TÜRKİYE CUMHURİYETİ
ÇUKUROVA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
TÜRK DİLİ VE EDEBİYATI ANABİLİM DALI

BİLGİSAYAR DESTEKLİ SÖZLÜK BİLİMİ ÇALIŞMALARINDA DERLEME SÖZLÜĞÜ VERİ TABANI ÖRNEĞİ

B. Tahir TAHİROĞLU

DOKTORA TEZİ

Adana, 2010

TÜRKİYE CUMHURİYETİ
ÇUKUROVA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
TÜRK DİLİ VE EDEBİYATI ANABİLİM DALI

BİLGİSAYAR DESTEKLİ SÖZLÜK BİLİMİ ÇALIŞMALARINDA DERLEME SÖZLÜĞÜ VERİ TABANI ÖRNEĞİ

B. Tahir TAHİROĞLU

Danışman: Prof. Dr. Şükrü Halûk AKALIN

DOKTORA TEZİ

Adana, 2010

ÖZET

BİLGİSAYAR DESTEKLİ SÖZLÜK BİLİMİ ÇALIŞMALARINDA DERLEME SÖZLÜĞÜ VERİ TABANI ÖRNEĞİ

B. Tahir TAHİROĞLU

Doktora Tezi, Türk Dili ve Edebiyatı Ana Bilim Dalı

Danışman: Prof. Dr. Şükrü Halûk AKALIN

Temmuz, 2010, XII+207 sayfa

Bu çalışmada Anadolu ağızlarının söz varlığını içeren 12 ciltlik Derleme Sözlüğü'nün çevrim içi sorgulanabilir veri tabanı hazırlanmıştır. Sözlüğün madde başı, tanımlamalarını içeren anlam bölümü, örnek cümlelerin verildiği tanık bölümü, madde başlarının değişkeleri (varyantları) tablosu ve derleme yeri verisini içeren yer tablosu sözlükteki yapı esas alınarak oluşturulmuştur.

124.692 adet madde başından oluşan sözlüğün birbiriyle ilişkilendirilmiş veri tabanı modeli sorgu arayüzünde gelişmiş sorgulama seçenekleriyle desteklenmiştir. Anlam tablosundaki eşdizim örüntüleri, kavramsal sorgulama, anlamlarda anahtar sözcükler , sıklık listesi ve en çok derleme yapılan yer bilgisini veren derleme yeri listesi de çıkarılmıştır.

Doğal Dil İşlemenin çeşitli uygulama alanları, derlem dil biliminin tanımı, sözlük ve sözlük tekniği kavramları da tanıtılmaya çalışılmıştır.

Anahtar Sözcükler: Derleme Sözlüğü, Veri Tabanı, Dilsel Veri Tabanı, Doğal Dil İşleme, Derlem Dil Bilimi, Sıklık, Eşdizimlilik, Sözlük.

ABSTRACT

A SAMPLE OF DERLEME SÖZLÜĞÜ (COMPILATION DICTIONARY) DATABASE IN COMPUTER-ASSISTED LEXICOGRAPHY STUDIES

B. Tahir TAHİROĞLU

Ph. D. Thesis, Department of Turkish Language and Literature

Supervisor: Prof. Dr. Şükrü Halûk AKALIN

July 2010, XII+207 pages

In this thesis, we prepared a database based on Derleme Sözlüğü (Compilatin Dictionary) consisting of 12 volumes including Anatolian dialects vocabulary. Based on dictionary's structure, several table structures including lemma table that keeps the main words, variation table that keeps the lemmas's changing form, meaning table that keeps the definitions, sample table that keeps the lemmas usable forms and finally location table that keeps the lemmas's compilation data have been build.

The dictionary consisted 124.692 numbers of lemmas, have been build on relational table structure and also supported by a advanced search capability. In addition to searchable interfaces, based on meaning table, collocational structures and keywords have been extracted. Also, based on location table a location frequency list has been prepared for deep analysis. Thus, this dictionary's data have been converted to a digital form for futher investigation in Anatolian dialects.

Natural language processing and its several application fields and definition of corpus linguistics, dictionary and dictionary making techniques have been outlined for intruduction.

Keywords: Derleme Sözlüğü (Compilation Dictionary), Database, Linguistic Database, Natural Language Processing, Corpus Linguistics, Frequency, Collocationality, Dictionary.

ÖN SÖZ

1990'lı yılların sonundan itibaren internet ve çevrim içi "sanal" ortamlar günlük yaşamdan başlayarak yaşamın her alanını etkilemiş, bu etkilenmeden akademik çalışmalar da var olanın yayınlanmasından yeni çalışmaların üretilmesine kadar çevrim içi uygulamaları gerek yerel gerek uluslar arası düzeyde giderek artan yoğunlukta kullanmaya başlamıştır. E-öğrenme, e-kitap, e-devlet vb. gibi elektronik kavramıyla bütünleşmiş yeni sözlerin ortaya çıkması, değişen yaşam koşullarının dile yansıması olarak yukarıda belirtilen etkilenmenin örneklerini göstermektedir.

XX. yüzyılın ikinci yarısından günümüze kadar gelen ve hâlen yoğun bir biçimde sürdürülmekte olan DDİ araştırmalarının ve makineli çeviriyle başlayan insan dilini ve karmaşıklığını modelleme çalışmalarının bir yanı da basılı sözlüklerden yardım almaya dayanmaktadır. Sözlük kullanıcısının insan olduğunun düşünüldüğü günlerden, insanlar için hazırlanmış sözlüklerin makinelerin "anlayabileceği" biçimlere dönüştürüldüğü günümüz çevrim içi ortamlarının hazırlandığı, sözlüklerin malzemesininin tutulacağı veri tabanlarını ve milyonlarca hatta milyar düzeyinde gerçekleşmiş sözcüksel ortamların nasıl hazırlanacağından, ne tür işaretlemelere tabi tutulacağına varıncaya kadar yeni bir sözlük bilimi araştırma ve uygulama alanının doğmasına tanık olunan günlere gelinmiştir. Kuşkusuz, bilgisayar biliminin ve donanımın gelişmesinin bunda başat rolü vardır.

Dil biliminin uygulamalı bir alanı olan derlem dil bilimi, hem dil bilimsel kuramların oluşturulmasında hem de var olan kuramların sınanmasını sağlayacak dil verilerinin sayısallaştırılarak veri tabanlarının oluşturulmasını temel olarak konusu içine almaktadır. Genel olarak derlem işaretleme standartlarının geliştirilmesi yanında, sözlüğe kaynaklık edecek verilerin belirli standartlar çerçevesinde bir araya getirilmesi de yine derlem dil biliminin konusudur. Bu kapsamda derlem dil bilimi, derlem tabanlı elektronik sözlüklerin hazırlanması çalışmalarıyla yeni sözlük biçimlerinin oluşmasına da katkı sağlamaktadır.

Bu tez çalışması kapsamında, derlem dil bilimi, günümüz sözlükçülük teknikleri ve DDİ araştırmalarının genel hatları ve uygulama alanları örnek uygulamalarla verilmeye çalışılmış, bilgisayar destekli sözlük biliminde kullanılan ilişkisel veri tabanı örneği 12 ciltten ve 4402 sayfadan oluşan Derleme Sözlüğü'ne uygulanmıştır. Sözlüğün sorgulanabilirlik kapsamının genişletilmesi için 9 ayrı tabloda 124.692 madde başı

bilgisi bölümlenmiştir. Böylece sözlüğe dair her bir karakterden cümle düzeyine kadar birimler, zaman sınırı olmadan ulaşılabilir duruma getirilmiştir. Çalışmada anlam tablosundaki verinin sözcüksel yoğunluğu dikkate alınarak sıklığa dayalı anahtar kavram, eşdizimlilik örüntüleri ve derleme yeri sıklıkları da çıkarılmıştır.

Derleme Sözlüğü Veri Tabanı Örneği'nin, Anadolu ağızlarının araştırılmasında çeşitli kolaylıklar sağlayacağı düşünülmektedir. Bu veri tabanından hareketle; sözlük hazırlama tekniklerinin çeşitleneceğini, ağızların söz varlığının daha geniş ağlarda tutulduğu etkileşimli sözlük uygulamalarının hazırlanması aşamalarında bu çalışmanın veri sağlamada önemli olduğunu düşünmekteyiz.

Çalışma boyunca katkılarını esirgemeyen değerli hocam Prof. Dr. Şükrü Halûk Akalın'a teşekkürü bir borç bilirim. Ayrıca, çalışmanın veri tabanının hazırlanması bölümünde görüşlerinden yaralandığım Sinan Yalçınkaya'ya, değerli arkadaşlarım Bülent Özkan ve Fatih Karaoğlan'a da teşekkürlerimi sunuyorum.

B. Tahir TAHİROĞLU
ADANA/2010

İÇİNDEKİLER

ÖZET.....i
ABSTRACT.....ii
ÖN SÖZ.....iii
KISALTMALAR LİSTESİ.....viii
TABLOLAR LİSTESİ.....ix
ŞEKİLLER LİSTESİ.....x

BİRİNCİ BÖLÜM

GİRİŞ.....1
1.1.Amaç ve Kapsam.....1
1.2. Materyal ve Yöntem.....1
1.3. Sözlük, Sözlük Bilimi ve Tarihçesi.....1
1.3.1. Sözlük Tekniği.....8
1.3.2. Sözlük Verisi.....13
1.3.3. Terim Sözlükçülüğü.....15
1.3.4. İnternet Sözlükçülüğü.....17
1.4. Doğal Dil İşleme, Yöntem ve Uygulama Alanları.....38
1.4.1. Doğal Dil İşleme ve Bilgisayarlı Dil Bilimi Terimi.....38
1.4.2. Doğal Dil İşlemede Yöntemler.....40
1.4.2.1. Kural Tabanlı Yöntem.....40
1.4.2.2. İstatistiksel Yöntem.....55
1.4.2.3. Melez Yöntem.....59
1.4.3. Doğal Dil İşlemede Uygulama Alanları.....60
1.4.3.1. Belge Bölütleme.....60
1.4.3.2. Bilgiye Erişim.....60
1.4.3.3. Bilgi Çıkarımı.....62
1.4.3.4. Belge Sınıflama.....64
0.4.3.5. Varlık Adı Tanıma.....67
1.4.3.6. Özetleme.....70
0.4.3.7. Soru Yanıtlama.....73
1.5. Sözlük Biliminde Bilgisayarlı Yaklaşımlar.....75
1.5.1. Bilgisayarlı Sözlük Bilimi ve Bilgisayar Destekli Sözlük Bilimi Terimleri.75

1.5.2. Sözlük Birimi Elde Etme 75
1.5.2.1. Klasik Yöntemle Elde Etme 75
0.5.2.2. Derleme Dayalı Yöntemle Elde Etme 76
0.5.3. Bağlamlı Dizinler 80
1.5.4. Sıklık Kavramı 81
1.5.5. Otomatik Terim Çıkarımı 84
1.5.6. Çok Anlamlılık Çözümlemesi 87
1.6. Derlem Dil Bilimi 90
1.6.1. Derlem Kavramı 90
1.6.2. Derlem Araştırmalarının Tarihçesi 92
1.6.3. Derlem Türleri 93
1.6.4. Başlıca Derlemler 94
1.6.5. Derlem Oluşturma Yöntemleri 97
1.6.6. Derlem ve İstatistiksel Çıkarım 100
1.6.7. İnternet Üzerinde Derlem Çalışmaları 102
1.6.8. Türkçe için Derlem Çalışmaları 107
1.7. Veri Tabanı 111
1.7.1. Veri, Enformasyon, Bilgi ve Veri Tabanı Kavramı 111
1.7.2. Dilsel Veri Tabanları 115

İKİNCİ BÖLÜM

DERLEME SÖZLÜĞÜ VERİ TABANI MODELİ 124
2.1. Hazırlama Yöntemi ve Kullanılan Teknoloji 124
2.1.2. Veri Tabanı Yöntemi 128

ÜÇÜNCÜ BÖLÜM

DERLEME SÖZLÜĞÜ VERİ TABANI SORGULAMA YÖNTEMİ 134
3.1. Madde Başında Sorgulama 134
3.2. Anlam İçinde Sorgulama 138
3.3. Kavram Sorgulaması 146
3.4. Derleme Yeri Sorgulaması 159

DÖRDÜNCÜ BÖLÜM

DERLEME SÖZLÜĞÜ'NÜN SÖZLÜKSEL GÖRÜNÜMLERİ 169
4.1. Eşdizimlilik Görünümleri 169
4.2. Anlam Tablosunda Anahtar Sözcük Görünümleri 182
4.3. Anlam Tablosu Söz Sıklıkları 189

SONUÇ 196
KAYNAKÇA 198
ÖZ GEÇMİŞ 207
EK- DERLEME SÖZLÜĞÜ VERİ TABANI YAZILIMI CD'Sİ

KISALTMALAR LİSTESİ

BTS : Büyük Türkçe Sözlük

DDİ : Doğal Dil İşleme

DS : Derleme Sözlüğü

GTS : Güncel Türkçe Sözlük

MİS : Makinece İşlenir Sözlük

MOS : Makinece Okunur Sözlük

MUC : Message Understanding Conference

RSS : Rich Site Summary

SDM : Sonlu Durum Makinelerini

SGML : Standard Generalized Markup Language

VTYS : Veri Tabanı Yönetim Sistemi

VAT : Varlık Adı Tanıma

XML : Extensible Markup Language

age. **:** Adı geçen eser

vb. : Ve benzeri

vd. : Ve diğerleri

TABLOLAR LİSTESİ

Tablo 1. Cobuild Concordance and Collocations Sampler Çevrim içi Eşdizim Çözümlemesi 105
Tablo 2. ODTÜ Türkçe Derlemi Metin Türleri Dağılımı 108
Tablo 3. TurCo'da Sözcüklerin Dağılım Oranları 110
Tablo 4. 30 Satırlık Anlam Tablosu İçeriği 132
Tablo 5. Çok Anlamlı Sözler Sonuç Tablosu Örneği. 139
Tablo 6. Çok Anlamlı Sözler Ayrıntılı Sorgu Sonuç Örneği. 142
Tablo 7. Anlam Tablosu Söz Sıklığı Örneği. 147
Tablo 8. "insan." Sorgusunun Bölge ve İllere Göre Dağılımı. 153
Tablo 9. "kişi." Sorgusunun Bölge ve İllere Göre Dağılımı 155
Tablo 10. "madde." Sorgusunun Bölge ve İllere Göre Dağılımı 156
Tablo 11. "renk." Sorgusunun Bölge ve İllere Göre Dağılımı 158
Tablo 12. Anlam-Derleme Yeri Sorgu Sonucu Örneği 159
Tablo 13. Derleme Yeri Sıklık Listesi Örneği 166
Tablo 14. Derleme Sözlüğü Anlam Tablosu 2 Birimli n-gramlar 170
Tablo 15. Derleme Sözlüğü Anlam Tablosu 3 Birimli n-gramlar 176
Tablo 16. Derleme Sözlüğü Anlam Tablosu Anahtar Sözcük Listesi 183
Tablo 17. Derleme Sözlüğü Anlam Tablosu Sıklık Listesi 190

ŞEKİLLER LİSTESİ

Şekil 1. Longman Çevimiçi Sözlük Ana sayfa Sorgu Arayüzü 19
Şekil 2. Longman Sözlüğü Ara Sonuç Görünümü..... 20
Şekil 4. Longman Metne Duyarlı Sözlük uygulaması 21
Şekil 5. Wordreference Çok Dilli Sözlük Ana Sorgulama Sayfası 23
Şekil 6. lexicool.com Sözlük Bilgisi Sorgulama Formu 24
Şekil 7. dict.info.com İngilizce-Türkçe Sorgu Sonuç Sayfası..... 24
Şekil 8. Seslisözlük Ana sayfası 26
Şekil 9. Seslisözlük Uygulamaları Örneği 27
Şekil 10. Seslisözlük RSS Uygulaması..... 27
Şekil 11. TDK GTS Sorgulama Arayüzü..... 28
Şekil 12. Büyük Türkçe Sözlük (BTS) Sorgu Formu 29
Şekil 13. BTS Sonuç Gösterim Sayfası..... 29
Şekil 14. Türk Lehçeleri Sözlüğü Sorgu Formu ve Sonuç Tablosu 30
Şekil 15. Ekşisözlük Ana Sayfası 31
Şekil 16. Google Zeitgeist Ana sayfası 34
Şekil 17. Google Zeitgeist Sorgu Kalıpları Örnekleri..... 34
Şekil 18. Google Insight for Search Sorgu Formu..... 35
Şekil 19. Google Insight for Search - "bilgisayar" Sözcüğünün Dünya Ölçeğinde Dağılımı..... 35
Şekil 20. Google Insight for Search'de "bilgisayar" Sözcüğü ve Yükselen Arama Sonuçları..... 35
Şekil 21. The Golobal Language Monitor (GLM) Ana sayfası..... 37
Şekil 22. Düzenli İfadeler, Dil, İlişki ve SDM Ağı İlişkisi 42
Şekil 23. İspanyolca cantar, canto ve carta Sözcükleri İçin İlişki Modellemesi 43
Şekil 24. SDM Örneği..... 43
Şekil 25. Yüzeysel Bölümleme ve Sözlüksel İşlevsel Dil Bilgisi Gösterimi..... 45
Şekil 26. Türkiye Türkçesi Biçim birimsel Çözümleyicide Sözlüksel ve Yüzeysel Düzeylerin Gösterimi..... 48
Şekil 27. Ek-fiil Kümesi için Bir SDM..... 49
Şekil 28. Fiil Zaman Ekleri İçin Sağdan Sola Bir SDM 50
Şekil 29. Programa Dönüştürülmüş Biçim Birimsel Çözümleyici..... 51

Şekil 30. VisualText Dil Bilgisi Geliştirme Ortamı. 52
Şekil 31. VisualText Dil Bilgisi Editörü. 53
Şekil 32. NooJ Dil Bilgisi Geliştirme Ortamı. 54
Şekil 33. NooJ Sisteminde Bağlamlı Dizin 54
Şekil 34. Dil Bilgisel Bağlılık Gösterimi 58
Şekil 35. GATE 5.1 Uygulama Geliştirme Ortamı 63
Şekil 36. Belge Sınıflama Aşamaları 66
Şekil 37. Clusty Arama Motoru Sonuç Sayfası 67
Şekil 38. Lingpipe Metin Giriş Formu 69
Şekil 39. Lingpipe VAT İçin XML Çıktısı 70
Şekil 40. LexRank Uygulaması 72
Şekil 41. AnswerBus Çevrimiçi Soru Yanıtlama Sistemi Ana sayfası 74
Şekil 42. Derlem, Derleme Eklenen Dış Veri ve DDİ Sistemi İlişkisi 77
Şekil 43. Soyutlama Düzeyi Gösterimi 77
Şekil 44. Sözlük Birimi Elde Etme Sistemi 78
Şekil 45. AntConc Yazılımı Bağlamlı Dizin Görünümü 80
Şekil 46. NooJ Sistemi Tek Birimli Sözcük Sıklığı Tablosu 83
Şekil 47. NooJ Sistemi İki Birimli Sözcük Sıklığı Tablosu. 84
Şekil 48. Eurodicautom Terim Bankası Sistemi Sorgu Arayüzü. 85
Şekil 50. Teleport Yazılımı Ekran Görüntüsü. 103
Şekil 51. AHDS Derlem Geliştirme Kılavuzu 104
Şekil 52. Cobuild Concordance and Collocations Sampler Çevrimiçi Derlem Sorgulayıcısı 105
Şekil 53. WebAsCorpus Bağlamlı Dizin Oluşturucusu 106
Şekil 49. ODTÜ Türkçe Derlemi İşaretleme Yazılımı İşaretleme Formu 109
Şekil 54. Bilginin Elde Edilmesi Süreçleri 112
Şekil 55. VTYS ve Veri Tabanı 114
Şekil 56. IMS Corpus Query System Sorgu Sonuçları 120
Şekil 57. IMS Corpus Query System Bağlamlı Dizini 120
Şekil 58. WORDNET'te Sorgulama Sonuçları 122
Şekil 59. Derleme Sözlüğü Birim Yerleşimi 125
Şekil 60. Derleme Sözlüğü Veri Tabanı İlişkiler Tablosu 128
Şekil 61. Derleme Sözlüğü Veri Tabanı Tablolarının VTYS'de görünümü. 131
Şekil 62. Derleme Sözlüğü Sorgulama Ana sayfası 134

Şekil 63. Madde Başı “eşit” Sorgusu 135
Şekil 64. Gelişmiş Arama Sorgu Arayüzü 136
Şekil 65. Madde Başı Sorgusu Gelişmiş Özellikler 137
Şekil 66. Anlam Sorgusu Sonuç Sayfası 138
Şekil 67. Çok Anlamlı Sözler Ayrıntılı Sorgu Formu 142
Tablo 6. Çok Anlamlı Sözler Ayrıntılı Sorgu Sonuç Örneği. 142
Şekil 68. Anlam-İl Gruplama Tablosu Sorgu Sayfası 146
Şekil 69. WordSmith Tool Anahtar Sözcük Hesaplama Ayarları. 183

BİRİNCİ BÖLÜM

GİRİŞ

1.1. Amaç ve Kapsam

Bu çalışmada, 124.692 madde başı söz içeren Derleme Sözlüğü'nün ilişkisel veri tabanı hazırlanıp çevrim içi yayınının gerçekleştirilmesi amaçlanmıştır. Derleme Sözlüğü Türkiye'de halk ağzından derlenen sözlerin bulunduğu 12 ciltlik bir sözlüktür. Sözlüğün çevrim içi yayınıyla ve aynı zamanda; madde başı ve madde içini oluşturan söz, anlam, değişke, tanıklar ve derleme yeri sorgulamalarını barındıran sözlük veri tabanıyla, Türk diline ve ağızlarına dayanan olası metin madenciliği, bilgi süzme uygulamaları ve anlam bilimi çalışmaları için ontolojilerin oluşturulmasında bir veri kaynağının oluşturulması da amaçlanmıştır.

Bilgisayar destekli sözlük biliminin, sözlük bilimi çalışmalarındaki konumu, yararları ve Doğal Dil İşleme (DDİ) çalışmalarıyla olan ilişkisi de çalışmada kuramsal bölümü oluşturacaktır. Dünyada yürütülen derlem dil bilimi çalışmalarının sözlüklerin hazırlanmasındaki yeri ve durumu, kimi istatistiksel uygulamalar ile otomatik terim belirleme, dilsel veri tabanları da çalışmanın kapsamı içinde bulunmaktadır.

1.2. Materyal ve Yöntem

Türk Dil Kurumu yayını olan 12 ciltlik Derleme Sözlüğü çalışmanın yürütüleceği taban olarak kullanılmıştır. Çalışmada, ilişkisel veri tabanı yöntemi ile sözlükteki sözler kaydedilip, ileri düzey düzenli ifadeler ile kaydedilen bu sözlerin düzenlilikleri ve sözlüğün madde başı yapıları modellenmiştir. MySql açık kaynak kodlu veri tabanı yönetim sistemi ve SQL standart sorgu dili kullanılmış, tablolanan veriden hazırlanan seçenekli sorgu arayüzüyle istenilen sonuçlar çıkarılmıştır.

