DERLEM, BİLGİSAYAR DESTEKLİ SÖZLÜK BİLİMİ, EŞ DİZİMLİLİK VE OTOMATİK TERİM ÇIKARIMI

Bekir Tahir TAHİROĞLU*

Özet

Bilişim teknolojilerinde son yirmi yılda yaşanan gelişmeler dil bilimi çalışma ve uygulamalarını artan bir biçimde etkilemektedir. Bilgisayar destekli dil bilimi, doğal dil işleme (DDİ), derlem dil bilimi gibi görece yeni terimler özellikle metin çözümleme başta olmak üzere dil birimlerinin otomatik belirlenimi, çıkarımı ve bu uygulamalar üzerinden yoruma dayalı çalışmalar için vazgeçilmez nitelikler taşımaktadır.

Genel Ağ'da (İnternet) milyonlarca sayısallaşmış metinlerden belgelerden bilgi çıkarımında, arama-sorgulama uygulamalarında terimler önemli göstergeler, ipuçlarıdır. Terimlerin otomatik çıkarımı, elle yapılması mümkün olmayan yığınlarca sayısal metnin kavram çözümünde kullanılan yöntemler arasındadır. Bu bakımdan terime dayalı çözümlemeler yalnızca dil bilgisel çalışmalarda değil örneğin biomedikal uygulamalardan büyük kavram ağaçlarının hazırlanmasına kadar genişleyen bir alanda kullanılmaktadır.

Bu çalışmada TÜBİTAK popüler kitaplarından Şaşırtan Varsayım adlı çeviri yayının eş dizimli terimler açısından olasılık-istatistik yöntemler kullanan iki yazılımca çözümlemesi yapılmıştır. Sonuç olarak otuz dokuz eş dizimli birimin terim olarak değerlendirilebilecek adaylar olduğu belirlenmiştir.

Anahtar kelimeler: *Derlem, Doğal Dil İşleme, Bilgisayar Destekli Dil Bilimi, Otomatik Terim Çıkarımı, Sözlük Bilimi*

Corpus, Computer-assisted Lexicography, Collocation and Automatic Term Extraction

* Uzman, Çukurova Üniversitesi, Türkoloji Araştırma ve Uygulama Merkezi

Abstract

Developments in information technologies have effected linguistic studies in terms of varies linguistic techniques in the recent twenty years. Computer-assited linguistics, natural language processing (NLP), corpus linguistics are significant terms in recent linguistic literatures. These terms and their application areas convey important meanings in text mining based inferences and considerations about language itself.

Terms are also important culprit in text analyzing by means of internet. Internet contains millions amount of text conveying valuable linguistic knowledge, and those texts are not only used in linguistic survey but is also used from biomedical works to concept extraction studies. Corpus based linguistics and other computer-related fields are expanding their application areas.

In this artcile, TÜBİTAK's (The Scientific and Technological Research Council of Turkey) book called Şaşırtan Varsayım has been investigated about collocational terms. Thirty-nine terms is meaningful collocational candidate term. Sophisticated softwares which is capable of probabilistic analyzing properties has been used in the work.

Key words: *Corpus, Natural Language Processing, Computer-assisted linguistics, Automatic Term Extraction, Lexicography*

❑

Giriş

Modern dil biliminin kurucusu F. de Saussure'ün ders notlarından oluşan ve birçok dile çevrilen *Genel Dilbilim Dersleri* adlı yapıtında dilin nitelikleri üzerine kapsamlı kuramsal açıklamalar yer almaktadır. 20. yüzyılın ikinci yarısının hemen başında, *Sözdizimsel Yapılar, Dil ve Zihin* ve diğer çığır açıcı yapıtlarıyla Chomsky, Saussure'den sonra dil biliminde ikinci büyük kuramsal dönemi sürdürmüştür. Chomsky ile birlikte dilin nitelikleri mentalistik ve felsefi yöntemlerle açıklanmaya çalışılmış, var olanın dışında olası dilsel üretimler de üretici-dönüşümsel kuram doğrultusunda incelenmiştir.

Doğal Dil İşleme (Natural Language Processing) (DDİ) çalışmalarının günümüzde bilgisayar bilimi ve yapay zekâ çalışmalarında giderek bir uzmanlık alanı olduğu bilinmekle birlikte, dille ilgilenen araştırmacıların bu alanın çalışma yöntem ve sorunlarını dikkate almaları, geliştiren yöntemleri incelemeleri son yıllarda iyice önem kazanmıştır. DDİ'de geliştirilen bir biçim bilimsel çözümleyici, söz dizimsel çözümleyici, anlam belirsizliğini gideren yazılımların geliştirme sürecindeki karmaşık yapı ve sıra dizgesel (algoritmaya dayalı) düzen dikkate değerdir. Geleneksel dil bilgisi incelemelerinin ses bilimi, biçim bilimi, söz dizimi ve anlam bilimsel düzeni, DDİ'de de aynı sıra içinde ele alınmakta, yazılan kitaplar bir dil bilgisi kitabından farklı olmamaktadır. Dolayısıyla, burada ayrıntısına girilmeyecek daha birçok gelişme, DDİ ile dil bilimi sınırını ortadan kaldırmaya başlamıştır denilebilir.

Dil felsefesi ve dilin zihinsel üretim durumlarının açıklanması sürerken, özellikle 90'lı yıllardan başlayarak bilişim ve bilgi teknolojilerindeki çok hızlı gelişmeler dil biliminde, özellikle araştırma yöntemlerinde kendini göstermiştir. "Var olanın" ya da "doğal olarak oluşan"ın (naturally-occuring) önemi bu süreç içinde oldukça artmıştır. Günümüzde dil biliminin uygulamalı alanında yer alan sözlük bilgisi ya da sözlük biliminde kullanılan başlıca yöntem derleme dayalı bilgi çıkarımıdır.

