen W sözcük dizisi aranmaktadır. Bayes kuralının konuşma tanımaya uyarlanmış bu biçiminde geçen P(W), W sözcük dizisinin (tümcenin) üretilme olasılığı şeklinde tanımlanır (Jelinek, 1976, s. 538). P(W), Bayes kuralında önsel olasılıklara karşılık gelir. Bir dil modelinin amacı bu önsel olasılıkları kestirmektir (Bahl vd., 1989, s. 1001).

Dili modellemek, yapay problemler için sonlu durum makineleri ve bağlam bağımsız dilbilgisi kullanarak kolaylıkla gerçekleştirilebilirken, doğal dil problemlerinde P(W) olasılığını kestirmek çok daha zordur (Bahl vd., 1983, s. 180).

P(W) aslen birleşik olasılıktır ve bu olasılık W sözcük dizisinin tümünün birlikte gözlenme olasılığı şeklinde ifade edilebilir. Dizideki bütün sözcüklerin bir arada gözlendiği birçok örneği bulmayı gerektirdiği için pratikte hesaplanması mümkün değildir. Bu nedenle birleşik olasılığa yakınsayan Bayes zincir kuralı (chain rule) kullanılır:

$$P(W_1, W_2, \dots, W_m) = P(W_m|W_1, \dots, W_{m-1})P(W_1, \dots, W_{m-1})$$
$$= P(W_m|W_1, \dots, W_{m-1})P(W_{m-1}|W_1, \dots, W_{m-2})P(W_1, \dots, W_{m-2})$$
$$\dots$$
$$\prod_{i=1}^{m} P\left(W_i \middle| \cap_{j=1}^{i-1} W_j\right) \quad \textbf{(7.2)}$$

Dikkat edilirse, Bayes zincir kuralını uygulamanın da pratik hesaplama problemini tam olarak çözemediği görülür. Bu noktada, bir sözcüğün gözlenme olasılığının ondan önceki birkaç sözcükten etkilendiğini öne süren Markov kabullenmesi yapılırsa şu denkleme ulaşılır:

$$\prod_{i=1}^{m} P\left(W_i \middle| \cap_{j=i-n+1}^{i-1} W_j\right) \quad \textbf{(7.3)}$$

Bu denklemdeki n değişkeni odak sözcüğün (W_i) gözlenme olasılığının kendinden önceki kaç sözcükten etkilendiğini belirleyen parametredir ve n-gram dil modelindeki n'ye karşılık gelir. n'nin 1 olduğu dil modeli tekli (unigram), 2 olduğu dil modeli ikili (bigram) ve 3 olduğu dil modeli üçlü (trigram) olarak adlandırılır. Son denklemin en çok kullanılan bu üç dil modeli için uyarlanmış biçimleri şöyledir:

$$tekli\ model: \prod_{i=1}^{m} P(W_i)$$
$$ikili\ model: \prod_{i=1}^{m} P(W_i|W_{i-1})$$
$$\text{üçlü}\ model: \prod_{i=1}^{m} P(W_i|W_{i-1}, W_{i-2}) \quad \textbf{(7.4)}$$

$P(W_i|W_{i-1})$ koşullu olasılığını hesaplarken en çok olabilirlik kestiricisi (Maximum Likelihood Estimate) kullanılır. Buna göre $P(W_i|W_{i-1})$ koşullu olasılığı şöyle verilir:

$$P(W_i|W_{i-1}) = \frac{C(W_{i-1}, W_i)}{C(W_{i-1})} \quad \textbf{(7.5)}$$

Burada $C(W_{i-1}, W_i)$ ifadesi önce W_{i-1} sonra W_i sözcüklerinin gözlenme sayısı iken $C(W_{i-1})$ ifadesi W_{i-1} sözcüğünün, öncesinde veya sonrasında hangi sözcüğün bulunduğu fark etmeksizin, gözlenme sayısını simgeler.

Tablo 7.1'de kurgusal bir dilden elde edildiği varsayılan bir derlem görülmektedir. Bu dilin sözlüğünde yalnızca beş sözcük bulunsun: a, b, c, d ve e. Derlemdeki her bir satır bir tümceye karşılık gelir.

Tablo 7.1. *Kurgusal Bir Derlem*

aabaeb
cbddade
deeabee
cdbcabb
bbeeba
addc
ccedcde
ecb
baedc
ecеceedd

Tablo 7.1'deki derleme göre, sözgelimi "baba" ve "dede" tümcelerinin gözlenme olasılıkları ikili dil modeline dayanarak hesaplanmak istensin:

$$P(baba) = \prod_{i=1}^{4} P(W_i|W_{i-1})$$

$$= P(W_1)P(W_2|W_1)P(W_3|W_2)P(W_4|W_3)$$

$$\frac{C(b)}{C(.)}\frac{C(ba)}{C(b)}\frac{C(ab)}{C(a)}\frac{C(ba)}{C(b)} = \frac{12}{60}\frac{3}{12}\frac{3}{9}\frac{3}{12} = \frac{1}{240} \cong 0.0042$$

$$P(dede) = \prod_{i=1}^{4} P(W_i|W_{i-1})$$

$$= P(W_1)P(W_2|W_1)P(W_3|W_2)P(W_4|W_3)$$

$$\frac{C(d)}{C(.)}\frac{C(de)}{C(d)}\frac{C(ed)}{C(e)}\frac{C(de)}{C(d)} = \frac{12}{60}\frac{3}{12}\frac{3}{16}\frac{3}{12} = \frac{3}{1280} \cong 0.0023$$

Yukarıda verilen örnekte gerçekleştirilen sayım işlemi (C) uygulamada eğitim (train) aşamasına karşılık gelir. Pratik olması için bu sayım işlemi derlem üzerinde bir

kez gerçekleştirilir ve saklanır. Modelden daha önce karşılaşmadığı bir ifadenin olasılığını talep etmek de sınama (test) aşamasını karşılar.

İstatistiksel dil modelleri farklı hesaplama birimlerine dayalı olarak oluşturulabilir. En sık kullanılan hesaplama birimleri sözcük ve harflerdir. Ancak hece ve biçimbirimler de hesaplama birimi olarak kullanılabilmektedir.

7.1.1. Dil modelinin değerlendirilmesi

Dil modellerini değerlendirmek için en çok kullanılan ölçütler düzensizlik (entropy) ve şaşırmadır (perplexity) (Tür, 2000, s. 16). Kesikli bir rastgele değişkeni ele alalım. Bu, sonlu sayıda değer alabilen bir değişkendir. Söz gelimi kesikli rastgele değişken (K) tümcenin ilk kurucu bileşeni olma şeklinde tanımlansın. Olası kurucu bileşenler K'nın alabileceği değerler listesidir: özne, tümleç, eklenti ve eylem. Burada olasılık, K'nın belli bir değeri alıp almayacağına ilişkin sorulan sorunun ayrıntılı bir yanıtıdır. Örneğin, "K'nın özne olma olasılığı 0,37'dir" veya "K'nın eylem olma olasılığı sıfırdır" vb. ifadeler kullanabiliriz. K'nın bütün olası değerleri olaylara, değerlerin olasılıklarından oluşan liste de olasılık dağılımına denk düşer. Olasılık aslen bir değişkenin bir değeri almasına ilişkin belirsizliğin ölçüsüdür. "K'nın özne olma olasılığı 1'dir." dendiğinde aslında belirsizliğin sıfır olduğu ima edilir ve bu da kesinlik anlamına gelir. Düzensizlik ise değişkenin bir değeri alma olasılığının 2 tabanındaki logaritmasının -1 ile çarpımına eşittir[12]. Kesikli bir rastgele değişkenin tek değeri alma olasılığı o değer için bir tür belirsizlik olarak düşünülürse, değişkenin alabileceği bütün değerlerin düzensizliklerinin matematiksel beklentisi dağılımın düzensizliğine karşılık gelir (Şamilov, 2015, s. 3). Düzensizliğin temel denklemi şöyledir:

$$H(X) = -\sum P(X) log_2 P(X) \qquad \textbf{(7.6)}$$

Şaşırma, dil modellerini karşılaştırmak için sıkça kullanılan bir ölçüdür. Şaşırmanın k olması, her adımda seçilebilecek k adet eşit olasılıklı seçenek anlamına gelir ve aşağıdaki denklemle ifade edilir:

[12] Düzensizliğin matematiksel kanıtı Şamilov'dan (2015, s.6) incelenebilir.

$$Per(X) = 2^{H(X)} \quad \textbf{(7.7)}$$

İki dil modelini karşılaştırırken şaşırma ölçüsü daha küçük olan modelin daha başarılı olduğu kabul edilir.

7.1.2. Kullanılan dil modelleri

Bu tezde farklı hesaplama birimlerinin kullanıldığı ve iki grupta toplanan beş adet dil modeli oluşturuldu. Bu modelleri elde etmek için imlenmiş dilbilimsel veriye ihtiyaç bulunmaktadır. Fakat tezde söz konusu ihtiyacı karşılamak amacıyla gözetimsiz yöntemler kullanılmıştır. Başka bir deyişle, tezde üretilen hiçbir dil modeli için imlenmiş verilere başvurulmamıştır.

Bütün dil modelleri tekli, ikili ve üçlü olarak oluşturulmuş ve toplam 15 dil modeli biçimi elde edilmiştir. Dil modelleri için bir sezgisel varsayım uygulanmıştır. Buna göre, eğer model, verilen ifade için belli bir düzeyde birden çok olasılık üretiyorsa bunlardan en büyüğü seçilir ve ilgili olasılık bir üst düzey için kullanılır. Örneğin, "Ata bindi." tümcesinin sözcükleri biçimbilimsel belirsizlik (sözcük düzeyinde) içermektedir. Bunun sonucu olarak ilgili tümce için birçok yorum ortaya çıkar. Dil modeli bu yorumların her biri için ayrı olasılıklar hesaplayabilir. Bir üst düzeyin (yüzeybiçim), başka bir deyişle "Ata bindi." tümcesinin olasılığı bu olasılıkların en büyüğü olacaktır.

7.1.2.1. Sözcük tabanlı dil modelleri

Sözcük tabanlı dil modelleri, biçimbirim, gövde, son ek ve kanal olmak üzere dört adettir. Bu modeller tümcelerin biçimbilimsel çözümlemesi yapılmış sözcüklerinden elde edildiği için sözcük tabanlı dil modeli adı verilmiştir.

Biçimbilimsel çözümleme esas olarak bir sözcüğün biçimbilimsel yapı taşları olan biçimbirimleri ayrıştırmayı hedefler. Bu işlem sonucunda elde edilen temel ürünler biçimbirimlerdir. Biçimbirim dil modeli sözcüklerin bütün biçimbirimlerinin oluşturduğu silsileden elde edilmektedir. Biçimbilimsel çözümleme ile bazı yararlı yan

ürünler de ortaya çıkar. Bunlardan ikisi gövde ve son ektir. Gövde, bir sözcüğün çekim eklerinden arındırılmış biçimidir. Bu birim, sondan eklemeli bir dil olan Türkçe için önemli bir enformasyon içerir. Simge dil modeline[13] kıyasla daha az seyrek bir dil modeli olması beklenir. Çünkü simge dil modelinde örneğin *gel* eyleminin bütün çekimleri (geliyor, geliyorlar, geldi, geldiniz, vs.) hesaplanırken, gövde dil modelinde bu eylem yalnızca *gel* gövdesi ile temsil edilir. Dolayısıyla gövde dil modelinde eğitim derleminde *gel* gövdesine bir kez rastlanılması ilgili boşluğu doldurmaya yetecekken, seyrekliği önlemek için simge dil modelinde bütün çekimlerin gözlenmesi gerekir. Bu da tek bir eylem gövdesinden binlerce yüzey biçim üretilebilen Türkçe gibi bir dil için çok zordur.