1.3. Sözlük, Sözlük Bilimi ve Tarihçesi

Sözlük terimini ilk kez John Garland 1125'te kullanmıştır. İlk sözlüğün hangi dilde ve kim tarafından hazırlandığı bilinmemektedir. Aksan, günümüzdekilere benzer

ilk sözlüğün, İskenderiye müzesi kütüphanecisi Bizanslı Aristophanes'in M.Ö. II. yüzyılda hazırladığı sözlük olduğunu dile getirmiştir (Aksan, 1998: 69).

Aksan, sözlüğü, "bir dilin (ya da birden çok dilin) söz varlığını, söyleyiş biçimleriyle, yazılışlarıyla veren, bağımsız biçimleri temel alarak bunların başka ögelerle kurdukları söz ögeleriyle birlikte anlamlarını, değişik kullanımlarını gösteren söz varlığı kitabı" olarak tanımlamaktadır (age., 1998:75).

Sterkenburg, sözlüğün çağrıştırdığı genel anlam olarak bir "prototipik sözlük"ün varlığının söz konusu olduğunu vurgulayarak, prototipik sözlükten alfabetik ve tek dilli nitelikte olanlarının anlaşıldığı belirtilir. Prototipik sözlük ansiklopedik olmaktan çok dilsel, dil bilimsel bir sözlüktür. Sterkenburg'un bu özellikli bir sözlükle kastettiği ilk anda anlaşılan anlamıyla sözlüktür: Genel amaçlı, tek dilli bir sözcükler listesi. Geeraerts ise tek dilli genel amaçlı sözlüklerin başlıca onomastik olmaktan çok semasiyolojik olduğunu, ölçünlü dildeki ögeleri temel aldığını ve pedagojik bir amaçla yazıldığını belirterek, tek dilli olan bu sözlüklerin eleştirel ve akademik bir amaçla yazılmadıklarını dile geitrmiştir (Sterkenburg, 2003:4).

Kirkness da sözlüğü bir prototipik sözcüksel kaynak olarak görmektedir. Sözlük bilgileri basılı biçimlerde ya da günümüzde giderek artan sayıda elektronik ortamlarda sunulmaktadır. Sözlükteki bilgi sözcüklerle ilgili olduğundan, sözlük temel olarak sözcüğe dayanır. Kirkness'a göre sözlük, sözcük hakkında bir kitap ya da bir bankadır (Kirkness, 2004: 59).

Ansiklopedi ile sözlük arasındaki fark göz önüne alındığında, ansiklopedinin dil dışı bağlama (extralinguistic) işaret ettiği ve bir tür olgular kitabı özelliği taşıdığı görülmektedir. Kirkness, ansiklopedinin sözlük gibi iki ya da çok dilli biçimlerinin sözlüklerdeki gibi olmadığını, ansiklopedilerin çevirilerinin yapılabileceğini ancak sözlüklerin çeviri biçimlerinin bulunmadığını belirtmektedir (age., 2004: 59).

Önemli sözlük bilimcilerden L. Zgusta'ya göre sözlük, toplumsal olarak kabul edilmiş dilsel formlar ve anlamların yanında işlevlerinin de sistematik olarak sınıflandırıldığı bir listedir. Zgusta, sözlüğün oluşturulma sürecinde sözlük bilimcinin eğitiminin iyi olup olmamasının da sözlüğün bütünlüğünü, yapısal özelliklerini etkilediğini belirtmiştir (Sterkenburg, 2003:4).

Jackson'a göre sözlük, toplumsal yapının bir parçasıdır. Herkes bir anlamda sözlüğü otorite olarak gördüğünden sözlüğün belirttiğini dikkate alma eğilimindedir. Jackson, sözlüğün olmadığı bir dilsel tartışma ortamının düşünülemeyeceğini belirtir. Sözlük, dil hakkında bir kitap, sözcüklerle ilgili bir referanstır. Ansiklopedinin bir

kuzeni olmakla birlikte, ansiklopedi nesneler, insanlar, yerler ve düşüncelerle yani "gerçek dünya"yla ilgiliyken sözlük dille ilgilidir (Jackson, 2002:21).

Sözlükleri hazırlayan sözlük bilimcilerin, kullanıcıların sözlükte neleri bulmayı bekleyebileceklerini göz önüne almaları gerektiğini vurgulayan Jackson, sözlüğün başvuru kaynağı olması yanında kısmen bir söz varlığı kütüğü olduğunu da belirtir. Dil bilgisi ve sözlüğün karşılıklı olarak birbirlerini tamamlamaları da dilin karakteristiklerinin betimlenmesi bakımından önemlidir. Sözlük, dil bilgisinin ortaya koyduğu terimleri ve ögeleri de tanımlamakla bir işlevsel bileşen özelliği de taşır (age., 2002: 22).

Ilson, sözlüğü, sözlük bilgisel sözlük (lexicographic dictionary) olarak adlandırılmış ve sözcüksel olarak ilişkili bilgileri (lexically relevant information) sağlayan bir kaynak olarak tanımlanmıştır. Sözcüksel bilgi Ilson'a göre, sözcüksel olarak ilişkili birimlerin (lexically relevant units) sesletim, anlam vb. niteliklerini içeren bilgilerdir (Ilson, 2002: 331).

Kocaman'a göre, sözlüğün, bir dilin sözcüklerini alfabetik sırada veren yapıtlar olarak tanımlanması eksiktir, çünkü kavramlalar dizinini veren sözlükler (thesaurus), uyak sözlükleri, ters dizimli sözlükler gibi özel amaçla yapılanmış sözlükler bu tanımların dışında bırakılmışlardır. Ayrıca sözlüklerde yer alan dil bilgisel ya da kullanıma yönelik açıklamalar da günümüz sözlük biliminde gittikçe yaygınlaşan bir yöntem olmakta ve genel sözlük tanımını değiştirmektedir. Artan bilgi çeşitliliği ve disiplinlerarasılık, sözlük bilimi yöntemlerinin değişik bilim alanlarıyla desteklenmesini sağlamaya başlamıştır. Dil biliminin kuramsal yanı sözlük bilimi için de geçerlidir. Bir üst yapı olarak sözlük kuramının oluşturulması ve bunun uygulanması sözlüklerin planlanmasından basımında kadar olan süreci kapsamaktadır, bunun için de bilgisayar destekli dil bilimi, antropoloji ve felsefe gibi uzmanlık alanlarının ortaya koyduğu bilgilerden de yararlanılması gerekmektedir (Kocaman, 1998: 111).

Yukarıda verilen bilgi ve tanımlamalardan sözlüğün temel olarak alfabetik bir liste özelliği taşıdığı görülür. Bu listenin sıralama biçimleri sözlüğün amacına ve türüne göre değişmekle birlikte madde başı ve madde içi bölümlerin bulunduğu sözlüksel bilginin gelişigüzel olmayan bir düzende verilerek bir "sözlüksel sıradüzenin" oluşturulduğu açıktır. Sözlüğün, biçimsel nitelikleri yanında kültürel nitelikleri de önemlidir. Akalın, sözlüğün temelde dillerin ve insanlığın "belleği" olduğunu belirterek, sözlükler aracılığıyla, binlerce yıllık bilgi birikiminin hen bugünün hem de yarının insanına taşındığını vurgulamıştır (Akalın, 2010: 163).

Bayrav, sözlüğün ortaya çıkış sürecinde özellikle iki dilli bir coğrafyanın etken olduğunu belirtir. Bu bölgelerde bugünkü iki dilli sözlüklerin ilk örneği sayılabilecek söz listelerini göstermek mümkündür. Kimi sözlerin yazı dilindeki biçimlerini gösteren ya da yanlış söyleyiş biçimlerini düzelten listelere örnek olarak Reichenau sözlüğü, Appendix Probi ve 955-1025 tarihlerinde Aelfric'in Latince-Saksonca sözlüğü verilebilir (Bayrav, 1998: 88-89).

Rönesansla birlikte insanın bilgi dünyasının ve nesnelerle olan ilişkisinin artmaya başlaması ve özellikle Amerika'nın keşfi yeni birçok bilginin üretimini ve tanımlanmasını sağlamıştır. Ulaşım olanaklarının artması üretilen nesnenin dolaşıma çıkmasıyla birlikte ticaret nesnesi olarak çeşitli ürünlerin yabancı dillerde tanıtılması gereği de ortaya çıkmıştır. Bir yandan ticaret diğer yandan din adamlarının çeşitli amaçlarla yerel dilleri öğrenme gereksinimleri Rönesanasla birlikte çok dilli sözlüklerin hazırlanmasını hızlandırmıştır. Edebiyat, felsefe ve kutsal metinler üzerindeki çalışmalar bu dönem Hümanizm akımının amacına uygun olarak çeşitli dil bilgisel araştırma çabalarına dönüşmüştür. Bu çalışmalar klasik Latince, eski Yunanca ve İbranice gibi dillerin öğrenilmesini gerektirdiğinden bir süre sonra karşılaştırmalı tarihsel dil bilgisi ve ses bilgisi yönteminin temelleri atılmış olmaktadır. Ambrogio Calepino'nun 1542-1592 tarihleri arasında 18 kez basılan çok dilli bir sözlüğü bulunmaktadır. Bir terimin yedi dildeki karşılıklarını veren yapıtın 1586 baskısı 10, 1590 baskısı ise 11 dili kapsar. İlk İngilizce-Fransızca sözlüğü 1632'de Randle Cotgrave hazırlamıştır. 1550'de İtalya'da Crusca Akademisi'nin sözlüğü yayımlanmıştır (age., 1998: 90-91).

Bayrav, 18. yüzyıl Avrupa'sında hazırlanan sözlüklerin belirgin niteliklerinin "Tarafsız birer öğretim aracı olacak yerde, zamanlarındaki yaygın fikirleri ve yerleşmiş kurumları yıkmaya, yeni görüş ve düzenler getirmeye uğraşan polemik yazılar." olduğunu belirterek, Bayle'in *Sözlük*'ü ile Voltair'in *Felsefe Sözlüğü*'nü bu yapıtlara örnek gösterir. Bu sözlükler madde içinde dönemin görüşüne sakıncalı söylemlere yer vermeseler de maddelerin bir araya getirilişinde çeşitli görüş ve inanışları çürütmeyi hedeflemişlerdir (age., 1998: 92).

Aksan, ilk sözlüklerin hazırlanmasının ardındaki gereğin, ulusların birbirleriyle karşılaşma sürecinde yabancı dil öğrenme gereksiniminin ortaya çıkması olarak ele almaktadır (age., 1998, 70). Sözlük bilimi çalışmalarının günümüz derlem tabanlı dil bilimi çalışmalarındaki veriye dayanan yöntemle birleşmesi özellikle Batı'da yürütülen uygulamalı dil bilimi çalışmalarının önemli bir parçası durumundadır.

Sterkenburg, 1990'lı yıllara kadar ilk sözlüklerin oluşumunda iki dilli sözlüklerden tek dilli sözlüğe doğru bir geçiş olduğunu, bu saptamanın ortaya çıkışında Babillilerin ve Akadların Sümerceyi anlayabilmek için bir çeşit eş anlamlılar sözlüğünü kil tabletlere yazmalarının rol oynadığını belirtir. Babillilerin sözlükleri pratik kullanıma yönelik eğitimsel (pedagojik) niteliklidir. Boisson ve arkadaşları da Mezopotamya'da yaptıkları araştırmada, tek dilli sözlüklerden iki dilli sözlüklere doğru bir gidişin olduğunu saptamışlardır. Bu durumun ortaya çıkmasında, büyük uygarlıkların kendi sözlükleri üzerinde yoğunlaşması, dolayısıyla kendi merkezli bir sözlük geleneğinin oluşturularak çevre kültürler üzerine odaklanılmaması rol oynamış olabilir. Avrupa sözlükçülük geleneğindeyse bunun tersi bir durum söz konusudur. Avrupa'da sözlükler iki dillidir, çünkü Avrupa uygarlıkları daha çok yabancı dillerdeki metinler üzerinde yoğunlaşmışlardır (Sterkenburg, 2003:9).

Sözlüklerin ortaya çıkışında sosyal olguların önemli bir payı vardır. Bu olgular arasında dinsel metinlerin çevirileri ve öğretim amaçlı uygulamalar başta gelmektedir. Hausmann, M.S. ikinci bin yılda sözlükçülüğün gelişiminde dinin büyük bir motivasyon kaynağı olduğunu belirtmiştir. Örneğin Hindistan'da rahiplerin Sanskritçe metinlerin çözümünde sözlüğe gereksinim duymalarında olduğu gibi, Çin'de Konfüçyüs'ün çalışmalarını okumada olduğu gibi her zaman bir sözlüğün hazırlanması için bir motivasyon unsuru bulunmaktadır. Ortaçağ Avrupa'sında sözlükçülüğün gelişiminde din olgusu karşımıza çıkar. Rahiplerin ana dilleri dışında özellikle Latince ve Grekçeyi öğrenmeleri, dinsel metinlerde geçen Latince sözcüklerin anlamlarına ulaşma çabaları ve gereksinimleri, sözlükçülüğün bu bölgede de gelişimini sağlamıştır. Metinlerdeki satır arası yorumlar, bu dönem metinlerinde sıkça kullanılmıştır (age., 2003:9).

Batı'da önemli sayılan bir çalışma Pallas adlı Alman gezginin Rusya'daki dillerle ilgili çalışmasıdır. 285 kavramın temel alındığı sözlükte 200 Avrupa ve Asya dilindeki karşılıkları verilmiştir. Sözlük 1787'de Linguarum Totius Orbis, Vocabularia Comparativa adıyla basılmıştır. Aksan'a göre bu sözlük, çok dilli sözlük biçiminin ilk örneğidir (Aksan, 1998: 70).

Sözlük çalışmaları tarihinde önemli yapıtlardan biri olan ve 500 kadar dilde dua karşılıklarını içeren ve Adelung ile Vater'in hazırladığı Mihridates, 1806-1817 yılları arasında yayımlanmıştır. Özellikle Almanya'da Grimm kardeşlerin (Jacob Grimm, Wilhelm Grimm) 1854'te yayımlanmaya başlanan Alman Sözlüğü toplam 16 ciltten oluşmuş ve yayımı 1961'de bitirilmiştir. Bu sözlük tarihsel sözlük niteliğinde olmakla

birlikte, karşılaştırmalı olarak lehçelerin söz varlığını da kapsamaktadır. Fransa'da Emile Littre'nin 30 yılda hazırladığı Fransız Dilinin Sözlüğü 1873'te basılmıştır. 4 cilt ve bir de ek ciltten oluşan sözlükte sözcüklerin etimolojisi de bulunmaktadır (age., 1998: 71).

Sözlük bilimi teriminin yanında, Türkçede, "sözlükçülük"ün, sözlüğün planlanmasından içeriğinin oluşturulmasına kadar geçen sürede ele alınan yöntemin adını belirtmesi açısından çok daha önceleri kullanılmaya başlandığını söyleyebiliriz. Sözlük bilimi terimi, Büyük Türkçe Sözlük (BTS) veri tabanında "Sözlük yazma ve hazırlama işi, leksikoloji." biçiminde, Gramer Terimleri Sözlüğü'ndeyse "Bir dilin veya karşılaştırmalı olarak çeşitli dillerin söz varlığını sözlük biçiminde ortaya koyma yöntemlerini ve uygulama yollarını gösteren dil bilimi dalı." olarak tanımlanmıştır. Bu tanımlardan terimin, leksikoloji (lexicology) karşılığı olarak ele alındığı görülmektedir.[1]

Sözlükçülük terimi için Fransızca "lexicographie" karşılığı olarak veri tabanında yer alan Dil bilimi Terimleri Sözlüğü'nde, "Sözlük nevinden eserler meydana getirme işi." ve Güncel Türkçe Sözlükte (GTS) "Sözlük yazma veya hazırlama işi, sözlük yazarlığı, lügatçilik." tanımları verilmiştir. Görüleceği gibi, genel olarak sözlüğün oluşturulmasıyla ilgili işlerin ortak adı sözlük bilimi ve sözlükçülüktür.[2] Akalın, dil bilgisi kaynaklarında "sözlük bilimi", "sözlükçülük", "sözlükbilim", "sözlükbilgisi", "sözcükbilgisi" gibi terimlerin bulunduğunu belirterek, terim karmaşasına yol açmamak için leksikografi için "sözlükçülük" teriminin kulllanılmasının uygun olacağını dile getirmiştir Akalın, 2010: 164).

Aksan, Türklerde gelişmiş bir sözlükçülük geleneği olduğunu belirterek bunun en önemli kanıtı olarak XI. yüzyılda yazılmış olan Kaşgarlı Mahmut'un Divanü Lügati't-Türk adlı eserini gösterir (Aksan, 2004: 372). Sözlük Arapça olarak yazılmış yaklaşık 7500 maddeden oluşur. Yazılış amacı Türkçenin söz varlığını tanıtmak ve öğretmektir. Kaşgarlı Mahmut Divanü Lugati't-Türk'te madde başlarını Arap sözlükçülük geleneğine göre sırlamıştır. Madde başlarının geçtikleri cümleleri ve şiirlerden örnekleri de tanık olarak alan Kaşgarlı'nın eserini yaklaşık 800 yıl sonra Ali Emiri Efendi Arap harfleriyle üç cilt olarak basmıştır. Sözlüğün Yeni Uygurca ve Özbekçe çevirileri yapılmış, 1928'de eserde yer alan sözleri kapsayan bir sözlüğü Brockelmann hazırlamıştır. Bu sözlük 1990'da bir tıpkıbasımla tekrar yayımlanmıştır. 1972'de TDK sözlüğün bir madde başı

[1] http://tdkterim.gov.tr/bts/?kategori=veritbn&kelimesec=288264
[2] http://tdkterim.gov.tr/bts/?kategori=veritbn&kelimesec=288274

dizinini yayımlamıştır. R. Dankoff ve J. Kelly, Ölmez'e göre sözlüğün daha sağlam bir çevrisini İngilizce olarak hazırlamışlardır. Memlukluların Mısırda yönetime hakim olmalarından sonra bölgede Türkçenin öneminin artmasıyla Türkçe sözlük ve gramerlerin hazırlanması da hız kazanmış bu bağlamda Tercümân-ı Türkî ve Arabî, Kitabü'l-idrâk ve Ettuhfetü'z-zekiyye gibi önemli sözlükler oluşturulmuştur (Ölmez, 1994: 88, 90).

Türklerde sözlük yazma geleneği, XIX. yüzyıla kadar Mustafa bin Şemsettin'in Ahterî adlı sözlüğü 1545, Mehmet bin Mustafa'nın XVI. yüzyılda yazdığı Vankulu sözlüğü, Mütercim Asım'ın 1814-1818 yıllarında Fîruzâbâdî'nin eserinden çevirdiği Arapça-Türkçe sözlük Kamusü'l-Muhît ile Farsça-Türkçe Burhan-ı Kâtı gibi sözlüklerle gelişimini sürdürmüştür. Cumhuriyet dönemi öncesi yayımlanan sözlüklerden Ahmet Vefik Paşanın 1876'da hazırladığı 2 ciltlik Lehçe-i Osmanî adlı eserinin ilk cildi Arapça ve Farsça, ikinci cildi Türkçe sözlere ayrılmıştır. Türk sözlükçülüğünün önemli bir dönüm noktasında, Şemsettin Sami 1900 yılında yayımladığı Kâmus-ı Türki'sinde, diğer sözlük yazarlarından farklı olarak eserinde Türkçenin söz varlığını hazırladığını belirtmiş, Türkçe söz varlığını öne çıkarmak için çalışmıştır. 1927-1928'de Hüseyin Kazım Kadri'nin Büyük Türk Lugati'nin ilk iki cildi yayımlanmıştır. Çeşitli Türk lehçelerindeki sözlerle yabancı sözleri barındıran sözlüğün ilk iki cildi Arap harfleriyle basılmış, yazarının ölümünden sonra 3. ve 4. ciltleri 1943 ve 1945'te Latin harfleriyle yayımlanmıştır (Aksan, 2004:372, 373).

Cumhuriyet dönemi ve Dil Devrimiyle birlikte Türkçenin kendi söz varlığının belirlenmesi ve işlenmesine yönelik çalışmalar hız kazanmıştır. Bu tez çalışmasının konusu olan Derleme Sözlüğü için ilk önce Anadolu ağızlarının söz varlığının derlenmesi amacıyla bir komisyon kurulmuş ve bütün Anadolu bir söz derleme ağına dönüştürülmüştür. Bu çalışmaların ürünü olarak ilk önce 153.000 fişlik veriyle hazırlanan ve 6 cilt olan Türkiye'de Halk Ağzından Söz Derleme Dergisi oluşturulmuştur. Bu dergi bir sözlük niteliğinde olup 6. cildi ölçünlü dilden halk ağzına bir dizin biçimindedir. Bu çalışmaya dayanarak daha geniş bir sözlük 1963-1982 yılları arasında çıkarılan 11 ciltlik Derleme Sözlüğü'dür, sözlüğün 12. cildi ek maddeleri içerir ve alfabetiktir. Tarihsel sözlük niteliğinde yazılı dilde bulunmayıp eski eserlerde bulunan sözlerin belirlenimi için 1932'de başlatılan çalışmalar, ilk ürününü 2 cilt olarak Tarama Dergisi adıyla 1934'te vermiştir. 1941 yılına kadar XIII-XIX yüzyıllar arasında basılan 120 eser taranmış eser sayısı daha sonra 227'ye çıkarılmıştır. Bu çalışmalar son

olarak Tanıklarıyla Tarama Sözlüğü adıyla 1951-1976 yıllarında 8 cilt olarak yayımlanmıştır (Boz, 2006: 12).