Bilgisayarlı dil biliminin ya da bilgisayar destekli dil bilimin araştırma ve özellikle uygulama yöntemlerinden yararlanılarak ortaya konmuş "dil ürünleri" ya da "dil teknolojileri" arasında metin özetleme, çeviri yazılımları, elektronik sözlükler başta gelmektedir.

Dil biliminde kullanılan olasılık kuramlarıyla *cümle bölümleme yazılımlarında* (parsing software) başarım oranı çok daha yukarılara taşınmıştır. Otomatik biçim bilimsel çözümlemenin ardından, Türkçe için cümle düzeyindeki otomatik çözümleme yaklaşım ve çalışmaları da sürdürülmektedir.

Bugün gelinen noktada, klasik dil bilgisel çalışma yöntemlerinin bütünüyle terkedilmeye başlanması yanında, dil bilimcinin sezgisel bilgisine dayalı çıkarımlarının gerek dil bilgisi yapıtlarının hazırlanmasında gerekse sözlüklerin yazılmasında ne kadar doğruluk taşıdığı da sorgulanır olmuştur. Kısaca dil bilimi ve dil bilgisel yöntem, gözlemlenebilen, istatistiksel ve sayısallaştırılıp sonradan veri tabanı olarak kullanılabilecek "veri" temelli uygulamalar yönünde hızla ilerlemekte, bilgi teknolojilerindeki hıza paralel olarak gelişmeye ve değişime açık bir yapıya bürünmektedir.

Bu çalışmamızda, kullandığımız başlıca iki yazılımla, TÜBİTAK Popüler Bilim Kitapları serisinden *Şaşırtan Varsayım* adlı kitabı, içerdiği eş dizim nitelikli terimler açısından incelenmiştir.

Materyal ve Yöntem

Çalışmada, bilimsel nitelikli çeviri bir metin üzerinde duruldu. Metin, Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) yayınlarından Nobel ödüllü yazar Francis Crick'in *Şaşırtan Varsayım* adlı kitabıdır. 332 sayfalık kitap, yaklaşık 19 bin sözcükten oluşmaktadır.

Nörobiyolojinin kimi konularını ele alan kitapta, çeviri terimlerin neler olabileceği, *Güncel Türkçe Sözlük*'te (GTS) yer almayan terimlerin de bulunabileceği düşüncesiyle de incelenmeye çalışılmıştır. Çalışmanın ilerleyen bölümlerinde de söz edildiği üzere, eş dizimli yapılar çıkarılmış, türemiş tek birimler inceleme dışı bırakılmıştır. Kitap taranıp Optik Karakter Tanıma (OKT) ile düz metin biçiminde saklanmıştır.

Laurence Anthony'nin AntConc (sürüm 3.1.302) adlı metin işleme yazılımıyla, Wirote Aroonmanakun'un özel olarak eş dizimlilikleri çıkarmak için hazırlanmış olan *Collocation Extract* (sürüm 3.06) adlı yazılımı kullanılmıştır. AntConc'ta, DDİ'de n-gram olarak adlandırılan birliktelikler sıklıklarına göre çıkarılmakta; Collocation Extract'te ise daha özelleşmiş istatistiksel yöntemler kullanılarak en olası

birliktelikler bulunabilmektedir. Her iki yazılımda ücretsiz olarak araştırmacıların kullanımına sunulmuştur. Bunun yanında, her iki yazılımda Türkçe karakter seti desteği sunmaktadır.

Dil bilgisi çalışmalarında sıkça kullanılan fişleme ve sınıflandırma yöntemi tarihteki yerini almıştır diyebiliriz. Bu bakımdan bu çalışmada önümüzdeki yıllarda dil bilgisel çalışmalarda temel araştırma yöntemi olarak kullanılacağını düşündüğümüz bilgisayarla işleme adını verebileceğimiz yöntem kullanılmıştır. Kullanılan yazılımlar: **AntConc: (**http://www.antlab.sci.waseda.ac.jp/software.html) ve **Collocation Extract: (**http://pioneer.chula.ac.th/~awirote/colloc/) adlı yazılımlardır.

"Terim" Kavramı

"Özel ve belirli bir kavramı karşılayan sözcüklerden olan terimler, herhangi bir bilim, sanat, meslek vb. alanlarda, bu alanlarla ilgili kimseleri tek bir dil göstergesine bağlayan kullanımlardır" (Özkan 2006: 269). Bu anlamda, bilimsel literatürde geçen *radyum, nöron, hücre çekirdeği, eklemleme, art zamanlı* vb. sözler terim niteliklidir (Aksan 1998: 40).

Vardar (1988) terimi, "özel bir bilgi ya da etkinlik alanına, bir bilim, uygulayım ya da uzmanlık dalına özgü sözcük" olarak tanımlarken terimlerin uzmanlar arasında iletişim sağlayıcı, etkin ögeler olduğunu söyler. Terimler bu yönüyle alanın uzmanlarınca kullanılan özel söz varlığı ögeleri olarak da nitelenebilirler.

Bir dilin konuşurları ve o dili bilim dili olarak kullanan okur-yazarca ortak kullanılan söz varlığı birimleri –örneğin: kitap, dergi, ağaç, dere, dağ vb.- yanında, *konu-duyarlı* birimler, yine genel söz varlığı içinde; ama özel bir alan içinde kabul edilebilecek terimleri oluştururlar.

