5.3 Elektronik kaynaklar

(a) elektronik sözlükler, (b) metin bütünceleri.

5.3.1 Elektronik sözlükler

CD-ROM formatında bulunan başlıca İngilizce tekdilli (monolingual) sözlükler şunlardır:

1. *Oxford English Dictionary* (2. baskı) 20 cilt
2. Masa tarzı *Collins English Dictionary*, *The Concise Oxford Dictionary* (9. baskı) ve *Longman Dictionary of the English Language*.
3. Öğrenim sözlüğü olarak *Oxford Advanced Learner's Dictionary, Longman Dictionary of Contemporary English* ve *Collins COBUILD Dictionary*.

Elektronik sözlüklerin avantajları nelerdir?

- Bu sözlüklerde ayrıntılı arama yapılabilir.
- Aramalarda korelasyonlara girilebilir.
- Elektronik sözlükler basılı sözlüklere göre fazladan bilgiler içerir, örneğin bu sözlüklerde sözlüksel alanlar için başlıklar bulunabilir.
- Elektronik sözlüklerin basılı olanlar gibi maddebaşlarının alfabetik olarak sıralanmaları gerekmez.
- Elektronik sözlüklerin basılı sözlüklerde olduğu gibi yer sorunu bulunmamaktadır. (bkz. McArthur 1998). CD-ROMlarda yer sorunu yoktur. *Collins COBUILD on CD-ROM* (1994), 5 milyon sözcüklü bir veri tabanı içermektedir.

Elektronik sözlüklerin dezavantajları nelerdir?

- Sözcükler hakkındaki bazı bilgiler ya tamamiyle kaybolmuş ya da uyumlu bir biçimde girilmemiştir. Örneğin tanımlamalar çoğunlukla sözcüklenmiştir. Bazı sözcüklerin çözümlemeli tanımlamaları varken bazılarının karşısında yalnızca eşanlamlıları verilmiştir.
- Başlıklandırmalar genellikle sistematik değildir ya da yanlışlar içermektedir.
- Sözcüklerin biçimbilimsel yapısı hakkında çok az bilgi vardır.
- Elektronik sözlüklerin basılı sözlüklerin olduğu gibi bilgisayara aktarılmış biçimleri olmaları, basılı sözlüğün bilgisayar için modife edilmemesi, eksiklik ve yanlışlıkların aynen korunması sonucunu doğurmaktadır.

Elektronik sözlüklerin başka alanlara katkısı

Elektronik araçlar yalnızca sözvarlığının incelenmesi için yeni olasılıklar sunmaz, sözvarlığının betimlenmesi için de yeni olasılıklar sunar. Özellikle yapay zeka ve doğal dil mühendisliği alanında çalışanlar makineli çeviri, konuşma sentezleme, otomatik metin çözümleme gibi alanlarda elektronik araçlara başvurulmaktadır (bkz. Atkins and Zampolli, eds, 1994, Walker ve diğ., eds, 1995, Wilks ve diğ. 1996.)

Elektronik biçimdeki sözlüklere **Machine Readable Dictionaries** (MRD) denir ve bunlardan biri *Longman Dictionary of Contemporary English*'tir (1978). Bu sözlük, sözlüksel kaynak olarak birçok projede kullanılmıştır (bkz. Wilks ve diğ. 1996).

Veritabanları

Sözlüksel Veritabanları (lexical database)

Sözlüksel Bilgi Tabanı (Lexical Knowledge Base)

ek bilgi 5

Bu türden bir veritabanı **WordNet**tir. Bu proje, Princeton Üniversitesinde George Miller'in başkanlığında 1980'li yılların ortalarında başlamıştır. WordNet "online database" veya "online sözlüksel gönderim sistemi" olarak tanımlanmıştır. Bu veritabanında adlar, eylemler ve sıfatlar İngilizcenin, her biri tek bir temel sözlüksel kavrama dayanan eşanlamlılar seti olarak düzenlenmiştir. Eşanlamlıları içeren setler birkaç farklı yolla birbirlerine bağlanmıştır. Burada sözcük biçimleri değil, Miller'in kendisinin geliştirdiği bir yöntemle sözcük anlamları önemlidir. Bu proje ile ilgili geniş bilgi şu adreste bulunabilir: http://www.cogsci.princeton.edu/~wn/

araştırma 5

Türkçe için oluşturulmuş tekdilli, çokdilli elektronik sözlükler nelerdir? Bu sözlüklerin hangileri web tabanlıdır? Araştırınız.

