

Türkçe Söylem İşaretleme Çalışması: ODTÜ-MEDİD

Deniz Zeyrek

Orta Doğu Teknik Üniversitesi

İşaretlenmiş derlem, dilbilimde kuramsal ilerlemelere yardımcı olabileceği gibi, dil teknolojilerine de katkıda bulunur. Şu ana kadar işaretlenmiş geniş çaplı tek Türkçe derlem OTD [ODTÜ Türkçe Derlem]'dir (Say ve diğ. 2002). OTD'de, metin türü, yazar, yayın yılı gibi bilgiler işaretlenmiştir. Bu yazıda, 2007 yılı Ekim ayında başlanan ve söylem bağlaçlarını işaretleyerek OTD'yi söylem düzeyine çıkarmayı amaçlayan ODTÜ Metin Düzeyinde İşaretlenmiş Derlem (ODTÜ-MEDİD) projesi'nin hedefleri ve kaydettiği ilerlemeler özetlenecektir.¹

1.0 Artalan: Söylem

Söylemi, tümce ötesi ve tümceler arası dilsel ilişkilerin, yapı ve örüntülerin gösterildiği bir bütün olarak alıyoruz. Söylem terimini ise, sözlü ve yazılı dil arasında bir fark gözetmeden kullanıyoruz. Söylem çalışmalarının kabaca ikiye ayrıldığını söyleyebiliriz. Bir yaklaşım çerçevesinde, söylemdeki küçük ölçekli yapısal ilişkiler ve anlamlar araştırılmakta (örn., Webber ve diğ. 2003, Grosz ve diğ. 1995, Polanyi 1988), diğer yaklaşım çerçevesinde ise anlatı, makale, haber gibi büyük ölçekli bütünlerin alt birimleri arasındaki yapısal ve anlamsal ilişkiler üzerine yoğunlaşmaktadır (örn. Dijk 1977, 1980, Beaugrande ve Dressler, 1980). Söylem çalışmaları, söylemin sistemli bir dilbilgisi yapısına sahip olduğunu ve belli yapıtaşları üzerine kurulduğunu ortaya çıkarmıştır. Bağlılık ve tutarlılık, gönderim, sezdirim ve çıkarım, yüklem zaman, kiplik, görünüş gibi özellikleri bunların en önemlileridir. Ayrıca, bilgi yapısının söylemi biçimlendiren bir mihenk taşı olduğu anlaşılmıştır. Söylemin sadece yapısal özelliklerden değil, dikkat ve niyet bileşenlerine sahip olduğu da belirtilmiştir (Grosz ve Sidner 1986). (bkz Turan ve diğ., basımda.)

M.A.K. Halliday, söylemin yapısını irdelediği öncü çalışmalarında, tutarlılık için bağlaşıklık şart olduğunu (fakat bunun yetmediğini) yazar (1985: 318). Halliday'e ve diğer pek çok araştırmacıya göre bağlaşıklık ve tutarlılığı en açık olarak ortaya koyan öğeler bağlaçlardır. Esasen bağlaşıklık ve tutarlılık, örneğin sözlüksel bağlantılarla da sağlanabilir (eşadlılık, eşanlamlılık, zıt anlamlılık gibi). Ancak bağlaçlar, söylemdeki bağlaşıklık ve tutarlılık bağlarını her hangi bir kuşkuyla yer vermeden, açıkça ortaya koyan sözcükler olduklarından çoğu söylem kuramında önemli bir yer tutar.

2.0 ODTÜ-MEDİD'in Amaçları ve Dilbilimsel Yaklaşımı

ODTÜ-MEDİD projesinin genel amacı, söylem tutarlılığını küçük ölçekli düzlemde araştırmaktır. Projede söylem bağlaçları (SB) ve söylem bağlaçlarının bağladıkları metin aralıkları işaretlenmektedir. Proje PDTB'in (Penn Discourse TreeBank, Prasad ve diğ. 2007) ilkelerini paylaşmakta ancak Türkçe'nin özellikleri nedeniyle kimi noktalarda PDTB'den ayrılmaktadır.

