Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi

Afyon Kocatepe University Journal of Science and Engineering

AKÜ FEMÜBİD **15** (2015) 035101 (275-285) AKU J. Sci. Eng. **15** (2015) 035101 (275-285)
DOI: 10.5578/fmbd.10354

Araştırma Makalesi / Research Article

Türkçe Tümce Çözümlemede Vektör Yaklaşımı

İlknur Dönmez[1], Eşref Adalı[2]
[1,2] İstanbul Teknik Üniversitesi, Bilgisayar ve Bilişim Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul.
e-posta: buyukkuscu@itu.edu.tr; adali@itu.edu.tr

Geliş Tarihi: 03.07.2015; Kabul Tarihi: 04.11.2015

Özet

Anahtar kelimeler
Dilbilgisi çözümlemesi, anlamsal çözümleme, öbek-kavramın yüklem uyumluluğu, cümlenin vektör temsili.

Tümcenin anlamsal ve dilbilgisi açısından çözümlenmesi Doğal Dil İşleme (DDİ)'nin ana konulardan biridir. Çalışmamızda, tümcedeki temel dilbilgisi ve anlamsal yanlışları saptamak için yüklemi temel alan yeni bir yöntem önerilmektedir. Türkçe tümcede yüklem özne ve zaman bilgisi içerir. Ayrıca yüklem, o tümcenin hangi öbeklerden oluşabileceği konusunda da belirleyicidir. Örneğin, "büyümek" yüklemi tümce içinde nesne almazken, -de ekiyle biten dolaylı tümleç öbeğini alır. Yüklem ayrıca her bir öbeğin içereceği kavram hakkında da bilgi içermektedir. Örneğin "düşünmek" yüklemi insanlara özgüdür. Dolayısıyla özne olarak insan kavramıyla ilişkilidir. Bu saptamalardan yola çıkarak çalışmamızda, tümcelerin öbekleri bulunmuş; her bir öbeğin hangi kavramla ilişkili olduğu belirlenmiş ve tümcenin dilbilgisi çözümlemesini ve anlam çözümlemesini yapan bir model tasarlanmıştır.

Vectorial Approach For Analysing Turkish Sentence

Abstract

Keywords
Grammatical and semantic analysis, phrase-concept and verb compatibility, vector representation of sentence.

The grammatical and semantic analysis of the sentence is one of the main subjects of Natural Language Processing (NLP). In this paper, we present a novel method to detect basic grammatical and semantic disorders by concentrating on the predicate. In Turkish, the predicate includes information about the subject and tense. The predicate also helps identify the phrases which make up the sentence. For example, "büyümek" (to grow) does not take an object, but it can take a locative phrase ending with the suffix "-de". The predicate is also informative about the semantic concept of a phrase. For example "düşünmek" (to think) is specifically an action performed by a human, so the subject will be related with the concept of a human. With these properties considered, a model has been designed to find phrases in a sentence, identify their relations to specific concepts, and analyze the sentences grammatically and semantically.

© Afyon Kocatepe Üniversitesi

1. Giriş

1970'lerden sonra bilgi teknolojilerindeki gelişmelerin sonucu olarak Doğal Dil İşleme (DDİ) alanındaki araştırmaların yoğunlaştığı görülmektedir. Tümcelerin anlamsal ve dilbilgisi açısından çözümlenmesi konusundaki çalışmalar da bu bağlamda sürdürülmektedir. Dilbilgisi ve anlamsal yanlışları düzeltmek için kural tabanlı, istatistiksel ve karma yöntemler kullanılmaktadır. Atwell (1987) ve Izumi (2003)'nin çalışmaları istatistiksel yöntemlere; Bigert ve Knutsson (2000), Ehsan ve Faili (2013) ve Wang (2014)'ın çalışmaları karma yöntemlere örnektir.

Türkçe dilbilgisi çözümleme çalışmaları tümceleri sözcüklere ayırma, sözcük kökü bulma (Cebiroğlu, 2002), biçim bilimsel çözümleme (Oflazer, 2003), tümce içi bağlılık ilişkilerini bulma (Eryiğit ve ark. 2008) çalışmaları olarak kümelenebilir. Türkçe tümcelerin dilbilgisi açısından çözümlenmesi konusunda Doğan ve Karaağaç (2012), İşgüder ve Adalı (2014) ve Aygül (2014)'ün çalışmaları bulunmaktadır. Bu çalışmalar kapsamında tümceler söz gruplarına bölünmekte ve sözcüklere eklenen hal ekleri incelenmektedir. Türkçe metinlerdeki yazım yanlışlarını otomatik bulmaya yönelik çalışmaların varlığı da bilinmektedir (Delibaş, 2008), (Dilsiz, 2015). Bu çalışmalarda istatistiksel yöntemler kullanılmıştır. Son zamanlarda tümce

içindeki sözcükler arasındaki kavram ilişkilerini inceleyen çalışmalara da tanık olunmuştur (Güngör ve Güngör 2007), (Amasyalı, 2012).

DDİ yöntemleri kullanılarak dilbilgisinin çözümlenmesi alanında çalışmalar halen devam etmektedir. Çalışmamızda dilbilgisi analiz sorununa anlamsal ve yüklem-öbek uyumu açısından bakan kural tabanlı bir yöntem önerilmektedir. Bu çalışmanın çıkış noktası Gottlob Frege'ın tanımıyla "tümcedeki anlamsal rolleri ilişkilendiren ve işlev olarak bağlayan yüklemdir" (Haaparanta ve Hintikka 2012). Ancak Türkçe bir tümcede yüklemin etkisi daha büyüktür çünkü yüklem eylemin öznesini, zamanını, ilişkili olduğu öbeği ve kavramı belirler. Yüklem uyumu ölçütlerine uygun olarak yazılmış bir tümcenin dilbilgisi açısından doğru, ölçütleri sağlamayan tümcelerinse yanlış olacağı görüşümüzdür ve bu çalışma bu ilkeler çerçevesinde geliştirilmiştir.

Çalışmamızda ilk olarak bir tümce, Tablo-1'de görüldüğü üzere sekiz öbek kümesine ayrılmıştır. Belirlenen her bir öbeğin yüklemle uyumluluğu denetlenmiştir. Örneğin, "*Ayşe kalemini Fatma'ya uzattı.*" tümcesinde "*uzattı*" yüklemiyle uyumsuz bir öbek bulunmamaktadır. Bir başka örnek olarak, "*Ayşe, okuldan bulundu.*" tümcesinde bulunmak yüklemi "-den" alan bir ayrılma öbeğiyle birlikte kullanılamaz. Buna karşın "–de" alan bir bulunma öbeğiyle birlikte kullanılır; dolayısıyla "*okuldan*" öbeği sorunlu bir öbektir.