Son ek dil modeli, sözcüklerin sonunda yer alan eklerden oluşan diziden elde edilir. Eğer sözcük tek biçimbirimden oluşuyorsa "0" şeklinde yapay bir son ek varmış gibi düşünülür. Sözcüğün sonundaki ek, Türkçe için konuşursak, yapım eki veya çekim eki olabilir. Bu tezde sözdizimsel olarak bir etkisi olmayan saf yapım ekleri gövdenin içinde bırakıldığından sözcüğün sonundaki ekin sözdizimsel bir role sahip olduğu varsayılır. İşte bu dil modeli bu varsayımın sınanması için kurulmuştur.

Kanal dil modeli biçimbilimsel çözümleyicinin sözcük için ürettiği biçimbilimsel kanal verilerinden elde edilir.

7.1.2.2. ***Tümce tabanlı dil modeli***

Tümce tabanlı dil modeli olarak rol adlı model üretilmiştir. Bu model tümce çözümlemesi sonucunda ortaya çıkan sözdizim ağaçlarından elde edildiği için tümce tabanlı dil modeli adı verilmiştir.

Rol dil modeli, verilen tümcenin kurucu bileşenlerinin türlerinden oluşan diziden elde edilir. Örneğin "Dün güzel bir maç izledim." tümcesi için kurucu bileşenler dizisi, eklenti (dün), tümleç (güzel bir maç) ve eylemdir (izledim).

[13] Simge dil modeli bir tümcenin bütün sözdizimsel yorumları için aynı değeri üretir. Bu nedenle ilgili model tez kapsamında incelenmedi.

7.1.3. Bulgular

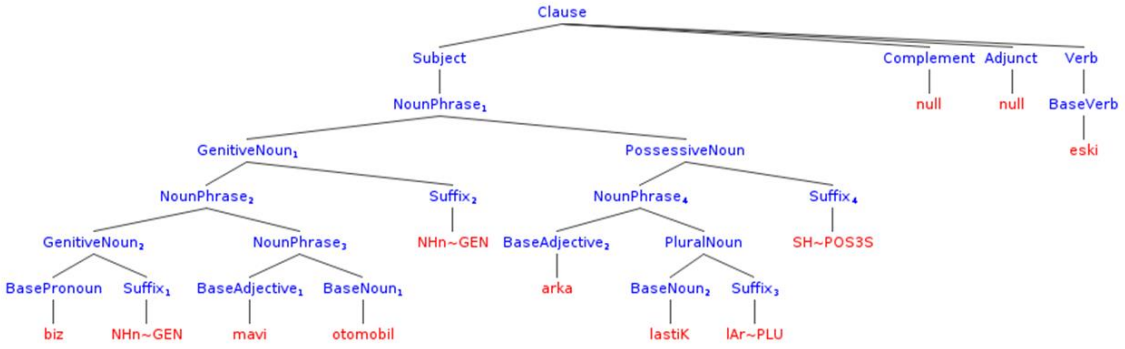
Tablo 7.2'de, istatistiksel dil modellerinin aldığı şaşırma değerleri sunulmaktadır. Şaşırmanın hesaplandığı veri Bölüm 6'da tanıtılan metin koleksiyonundan çekilen rastgele 100.000 adet tümceden oluşmaktadır.

Tablo 7.2. *Dil Modelleri ve Şaşırma Değerleri*

Model	Şaşırma
Gövde_1	2432
Gövde_2	401
Gövde_3	8170
Son Ek_1	19
Son Ek_2	8
Son Ek_3	9
Biçimbirim_1	30
Biçimbirim_2	7
Biçimbirim_3	7
Kanal_1	34
Kanal_2	13
Kanal_3	18
Rol_1	5
Rol_2	2
Rol_3	2

7.2. Olasılıksal Bağlam Bağımsız Dilbilgisi ve Öbek Olasılık Modeli

Olasılıksal Bağlam Bağımsız Dilbilgisi (probabilistic context free grammar, OBBD), bir tümcenin sözdizim ağaçlarına olasılıksal bağlam-bağımsız dilbilgisi esasına göre olasılıklar atayan ve alanyazında sıkça yararlanılan bir modeldir. Tezde önerilen öbek olasılık modeli ise, bir sözdizim ağacındaki öbek yapıların baş ve kuyruk kısımlarını temsil eden sözlükbirimlerin art arda gelme olasılığını hesaplar. Bu model, sözcükler arası uzaklıktan bağımsız olan bir tür anlamsal ilişki belirleyicidir. Aşağıdaki şekilde bu iki modeli örneklemek için "Bizim mavi otomobilin arka lastikleri eskidi" tümcesinin bir çözümlemesi görülmektedir:



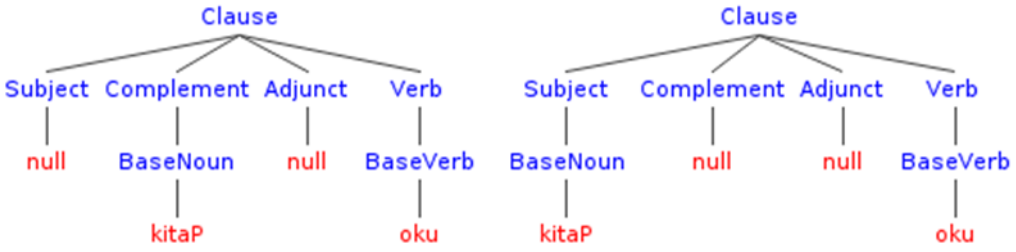
Şekil 7.1. OBBD ve Öbek Olasılık Modeli İçin Bir Örnek Sözdizimsel Çözümleme

OBBD modelinde ağacın yaprakları ve düğümleri arasındaki her bir ikili geçişin meydana gelme olasılığı hesaplanır ve bütün olasılıklar çarpılarak ağacın bütününün yapısal olasılığı bulunur. Öbek olasılık modelinde ise sözdizim ağacındaki öbek yapıların baş ve kuyruk kısımlarını temsil eden sözlükbirimlerin art arda gelme olasılığını hesaplanır. Örneğin NounPhrase_1 öbeği için *otomobil* ve *lastik* sözlükbirimlerinin art arda gelme olasılığı arada başka birimler olsa bile elde edilebilir. Bu model, birimler için uzaklıktan bağımsız bir tür anlamsal ilişki belirleyicidir.

7.3. İlişkisel Modeller

Üçüncü grup, bir tümce için üretilen her bir tümcecik üzerinde kurucu bileşenlerin ve bazı yapısal ögelerin marjinal, birleşik ve koşullu olasılıklarının tümcecik sayısının etkisini ortadan kaldıracak şekilde normalleştirilerek tümceyi temsil edecek birer model olmasıyla meydana gelir. Bu modeller şunlardır: birleşik durum ekli tümleç-eylem, birleşik özne-eylem, eylem, koşullu durum ekli tümleç-eylem, koşullu özne-eylem, özne, tümleç-eylem.

Örneğin “Kitap okudu” tümcesinin iki yapısal yorumu vardır:



Şekil 7.2. *İlişkisel Modeller İçin Örnek Sözdizimsel Çözümlemeler*

Birleşik özne-eylem ilişkisel modeli kitap ile oku arasındaki özne-eylem ilişkisi olasılığını belirlerken, tümleç-eylem ilişkisel modeli kitap ile oku arasındaki tümleç-eylem ilişkisi olasılığını belirler.

8. SÖZDİZİMSEL BELİRSİZLİK GİDERME

Bu bölümde tezin temel amacı olan sözdizimsel belirsizlik giderme (SBG; syntactic disambiguation) için önerilen modellerin sınandığı çeşitli deneyler ve sonuçları ayrıntılı şekilde sunulacaktır. Buna geçmeden önce, doğal dil işlemede belirsizlik kavramının tanımı ve türleri verilecek, ardından SBG ile ilgili alanyazın çalışmalarına değinilecek ve daha sonra ise tezde önerilen yöntem açıklanacaktır.

Hong (2014, s. 10), dildeki belirsizlik (ambiguity) için kapsamlı bir tanım verir: "Bir sözcük, terim, notasyon, işaret, sembol, öbek, tümce veya iletişim amaçlı kullanılan herhangi bir biçim, eğer birden çok şekilde yorumlanabiliyorsa belirsizdir." Belirsizlik biçimsel diller (formal languages) açısından istenmeyen bir durumdur çünkü bir dizgeyi biçimsel dil aracılığıyla yönetirken kesinlik ve belirlilik sağlamak esastır. Aksi takdirde dizgenin kararlı biçimde çalışması mümkün olmaz. Öte yandan doğal dillerde belirsizlik yaygın bir olgu olarak karşımıza çıkar. Buna, yazımsal düzeyde (kar: kar-kâr), sözcük düzeyinde (saç: ad/eylem), öbek düzeyinde (kırmızı kalem kutusu: kalem kırmızıdır/kutu kırmızıdır) ve tümce düzeyinde (kedi gördü: özne kedi/özne <boş>) çok sayıda örnek verilebilir. Doğal dil konuşucuları biçimsel dillerde olduğu gibi belirsizlikleri gidermeye çalışmaz. Aksine, en az çaba ilkesi gereği, olabildiğince ekonomik davranarak belirsizliğe yol açar. Elbette konuşucuların amacı belirsizlik üretmek değildir. Ancak belirsizlik dilin karakteristik özelliklerinin yanı sıra yazımda daha kolay sembolleri seçme, daha az sözcük kullanma, daha kısa tümceler kurma gibi davranışlardan etkilenmektedir. Sonuç olarak, belirsizlik, dilde farklı düzeylerde gözlenebilir, konuşucuların tercihlerinden etkilenir ve dilin biçimsel, yapısal ve anlamsal özelliklerinin doğal bir sonucudur.

Sözlüksel belirsizlik (lexical ambiguity), bir sözcüğün sözlükte bulunan farklı tanımları arasından geçerli olanını seçme problemidir. Sözlüksel belirsizlik bağlam-bağımlıdır. Öyle ki aynı dilbilimsel öge (bir sözcük, öbek ya da tümce) bir bağlamda belirsizken başka bir bağlamda belirsiz olmayabilir (Hong, 2014, s. 10).