Söylem Bağlacı ve Üyelerinin Tanımı: Yapısal ve Anlamsal ölçütler

ODTÜ-MEDİD, yapısal yönden SB'lerin sadece ve sadece iki üye alan yüklem olduğu görüşünden yola çıkar. Bu metin aralıklarına üye denmektedir. Türkçe'de SB'ler yana sıralama bağlaçları (örn., *ve*, *ama*, *fakat*), alta sıralama bağlaçları (örn., *ilgeç* ve *ulaçlar*) ve söylem belirteçlerinden (örn., *ne var ki*, *nihayet*, *en sonunda*) oluşmaktadır. Tümce belirteçleri, yani sadece bir tümceyi üye olarak alan sözcük/sözcük öbekleri (örn., *maalesef*,

¹ Projeyi destekleyen TÜBİTAK'a, ayrıca katkılarından dolayı Işın Demirşahin, Ayışığı B. Sevdik-Çallı, Ümit D. Turan, Cem Bozşahin, Ruket Çakıcı, Berfin Aktaş, Hale Ögel-Balaban ve İhsan Yalçınkaya'ya teşekkür ederim. 23. Dilbilim Kurultayı'nda bildiri dinleyerek soru ve yorumlarıyla katkıda bulunan dinleyicilere de teşekkür borçluyum.

*ne yazık ki, allah*tan) ve söylemi düzenleme amacı taşıyan kelimeler (örn., *şey*) SB sınıfına girmez. Anlam yönünden SB'ler, zıtlık, düzeltme, örnekleme, neden-sonuç gibi anlam bağları kuran tümce ya da tümcecikleri birbiriyle ilişkilendiren sözcük/sözcük öbekleridir. Önemli bir ölçütümüz, üye olarak seçilecek metin aralıklarının soyut nesne, yani olay, olgu, durum, gerçek, olasılık, soru, dilek, emir, önerme olmaları gerekliliğidir (Asher 1993). Bu sınıflamaya girmeyen bağlaçlar ODTÜ-MEDİD'de işaretlenmemektedir. Örneğin, *Ahmet'in Zeynep'le iyi anlaşmasının nedeni sanata ve estetiğe değer vermesidir* tümcesinde ve bağlacı, iki ad öbeğini bağladığından SB görevini görmez.

SB'lerin üyelerine, 1. üye ve 2. üye denmektedir. Bu terimler üyelerin metinde soldan sağa sıralanmalarıyla ilgili değildir. Bağlacı içeren üye her zaman 2. üye, diğer üye 1. üyedir. Üyelerin belirlenmesi için en önemli ölçüt, bağlaç anlamıyla ilişkili görülen metin aralıklarını bulmaktır. İkinci ölçüt ise, üye olarak belirlenen metin aralığının kapsamıdır. Metin aralığı kapsamının belirlenmesi için anlamsal bir ilke olan yeterlik ilkesi göz önünde bulundurulmaktadır, yani, bağlaç anlamını tam olarak vermeye yetecek en kısa metin aralığı üye olarak seçilmektedir.

İlkelere göre gerçekleştirilmiş işaretlemelerden bazıları aşağıda örneklenmiştir. Örneklerde bağlaçlar alt çizgiyle, 1. üye yana yatık, 2. üye ise kalın harflerle gösterilmektedir. Örneklerden sonra parantez içinde OTD dosya numarası verilmiştir. İlk örnek, yana sıralamalı bağlaç *ve*'yi ve üyelerini göstermektedir (örnek 1). Bu örnek, üye olarak bir tümcecigi alabilen *inan-* eylemini içermektedir. Yeterlik ilkesi nedeniyle, *inan-* eyleminin adlaşmış üyeleri *ve* bağlacının üyeleri olarak seçilmiş, *inan-* eylemi işaretleme dışında bırakılmıştır.

- (1) *Romanı ilerletemediğine, kimin katil, kimin maktul olacağına bir türlü karar veremediğine* ve ayrıca **roman anlayışının detektif romanlarına asla uygulanamayacağına** inanmaktadır. (00026231)

Sonraki örnek (2), *rağmen* bağlacının işaretlemesini göstermektedir. İlk olarak, bu bağlacın bağladığı metin aralıklarının aynı özneyi paylaştığı görülmektedir. Bu nedenle özne (Rukiye) işaretleme dışında bırakılmıştır. İkinci olarak, 2. üyenin biçimbirimsel özelliği, yani adlaşmış bir eylemden oluşan bir tümcecik olması projemiz için önemlidir. Bu tür yan tümceciklerden oluşan üye/ler Türkçe'de pek çok bağlacın sergilediği bir yapıdır. Böyle sık görülen bir yapının gözden kaçırılmaması için adlaşmış yapıların bağlaç üyesi olarak seçilmesine karar verilmiştir. Bu karar, ODTÜ-MEDİD'i PTDB'den ayıran bir özelliktir.