İkinci aşamada, bulunan her bir öbeğin içerdiği kavram belirlenmiş; böylece tümcenin öbek-kavram çiftleri oluşturulmuştur. Daha sonra bu öbek-kavram çiftleri yüklemle karşılaştırılarak uyumluluğu denetlenmiştir. Örneğin "*Masa, üzerindeki yemeği düşündü*." tümcesinde, eşya kavramına sahip bir özne, yemek kavramına sahip bir belirtili nesne öbeği bulunmaktadır ve yüklemi *düşünmek*tir. Eşyaların düşünemediği bilindiğine göre bu tümce mantıksız bir tümcedir. Oysa tümce "*Ali masayı düşündü*." şeklinde olsaydı, "*düşünmek*" yüklemi insanlara özgü olduğu için tümce sorunsuz olurdu.

Verilen örneklerden görüldüğü üzere yüklemle uyum sadece kavrama bağlı olmayıp bu kavramın hangi öbekte olduğuna da bağlıdır. Örneğin sıvı kavramını içeren "*su*" sözcüğü için; "*Su yavaşça akıyordu.*", tümcesinde su "*akmak"* yüklemiyle özne olarak uyumludur. "*Yolda yürürken suya düştü.*", tümcesindeki su "*düştü"* yüklemiyle yönelme öbeği olarak uyumludur.

Anlatım bozukluğu sorununa yüklem ve yüklem dışındaki kısımların uyumluluğu açısından baktığımız için, yüklemin aldığı eklerin etkisi de incelenmiştir (Bybee, 1985), (Banguoğlu, 1986), (Korkmaz, 1994). Çatı ekleri yüklemin aldığı öbekleri doğrudan etkilemektedir. Yüklemin aldığı şahıs ekinin tümcenin öznesi ile uyumlu olması ve yüklemin aldığı zaman ekinin tümce içindeki zaman bildiren öbekle uyumlu olması gerekmektedir. Bu üç çeşit ekin öbeklerle uyumluluğu, farklı modellerle gösterilerek anlatım bozukluğu bulmada ne kadar katkı sağladığı değerlendirilmiştir.

Çalışmamız, ilk aşamada bileşik ve basit tümceleri kapsayacak şekilde tasarlanmıştır ve tümce içindeki iç tümceleri kapsamamaktadır. Örneğin "elinde çiçekle gelen adam" öbeği tek bir öbek-kavram çifti ile ifade edilmektedir. Bu örnekte elinde çiçekle gelen adam insan kavramına sahip öznedir.

2. Öbekler ve Kavramlar

Çalışmamızda dilbilgisi ve anlam çözümlemesi "*öbek*" ve "*kavram*" olarak iki ana başlık altında incelenmektedir.

2.1. Öbek

Yüklemle uyumluluğu denetlenecek öbekler, Tablo-1'deki gibi belirlenmiştir (Aliyeva, 2004), (Kalkan, 2006). Bu sekiz öbeğin yedi tanesi ismin durum ekleriyle elde edilmiştir. Bu eklerin oluşturduğu öbekler dışında çalışmamız kapsamında -le, -la ekli araç öbeği de kullanılmıştır. Ayrıca ek almayan zarf öbeği de eklenmiştir.

Tablo-1. Tümce Çözümlemede Kullanılan Öbekler

Öbek	Ek
1-Özne	-
2-Belirtisiz Nesne	-
3-Belirtili Nesne	-(y)ı, -(y)i, -(y)u, -(y)ü
4-Yönelme Öbeği	-(y)e, -(y)a
5-Bulunma Öbeği	-de, -da, -te, -ta
6-Ayrılma Öbeği	-den, -dan, -ten, -tan
7-Araç Öbeği	-le, -la
8-Zarf Öbeği	-

Türkçedeki hal ekleri ve öbekler arasındaki ilişkilerin biçimsel dil temsil yöntemiyle gösterimi Tablo-2'de verilmiştir. Bu temsil P üretim kuralı, N sözdizim değişkeni, T uç sembol ve S başlangıç sembolü olmak üzere {P, N, T, S} şeklindedir (Backus, 1959), (Chomsky, 2002). Tablo-2'de görüldüğü üzere "i", "e", "de", "den", "le" hal ekleri, isim öbeği, isim, zarf öbeği ve zarf uç sembollerdir. S, P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, X ve Z sembolleri ise sözdizim değişkenleridir. λ boş karakteri ifade etmektedir.

$P = \{P_i \mid 1 \leq i \leq 8\}$ kümesi üzerindeki tüm permütasyonların kümesi ∏ olsun. Türetim kuralı $p \epsilon \prod$ için S, $S \leftarrow p\ Z$ şeklinde ifade edilir.

Tablo-2. Türkçedeki hal ekleri ve öbekler arasındaki ilişkilerin biçimsel dil temsili

$S \leftarrow p\ Z$	(S: Basit Tümce)
$P_1 \leftarrow X \mid \lambda$	(P_1: özne öbeği)
$P_2 \leftarrow X \mid \lambda$	(P_2: X'i gösterme durumu)
$P_3 \leftarrow Xi \mid \lambda$	(P_3: X'i gösterme durumu)
$P_4 \leftarrow Xe \mid \lambda$	(P_4: X'e yönelme durumu)
$P_5 \leftarrow Xde \mid \lambda$	(P_5: X'de bulunma durumu)
$P_6 \leftarrow Xden \mid \lambda$	(P_6: X'den ayrılma durumu)
$P_7 \leftarrow Xle \mid \lambda$	(P_7: X'i kullanma durumu)
$P_8 \leftarrow$ zarf öbeği | zarf | λ	
X ← isim öbeği | isim	
Z ← yüklem	

Türkçede öbek sırası esnektir dolayısıyla S tümcesi yukarda bahsedilen P_1, P_2, P_3, P_4, P_5, P_6, P_7 ve P_8 sözdizim değişkenlerinin tüm permütasyonları ve Z şeklinde oluşturulabilir. Türkçedeki öbek sıralamasındaki serbestlikle ilgili çalışma Hoffman (1995) tarafından yapılmıştır.

Tümce öbeklerine ayrılırken tümcenin bağlılık çözümlemesi sonuçları ve tümcenin öbekleri içeren biçimsel temsili kullanılmıştır. Çalışmamız kapsamında tümcenin içerdiği sekiz tip öbeği bulan "Öbek Bulucu" programı geliştirilmiştir.