Sözlüksel belirsizlik gibi biçimbilimsel belirsizlik de (morphological ambiguity) sözcük düzeyinde ortaya çıkan bir belirsizlik türüdür. Bu belirsizliği özellikle zengin

biçimbilimsel özelliklere sahip olan diller için tanımlamak daha kolaydır. Türkçede iyelik eki ile ilgi durum ekinin biçimciklerindeki kesişme, sonu ünsüz bir harfle biten ad gövdelerinde biçimbilimsel belirsizlik üretir. Örneğin, "kitabın" sözcüğü tek başına ele alındığında Ad+İyelik Eki veya Ad+İlgi Durumu biçiminde çözümlenebilir. Bu örnekte "ın" biçimciği hem iyelik ekinin hem de ilgi durum ekinin üyesidir.

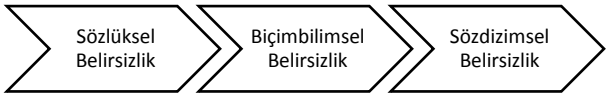
Cruse (1986), bir belirsizlik sınıflandırması önermiştir. Bu sınıflandırmaya göre tümce düzeyinde dört tip belirsizlik gözlenebilir:

1. Tam sözdizimsel belirsizlik: *Heyecanlı öğrenciler ve aileleri* – 1. yorum: Heyecanlı öğrenciler ve onların heyecanlı aileleri; 2. yorum: Heyecanlı öğrenciler ve onların heyecanlı olmayan (ya da olup olmadıkları bilinmeyen) aileleri.
2. Yarı sözdizimsel belirsizlik: *Kırmızı kalem* – 1. yorum: Kalemin dış kaplaması kırmızı renkte; 2. yorum: Kalemin ucu veya mürekkebi kırmızı renkte.
3. Sözlüksel-sözdizimsel belirsizlik: *Yemek istedi* – 1. yorum: Bir şey yemek istedi; 2. yorum: Bir yemek istedi.
4. Tam sözlüksel belirsizlik: *Cildini tazelemiş* – 1. yorum: Kitabın cildini tazelemiş; 2. yorum: Hastanın cildini tazelemiş.

Small vd. (1987, s. 3), sözlüksel belirsizliği sözdizimsel ve anlambilimsel olmak üzere iki türe ayırır. Buna göre sözdizimsel belirsizlik (syntactic ambiguity), ad, eylem vb. sözcük kategorileri ile ilişkili olan belirsizliktir. Örneğin, *kap* sözcüğü hem ad hem de eylem olarak gözlenebilir. Small vd. (1987, s. 3) anlambilimsel belirsizliği de (semantic ambiguity) kendi içinde iki türe ayırır: çokanlamlılık ve eşseslilik. Çokanlamlılık (polysemy) bir sözcüğün birbiriyle ilişkili birden çok anlama sahip olmasıdır. Örneğin, uçmak eylemi "Yavru serçe yuvadan uçtu." ve "Tabaktaki meyveler uçmuş." tümcelerinde birbiriyle ilişkili anlamlarda kullanılır. Buradaki ilişki eğretileme (metafor) yöntemiyle kurulmuştur: Birinci tümcedeki anlam bir canlının vücudunu kullanarak havalanması ve hareket etmesi iken ikinci tümcedeki anlam birinci anlamdan eğretileme yoluyla (meyveler uçabilen bir canlıya benzetilerek) elde edilir. Eşseslilik (homonymy) ise birbiriyle ilişkili olmayan anlamlara sahip sözcüklerin aynı biçimsel gösterimde olmasıdır. Örneğin, *bar* sözcüğü için eşseslilik kaynaklı anlam

ayrımları vardır: “Ağırlıkları bara geçirdi.”, “Barın karşısındaki otelde kalıyorum.”, “Cihazın basıncı iki bar oldu.” vb.

Small vd. (1987, s. 3) yapısal belirsizliği (structural ambiguity), öbek yapı ağaçları üzerindeki belirsizlik şeklinde tanımlar ve sözdizimsel belirsizlikten farklı olarak değerlendirilmesi gerektiğini belirtir. Bu tezde ise sözdizimsel belirsizlik yapısal belirsizliği de içine alan geniş bir kavram olarak yorumlanmıştır. Türkçe için konuşmak gerekirse, bir tümcenin belirsizliği onu oluşturan sözcüklerdeki biçimbilimsel belirsizlikten etkilenir. Çok açıktır ki özellikle çekim ekleri sözdizimsel yapıların kurulmasında rol alır ve bunların hangisinin ne şekilde seçileceği de biçimbilimsel belirsizliğin kapsamındadır. Aynı şekilde biçimbilimsel belirsizlik de sözlüksel belirsizlikten etkilenir. Öyleyse sözdizimsel belirsizlik, anılan belirsizliklerin oluşturduğu zincirin sonunda yer almalıdır (bk. **Şekil 8.1**). Zincirin sağ tarafı tümce düzeyinin ötesindeki belirsizliklere doğru uzanmaktadır. Ancak bunlar tezin kapsamı dışındadır.



Şekil 8.1. *Belirsizlik Türleri Arasındaki İlişki*

SBG, alanyazında, bir tümceye ait birden çok yapısal yorum arasından verilen bağlam için geçerli olan yapıyı belirlemek şeklinde ele alınmaktadır. Eğer bağlam bilinmiyorsa yapısal olarak mümkün olan bütün yorumlar geçerli olacaktır. Örneğin “Kitabını aldım.” tümcesi için bağlam bilgisi verilmesin. Bu durumda tümcede tamlayanın düşürülmesinin yol açtığı *senin kitabın/onun kitabı* belirsizliği çözümlenemez ve her iki yorum da geçerli olur.

8.1. Önceki Çalışmalar

SBG konusundaki öncü çalışmalardan biri Basili vd. (1991) tarafından yapılmış olup bu çalışmada saf istatistiksel analiz yoluyla türetilen sözcük birliklerinin SBG problemiyle başa çıktığı iddia edilmiştir. Sözcük birliklerinin iki yöntemle elde edildiği çalışmada, birinci yöntemde ad, eylem, sıfat gibi içerik sözcüklerinin sıralı çiftleri toplanmış; ikinci yöntemde ise bir sözdizimsel çözümleyicinin çıkarttığı sözcük çiftleri sıralanmıştır. Birinci yöntemde ±5 pencere kullanılırken ikinci yöntemde sözdizimsel belirsizlikle karşılaşıldığında belirsizlik içeren birliklerin tümü üretilmiştir. Çalışmada ayrıca etki alanı (domain) bilgisine de değinilmiştir: "Örneğin ağaç sözcüğü bilgisayar etki alanında ayrıştırma, gramer, karar gibi sözcüklerle ilişkiliyken, başka bir etki alanında zeytin veya yılbaşı sözcükleriyle ilişkili olabilir." Bu makalede yalnızca verilen dilbilimsel etki alanından toplanmış sınırlı veriden faydalanmak için sözcüklere anlambilimsel etiketler (ACTIVITY, ARTIFACT, HUMAN_ENTITY, MATERIAL vb.) atanmış ve bu etiketler toplanan birlikleri kümelemek amacıyla kullanılmıştır. Sözdizimsel belirsizliği nasıl çözdüklerini şu örnekle açıklamışlardır: "(1) To sell wine in bottles (2) To sell wine to the customers. Birinci tümcede sell-in-ARTIFACT birlikte gözlenme olasılığı wine-in-ARTIFACT'a göre önemli derecede düşüktür. İkinci tümcede sell-to-HUMAN_ENTITY birlikte gözlenme olasılığı wine-to-HUMAN_ENTITY'ye göre önemli derecede yüksektir." Önceki çalışmalardan farklı olarak sözcük birliklerini belirlemede basit birlikte gözlenmeler yerine sözdizimsel birlikler (verb-object: sell-wine; noun-preposition-noun: wine in barrel) olarak adlandırdıkları çiftleri ve üçlüleri kullanmışlardır. Özellikle ilgeçleri hesaba katmışlardır. Bu konuyla ilgili şu yorumu yapmaktadırlar: "sözdizimsel bağlantı tipi ve varsa bağlayıcı ilgeç, birlikte gözlenen sözcüklerin anlambilimsel ilişkisinin doğası üzerinde önemli kısıtlamalara işaret eder: 'to sell flats ile to sell...in flats' farklı sözdizimsel birliklerdir."

Lee vd. (2003) tarafından yapılan çalışma, E-TRAN 2001 isimli İngilizce-Korece etki alanı-bağımsız (domain-independent) makine çeviri sisteminde kullanılmak üzere geliştirilen kural tabanlı SBG yöntemini açıklamaktadır. Makalede, PATI (Parser's Ambiguity Type Information) adını verdikleri bir bilgi yapısını önermişlerdir. Bu yapı

sözdizimsel çözümleyicinin ürettiği aday ağaçlardaki belirsizlik tiplerini tanılamak için kullanılmaktadır. PATI, eğitim derleminden gelen kural tercih skorlarını hesaplamakta ve elde edilen bilgiyi dilbilgisi kurallarına dönüştürmektedir. Bir tümceden türetilen birden çok sözdizim ağacı içinden doğru olanı seçmek doğru ağaçtaki kurallar kümesi ile doğru olmayan ağaçlardaki kurallar kümesi arasında bir önceliklendirme bilgisi üretmektedir. Çalışmanın temel amacı bu bilgiyi eğitim derleminden çıkarmaktır. Sonuç olarak SBG başarımının arttığı rapor edilmiştir. Ayrıca çalışmanın ek bölümünde verilen belirsizlik sınıflandırması incelemeye değerdir.

Toutanova vd. (2005) yaptıkları çalışmada doğru tümce çözümlemesini seçme amaçlı olasılıksal modeller geliştirmiştir. Modeller baş odaklı öbek yapı dilbilgisine (Head-Driven Phrase Structure Grammar: HPSG) dayanan Redwoods ağaç yapılı derleminden elde edilmiştir. Çalışmada mevcut olasılıksal yöntemlere ek olarak yeni yöntemler de önerilmektedir. Çalışmanın sonucunda elde edilen en iyi başarım değeri %76 tam tümce doğruluğudur.

İncelediğimiz kadarıyla Türkçenin konu edildiği gözetimli ya da gözetimsiz bir SBG çalışması bulunmamaktadır. Bir sonraki başlıkta bu tezin temel çalışması olan gözetimsiz SBG yöntemi tanıtılacaktır.

8.2. Önerilen Yöntem

SBG için önerdiğimiz yöntem, TMoST'un tümce için ürettiği sözdizim ağaçlarının sıralanmasında Bölüm 7'de tanıtılan modellerin puanlayıcı olarak kullanılmasından ibarettir. Modeller bir sözdizim ağacı verildiğinde ağacı oluşturan biçimbilimsel, sözdizimsel ve ilişkisel birçok işareti değerlendirerek bir olasılık üretir. Modellerin elde edilmesinde dilbilgisel olarak işaretlenmemiş metinler (metin koleksiyonu) kullanıldığı için gerçekleştirilen SBG işlemi gözetimsiz olmaktadır.