Tıpta *sinir hücresi*, botanikte *çanakyaprak*, dil biliminde *biçim birimi*, *derlem* örneklerinde görüldüğü gibi, terimler genelde tek anlamlı birimlerdir. Terimlerin bu anlamsal özellikleri dışında türetilmeleri de çeşitli yollarla yapılmaktadır: **a)** Var olan bir sözlüksel birime yeni anlamlar yükleme: Diş hekimliğinde *köprü* gibi. **b)** *Bilgisayar, damariçi, dil bilimi, duygu durumu* gibi örneklerde birleştirme yoluyla **c)** Çeviri yoluyla: *ısıölçer, sinir hücresi, çevrimiçi*. **d)** Türetme: *saplantı, zorlantı, duygulanım, duyarga, benzeşme* (Aksan 1998: 41)

Terim bilimi (terminology), " Terimleri inceleyen, terim yaratımıyla ilgili sorunları ele alan uyulamalı dil bilimi dalıdır." (Vardar 1988: 200). Terimlerin taranarak bir araya getirilmesi ve sözlüklerinin hazırlanması da bu dalın uğraşı alanında yer alır. Son yıllarda *bilgisayarlı terminoloji* (computational terminology) adıyla *bilgi çıkarım* (information extraction) teknolojilerinin kullanılmaya başlandığı görülmektedir. Konuyla ilgili olarak 2004 yılında bir de çalıştay düzenlenmiştir (http://www.new.biomath.jussieu.fr/~pz/computerm2004.html).

Derlem ve Eş Dizimlilik

Derlem (corpus), "Basit anlamda, metinler bütünü; daha yaygın ve geniş

kullanımıyla *bilgisayarca okunur* (machine-readable) metin bütünü; daha da dar anlamıyla sınırlı boyutta, bir dilin bütününü ya da bir değişkesini büyük oranda yansıtabilen bilgisayarca okunur metinlerdir." (McEnery ve Wilson 2004: 197).

"Derlem, büyük ve belirli prensipler doğrultusunda oluşturulmuş metinlerdir." (Biber vd. 2000: 12). Görüleceği üzere, Türkçe dil bilimi yazınına son yıllarda *bütünce* olarak giren, ancak özellikle DDİ araştırmalarıyla uğraşan akademisyenlerce *derlem* olarak yaygınlaştırılan bu kavram, kabaca metinler bütünü anlamına gelmektedir. *corpus* ve onun çoğulu *corpora* terimleri batı literatüründe kullanılan terimlerdir.

Derlem, metinler bütünü olmakla birlikte, bu bütün bir "havuz" olarak düşünülmemelidir, çünkü rastgele seçilmiş, her şeyin birbiri içinde olduğu karmaşık bir bütün değildir derlem. 1930'lu yıllarda Amerikan yapısal dil biliminde de kullanılan korpus (corpus) terimi, o yıllar için -bilgisayarların olmadığı düşünülürse- elle toplanmış ve sınıflandırılmış notlardan oluşan bütünlerdir. Bu açıdan, terimin bu ilk kullanılmaya başlandığı yıllardaki durumunu çağrıştıracak bir kullanımından çok, derlem terim olarak, yukarıdaki tanımlarda geçen "bilgisayarca okunur" anlamını taşır. Gerçekten de, günümüzde insanın anlayabileceği ancak bilgisayar diline dönüştürülmemiş yapılar derlem olarak kabul edilmemektedir. Geliştirilen çeşitli belge işaretleme dilleri sayesinde (Bunlardan bugün en çok kullanılanı SGML'den türetilen XML'dir.) yapılandırılmamış bir metin, hemen her ögesi, birimi, üzerinde birçok sorgulama, sıralama işleminin yapılabileceği bir yapıya kavuşturulabilmektedir. İşte bilgisayar dilince bu tür işaretlenmiş metinler gerçek derlemler olarak değerlendirilmektedir.

Derlem çalışmalarının öncüleri Henry Kucera ve Nelson Francis'tir. Amerika'da Brown Üniversitesinde 1967'de oluşturulan derlem *Brown Derlemi* olarak adlandırılmış, bu derlem üzerine Amerikan İngilizcesinin ayrıntılı istatiksel çözümlemeleri gerçekleştirilmiştir. Bu derlemi, Oxford Üniversitesinde oluşturulan ve 100 milyon işaretlenmiş sözcüklük İngiliz Ulusal Derlem'i (British National Corpus-BNC) izlemiştir. Şu anda Amerikan Ulusal Derlem'i (American National Corpus-ANC) hazırlanmakta, 100 milyon işaretlemiş sözcüğün hedeflendiği çalışmanın 22 milyon sözcüklük bölümünün işaretlendiği bilinmektedir.[1]

Türkiye'de derlem oluşturma çalışmaları oldukça yenidir. ODTÜ Enformatik Enstitüsü tarafından oluşturulan işaretlenmesi yapılmış derlemin boyutu 2 milyon sözcüktür. Son olarak Dokuz Eylül Üniversitesi Bilgisayar Mühendisliği Bölümü bir proje olarak derlem oluşturma çalışmalarını sürdürmektedir. Bunların yanı sıra Türkçe için bireysel çabalarla oluşturulmuş ve amaca uygun olarak işaretlenmiş derlemler de vardır.[2]

Derlem hazırlamada kullanılan birçok yazılım bulunmakta, bunların bir bölümü ücretsiz olarak dağıtılmaktadır.[3] Burada ayrıntısına girilmeyecek olan bu yazılımlarla,

1 http://americannationalcorpus.org/

2 Çukurova Üniversitesi, Eğitim Fakültesi, Türkçe Eğitimi Bölümünde yaklaşık 12.5 milyon sözcük içeren kabaca işaretlenmiş ve sözcük temelli sorgulama yapılabilen bir derlem üzerinde çalışılmaktadır. Bu derlem çalışmasının ilk ürünü "Türkçede Belirteçlerin Fiillerle Birliktelik Kullanımları ve Eşdizimliliği" konulu bir doktora tezidir (http://sosyalbilimler.cukurova.edu.tr/tez/1082/).