ek bilgi 6

Türkçe için oluşturulmuş web tabanlı tekdilli, çokdilli elektronik sözlükler

Tekdilli Sözlükler

- http://www.nlp.cs.bilkent.edu.tr/Sozluk/ (Bilkent Üniversitesi Bilgisayar Mühendisligi Bölümü)
- http://turkcesozluk.halici.com.tr/arama.asp (Halıcı Yazılım)
- http://www.karpuz.com/onlinesozluk.htm (Karpuz.com)
- http://sozluk.mikrobeta.com.tr/ (Mikrobeta Ltd. Şirketi)

Çokdilli Sözlükler

- http://www.hazar.com/ (İng., Alm., Fr., İt., Dan., İsp. ve Tür.)
- http://software.estr.com/ (İng.-Tür. / Tür.-İng.)
- http://www.seslisozluk.com/ (İng.-Tür. / Tür.-İng.)
- http://www.tur.net/sozluk/ (Tür.-İng. / Tür.-İt.)
- http://www.langtolang.com/ (İng., Alm., Fr., İt. vb. ile Türkçe)

Terim Sözlükleri

- http://www.kadifeli.com/cgi-bin/compdict.pl (Bilgisayar tetimleri)
- http://www.zargan.com/ (Zargan Terim Bankası)
- http://www.tbd.org.tr/sozluk.html (TBD Bilişim Terimleri Sözlüğü)

Halk Dilinde Türkçe Sağlık Deyişleri Sözlüğü
http://www.hipokrat.org/hnet/deginme/sozluk/halkdeyisINDEX.html

Reçete Yazımında Kullanılan Terim ve Sözlere Ait Başlıca Kısaltmalar
http://www.hipokrat.org/hnet/deginme/sozluk/recete.html

ek bilgi 7

İngilizce için oluşturulmuş web tabanlı elektronik sözlükler ve sözlüksel kaynaklar

Sözlüksel kaynaklar ve sözlüksel araştırmalar

- www.notredam.ac.jp/cgi-bin/wn/ (WordNet 1.6 Vocabulary Helper)
- crl.nmsu.edu/Resources/clr.htm/ (New Mexico Eyalet Üniversitesinde sözlüksel kaynaklara linkler veren site)
- www.clres.com/siglex.html/ (sözlüksel kaynaklara linkler veren site)
- www.kun.nl/celex/ (Almanca, İngilizce ve Hollandaca ile sınırlı sözlüksel veri tabanı sunan site)

Sözlükler

- www.math.uni-paderborn.de/dictionaries/ Dictionaries.html/ (Paderborn Üniversitesinde tek dilli iki dilli ve özel sözlüklere ve tezaruslara linkler veren Index of Online Dictionaries adlı site)
- www.falstaff.bucknell.edu/rbeard/diction.html/ (Bucknell Üniversitesinde Robert Beard tarafından geliştirilen ve 150 farklı dilde 800 sözlüğe link veren A Web of On-line Dictionaries adlı site)
- www.onelook.com/ (449dan fazla sözlüğe endeksli olduğu iddia edilen OneLook Dictionaries adlı site)
- www.oed.com (Oxford English Dictionary)
- www.m-w.com/home.htm (Merriam-Webster sözlüğü)
- www.linguistics.ruhr-uni-bochum.de/ccsd/ (Bochum'da Ruhr Üniversitesinde kurulan Collins COBUILD Student's Dictionary adlı online)
- www.ims.uni-stuttgart.de/euralex/
- www.up.ac.za/academic/libarts/afrilang/homelex.html
- www.anu.edu.au/linguistics/alex/
- www.anu.edu.au/linguistics/alex/asialex.html
- //polyglott.Iss.wisc.edu/dsna/index.html/
- www.ex.ac.uk/drc/ veya www.ling.mq.edu.au/drc