Deneyler için SBG işinin başarımını ölçmek amacıyla altın standart veri olarak AUT'den (bk. **Bölüm 6.2.2**) yararlanıldı. Deneylerde dört tür başarım ölçüsü kullanıldı: ortalama, bağıl ortanca, benzerlik ve bağıntı. Ortalama ve bağıl ortanca, modelin tümceler için gerçekleştirdiği ağaç sıralamalarını esas alır. Model, tümce ağaçlarını en

olası ağaçtan en az olası ağaca doğru sıralar. Bu sıralamada, geçerli ağacın (altın standart) bulunduğu konum, sıralama skorunu verir. Ortalama başarımı her bir tümce için üretilen bu sıralama skorlarının aritmetik ortalamasıdır:

$$Ortalama = \frac{\sum_{i=1}^{T} S_i}{T} \quad \textbf{(8.1)}$$

Denklem 8.1'de S_i i. tümce için elde edilen sıralama skorunu, T ise toplam tümce sayısını simgeler. Ortalamada, tümceden üretilen ağaç sayısı ne olursa olsun, geçerli ağacın sıralamadaki konumu dikkate alındığı için, bu başarım, tümce ağaç sayısından etkilenir. Bağıl ortancada ise sözü edilen bu etki, her bir tümce için elde edilen sıralama skorunun tümcenin ağaç sayısına bölünmesiyle ortadan kaldırılmaktadır. Tümceler için elde edilen bu bağıl skorların ortancası alınarak bütün tümce derlemi için tek bir skor elde edilir. Elde edilen bu skor sıralama cinsinden bir ölçü olmadığı için tümce ağaç sayılarının ortancası ile çarpılır:

$$Bağıl\ Ortanca = Ortanca(\frac{S_i}{A_i}) x Ortanca(A_i) \quad \textbf{(8.2)}$$

Denklem 8.2'de S_i i. tümce için elde edilen sıralama skorunu, A_i ise i. tümce için üretilen sözdizim ağacı sayısını simgeler. Bağıl ortanca başarımında ortalama yerine ortanca kullanılmasının nedeni ölçümün sağlam (robust) hâle getirilmek istenmesidir.

Benzerlik başarımı, ortalama ve bağıl ortanca başarımlarından farklı olarak bir sıralama skoru ile hesaplanmaz. Bunun yerine modelin önerdiği en olası ağaç ile geçerli ağaç arasındaki benzerlik ölçülür. İdeal durumda model zaten geçerli ağacı birinci sırada önereceği için benzerlik 1 olacaktır. Diğer durumlar için ise 0 ile 1 arasında skorlar elde edilir:

$$Benzerlik = \frac{\sum_{i=1}^{T} Benzerlik(Ç_{i,1}, G_i)}{T} \quad \textbf{(8.3)}$$

Denklem 8.3'te $Ç_{i,1}$ i. tümce için dizgenin önerdiği birinci çözümlemeyi ifade eden sözdizim ağacını, G_i i. tümce için geçerli (gold) sözdizim ağacını, T ise toplam tümce sayısını simgeler. Bu başarım yalnızca en iyi öneriyi değerlendirdiği için katı bir ölçüt gibi görünse de dereceli bir benzerlik değeri üretmesi bakımından modeli yorumlamayı kolaylaştırır.

Son başarım ölçüsü olan bağıntı, melez bir ölçü olup, modelin sıraladığı ağaçların benzerlik dizilimi ile ideal benzerlik dizilimi arasındaki bağıntıdır (korelasyon) ve -1 ile 1 arasında değişir:

$$Bağıntı = \frac{\sum_{i=1}^{T} Bağıntı(O_i, I_i)}{T} \quad \textbf{(8.4)}$$

Denklem 8.4'te O_i, her bir sözdizim ağacının geçerli (gold) ağaç ile benzerliklerinden oluşan listenin dizgenin i. tümce için ürettiği orijinal sıralaması, I_i ise aynı listenin ideal sıralamasını simgeler. İdeal sıralamada her bir tümce için tümce ağaçları geçerli ağaca olan benzerlikleri çoktan aza olacak şekilde dizilir. T ise toplam tümce sayısını ifade eder.

8.3. Deney Sonuçları

Tablo 8.1'de modellerin tekil değerlendirme başarımları görülmektedir. 1, 2, 3 rakamlarıyla biten modeller dil modelleri olup; 1, tekli dil modelini, 2, ikili dil modelini ve 3, üçlü dil modelini temsil eder. * sembolü ile başlayan modeller ikinci grup modeller iken, # sembolü ile başlayan modeller üçüncü grup modellerdir. Bu incelemede modeller tek tek ele alındığı için AUT verisi üzerinde herhangi bir eğitim/test gruplaması yapılmadı. **Tablo 8.1**'deki sıralama, bağıntı başarımına göre büyükten küçüğe doğru gerçekleştirilmiştir.

Tablo 8.1. *Tekil Değerlendirme Başarımları*

Model	Ortalama	B. Ortanca	Benzerlik	Bağıntı
Biçimbirim_3	4,61	3,33	0,65	0,30
***OBBD**	5,13	4,00	0,64	0,29
Biçimbirim_2	4,70	3,56	0,64	0,29
Kanal_1	5,26	4,00	0,61	0,24
Kanal_2	5,59	4,21	0,61	0,23
Biçimbirim_1	5,62	4,00	0,63	0,23
Kanal_3	6,18	5,33	0,60	0,19
Son ek_1	6,86	5,71	0,60	0,17
Son ek_2	6,98	6,00	0,58	0,15
Son ek_3	7,00	6,00	0,58	0,15

Rol_3	7,40	6,74	0,57	0,12
Rol_1	7,70	7,20	0,57	0,11
Gövde_1	7,66	7,00	0,58	0,10
#Tümleç-eylem	8,22	8,00	0,58	0,09
Gövde_2	7,69	6,67	0,57	0,08
Rol_2	8,00	7,33	0,56	0,07
Gövde_3	7,86	7,11	0,56	0,07
#Birleşik durum ekli tümleç-eylem	8,38	8,00	0,56	0,06
#Koşullu durum ekli tümleç-eylem	8,71	8,00	0,55	0,04
#Eylem	8,40	8,00	0,54	0,03
***Öbek**	8,43	8,00	0,54	0,03
#Birleşik özne-eylem	8,57	8,00	0,54	0,02
#Koşullu özne-eylem	8,98	8,67	0,54	0,00
#Özne	9,11	9,14	0,54	-0,03

Tablo 8.1'de görüldüğü gibi üst sıralarda çoğunlukla dil modelleri gözlenmektedir. Yanı sıra, üçüncü grup modeller diğerlerine göre daha başarısızdır. İkinci grup modellerden *OBBD üst sıralarda yer alırken, *Öbek modeli oldukça başarısızdır.

8.3.1. Model birleştirme

Modellerin birlikte kullanılması tek tek uygulanmalarından daha başarılı sonuçlar vermiştir. Model birleştirme için basit bir yöntem olan ileri yönlü öznitelik seçimi (forward feature selection) uyarlanmıştır. Bu uyarlamada modeller öznitelikler olarak ele alınmış ve her bir başarım ölçüsü için ayrı ayrı en iyi alt küme aranmıştır.

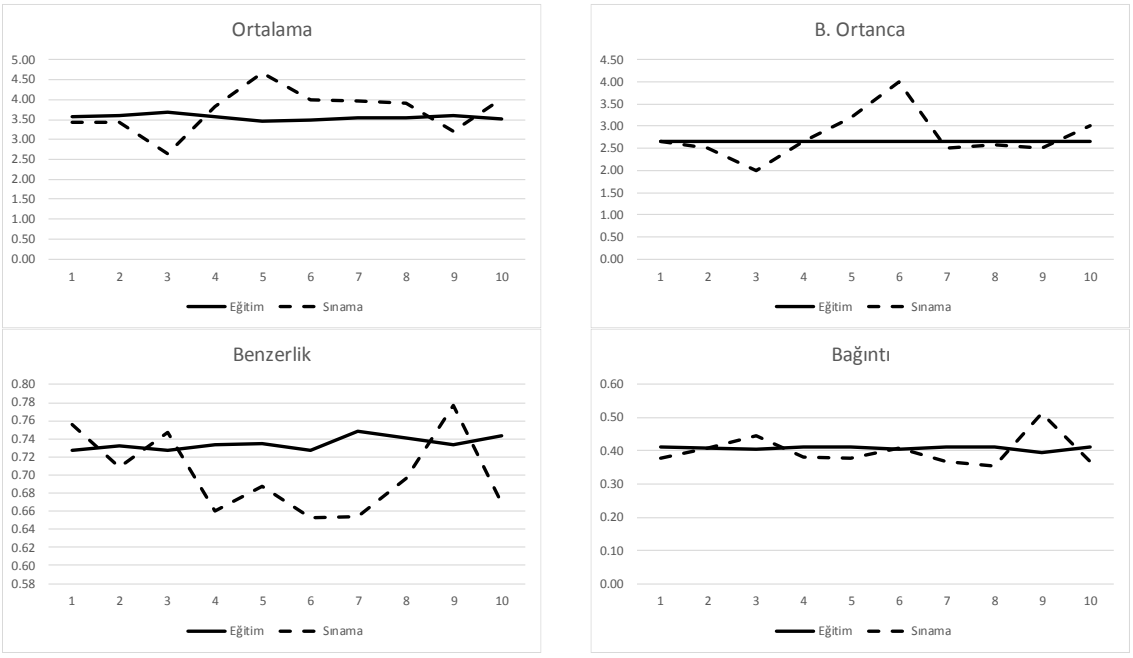
Öznitelik seçiminde veriden kaynaklı tesadüfi hataların önüne geçmek için 10 gruplu çapraz doğrulama yapılmıştır. **Tablo 8.2**'de öznitelik seçimlerinden elde edilen sonuçlar görülmektedir.

Tablo 8.2. *Öznitelik Seçimi Sonuçları*

		Ortalama*	B. Ortanca*	Benzerlik*	Bağıntı*
Eğitim	**Ortalama**	3,55	4,14	3,81	3,67
	B. Ortanca	2,67	2,67	2,67	2,67
	Benzerlik	0,70	0,68	0,73	0,72
	Bağıntı	0,39	0,34	0,39	0,41
Sınama	**Ortalama**	3,71	4,17	4,04	**3,70**
	B. Ortanca	2,53	2,76	2,85	**2,49**
	Benzerlik	0,69	0,67	0,70	**0,72**
	Bağıntı	0,37	0,34	0,36	**0,40**

Tablo 8.2'de sütunlar öznitelik seçimi yapılırken temel alınan ölçüyü, satırlar ise eğitim ve sınama için tekrarlı olmak üzere öznitelik seçiminin sonucu olarak hesaplanan ölçüleri gösterir. Örnekle ifade edilecek olursa, Ortalama* sütunu, ortalama ölçüsüne göre yapılan eniyileme sonucunda seçilen özniteliklerin her bir çapraz doğrulama grubu için ürettiği değerlerin ortalamasıdır. Bu tabloda sınama değerlerine odaklanmak gerekir. En iyi ve aynı zamanda eğitim sonuçlarıyla en tutarlı sınama değerleri Bağıntı* ölçüsü ile elde edilmiştir. Bu sonucu bağıntı ölçüsünün sıralamayı benzerliklere dayalı olarak bir bütün halinde değerlendirmesi özelliği ile ilişkilendirebiliriz.

Tablo 8.2'de başarım ölçülerinin çapraz doğrulama grupları üzerindeki genel eğilimleri özetlenmiştir. **Şekil 8.2**'de ise başarım ölçülerinden elde edilen sınama değerlerinin değişkenliği 10 çapraz doğrulama grubu üzerinde incelenmektedir. **Tablo 8.2**'den farklı olarak eniyilenen ölçü ile incelenen ölçü aynıdır.