3 http://www.tei-c.org/Software/

kullanıcı dostu bir arayüz aracılığıyla metinler biçim bilimsel, söz dizimsel, anlam bilimsel ve söylem özellikleri ve metnin üst-bilgileri bakımından kolayca işaretlenmektedir.

Dil biliminde, *derlem-tabanlı (*corpus-based*)* yöntemler ağırlık kazanmakta, dil öğretiminde en çok kullanılan metin türlerinden hareketle söz varlığı belirleme çalışmaları artan bir hızda sürdürülmektedir. Derlemin dil öğretiminde nasıl kullanılacağı, hangi yapıların öncelikle öğretilmesi gerektiği, dil öğrenicisine yönelik sözlüklerde sözcük seçimi gibi konular, derlem-tabanlı, gerçekleşmiş dil kesitlerinden yararlanılarak işlenmektedir.

Derlem dil bilimiyle (DD) ilgili olarak yazılan kitap sayısı her geçen gün artmaktadır. İngilizcede yazılan kitap sayısı yüzü aşmıştır. Türkçede için şu anda henüz bu uzmanlık konusuyla ilgili kitap bulunmamaktadır.

Sonuç olarak, DD, başlı başına bir uzmanlık alanı olup, dünyada bu alanda çalışma yapan dil bilimci, bilişimcilerin sayısı artmaktadır. Bu bağlamda Türkiye'de de bu alanla ilgili tanıtıcı, bilgilendirici yayınların sayısının artırılması, çevirilerin yapılması ve uzmanların yetiştirilmesi ivedilikle gerekmektedir.

Eş Dizimlilik

Eş dizimlilik (collocation) "iki ya da daha çok sayıda dil biriminin genellikle aynı dizimlerde yer alması" olarak tanımlanmaktadır (Vardar 1988: 98). Kavram üzerinde yeterince çalışma yapılmadığından kuramsal olarak birtakım sorunlar yaşanmaktadır. Dil biliminde, dizimsel (syntagmatic) ve dizisel (paradigmatic) yapılar olarak kabul edilen dil dizgesinin iki düzleminde, eş dizim, dizimsel bölümde yer alan en az iki birimin bir arada anlamlı bir biçimde bulunmasıdır. *kapıyı açmak, kafayı yemek, dışarı çıkmak, mavi tren, bıçak bilemek, akşam eve geç gelmek, başı ağrımak* gibi, kimi deyim olarak adlandırılan yapılar birer eş dizimsel yapıdır. Çizgisel olarak aynı çerçeve içinde bulunurlar.

Dil biliminde eş dizim olarak adlandırılan bu yapılar DDİ'de *n-gram* yapıları olarak ele alınır. N-gramlar *söz katarlarının* (string) belirlenmesinde, yazıda ya da sözde geçen bir birimin bir önceki ve bir sonraki birimle olan yakınlığının, çeşitli olasılık formülleri yardımıyla bulunmasında önemlidir. *bi-gram, trigram* gibi terimler ikili, üçlü *n-gram*'ları karşılamak için kullanılmaktadır. Basit n-gramlar yanında karmaşıkları HMM (Gizli Markov Modeli-Hidden Markov Model), *Maximum Likelihood Estimation* (MLE) gibi başlıca olasılık yöntemleriyle ham bir derlemden çıkarılmaya çalışılmaktadır (Jurafsky ve Martin 2000: 194-198).

Eş dizimlilik konusu, temelde bir dil bilimsel olasılık konusudur. Bir derlemden çıkarılan eş dizimli yapıların niteliklerinin açıklaması, kurucu birimlerin türleri, dil bilgisel bir arada bulunurluk formülleri sonradan verilebilir. Bugün bir ham (yapılandırılmamış/işaretlenmemiş) derlemden eş dizim çıkarmak için dil bilgisel yöntemlerden çok istatistiksel yöntemler uygulanmaktadır.[4]

4 Eş dizimlilik, birliktelik kullanımı ve derlem-tabanlı bir uygulama için bk. Bk. Özkan, Bülent (2007), "Türkiye Türkçesinde Belirteçlerin Fiillerle Birliktelik Kullanımları ve Eşdizimliliği", Çukurova Üniversitesi, Sosyal Bilimler Enstitüsü, Türk Dili ve Edebiyatı Anabilim Dalı, Yayınlanmamış Doktora Tezi, Adana. (Danışman: Prof. Dr. Mehmet Özmen).

Bilgisayar Destekli Sözlük Bilimi

Sözlükler, genel olarak 'bir dilin eş zamanlı ya da art zamanlı söz varlığı bilgisini daha çok basılı biçimde sergileyen yapıtlar' olarak tanımlanabilir.