- (2) Rukiye, **kendisinden üç yaş ufak olmasına** rağmen, *erkek kardeşini kendi oğlu sanıyordu*, ... (00062211)

Aşağıdaki örnekte ise, yana sıralamalı bağlaç *ama* ve bağlacın üyeleri verilmiştir. Bu örnek, bağlaca bitişik olmayan bir tümcenin 1. üye olarak seçilebileceğini gösterir. Örnekte [1], [3] ile zıtlık kurduğundan sol üye olarak seçilmiştir. Soldaki diğer tümce [2], anlamsal olarak [1]'in bir tür açılımıdır; bu nedenle işaretlenmemiştir. Bu yapıya OTD'de *aksine, aksi halde, yine de, ayrıca* bağlaçlarında sıkça rastlanmaktadır. Hatta bazı örneklerde 1. üye bağlaçtan 3-4 tümce uzakta olabilmektedir.

- (3) *Aşıklı'ya yerleşenler ilk geldiklerinde kerpiç mimariyi biliyor.* [1] Yapılarını kerpiçten yapıyorlar, [2] ama sonra taşı kullanmayı öğreniyorlar. [3] (00013112)

3.0 İşaretleme Ön Çalışması ve İşaretleme Süreci

OTD'de bulunan metin türleri bilgisi Tablo 1'de verilmiştir. Ön çalışma olarak, OTD 4 örnekleme ayrılmıştır. Her örneklem OTD'deki metin türlerini dosya ve sözcük sayısı bakımından eşit derecede temsil etmektedir. Proje bitiminde örneklem 1'in işaretlenmiş olması hedeflenmektedir.

Tablo 1: OTD’de ve ¼ örnekleme de metin türü bilgisi

Tür	Dosya sayısı	Yüzde	¼ Örneklem I’de dosya sayısı	Yüzde
Roman	123	15.63%	31	15.74%
Öykü	114	14.49%	28	14.21%
Araştırma-inceleme	49	6.23%	13	6.60%
Makale	38	4.83%	9	4.57%
Gezi	19	2.41%	5	2.54%
Söyleşi	7	0.89%	2	1.02%
Anı-Günce-Hatıra	18	2.29%	4	2.03%
Haber	419	53.24%	105	53.30%
TOPLAM	787	100.00%	197	
Ortalama sözcük sayısı	2,000,000		400,000	

İşaretleme süreci başlamadan önce işaretleme ilkeleri hazırlanarak proje ekibine tanıtıldı. İşaretleme ilkeleri işaretleyicilerin dil sezgilerine müdahale etmeden, genel kurallara değinmekte ve gerekirse yeni kurallarla güncellenmektedir. İlkelerin belirlenmesiyle eş zamanlı olarak seminerler düzenlenerek alanyazıdaki söylem yazıları ve Türkçe’ye özgü konular proje ekibince tartışıldı. Sekiz aylık bir zaman dilimine yayılan bu süreçten sonra işaretleme aracı geliştirildi. Projenin ikinci yılından itibaren işaretlemeyle başlandı. Şu anda işaretleme, en az 3 işaretleyicinin belirlenen bağlaçları işaretleme, işaretleyiciler arası uyum istatistiği yapılması ve uyumsuz noktaların proje ekibince çözümlenmesi şeklinde devam etmektedir.