Şekil 8.2. Başarım Ölçülerinin Çapraz Doğrulama Grupları Üzerinde Değişimi

Şekil 8.2'de görüldüğü gibi sınama değerleri için çapraz doğrulama grupları arasında en az değişkenlik bağıntı ölçüsü için elde edilmiştir. Bu sonuç, bağıntının veri değişimine karşı daha esnek, dolayısıyla ezberlemeye karşı daha dirençli bir ölçü olduğu şeklinde yorumlanabilir.

Tablo 8.3'te öznitelik seçiminde temel alınan ölçüye göre modellerin çapraz doğrulama gruplarında kaç kere seçildiği gösterilmektedir.

Tablo 8.3. *Modellerin Seçilme Sayıları*

	Ortalama	**B. Ortanca**	**Benzerlik**	**Bağıntı**
#Birleşik durum ekli tümleç-eylem	1	0	6	10
#Birleşik özne-eylem	7	0	9	10
#Eylem	9	0	10	3
#Koşullu durum ekli tümleç-eylem	0	0	1	0
#Koşullu özne-eylem	8	0	8	5
#Özne	0	0	6	0
#Tümleç-eylem	1	1	10	0
***OBBD**	10	2	9	10
***Öbek**	1	0	9	6
Biçimbirim_1	0	0	4	0

Biçimbirim_2	6	8	9	0
Biçimbirim_3	10	10	10	10
Gövde_1	8	8	10	10
Gövde_2	0	0	7	0
Gövde_3	0	0	2	0
Kanal_1	10	2	10	10
Kanal_2	2	0	4	0
Kanal_3	0	0	0	0
Rol_1	7	0	6	0
Rol_2	1	0	5	9
Rol_3	10	0	9	10
Son ek_1	10	0	8	10
Son ek_2	7	0	9	10
Son ek_3	2	0	2	2

Tablo 8.3 incelendiğinde en az öznitelik seçiminin bağıl ortanca ölçüsünde gerçekleştiği fark edilecektir. Bu, ilgili ölçünün bir tür ortanca hesaplaması olmasından kaynaklanmaktadır. Ortanca, bir dizinin tam ortasındaki değeri aldığı için düşük hassasiyete sahiptir. Bu da öznitelik seçiminde alt kümeleri ortanca ölçüsü üzerinden değerlendirirken farklılıkların az olmasına yol açar. Böylece ileri yönlü öznitelik seçiminde az sayıda öznitelikle aynı başarımı elde etmek çok daha olası olur.

Bu noktada başarım ölçülerini karşılaştırmak ve içlerinden birini seçmek gerekirse, ortalama, tümcelerin ağaç sayılarından etkilendiği için elenir; bağıl ortanca, ortalamaya göre daha sağlam olmasına rağmen hassas olmadığı için elenir; benzerlik, çapraz doğrulama grupları arasında sınama değerlerinin değişkenliği çok olduğu için elenir. Sonuç olarak, eldeki veriler ışığında en uygun başarım ölçüsünün bağıntı olduğu söylenebilir.

Bağıntı ölçüsü temelli öznitelik seçiminde en çok seçilen öznitelikler, 10 grubun yarısından çoğunda gözlenme şartıyla, en çok seçilenden en az seçilene doğru, şöyle sıralanabilir: #Birleşik durum ekli tümleç-eylem, #Birleşik özne-eylem, *OBBD, Biçimbirim_3, Gövde_1, Kanal_1, Rol_3, Son ek_1, Son ek_2, Rol_2, *Öbek. Bu liste dikkatle incelenecek olursa tekil değerlendirme sıralamasında (bk. **Tablo 8.1**) sonlarda yer alan üçüncü grup modellerden (# ile başlayanlar) ikisinin öznitelik seçiminde en çok gözlenen ilk iki öznitelik olduğu fark edilecektir. Buradan, yalnızca tekil

değerlendirme sonuçlarına bakarak modeller için mutlak bir yorum yapılamayacağı söylenebilir.

8.3.2. Model ağırlıklandırma

Öznitelik seçimi ile bir tür model birleştirme yapılmış ve modelleri tek başına kullanmaktan çok daha iyi bir performans elde edilmiştir. Bu performans artışı sağlanırken çapraz doğrulama ile ezberlemenin önüne geçilmeye çalışılmıştır. İleri yönlü öznitelik seçimi açgözlü (greedy) bir yöntem olduğu için en basit ve en az modelden oluşan model birleştirmeyi üretir fakat bu birleştirmede modellerin ağırlıkları eşittir. Ağırlıklı oylama ile modellere ağırlık atamak performansı daha da artırabilir. Ağırlıklı oylamada her bir modele bir ağırlık değeri atanır ve modelin ürettiği log-olasılık bu ağırlık değeri ile çarpılarak sonuçlar doğrusal biçimde birleştirilir. Ağırlıklı oylama ile model birleştirmenin genel denklemi aşağıda verilmiştir:

$$\sum_{i=1}^{n} w_i m_i \ \text{öyle ki} \ \sum w_i = 1 \ \ ve \ \forall w_i \geq 0 \qquad \textbf{(8.5)}$$

Ağırlıklı oylamanın temel parametreleri olan ağırlıkların belirlenmesi için çeşitli yollar izlenebilir. İlk olarak öznitelik seçiminden elde edilen, her bir modelin çapraz doğrulama gruplarında kaç kere seçildiği bilgisi bir oylama olarak yorumlandı. Bu ağırlıklar **Tablo 8.3**'ün bağıntı sütunundaki seçim sayılarıdır. İlgili ağırlıklar kullanıldığında başarım değerleri şöyle elde edilmiştir: ortalama: 3,65, bağıl ortanca: 2,67, benzerlik: 0,72, bağıntı: 0.41.

İkinci ağırlık belirleme yolu olarak genetik algoritmalar yaklaşımına başvurulmuştur. Genetik algoritmadaki birey veya kromozom kavramı, bu problemde, toplamları 1 olan ve pozitif gerçek sayılardan oluşan bir ağırlık dizisine denk düşer. Ağırlık dizisindeki her bir ağırlık birer gen olarak yorumlanır. Deneyde tek nokta çaprazlama kullanılarak iki ağırlık dizisinden iki yeni ağırlık dizisi daha elde edilmektedir. Bu yeni ağırlık dizilerinin toplamı genellikle 1 olmaz. Bu nedenle her çaprazlama ürünü için bir düzeltme işlemi gerçekleştirilir. Çaprazlama sonrasında belli sayıda gen için mutasyon gerçekleştirilir. Mutasyon ilgili ağırlığın küçük bir miktar

azaltılması ya da artırılmasıdır. Bu işlem sonucunda da ilgili ağırlık dizisi üzerinde bir düzeltme işlemine ihtiyaç vardır. Değerlendirme ölçütü olan uygunluk fonksiyonu, bu problemde, ağırlık dizisinin model birleştirmeye uygulanmasıyla elde edilen AUT verileri için hesaplanan başarım ölçüleridir. Dolayısı ile genetik algoritma yaklaşımında da öznitelik seçiminde olduğu gibi her başarı ölçüsü için ayrı deneyler gerçekleştirilmiştir.

8.3.3. Genel değerlendirme

Tablo 8.4'te SBG problemi için tanımlanan dört başarım ölçüsüne göre taban, önerilen gözetimsiz yöntemler ve kâhin sıralayıcıdan elde edilen sonuçlar özetlenmektedir.

Tablo 8.4. *Genel Değerlendirme*

		Ortalama	B. Ortanca	Benzerlik	Bağıntı
Taban	**Rastgele Sıralama**	8,76	8,36	0,5498	0,0137
Deneyler	**En İyi Tek Model (Biçimbirim_3)**	4,61	3,33	0,6500	0,3000
	Öznitelik Seçimi	3,70	2,49	0,7195	0,4001
	GA ile Ağırlıklandırma	**3,63**	**2,51**	**0,7280**	**0,4071**
Kâhin	**Öbek Belirleme (ÖB)**	3,61	2,67	0,7420	0,4526
	Biçimbilimsel Belirsizlik Giderme (BBG)	2,11	1,45	0,7752	0,3289
	ÖB + BBG	1,25	1,14	0,9175	0,5763

Tablo 8.4'te görüldüğü gibi, taban (baz) olarak alınan rastgele sözdizim ağacı sıralamasının bağıntı başarımı 0,0137 iken önerilen modellerden en iyi performansı gösteren biçimbirim_3 modeli bunu 0,30'a çıkarmaktadır. Ardından öznitelik seçimiyle 0,4001'e ve GA ile ağırlıklandırma ile 0,4071'e ulaşmaktadır. Kâhin sıralayıcılar kullanıldığında varılabilecek tavan değerler ÖB ile 0,4526, BBG ile 0,3289 ve her ikisi birlikte 0,5763'tür. 0,5763'ün bağıl ortanca karşılığı olan 1,14 değeri, bu sıralama gerçekleştiğinde doğru sözdizim ağacının sıralamadaki konumunun tümcelerin yarısı için en çok 1,14 olacağını ifade eder. Bu, çok iyi bir performans anlamına gelmekte olup

bu noktaya basit sezgisel yöntemler ve ilkel kurallarla tanımlanmış biçimbilimsel belirsizlik giderme ve öbek belirleme modülleriyle ulaşmak mümkün olabilir. Bu tezde ilgili varsayımlar sınanmamıştır ancak sonraki çalışmalarda, önerilen yöntemlerle elde edilen başarımlardan daha ileriye gitme olasılığı hakkında bir fikir vermektedir.

8.3.4. Biçimbilimsel belirsizlik giderme

Biçimbilimsel belirsizlik giderme, biçimbilimsel yapısı birden çok şekilde yorumlanabilen sözcükler için ilgili yorumların arasından geçerli olanı belirleme işidir. Bu iş SBG'nin bir yan ürünü olarak değerlendirilebilir. Sözdizim ağaçları sıralaması aynı zamanda biçimbilimsel seçimler için de bir sıralama vermektedir. **Tablo 8.5**'te SBG için yapılmış deneylerdeki sözdizim ağacı sıralamaları bu amaçla incelenmiş ve biçimbilimsel belirsizlik gidermeye ilişkin başarımlar elde edilmiştir. Simge sütunu biçimbilimsel yapısı doğru olarak belirlenmiş simgelerin sayısının toplam simge sayısına oranını, tümce sütunu ise bütün simgelerinin biçimbilimsel yapısı doğru olarak belirlenmiş tümcelerin sayısının toplam tümce sayısına oranını göstermektedir.