Türkçede daha çok "sözlükçülük" olarak adlandırılan sözlük hazırlama çalışmaları için günümüzde *sözlük bilgisi* ve *sözlük bilimi* terimleri kullanılmaktadır. Bu iki terim birbirinden farklı içeriklere sahiptir. *Sözlük bilgisi* (lexicograhy), "sözlük hazırlama teknikleri, ilkeleri ve uygulama ile ilgili koşulları ele alan ve sözlük biliminin alt alanı olan dal" olarak tanımlanmaktadır (Hengirmen 1999: 341). Sözlük bilgisi tıpkı dil bilgisinin dil bilimin bir alt uygulama alanı olarak düşünülebileceği gibi sözlük bilimin verilerinin uygulandığı alan olarak da düşünülebilir. *Sözlük bilimi* (lexicology) ise, "sözlük hazırlamanın kuramsal yönüyle ilgilenen, sözlüğün kapsamı, içeriği, sınırı üzerine kuramsal sonuçlara yönelik çalışmaların yapıldığı dilbilim alanı"dır (Hengirmen 1999: 341). Görüleceği üzere, sözlük bilgisi daha çok "uygulamaya", sözlük bilimi ise "kurama" dayalı alandır. Bu iki alanın bilgilerinin bir araya gelmesiyle uygulamada somut olarak kullanılabilen dilsel ürünler ortaya konulmaktadır.

Geniş bir salonu dolduran hantal bilgisayarların yerine taşınabilir bilgisayarların hemen her yerde kullanılabildiği günümüzde, bilgisayarlar uygulamalı çalışmaların yürütüldüğü bilim dallarında her geçen gün önemini arttırmaktadır. Öyle ki "bilgisayar destekli" terimi doğrudan "bilgisayarlı" sözcüğüyle yer değiştirmekte, bu durum özellikle dil bilimi için daha da belirginleşmektedir. DDİ'nin bilgisayarsız yapılamayacağı gerçeği bu durumu ortaya açıkça koymaktadır. Bilgi teknolojilerindeki sürekli gelişim ve değişim sonucunda Genel Ağ'ın (İnternet'in) ilk günlerindeki tanıtım amaçlı bilgi sunumu, yerini uzunca bir süreden beri etkileşimli-dinamik bilgilendirmeye bırakmıştır. Veri tabanına dayalı, kanıtlanabilir, kolayca dönüştürülebilir bilgileşim ortamında sözlüklerin hazırlanmasında da büyük değişimler ve dönüşümler yaşanmaktadır.

Bilgisayarlı Sözlük Bilimi'nde (computational lexicology), yararlanılan temel kaynaklardan biri derlemdir. İşaretlenmiş bir derlemden elde edilen söz türü bilgileri *sözlüksel veri tabanları*nda (lexical databases) ilişkisel bir biçimde depolanmaktadır. Anlamsal olarak da işaretlenmiş derlemlerden çok daha fazla bilgi bu tür veri tabanlarına aktarılarak, sonradan karşılaşılacak metinlerden *sözlüksel girdi* (lexical entry) bilgileri var olanla karşılaştırılabilmektedir. DDİ'de, metinde geçen cümle, paragraf ve daha büyük dilsel birimlerin (sözce, söylem) anlaşılması, kestirilmesi için de başvurulabilinecek bir söz varlığı üst bilgisine (meta bilgi) gereksinim duyulmaktadır. Bu türden bilginin derlenip bir araya getirildiği yapıların adına *söz dağarcığı-leksikon* (lexicon) denilmektedir. Leksikonların oluşturulmasında var olan sözlüklerin elektronik biçimi ya da bilgisayarca okunur biçiminden (MRD=Machine Readable Dictionary) yararlanılmaktadır. Bu türden sözlükler içinde adı en çok anılan sözlük "Longman Dictionary of Contemporary English" adlı sözlüktür.

Bilgisayarlı bir leksikonda kurulan sözlüksel girişin yapısında *1. Söz türü, 2. Anlam sayısı, 3. Alt sınıflama (çatı bilgisi), 4. Anlamsal özellikler* bulunur. Temel olarak bu 4 bilgi dışında, *sesletim bilgisi, bağlam ve stilistik bilgisi, köken bilgisi,*

kullanım bilgisi, kısaltma bilgileri de bilgisayarın bir metinden sözlüksel yapıyı çıkarabilmesi için gerekebilmektedir.

Örnek bir sözlüksel girdi bilgisi şu şekilde verilebilir:

Sözcük: *menekşe*

1- **Söz türü:** İsim

2- **Anlamsal özellik:** [+somut]

3- **Bağlam:** Bitki

Sözcük: *Olgu*

1- **Söz türü:** İsim

2- **Anlamsal özellik:** [+soyut]

3- **Bağlam:** Felsefe

Son yıllarda, bilgisayarlı sözlük biliminin kendi içinde çeşitli uygulama alanlarına ayrılmakta olduğu görülmektedir. Sözlük bilimcinin uygulamada, sözlük yazımında ya da madde başı niteliklerinin yazılmasında kullanabileceği çeşitli yazılımların gerçeklenmesi ayrı bir uğraşı alanıyken, olası madde başı adaylarının belirlenmesinin otomatikleştirilmesi diğer bir uğraşı kolu olarak görülmektedir.