1.3.1. Sözlük Tekniği

Sözlüklerin yapısal olarak üç bölümden oluştuğu görülür: *Dış yapı* (outside), *büyük yapı* (macro structure) ve *küçük yapı* (micro structure). Dış yapıda, sözlüğün kullanımıyla ilgili ek bilgiler yer alır. Sözlüğün yeniden basımı dolayısıyla değişen madde başları varsa bu madde başlarına yönelik editör yorumları yer alabileceği gibi sıklık tabloları, kullanılan işaretler (kodlar), madde başlarının yerleşimi gibi bir kılavuz bu dış yapıyı oluşturur. Dolayısıyla aslında bu yapı sözlüğün bir *üst dil*idir (metalanguage). Dış yapıda sözlükte yer alan kullanım bilimsel, söz dizimsel ve biçim bilimsel ögelerin sözlükte nasıl verildiğiyle ilgili bilgilerin yanında resimle gösterilen birimler de bulunabilir. Dış yapı tek dilli prototipik sözlüklerde ortak olarak bulunmakla birlikte, bazı sözlükçülük geleneğinde sözlüğün girişinde bu bilgiler sözlüğün son sayfalarında ek bilgiler olarak verilebilir. Sözlüğün büyük yapısında sözlük nesnesi olarak madde başları ele alınır. Bu birimler biçim bilimsel olarak ek almamış nötr birimler biçiminde alfabetik olarak sıralanır. Sözlük bilimci, madde başlarının nasıl seçileceğini hedeflenilen kullanıcıya göre belirlemek zorundadır. Kısaltılmamış geniş hacimli sözlüklerde madde başlarının sayısı olabildiğince büyük tutulur. Tek ciltlik görece kısa ya da küçük sözlüklerde dile yeni giren birimler, teknoloji sözcükleri madde başı olarak seçilirler. Büyük yapının oluşturulmasındaki en önemli ölçüt yine kullanıcı odaklıdır: Kullanıcı aradığı birimi olabildiğince hızlı bulabilmelidir. Küçük yapı, madde başlarıyla ilgili verilen bilgiler üzerinedir. Bu bilgiler biçimsel, anlamsal ve kullanım bilimsel olarak düzenlenir. Biçimsel olarak madde başlarının sesletimi, yazımsal özellikleri yanında söz türüne ait nitelikleri ve geçişlilik, ettirgenlik gibi *alt kategori*k (sub-category) özellikleri verilebilir. Madde başının kök ve türemiş biçimlerinin gösterimi de bu biçimsel sırada yer alabilir. Anlam bilimsel olarak madde başının ilk anlamı ve yan anlamlarıyla yakın ve eş anlamlıları belirtilir. Kullanım bilimsel özellikleriyle madde başlarının değerleri gösterilir. Kullanıldıkları konu, biçemleri tanıklarıyla verilip bu bilginin desteklenmesi için söz dizimsel nitelik göstericileri olan *sözcüksel eşdizimlilik* (collocation), *dil bilgisel eşdizimlilik* (colligation) özellikleri gösterilebilir. Bu özelliklerin sıralanışında yine hazırlanma amacı olarak belirlenen kullanıcı kitlesinin nitelikleri göz önünde tutulur (Kirkness, 2004: 63-64).

Sözlük tekniği açısından Boz, sözlüğün hazırlanması sırasında karşılaşılan sorunları; 1. Sözlük hazırlamada seçilecek yöntem, 2. Amaç ve hedef kitlenin belirlenmesi, 3. Madde başlarının belirlenmesi, 4. Madde başlarının sıralanması, 5. Madde başlarının tanımlanması, 6. Madde başları için örnek cümlelerin belirlenmesi ve diğer sorunlar olmak üzere başlıca 7 sorun kategorisinde ele alır. Hazırlamada elle hazırlanan fişlerin yerine mekanik ve elektronik yöntemlerin kullanılabileceğini, buna örnek olarak da gamma 60 adlı araçtan yararlanılarak günlük 100 bin civarında fişin hazırlandığını belirtmektedir. Sözlüğün amacının belirlenmesi içindeki söz varlığının sınırları açısından bir planlamayı gerektirir. Örneğin bir ilköğretim sözlüğünün hazırlanması, genel sözlüğe göre daha kısıtlı ve öğrencinin gereksinimlerini karşılayacak nitelikte olmalıdır. Boz'a göre sözlüğün amacı sadece kullanıcı yönelimli değil kültürel bir değeri olduğu için de aynı zamanda en büyük hedef kitlesi toplumun tamamıdır. Boz, madde başlarının belirlenmesinde ölçütün hedef kitleye göre değişebileceğini belirterek, bilinirlik ve sıklık açısından elimizde güvenilir bir kaynağın bulunmadığını vurgulamıştır. Boz'un bu sorunları konu ettiği Sözlük TDK'nin Türkçe Sözlük'üdür ve sıklık açısından "müsmir" ve "vahi" sözlerinin bilinirlik ve yaygınlığının tartışılabileceğini dile getirmiştir. Söz varlığını en geniş biçimde vermeyi amaçlayan bir sözlüğün amaç olarak söz varlığını geniş tutmasının bu sözlerin sözlükte yer almasını açıkladığını da belirtmiştir. Madde başlarının genel olarak alfabetik sıralanmasının arananı bulma açısından kolaylık getirmesi iyi bir özellik olarak düşünüldüğü kadar, Boz'a göre aynı kökten gelen örneğin bilgi, bilgiç, bilgin, bilgisiz, bilmek sözleri, bil-, bil-gi, bil-gi-siz, bil-giç, bil-gin biçiminde türemeye dayalı bir sıralamayla gösterilebilir. Hem alfabetik hem de türeme sırasıyla hazırlanan bir sözlük kullanıcının değişik arama gereksinimlerine seslenebilir. Madde başlarının tanımlanması sırasında anlamların hangi ölçüte göre sıralanacağı tartışmalıdır. Genel eğilim sık kullanılandan seyrek kullanılana doğru sıralamadır. Burada, sözlerde zamanla yaşanan anlam genişlemesi, daralması, iyileşmesi ve kötüleşmesi gibi kimi anlam bilimsel olaylar sözlük yazarı ya da yazarlarının üzerinde titizlikle durması gereken bir konudur. Sözlüğün dayanacağı verinin sınırları geniş tutulmalıdır, bu verilerden seçilecek tanıkların kalitesi de veri tabanının genişliği ve çeşitliliğiyle doğru orantılıdır. Boz'un diğer sorunlar başlığı altında verdiği uzman sorunu, sözlük planlaması aşamasında, araştırmacının niteliği ve bilimsel yeterliliği açısından oldukça önemlidir. Günümüz dil biliminin özellikle Batı'da görülen çok disiplinli alanlara kayması, disiplinler arası çalışmanın ortaya çıkmasıyla bir anlamda dil biliminin uzmanlık alanlarının çeşitlenmesini de sağlamıştır. Bu

çeşitlenmede bilgi teknolojilerinin sağladığı yığın bilginin sınıflanmasında yetişmiş eleman gücünün payı büyüktür. Türkiye'de bazı Türk dili ve edebiyatı, dil bilimi bölümlerinde sözlükbilim adı altında dersler verilebilmekle birlikte, bu dersin konusu olan alanın ülkemizde yeterince tanındığı söylenemez. Geniş kapsamlı ve nitelikli sözlüklerin kurumlaşmış bir uzmanlar ekibinin işi olduğu açıktır. Bu ekipte bilgisayar bilimci, dil bilimci ve filologların bulunması gerekmektedir (Boz, 2006: 24-31).

Bayrav, teknik terim sözlükleri gibi bilimsel alanlara yönelik sözlüklerin dışında yer alan genel sözlüklerin çoğunlukla ortak dili kapsadığını belirterek bu ortak dilin sınırlarının belirlenmesinin güç olduğunu belirtmektedir. Sözlüğe girecek olan yeni birimlerin (neolojizm) ya da bir terimin ortak dille olan bağlantısının ölçütlerinin neler olduğu tartışmalıdır. Kişilerin söz varlıkları ilgilendikleri konularla sınırlı olduğu gibi, ölçünlü dille olan yazıya dayalı ilişkileri de kişisel söz varlıklarının sınırlarını belirler. Bayrav'a göre sözlükler bir dildeki bütün birimleri eksiksiz bir araya getiremeyen yapıtlardır. Sözlük bilimcilerin sözlük verisini değerlendirmedeki zorluklarından biri sözlük biriminin ne olduğuyla ilgili üzerinde tam olarak anlaşılmış bir standardın bulunmamasıdır (Bayrav, 1998: 93-94).

Sözlük birimlerin tanımlanmasında Bayrav, a) otonim (autonyme) tanım ve b) belirtilenin tanımı olmak üzere iki ayrım yapar. Örnek olarak "fırın" sözcüğünü ele alarak "fırın"ın, birincisi cins ad diğeri de ekmek yapılan ve pişirilen yer adı olarak iki kullanımını verir. Gösteren tek bir gösterilenle ilişkiliyse yani tek anlamlıysa sorun yoktur, ancak çok anlamlı bir gösteren varsa bu anlamın madde başında neye göre sıralanacağı tartışma konusudur. Bayrav, sözlük tekniği açısından madde başlarının tanımında, dairesel bir niteliğin yer almasından kaçınılması gerektiğini, kısaca, bir tanımın tanımlanan birimin eş anlamlısıyla yapılmaması gerektiğini belirterek bu hatanın pek çok sözlükte görüldüğünü vurgular (age., 1998: 95-96).

DDİ sistemlerinde genel olarak sözlüksel çerçeve, sözlüksel bilgi tabanı göz ardı edilmiş, metinlerin ve sözcüklerin yüzeysel yapıları ***bağlamdan bağımsız gramerler*** (contex-free grammar) gibi kural tabanlı uygulamalarla işlenmeye çalışılmıştır. Wilks ve arkadaşları, sözlüklerin düzenlenmesinde, dil bilimsel temelli kuramlar ve psikolojik temelli kuramlar olarak iki farklı kuram çerçevesi belirlemişlerdir. Dil bilimsel temelli kuramlar kendilerini kanıtlamak için sözlüğün yapısına müdahale ederler. Bunun bilgisayar uygulamalarındaki örneği UNITRAN adlı çeviri sistemidir. Bu sistemin dayandığı kuram Jakendoff'un 1983'te ortaya koyduğu sözlüksel kuramsal yapıdır. Psikolojik temelli yaklaşım, insanların sözcük çağrışım sistemleri ve kullanımıyla ilgili

deneysel bilgilerden oluşmaktadır. Bu yaklaşımın ortaya çıkardığı sistem WordNet'tir. WordNet'te, çağrışım testleriyle sözcükler arasındaki anlamsal ilişkiler tanımlanıp her bir kavram alanı bir set olarak belirlenmiştir. Reaksiyon zamanı testi ile insan belleğinin sözcük işleme düzeneği modellenmeye çalışılmıştır (Wilks, 1995: 122-126).

Wilks ve arkadaşları *makinece okunur sözlük*lerle (Machine readable dictionary-MRD) (MOS) *makinece işlenir sözlük* (Machine tractable dictionary-MTD) (MİŞ) arasında ayrım yaparlar. Buna göre, MİŞ'ler MOS'ların daha da rafine edilmiş biçimleridir. Yani makinenin bir sözlüğü okuması, onun basılı biçiminin elektronik ortama aktarılmış biçimiyle sınırlıyken, MİŞ'ler DDİ uygulamaları için sözlüğün yapılandırılmış ve dönüştürülmüş biçimleridir.

Sözlüğün kendisi dolayısıyla MOS'lar içlerinde örtük olarak aslında yapısal bilgiyi tutarlar. Bu bilgiler, algoritmik bir yaklaşımla yani yazılımsal bir süreçle açığa çıkarılabilirler. Bilgisayar destekli dil bilimi açısından bu süreç, sözlük için kodlanmış olan dilsel verinin paketini açmak, örüntüleri açık seçik ortaya koymak olarak anlaşılır. Bir sözlüğün otomatik olarak işlenmesi yaklaşımında Wilks ve arkadaşları iki ayrım yapmışlardır: 1) Otomatik işleme, tanımlardaki söz katarlarını (string) ve kodları daha sonradan rastgele erişime olanak tanıyacak biçime sokulması, buna sözlüğün yapısının açıklaştırılması denebilir. 2) Tanımlardaki anlamca bulanık ögelerin belirlenmesi ve bu ögelerle tanımlar arasındaki ilişkilerin çıkarılmasıdır. Bu işlemeye de örtük bilginin ortaya çıkarılması adı verilebilir.

Bir sözlüğün anlamca işlenmesi sırasında sözlükteki bilgilerin nasıl bir biçimde ve sırada düzenlendiğini bilmek kadar, bu düzenliliklerin örüntüsünü ve ipuçlarını çıkarmak da önemlidir. Wilks ve arkadaşları, sözlüğün anlam bilimsel yapısından yola çıkarak üç varsayım öner sürerler: *Yeterlilik* (sufficiency), *çıkarımlanabilirlik* (extricability) ve *önyükleme* (bootstrapping). Yeterlilik, ele alınan sözlüğün güvenilir bilgi taşımasıyla ilgilidir. Sözlüklerin anlamsal bilgisinin bağlamdan bağımsız bilgi içermeleri nedeniyle yetersiz olduğu göz önüne alındığında, sözlüğün diğer birimlerinden hareketle istenilen anlamsal yapının ortaya çıkarılması süreci, Wilks ve sözlüğün yeterlilik özelliğiyle açıklanabilir. Yani, sözlük verisinin kendi içinde birbirini tamamlaması yeterlilik özelliğiyle ilgilidir. Çıkarımlanabilirlik, sözlüğün DDİ için uygun olup olmaması yani işlemsel süreçlerin sözlüğe ne kadar uygulanabildiğidir. Önyüklemeyse, herhangi bir otomatik işlemden önce sözlük üzerinde gerekirse elle yürütülen ön işlemlerdir (Wilks, 1995: 138, 141-142).

Boagards, sözlüğü, olgular hakkında sadece bir alfabetik yapı olarak değil, aynı zamanda bir dilsel başvuru aracı olarak görür. Sözlüklerin kullanım amacına bakıldığında, en çok, sözlüğün okumaya yardımcı bir araç niteliği göze çarpar. Daha az oranda da yazmaya yardımcı araç olarak görülür. Dinleme ve konuşma eğitiminde de okumaya göre sözlüğün kullanımı azalmaktadır. Dil bilgisel ve köken bilgisel araştırma amaçlarıyla sözlük kullanımı diğer kullanım amaçlarına göre en alt düzeyde kalmaktadır (Boogards, 2003: 26-27).

Sözlüklerin eleştirisini yapmanın, sözlük kullanıcısının gereksinimlerini göz önüne almanın ve uygun özel sözlüklerin hazırlanması aşamalarının tanımlanması için kullanılan terim, ***üst sözlük bilgisi*** ya da ***üst sözlükçülük*** (meta-lexicography) olarak adlandırılır. Zöpgen'e göre üst sözlükçülük halihazırdaki sözlüksel bilgi yerine daha çok kullanıcının somut gereksinimini ve isteklerini dikkate almalıdır (age., 2003: 28).

Sözlüğün oluşturulmasında, içerik tasarımı, belirlenen hadef kitle doğrultusunda ana çerçevesini kazanır. Seçilen madde başlarının hedef kitle (kullanıcı) düzeyinde belirlenmesi emek yoğun bir süreci içerdiği gibi, madde başı tanımlamalarının yapılması da bir o kadar yoğun anlam bilimsel bir ilişki ağının çözümlenmesini gerektirir. Geeraerts, anlam tanımlama ve tanımlamalar arası ilişkileri başlıca iki ana başlık altında tartışmıştır: Birinci sorun semasiyolojik tanımlama ölçütüdür. Bu tanımlamada tanımlanan sözcükle ilgili anlamsal içerik sunulur. Çok anlamlı birimler madde başının küçük yapısını oluştururular. Diğer ölçüt onomastik ölçüttür. Bu ölçütte, sözcük, bağlı bulunduğu kavram alanı ile ilişkilendirilir. Kavram alanı ile ilgili birimler bu alanın içine alınırlar. Günümüzde sözlükler semasiyolojik tabanlı tanımlamaları ilk sırada sunup, onomastik özelliği küçük yapının içinde vermeye başlamışlardır. Onomastik yapı, sözcüğün konu alanıyla da ilgili olabilir. Bir tıp teriminin semasiyolojik tanımından sonra uygun kısaltma ile onomastik kullanımı verilebilir. Tanımlamada, madde başından hemen sonra verilen içerik betimleyici niteliktedir. Bunun literatürdeki adı gösterimsel anlam (denotational meaning) olarak belirlenmiştir. Geeraerts, betimleyici tanımlamada üç tür anlam kategorisi olduğunu belirtmiştir: Duygusal anlam (emotive meaning), dil bilgisel anlam (grammatical meaning) ve kullanımbilimsel anlam (pragmatic meaning). Duygusal anlam alaycı ya da kaba kullanımla ilgilidir. Dil bilgisel anlam için yazar, İngilizce "that" örneğini vererek bu sözcüğün bir nesneyi karşılamadığını, bir sözdizimsel işlev görünümü oluşturduğunu belirtir. Kısaca dil bilgisel anlam işleve yöneliktir. Kullanım bilimsel anlam örneği olarak Türkçeden "hoş geldiniz, iyi günler" gibi karşılamada kullanılan selamlama

işlevini yerine getiren sözleri örnek verebiliriz. Kullanım bilimsel anlam, dil biliminde söz edimleri kuramının bir parçası olup, sözcüklerin belirtme değerinden çok bağlam ifade etme kapasiteleriyle ilgilidir. Tanımlama biçimlerinden dar ya da yoğun tanım (intensional definition) Geeraerts'e göre, bir kategori üyelerinin genel nitelikleriyle ilgilidir. Örneğin kuşun uçması, tüylerinin olması bu tanıma girerken, geniş tanım (extensional definition) kuş tüylerinin adlarının verildiği bir tanım çeşididir (Geeraerts, 2003: 84, 87, 88).

1.3.2. Sözlük Verisi

Hazırlanacak sözlüğün dayandığı veriyi Čermák birincil ve ikincil kaynaklar olmak üzere ikiye ayırır. Birincil kaynaklar bir arşiv ve derlemden, ikincil kaynaklar ise alan çalışmaları, daha önce yayımlanan sözlükler, ansiklopediler ve internettir. Sözlüğe alınacak ögeler için bazen birincil kaynaklar yetersiz kaldığında, tanımların daha açıklayıcı ve belirsizlikten uzak olması için özel teknik alanlara gereksinim duyulabilir. Tek dilli sözlükler genel olarak birçok kaynağı, örneğin derlem ya da diğer sözlükleri tek bir ana kaynak olarak kullanırlar. Čermák, ikincil kaynakların daha çok pragmatik (faydalanmacı) amaçlı kullanıldığını, yeni bir sözlüğün hazırlanması amaçlandığında sözlük bilimcilerin, genelde daha önceki sözlükleri ya da önceki basımları göz önüne aldıklarını belirtmektedir. Bu noktada sözlük bilimci, var olan birimlerin güncellenmesi, sözlükten çıkarılacakların ya da değişeceklerin belirlenmesi için ikincil kaynakları etkili biçimde kullanabilir. Bir dilin sözlüğünün ilk kez hazırlanması durumunda, ikincil kaynaklar yeterli olmayabilir. Böyle bir durumda görüşme ve anket teknikleri iyi bilinen sözlük bilimi bulgulama yöntemlerindendir. Bu yöntemin en önemli sakıncası zaman alması ve pahalı olmasıdır. Yazılı bir geleneği olmayan dillerin sözlüklerinin hazırlanmasında, o dilin konuşucuları arasında bir sözlük bilimci olmadığından dışarıdan bir sözlük bilimcinin gerekli malzemeyi derlemesi gerekir. Standart sözlüklerin hazırlanmasında görüşme ve anket tekniği, asıl olmasa da özellikle ağız sözlüklerinin hazırlanmasında ilk başvurulan yöntem olmaktadır (Čermák, 2003: 18-20).

İkincil kaynak olarak Yahoo, Google gibi internet arama motorlarının kullanımı özellikle yeni kavramların ve sözcük kullanımlarının araştırıldığı çalışmalarda değerli veriler sağlamaktadır. Bu internet arama motorlarındaki sorgu sonuçları kimi belirsiz ve ikili sonuçlar çıkartabilmektedir. İnternetteki sözlük bilgisinin ortaya çıkaracağı

tartışmalı, standart dışı kullanımlar sözlükte ayrı bir kullanım kutucuğu ya da bölümü olarak gösterilebilir.

Kaynakların kullanımı açısından diğer bir önemli nokta verinin büyüklüğü ve yeterliliği sorunudur. Hapax legomenon adı verilen ve derlenen verilerde yalnız bir kez kullanılan ögeler ile verinin çok büyük büyutlar tuttuğu derlemlerdeki şişmiş sözcük sıklıkları en önemli sorunlardandır. Hapax legomenon sorunu tarihsel dönem çalışan sözlük bilimciler için çok daha büyük bir sorundur, çünkü tarihsel döneme ait kayıtların içeriği ve büyüklükleri oldukça kısıtlı olabilmektedir. Derlem dil biliminin sağladığı olanaklar sayesinde aynı dil için hazırlanmış farklı konu ve büyüklükteki derlemler sözlük bilimci için aranan sözlük birimler üzerinden karşılaştırma seçeneği sunarlar. Aşırı veri yığılmasının ve bir kez geçen öge dezavantajının çözümünde sözlüğün amacı ve madde başı olacak sözcüğün tipi önemlidir. Bütün bu durumlarda asıl olan ilk olarak birimin sıklığı temel olarak dikkate alınmalıdır (age., 2003: 18-20).

Čermák, sözlüklerin oluşturulması için yapılandırılacak derlemlerin hazırlanmasında hangi dil katmanının hedeflendiğinin önceden belirlenmesi gerektiğini belirterek, sözlüğün özel amacı için özel bir derlem temsilinin varlığının gerekliliğini dile getirmektedir. Bunun için dengeli bir metin ve kayıt sayısına ulaşılması gerekmektedir. Dengeli bir kaynak için iki farklı dil katmanı için iki ayrı boyuttan söz edilebilir. Birinci boyut geneldir yani genel konuşucu-okurun kullandığı boyut, diğeri ise yazılı ve sözlü özelliği olan dil boyutudur. Sözlükler genel olarak yazılı boyutu esas alırlar. Sözlük bilimci, çalışılan derlemden bilgisayar aracılığıyla elde edilen ***bağlamlı dizin***i (concordance) bir fiş biçiminde düşünebilir. Bağlamlı dizindeki her bir satır, aslında etiketlenmesi yapılmış bir fiştir. Sözlük bilime yönelik oluşturulan veri tabanlarını da ikiye ayırır. Birincisi ikincil kaynaklar olarak daha önce sözü edilen verilerde kullanılan MOS, ikincisi ise sözcüksel veri tabanlarıdır. İkinci veri tabanı tipi geleneksel sözlüklerde gösterilmeyen açıklıkta bir etiketleme sistemine sahip olabilir. Bunda bu veri tabanı tipinin özellikle DDİ uygulamalarına girdi sağlamasının etkisi vardır (Čermák, 2003: 21-24).