Tablo 8.5. *Yan Ürün Olarak Biçimbilimsel Belirsizlik Giderme*

		Simge	**Tümce**
Taban	Rastgele Sıralama	0,7611	0,1878
Deneyler	En İyi Tek Model (Biçimbirim_3)	0,8869	0,5242
	Öznitelik Seçimi	0,9076	0,6097
	GA ile Ağırlıklandırma	**0,9097**	**0,6189**
Kâhin	Öbek Belirleme (ÖB)	0,8194	0,2933
	Biçimbilimsel Belirsizlik Giderme (BBG)	0,9966	0,9954
	ÖB + BBG	0,9643	0,8684

8.3.5. Öbek belirleme

Öbek belirleme, bu tezde, verilen bir tümceyi kurucu bileşenlerine ayırma işi olarak tanımlanmıştır. Kurucu bileşenler özne, tümleç, eklenti ve eylemdir. Öbek

belirlemede bileşenlerin sınırları ve türleri doğru olarak belirlenmeye çalışılmaktadır. Biçimbilimsel belirsizlik giderme gibi öbek belirleme de s SBG'nin bir yan ürünü olarak değerlendirilebilir. Sözdizim ağaçları sıralaması aynı zamanda öbek belirleme seçimleri için de bir sıralama vermektedir. **Tablo 8.6**'da SBG için yapılmış deneylerdeki sözdizim ağacı sıralamaları bu amaçla incelenmiş ve öbek belirlemeye ilişkin başarım elde edilmiştir. Tümce sütunu bütün öbekleri doğru olarak belirlenmiş tümcelerin sayısının toplam tümce sayısına oranını göstermektedir.

Tablo 8.6. *Yan Ürün Olarak Öbek Belirleme*

		Tümce
Taban	Rastgele Sıralama	0,3834
Deneyler	En İyi Tek Model (Biçimbirim_2 ve 3)	0,4942
	Öznitelik Seçimi	0,5774
	GA ile Ağırlıklandırma	**0,5912**
Kâhin	Öbek Belirleme (ÖB)	0,9954
	Biçimbilimsel Belirsizlik Giderme (BBG)	0,6189
	ÖB + BBG	0,9954

9. TARTIŞMA

Bu tezde Türkçe için sözdizimsel belirsizlik giderme problemi ele alınmıştır. Bunun için sözdizimsel çözümleyici, biçimbilimsel çözümleyici (Morfolog) ve sözlükçe (TrLex) gibi özgün altyapı ögeleri tasarlanmış ve bunları eşgüdümlü biçimde yöneten bir dizge oluşturulmuştur. TMoST adını verdiğimiz bu dizge, mevcut eksiklikler giderildikten sonra araştırmacıların kullanımına açılması düşünülen kapsamlı bir Java kütüphanesidir.

Derlemler, dil modelleme ve önerilen yöntemleri sınama gibi işlerde kullanıldıkları için, birçok DDİ çalışmasında belirleyici unsurdur. Bu anlamda Türkçe düşük kaynaklı dillerden biri olarak yorumlanabilir. Özellikle sözdizim çözümlemesi konusunda ihtiyaç duyulan AYD açısından bakıldığında Türkçe çok geridedir. Bildiğimiz kadarıyla kullanılabilecek tek AYD, bağımlılık dilbilgisi esasına göre işaretlenmiş olan OSTAD'dır. TMoST daha çok öbek yapı dilbilgisine benzeyen bir çözümleme yöntemi kullandığı için yeni bir AYD oluşturmak gerekmiştir. AUT adını verdiğimiz bu derlem dil modellemeye yetecek sayıda tümce içermemektedir. Bunun nedeni, dili modellemek için geniş kapsamlı bir derlemden elde edilecek bir örnekleme ihtiyaç duyulması ve örneklemdeki tümcelerin sözdizimsel olarak işaretlenmesinin uzun bir süre gerektirmesidir. Bütün bu zorunlulukların bir sonucu olarak tezde dil modelleme amacıyla ham metinler kullanılmıştır. AUT ise deneylerde sınama amaçlı olarak değerlendirilmiştir.

AYD, işaretleyicilerin doğruluk oranına bağlı olmak üzere, tümceler için verilen bağlamda geçerli olan sözdizimsel yorumlardan oluşur. Böylece ilgili yorumlara dayanarak bilgi çıkarımı ve modelleme yapılabilir. Ham metinler ise dilin taşıdığı belirsizliği yansıtmaları nedeniyle gürültü içerir. Belli koşullarda ve bazı yöntemler kullanılarak bu gürültüyü azaltmak ve dili modellemek mümkün olabilir. Bunu sağlayan yöntemler gözetimsiz yöntemlerdir. Bu tezde doğrudan bir gözetimsiz veri işleme veya bilgi çıkarma yöntemine başvurulmamıştır. Bunun yerine, TMoST'un sözdizimsel çözümleme bileşenine gömülü olan dilbilgisi kuralları, yapı nesneleri mekanizmasıyla

birlikte bir filtreleyici gibi davranmış ve ham metinlerdeki Türkçeye aykırı yapılar elenerek dolaylı biçimde gözetimsiz dil modellemeye imkân sağlanmıştır.

Tezde tanıttığımız sözdizimsel çözümleyici özgün bir sözdizim ağacı üretmektedir. Bu ağaç yapısı, öbek yapıları ayrıntılı biçimde göstermesi açısından öbek yapı dilbilgisine; özne, tümleç gibi işlevsel ögeleri içermesi nedeniyle de bağımlılık dilbilgisine benzemektedir. Önerilen bu yapının Türkçeye uygun bir gösterim biçimi olduğunu düşünüyoruz. Bildiğimiz kadarıyla Türkçenin sözdizimsel gösteriminde yalnızca bağımlılık dilbilgisi yaklaşımından yararlanılmaktadır. Tezde iki gösterim arasındaki farklılık incelenmemiş olsa da bu başka bir çalışmanın konusu olabilir.

Bölüm 8'de sözdizimsel çözümleme için Türkçeye uygun yeni bir kavram önerdik: dizimbirim (syntheme). Dizimbirim kavramı sayesinde bir sözcük tabanıyla (Base) bir öbek (Phrase) denk biçimde işlenebilmektedir. Dizimbirimler Türkçenin bağımlılık dilbilgisi ile çözümlenmesinde kullanılan IG kavramına koşut olarak değerlendirilebilir. Ancak bunlar daha karmaşık ve soyut bir yapıdır. Buna ek olarak sözdizimsel çözümleyici için eylemler de ayrı bir öneme sahiptir. Bu önem dizgeye sirayet etmiş olup bir tümce çözümlemesinde gözlenen her bir etkin eylem başlangıçta birer potansiyel eylem öbeği (Clause yapısı) kurmakta ve çözümleme eylem öbeklerinin etrafında şekillenmektedir.

TMoST'un alt dizgelerinden biri olan Morfolog özgün bir biçimbilimsel çözümleyici olup, kapsamlı bir çalışmanın sonucunda ortaya çıkmıştır. Özellikle sözdizimsel çözümlemede ihtiyaç duyulan tanecikselliği sağlama açısından önemli bir role sahiptir. Alanyazında Zemberek, TrMorph gibi kullanıma açık çözümleyiciler olmasına rağmen yeni bir biçimbilimsel çözümleyici tasarlamamızın nedeni, sözü edilen örneklerin tamamen şeffaf bileşenlere sahip olmaması veya geliştirildikleri platformların Java ile uyumsuz olmasıdır. Buna ilave olarak, mevcut çalışmalarda faydalanılan dilbilgisel kaynaklar oldukça sınırlıdır. Biz bu tezde dilbilgisel doğruluğa ve Türkçenin karakteristik durumlarına özel olarak dikkat etmeye çalıştık. Bu anlamda, biçimbilimsel çözümleyicinin en önemli kısımlarından biri olan sözlükçe çok ayrıntılı ve emek-yoğun bir çalışmanın sonunda hazırlanmıştır. Biçimbilimsel çözümleyicinin

başka bir önemli bileşeni olan morfotaktikleri oluşturan ekler ise yazarın yüksek lisans tez çalışmasından (Aslan, 2008) elde edilmiştir.

Tezde en ayrıntılı çalışma TrLex üzerinde gerçekleştirilmiştir. TrLex 110.960 adet girdiden oluşan kaynak sözlüğün işlenmesiyle elde edilmiş bir biçimbilimsel sözlükçedir. Bu sözlükçe biçim, yapı, anlam vb. birçok bilgiden oluşmakta olup kısıtlı bir sürümü araştırmacıların kullanımına açılacaktır.

Bölüm 6'da tanıtılan metin koleksiyonu, belirtildiği gibi, Türkçeyi modellemek için kullanılan metinlerdir. 42.630.365 filtrelenmiş tümceden oluşan bu koleksiyon doğrudan paylaşılmayacaktır; ancak koleksiyondan elde edilen modellerin paylaşılması mümkündür. AUT ise araştırmacıların kullanımına sunulacak olan materyallerden biridir.

Tezde istatistiksel dil modeli, OBBD ve öbek modeli ile ilişkisel modeller olmak üzere üç grup model üretilmiş ve deneylerde kullanılmıştır. Dil modelleri beş adet olup, biçimbirim, gövde, son ek, kanal ve rol adlı modellerdir. Özellikle biçimbirim modeli, TMoST'un çözümleme mekanizmasının merkezinde de biçimbirimlerin yer alması nedeniyle önemlidir. Aşağıda bu modelin sözdizimsel belirsizlik gidermede açısından önemi ayrıntılı olarak tartışılacaktır. Bu dil modellerinden kanal ve rol adlı modeller TMoST'un kendine özgü yapısının bir yansımasıdır ve özgün oldukları iddia edilebilir. Biçimbirim, gövde ve son ek modelleri ise halihazırda bilinen kavramlar olup dil modeli olarak kullanılmış olmaları muhtemeldir. Bununla birlikte, yaptığımız araştırmalarda Türkçe için bu modellerin kullanıldığı çalışmalara rastlayamadık.

OBBD modeli bir sözdizim ağacının yapısının olasılığını hesaplaması nedeniyle faydalı bir araçtır. Buna ek olarak önerdiğimiz öbek modeli ise sözlükbirimler arasındaki ilişkiyi modellediği için yeni bir yaklaşım olarak değerlendirilebilir. Son grup olan ilişkisel modeller tümce içinde roller arası ilişkileri modellemektedir. Bu model grubu da tamamen özgündür.

Sözdizimsel belirsizlik giderme problemi, verilen bir tümce için sözdizimsel çözümleyicinin ürettiği sözdizim ağaçlarının uygun biçimde sıralanması şeklinde tanımlanmıştır. Ağaçları uygun biçimde sıralamak ile kastedilen şey, verilen bağlamda

geçerli olan ağacın en üstte yer aldığı, bunu takip eden ağaçların da en üstteki ağaca olan yapısal benzerliklerine göre sıralandığı düzendir.

Deneylerde dört tür başarım ölçüsü kullanılmıştır. Ortalama adlı ölçü ilk akla gelen en temel yöntemdir. Diğer üç ölçü (bağıl ortanca, benzerlik ve bağıntı) ise tez kapsamında geliştirilmiştir. Bu ölçüler arasında özellikle bağıntı adlı başarım ölçüsünün üzerinde durmak gerekir. Bu ölçü sıralamayı benzerliklere dayalı olarak bir bütün halinde değerlendirir. Sıra başarımları yalnızca geçerli ağacın konumuyla ilgilendiği için, benzerlik başarımı da modelin önerdiği en olası ağacı dikkate aldığı için bütünsel bir değerlendirme sunmazlar. Bağıntı ise sıra ve benzerlik başarımlarını birleştirerek hem sıralamayı hem de benzerlikleri dikkate alır. Bir model tesadüfi olarak geçerli ağacı birinci sırada önerdiğinde ortalama ve benzerlik başarımları en yüksek değerlerini alacaktır. Ancak bağıntı diğer ağaçların sıralamasıyla da ilgilendiği için bu tesadüfi başarıdan çok az etkilenecektir. Bu nedenlerle sözdizimsel belirsizlik gidermede bir değerlendirme ölçüsü olarak bağıntıyı öneriyoruz.