Daha geniş sözlükler hazırlamak için, bilgisayarların dilsel bilginin ya da bir yapının var olan sözlüklerden, metinlerden yararlanılarak kullanılması DDİ'nin görece yeni bir alanıdır. İstatistik ve olasılık kuramlarının bu alana uygulanmasıyla var olan sözlük kaynaklarından bilgisayarca okunur sözlüklerin hazırlanması için çok önemli gelişmeler yaşanacağı açıktır. Bugünkü aşamada güçlü yazılımlar kullanarak büyük bir başarı oranıyla geniş söz listelerinin elde edilmesi sıradan bir iş durumuna gelmiştir; ancak yine de insan denetimine gereksinin duyulmaktadır. Yine, bugünkü bilgilerimiz öngörüsünde, insan olmadan sadece bilgisayarın olduğu bir mekanizma kurmayı düşünenler için önlerinde uzun bir yol durmaktadır (O'grady vd. 1997: 648-650).

Otomatik Terim Belirleme

Belirli bir bilim alanında konuya duyarlı ya da o alanı temsil eden birimlerin yani terimlerin belirlenmesi dil bilimi için önemli olduğu kadar DDİ ve *bilgiye erişim* (information retrieval) çalışmalarında vazgeçilmez nitelik taşımaktadır. Google başta olmak üzere bilginin aranıp çıktının çeşitli biçimlerde sunulduğu "arama-çıkarma" ortamlarında, 2000'li yılların başlarından itibaren bilgi çöplüğü ya da gürültüsünden arındırılmış, hedef aramaya yönelik duyarlı sonuçların elde edilmesi üzerinde durulmaktadır. *Anlamsal ağ (semantic web)* kavramı içinde yer alan çalışmalarda, İnternet'te yer alan bilimsel bilginin daha duyarlı sorgulanmasında, terimler temsil edici olarak yer almaktadır. Böylelikle anlamsal olarak ya da konuya duyarlı sözlerle sınıflanmış, sınırlandırılmış Genel Ağ, kişilerin aradığını daha zahmetsizce bulduğu bir ortam olacaktır.

Otomatik terimleme adı altında özetleyebileceğimiz çalışmalarda güdülen iki amaç vardır:

1- Var olan terimleri tanıma ve çıkarma,

2- Yeni ve olası terimleri tanıma ve çıkarma (Valderrábanos vd.).

Önceden hazırlanmış işaretlenmemiş bilimsel metin bütünlerinden terim belirlemek için kullanılan yöntemler: Kural tabanlı ve istatistiksel yöntem olmak üzere ikiye ayrılır. Kural tabanlı yöntemde, işlenen dilde daha önceden terim olan birimlerin türetim kuralları belirlenir. Bu yöntem, dil bilgisel yöntem olarak da adlandırılır ve özellikle yeni terimlerin belirlenmesinde etkili olduğu bilinmektedir. Metin bütününde sıklığı yüksek birimlerin terimlenmesi görece daha kolaydır. Bununla birlikte sıklığı düşük yeni terimlerin gözden kaçırılması sorunu karşımıza çıkmaktadır.[5]

İkili ve üçlü ögelerin belirlenmesinde Goldmann ve Wehrli, söz dizimsel kural tabanlı yöntemin söz dizimsel işaretleme ile yapılmış derlemde, istatistiksel yönteme üstün olduğunu belirlemiştir. Çalışmalarını Fransızca Liberation gazetesinin bir milyon sözcüklük derlemi üzerine yürütmüşler ve 170.000 eş dizimli öge belirlemişlerdir. Geliştirdikleri sistemin adı FipsCo'dur.[6]

Başta İngilizce olmak üzere diğer birkaç dil için geliştirilen çevrimiçi terim belirleme yazılımı "Gensen Web"'dir. Bu yazılımla İnternet adresi verilen ya da doğrudan girilen metinlerin çözümlemesi yapılabilmektedir.[7]

Türkçe için bu çalışmanın yapıldığı sırada hazırlanmış bir otomatik terim belirleme yazılımı bulunmadığı gibi, literatürümüzde bu alana yönelik herhangi bir Türkçe makale de bulunmamaktadır.

Terimsel yapılar belge türlerinin belirlenmesinde de kullanılırlar. "Bir belge, birbirinden anlamsal olarak bağımsız olan terimler vektörüyle ifade edilir. Bu vektördeki bir terimin göreceli ağırlığı terimin derlem ve belge sıklığına bağlı olarak çeşitli şekillerde hesaplanabilir. Terimler vektörünün taşıdığı ağırlık değerlerinin belirlenmesinde pek çok yöntem kullanılır. Fakat bu yöntemlerin hemen hemen hepsi iki önemli noktaya dayanır: Bir terim, bir belgenin içinde ne kadar sık geçerse, o belgenin bir kategoriye atanmasında o kadar etkili olur ki, buna kısaca *terim sıklığı* (tf) denir. Ne kadar çok farklı belgede bulunursa, o sözcüğün ayırt edici özelliği o kadar azalır ki, buna da *ters belge sıklığı* adını vermekteyiz" (Sevet ve Bolat 2006: 176).

Çalışmamızda kullandığımız yöntem istatistiğe dayalı yöntemdir. Kullandığımız yazılım (Collocation Extract) üç olasılık yöntemi kullanmaktadır: 1- Dunning's log likelihood, 2- Pearson's Chi-square, 3- Mutual Information.[8] Seçtiğimiz kitapta uygulanan varsayım Dunning's log likelihood varsayımıdır.

5 http://www.cervantes.es/seg_nivel/lect_ens/oesi/liquid02/documet_pdf/automatic_terminolog_extraction_validation_liquid_approach.pdf

6 http://www.federation-nlp.uqam.ca/publications/01/goldman.pdf

7 http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb_eng.html

8 Yöntemlerle ilgili daha ayrıntılı bilgi için bkz.: http://pioneer.chula.ac.th/~awirote/colloc/statmethod1.htm.

5. Metinden Çıkarılan Eş Dizimli Terim Adayları

Aşağıdaki tabloda, log likelihood sayısı en yüksek eş dizimli birimler yer darlığı nedeniyle iki tablo biçiminde verilmiştir. Terim olarak kullanılan ya da kullanılabilecek birimler kalın ve italik olarak işaretlenmiştir.