İngilizce için hazırlanan ilk sözlüklerin (Jhonson, Murray ve Webster) anlam dizimleri tarihseldir. Sözcüğün en eski anlamı önce, yeni anlamları sonra verilmiştir. Bilgisayar uygulamaları için bu sıralama Hanks'e göre uygun değildir. Sözcüğün ilk ve köken bilgisel olarak eski anlamının günümüz metin işleme uygulamaları için değeri bulunmamaktadır. İnsanlar için hazırlanan sözlüklerin bilgisayar uygulamalarında sorun oluşturabilecek yanlarını Hanks şu biçimde açıklamıştır: "1) İnsanlar için hazırlanan

sözlüklerde anlamlar çok iyi ayrımlanmıştır. Bu durum bilgisayar uygulamaları için her zaman kolay işlenir bir ortam sağlamamaktadır. Önemli ve önemsiz anlamlar arasında ayrımlama yapacak bir algoritma geliştirmek güçtür. 2) Sözcüklerin ayrı anlamları, yine bu anlamlarla ilişkili sözdizimsel yapılarla açık biçimde ilişkilendirilmemiştir. 3) Sözlüklerde genellikle sözcük ve anlam sıklıkları bulunmamaktadır." (Hanks, 2003: 60). Tüm bu olumsuzluklar günümüzde sürmekle birlikte, bilgisayar destekli dil bilimi uygulama alanlarının gelişimiyle, halihazırdaki bu sözlükler yeniden farklı amaçlar için işlenmeye başlanmıştır. İnsanlar için hazırlanan sözlüklerde, insanın sağduyusuna ya da gerçek dünya bilgisinin önceden varsayılması sonucunda madde başları ve anlamları bir liste biçiminde anlamın yapısal kurucuları gösterilmeden verilmekte bu da makineler için büyük bir sorun yaratmaktadır. Çünkü makineler insanlar gibi genellemeler yapamazlar, mutlaka metinsel ipuçlarına gereksinim duyarlar. 1985'te Choueka ve Luisgnan anlamın oluşumuna bir referans olarak ilk kez *co-text* adı verilen kavramı ortaya atmışlardır. Böylece sözcükler arasındaki en önemli istatistiksel ilişkilere yönelik çalışmaların önü açılmıştır. Aslında sözcüğün anlamının kurucuları olarak çevresindeki birimlerin önemini Firth 1957'de "Bir sözcüğü eşlik ettiklerinden bilirsiniz." ifadesiyle belirtmiştir. Böylece eşdizimlilik kavramının gelişmesi ve makineler için önemli anlam belirleyicileri olacak metinsel ipucu sağlayıcılarından olması, eşdizimliliğe dayalı sözlük bilimsel çalışmaların hızını artırmıştır (age., 2003: 60).

1.3.3. Terim Sözlükçülüğü

Bowker, sözlükçülükte terim sözlükçülüğü konusuyla özelleşmiş sözlük hazırlama disiplinini ilişkilendirmektedir. Özel alan sözlüklerinin terminolojiyle ilgili olmakla birlikte, terim sözlükçülüğünde sözcüğün tanımının esas olduğunu, terminolojide ise daha çok onomastik niteliklerin yani daha çok konuya duyarlı kavramların, kavram alanlarının çıkarılıp sözcük düzeyine indirgenmesinin esas olduğunu belirtir (Bowker, 2003: 155).

Terim sözlükleri alana duyarlı, konu alanına yönelik sözlüklerdir. Genel sözlüklerin hazırlanmasında başta gelen kural olarak terim sözlükleri de bir belirli hedef ve kitle doğrultusunda hazırlanırlar. Bu sözlüklerin başlıca amaçları, aynı konuda çalışan uzmanlar arasında iletişimi kolaylaştırmaktır. Konu alanının uzmanı olan ve olmayanlar için hazırlanan sözlüğün içerik örüntüsü ve taşıdığı ansiklopedik bilgi de değişmektedir. Özel amaçlı sözlüklerin büyük ve küçük yapı özellikleri de genel

sözlüklerden farklılıklar gösterir. Özel amaçlı sözlükler alfabetik olmakla birlikte, Bowker'a göre sistematik olarak kavramsal ilişkilerin bir arada verildiği bir büyük yapı bu tür sözlükler açısından daha uygundur. Sözlüğe başvuran açısından aradığı terimle ilgili genel çerçevenin görülmesi iyi bir terim sözlüğünün özelliğidir. Sistematik sıralama en çok fen bilimlerine yönelik sözlüklerde görülmekte, bu sıralamanın gelişiminde tarihsel olarak zooloji ve botanik bilimlerinde ortaya konan ilk sınıflama çalışmalarının rolü büyüktür (age., 2003: 156-158).

Terim sözlüğünün küçük yapısı da hedef kitleye göre değişir. Uzmanlık alanının dışına seslenen bir sözlük daha fazla, konu uzmanına göre hazırlanan bir sözlük ise daha az bilgi verecektir. Çok dilli terim sözlükleri bilgi yoğunluğu açısından daha az yüklü olmakta, terimin sadece karşılığını verme amacı güdülmektedir. (age., 2003: 159).

Elektronik kaynaklar olarak terim sözlüklerinin önemli biçimleri terim bankalarıdır. Bunlar 1960'lı yıllarda çeşitli bilgisayar uygulamalarının ve özellikle makineli çevirilerin gerçekleştirilebilmesi amacıyla hazırlanmış ayrıntılı yapılara sahiptir. Ayrıca, terim sözlüklerinin genel sözlüklerde olduğu gibi elektronik ortamda (CD, internet) sunulması ve madde başlarına gerek metin kutusu arama yolu gerekse bağlantılı tıklama yoluyla anında erişim sağlanması, sözlük tekniği açısından alfabetik yapının varlığını ve katılığını tartışılır duruma getirmiştir. Terim bankaları bu açıdan oldukça dinamiktir ve güncellenmeleri çok kolaydır. Bowker'a göre, terim sözlüğü alanında çalışacak sözlük bilimci, çalışacağı alanla ilgili konunun önceden kavramsal haritasını genel olarak çıkarmalı ve alanla ilgili ilk okumalarını yani genel bilgi amaçlı çıkarıma gidecek çalışmalarını tamamlamalıdır. Böylece hem alana olan kavramsal bilgi yakınlığı artacak hem de alanın uzmanlarıyla olan işbirliği ve iletişim gücü kuvvetlenecektir (age., 2003: 161).

Sinclair'e göre sözlük "Bir dilin ya da bir uygun tutarlı alt dilin söz varlığını betimler." Burada sözü edilen tutarlı alt dil, dilin çeşitli değişkelerinin ya da varyasyonlarının ölçülü ve dengeli biçimde bir araya getirilmiş bir derlemidir. Sinclair, sözlüğün böylece aslında bir derlemi de betimlediğini belirtmektedir. Sözlük bilimci, sözlük projesinin başlangıcında bir derleme güvenirlilik açısından nasıl yaklaşacağını test etmelidir. Çünkü günümüzde sözlüğün dayandığı ana malzeme Sinclair'e göre derlemlerdir. Yine Sinclair'e göre derlemden elde edilecek tekrar bilgisi oldukça yol gösterici olmaktadır, çünkü bireysel olarak bir söz varlığı biriminin bütün tekrarlı kullanımlarını bilmek söz konusu değildir. Bu yüzden derlem, söz varlığının tekrarlı görünümleri açısından son derece önemlidir. Sözlük için hazırlanacak derlemde

metinsel bütünlük aranmalıdır. Derlemde metinsel bütünlüğün iki sağlayısıcı vardır: 1) Orijinal metnin saklanması 2) Metnin dijital bir kopyasının elde edilmesi. Metnin daha sonra arandığında ulaşılabilmesini sağlayan bir de metin kimliği, her bir metin için tanıma işareti olarak kullanılmalıdır. Konuşma dilinin elektronik olarak kaydı geniş ölçüde elde edilebilirken bu kayıtların çeviriyazılarının yapılabilirliği sorunludur ancak çeviriyazısı yapılmış elektronik metin, bazı özelliklerinin sonradan işlenmesi ve düzeltilmesi gerekse de sözlüğün ana verisi olduğundan arşiv malzemesi olarak değerlidir (Sinclair, 2003: 167, 170-171).

1.3.4. İnternet Sözlükçülüğü

Çevrim içi kavramı TDK Bilişim Terimleri Sözlüğünde İngilizce *online* sözcüğünün karşılığı olarak "Giriş verilerinin oluştuğu yerde, anında girildiği, çıkış verilerinin gereksenen yerde, anında kullanıma sokulduğu herhangi bir donanım vb. kaynağın bu niteliğine ya da bu tür kullanım biçimine değgin."[3] biçimindeki tanımında, sürecin "anında" olarak vurgulanması görülür. Türkçede *gerçek zamanlı* (real-time) sözcüğüyle bu kullanım eş anlamlı olarak görülmekle birlikte, bu iki kavramın kullanıldığı bağlamlar farklıdır. İnternetin gelişimiyle birlikte hemen her alanda olduğu gibi kitap ortamından elektronik ortama geçiş, sözlüklerin internet üzerinde biçimlenmesini de gerektirmiştir.

Matbaanın bulunuşundan günümüze gelinen noktada basım ve yayım işlerinin bütün dünya üzerinde genel bir ağ yapısıyla saniyeler içinde gerçekleştirilmesi, dil konuşucusu ve kullanıcısı için yeni teknolojileri ve sorgulama farkındalığı denilebilecek bir durumu gündeme getirmiştir. Günümüz internetinin önde gelen arama motoru Google, kurulduğu 1998'den bu yana geliştirdiği alt yapısı ve sunduğu hizmetlerle bilginin işlenmiş biçiminin kullanıcıya ulaştırıldığı basit ve işlevsel arayüzüyle arama kavramının çağrışım alanını değiştirmeye başlamıştır. Basit bir sorgu arayüzüyle karşılaşan dil kullanıcısı genellikle söz ya da söz öbeklerini ararken karşılaştığı metinlerle kurduğu ilişkide bilmediği sözlerle karşılaşma olasılığını da farkında olmadan artırmaktadır. İnternet ne tür bir ortam sağlarsa sağlasın, içeriği her zaman metin ve yazı tabanlı olacaktır. Dolayısıyla aslında internet ortamı dil kullanıcısının her zamankinden fazla dille karşılaştığı bir ortam yaratmakla birlikte bu ortam tek düze değil çok boyutlu bir alandır. Resmi yazışmalardan söyleşilere (chat), belge

[3] http://tdkterim.gov.tr/bts/?kategori=veritbn&kelimesec=73633

barındırmaktan e-posta ve oyunlara kadar hayatın hemen bütün alanlarının dilsel içeriği internetle yeniden boyutlandırılmış ve özellikle metin ve yazı önceden olmadığı kadar biçimlendirilebilir duruma gelmiştir (Yıldırım ve Tahiroğlu, 2006: 363).

Yazıyla daha çok ilişki kurmaya başlayan internet kullanıcıları sözlüklere her zamankinden daha çok gereksinim duyar duruma gelmişlerdir. Tipik bir internet kullanıcısı genellikle günlük internet kullanımının başlangıcını haber ya da gazete sitelerine ayırmaktadır. İnternet haberciliğinde günlük olayların anlık olarak verilmesi, sözlerin çevrimini de teknolojiyle birlikte son derece hızlandırmıştır. Haber metinlerinin internetin dili olan "html" biçiminde saklanması, arşivlenmesi ve bu metinlere erişebilmek üzere haber içi arama, arşiv sorgulama sistemlerinin geliştirilmesi, internetteki dilsel kullanım ve dilsel olana ulaşım koşullarını da sürekli değiştirmektedir. Kısacası bir internet kullanıcısı dille kurduğu ekran okuması bağına bir de arama-sorgulama bağı eklemekte, bu da kullanıcıyı basılı sözlük biçiminden anında ulaşılan çevrim içi sözlüğe ulaştıran dolaylı yollardan birini oluşturmaktadır.

Sözlüklerin basılı biçimlerinden elektronik ortama aktarılma aşamasından sonra elde edilen yoğun disk biçimiyle, internetteki biçimleri arasında arayüz farklılıklarının olduğu görülür. Yoğun disk içeriği internet üzerindeki bir ilişkisel veri tabanı olarak yapılandırılmış ortama aktarılmaktadır. Yoğun disk ortamındaki sözlükler bir masaüstü ya da dizüstü bilgisayarda internet bağlantısı olmadan (kimi sözlüklerde internet bağlantısıyla güncelleme olanağı sağlanmaktadır) çalıştırılabilirken internet sözlükleri, arayüzlerini bir veri tabanı sunucusundan aldıkları veriyle oluştururlar. Bu yapı sorgu arayüzleri çeşitliliğine karşın veriler hemen her zaman bir veri tabanında sorgulanmaya uygun bir yapılandırma içerisinde tutulurlar.

Tek dilli standart sözlüklerin internet üzerindeki sunumlarına bakıldığında bu sözlüklerin genel olarak tek bir arayüz üzerinde ayrıntılı sorgulama olmadan bir sorgu kutusu ile kullanıcıya ulaştıkları görülür. Bu tip sözlükler piyasada marka olmuş ve birçok basımı bulunan Longman, Cambridge, Oxford ve Webster gibi içinde kurumsal yapıları da barındıran İngilizce hazırlanan sözlüklerdir. Ana sayfada seçilmiş o güne özgü sözcük bildirimlerinin yanında, sözlüğün nasıl kullanılacağının anlatıldığı sayfalara bağlantılar içeren sözlük arayüzü yapıları bulunur. Sorgu girişinin bulunduğu sayfaların tasarımında görülen basit bir işlevsel yapı, aslında sözlük bilgisinin bütün ayrıntılarıyla tutulduğu karmaşık bir veritabanı yapısının hem gizlenmesini sağlamakta hem de sorgu sonuçlarının ortaya koyacağı işlevselliği hazırlamaktadır. Örneğin Longman çevrimiçi sözlük ana sayfası (http://www.ldoceonline.com/), büyük puntolarla

kullanıcının doğrudan arama gereksinimini karşılayıcı ve dikkat çekici bir görsellik sunarak, kullanıcıyı öncelikle sorgulamaya yönlendirmektedir.



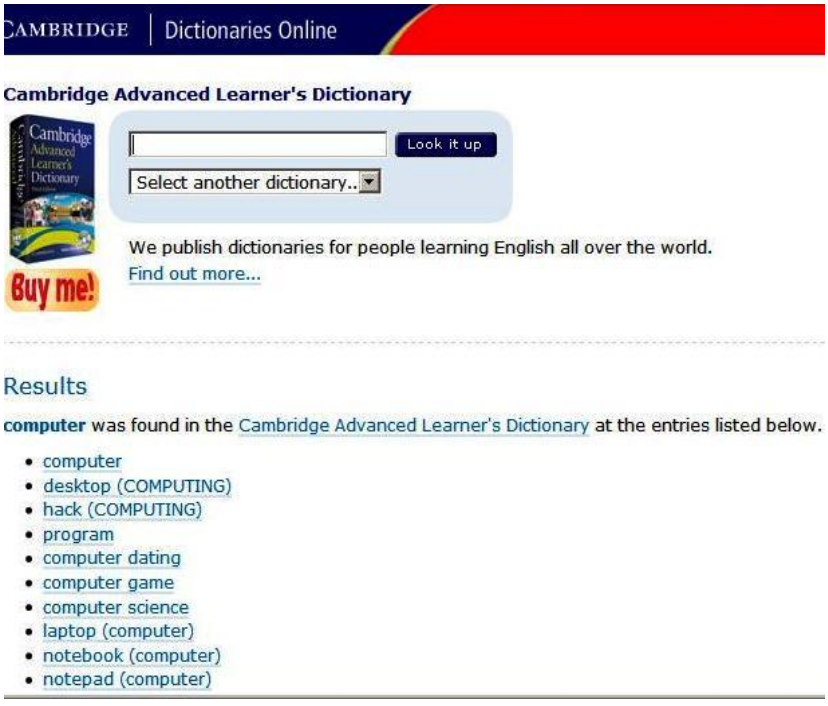
Şekil 1. Longman Çevimiçi Sözlük Ana sayfa Sorgu Arayüzü (http://www.ldoceonline.com/)

Genel web sayfası tasarımında standart içerik sunumuna paralel olarak sözlük içeriğinin sunumu da günümüz internet dillerinin gelişimi doğrultusunda (html, asp, asp.net, php gibi yorumlanan diller ve javascript, ajax. css gibi tarayıcının yorumladığı diller) profesyonel çalışma ekipleriyle gün geçtikçe işlevsel fakat karmaşık sorgulama süreçlerinin oluşturulması yönündedir. Bir web sayfasının tasarım kuralları genel olarak sözlük arayüzlerinin tasarımında da geçerli olmakla birlikte, arayüzün veritabanıyla olan bağlantısının hızlı bir biçimde sağlanması ve üretilen sonuçların da daha verimli ve hızlı dökümünün yapılması göz önünde tutulmaktadır. Bu amaçla sayfalarda büyük boyutlu resimlerin kullanılmasından kaçınılmakta, arayüzü sunan sunucu makineyle veri tabanını sunan makine ayrı olarak işlemleri yürütmektedir.

Sözlük verisinin sonuç olarak üretildiği sonuç gösterim sayfalarında genel olarak aranan söz ya da sözün içinde geçtiği birimler kendi aralarında bağlantılandırılarak bir ara sonuç, en sonunda aranan sözün tam olarak kendisi gösterilmektedir. Aşağıda Longman ve Cabridge sözlüklerinin çevrimiçi sürümlerinin ara sonuç sayfaları gösterilmiştir.



Şekil 2. Longman Sözlüğü Ara Sonuç Görünümü
(http://www.ldoceonline.com/search/?q=computer)



Şekil 3. Cambridge Sözlüğü Ara Sonuç Görünümü
(http://dictionary.cambridge.org/search/british/?q=computer&x=0&y=0)

Yukarıdaki şekillerde sorgulanan sözün tam olarak içeriğinin doğrudan verilmesi yerine sözle birlikte geçebilecek diğer birimler gösterilerek kullanıcıya bir söz

seçenekleri tablosu sunulmuş olmaktadır. Bu biçimde bir gösterim yöntemiyle sözlük okurunun söz dağarcığının geliştirilmesine, hatta sözlük kullanımının özendirilmesine katkısının olacağı söylenebilir. Bu noktada sözlüğün basılı biçiminde kullanıcının sözler arasında gezinti olanakları kısıtlıyken çevrimiçi biçiminde sözlükten faydalanma olanakları artmakta aynı zamanda sözlük içeriğinin metin yazım süreçlerinde kullanılması, taşınması ve buna dayalı olarak yeniden bilgi üretimi çok daha hızlı işlemektedir.

Son yıllarda internet tarayıcılarının (browser) gelişen ve artan yetenekleriyle birlikte, sayfa içeriğinin sunum biçiminde özellikle ayrı bir bilginin ayrı bir küçük (kutu) pencere (pop-up) içinde gösterilmesi yaygınlaşmıştır. Özellikle ansiklopedi sayfalarında yer alan madde içi bilginin yine o maddeyle ilgili hatırlatıcı bilgisinin verilmesinde ve resmi kurum sitelerindeki duyuruların gösterilmesinde kutu pencere özelliğinden yararlanılmaktadır. Bu özellikle yakından ilgili olarak sözlükteki madde başlarının içeriklerinde geçen tanım bilgisindeki her bir söz yine aynı sözlüğün veritabanıyla bağlantılandırılabilmektedir. İlk örneklerinin yabancı haber siteleri ve ansiklopedi sitelerinde görülen bu uygulama, günümüzde büyük sözlük yayımcılarının sitesindeki uygulamalarda yer almaktadır. Bütün sözlük metni, sözlüğe duyarlı biçime getirilerek, elektronik metin okuma süreçleri kapsamlı bir değişime girmiş bulunmaktadır.



Şekil 4. Longman Metne Duyarlı Sözlük uygulaması.
(http://www.ldoceonline.com/popup/popupmode.html?search_str=electronic)

Şekildeki sözlükte, “computer” sözcüğü sorgulanmış, tanım metninde geçen “elektronic” sözcüğü çift tıklanarak sözcüğün tanım içeriği kutu pencerede görüntülenmiştir. Metne duyarlı bu uygulama Cambridge sözlüğünde de bulunmakla birlikte ayrı bir sözlük ara sonuç sayfası çağrılarak gösterilmektedir.

Sorgulamadaki ara sonuçlarından sonra gösterilen tam sonuç sayfasında tek dilli internet sözlüklerinde özellikle yabancı dil öğretiminde kullanılan sözlüklerde sesletim özelliği de yer alırken resimle desteklenen sonuçlara da yer verilmektedir.

İnternet üzerinde yer alan çok dilli sözlük yapılarına bakıldığında bu tür sözlüklerin temel olarak bir hedef ve birçok amaç dil için sorgu sayfaları biçiminde düzenlendiği görülür. Veri tabanlarında barındırdıkları dil kadar veri bulunurken ana sayfaları da bu dillerden birinin çevrisini içeren başka bir ana sayfa tasarımı da içerirler. Dolayısıyla bu sözlükler aslında ana dillerinin seçenekli sunulduğu bir tür basit çeviri sayfaları niteliğindedirler. Örneğin wordreference.com alan adlı sitede İngilizce ana sayfa tasarımında sorgulanacak sözcüğün temel alındığı bir metin kutusu ve İspanyolca, Fransızca, İtalyanca, Almanca radyo düğmesi seçeneklerinin alt grupları olarak karşılıklı çeviri seçenekleri verilmektedir. Sözlükte İngilizceden Türkçeye sözlük bölümü de yer almaktadır. Bu sözlük sisteminin bir özelliği de özellikle son yıllarda yaygınlaşan forum siteleri gibi bir tartışma ve paylaşım alanına sahip olmasıdır. Sorgulanan sözlerle ilgili sözlüğün bir forum alanı bulunmakta, kullanıcılar sistemin verdiği tanımlar ya da sonuçlarla ilgili düşüncelerini ve eleştirilerini paylaşabilmektedir. Böylelikle aslında çevrimiçi sözlük yapılarının basılı biçimlerden en büyük farkı da anında geribildirim olanağının ortaya çıkması ve böylelikle bir tür katılımlı sözlük alanlarının oluşmuş olmasıdır. Wordreference de diğer çevrimiçi sözlük yapıları gibi bir ara sonuç gösterim sayfasına sahiptir. Sorgulanan sözcüğün sesletimi de bulunmakla birlikte sesletim dili İngiliz ve Amerikan İngilizcesidir.



Şekil 5. Wordreference Çok Dilli Sözlük Ana Sorgulama Sayfası.

(http://wordreference.com/)

Çok dilli çevrimiçi sözlük örneği olarak lexicool.com alan adlı site bir çevrimiçi sözlük ve dil araçları veri tabanı sunmaktadır. Ana sayfasında hedef ve amaç dilde konu alanlarına duyarlı bir açılır menü listesi verilmekte, listeden seçilen alanla ilgili sözlüklere bağlantılar yapılabilmektedir. Bu sözlük tipi ara kaynak olarak adlandırılabilecek niteliktedir. Ulaşılmak istenen sözlerle ilgili 6000 adet sözlüğün kaynak bilgilerini ve bağlantılarını içererek sözlük kullanıcısını sorguladığı konuyla ilgili popüler bağlantıya yönlendirmektedir. Ara sonuç sayfası bilgilerinde, bağlantısı verilen sözlüğün madde başı sayıları ve sözlüğün bağlantısının sağlam olup olmadığıyla ilgili bildirimlerin yapılabileceği ikonlar da bulunmaktadır. Bu sözlükte de listeli menüde yer alan konularla ilgili Türkçe sözlük kaynakları bilgisi yer almaktadır. Bağlantılara yorumlar eklenebildiği için bu sözlük kaynağı sistemi de etkileşimli bir yapıdadır.