Modellerin tekil değerlendirmesi incelendiğinde biçimbirim dil modelinin özellikle üçlü ve ikili sürümünün çok başarılı olduğu görülmektedir. Eğer sözdizimsel belirsizlik giderme için en yalın modelin ne olduğu sorulursa, bu veriler ışığında yanıt, biçimbirime dayalı dil modeli olacaktır. Örneğin biçimbirim üçlüleri için deneylerden elde edilen bağıntı ölçüsü başarım değeri 0,30'dur. Rastgele sıralamadan elde edilen 0,0137 ile karşılaştırıldığında bu skor oldukça iyidir. Tekil değerlendirmede biçimbirimi, OBBD ve kanal dil modeli takip etmektedir. OBBD modelinin başarılı olması alanyazında sıklıkla kullanılan bir model olmasıyla koşutluk göstermektedir. Sözdizim ağaçlarının yapısal olasılıklarını üreten bu model sözdizimsel belirsizlik gidermede tek başına ele alındığında oldukça başarılıdır. Kanal dil modeli ise biçimbirim dil modeline göre çok daha karmaşık bir modeldir. Biçimbirim modeli biçimbirimlerin sözlüksel formları ve biçimbilimsel etiketlerinden oluşmakta iken kanal modeli bir sözcükte gözlenen bütün biçimbilimsel etiketler dizisinden oluşur. Bu nedenle kanal modelinde birlilerin ikililere ve ikililerin de üçlülere göre daha başarılı olması tesadüf olmamalıdır. Dil modellerinde hesaplama birimi olarak alınan öge karmaşıklaştıkça birimlerin art arda gözlenmesi için daha büyük derleme ihtiyaç

duyulmaktadır. Sonuç olarak, veride seyreklik probleminden uzaklaşmak için yalın modellere başvurulmalıdır.

Deneyler sonucunda modellerin birlikte kullanılmasının performansı artırdığı görülmektedir. En iyi model kombinasyonlarını belirlemek için her bir model birer öznitelik gibi düşünülerek ileri yönlü öznitelik seçimi gerçekleştirilmiştir. Bunun sonucunda en iyi bağıntı başarımı (0,4001) yine bağıntı ölçüsü en iyilenerek elde edilmiştir. Sözü edilen bu model birleştirmede aşırı uyumdan kaçınmak için 10 gruplu çapraz doğrulama uygulanmıştır.

Şekil 8.2'de her bir başarım ölçüsünün 10 gruplu çapraz doğrulamada eğitim ve sınama verilerinde nasıl bir değişkenlik gösterdiği verilmiştir. Buna göre, en az değişim gösteren ölçü bağıntı ölçüsüdür.

Model birleştirmede modeller eşit ağırlıklarla bir araya getirilmiştir. Model ağırlıklandırma ise her bir modelin farklı ağırlıklar alabildiği bir deney kurgusudur. Bunun sonucunda elde edilen en iyi bağıntı değeri 0,4071 olmuştur. Bu değer model birleştirme ile elde edilen 0,4001 ile karşılaştırıldığında büyük bir ilerleme olarak yorumlanmayacaktır. Bunun sonucu olarak, modellere ağırlıklar atamanın yapılan deneyler için kayda değer bir getiri sağlamadığı söylenebilir.

Genel değerlendirmede (bk. **Tablo 8.4**) ayrıca, kâhin sıralayıcılardan elde edilen başarım değerleri de verilmiştir. Buna göre, öbek belirleme mükemmel şekilde çalıştığında, sözdizimsel belirsizlik giderme için 0,4526 bağıntı başarımı, biçimbilimsel belirsizlik giderme mükemmel şekilde çalıştığında ise 0,3289 bağıntı başarımı elde edilmektedir. Her iki kâhin birlikte iş gördüğünde ise 0,5763 değerine ulaşılmaktadır. Buradan çıkan sonuç sürprizdir. Çünkü biçimbilimsel belirsizlik gidermenin sözdizimsel belirsizlik giderme üzerinde daha etkili olması beklenir. Bu sonuç tezin araştırma sorularından biri olmayıp deneyler sonucunda ortaya çıkan ilginç bir bulgudur ve başka çalışmalarda araştırılmalıdır. Her iki kâhin sıralayıcının birlikte kullanılmasıyla elde edilen 0,5763 değeri TMoST dizgesine öbek belirleme ve biçimbilimsel belirsizlik giderme bileşenleri eklendiğinde ulaşılabilecek olan tavan değerdir. Bunun ötesine geçmek için başka araçlara ihtiyaç vardır. Ancak bu değerin

ortalama ve bağıl ortanca karşılıklarına bakacak olursak, sözü edilen tavan değerinin bu iki ölçü için ideal değer olan 1'e oldukça yaklaştığı görülmektedir.

Sözdizimsel belirsizlik giderme ile elde edilen sözdizim ağacı sıralamasından birçok bilgi çıkarılabilir. Bunlardan ikisi biçimbilimsel belirsizlik giderme ve öbek belirlemedir. Bunlar sözdizimsel belirsizlik giderme sürecinin birer yan ürünü olup bu sürecin hatalarını yansıtmaktadır. Buna göre model ağırlıklandırma ile elde edilen biçimbilimsel belirsizlik giderme başarımı 0,6189 tam tümce doğruluğudur. Bu da tezde önerilen modeller uygun kombinasyonlarla ve uygun ağırlıklarla bir araya getirildiğinde ve elde edilen sözdizim ağacı sıralamaları biçimbilimsel belirsizlik gidermeye uyarlandığında tümcelerin yaklaşık %62'sinin bütün simgelerinin biçimbilimsel çözümlemesinin doğru biçimde gerçekleştirildiği anlamına gelir. Sıralamalar öbek belirlemeye uyarlandığında ise tümcelerin yaklaşık %59'u doğru biçimde öbeklerine ayrılmış olur.

Tezde gerçekleştirilen çalışmaların sonucunda şu önerileri yapabiliriz:

1. Tümce, biçimbirimler dizisi olarak ele alınmalıdır. Bu anlamda biçimbirimleri doğru ve eksiksiz biçimde belirlemek çok önemlidir.
2. Sözdizimsel çözümleme eylem merkezli gerçekleştirilmelidir.
3. Biçimbilimsel yapılarla sözdizimsel yapıların etkileşmesi için dizimbirim adını verdiğimiz ara yapılara benzer ögelere ihtiyaç vardır.
4. Sözdizimsel belirsizlik gidermede eğer model ya da modellerin performansı en ayrıntılı biçimde ölçülmek isteniyorsa başarım ölçüsü olarak modelin önerdiği sıralama ile ideal sıralama arasındaki korelasyonu ölçen bağıntı ölçüsü kullanılmalıdır.
5. Sözdizimsel belirsizlik gidermede tek bir model kullanılacaksa bu model olabildiğince yalın bir model olmalıdır.
6. Model birleştirme yapılacaksa mutlaka çapraz doğrulama uygulanmalıdır. Çünkü modeller arasındaki etkileşim ile veride yer alan yapı farklılıklarının bir araya gelmesi sonuçları yorumlamayı imkânsız hâle getirebilir.
7. Model ağırlıklandırma eğer başarımı kayda değer biçimde artırmıyorsa tercih edilmemelidir.

KAYNAKÇA

Akın, M. D. ve Akın, A. A. (2007). Türk dilleri için açık kaynaklı doğal dil işleme kütüphanesi. *Elektrik Mühendisliği Dergisi,* 431, 38-44.

Alagözlü, N. (2016). Eğitimsel dilbilim kapsamında küçük ölçekli dilbilim ve dil öğretimi: kavramları ve katkıları, *Türkbilig*, 32, 181-208.

Altun, H. O. (2010) Düzeltme işareti ve Türkçede yazıldığı gibi okunmayan kelimeler. *Atatürk Üniversitesi Türkiyat Araştırmaları Enstitüsü Dergisi*, 17(43), 167-179.

Arısoy, E. and Arslan, L. M. (2005). Turkish dictation system for broadcast news applications. *2005 13th European Signal Processing Conference*, Antalya, Sep. 4-Sep. 8, 2005, pp. 1-4.

Aslan, Ö. (2008). *Türkçe kelimelerin biçim birimlerine ayrılması için kullanılacak standart biçim birimi kümesinin oluşturulması*, Yayımlanmamış Yüksek Lisans Tezi. Muğla: Muğla Üniversitesi.

Atalay, N. B., Oflazer, K. and Say, B. (2003). The annotation process in the Turkish treebank. *Proc. of the 4th Intern. Workshop on Linguistically Interpreteted Corpora*, LINC, pp. 38.

Bahl, L. R., Jelinek, F., and Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, (2), pp. 179-190.

Bahl, L. R., Brown, P. F., de Souza, P. V. and Mercer, R. L. (1989). A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(7), pp. 1001-1008.

Basili, R., Pazienza, M. T. and Velardi, P. (1991). Using word association for syntactic disambiguation. *Congress of the Italian Association for Artificial Intelligence,* Berlin: Springer, Oct. 29-Oct. 31, 1991, pp. 249-260. E. Ardizzone, S. Gaglio and F. Sorbello (Eds.).

Booij, G. (2005). Compounding and derivation. W. U. Dressler, D. Kastovsky, O. E. Pfeiffer and F. Rainer (Eds.), In *Morphology and its demarcations* (pp. 109-132). Amsterdam: John Benjamin Publishing Company.

Bozşahin, C. ve Zeyrek, D. (2000). *Dilbilgisi, bilişim ve bilişsel bilim*. Dilbilim Araştırmaları. İstanbul: Boğaziçi Üniversitesi Yayınları.

Cebiroğlu, G. ve Adalı, E. (2002). Sözlüksüz köke ulaşma yöntemi. *Proceedings of the 19th TBD Bilişim Kurultayı*, İstanbul, Turkey. pp. 155-160.

Chomsky, N. (1957). *Syntactic structures*. The Hague/Paris: Mouton.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. New York: Harper & Row.

Clements, G. N. and Sezer, E. (1982). Vowel and consonant disharmony in Turkish. H. van der Hulst and N. Smith (Eds.), In *The structure of phonological representations* 2: pp. 213-255.

Çöltekin, Ç. (2010). A freely available morphological analyzer for Turkish. *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC2010*, Valletta, Malta. pp. 820-827.

Cruse, D. A. (1986). *Lexical semantics*. New York: Cambridge University Press.

Deny, J. (2000). *Türk dili gramerinin temel kuralları (Türkiye Türkçesi)*. (Çev: O. Şahin). Ankara: TDK Yayınları.

Ediskun, H. (2005). *Türk dilbilgisi*. İstanbul: Remzi Kitabevi.