SÖZCÜK 1	SIKLIĞI	SÖZCÜK 2	SIKLIĞI	SÖZCÜK 1 VE 2 SIKLIĞI	LOG LIKELIHOOD SKORU
ya	382	da	799	328	2925.9345
çok	643	sayıda	127	84	674.5713
belli	183	bir	3066	130	659.74062
kısa	112	dönemli	64	49	608.43551
görüş	***105***	***alanının***	***64***	***47***	***583.11823***
en	238	azından	43	42	494.46023
pek	167	çok	643	65	423.41178
tam	103	olarak	436	50	393.74588
ne	248	kadar	178	49	376.19847
tek	202	bir	3066	96	373.92134
beyin	***255***	***kabuğunun***	***40***	***34***	***367.8502***
bir	3066	biçimde	193	91	352.46639
o	240	halde	47	33	336.79229
hem	93	de	723	47	325.59359
her	327	iki	239	49	314.074
aynı	172	anda	78	32	299.64732
çoğu	100	kez	64	27	286.68645
şaşırtan	21	varsayım	24	17	275.09628
daha	605	çok	643	69	263.13518
new	19	york	16	15	263.05885
university	22	press	50	18	255.35108
açık	70	seçik	19	18	253.91139
dönemli	***64***	***bellek***	***56***	***22***	***251.24683***
görme	***277***	***sistemi***	***47***	***27***	***249.43547***
değil	180	de	723	47	249.14395
johnson	16	laird	18	14	241.56319
y	35	z	20	16	239.09409
son	54	derece	22	17	233.66947
akla	46	uygun	67	20	230.6451
olursa	13	olsun	20	13	229.24038
bradford	13	books	21	13	227.22805
sinir	***159***	***hücrelerinin***	***29***	***20***	***219.0938***
böyle	129	bir	3066	58	217.11462
books	21	mİt	15	13	215.44818
bundan	50	dolayı	33	17	214.7117
van	12	essen	11	11	212.66011
mit	15	press	50	14	206.04667

kabuk	189	bölgesi	73	24	205.44386
birden	47	fazla	69	18	199.82154
bir	3066	başka	261	74	199.33624
of	104	the	129	23	194.49793
görüş	105	alanındaki	18	15	187.75847
görme	***277***	***sisteminin***	***51***	***22***	***185.39902***
sinirsel	***64***	***ağlar***	***19***	***14***	***183.55298***
olmasına	13	karşın	38	12	182.98529
belki	56	de	723	27	182.82545
lgn	61	den	45	16	178.6234
başka	261	deyişle	21	17	178.3603
bu	1300	nedenle	36	26	177.08048
sci	10	usa	9	9	176.73449
tepki	124	gösterir	36	17	175.65651
primat	18	beyninin	52	13	175.38953
cerebral	13	cortex	34	11	165.59422
yine	56	de	723	25	164.10772
sinir	***159***	***hücresi***	***23***	***15***	***161.03934***
küçük	126	bir	3066	48	160.34747
yavaş	***37***	***dalgalı***	***10***	***10***	***158.30931***
beyin	***255***	***kabuğunda***	***24***	***16***	***157.29045***
her	*327*	*biri*	*90*	*23*	*156.86547*
görme	***277***	***kabuğu***	***38***	***18***	***155.88162***
acad	8	sci	10	8	154.76868
öte	10	yandan	17	9	153.244
bir	3066	süre	53	33	152.5993
normal	45	olarak	436	20	151.355
olmakla	14	birlikte	54	11	150.423
daha	605	az	137	29	149.82453
özgür	18	İrade	11	9	147.86968
ve	1272	arkadaşları	21	19	147.85833
tek	202	tek	202	24	147.76204
bir	3066	miktar	22	22	147.53162
sinirsel	64	karşılığı	14	11	146.28466
pazar	7	vaazı	7	7	146.03168
görme	***277***	***ruh bilimi***	***18***	***14***	***142.7847***
daha	605	fazla	69	23	142.40482
terimler	13	sözlüğü	9	8	141.17463
bazı	167	durumlarda	26	14	140.63917
natl	7	acad	8	7	140.02084
ölçüde	12	spekülasyon	10	8	139.49253
enine	8	boyuna	19	8	138.91281
tepki	124	gösteren	35	14	138.20064
ten	9	pazar	7	7	136.49698
çok	643	kısa	112	26	136.32103
gerçekten	49	de	723	21	135.51201

en	238	iyi	110	20	134.09507
koku	8	alma	8	7	133.99254
biraz	81	daha	605	23	133.74766
dikkat	121	ediniz	10	10	132.47646
ana	73	hatları	12	10	132.37683
çok	643	az	137	27	131.94756
görüş	***105***	***alanında***	***15***	***11***	***131.73425***
uzun	82	dönemli	64	14	130.57053
proc	12	natl	7	7	129.73105
san	12	diego	7	7	129.73105
görsel	***252***	***farkındalık***	***46***	***16***	***128.91248***
herhangi	23	bir	3066	21	127.3868
sinir	***159***	***hücresinin***	***19***	***12***	***127.36219***
bir	3066	sürü	21	20	126.15266
şimdiye	9	dek	27	8	125.68275
bir	3066	nöronun	77	34	125.50742
aynı	172	zamanda	19	12	125.40615
daha	605	iyi	110	24	125.38992
hareket	112	eden	18	11	123.58519
lgn	61	nin	48	12	122.99915
da	799	olsa	79	23	122.31641
bir	3066	parçası	41	26	121.62071

Sonuç

İki birimin bir arada istatistiksel olarak anlamlı bulunurluğunun adı olan eş dizimin, incelediğimiz metinde ortaya çıkardığı kimi terim ya da terim adayları aşağıda gösterilmiştir.