Şekil 6. lexicool.com Sözlük Bilgisi Sorgulama Formu (http://www.lexicool.com/).

Çok dilli sözlüklerin hazırlanmasında önce son yıllarda "yetenekleri" artan iki dilli sözlükler de özellikle yabancı dil öğrencilerinin sıkça kullandıkları ineternetteki çevrimiçi sözlük yapılarıdır. İki dilli sözlüklerin bilgilerini içeren başlıca kaynaklardan biri dicts.info alan adında yayın yapan ve 60'tan fazla sözlüğe İngilizce temelli sorgulamaya olanak tanıyan sistemdir. Aralarında Türkçenin de bulunduğu bu sistemde sorgulanan sözcüğün İngilizce kavram alanı ve kavram alanına göre hedef dildeki karşılığı tablolu liste biçiminde sunulmaktadır. Sözcüğün sesletimi de diğer çevrimiçi çok dilli sözlük veri tabanlarında olduğu gibi dinlenebilmektedir. Bu sistem, sonuç sayfalarında çeşitli çevrimiçi dil kaynaklarına destek nitelikli olarak bağlantılar da içermektedir.



Şekil 7. dict.info.com İngilizce-Türkçe Sorgu Sonuç Sayfası. (http://www.dicts.info/ud.php?11=turkish&12=turkish)

İngilizcenin birçok alanda kullanılan yaygın bir dil olması, bu dili öğrenmek isteyen büyük bir kitlenin dil öğrenme gereksinimlerinin karşılanabilmesi amacıyla son yıllarda özellikle iki dilli çevrim içi sözlük yapılarının ve veri tabanlarının sürekli geliştirilmesini sağlamıştır. Yukarıda ekran görüntüleriyle verilen örneklerde de görüldüğü gibi, çevrimiçi sözlük sayfalarının tasarımı artık basılı sözlüklerin hazırlanmasında olduğu gibi uzmanlık gerektiren bir alan niteliğini kazanmıştır. Basılı malzemenin sayısal ortama aktarımı, internetin yaygınlık kazanmaya başladığı ilk yıllardaki gösterim nitelikli yapıları, internet teknolojilerinin gelişimiyle geride kalmıştır. Arayan ve sorgulayan her kullanıcının genel gereksinim düzeylerinin araştırılıp gerektiğinde istatistiksel gereksinin çözümlemelerinin de yapılması hazırlanacak veri tabanlarının işlevselliği açısından önem kazanmıştır. İnternetle birlikte farklı bir metin okuma sürecinin başladığı söylenebilir. Klasik okuma sürecinden farklı olarak elektronik metin ve bu metinle bağlantılı sözlük biçimleri, okuyucunun söz dağarcığının yetmediği durumları anında doldurabilen bununla kalmayıp paylaşılmış bir okuma ortamıyla birlikte kullanım sıklığı çok düşük sözcüklerin bile yaygınlaşması ortaya çıkabilmektedir.

Türkçe internet ortamında, son yılların önde gelen çevrim içi sözlük sitesi seslisözlük, İngilizce-Türkçe ve Türkçe-İngilizce sözcük çevirisine dayalı bir sistem olmakla birlikte Almanca, Yunanca, Fransızca ve Hollanda diline çeviriler de yapmaktadır. 1999'da açılan siteye son yıllarda kullanıcılarına gelişmiş sorgu seçeneklerini eklenmiştir. Türkçe karakter sorununun aşılabilmesi için küçük bir uygulamayla, yabancı klavye kullanan kullanıcı için sorgu kutusuna ç, ş, ı, ğ gibi karakterlerin girilebilmesi sağlanmıştır. Bu yaklaşım TDK Büyük Türkçe Sözlük için de hazırlanmıştır. Sözlüğün ana sayfasında günün sözcüğü uygulaması yer almakta, karşılığı ya da tanımı verilen sözcüğün tanım bilgileri de veri tabanıyla bağlantılı bir biçimde düzenlenmiş olup tanımda geçen herhangi bir sözcük tıklandığında, tıklanılan sözcük madde başı durumuna getirilerek sorgulanabilmektedir. Tanımda yer alan çekimli sözcüklerin tıklamayla madde başında sorgulanabilir duruma getirilmesinde güçlükler olduğu bilinmektedir. Bu sistemde de örneğin "sonuçlarını" gibi bir sözcük tıklandığında ara bir uygulama olarak çalışabilecek bir biçim birimsel çözümleyicinin bulunmadığı görülmektedir. Dolayısıyla sözlükte "sonuç" biçiminde madde başı olarak yer alacak bu sözcüğün eklerinin çözümü yapılamayıp "sonuç" sözcüğüne bağlantı kurulamamaktadır. Seslisözlük, sesletim uygulamasını da desteklemekle birlikte bunun için kullanıcıların sisteme üyelikleri gerekmektedir. Kullanıcının sözcük arama geçmişi

de çerezler olarak bilinen "cookiler" aracılığıyla tarayıcıdan alınan bilgiler doğrultusunda gerçekleştirilmektedir. Sözlüğün yeterliliklerin ve veriminin artırılması için kullanıcıyla iletişime dayalı uygulama da ana sayfada yer almaktadır. Bu da, bilişim uygulamalarında tasarımın ana konusu olan kullanıcı gereksinimlerinin göz önüne alınmasıyla, geliştirilmesi devam eden bir sistem olduğunu göstermektedir.

Son yıllarda metin yoğun web sayfalarında kullanıcının sorgulamasına geribildirim düzeneği olarak düşünülebilecek punto büyütme uygulamalarının geliştirildiği görülmektedir. Geniş bir web sitesinin binlerce alt sayfası ve buna bağlı büyük boyutlu bir veri tabanı bulunabilir. Metinsel bilginin bu kadar yoğun olması özellikle günlük haber sitelerinin arşiv sorgulamalarından çıkarılacak bilginin değerini artırmaktadır. Böylelikle aslında eğilim araştırması yönündeki çalışmaların bir ucu, kullanıcının söz sorgularına dayalı bir özel sözlük uygulamasının kendi içinde otomatik olarak gelişmesi yolunu da açmaktadır. Bu durum sözlük metni olmayıp çeşitli konularda yayın yapan düz metinli kaynaklar olabileceği gibi sadece sözlüksel bilgi tutan ve bu bilginin üzerinde uygulama yürüten siteler için de geçerlidir. Kullanıcı sorgusuna dayalı sözlük uygulaması diyebileceğimiz bu uygulamalarla; genel standart sözlüğe, çeşitli metin çözümleme ve uygulama yöntemlerinden elde edilen başta sıklık bilgisinin yer aldığı bilgiler yanında, web siteleri üzerinden geribildirime dayalı ve otomatik olarak ortaya çıkacak bilgiler de kaynak sağlamış olacaktır. Seslisözlük ana sayfasında en çok sorgulanan 12 sözcük bir bağlantılı metin kutusu içinde yer alırken bu sözcüklerin de en çok sorgulananları diğerlerinden daha büyük puntolarla gösterilmektedir. Bu yaklaşım Türkçe internet sözlük uygulamalarında yeni bir yaklaşım olup gelişen internet yazılım teknolojileriyle daha da çeşitlenmesi beklenebilir.



Şekil 8. Seslisözlük Ana sayfası (http://www.seslisozluk.com/).



Şekil 9. Seslisözlük Uygulamaları Örneği (http://www.seslisozluk.com/)

Haber kaynakları ve blog adı verilen "günlük" sitelerinde sıkça kullanılmaya başlanan RSS (Rich Site Summary) uygulamalarıyla en güncel bilgiler internet kullanıcılarına sağlanmaya başlamıştır. Zengin site özeti ya da içeriği olarak karşılık verilebilecek ("besleme" olarak da karşılığı yaygınlaşmaktadır.) RSS uygulamalarındaki bilgiler, XML diline dayalı .rss ya da .xml uzantılı dosya biçimlerinde tutulurlar. Bir RSS okuyucu ile RSS sağlayan siteye bağlanılabileceği gibi internet tarayıcıları da RSS kaynağı olan sitelerdeki uygulamaları otomatik olarak tanıyarak, sitedeki en güncel bilgilere abone olmayı sağlayabilmektedir. Seslisözlük de sözlüğe en son eklenen maddeleri RSS teknolojisiyle yayımlayabilmektedir. Böylece kullanıcı siteyi ziyaret etmese bile tarayıcısındaki özellikten yararlanarak en son eklenen madde başı ve tanımları izleyebilmektedir. Bu uygulamanın gerek genel gerek diğer sözlük biçimlerinde de başka siteler aracılığıyla desteklenip yaygınlaşacağı öngörülebilir.



Şekil 10. Seslisözlük RSS Uygulaması (http://www.seslisozluk.com/).

Türkçede tek dilli standart sözlüklerin başında yer alan TDK'nin hazırladığı ve 10. baskısı yapılan Türkçe Sözlük; internet üzerinde ilk önce Güncel Türkçe Sözlük (GTS) adıyla çevrimiçi olarak yayımlanmaya başlanmış, 2008 yılının sonlarından itibaren bu sözlük, Büyük Türkçe Sözlük (BTS) adlı projenin bir parçası olarak yayımlanmaya başlanmıştır. TDK'nin ana sayfasında, GTS için ayrı bir arama kutucuğu da sorgulama için kullanılmaya devam etmektedir.



Şekil 11. TDK GTS Sorgulama Arayüzü
(http://www.tdk.gov.tr/TR/Genel/Ana.aspx?F6E10F8892433CFFAAF6AA849816B2E F4376734BED947CDE)

GTS'de kullanılan teknoloji önceleri sunucu taraflı (server side) bir Microsoft firması teknolojisi olan asp (active server pages) olarak kullanılırken son yıllarda bu teknolojinin üst versiyonu olan aspx teknolojisi TDK'nin sunucularında bütün sözlükleri için kullandığı bir sunucu teknolojisi olmuştur. GTS'nin sorgu arayüzü ayrıntılı sorgulama olmaksızın tek bir arayüzden aranan sözcükle ilgili bütün ayrıntılar basılı sözlük bilgisine dayanarak sonuç üretmektedir. Sözlerle ilgili atasözü, deyimler ve birleşik yapılar da bağlantılı tablo biçiminde olup, sunucuya belirli bir tarihten itibaren söz sorgulaması için kaç kez ulaşıldığını da belirten alt sayfa bilgisi yer almaktadır. Sözlük kullanıcısıyla etkileşimin sağlanması ve sözlüğe yönelik katkı ve eleştiriler için e-posta uygulaması da hizmet vermektedir.

TDK'nin Kişi Adları Sözlüğü, Eş ve Yakın Anlamlı Kelimeler Sözlüğü, Türk Lehçeleri Sözlüğü, Yabancı Kelimeler Sözlüğü ve Bilim ve Sanat Terimleri Sözlüğü çevrimiçi olarak yayınlarını sürdürmektedir. Büyük Türkçe Sözlük (BTS), sözü edilen bu sözlükleri kapsayan ve bir çatı altında toplayan bir veri tabanı yapısına sahiptir.

BTS'de terim, ad, söz ve deyim olmak üzere toplam 616.767 birim bulunmakta, sorgulanan sözcükler bu birimler arasından hızlı biçimde ara sonuç sayfasında görüntülenmektedir.



Şekil 12. Büyük Türkçe Sözlük (BTS) Sorgu Formu (http://tdkterim.gov.tr/bts/).

BTS'nin ara sonuç sayfalarında sorgulanan sözcüğün veri tabanına bağlantısı yapacağı bir kimlik (id) numarasıyla sağlanmaktadır. Örneğin "bilgisayar" sözcüğünün sorgu sonucuna ulaştıracak ara sonuç sayfasının URL'si verilmekte, bu bağlantı bilgisinde sorgulanmak istenen sözcüğün madde başlarında tam olarak geçen birimlerden seçilmesi istenmektedir. http://tdkterim.gov.tr/bts/?kategori=veritbn&kelimesec=45190 adresi, "bilgisayar" sözcüğünün ara sonuç sayfasından tam sonuç sayfasına geçişini sağlamakta, sonda yer alan numara birimin veri tabanında kayıtlı kimlik numarasını göstermektedir. Bu numarayla birlikte veri tabanında bu birimin alt özellikleri olan metin bilgilerine, geçtiği kaynağa ait özellikler alt alta sıralanmaktadır.



Şekil 13. BTS Sonuç Gösterim Sayfası
(http://tdkterim.gov.tr/bts/?kategori=verilst&kelime=bilgisayar&ayn=tam)

TDK ve Dokuz Eylül Üniversitesi Dil Bilimi Bölümünün işbirliğiyle gerçekleştirilen Eş ve Yakın Anlamlı Kelimeler Sözlüğü'nde, sorgulanan sözcük bilgisine ulaşılması için POST form verisi gönderme metodu kullanılmıştır. Bu yöntemde sorgu formunda yer alan metin kutusuna ait değişkenin taşıdığı veri, adres çubuğunda görüntülenmeden veri tabanına gönderilmektedir. Sözlüğün ara sonuç sayfası olmadığından sonuçlar doğrudan görüntülenmekte, sözcüklerin birbirlerine gönderimi bulunmamaktadır.

TDK'nin Kişi Adları Sözlüğü'nde toplam 10842 ad yer almakta, sorgu sayfasında kız /erkek adı ayrı sorgulama ve benzer/aynı sorgu seçeneği sunularak gelişmiş sorgu formu biçiminde hizmet vermektedir. Ara sonuç sayfasının bulunduğu sözlük ada ve anlama göre sıralama veri tabanı sonuçlarını da içermektedir.

TDK ve Ahmet Yesevi Üniversitesi işbirliğiyle Kültür Bakanlığı'nın yayımladığı Türk Lehçeleri Sözlüğü'nün elektronik ortama aktarılmasıyla elde edilen veri tabanına dayalı Türk Lehçeleri Sözlüğü'nde; sorgu formunda lehçeler arası değiştirim seçenekleri verilmekte, sorgulanan sözcüğün örneğin Türkiye Türkçesindeki biçiminin yanında diğer lehçelerdeki karşılıkları tanım olmadan gösterilmektedir. Sorgulanan her sözcüğün Rusça karşılığı da diğer lehçelerin konuşulduğu ülkelerin bayraklarının küçük birer resmiyle sunulmaktadır.



Şekil 14. Türk Lehçeleri Sözlüğü Sorgu Formu ve Sonuç Tablosu (http://www.tdk.org.tr/lehceler/).

İnternet üzerinde standart sözlük biçimlerinin öncelikli olarak basılı biçimlerinin elektronik ortama aktarılmasıyla ortaya çıkan ve yukarıda anlatılan biçimlerinin yanında, standart sözlüklerin elektronik biçimleri gibi bağlantılı listeleri de içeren ve

“katılımcı sözlük” adıyla yayınlanmaya başlayan sözlüklerin Türkiye'deki ilk örneği 1999'da yayınlanmaya başlayan Ekşisözlük'tür. Bu tür sözlük yapılarında sözlüğün madde başı yapılarının altında yer alan ve teoride sonsuz olarak veri girilebilen alanlarına entry ya da girdi adı verilmektedir. Katılımcı sözlükler bir ya da birkaç yönetici etrafında kurulup sözlük yazarlarının belirli periyotlarla sisteme üyeliklerinin kabul edildiği bir şema sunarlar. Site yöneticileri sözlükle ilgili her türlü verinin akışını kontrol etme yetkisine sahip olup kimi hukuksal sorunlar yaratabilecek girdiler için sözlüğün yayın hayatını tehdit edebilecek durumlarda sorumlu olabilmektedirler.



Şekil 15. Ekşisözlük Ana Sayfası. (http://sozluk.sourtimes.org/)

Bu sözlüklerin yapılarına bakıldığında sözlükte yer alacak madde başları sol sütunda giriş zamanına göre sıralanmakta madde başlarının giriş verileri de sağ geniş ekranda gösterilmektedir. Hemen bütün katılımcı sözlüklerde gelişmiş bir arama yapısı da bulunmaktadır. Yapısal olarak kullanıcılar okuyucu olarak giriş yapabilmekte, ayrıca yazarlara özgü daha ayrıntılı sorgulama yapıları da içermektedirler. Bu tür sözlüklerin madde başı konularının oluşumunda, gündemde yer alan olay ya da kişilerin etkisinin olduğu görülmektedir. Türkiye'de en çok kullanıcıya sahip Ekşiszölük'e günlük madde başı girişi ortalama 2000 civarındadır. Sözlüğün içerik olarak hemen her tür konuda, sözcük düzeyinden cümle yapısına kadar madde başlarının oluştuğu görülmektedir. Aşağıda bu sözlükten madde başlarından örnekler yer almaktadır:

bilgisayar:

1. bildigim kadariyla altan erbulak'ın (yanlış hatırlıyor olabilirim) bulduğu sözcük. computer yani, gayet güzel bir sözcüktür. hayir efendim, bundan gayet emin olan distantin soyledigi kadari ile aydın köksal bulmu$ bu sozcugu. (guru, 22.06.1999 ~ 01.02.2005 20:08)

3. walla iyi güsel bi makina ama i$te ne yazikki duydulari ve ce$itli vurgulari ya$atamiyosun

kar$indakine.(elmyra, 06.07.1999)

4. iki kere ikinin dort ettigini aninda bulan alet (emirhan, 19.07.1999)

5. akil hastasi olmama neden olacak alet (ahmet corleone, 11.02.2000 16:28)

sözlük:

1. eksi olmayanlari da en az olani kadar ise yarayan kitap bicimli sey. ayni ya da farkli dillerdeki anlami bilinmeyen kelime, kelimeler, kaliplar vesairenin anlamini barindirir icinde. (bkz: gozluk) ((yok, 28.03.1999 ~ 27.03.2001 21:28))

2. bi dildeki kelimelerin baska bi dildeki karsiliklarini, ya da ayni dilde tanimlarini alfabetik sirayla iceren eser, basvuru kaynagi, basucu yapiti. (phobios, 23.01.2000 03:15)

11. eksi sozluk lafının yerine kullandığımız kelime. (multiple, 15.03.2002 15:29)

Yukarıda yer alan iki madde başında (sözlükte başlık olarak adlandırılmaktadır) tanım nitelikli girişlerin yanında espri nitelikli girişler de görülmektedir. Katılımcı sözlüklerde yazarlar gerçek adları yerine nick ya da nickname olarak belirtilen takma ad kullanırlar, böylece konu hakkında daha serbest, kimliklerinin ortaya çıkma kaygısından uzak biçimde girişlerini, düşüncelerini, tanımlarını yazabilmektedirler.

Madde başlarının seçiminde Türkçenin yanında İngilizce yazılanlar da yer alır. Standart, klasik sözlük yapılarından katılımcı sözlük biçimlerini ayıran en önemli özellik teknolojik olarak yapı özelliğinden çok madde başı seçimi ve biçem (üslup) farkıdır. Standart sözlüklerde kullanılan ölçünlü dil katılımcı sözlüklerde yer yer bırakılmakta bunun yanında ölçünlü dilde yazılan bir metin tanımı desteklemek amacıyla olduğu biçimiyle kullanılabilmektedir. Sözlük'te büyük harflerin ve Türkçe ç, ı, i, ö, ğ, ü, ş, İ karakterlerinin kullanılmaması dikkati çekmektedir. Konuşma diline dayalı ifadeler de sözlükte yer yer bulunmaktadır. Aşağıda örneği verilen sözlük alıntısında Türkçe karakterlerin yer almadığı görülmektedir.

"bilgisayar gunumuz burolarinda kullanilan, en cefakar ve de maassiz calisan, elektrikle beslenen, sendikasiz ama sigortali eleman (tharaxi, 05.07.2001 14:52)."

"sozlugun bolcana ihtiva ettigini dusundugum bir madde (madarc, 26.06.2000 05.38) (Yıldırım ve Tahiroğlu, 2006: 372-373).

Türkçe yayınlanan katılımcı sözlükler arasında Ekşisözlük dışında itüsözlük, uludağsözlük, 3harf, zamanesözlük gibi bu tür sözlüklerin konu olarak "genel" nitelikli olanlarının yanında, futbol kulüplerinin taraftarlarının hazırladığı 1903sözlük, cimbomsözlük, fenersözlük gibi özel konular için hazırlanan katılımcı sözlükler de son yıllarda yayınlanmaya başlamıştır. Bu sözlükler arasında itüsözlük metin verisiyle birlikte görsel ögelere de yer vermekte, aranan madde başıyla ilgili görsel (resim verisi) malzeme görüntülenebilmektedir. Sözlüklerin yukarıda yapı özelliği olarak anlatılan Ekşisözlük formatına dayandığı görülmekle birlikte içerik farkları bulunmaktadır.

İnternet metin yoğun bir ortam sunmasıyla açık, yapılandırılmamış bir derlem olarak kabul edilebilir. Derlemlerin özellikle dil bilgisel çözümlemede giderek önemini artırması; sadece internetin bile başlıca bir veri tabanı olarak düşünülmesini sağlayarak sözcüklerden yola çıkılarak kullanıcı eğilimlerinin ölçülmesini ve değerlendirilmesini sağlamada çok önemli bir duruma gelmesini sağlamıştır. Ticari olarak müşterilerinin nelerle ilgilendiğini, hangi ürünü tercih ettiklerinin ortaya çıkarılmasını amaçlayan şirketlerin yanı sıra, özel olarak sözlük bilimi araştırmacılarının bu eğilim çözümlemelerinden elde edilen bilgiyi de kullanmaları sözlüklerin günceli yakalaması açısından önemli görünmektedir.

Hemen bütün dünya dilleri için kendi içinde bir altyapı sağlayan Google firması, barındırdığı çok büyük boyutlu veri ve depolama ortamıyla kullanıcıların arama eğilimlerini yayınladığı sayfalarla güncel biçimde bu bilgileri paylaşmaktadır. Google Trends adıyla paylaşılan sayfalar son bir yılda Google Insights for Search olarak daha ayrıntılı ve görsel sonuçlara yer vermektedir. Google arama motoru için metin verileri resimli ya da ortam (medya) verisinden daha değerlidir. Çünkü Google'ın web sayfalarından veri toplayan robotları, site içinde yer alan her bir bağlantıyı ve ardındaki veriyi çeşitli kategorilerde bölümlemekte ve ana veri tabanlarına bu veriye dayalı sonuçları kaydetmektedir. Sayfalarda yer alan resimlerdeki nesnelerin tanınması için henüz çok başarılı uygulamalar bulunmamakta, bu haliyle internet metin ve sözcük ağırlıklı olarak büyümeye devam etmektedir. Google yıllık olarak en çok aranan sözcükleri ve sözcük gruplarına dayalı sonuçları Google Zeitgeist sayfalarıyla duyurmaktadır. Zeitgeist Almanca "zamanın ruhu" anlamına gelen bir sözcük olup, yılın her günü derlenen internet verisinden ortaya çıkan ülkeye dayalı internet araması sonuçları olarak Google'la birlikte elektronik ortam sözcüğü durumuna gelmiştir.