Emekligil, E., Arslan, S. and Agin, O. (2016). A bank information extraction system based on named entity recognition with CRFs from noisy customer order texts in Turkish. *International Conference on Knowledge Engineering and the Semantic Web*, Springer International Publishing, pp. 93-102.

Eryigit, G., Nivre, J., and Oflazer, K. (2008). Dependency parsing of Turkish. *Computational Linguistics*, 34(3), 357-389.

Eryigit, G. (2014). ITU Turkish NLP web service. *The European Chapter of the ACL*, EACL, pp. 1-4.

Eryigit, G., Adali, K., Torunoglu-Selamet, D., Sulubacak, U. and Pamay, T. (2015). Annotation and extraction of multiword expressions in Turkish treebanks. *NAACL: North American Chapter of the ACL*, pp. 70-76.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M. and Soria, C. (2006). Lexical markup framework (LMF). *International Conference on Language Resources and Evaluation-LREC 2006*.

Gedizli, M. (2012). Türkçede tek şekilli ve çok işlevli yapim ekleri. *Electronic Turkish Studies*, 7(4), 3351-3369.

Grishman, R., Macleod, C. and Meyers, A. (1994). COMLEX syntax: Building a computational lexicon. *Proceedings of the 15th conference on Computational linguistics-Volume 1*, Association for Computational Linguistics, pp. 268-272.

Grønbech, K. (2011) *Türkçenin yapısı*. (Çev: M. Akalın). Ankara: TDK Yayınları.

Güngör, T. (2003) Lexical and morphological statistics for Turkish. *Proceedings of TAINN*, pp. 409-412.

Güngördü, Z. (1993). *A lexical-functional grammar for Turkish*. Yayımlanmamış Yüksek Lisans Tezi. Ankara: Bilkent Üniversitesi.

Günşen, A. (2006). Göster- ve görset-/ körset- fiillerinin yapısı üzerine. *Erciyes Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 20. Yıl özel sayısı, 2006/1, 35-49.

Hankamer, J. (1989). Morphological parsing and the lexicon. W. Marslen-Wilson, (Ed.). In *Lexical Representation and Process*, (pp. 392-408), Cambridge, MA: MIT Press.

Hayashi, Y. and Ishida, T. (2006). A dictionary model for unifying machine readable dictionaries and computational concept lexicons. *Proceedings of Language Resources and Evaluation Conference 2006*, pp. 1-6.

Hong, J. F. (2014). *Verb sense discovery in Mandarin Chinese - A corpus based knowledge-intensive approach*. Berlin: Springer.

İlgen, B., Adali, E. and Tantuğ, A. C. (2012). The impact of collocational features in Turkish word sense disambiguation. *16th International Conference on Intelligent Engineering Systems, INES*, pp. 527-530. IEEE.

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4), 532-556.

Johanson, L. (2014). *Türkçe dil ilişkilerinde yapısal etkenler*. (Çev: N. Demir). Ankara: TDK Yayınları.

Kapusuz, E. (2006). *Stemming Turkish Words Using Snowball*. http://snowball.tartarus.org/algorithms/turkish/accompanying_paper.doc (Erişim: 13.05.2015)

Karaağaç, G. (2013). *Dil bilimi terimleri sözlüğü*. Ankara: Türk Dil Kurumu Yayınları.

Karaca, V. İ. (2012). Türkiye Türkçesindeki alıntı sözcüklerde görülen ses olayları üzerine bir inceleme. *Electronic Turkish Studies*, 7(4).

Kıran, Z., ve Kıran, A. (2013). *Dilbilime Giriş*. Ankara: Seçkin Yayıncılık.

Koskenniemi, K. (1983) *Two level morphology: A general computational model for word-form recognition and production*. Yayımlanmamış Doktora Tezi Helsinki: University of Helsinki.

Köksal, A. (1981). Tümüyle özdevimli deneysel bir belge dizinleme ve erişim dizgesi. *TURDER, TBD 3. Ulusal Bilişim Kurultayı*, s. 37-44.

Lee, J. W., Kim, S. D., Chae, J., Lee, J. and Kim, D. H. (2003). English syntactic disambiguation using Parser's Ambiguity Type information. *ETRI journal*, 25(4), 219-230.

Lieber, R. (1980). *On the organization of the lexicon*. Doktora Tezi. Massachusetts Institute of Technology.

Litkowski, K. (2005). Computational lexicons and dictionaries. K. Brown (Ed.). In *Encyclopedia of Language and Linguistics*, (pp. 753-761). Oxford: Elsevier Publishers.

Logacev, Ö. Ü., Fuchs, S. and Żygis, M. (2014). Soft 'g'in Turkish: Evidence for Sound Change in Progress*. Journal of the International Phonetic Association*. 1-24.

Marcus, M. P., Marcinkiewicz, M. A. and Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2), 313-330.

Molinero, M., Sagot, B. and Nicolas, L. (2009). A morphological and syntactic wide-coverage lexicon for Spanish: The Leffe. *RANLP 2009-Recent Advances in Natural Language Processing*. pp. 264-269.

Oflazer, K. (1994). Two-level description of Turkish morphology. *Literary and Linguistic Computing* 9(2), 137-148.

Oflazer, K., Say, B., Hakkani-Tür, D. Z. and Tür, G. (2003). Building a Turkish treebank. *Treebanks*, 261-277.

Oflazer, K. (2014). Turkish and its challenges for language processing. *Language resources and evaluation*, 48(4), 639-653.

Özsoy, A. S. (2004). *Türkçe'nin yapısı* (Vol. 1). İstanbul: Boğaziçi Üniversitesi Yayınevi.

Parlak, S. and Saraclar, M. (2009). Spoken information retrieval for Turkish broadcast news. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 782-783. ACM.

Petasis, G., Karkaletsis, V., Farmakiotou, D., Androutsopoulos, I. and Spyropoulo, C. D. (2001). A Greek morphological lexicon and its exploitation by natural language processing applications. *Panhellenic Conference on Informatics*, pp. 401-419. Berlin: Springer.

Porter, M.F. (2001). *Snowball: A language for stemming algorithms*. http://snowball.tartarus.org/texts/introduction.html (Erişim: 13.05.2015).

Ralli, A. (2010). Compounding versus derivation. *The Benjamins Handbook of Compounding*. 434-456. Philadelphia: Jonh Benjamins Publishing Company.

Rosner, M., Caruana, J. and Fabri, R. (1998). Maltilex: A computational lexicon for maltese. *Proceedings of the Workshop on Computational Approaches to Semitic Languages*, pp. 97-101. Association for Computational Linguistics.

Ruimy, N., Monachini, M., Distante, R., Guazzini, E., Molino, S., Ulivieri, M. and Zampolli, A. (2002). CLIPS, a Multi-level Italian Computational Lexicon: a Glimpse to Data. pp. 792-799. *Proceedings of Language Resources and Evaluation Conference*.

Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. *7th international conference on Language Resources and Evaluation*, LREC 2010, pp. 2744-2751.

Sak, H., Saraclar, M. and Güngör, T. (2010). Morphology-based and sub-word language modeling for Turkish speech recognition. *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 5402-5405. IEEE.

Sanders, A. F. and Sanders, R. H. (1989). Syntactic parsing: A survey. *COMP. HUM.*, 23(1), 13-30.

Say, B., Zeyrek, D., Oflazer, K. ve Özge, U. (2002). Development of a Corpus and a Treebank for Present day Written Turkish. *Proceedings of the Eleventh International Conference of Turkish Linguistics*, pp. 183-192. Eastern Mediterranean University, Cyprus.

Şahin, H. (2006). Türkçe'de ön ek. *Uludağ Üniversitesi Fen-Edebiyat Fakültesi Sosyal Bilimler Dergisi*, 10, 65-77.

Şamilov, A. (2015). *Entropi, informasyon ve entropi optimizasyon*. Ankara: Nobel Yayınları.

Sever, H. and Tonta, Y. (2006). Truncation of content terms for Turkish, *Conference on Intelligent Text Processing and Computational Linguistics*, CICLing.

Small, S. I., Cottrell, G. W. and Tanenhaus, M. K. (1987). Lexical ambiguity resolution. San Mateo, CA: Morgan Kaufman.

Tadić, M. and Fulgosi, S. (2003). Building the Croatian morphological lexicon. *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, pp. 41-46. Association for Computational Linguistics.

Taylan, E. E. (1984). *The function of word order in Turkish grammar* (Vol. 106). Los Angeles: Univ of California Press.

Toutanova, K., Manning, C. D., Flickinger, D. and Oepen, S. (2005). Stochastic HPSG parse disambiguation using the Redwoods corpus. *Research on Language & Computation*, 3(1), 83-105.

Tür, G. (2000). A statistical information extraction system. Yayımlanmamış Doktora Tezi. Ankara: Bilkent Üniversitesi.

Türk Dil Kurumu. (2005). *Türkçe Sözlük*. Ankara: Türk Dil Kurumu Yayınları.

Urešová, Z. (2009). Building the PDT-VALLEX valency lexicon. *On-line proceedings of the fifth Corpus Linguistics Conference*. University of Liverpool.

ÖZGEÇMİŞ

Adı-Soyadı : Özkan ASLAN

Yabancı Dil : İngilizce

Doğum Yeri ve Yılı : Afyonkarahisar/1982

E-Posta : euzkan@gmail.com

Eğitim ve Mesleki Geçmişi:

- 2003, Lisans, Selçuk Üniversitesi, Eğitim Fakültesi, Bilgisayar ve Öğretim Teknolojileri Öğretmenliği Anabilim Dalı
- 2003-2005, Bilgisayar Öğretmeni, İzmir/Ödemiş Atatürk İlköğretim Okulu
- 2005, Bilgisayar Öğretmeni, Muğla/Ula Atatürk İlköğretim Okulu
- 2005-2008, Araştırma Görevlisi, Muğla Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı
- 2008, Yüksek Lisans, Muğla Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı
- 2010, Araştırma Görevlisi, Muğla Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Anabilim Dalı
- 2010-2017, Araştırma Görevlisi, Anadolu Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Anabilim Dalı

Yayınları ve Bilimsel/Sanatsal Faaliyetleri:

- Aslan, Ö. (2008). Türkçe kelimelerin biçim birimlerine ayrılması için kullanılacak standart biçim birimi kümesinin oluşturulması, Yayımlanmamış Yüksek Lisans Tezi. Muğla: Muğla Üniversitesi.
- Aslan, Ö., Dinçer, B. T. ve Pembeci, İ. (2013). Türkçede öbek yapıların imlenmesi: bir derlem çalışması. *27. Ulusal Dilbilim Kurultayı*, Antalya.
- Turan, Ü. D., Aslan, Ö. ve Corga, E. (2014). Tümce-başında kullanılan eylemler: Derlem tabanlı bir çözümleme. *Anadolu Üniversitesi Sosyal Bilimler Dergisi*, Cilt: 14. Eğitim Özel Sayısı, Eskişehir.
- Aslan, Ö., Kantar, Y. M. ve Usta, İ. (2015). Genetic algorithms for solving portfolio allocation models based on relative-entropy, mean and variance. *Journal of Scientific Research and Development.* 2(12): 7-12.