2.2. Kavramlar

Tümce sekiz öbeğe göre ayrıştırıldıktan sonra her bir öbek için kavram çözümlemesi yapılmıştır. Her bir öbeğin hangi kavramla ilgili olduğu saptanmıştır. Varlıkların kavramsal olarak sınıflandırılması son zamanlarda ontolojilerle birlikte sıklıkla karşımıza çıkmaktadır. Lakoff (1990)'a göre varlıkları sınıflandırmak, fiziksel çevreye, kültüre bağlı farklı değişkenleri içeren kapsamlı bir süreçtir.

Çalışmamızda tümceleri kavram uyumluluklarına göre inceleyebilmek için, üç önemli nokta göz önünde bulundurularak kavram seçimi yapılmıştır. Bunlardan ilki tümceyi hangi kavramların en iyi temsil edebileceği, ikincisi kavramların birbirinden yeterince ayrık olması ve üçüncüsüyse hangi kavramların yüklem uyumluluğuna göre belirleyici olduğudur. Kavramlar seçilirken WordNet (Fellbaum, 1998) gibi değişik ontolojiler incelenmiş ve hiyerarşik yapının en tepesinden en uygun kavramlar seçilmiştir. Çalışmamızda kullanılan kavramlar Tablo-3'de gösterilmektedir.

Tablo-3. Kavramlar

1-Zaman	2-Yer	3-Gaz
4-Sıvı	5-İçecek	6-Yemek
7-Zaman	8-Vasıta	9-Eşya
10-İnsan	11-Hayvan	12-Bitki
13-Organizasyon	14-Soyut	15-Ölçü
16-Eylem	17-Bilinmiyor	

Kavram başlıkları belirlendikten sonra, her bir öbeğin hangi kavram grubunda olduğunu anlamak için "kavram listeleri" oluşturulmuştur. Bu listeler, kavramı içeren isim ve isim öbeklerini bulundurur. Bu listeler oluşturulurken BalkaNet'den (Stamou ve ark. 2002), (Bilgin ve ark. 2004), Türk Dil Kurumunun terim sözlüklerinden, klasik sözlüklerin tanımlarda geçen açıklamalardan ve özel terim sözlüklerinden yararlanılmıştır (Baytop, 1994).

Yüklemler için de ayrı bir sınıflandırma yapılmıştır. İsim öbekleri için belirlenen kavram sayısı, yüklemler için oluşturulan öbek-kavram sınıfları sayısıyla doğrudan bağlantılıdır. Çalışmamızda her bir öbek için 17 kavram kullanılmıştır. Her bir

yüklem 17 kavram x 8 öbek çiftinden hangisiyle ilişkilendirilebileceğine göre sınıflandırılmıştır ve böylece 136 yüklem sınıfı elde edilmiştir. Örneğin, zaman kavramını nesne olarak alan yüklemler, insan kavramını nesne olarak alan yüklemler, yer kavramını yönelme olarak alan yüklemler gibi yüklem grupları oluşturulmuştur. Yüklem sınıflarına örnek olarak, öznesi insan olan yüklemler sınıfında "okumak, düşünmek, yemek, ağlamak, yakınmak" gibi yüklemler bulunmaktadır. Belirtili nesnesi bitki olan yüklemler sınıfında "sulamak, koparmak, ekmek, dikmek" gibi yüklemler bulunmaktadır.

Yüklemin hangi durum ekini aldığına göre yüklem grupları listeleri Türk Dil Kurumundan sağlanmıştır (Kurumu, 2010). Türk Dil Kurumundan alınan yaklaşık 7500 yüklemden kökü fiil olan ve çatı eki almamış yaklaşık 1000 tanesi tek tek incelenerek hangi öbek-kavram sınıfına gireceği çalışma kapsamında belirlenmiştir.

Yüklemlerin çatı eki aldığında nasıl davranacağı da çalışma kapsamında kurallarla belirlenmiştir.

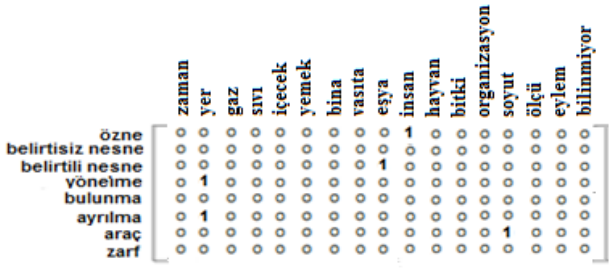
3. Model

Çalışmamızda yüklemle öbekler arasındaki ilişkileri tanımlayacak modelleme çalışmaları yapılmıştır. Bu modelleme çalışmaları ilgili bölümlerde açıklanmaktadır.

3.1. Tümcelerin Matris Gösterimi

Bildiğimiz gibi bilişim ortamlarının yaygın kullanımı ile birlikte, elde edilen büyük ölçekli verinin işlenmesi günümüzün en büyük gereksinimlerinden biridir (Cambria ve ark. 2013), (Procter ve ark. 2013). Bu konu doğal dil işleme alanında da büyük önem taşımaktadır. Doğal dil işleme çevrelerince öncelikle gereksinim duyulan konulardan biri metin verisinin, tümcelerin ya da sözcüklerin şu anki ayrık temsili yerine daha az boyutlu vektörle temsillerinin yapılabilmesidir.

Çalışmamızda, tümce 8x17'lık öbek-kavram matrisiyle ya da 136x1'lik öbek-kavram vektörüyle ifade edilmektedir. Şekil-1'de görüldüğü üzere "Ayşe, kırılan kalemi sevdiği evinden okula sevinçle götürdü." tümcesi için; "*Ayşe*" "*özne-insan*" çiftiyle eşleşmekte, "*kırılan kalemi*" "*belirtili nesne öbeği-eşya*" çiftiyle eşleşmekte, "*sevdiği evinden*" "*ayrılma öbeği-yer*" çiftiyle eşleşmekte, "*okula*" "*yönelme öbeği-yer*" çiftiyle eşleşmekte ve son olarak "*sevinçle*" "*araç öbeği-soyut*" çiftiyle eşleşmektedir.



Şekil 1. Tümcenin matris temsili

Matris elemanları 0 ya da 1 değerlerini almaktadır. Öbek-kavram çiftinin varlığı 1 ile bu çiftin tümcede bulunmaması 0 ile gösterilmektedir. Bu matris temsili sayesinde, O(1) karmaşıklığında dilbilgisi ve anlamsal hata tespiti yapılabilmektedir.