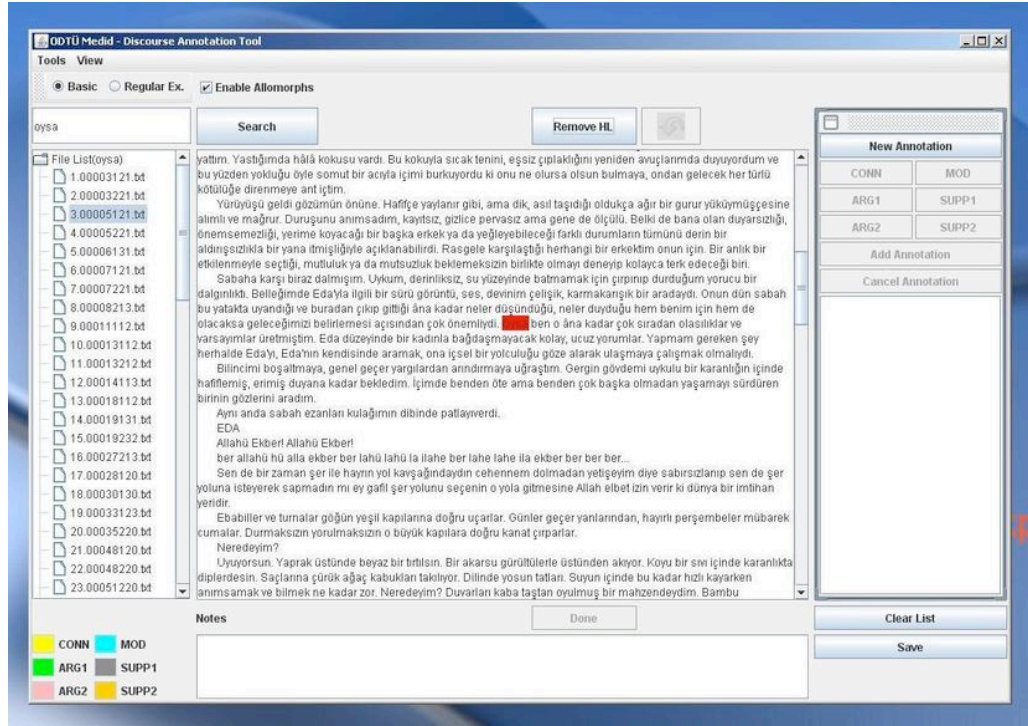
İşaretleme Aracı

İşaretleme için kullanılan araç, Türkçe için geliştirilmiştir.² Bu yazılım, gerek yalın sözcüklerden oluşan *ama*, *fakat*, *oysa* gibi bağlaçlar, gerek *hem ... hem* gibi kesintili bağlaçlar, gerekse 2. üyede adlaştırma ekleri bulunan *için* türünden bağlaçlar için sorgulama yapma olanağı vermektedir. Ünlü uyumu ve ünsüz ötümsüzleşmesi nedeniyle farklılık gösteren adlaştırma eklerinin tek sorguda aranması, D, A ve H harfleriyle mümkündür. Örneğin, -*mAsH* için sorgusu ile *için* bağlacının -*mesi/-ması* içeren üyesi bulunabilir. Bu sorgunun yapılabilmesi, doğru bağlacın bulunması için önemlidir. Örneğin, amaç anlamındaki *için* ile neden anlamındaki *için* bağlacı ayrı ayrı sorgulanabilmektedir. (krş. *Ailesi, Ayşe’nin başarılı olması için ona her türlü desteği sağlıyor – Ali toplantıya zamanında gelmediği için önemli bir bilgiyi öğrenemedi.*)

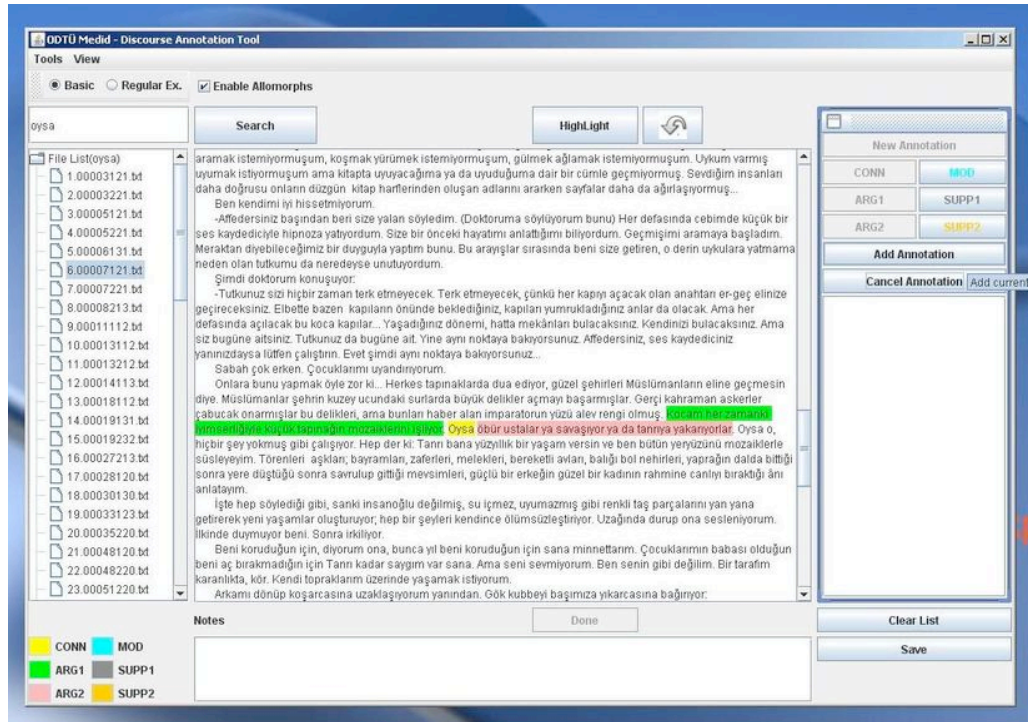
Resim 1’de işaretleme aracı *oysa* sorgusu sonucunda elde edilen dosyalar ekranının solunda ve bir dosyanın içerdiği metin ekranının ortasında görülmektedir. Resim 2’de, *oysa* bağlacı ve üyelerinin işaretleme gösterilmektedir. Sağdaki kutucuk içindeki uygun seçenek düğmeleri aracılığıyla, bağlaç, 1. ve 2. üye, bağlaç niteleyeni ve destekleyici metin aralıkları seçilebilmektedir. Ekranın en altındaki notlar bölümüne ise çeşitli konular yazılabilir.