Şekil 16. Google Zeitgeist Ana sayfası
(http://www.google.com/intl/en/press/zeitgeist2008/)

Yukarıdaki şekilde görüldüğü gibi 2008 yılına ait sonuçlar içinde dünya ölçeğinde arama sonuçlarında ilk 10 sıralanmıştır. Kullanıcıların ilgisini çekecek bir sınıflamanın olduğu, sayfa solunda yer alan spor, gösteri dünyası ve politik bölümlerinden anlaşılmaktadır. Sonuçlar arasında Türkiye'den elde edilen sonuçlar yer almamakta, sorgular içinde geçen sözcüklerde özellikle nasıl, kim ve nedir sorularına dayalı bilgiler ve kalıplara da yer verilmektedir.



Şekil 17. Google Zeitgeist Sorgu Kalıpları Örnekleri
(http://www.google.com/intl/en/press/zeitgeist2008/mind.html)

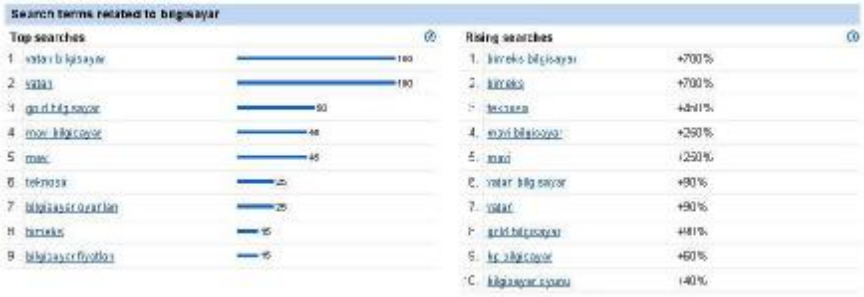
Türkçe arama sonuçlarının grafikle gösterimini de sunan Google Insight fo Search 2004 yılından günümüze tüm arama kalıplarını barındırmaktadır. Şu anda üç seçenekli arama formunda terim, yer ve zaman dilimlerine göre sorgulama yapılmakta, sonuçlar harita destekli olarak görüntülenebilmektedir. Sonuç sayfasında, sorgulanan sözcüğün tüm sorgulamalar içinde yükselen eğiliminin yüzdesi, tüm dünya ölçeğinde hangi ülkelerden sözcükle ilgili sorgulama yapıldığı, bu sözcüğün haber kaynaklarındaki eğilim grafiği verilmektedir.



Şekil 18. Google Insight for Search Sorgu Formu
(http://www.google.com/insights/search/#)



Şekil 19. Google Insight for Search - "bilgisayar" Sözcüğünün Dünya Ölçeğinde Dağılımı
(http://www.google.com/insights/search/#q=bilgisayar&cmpt=q)



Şekil 20. Google Insight for Search'de "bilgisayar" Sözcüğü ve Yükselen Arama Sonuçları (http://www.google.com/insights/search/#q=bilgisayar&cmpt=q)

Yukarıdaki şekillerden de görüleceği gibi tek bir sözcükle internet üzerinden kullanıcıların arama eğilimleri ölçülebilmektedir. Sorgu formuna 2'nin üzerinde sözcükle de sorgu yapılamasını sağlayan ek giriş kutuları bulunmakta, böylelikle aralarında sorgulamayı yapan kullanıcının sezgisine göre ilişkisi olan birimlere ait sonuçlar da görülebilmektedir.

Google gibi dünya çapında hemen her kullanıcının bilebildiği, kurumların yanısıra *veri madenciliği* (data mining) üzerinde özelleşmiş kurumların da kendi veri tabanlarına dayanan uygulamaları bulunmaktadır. Bu kurumlar özellikle ticari firmalara bilgiler sağlamakta, yayınladıkları web sayfalarında metin işleme ve çözümlemeye yönelik paket programlarını sunmaktadırlar.

Tüm bu uygulamalardan sözlük bilimi ve otomatik metin çözümlemeyle uğraşan araştırmacılar da yararlanabilmektedirler. Sözlük bilimi ve internetin bütünleşik kullanımı yukarıda verilen örneklerden de görülebileceği gibi gerek yeni sözlüksel birimlerin belirlenmesinde gerekse sözlük yayıncılığının görsel gelişiminde önümüzdeki yıllarda köklü değişimlerin beklenebileceğini ortaya koymaktadır.

Google'ın yukarıda verilen eğilim çözümleme yöntemiyle birlikte özellikle internette yayınlanan çevrimiçi metinlerin çözümlenip sözcüklere dayalı bir eğilim yoklamasının yapılmasına olanak tanıyan The Global Language Monitor (GLM) adlı sistem, İngilizce yayın yapan yazılı ve görsel basını izlemektedir. Sistemi kuran kuruluşta dil bilimciler ve sözlükbilimciler birlikte çalışarak İngilizce için 2006 yılından başlayarak tüm yıl boyunca yaşanan gelişmelere bağlı olarak ortaya çıkan "yılın sözünü" ve dile yeni giren sözcüklerin ortaya çıkarılması sağlanmaktadır. 2003'te Silikon Vadisi'nde kurulan sistemi, yine internet üzerinde yayın yapan www.yourdictionary.com'u kuran ekip desteklemiş ve ilk olarak amatörce sözcük sayımları yapan Paul J.J. Payack bu sistemden sorumlu başkan olarak ilk "yılın sözü" çalışmasını 2006'da yayınlamıştır. Sistem dünyada belli başlı basım evleri ve çevrimiçi yayıncılık yapan kuruluşlarla bağlantı içinde bulunmaktadır. GLM çevrimiçi yayınlanan gazete ve dergilerin yanında, akademik nitelikli Lexis-Nexis gibi veri tabanlarını da kullanmaktadır.

Tahmini Nitelikler Göstergesi (Predictive Quantities Indicator-PQI) olarak çevrilebilecek bir çözümleme algoritması kullanan sistem, sözcük sıklıklarını temel almakla birlikte kullanılan algoritmanın ayrıntıları açıklanmamaktadır. Veriler politik, iş dünyası ve güncel konular olmak üzere üç başlık altında ele alınmakta her bir başlığa yönelik algoritma ve çözümleme yöntemi geliştirilmektedir. Bu algoritmalara

dayanılarak yapılan çözümlemede 2007 yılının ilk 10 sözcüğü içinde birinci sırada "hybrid" sözcüğü, yine aynı yılın eşdizimli diyebileceğimiz ifadesi ise "climate change" olarak belirlenmiştir. GLM'nin kullandığı yöntemin güvenilir olmadığına işaret eden Geofrey Nunberg ve Jesse Sheid Lower gibi araştırmacılar, sisteme girecek sözcüklerin ölçütlerinin genel olarak kesinleşmiş olmadığını, eldeki derlemlerle karşılaştırılmalı bir ölçütün olanaklı olmadığını savunarak sistemin sonuçlarının güvenirliği konusundaki kuşkularını dile getirmişlerdir. Kullanılan algoritmanın özel olarak saklanması, olası yeni sözcük birimlerin nasıl belirlendiği konusunu da tartışmalı bırakmaktadır. GLM, derlenen veriler üzerinde PQI algoritmasını dünya çapında dil dizini, ağırlıklı dizin, uzun dönemli ve kısa dönemli faktör çözümlemesi, devinim (momentum) ve vektörel hız adı verilen tekniklerle uygulamaktadır.[4]



Şekil 21. The Golobal Language Monitor (GLM) Ana sayfası (http://www.languagemonitor.com/)

[4] http://en.wikipedia.org/wiki/Global_Language_Monitor

1.4. Doğal Dil İşleme, Yöntem ve Uygulama Alanları

1.4.1. Doğal Dil İşleme ve Bilgisayarlı Dil Bilimi Terimi

Doğal Dil İşleme (DDİ) (Natural language processing-NLP) terimi son yıllarda özellikle bilgisayar bilimi ve mühendisliği alanında Türkçe karşılık olarak kullanılmaktadır. Bilgisayar biliminde insan-makine etkileşiminin sağlanmasında insan dilini makinenin en iyi biçimde "yorumlayabilmesi" için ortaya konan çalışmaların tümü -*Bilgi çıkarımı* (information extraction), *bilgi süzme* (informatin retrieval) bu çalışmalar içinde yer alır- DDİ'nin çalışma alanına girmektedir.

Baker vd. (2006:121) DDİ'yi şu biçimde tanımlamaktadırlar: "Bilgisayarlı dil biliminin ortaya koyduğu sorunlar için kullanılan genel terim. Bilgisayarları dilin kimi yönlerini ele almak için araç olarak kullanmak, insanın başka türlü yapmak zorunda kalacağı yüklü işlemleri otomatik olarak yaptırmak gibi sorunlar bilgisayarlı dil biliminin konusu içinde yer alır. *Metin hizalama* (alignment)[5], *söz türü etiketleme* (part of speech tagging), *varlık adı tanıma* (named entity recognition)[6] ve makineli çeviri sistemleri DDİ uygulamalarının örneklerindendir. DDİ'de birçok modern yaklaşım bir derlem kullanmayı gerektirir. Bir derlemin otomatik çözümlemesi, DDİ'de kullanılacak uygulamalara taban olacak biçimde, o dil hakkındaki bilginin yapılandırılması anlamında gelir."

DDİ'yi Oflazer , doğal dillerin yapısını çözmek, yorumlamak, yeniden üretmek amacıyla bilgisayar sistemlerinin geliştirildiği mühendislik alanı olarak ele almaktadır. 1950 ve 1960'lı yıllarda yapay zekanın bir alt dalı olarak ele alınırken, gelişen teknik ve donanım yapılarıyla birlikte bugün başlı başına bir bilim disiplini durumuna gelmiştir. DDİ, bilişsel psikolojiden, *bilgi gösterimi*ne (knowledge representation), kuramsal dil biliminden biçimsel diller kuramına kadar uzanan geniş bir alanın içine aldığı disiplinler arası özellik taşır (Oflazer, 2006: 814).

Bilgisayarlı dil bilimi (computational linguistics) ve doğal dil işleme terimlerinin eşanlamlı olarak kullanılabileceğini söylemek güçtür. Son yıllarda bu iki terim arasındaki anlam farkı belirginleşmektedir. Doğal dil işleme, dil bilimsel çözümlemenin daha çok makinece anlaşılabilir, bilgisayar biliminin mühendislik yaklaşımı içinde yer alırken, bilgisayarlı dil bilimi dil biliminin sınırları içinde ele alınmaya başlanmıştır.

[5] Metin hizalama, makineli çeviri sistemlerinde kullanılan paralel derlem adı verilen derlem türünde, aynı addaki farklı dillerdeki metinlerin satırlarının eşleştirilmesidir.
[6] Özel adların otomatik tanınması çalışmaları için kullanılan terimdir.

Bilgisayarlı dil bilimi, dil üzerine yazılımların kullanıldığı çalışmaları kapsayan dil biliminin alt çalışma alanıdır. Bu alan; yapay zeka, bilgisayar bilimi ve dil kuramını içeren disiplinler arası bir yaklaşımı gerektirmektedir. ***Bilgi tabanlı*** (data-driven) ya da ***veri tabanlı*** (information-driven) yöntemler ağırlıklı olarak kullanılırken, özellikle hazırlanmış bir derlemin veri tabanlı yaklaşımda kullanılması ağırlık kazanmıştır. Bilgisayarlı dil bilimciler, makineli çeviri, bilgi süzme, ***söz tanıma*** (speech recognition), ***yanıtlama sistemleri*** (voice response systems), internet arama motorları, ***otomatik içerik çözümleme*** (automated content analysis) sistemlerinin oluşturulmasında görev alırlar (Baker vd., 2006:41).

Dil bilimi sözcüğüne eklenen "bilgisayarlı" sözcüğü yıllardır bu alanda yürütülen çalışmaların çeşitlenmesi sonucunda kimi karışıklıklara, değişik yorumlara neden olmuştur. Gazdar ve Mellish (1989), doğal dil işleme terimini bilgisayarca okunur metinlerden (computer-readable texts) dizinler çıkarma ya da ***bağlamlı dizin***ler (concordance) elde etme çalışmaları alanı olarak dar kapsamda ele almazlar. Grishman (1986), bilgisayarlı dil bilimini doğal dili anlamak ya da üretmek amacıyla bilgisayar sistemleri üzerine çalışılan bir alan olarak görür. Bennet vd. dil bilimi ve bilgisayar bilimi arasında yapılan "evlilik" sonucu oluşmuş bir alan olarak bilgisayarlı dil bilimi sözünü kullanırlar. Bu noktada bir başka bakış açısı da Karlgen'den gelmiştir. Karlgen, bilgisayarlı dil bilimini dil bilim bakış açısıyla bilgisayarı kullanmak olarak ele alırken doğal dil işlemeye bilgisayar bilimi açısından yaklaşır. Doğal dil işleme arayüzleri hazırlamak ya da diğer sistemlerle ilişkisini sağlama çalışmaları ileri düzey mühendislik gerektirirken, bilgisayarlı dil biliminde, cümle çözümleme (parsing) veya dil bilgisi kurallarını sınama gibi kimi işlemlerle daha çok ilgilenilmektedir. Son olarak Ramsay, bilgisayarlı dil bilimini ***bilişsel bilim*** (cognitive science) alanında görürken, doğal dil işlemeyi mühendislik alanı olarak ele alır (Ooi, 1998: 24).

Bilgisayarlı dil bilimi terimi için Türkçe bilim yazınında kimi çalışmalarda ***hesaplamalı dil bilim*** karşılığının kullanıldığı görülür. Altan ve Orhan[7] "Anlam Belirsizliği İçeren Türkçe Sözcüklerin Hesaplamalı Dilbilim Uygulamalarıyla Belirginleştirmesi" başlıklı makalelerinde, hem başlıkta hem de anahtar sözcüklerde ***hesaplamalı*** sözcüğünü kullanmışlardır. Genel olarak Türkçe bilim yazınının bu alanda oluşmadığını, terimlerin standartlaşmadığını görmekteyiz.

[7] http://turkoloji.cu.edu.tr/DILBILIM/anlam_belirsizligi.pdf

Yukarıda sözü edilen bilgisayarlı dil bilimi terimi yerine, son yıllarda dil biliminde kullanılan klasik araştırma ve derleme yöntemlerine yardımcı teknikler geliştirmek ve alan çalışmalarının hızını artırmak amacıyla yazılım projeleri hazırlamak için yardımcı çalışma dalı olarak *bilgisayar destekli dil bilimi* teriminin de kullanıldığı görülür. Bu çalışmanın başlığının seçilmesindeki ana etken, bilgisayar sözlükçülüğü ya da bilgisayarlı sözlük bilimi araştırma ve uygulama alanının, bilgisayar destekli sözlük bilimi çalışmalarından temelde farklı olmalarıdır. Bunun yanında, DDİ ile ilişkisi bakımından disiplinler arası niteliğini koruduğu söylenebilir.

1.4.2. Doğal Dil İşlemede Yöntemler

DDİ'de son yıllarda istatistiksel yöntemler ağırlık kazanmakla birlikte, doğal dilin yapısal düzenliliklerinin otomatik çözümlenmesinde kullanılan yöntemler uzunca bir süre dil bilgisine dayanan kural tabanlı yöntemler olmuştur.

1.4.2.1. Kural Tabanlı Yöntem

Doğal dil konuşucusunun bir dil edincini otomatikleşmiş olarak üretme süreçleri birçok karmaşık yapının ürünüdür. Biyolojik olarak beynin dil için özelleşmiş bölümleri, bebeğin 3-4 yaş civarında artık somut kavramları anlatabilme ve daha sonraki yaşlarda da soyut kavramlara dayalı adlandırma, cümle kurma ve söyleme yönelik davranışlarını belirleme yeteneğinin hızlıca gelişmesini sağlamaktadır. Bir insan için beyinde oluşan çok karmaşık işlemlerin sonucu gerçekleşen dil yetisi elbette makineler için başta hiçbir anlam ifade etmeyen bir yapı görünümündedir. Bir bilgisayar sisteminin doğal dili "anlayabilmesi", her şeyden önce bütün görünümlü doğal dilin parçalara bölünerek çeşitli gösterimlerle (makine diliyle) kodlanmasını ve buna dayanarak girdilerin "yorumlanmasını" gerektirmektedir.

Bir dil bilgisi, Vardar ve arkadaşlarına göre (1998: 76) dört farklı düzeyi içine alan tanımlamayı içerir: "1. Bir dilin işleyişini ve sunduğu düzeni ortaya koyan, özellikle de biçimbilimle söz dizimini kapsayan inceleme. 2. Dilsel kullanımım kimi yönlerini kurala bağlamayı amaçlayan buyurucu ve kuralcı inceleme. 3. Üretici-dönüşümsel anlayışta, bir dilin konuşucu-dinleyicilerince geçerli sayılan cümleleri üretebilecek bir düzenek oluşturmak üzere dil bilimcinin kurduğu biçimsel dizge. 4. Konuşucu-dinleyicinin cümleleri üretmesini ve anlamasını sağlayan iç dizge ve bilgi;

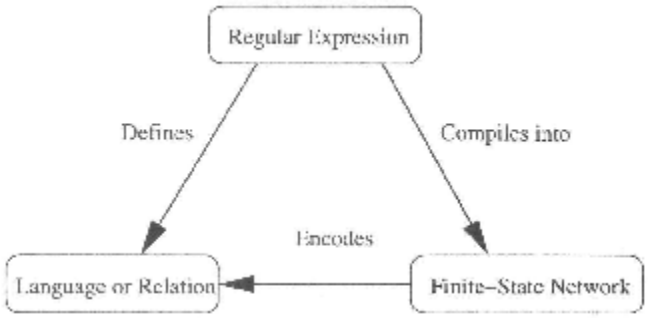
edinç". Dil bilgisi bir dili inceleme yaklaşımlarının adı olabileceği gibi, 4. tanımda da görüldüğü üzere bir iç yapının yani insanın dil yetisinin adı da olabilmektedir.

İnsanın dil yetisinin bir dil bilgisi etrafında ele alınması, dil bilgisel yapıların, seslerin dizgesinden başlayarak daha büyük birimlere doğru kavramsallaştırılarak sınırlarının belirlenmesi ve bu yapılırken "kurallı" yapıların ortaya konması, bir dilin makinece işlenir duruma getirilmesi açısından son derece önemlidir. Dil bilgisel kural Vardar ve arkadaşlarına göre (1998: 146), geleneksel dil çalışmalarının doğru-yanlış anlayışına göre yazımda, söyleyişte, biçim biliminde tek doğru biçim etrafında şekillenirken kural, "çağdaş dilbilimde genel bir kuram çerçevesinde oluşturulan varsayım; özellikle üretici-dönüşümsel dil bilgisinde, biçimsel bir dizge içinde doğru oluşturulmuş simge birleşimlerini belirten ilke." olarak görülmektedir. Doğal dil işleme açısından, istatistiksel ve olasılıksal yaklaşımlar dışında, bir metnin girdi olarak alınıp bu metne dayalı anlamlı sonuçların üretilmesinin beklendiği durumlarda kurallardan yola çıkılarak yapılan işlemler, ele alınan dilin başta sözcük ve sözcük birliktelikleri olmak üzere oluşturduğu cümle görünümlerinin ilkelerine dayanmaktadır. Dolayısıyla, DDİ bakış açısından bir dil bilgisel kural, hedef çıktının üretilebilmesi için tıpkı bir programla dilinde uyulması gereken prosedürler gibi, bir model olarak çıkarılması gereken, olabilecek bütün durumların bir gösterim dili içinde kodlaması yapılabilen, hem doğal dilin özelliklerini taşıyan hem de makine diliyle ifade edilen iki taraflı bir bütün olarak ele alınabilir.

DDİ'de kural tabanlı işlemler öncelikle dilsel düzenliliklerin bir veri yapısı olarak görülmesi anlamına gelmektedir. Biçimbilimsel ve sözdizimsel çözümlemelerde en sık kullanılan yapılar ***sonlu durum makineleri*** (Finite-state machine-FSA) olarak bilinen ve kendine özgü doğrusal bir grafla gösterilen yapılardır. Coleman (2005: 112), sonlu durum makinelerini (SDM), bir soyut hesaplama aracı olarak tanımlamakta ve konuşma dilini işlemede en önemli teknik olarak görmektedir.

Bilgisayar biliminde; ***düzenli ifadeler*** (regular expressions), ***formel diller*** (regular language), ***sembol*** (symbol), ***alfabe*** (alphabet), ***ilişki*** (relation), ***sözcük*** (word) ve ***dil*** (language) terimlerinin doğal dil için kullanılan terimlerden farklı anlamları bulunmaktadır. Beesley'e (2003:70-73) göre sembol a ve b gibi yazılabilir karakterler olabileceği gibi, bazı uygulamalarda isim (noun) ve çokluk (plural) kategorileri için atanmış soyut semboller de olabilir. Xerox'un uygulamalarında bu tür soyut semboller diğer a ve b gibi karakterlerden ayrılması için bir "+" işaretiyle atanmaktadırlar. Ayrıca + işaretinin bir diğer işlevi de, birleştirme için kullanılan "." operatöründen bu yapıları

ayırmaktır. Alfabe, sembollerden oluşan settir. Sözcük, bir alfabede yer alan sembollerin birleşmeleriyle oluşmuş karakter dizileridir (string). “bilgi”, “veri”, “jdklfırıdls” birer sözcük olup “” tırnak içinde yer alan boşluk da bir boş karakter (empty string) olarak doğal dildeki sözcükten farklı biçimde alfabede yer aldığı için “sözcük” kabul edilir. Dil ise bir sözcük setidir. Bir dil son durumlu yani başlangıcı ve bitişi önceden verili, belirli sayıda sözcükten oluşmuş bir yapı olabileceği gibi hiçbir sözcük de içermeyebilir. Formel diller işte bu sözcük setlerini içeren dillerdir. Düzenli diller terimi de, ***birbirine bağlama*** (concotenation), ***birleştirme*** (union) ve ***yineleme*** (iteration) ile tanımlanmış olası bütün formel dil alt setlerini anlatmak için kullanılan bir terimdir. Düzenli ifade, bir düzenli dili tanımlamada kullanılan ve SDM’ye çevrilen kendi içinde çeşitli karakterlerden ve bu karakterlerin bir araya gelerek bir sonlu durumu ifade eden yapılardır. Aşağıdaki şekilde Beesley, düzenli ifadeler, dil, ilişki ve SDM’yi gösteren yapıyı bir arada vermiştir.



Şekil 22. Düzenli İfadeler, Dil, İlişki ve SDM Ağı İlişkisi (Beesley, 2003: 73).