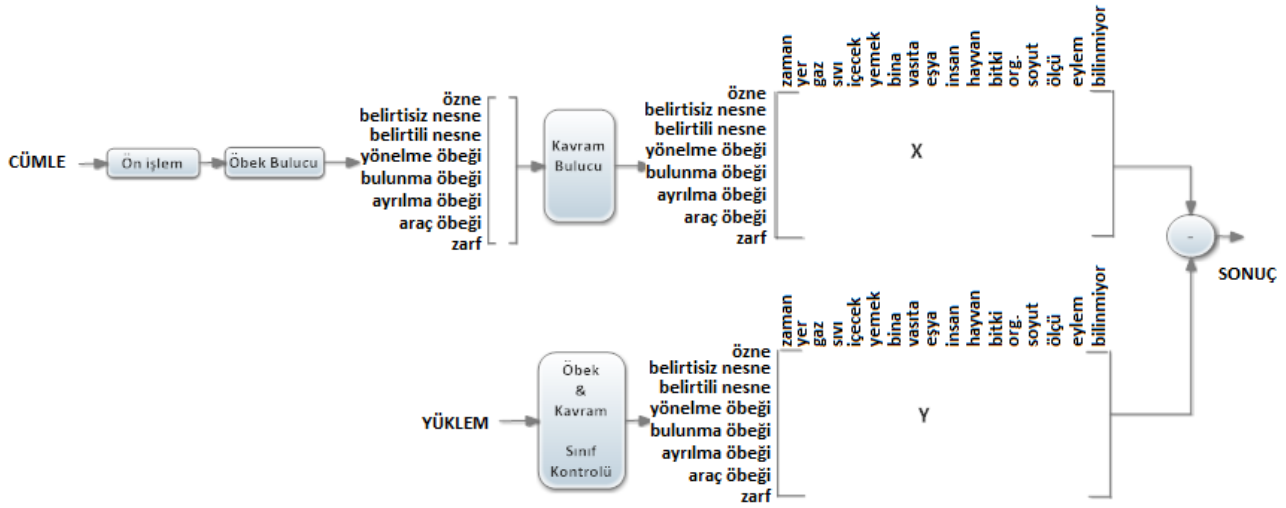
3.2. Ana Model

Çalışmamızda giriş verileri olacak tümceler öncelikle İTÜ Türkçe NLP Web servisinden (Eryiğit, 2014) geçirilmekte ve bağlılık çözümlemesi yapılmış tümceler elde edilmektedir.

Bağlılık çözümlemesi yapılmış giriş tümcelerimiz çalışma kapsamında geliştirilen "Öbek Bulucu" aracımızdan geçirilerek, varsa içerdiği sekiz öbek bulunmaktadır. Öbekler bulunurken bağlılık çözümlemesi sonuçları, bağlamdan bağımsız dil temsili ve durum ekleri kullanılmaktadır. Daha sonra "Öbek Bulucu" aracının sonuçları "Kavram Bulucu" aracından geçirilmektedir. Her bir öbeğin içerdiği kavram sınıfları kavram listelerinde taranarak belirlenmektedir. Sonuç olarak, tümce Şekil-2'de gösterilen ve "Kavram Bulucu" aracımızın çıktısı olan X matrisine dönüşmektedir.

Diğer taraftan, Tümcenin yüklemi, 136 ayrı öbek-kavram sınıflarında aratılmakta ve uyumlu olduğu öbek-kavram sınıfları 1, uyumlu olmadığı öbek-kavram sınıfları 0 olarak işaretlenerek Y matrisi oluşturulmaktadır. X matrisi tümcenin gözlenen

durumunu; Y matrisi ise tümcenin içermesi olası en genel halini ifade eder.



Şekil 2. Ana Model.

Eğer belli bir öbek-kavram çifti için Y elemanı 0, aynı öbek kavram çiftindeki X elemanı 1 ise sorun vardır. Bu yüklemin bu öbek-kavram çiftini kabul etmediği halde, bu öbek-kavram çiftinin tümcede bulunduğu anlamına gelir. Tümcenin doğruluğu (1) formülünde görüldüğü üzere X matrisinin tümleyeni ve Y matrisi toplanarak bulunur.

$$F = X^{-1} \vee Y \qquad (1)$$

Çalışmamızda sadece öbeklerin yüklemle uyumluluğunun denetimi birinci modeli, öbek-kavram çiftlerinin yüklemle uyumluluğunun denetimi ikinci modeli oluşturmaktadır. Yüklemin aldığı eklerin etkisi de üst üste konarak toplamda beş ayrı model oluşturulmuştur. Böylece her bir etmenin katkısının ayrı ayrı değerlendirilmesi sağlanmıştır.

3.3. Çatı Eklerinin Etkisi

Ana modeldeki Y matrisi yüklem kökünün uyumlu olduğu öbek-kavram çiftlerini ve Y_n matrisi yüklemin çatı ekleriyle birlikte uyumlu olduğu öbek-kavram çiftlerini göstermektedir. Her bir çatı eki için Y matrisinin nasıl Y_n matrisine dönüşeceği dilbilgisi kurallarına göre bellidir. Örneğin edilgen bir yükleme sahip tümcede nesne sözde özneye dönüşür ve tümce nesne içermez. Edilgen Y_n matrisi elde edilirken matrislerin her satırı bir öbeği gösterdiği için Y matrisinin nesneyi içeren ikinci ve üçüncü satırı sıfırla çarpılmalıdır ve Şekil-5'de de görüldüğü gibi T_3 dönüşüm matrisinin ilgili öbeğe ait satırları sıfır olmalıdır.

Çalışmamızda üçüncü model olarak, Y matrisi çatı eklerinin etkisiyle Y_n matrisine dönüşmekte ve tümcenin doğruluğu (2) formülü ile ifade edilmektedir.

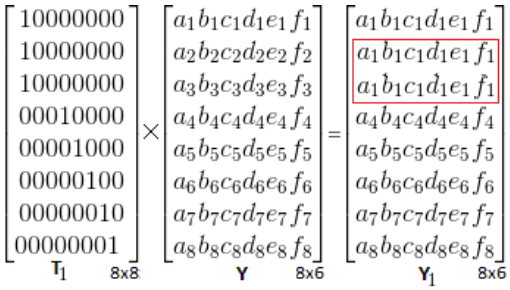
$$F = X^{-1} \vee Y_n \qquad (2)$$

Ana modelde oluşturduğumuz Y matrisinden yola çıkarak Y_n matrisinin nasıl elde edebileceği farklı çatı ekleri için anlatılmıştır.

Geçişli/Geçişsiz Yüklem: Geçişli bir yükleme sahip tümce nesne ve belirtili nesne alabilirken; geçişsiz bir yükleme sahip tümce nesne alamaz. Geçişli bir yüklem nesne-kavram sınıflarında bulunurken geçişsiz yüklem nesne-kavram sınıfında bulmaz. Örneğin, belirtisiz nesne-eşya alabilen yüklemler sınıfında geçişsiz bir yüklem olan *"büyümek"* yoktur.