² İşaretleme aracı, yüksek lisans tezi kapsamında Berfin Aktaş tarafından geliştirilmiştir.

Resim 1: İşaretleme aracında *oysa* sorgusu

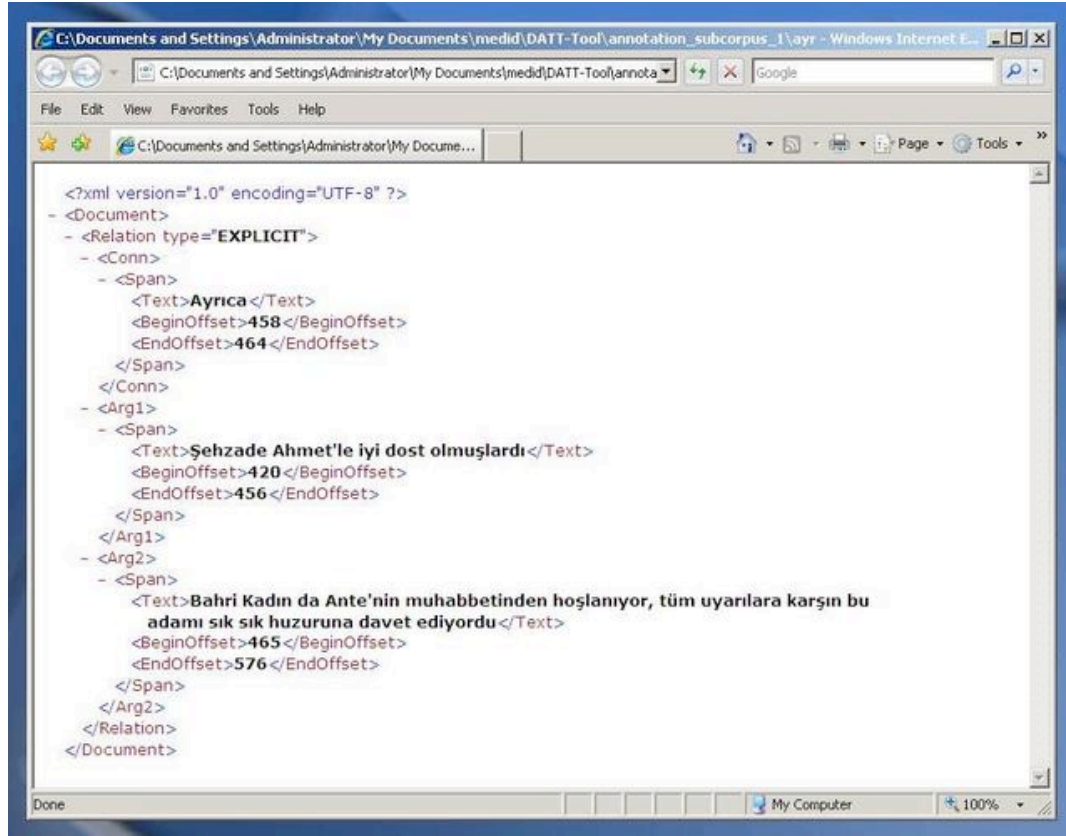


Resim 2: *Oysa* bağlacının işaretlenmesi



İşaretleme yazılımı, işaretleme sonucunda her bir dosya için bağlaç, üyelerin metin aralıkları ve bu aralıkların metin dosyalarındaki karakter yerlerinin kaydedildiği bir .xml dosyası oluşturur. Resim 3 bu dosyadan bir örnek sunmaktadır.