SDM’ler için kullanılan terimlerden bir diğeri ***ilişki*** (relation) terimidir. İlişki Karttunen’e göre (2003: 340), {<a,bb>,<cd,ε>} olarak gösterilen karakter çifleri arasında gerçekleşir. Bu ilişkide ilişki çiftinin ilk üyesi ***üst karakter*** (upper string) ikinci üyeye de ***alt karakter*** (lower string), karakter-karakter ilişkisinin ortaya çıkardığı modele de sırayla ***üst dil*** (upper language) ve ***alt dil*** (lower language) adı verilir. Yine, üst dil ***alan*** (domain), alt dil ise ilişkinin ***kapsam*** (range) bölümlerini oluşturur. {<a,bb>,<cd,ε>} örneğinde bir ilişki söz konusudur ve ilişkinin üst dili {a,cd}, alt dili {bb, ε} çiftlerinden oluşur.

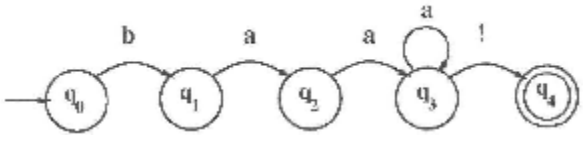
Beesley, ilişki kavramı için sözcük çiftleri (pairs of words) terimini kullanmış, örnek olarak yukarıda verilen soyut örnek yerine doğal dilden sözcükler içeren çiftler sunmuştur. Bu çiftler İngilizce sözcükleri içeren <”cat”, “cats”>, <”child”, “children”>, <”dog”, “dogs”> sözcükleridir. Beesley, bu çiflerin bir tür soyut sözlük oluşturduğunu,

çift üyelerden herhangi birine bir bakma (looking up) işlemi gerçekleştirildiğinde diğerine ulaşılabileceğini belirtmiştir. Bu noktada ilişkinin bakılan yönüne göre arama işlemi değişen adlar almaktadır. Örneğin yukarıda verilen "dog" sözcüğü ele alındığında ilişkinin sağından bir sözcük seçildiğinde "dogs" sözcüğüne ulaşılırsa bu tür aramaya bir ***bakma*** (looking up), "dogs" sözcüğü işleme alındığında "dog"a ulaşıldığında da bir ***geriye bakma*** (looking down) araması yapılmış olur. Beesley bu tür geriye bakmalaraın ***üretme*** (generating) olduğunu belirtmiştir. Xerox'un biçim bilimsel çözümleyicisinde milyonlarla ifade edilen ilişkili çift bulunmaktadır. Bu ilişkiler aşağıda İspanyolcadan verilen "canto" sözcüğünün sözcük biçimleri için hazırlanmış ve biçimbilgisel işaretlerin de (+verb, +PresInd, +1p, +Sg) bulunduğu tarzda ayrıntılandırılmaktadır (Beesley, 2003: 71).

```
<"cantar+Verb+PresInd+1P+Sg","canto">
<"cantar+Verb+PresInd+2P+Sg","cantas">
<"cantar+Verb+PresInd+3P+Sg","canta">
<"cantar+Verb+PresInd+1P+Pl","cantamos">
<"cantar+Verb+PresInd+2P+Pl","cantais">
<"cantar+Verb+PresInd+3P+Pl","cantan">
<"canto+Noun+Masc+Sg","canto">
<"canto+Noun+Masc+Pl","cantos">
<"carta+Noun+Fem+Sg","carta">
<"carta+Noun+Fem+Pl","cartas">
```

Şekil 23. İspanyolca cantar, canto ve carta Sözcükleri İçin İlişki Modellemesi (Beesley, 2003: 72).

SDM'ler basit bir karakter dizisinden oluşan sözcüklerin otomatik tanınmasında kullanılabileceği gibi karmaşık söz dizimsel çözümlemelerin otomatik işaretlenmesinde ve yazım hataların düzeltilmesi uygulamalarında da kullanılmaktadır. Basit bir SDM bir ***başlangıç durumu*** (start state) ile bir ***bitiş durumu*** (final state) arasındaki geçiş ağlarından oluşur. Aşağıdaki şekilde, Jurafsky ve Martin İngilizce meleme yansıma sözcüğünün SDM'sini göstermişlerdir.



Şekil 24. SDM Örneği (Jurafsky ve Martin, 2000: 34)

Bu şekilde q_0 ile gösterilen daire başlangıç durumunu, durumlar arasındaki yay biçimli oklar ise ***geçiş***leri (transition) göstermektedir. Makine, verilen karakter dizisine

uyan her bir karakterde bir sonraki duruma geçiş yapmakta, uygun olan her bir durum sonrasında bitiş durumuyla işlemini tamamlamaktadır. Bitiş durumları çift yuvarlaklı daire ile gösterilmektedir. Ayrıca her bir durum q_0, q_1, q_2 vb. gibi numaralandırılmaktadır. Bu örnekte "ba!" karakter dizisiyle başlayan makine q_3 durumunda a karakterini tekrarlayarak "baa!, baaa!, baaaa!, baaaaa!" olarak devam etmektedir ve burada alfabe a, b, ! sembollerini içermektedir. Bu da Σ = {a,b,!} olarak gösterilmektedir. Bu alfabenin oluşturacağı dil ise m model kısaltmasında L (m) = {baa!, baaa!, baaaa!...} biçiminde verilmiştir (Jurafsky ve Martin, 2000:34).

SDM'ler basit bir makine düzeyinde olsalar bile doğrusal bir düzlem izleyerek bir *ağ* (network) oluştururlar. Bu ağın parçaları durumlar, geçişlerdir. Ayrıca yukarıda geçişler için gösterilen yay biçimli kavislerin izlediği sıraya Karttunen *yol* (path) adının verildiğini belirterek, bir kavisin tek bir sembolle örneğin a sembolüyle ya da bir sembol çiftiyle örneğin a:b etiketlenebileceğini vurgular. a:b çiftinde kavisin sol tarafında yer alan a sembolü üst, sağ tarafında yer alan b ise alt sembol olarak adlandırılır. Bir SDM'de bütün kavisler tek bir sembolle etiketlenirse buna ***basit otomaton*** (simple automaton), en az bir sembol çiftiyle etiketlenirse yani kaviste bir çift bulunursa çiftin bulunduğu geçiş ağı ***dönüştürücü*** (transducer) adını almaktadır (Karttunen, 2003: 340).

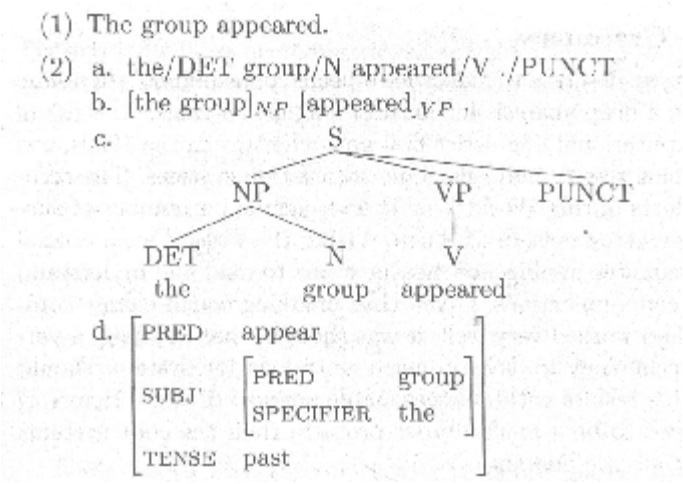
SDM gösterimi için özel semboller kullanılmaktadır. Roark ve Sproat (2007: 4), bir modeli M (aslında bir SMD'dir) ile vererek modelin barındırdığı yapıyı şu şekilde göstermişlerdir: M = (Q, s, F, Σ, δ). Burada Q durumların sonlu seti, s başlangıç durumu, F bitiş durumlarının seti, Σ alfabe, δ ise geçiş ilişkilerini göstermektedir.

SDM'ler dil kurallarının modellenmesinde ve özellikle biçim bilimsel çözümlemede sıklıkla kullanılırlar. Çeşitli bilgisayar dillerinde yazılan algoritmalar son olarak bu SDM'ye derlenerek çıktının üretilmesi sağlanır.

Butt ve King (2003: 129), dil bilgisi yazımının zor ve aynı zamanda yeterince tartışılmamış bir konu olduğunu belirtmektedirler. Klasik anlamda dil bilgisi kuramlarının dili betimlemek için ayrıntılı bilgiler sunabildiği bilinmekle birlikte, kuramın pratik uygulamalardaki değerinin anlaşılabilmesi dahası ölçülebilir bir biçime girebilmesi genel olarak DDİ için dil bilgisi yazım süreçlerinin ayrı bir konu olarak ele alınması gerekliliğini doğurmuştur. Bir sözdizimsel çözümleyici için çeşitli dil bilgisi kuramına dayalı biçimler (formalism) bulunmakla birlikte, bu biçimlerin makinede modellenmesi bir dil bilimcinin çalışmasından farklı olarak anında sınanmasını da

gerektirdiğinden, dil bilgisi yazan kişinin dil bilimsel bilgisi yanında makinede kullanılan veri biçimlerine ve işlevlerine de hakim olması gerekmektedir.

DDİ'de söz dizimsel yapıların çözümünde ***yüzeysel bölümleme*** (shallow parsing) için Chomsky'nin geliştirdiği bağlamdan bağımsız dil bilgisi kullanılırken, özellikle söz dizimsel kimi ikili sonuç üreten yapıların ele alınmasında cümle içi birimlerin ilişkilerini de içine alan ***sözlüksel işlevsel dil bilgisi*** (lexical-functional grammar) yaklaşımları denenmektedir. Butt ve King'de ***derin dil bilgisi*** (deep grammar) adıyla geçen yapılar ağaç diyagramları ve sözlüksel işlevsel dil bilgisi ayrımlamalarıdır. Sözlüksel işlevsel dil bilgisinde cümlenin öznesi ve nesnesi arasındaki ilişkiler ve özne yüklem uyumu da dikkate alınır (Butt ve King, 2003: 132).



Şekil 25. Yüzeysel Bölümleme ve Sözlüksel İşlevsel Dil Bilgisi Gösterimi (Butt ve King, 2003: 131).

Yukarıdaki şekilde, İngilizce "The group appeared" cümlesi 2(a ve b) örneğinde iki değişik gösterimle yüzeysel bölümlemeye alınmış, 2 (d) örneğinde sözlüksel işlevsel dil bilgisi gösterimi uyarınca daha ayrıntılı bir etiketleme yapılmıştır. Cümlenin zamanı (tense), yüklemi (pred) ve öznesi (subj) sol tarafta ana kategoriler olarak verilirken iç kategorilerde etiketlere İngilizceye özgü olarak bir tanımlık (specifier) eklenmiştir. Bu tür bir gösterim ***öznitelik değer matrisi*** (attribute-value matrice) olarak adlandırılmaktadır. Yukarıdaki gibi dil bilgisi gösterimlerinin çalışılan dilin özelliklerine göre değişeceği açıktır. Bu yüzden özellikle Türkçe için standartlaşmış gösterim araçlarına gereksinim Türkçe DDİ uygulamaları için değer kazanmaktadır.

Türkçenin otomatik biçimbilimsel çözümlemesine yönelik araştırmalar son yıllarda hız kazanmıştır. Başta Kemal Oflazer olmak üzere çeşitli üniversitelerde

biçimbirimsel çözümlemeye yönelik uygulamalar geliştirilmiş, hatta Türkçenin söyleyiş biçimlerinin modellendiği algoritmalar da çalışılan konular arasına girmiştir. Çağdaş Türk lehçeleri arasında makineli çeviri için özellikle Türkmence üzerinde çalışılmaktadır.

Türkçenin biçim birimsel çözümlemesinde *iki düzeyli biçim bilim* (two-level morphology) adı verilen yöntem de denen teknikler arasındadır. Bu yöntemde ***sözlüksel düzey*** (lexical surface) ve ***yüzeysel düzey*** (surface level) terimleriyle adlandırılan iki düzey söz konusudur. Yüzeysel düzeyde, biçim birimsel açıdan sözcük, yazımsal olarak temsil edildiği biçimde yani günlük kullanımda karşımıza çıkan biçimiyle görünürken sözlüksel düzeyde sözcüğün bütün işlevsel görünümleri parçalarına ayrılarak gösterilir. İki düzeyli biçim bilimde sözcüğün ses bilimsel gösterimi 4 kuralla temsil edilir:

1. `a:b => LC _ RC`
2. `a:b <= LC _ RC`
3. `a:b <=> LC _ RC`
4. `a:b /<= LC _ RC`

Yukarıdaki kurallarda, LC sol bağlam (left context), RC sağ bağlam (right context) kısaltmalarıdır. 1. kural a b olarak yalnızca sol ve sağ bağlam koşulunda gerçekleşir fakat bu gerekir şart değildir. 2. kuralda a, sol ve ve sağ bağlam açısından her zaman b olarak görülür. 3. kuralda, a her zaman b olarak görünür. Bu durum sadece sağ ve sol bağlamda görünür, başka bir bağlamda görünemez. 4. kuraldaysa, sağ ve sol bağlamda a hiçbir zaman b olarak görünmez. Bu kurallar bir SDM'de işlenecek yapılar olarak ele alınır ve SDM bütün kuralları sözlüksel ve yüzeysel yapı olarak pararlel biçimde kabul ya da reddeder. Yani kurala uyanlar kabul edilirken uymayanlar saf dışı bırakılırlar. Ayrıca bu kurallar yanında daha önceden belirlenmiş ***biçim birimi dizileri*** (morphotactics) bir SDM'de sözcük kökleri sözlüğü ve son eklerle temsil edilerek yukarıdaki kurallarla belirlenmiş ses bilimsel değişmeler, bu dizilerin sıralanışı esnasında ele alınırlar (Tantuğ vd., 2006: 187).

Tantuğ ve arkadaşları Türkmencenin iki düzeyli biçim bilimi için hazırladıkları SDM'de yukarıda belirtilen 4 sesbilimsel kural etrafında, Türkmencenin ses bilimsel varyasyonlarını temsil eden kural dizilerini şu biçimde çıkarmışlardır:

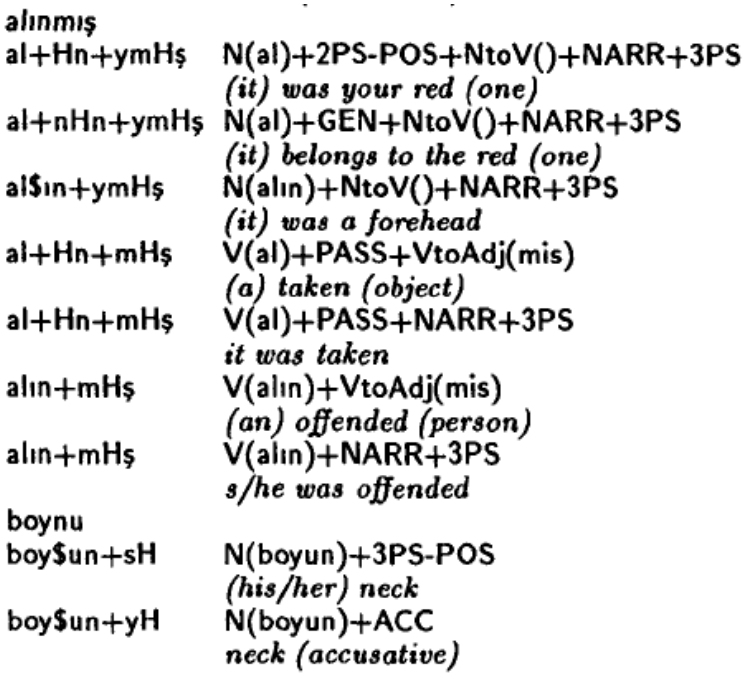
1. A:a => [:BACKV] [:CONS]* (%':%') (%+:0) [CONS: | :CONS | :0]* _
2. A:e => [:FRONTV] [:CONS]* (%':%') (%+:0) [CONS: | :CONS | :0]* _
3. V:a => [:BACKV] [CONS]* (%+:0) [CONS: | :CONS | :0]* _
4. V:E => [:FRONTV] [CONS]* (%+:0) [CONS: | :CONS | :0]* _
5. I:y => [:BACKV] [CONS]* (%':%') (%+:0) [CONS: | :CONS | :0]* _

6. I:i => [:FRONTV] [CONS]* (%':%') (%+:0) [CONS: | :CONS | :0]* _
7. H:i => [:FRUNROV] [CONS]* (%':%') (%+:0) [CONS: | :CONS | :0]* _
8. H:y => [:BKUNROV] [CONS]* (%':%') (%+:0) [CONS: | :CONS | :0]* _
9. H:u => [:BKROV] [CONS]* (%':%') (%+:0) [CONS: | :CONS | :0]* _
10. H:U => [:FRROV] [CONS]* (%':%') (%+:0) [CONS: | :CONS | :0]* _

Bu kurallar Xerox'un SDM aracının kabul ettiği gösterimler (notasyonlar) etrafında oluşturulmuştur. Xerox'un SDM'si biçim bilimsel çözümleyici olarak en sık kullanılan araçlar arasında yer almaktadır. Yukardaki kurallar Türkmence ünlü uyum kurallarını göstermektedir. Ünlülerin ön ve ardıl olmalarına göre alabileceği durumlar modellenmiştir (Tantuğ vd., 2006: 188).

Türkiye Türkçesi için iki düzeyli biçim birimsel çözümleyici geliştirme çalışmaları 1994'te Oflazer'in gerçekleştirdiği bir çözümleyici ile başlamıştır. Oflazer, çözümleyici geliştrime çalışmasında, Türkçenin sondan eklemeli bir biçim bilimsel yapıya sahip olduğunu bu eklemeli yapının sonlu olduğunu ancak yine de oldukça karmaşık bir görünüm sunduğunu belirtir. Bir ada gelen çeşitli yapım ekleriyle addan fiil türetme ve ve fiilden ad türetme gibi Türkçe için temel biçim birimsel görünümlerin yüzeysel düzeyde gerçekleşmesinde ses uyumlarının rolünü de ayrıca belirtmiştir. Bu açıdan Türkçe için hazırlanacak bir SDM'nin oldukça karmaşık bir yapıya sahip olacağı söylenebilir. 22 adet ses bilimsel kuralın hazırlandığı çalışmada biçim birimlerin dizilimleri yani morfotaktikleri de kodlanmıştır. Kuralların temel alacağı sözlükçe kısmı da bütün söz türlerini içine alacak biçimde 20950 sözden oluşmaktadır. Bu sözlerden 18500'ü ad ve sıfattan, 2450'si ise fiilden oluşmuş, bütün bu sözler kök biçimleriyle temsil edilmiş, 100 adet de ekler bölümü oluşturulmuştur[8] Aşağıdaki şekilde Türkiye Türkçesi için hazırlanan iki düzeyli biçimbirimsel çözümleyicide çeşitli sözlerin sözlüksel ve yüzeysel durumları gösterilmektedir.

[8] http://acl.ldc.upenn.edu/E/E93/E93-1066.pdf



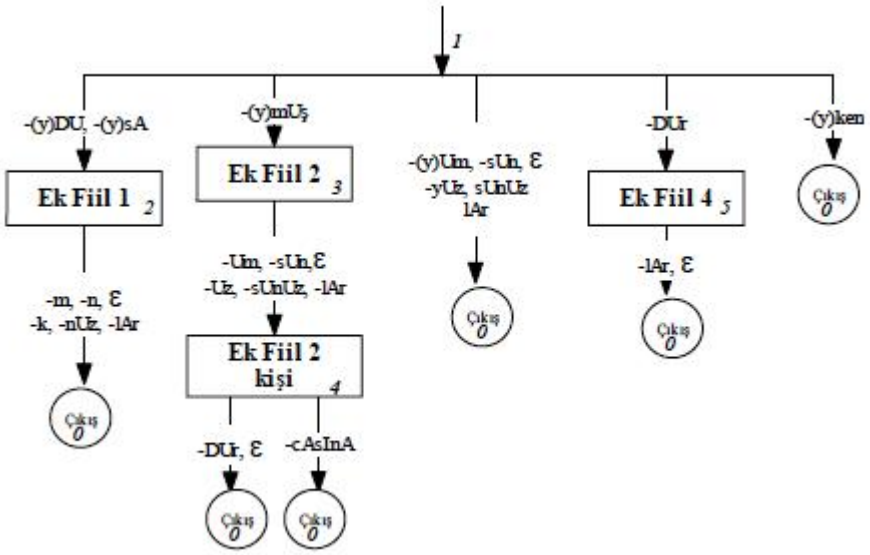
Şekil 26. Türkiye Türkçesi Biçim birimsel Çözümleyicide Sözlüksel ve Yüzeysel Düzeylerin Gösterimi [9]

Şekilde yer alana "alınmış" yüzeysel düzey için 7 sözlüksel düzeyin oluşturulduğu dikkati çekmektedir. SDM için "alınmış" sözcüğünün ad ve fiil mi durumundan başlayarak alabileceği bütün eklerin durumlarla temsil edilmesi gerekmektedir. Bu yapılmadan makinenin ayrıştırma yapabilmesi son derece zordur. Burada şunu da belirtmek gerekmektedir, literatürde *bağlamdan bağımsız dil bilgisi* (context-free grammar) olarak adlandırılan ve Chomsky ile başladığı belirtilen dil bilgisel yapıların salt bu biçimiyle özellikle çok anlamlı sözlerin çözümünde kural tabanlı bir yöntem olarak başarımının bir sınırı olacağı açıktır.

Biçim birimsel çözümlemede sözlerin eklerine ayrımlanması ve sonuçta köke indirgenmesi, kök sözlerin tutulduğu bir sözlük veritabanı yardımıyla yapılabilmektedir. Sözlük kullanılmadan da her bir ekin sözcükle kurduğu ilişkinin gösterimi ve kök sözcüğe ulaşılabilmesi için Cebiroğlu ve Adalı bir algoritma geliştirmişlerdir. Bu çalışmada ses bilimsel olarak hem ünlü hem de ünsüz uyumları bir SDM'de temsil edilmiştir. Hedef köklere ulaşmak için ekler 5 alt kümeye ayrılmış (yapım ekleri, ad ve fiil çekim ekleri ve ek fiil ekleri), kümeler farklı tablolar halinde bir veri tabanında saklanmıştır. Veri tabanında hızlı arama yapılabilmesi için eklerin ses olaylarıyla dönüş biçimleri de üretilerek kaydedilmiştir. Sağdan sola, sondan başa olmak üzere bir SDM

[9] http://acl.ldc.upenn.edu/E/E93/E93-1066.pdf

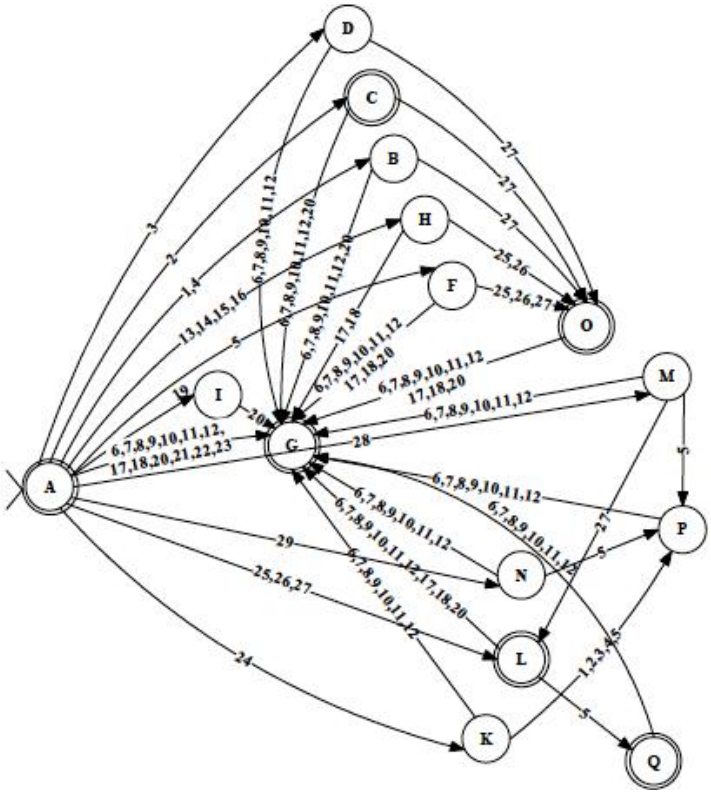
yapılarak, durumlar arası geçişlerin kontrolü için de ayrı bir modülle kontrol sağlanmıştır. Soldan sağa SDM'ler daha sonra bu makinede yer alan çıkış durumları dikkate alınarak sağdan sola yani sözcük sonundan başa doğru ters yönde bir SDM'ye alınmıştır. Bunun yanınsıra ekler ve ek kümeleri numaralandırılmıştır.