İsim Kökünden Türeyen Yüklemler veya İsim Yüklemler: Eğer tümcede yüklem kökü "-len, -leş" eklerini almış isim köklü fiilse veya sadece isimden oluşmuşsa tümce nesne içermez. Bu tarz yüklemlerin saptanması için uygulamamızda yüklemin kökü incelenmekte ve eğer yüklem kökü isim türünde ise Y matrisinin ikinci ve üçüncü satırları sıfırlanmaktadır.

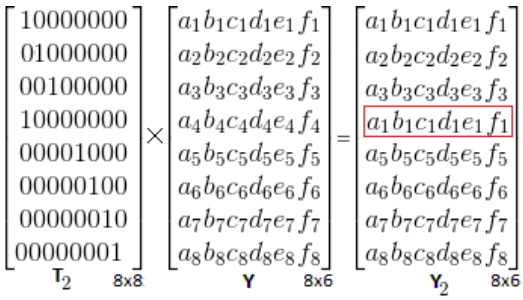
Oldurgan Yüklem: Oldurganlık ekiyle geçişsiz yüklem geçişli yükleme dönüşmektedir. Oldurganlık ekiyle birlikte geçişsiz haldeki özne, nesne olur. Bir başka değişle Y matrisinin birinci satırı, Y_1 matrisinsin ikinci ve üçüncü satırına yani belirtili ve belirtisiz nesneye geçmektedir. Şekil-3'deki gibi Y matrisi, T_1 dönüşüm matrisi ile çarpılarak, oldurgan Y_1 matrisi elde edilmektedir. Örnek: Ağaç büyüdü. → Ağacı büyüttü veya Ağaç büyüttü.



Şekil 3. Oldurganlık eki alan yüklemler için Y matrisinin dönüşümü

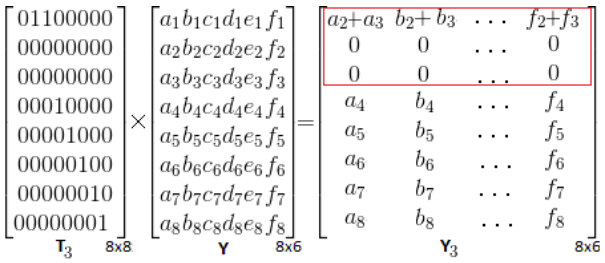
Gerçekte 8x17 boyutlu olan Y ve Y_1 matrisleri için 17 tane olan kavram sütunu 6 tane gösterilmiştir. Bu durumun nedeni matris çarpım ifadesinin şekille ifadesini kolaylaştırmaktır. Bu gösterim diğer Y ve dönüştürülmüş Y matrisleri için Şekil-4-5-6'da da kullanılmıştır.

Ettirgen Yüklem: Ettirgenlik ekiyle yüklem yönelme öbeği alabilecek hale gelir. Yönelme öbeğinin alabileceği kavramlar ise yüklemin eski halinin öznesinin alabileceği kavramlardır. Şekil-4'de görüldüğü üzere Y matrisi, T_2 dönüşüm matrisi ile çarpılarak ettirgen Y_2 matrisi elde edilmektedir. Sonuç olarak Y matrisinin birinci satırı Y_2 matrisinin dördüncü satırına geçer. Örnek: Ali ekmeği aldı. → Ali ekmeği Ayşe'ye aldırdı.



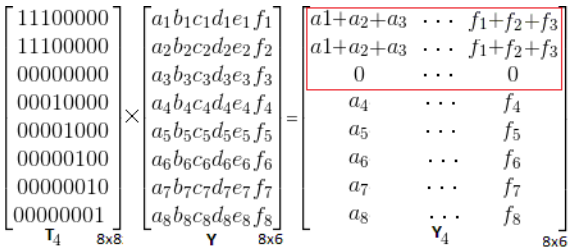
Şekil 4. Ettirgenlik eki alan yüklemler için Y matrisinin dönüşümü

Edilgen Yüklem: Edilgenlik eki almamışken nesne diye işaretlenen öbek, yüklemin edilgenlik eki aldığı tümcelerde özne satırına geçer (sözde özne). Şekil-5'de görüldüğü üzere Y matrisi ile T_3 matrisinin çarpımı Y_3 matrisini vermektedir. Örnek: Ali ekmeği aldı. → Ekmek alındı.



Şekil 5. Edilgenlik eki alan yüklemler için Y matrisinin dönüşümü

Dönüşlü Yüklem: Dönüşlü yüklemlerde olaydan etkilenen nesne ve olayı gerçekleştiren özne aynı olduğu için daha fazla bilgiyi Y matrisine aktarmak adına hem özne alanına hem de belirtisiz nesne alanına aynı değerler yazılmalıdır. Dönüşümlü yüklemin olduğu tümcelerde nesne belirtisiz nesne olarak kullanılır, belirtili nesne kullanılmaz. Örnek: "Ayşe tarandı." tümcesinde tarayan kişi Ayşe ve taranan kişi de Ayşe'dir. Şekil-6'da görüldüğü üzere Y matrisi ile T_4 matrisi çarpılarak Y_4 matrisi elde edilmektedir.



Şekil 6. Dönüşlü yüklemler için Y matrisinin dönüşümü

3.4. Zaman ve Şahıs Eklerinin Etkisi

Tümcedeki tüm öbeklerin yüklemle uyumluluğunun denetimini tamamlamak için zaman bildiren öbeğin yüklemle uyumluluğu ve şahıs bildiren öbeğin yüklemle uyumluluğu çalışmamız kapsamında incelenmiştir.

Türkçede yüklem aldığı şahıs ekleriyle tümcenin öznesi hakkında bilgi vermektedir ve yüklemin aldığı şahıs ekinin özne ile uyumlu olması beklenmektedir. Örneğin, "Ayşe ile ben okula

gittim.” tümcesinde yüklemin aldığı “–m” eki ile “Ayşe ile ben” öznesi arasında uyuşmazlık olduğu görülmektedir. Doğru olan “Ayşe ile ben okula gittik.” tümcesidir. Daha önce bahsedilen öbek-kavram denetimine ek olarak, yüklemin aldığı ekle şahıs bildiren öbeğinin uyumluluğunun denetimi dördüncü modeli oluşturmaktadır.