5.3.2 Metin bütünceleri

İlk bilgisaylı bütünce: "Brown Corpus"

İlk bilgisayarlı bütünce 1960'larda Brown Üniversitesinde Nelson Francis ve Henry Kucera tarafından yapılmıştır.

ek bilgi 8

Brown Corpus üzerine çalışmalar

İngiliz İngilicesi için "Brown Corpus"

1970'te LOB adı altında, aynı yılda (1961) yayınlanmış metinlere dayanarak *Brown Corpus* İngiliz İngilizcesi için de yapılmıştır. Bu da her iki İngilizce arasında çeşitli karşılaştırmalar yapma olanağı sağlamıştır.

Tarihsel bir çalışma: "Brown Corpus"un tekrarı (FROWN ve FLOB)

Aynı derlemeler 1991'de FROWN ve FLOB adı altında Freiburg Üniversitesinde tekrarlanmıştır. Böylece 30 yıl sonrasının İngilizcelerini de karşılaştırmaya katma olanağı doğmuştur. Bu derlemeler hep bir milyon sözcüklü olmuştur.

International Corpus of English (ICE) Projesi

Aynı hacimdeki derlemeler Sidney Greenbaum'un önderliğinde *Uluslararası Corpus of English* (ICE) projesini doğurmuştur.

COBUILD projesi

Sözlük araştırma için gerekli alan daha geniş veri için John Sinclair ve ekibi, Birmingham Üniversitesinde 1980'li yıllarda *COBUILD projesine* başlamıştır.

ek bilgi 9

Sözlüklerin kullandığı bütünceler

- *COBUILD Dictionary, Bank of English*'i kullanmıştır.
- *Oxford Advanced Learner's Dictionary* ve *Longman Dictionary of Contemporary English* başka bütüncelerin yanı sıra 100 milyon sözcüklü *British National Corpus*'u kullanmıştır.
- *Cambridge International Dictionary of English* 100 milyon sözcüklü *Cambridge Language Survey* bütüncesini kullanmıştır.
- Doğal konuşucular için hazırlanmış sözlükler bile böylesi bütünceleri kullanmışlardır: *Collins English Dictionary* (1998) ve *New Oxford Dictionary of English* (1998) gibi.

Bütüncedeki hangi sözcükler alınmalı?

Sinclair (1985), bir sözcük 7.3 milyon sözcüklü bir derlemede geçmiyorsa sözlüğe alınmaması gerektiğini belirtmektedir.

ek bilgi 10

Türkçe Üzerine Veritabanları ve Bütünceler

TURKISH ELECTRONIC LIVING LEXICON (TELL) -- Version 1.0
http://socrates.berkeley.edu:7037/cgi-bin/TELLsearch.cgi

ODTÜ Türkçe Derlem Projesi
http://www.ii.metu.edu.tr/~corpus/indextr.html

ODTÜ Enformatik Enstitüsü tarafından gerçekleştirilen projenin kapsamı Türkçenin 1990'lı yıllardaki yazılı kullanımlarından (gazete ve dergi makale ve haberleri, akademik makaleler, yazınsal yapıtlar, radyo oyunları ve konuşmaları vb.) oluşturulan 10 milyon sözcüklük bir seçkiyi uluslararası standartlara uygun bir biçimde dilbilim-ötesi ve dilbilimsel imlerle kodlayarak uygun erişim yazılımlarıyla beraber bir CD'lik bir derlem olarak araştırma ve geliştiricilere sunmaktır. Şubat 2002 tarihi ittibariyle derlemin içeriği 1,100.000 sözcüğe ulaşmıştır. Ayrıca imleme işleminin yarı otomatik yapılması için yazılım geliştirilmiştir. İşaretleyicilerin daha verimli çalışmasına olanak veren XML tabanlı bir editor/çözümleyici bir yaz projesi olarak geliştirilmiştir. Kullanıcılara yönelik arama ve sonuçların değerlendirilmesine olanak sağlayan derlem kullanıcı yazılımı geliştirilmektedir.