Resim 3: İşaretleme aracında karakter yerlerinin kaydedildiği .xml dosyası



ODTÜ-MEDİD'in değişik metin türlerinin OTD'deki dağılımını yansıtan 1. sürümü kullanıcı dostu bir arayüzle 2010 yılında akademisyenlerin hizmetine açılması planlanmaktadır.

4.0 Sonuç: Kazanımlar ve Yeni Araştırma Konuları

Proje bitiminde, Türkçe'ye söylem düzeyinde işaretlenmiş bir derlem kazandırılmış olacaktır. Bu ürün, orta ve uzun vadede çeşitli doğal dil işleme çalışmalarına kaynak oluşturabilir. Dilbilimsel yönüyle proje, söylem bağlaçlarını ve üyelerini işaretlemekle birlikte, söylem tutarlılığının sadece bağlaçlarla ortaya çıkarılabildiğini iddia etmemektedir. Projede belli bir kuram sıkı sıkıya izlenmeden, küçük ölçekli söylemde tutarlılık ilişkilerini ortaya çıkaracak yeterlikte işaret yapılması, ileride yapılacak söylem çalışmaları için özgün araştırma soruları üretilmesine yardımcı olacak, böylelikle söylem yapısının bilinmeyen özellikleri aydınlanacaktır. Projenin başlangıç noktası, tümce düzeyindeki yüklem-üye yapısının söylem düzeyinde de geçerli olacağı yaklaşımdır. Bu başlangıç noktasına karşın, projenin söylemde tümce sözdizimine benzeyen yapılar aradığı izlenimi doğmamalıdır. İşaretlemelelerde sözdizimsel bilgiler uygulamaya konsa bile, seçilen metin aralıklarının birbiriyle anlam bütünlüğü sağlaması konusu her türlü etkenin önünde yer almaktadır. Sözdizim ve söylemin nerede birbirinden ayrıldığı konusu bu projenin ortaya çıkardığı ve araştırmaya değer bir sorudur.

Kaynakça

Aktaş, Berfin. 2008. Computational Aspects of Discourse Annotation. Yayınlanmamış Yüksek Lisans Tezi, Bilişsel Bilimler Ana Bilim Dalı, Enformatik Enstitüsü, Orta Doğu Teknik Üniversitesi.

Asher, Nicolas. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers.

- Beaugrande, Robert de – Dressler, Wolfgang. 1980. *Introduction to Text Linguistics*. Norwood, N.J. :Ablex.
- Dijk, Teun A. van.1977. *Text and Context*. London: Longman.
- Dijk, Teun A. van. 1980. *Macrostructures*. Hillsdale, N. J.: Erlbaum.
- Grosz, Barbara J. – Sidner, Candace L. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics* 12: 175-204.
- Grosz, Barbara, Aravind Joshi, and Scott Weinstein. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse.*Computational Linguistics* 2/21: 203-225.
- Halliday, M.A.K. 1985. *An Introduction to Functional Grammar*. London, Great Britain, Victoria, Australia, Baltimore, Maryland, U.S.A.: Edward Arnold.
- Prasad, Rashmi, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Bonnie Webber. 2007. The Penn Discourse Tree Bank 2.0 Annotation Manual. The PDTB Research Group. www.seas.upenn.edu/~pdtdb.
- Polanyi, Livia. 1988. A Formal Model of the Structure of Discourse. *Journal of Pragmatics* 12, pp. 601-638.
- Say, Bilge, Zeyrek, Deniz, Oflazer, Kemal, Özge, Umut. 2002. Development of a Corpus and a TreeBank for Present-day Written Turkish. K. İmer ve G. Doğan (haz.). *Proceedings of the Eleventh International Conference of Turkish Linguistics*, Eastern Mediterranean University, Cyprus, August 2002. 183-192.
- Turan, Ümit Deniz, Zeyrek, Deniz, Bozşahin, Cem. (basımda). *Söylem ve Tutarlılık Bağları*. TDK Yayınları.
- Webber, Bonnie, Stone, Matthew, Joshi, Aravind, Knott, Alistair. 2003. Anaphora and Discourse Structure. *Computational Linguistics* 29(4), pp. 545-587.