Çalışmamızda son olarak, tümcenin zaman bildiren öbeğinin, yüklemin aldığı zaman ekiyle uyumluluğunun denetimi yapılmıştır. Bu inceleme yapılırken, zaman kavramına sahip öbekler yedi farklı grup “geçmiş zamanda bir an, şu an, gelecek zamanda bir an, sıklık, geçmiş zamanda bir zaman dilimi, bugün, gelecek zamanda bir zaman dilimi” olarak kümelenmiştir (Demirgüneş, 2008). Daha sonra bir ön çalışmayla tüm zaman ve kip ek çeşitlerinin yedi farklı zaman türü sınıfından hangileriyle birlikte kullanılabileceği belirlenmiştir.

3.5. Uygulama Çıktısı

Çalışmamız kapsamında oluşturduğumuz uygulama bir tümcelik girişleri ya da tümcelerden oluşan metinleri giriş olarak kabul etmektedir. Anlatım bozukluğu incelemesi tümce bazında yapılmaktadır.



Şekil 7. Uygulama Çıktısı Örneği

Şekil-7’deki örnekte görüldüğü üzere, uygulamamız sonunda tümce, öbekler ve bu öbeklerin içerdiği kavramlar halinde gösterilir. Eğer tümce içinde uyuşmazlık yaratacak bir öbek veya bir kavram varsa, uygulama sorunlu öbeği göstermektedir. Yüklemin aldığı zaman eki ile zaman öbeği arasında uyuşmazlık olup olmadığı veya yüklemin aldığı şahıs eki ile özne arasında uyuşmazlık olup olmadığı da ayrıca gösterilmektedir.

4. Değerlendirme

Günlük hayatımızdaki metinlerde anlatım bozuklukları çok sık karşımıza çıkmamaktadır. Sıklığı az da olsa eğitim görmüş ve dil konusunda başarılı insanlar bile özellikle uzun tümcelerde anlatım bozuklukları yapabilmektedir. Modelimizi değerlendirmeden önce değişik kaynaklardan

aldığımız tümceler üzerinde anlatım bozukluklarının günlük hayatımızda hangi sıklıkla görüldüğü incelenmiştir.

4.1. Anlatım Bozukluğu Olayının İncelenmesi

Film alt yazıları (Open-Subtitle), akademik yayınlar ve Vikipedi'den alınan 300 tümcelik eşit boyda günlük metin elle sınıflandırılmıştır. Tablo-4'de görüldüğü gibi bu metinler; tümcelerin içerdiği ortalama sözcük sayısı (SS), tümcelerin içerdiği ortalama ek sayısı (ES), tümcelerin içerdiği ortalama öbek sayısı (ÖS) ve tümce türüne göre incelenmiştir.

Tablo-4. Günlük Metin Analizi

Kaynak Metin (*)	SS	ES	ÖS	Tümce Türü
Open Subtitles	11,3	5,83	3,58	%36,37 Basit %58,32 Bileşik %5,31 Bağlı
Akademik Yayınlar	10	7,5	3,25	%27,21 Basit %60,55 Bileşik %12,24 Bağlı
Vikipedi	7,93	5,37	2,31	%18,74 Basit %56,12 Bileşik %25,14 Bağlı

(*) Tablodaki kaynaklar, İTÜ DDİ araştırma grubu tarafından hazırlanmış derlemlerdir.

Sonuç olarak günlük hayattaki metinlerde tümcedeki ortalama sözcük sayısı 9 ve ortalama ek sayısı 6 olarak bulunmuştur. Tümceler ortalama olarak yüklem dışında 3 öbek daha içermektedir. En sık karşılaşılan tümce tipi bileşik tümce ve en az rastlanan tümce tipi bağlı tümcedir.

Tablo-5'de anlatım bozukluğu çeşitlerinin, günlük hayatta kullanılan metinlerde ve İngilizceden Türkçeye Google çeviri aracıyla çevrilen metinlerde görülme sıklığı karşılaştırılmıştır. Tablo-5'de görüldüğü üzere, günlük hayatta kullanılan metinlerde sadece % 12,37 oranında anlatım bozukluğu görülmektedir ve yapısal anlatım bozuklukları, sözcüksel anlatım bozukluklarının yaklaşık 1,5 katıdır. Günlük metinlerdeki düşük anlatım bozukluğu oranının aksine, Google çevirisinde çeviri tümcelerinin %87'si anlatım bozukluğu içermektedir. Ayrıca bu tümcelerdeki yapısal anlatım bozuklukları sözcüksel anlatım bozukluklarının 7 katından fazladır. Sonuç olarak günlük metinlerin %87,63 düzgün ifadelerden oluşmaktadır. Google çevirinde, düzgün ifadelerin oranı %13 olarak bulunmuştur.

Tablo-5. Anlatım bozukluğu sıklığı

Anlatım Bozukluğu Türleri	Günlük Metin %	Google Çeviri %
Sözcük Temelli Hatalar (STH)		
Yanlış sözcük kullanımı	2,68	7
Yanlış sözcük sıralaması	1,33	3
Fazla sözcük kullanımı	0,61	0
STH Toplam oranı	4,62	10
Yapısal Hatalar (YH)		
Bağlaçlar ve iç tümcelerde hata	0,83	21
İsim tamlamaları hataları	0,92	6
Mantıksızlık	0,54	13
Öbek-kavram çiftlerinin yüklem uyuşmazlığı *	2,80	28
Kişi ve zaman öbeklerinin yüklem uyuşmazlığı *	1,91	9
Çoğul ekinin yanlış kullanımı *	0,75	0
YH Toplam oranı	7,75	77
Toplam anlatım bozukluğu oranı	12,37	87

Tablo-5'de görüldüğü gibi çalışmamız "*" ile işaretlenen "öbek-kavram çiftlerinin yüklem uyuşmazlığı", "Çoğul ekinin yanlış kullanımı" ve "kişi ve zaman öbeklerinin yüklem uyuşmazlığı" maddelerini kapsamaktadır. Bu da günlük metinlerdeki yapısal hataların %70,45'ine eşittir. Bir sonraki çalışmamızda "bağlaçlar ve iç tümceler" ile ilgili hataların da değerlendirilmesiyle birlikte günlük metinlerdeki yapısal hataların kapsanan kısmının %81,16'ya çıkarılması hedeflenmektedir.

4.2. Modellerin Değerlendirilmesi

Çalışmamızda her bir etmenin katkısını ayrı ayrı ölçebilmek amacıyla beş ayrı model oluşturulmuştur. M1 modelinde yüklemin tümce içindeki öbeklerle uyumluluğuna bakılmaktadır. M2 modelinde öbek-kavram çiftlerinin yüklemle uyumluluğuna bakılmaktadır. M3 modelinde, yüklemin aldığı çatı eklerinin etkisi ikinci modelle birleştirilmiştir. M4 modelinde, üçüncü modele ek olarak yüklemin aldığı kişi ekleriyle özne uyumluluğunun denetimi yapılmaktadır ve son olarak M5 modelinde yüklem zaman eklerinin zaman öbekleriyle uyumluluğunun denetimi dördüncü modele eklenmiştir.