39 söz, terimsel olarak anlamlı bulunmuştur. Kalın-italik olarak gösterilenler *Güncel Türkçe Sözlük*'te bulunan sözler olup, diğerleri *GTS*'de bulunmamaktadır. Tek bir kitap için ortaya çıkan bu sonucun, derlemin boyutu arttığında da anlamlı sonuçların sayı ve oranı da doğru orantılı olarak artacağı açıktır.

1. alış alanı
2. ardışık arama
3. ateşleme sıklığı
4. beyin kabuğu
5. beyin sapı
6. boğum hücresi
7. büyük birleşik
8. ***dalga boyu***
9. dikenli yıldız
10. ***doğal ayıklanma***

11. doğrusal olmayan
12. dönemli bellek
13. dördüncü katman
14. eski kabuk
15. görme bölgesi
16. görme kabuğu
17. görme ruh bilimi
18. görme sistemi
19. görsel bilgi
20. görsel farkındalık
21. görüş alanı
22. ışık duyargası
23. işlem birimi
24. işlem önermesi
25. işletim sistemi
26. kabuk bölgesi
27. kısa erişimli
28. kör nokta
29. nöron ağı
30. saklı birim
31. salınım devresi
32. sinir biyolojisi
33. sinir hücresi
34. sinir lifi
35. sinir sistemi
36. sinirsel ağ
37. üst tepecik
38. yavaş dalgalı
39. yeni kabuk

Kısaca yinelemek gerekirse, Türkçe için bilgisayarlı terim belirleme-çıkarma sistemlerinin gerçekleştirilmesi gerekmektedir. Bu çalışmaların tamamlanması, terimsel söz varlığının genişletilmesi ve belirlenen bir veri tabanı üzerinden yeni terimlerin tanımlanmasını sağlayacaktır. Bu, hem zaman kazanımı hem de doğru sonuçların hızlıca elde edilmesini olanaklı kılacaktır.

Kaynaklar

Aksan, Doğan (1998), *Her Yönüyle Dil Ana Çizgileriyle Dilbilim* (3. cilt), TDK.

Crick, Francis (2003), *Şaşırtan Varsayım* (Çev. Sabit Say), TÜBİTAK Yayını, Ankara.

Douglas Biber, Susan Conrad, Randi Reppen (2000), *Corpus Linguistics-Investigating Language Structure and Use,* Cambridge Üniversitesi Yayını.

Eker, Süer (2002), *Çağdaş Türk Dili*, Grafiker Yayınevi, Yayınları: 439, Ankara.

Hengirmen, Mehmet (1999), *Dilbilgisi ve Dilbilim Terimleri Sözlüğü,* Engin Yayınevi, Ankara.

Jurafsky, Daniel, James H. Martin (2000), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Prentice Hall Yayınları.

McEnery, Tony, Andrew Wilson (2004), *Corpus Linguistics- An Introduction,* 2. basım, Edinburg Üniversitesi Yayını.

O'grady, William, Michael Dobrovolsky, Mark Aronoff (1997), *Contemporary Linguistics-An Introduction*, St. Martin's Pres, New York.

Özkan, Bülent (2006), "Türkçede Dilbilgisel Terim Olarak 'olumlama' ve 'olumsuzlama'", *ÇÜ Sosyal Bilimler Enstitüsü Dergisi, Cilt 15, Sayı 1, 2006, s. 269-282.*

Özkan, Bülent (2007), "Türkiye Türkçesinde Belirteçlerin Fiillerle Birliktelik Kullanımları ve Eşdizimliliği", Çukurova Üniversitesi, Sosyal Bilimler Enstitüsü, Türk Dili ve Edebiyatı Anabilim Dalı, Yayınlanmamış Doktora Tezi, Adana. (Danışman: Prof. Dr. Mehmet Özmen).

Sever, Hayri, M. Zafer Bolat (2006), "Bilgi Süzme", *Türkiye Bilişim Ansiklopedisi*, Papatya Yayınları, İstanbul.

Vardar, Berke (1998), *Açıklamalı Dilbilim Terimleri Sözlüğü*, ABC Kitabevi Yayınları, İstanbul.

Zülfikar, Hamza (1991), *Terim Sorunları ve Terim Yapma Yolları*, TDK Yayınları: 569, Ankara.

Erişim:

http://americannationalcorpus.org/

http://www.tei-c.org/Software/

http://www.cervantes.es/seg_nivel/lect_ens/oesi/liquid02/documet_pdf/automatic_ter

minolog_extraction_validation_liquid_approach.pdf

http://www.federation-nlp.uqam.ca/publications/01/goldman.pdf

http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb_eng.html

http://tdk.org.tr/tdksozluk/sozara.htm

Kısaltmalar

DD : Derlem Dil bilimi

DDİ : Doğal Dil İşleme

GTS : *Güncel Türkçe Sözlük*

HMM : Hidden Markov Model

OKT (OCR) : Optik Karakter Tanıma

SGML : Standard Generalized Markup Language

XML : Extensible Markup Language