Türkçe için Biçimbirimsel ve Sözdizimsel Olarak İşaretlenmiş Ağaç Yapılı Derlem Projesi
http://www.ii.metu.edu.tr/~corpus/treebank/indextr.html

Proje, TÜBİTAK tarafından desteklenmekte ve ODTÜ Enformatik Enstitüsü ile Sabancı Üniversitesi tarafından gerçekleştirilmektedir. Biçimbirimsel ve sözdizimsel işaretlenmiş **ağaç yapılı metin derlemleri** (İng. treebank corpora) dilbilim ve bilgisayarlı doğal dil işleme üzerinde çalışan araştırmacıların hem bilgisayar ortamındaki böyle bir kaynağın yokluğunda gerçekleştirilemeyecek çalışmaları yapmalarını, hem de emek ve zamandan tasarruf etmelerini sağlamaktadır. Bu projede Türkçe metinlerin biçimbirimsel ve sözdizimsel işaretlenmesine yönelik işaretleme ve ayrıştırma öğelerinin belirlenmesi ve bu işaretlemenin bilgisayar aracılığıyla yapılması için gerekli çekirdek işlevleri ve kullanıcı arayüzlerini içeren programların geliştirilmesi ve bu programlar aracılığıyla yaklaşık 50,000 tümcenin öğelerinin bir ağaç yapısı çerçevesinde işaretlenerek Türkiye ve dünyadaki diğer araştırmacıların hizmetine sunulması amaçlanmaktadır.

ek bilgi 11

İngilizce için oluşturulmuş metin bütünceleri

- www.hit.uib.no/icame.html/ (Norveç'te klasik derlemelerin dağıtıldığı International Computer Archive of Modern and Medieval English (ICAME) için site)
- //thetis.bl.uk (The British National Corpus)
- www.cobuild.collins.co.uk/boesinfo.html/ (COBUILD sözlükleri için site)
- www.ucl.ac.uk/english-usage/ (Survey of English Usage için site)

5.3.3 Çözümleme araçları

Tanıklı Dizinleyiciler (concordancer)

I. *Sözcükleri bağlam içinde listeleme*

II. *Sıklık bilgisi*

III. *Eşdizimlilik hakkında bilgi*

ek bilgi 12

SARA Programı

Bu tür programlardan biri SARA programıdır. SARA programı "British National Corpus"u (website için bkz. http://thetis.bl.uk/) taramak için kullanılan bir program olarak sözcükleri içinde geçtikleri tam tümcelerle listelemektedir.

ek bilgi 13

Türkçe için bir sıklık çalışması

Türkçe için Pierce 1963'ün yapmış olduğu çalışmada konuşma dili ve yazı dili arasındaki ayrımlar görülmektedir:

	Konuşma dili		Yazı dili	
1	demek	8742	bir	5589
2	bir	4673	bu	2170
3	bu	3278	olmak	2053
4	o	3203	etmek	1944
5	ben	2764	**ve**	1736
7	olmak	2625	o	856
8	gelmek	2372	ne	685
9	**gitmek**	2372	**baş**	651
10	**sen**	1882	yapmak	650
11	**var**	1801	**için**	641
12	**şey**	1343	ben	607
13	almak	1281	**görmek**	569
14	yapmak	1264	gelmek	559
15	vermek	1216	**iki**	558
16	**ora-**	1175	vermek	546
17	**yok**	1175	**gibi**	495
18	etmek	1098	**bulmak**	486
19	**bakmak**	1080	**hareket**	478
20	**kız**	1072	almak	458

okuma

Pierce, John E. (1961). A frequency count of Turkish affixes, *Anthropological Linguistics* 3/9: 31-40.

Pierce, John E. (1962). Frequencies of occurrence of affixes in Written Turkish, *Anthropological Linguistics* 4/6: 30-41.

Pierce, John E. (1963). A statisitcal study of grammar and Lexicon in Turkish and Sahaptin (Klikitat, *International Journal of American Linguistics* 29/2: 96-106.

"Parser" / "tagger"lar (etiketleyici)

Yoğunluk ölçme araçları

$$\textbf{yoğunluk} = \frac{\text{metindeki farklı sözcük sayısı}}{\text{toplam sözcük sayısı}}$$