Tablo-6. Modellerin P, R, F ve A değerleri

	tp	fp	fn	tn	P	R	F	A
M1	99	71	14	23	0,58	0,87	0,69	0,58
M2	97	66	16	28	0,59	0,85	0,70	0,60
M3	98	61	15	33	0,61	0,86	0,72	0,63
M4	95	51	18	43	0,65	0,84	0,73	0,66
M5	95	47	18	47	0,66	0,84	0,74	0,68

Başarım değerlendirme aşamasında Türkçe Radyo Televizyon (TRT) Kurumu kayıtlarından elde edilen yapısal anlatım bozukluğuna sahip 94 tümce ve yine TRT kayıtlarından elde edilen 113 sorunsuz tümce kullanılmıştır. Her bir modelin hata bulmada gösterdiği başarı; Tutturma değeri (P), Bulma değeri (R), F-değeri (F) ve Doğruluk (A) değerleri olarak Tablo-6'de gösterilmektedir. Tablodan görüldüğü üzere öbek-yüklem uyumunun denetlendiği birinci model üzerine eklenen her bir etmen doğruluk değerimizi artırmıştır. M3 ve M4 modelleri %3 artış sağlarken, M2 ve M5 modelleri %2 artış sağlamıştır. Kullanılan başarı ölçüm değerlerinin nasıl hesaplandığı Tablo-7'de gösterilmektedir.

Tablo-7. Başarı Ölçüm Değerleri Hesabı

Toplam Örnek Tümce Sayısı	Gerçekte Doğru Olan Tümce	Gerçekte Sorunlu Olan Tümce
Modelin Bulduğu Doğru Tümce	**tp**: Gerçekte doğru olup doğru diye işaretlenen	**fp**: Gerçekte sorunlu olup doğru diye işaretlenen
Modelin Bulduğu Sorunlu Tümce	**fn**: Gerçekte doğru olup sorunlu olarak işaretlenen	**tn**: Gerçekte sorunlu olup sorunlu diye işaretlenen

Doğruluk değeri (Accuracy) =(tp+tn)/(tp+fp+fn+tn)

Tutturma değeri (Precision) = tp / (tp+fp)

Bulma değeri (Recall) = tp / (tp+fn)

F-değeri = 2*Tutturma*Bulma/(Tutturma+Bulma)

Tablo-5'deki anlatım bozukluğu sıklığı tablosundan görüldüğü üzere "yapısal anlatım bozukluklarından kapsadığımız * ile işaretli kısmın, yapısal anlatım bozukluklarının %70,45'ini oluşturması" bilgisi altında, elde ettiğimiz 0,68'lik doğruluk değeri oldukça yüksektir.

Akademik yayınlar arasında dilbilgisi analiz konusunda değişik çalışmalar mevcuttur fakat bu çalışmaların birbiriyle mantıklı bir şekilde kıyaslanması zordur. Bu çalışmalar aynı derlem üzerinde yapılsa dahi (ki biz Türkçe için bir çalışma yapmaktayız) eğitim verilerindeki hatalı tümcenin doğrusunun ne şekilde işaretlendiği ve neyin doğru olarak kabul edildiği (aynı tümce farklı şekillerde ifade edilebilir) başarı değerlerini etkilemektedir (Chodorow ve ark. 2012). Türkçe dilbilgisi alanında yapılan çalışmalar ise yazım hatalarının düzeltilmesi ile ilgili olduğundan çalışmamızla kesişmemektedir.

5. Sonuçlar

Çalışmamız yapısal anlatım bozukluklarının saptanması ve düzeltilmesiyle ilgilidir. Her ne kadar farklı kaynak dillerde çalışıldığı için kıyaslanması doğru olmasa da yapısal hatalar kapsamında İngilizce için yapılan çalışmalarda 0,60'lara varan doğruluk değerleri elde edilmiştir (Felice ve ark. 2014), (Lee ve Lee 2014), (Ng ve ark. 2014). Türkçe için ise yapısal anlatım bozukluklarının saptanması ile ilgili bir makale bulunamamıştır.

Türkçe yapısal anlatım bozukluklarının saptanması için %68 doğruluk değeri iyi bir başlangıç değeridir.

Çalışmamız ayrıca tümcenin öbeklerine ayrıldığı, içerdiği kavramların bulunduğu, içerdiği zaman türünün incelenip yüklemle kıyaslandığı, öznesinin tipinin, tekil ya da çoğul olduğunun incelendiği Türkçe tümce çözümleme kaynağı olmak hedefindedir.

6. İleriki Çalışmalar

Çalışmamızda, tümce öbek-kavram çiftleri olarak gösterilmektedir. Her öbek sadece bir birimi ifade etmektedir oysa birleşik tümcelerde öbekler içinde iç tümcelere sahip olabilmektedir. Örneğin "okula sevinçle gelen Ayşe" öznesi içinde farklı bir iç tümceyi içermektedir. İleriki aşamada, çalışmamız iç tümceleri ve bağlaçlarla bağlanmış tümceleri de içerecek şekilde genişletilecektir. İç tümceciklerin de kapsanması ile birlikte yapısal anlatım bozukluğu belirlemede 0,81 doğruluk değerlerine erişilmek hedeflenmektedir.

Çalışmamıza farklı bir açıdan bakıldığında, çalışmamız tümcelerin öbek-kavram şeklinde temsil

edilebilirliği yönünde bir pencere açmaktadır. Bu öbek-kavram temsilinin yanı sıra yüklemin anlamsal sınıfı da eklendiğinde bir tümcenin daha az boyutlu vektör temsili yapılabilecektir. Yüklemin anlamsal sınıfı ile kastedilen her bir yüklemi içerdiği anlama göre kümeleyen İngilizcedeki Levin (1993) sınıflarının Türkçeye uyarlanmış halidir. Levin sınıfları, iletişim bildiren fiiller, sahiplik değiştiren fiiller ve hareket içeren fiiller gibi fiil sınıflarını içermektedir. Bu şekilde tümceler vektör olarak temsil edilebildiğinde kural tabanlı bir yöntemle sadece vektör işlemleriyle metin benzerliklerinin saptanması ve metin gerektirimlerinin çıkarılması (Giampiccolo ve ark. 2007), (Padó ve ark. 2008), (Bentivogli ve ark. 2009) mümkün olacaktır.

Türkçenin düzenli tümce yapısı ve düzenli yüklem yapısı bu çalışmanın esin kaynağı olmasına karşın, öbek-kavram temsili tüm diller için kullanılabilecek bir yöntemdir.

Teşekkür

Çalışmanın başında Türkçe tümcelerin vektör temsili ile ilgili çalışma yapma konusunda fikir veren Doç. Dr. Deniz Yüret'e teşekkür ederim.

Kaynaklar

Dilsiz, S., 2015. Bulanık mantık ve yapay sinir ağları ile Türkçe yazım denetleyicisi. Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Eryiğit, G., 2014. İTÜ Turkish NLP web service. EACL 2014, 1.

Felice, M., et al., 2014. Grammatical error correction using hybrid systems and type filtering. CoNLL-2014, 15-16.

Lee, K. and Lee, G. G., 2014. POSTECH Grammatical Error Correction System in the CoNLL-2014 Shared Task. CoNLL-2014, 65.

Ng, H. T., et al., 2014. The conll-2013 shared task on grammatical error correction. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, 1-12.

Wang, Y., et al., 2014. Factored statistical machine translation for grammatical error correction. CoNLL-2014, 83.

Aygül, M., Karaalioğlu, G. and Amasyalı, M. F., 2014. Prediction of function tags of the simple Turkish sentences by conditional random fields. Sigma, 32, 23-30.

Isguder, G. G. and Adali, E., 2014. A pilot study on automatic inference rule discovery from Turkish text. In Application of Information and Communication Technologies, 1-5.

Cambria, E., et al., 2013. Big social data analysis. Big data computing, 401-414.

Ehsan, N. and Faili, H., 2013. Grammatical and context-sensitive error correction using a statistical machine translation framework. Software: Practice and Experience, 43, 187-206.

Procter, R., Vis, F. and Voss, A., 2013. Reading the riots on Twitter: methodological innovation for the analysis of big data. International Journal of Social Research Methodology, 16, 197-214.

Haaparanta, L. and Hintikka, J., 2012. Frege synthesized: essays on the philosophical and foundational work of Gottlob Frege. Springer Science & Business Media, Vol. 181.

Chodorow, M., et al., 2012. Problems in Evaluating Grammatical Error Detection Systems. COLING, 611-628.

Doğan, S. and Karaağaç, G., 2012. Dilbilgisel Bağdaşmazlık ve Anlamsal Tutarsızlık. İstanbul Aydın Üniversitesi Dergisi, 16, 25- 42.

Amasyalı, M. F., 2012. Kavramlar Arası Anlamsal İlişkilerin Türkçe Sözlük Tanımları Kullanılarak Otomatik Olarak Çıkartılması. EMO Bilimsel Dergi, 1, 1-14.

Kurumu, T. D., 2010. Büyük Türkçe sözlük. TDK. Ankara.

Bentivogli, L., et al., 2009. The fifth pascal recognizing textual entailment challenge. Proceedings of TAC, 9, 14-24.

Demirgüneş, S., 2008. Türkçedeki Zaman Belirteçlerinin Sınıflaması ve Dökümü Üzerine Bir Deneme Çalışması. Turkish Studies, 2, 278-293.

Padó, S., et al., 2008. Deciding entailment and contradiction with stochastic and edit distance-based alignment. In Proceedings of the Text Analysis Conference, Vol. 4.

Eryiğit, G., Nivre, J. and Oflazer, K., 2008. Dependency parsing of Turkish. Computational Linguistics, 34, 357-389.

Delibas, A., 2008. Doğal Dil İşleme İle Türkçe Yazım Hatalarının Denetlenmesi. Doktora Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Giampiccolo, D., et al., 2007. The third pascal recognizing textual entailment challenge. In Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing. Association for Computational Linguistics, 1-9.

Güngör, O. and Güngör, T., 2007. Türkçe bir sözlükteki tanımlardan kavramlar arasındaki üst-kavram ilişkilerinin çıkarılması. Akademik Bilişim Konferansı, 1, 1-13.

Kalkan, U., 2006. Türkiye Türkçesinde Hal (Durum) Kavramı ve Hal (Durum) Eklerinin İşlevleri.

Aliyeva, H., 2004. Türkiye Türkçesinde İsim Hâlleri ve Nesnenin İfade Vasıtaları. V. Uluslararası Türk Dili Kurultayı, 157-164.

Bilgin, O., Çetinoğlu, Ö. and Oflazer, K., 2004. Building a wordnet for Turkish. Romanian Journal of Information Science and Technology, 7, 163-172.

Izumi, E., et al., 2003. Automatic error detection in the Japanese learners' English spoken data. InProceedings of the 41st Annual Meeting on Association for Computational Linguistics, 2, 145-148.

Oflazer, K., 2003. Dependency parsing with an extended finite-state approach. Computational Linguistics, 29, 515-544.

Chomsky, N., 2002. Syntactic structures. Walter de Gruyter.

Stamou, S., et al., 2002. Balkanet: A multilingual semantic network for the balkan languages. In Proceedings of the International Wordnet Conference, 21-25.

Cebiroğlu, G., 2002. Sözlüksüz Köke Ulaşma Yöntemi. Doktora Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Bigert, J. and Knutsson, O., 2002. Robust error detection: A hybrid approach combining unsupervised error detection and linguistic knowledge. In Proc. 2nd Workshop Robust Methods in Analysis of Natural language Data, 10-19.

Fellbaum, C., 1998. WordNet. Blackwell Publishing Ltd.

Hoffman, B., 1995. The computational analysis of the syntax and interpretation of "free" word order in Turkish. IRCS Technical Reports Series, 130.

Baytop, T., 1994. Türkce bitki adlari sözlügü. Türk tarih kurumu, 578.

Korkmaz, Z., 1994. Türkçede eklerin kullanılış şekilleri ve ek kalıplaşması olayları. Türk Dil Kurumu, Vol. 598.

Levin, B., 1993. English verb classes and alternations: A preliminary investigation. University of Chicago press.

Lakoff, G., 1990. Women, fire, and dangerous things: What categories reveal about the mind. University of Chicago press, 1987-1987.

Atwell, E. S., 1987. How to detect grammatical errors in a text without parsing it. In Proceedings of the third conference on European chapter of the Association for Computational Linguistics, 38-45.

Banguoğlu, T., 1974. Türkçenin grameri. Baha Matbaası.

Bybee, J. L., 1985. Morphology: A study of the relation between meaning and form. John Benjamins Publishing, Vol. 9.

Backus, J. W., 1959. The syntax and semantics of the proposed international algebraic language of the Zurich ACM-GAMM conference. Proceedings of the International Comference on Information Processing.