Şekil 27. Ek-fiil Kümesi için Bir SDM[10]

Makine bu yapıda karşılaştığı her bik ek durumunda ok yönlerinin sonunda gösterilen daire durumunda eki tanıyarak başka bir duruma geçmeden işlemini tamamlayacaktır. ε ile gösterilen yapılar boş yapılardır.

[10] http://www3.itu.edu.tr/~gulsenc/papers/TBD_Gulsen_Makale.pdf

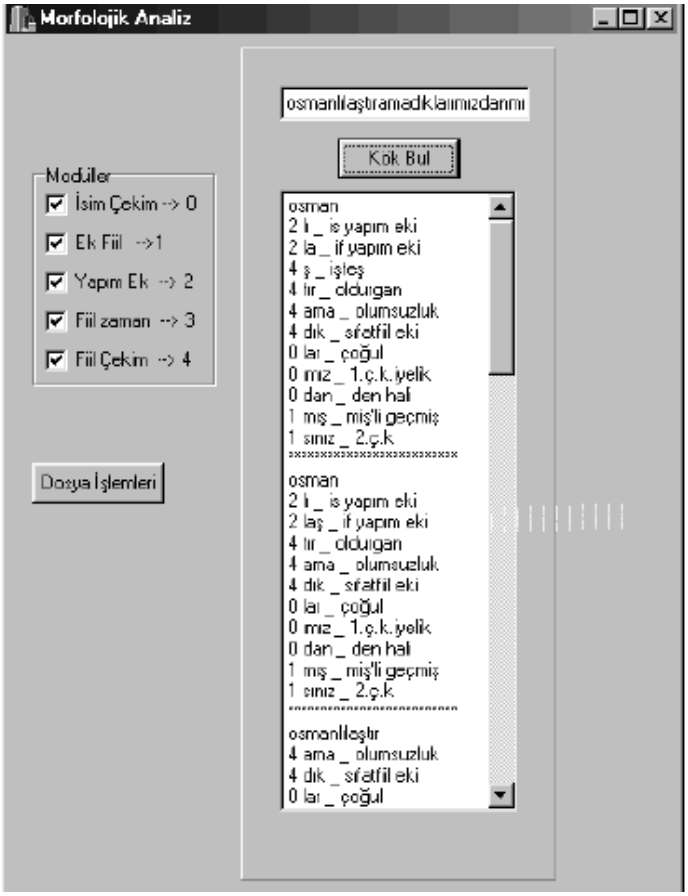


Şekil 28. Fiil Zaman Ekleri İçin Sağdan Sola Bir SDM[11]

Şekilde daireyle gösterilen durumlar ve durumlar arasında belirlenen eklerin numaraları yer almaktadır. Bu makine için okuyormuşum çekimli fiili örnek olarak verilmiş. oku-yor-muş-um çözümlemesinde ek numaraları olarak –yor :10, -muş:6, 27, -um: 1 belirlenmiştir. 1 numaralı ek bir A başlangıç durumundan daha önce belirlenmiş ek kümeleri sayılarına göre -um ekini B durumuna geçirirken bu makine kuralına göre 6 numaralı ekle B durumundan başka bir duruma geçiş olmadığından 27 numaralı ek ile 0 sonlu duruma geçilir. Yani bu söz bu noktada oku-yor-muş olarak tanınır. Sağdan sola devam edildiğinde 10 numaralı ek ile G durumuna gelinerek köke ulaşılmış olur ve söz şimdiki zamanın rivayeti olarak belirlenir.[12]

Bu makinenin bir programa dönüştürülmüş şekli aşağıda yer almaktadır. “osmanlılaştıramadıklarımızdan mısınız” cümlesinin çözümünü hazırlanan bu SDM yapmıştır.

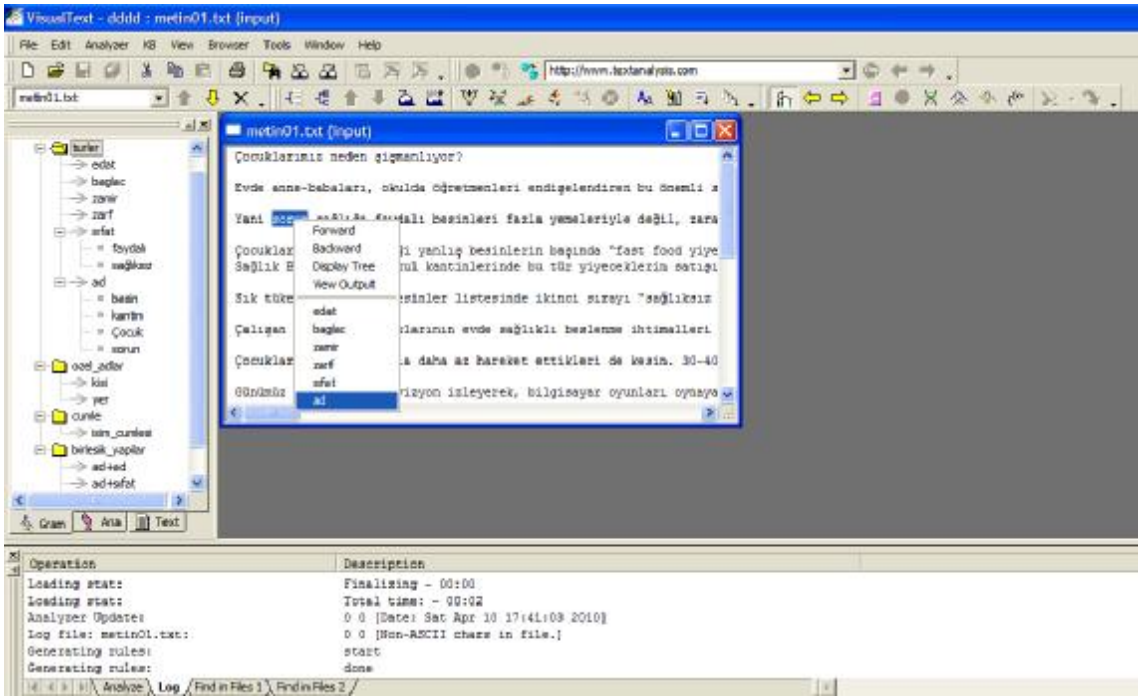
[11] http://www3.itu.edu.tr/~gulsenc/papers/TBD_Gulsen_Makale.pdf
[12] http://www3.itu.edu.tr/~gulsenc/papers/TBD_Gulsen_Makale.pdf



Şekil 29. Programa Dönüştürülmüş Biçim Birimsel Çözümleyici [13]

Metin işlemede, kural tabanlı yöntemler kullanarak derlemlerin oluşturulmasında ya da çeşitli dil bilimsel incelemede grafik arayüzlü uygulamalar kullanmak mümkündür. Bu uygulamalar arasında akademik kullanım için ücretsiz olan VisualText adlı yazılımla dil bilgisi kuralları oluşturmak ve bunları metinler üzerinde uygulamak mümkün olabilmektedir. Bu tür yazılımların her biri kendi iç işleyişine göre dillere sahiptir. Bu bağlamda VisualText, NLP++ adı verilen ve C diline dayanan bir yazılım olarak geliştirilmiştir. NLP++ ile görsel olarak geliştirilecek kurallar yanında, klasik programlama dili mantığına uygun olarak daha ince ve ayrıntılı kurallar oluşturarak örneğin anlam belirsizliği giderme gibi üst düzey uygulamalar geliştirilebilir. Yazılımda önce kuralların içinde tutulacağı üst düzey bir bilgi tabanı belirlenir. Bilgi tabanında yer alacak kurallar önceden tanımlanarak örnek metinlerde bu kuralların işaretlemesi yapılır. En basit biçimiyle söz türlerinin belirlenmesi için bir işaretleme aşağıdaki şekilde gösterilmiştir.

[13] http://www3.itu.edu.tr/~gulsenc/papers/TBD_Gulsen_Makale.pdf

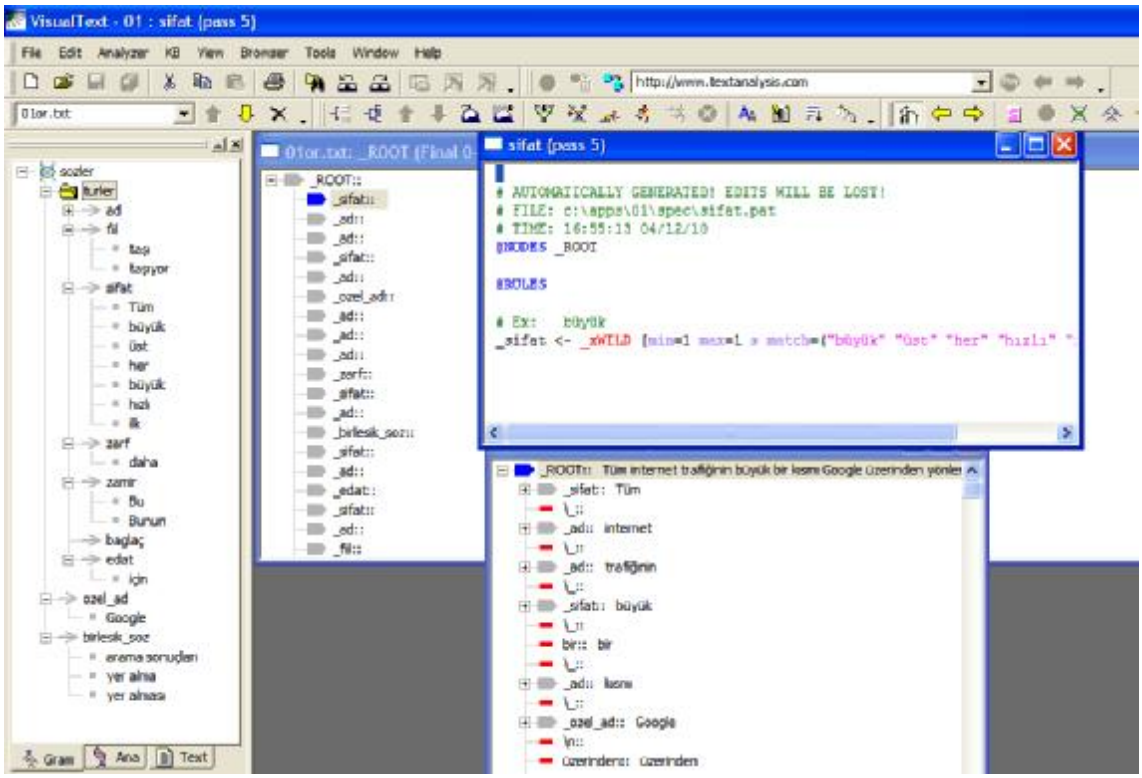


Şekil 30. VisualText Dil Bilgisi Geliştirme Ortamı.

VisualText, her bir kuralı bir düğüm (node) biçimine getirerek hiyerarşik ağaç modeli oluşturur. Bu modelde en üstte programın varsayılan düğümü root olarak adlandırılarak diğer düğümler bu düğümün altında yer alırlar. Bir düğümden diğerine geçiş kuralı elle belirlenebileceği gibi garfik ara yüzünde otomatik olarak da derlenebilmektedir. Aşağıdaki şekilde Türkçe için küçük bir metin üzerinde söz türlerine dayalı bir dil bilgisinin VisualText'te hiyerarşisi ve bu hiyerarşiye dayalı ağaç yapısı görülmektedir. Bu şekilde söz türleri yazılımın "pass" dosyaları adı verilen kural dosyalarında tutulur. Her bir kural bir düğümle temsil edilir ve düğümdeki kurallar sağdan sola olmak üzere düğüm başına birleştirilirler. Sıfat düğümüne birleştirilen kuralların yapısı şu biçimdedir:

```
_sifat <- _xWILD [min=1 max=1 s match=("büyük" "üst" "her" "hızlı"
"ilk" "Tüm")] @@
```

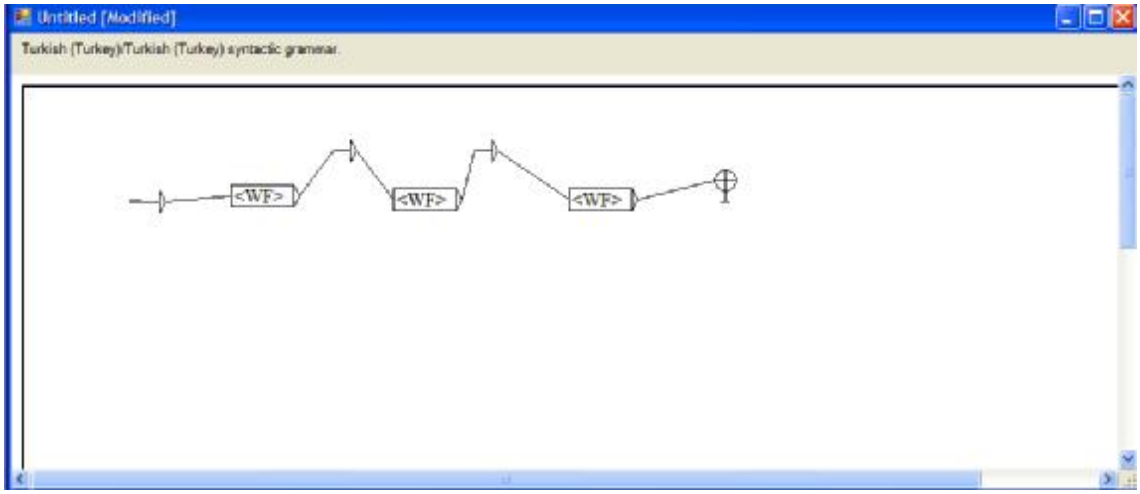
Burada match fonksiyonu ile metinde eşleşecek sıfatlar bir dizi olarak işlenir. Her bir kural @@ işaretiyle sonlandırılır. Şekilde _Root ana düğümünden sonra makinenin eşleştirdiği birimler de atanan düğüme göre metnin soldan sağa işlenişi yönünde belirlenmiştir. Bu dil bilgisi geliştirme ortamı gereksinimlere göre genişleyebilme yeteneğine sahip olup, örneğin sadece biçim birimsel yapıların ele alınacağı modüller geliştirilebilmektedir.



Şekil 31. VisualText Dil Bilgisi Editörü.

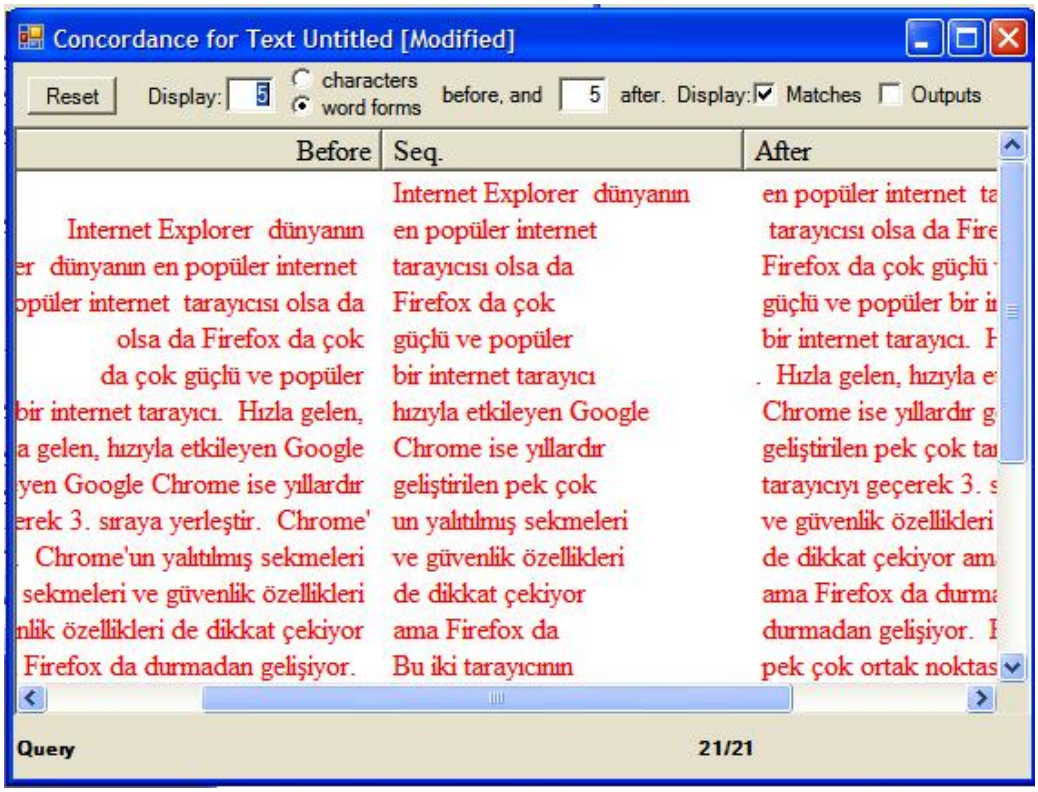
Grafik ara birimli dil bilgisi geliştirme ortamı olarak çok kullanılan NooJ sistemi de kural tabanlı işlemler için tasarlanmıştır. Gerek grafik ara birimi gerekse elle yazılan kurallar NooJ'da da bir SDM'ye otomatik olarak dönüştürülmektedir. Yazılım kendi içinde PERL genel amaçlı programla diline uygun düzenli ifadelerle işlem yürütebilme kapasitesine sahipken aynı zamanda NooJ için geliştirilen düzenli ifadeler seti de kullanılabilmektedir. Diğer yazılımlarda olduğu gibi NooJ'u kullanabilmek için yazılımın kendi iç mimarisini ve kullandığı programlama mantığını öğrenmek gerekmektedir. Sonuçta ortaya konulan ürünlerin hepsi kurallar etrafında çalışmakla birlikte, her yazılım kendi kural ve işaretleme seti uyarınca çalışmaktadır.

NooJ basit biçim birimsel çözümleyiciden karmaşık cümle çözümleyici kurallar oluşturulmasını desteklemektedir. Birden çok derlemin de hazırlanabildiği ortam, derlem işaretleme ve etiketleme olanakları da sunmaktadır. NooJ'da dil bilgisi kuralları için kullanılan grafik editörü bir başlangıç ve bir son durum arasında yazılımca belirlenen kimi sembollerle yazılan kuralları içerir. Örneğin herhangi bir sözcük biçimi için "WF" sembolü kullanılırken söz türlerinin işaretlenmesi mutlaka bir "+" operatörüyle başlamaktadır. Aşağıdaki şekilde herhangi iki sözcüğü bulabilen bir SDM kuralı grafik tabanlı olarak hazırlanmıştır.



Şekil 32. NooJ Dil Bilgisi Geliştirme Ortamı.

Yukarıdaki şekilde WF içeren kutucuklar durumları, boş oklar da boş karakter dizisini temsil etmektedir. Bu yapının sonucu aşağıdaki şekilde gösterilmiştir. Şekilde sağ ve sol bağlamla birlikte 3 sözcüklü yapılar ortadaki sütunda yer almaktadır.



Şekil 33. NooJ Sisteminde Bağlamlı Dizin.

1.4.2.2. İstatistiksel Yöntem

İstatistik; verilerden yola çıkılarak verilerde belirlenen değişkenler arasındaki ilişkilerin belirlenmesi, bu ilişkilerden yola çıkarak kararların verilmesi, dağınık verilerin düzenlenmesi ve grafiksel olarak gösterilmesini, olasılık kuramları yardımıyla tahminlerin yapılmasını ve elde edilen sonuçların test edilmesini sağlayan tekniklerin geliştirilmesiyle ilgilenen bilim dalıdır. Bir olgunun sayısal olarak ifade edilmesi ve bu ifadelerin başkalarına iletilmesi istatistiksel yöntemler yardımıyla sağlanmaktadır (Gürsakal, 2008:6).

Doğa bilimleri başta olmak üzere çeşitlibilim alanları istatistik yöntemlerinden yararlanmaktadır. Nicel verinin esas olduğu doğa bilimlerinde, istatistik doğrudan ulaşılan bir inceleme yöntemi olurken, sosyal bilimlerde değişkenlerin sınırlarının kesin olmaması ve verinin nitel yönlerinin baskın olması nedeniyle sosyal bilimlerde nitel verinin sayısallaştırılması başka bir ara aşama olarak görülebilir. Dil bilimsel veriden elde edilecek başta sözcükler olmak üzere "tekrarlanan" veriler istatistiksel ve olasılıksal çalışmalar için uygun bir çalışma ortamı oluşturmaktadır. Sözcüğün en alt biriminden cümle üstü birimlere kadar gerçekleşen dil olguları bir noktadan sonra zamanla tekrar tekrar gerçekleşirler, dolayısıyla gerçekleşen her bir dil olgusu sayısal bir ortama alınmak üzere gözlemlenebilir ve ölçülebilir bir nitelik kazanır. DDİ'nin çeşitli uygulamalarında da derlemler aracılığıyla bu gerçeklemiş dil olguları sayısallaştırılarak örneğin basit söz türü etiketlemeden karmaşık anlam bilimsel çözümlemeler için kullanılacak veri kümeleri durumuna getirilirler.

Son yıllarda istatistiksel DDİ çalışmalarının büyük bir bölümü AT&T Bell laboratuarları ve IBM firmasınca gerçekleştirilmiştir. Bu çalışmalar daha çok konuşmanın sinyal işleme tekniklerinin istatistiksel modellenmesine yöneliktir. Özellikle sözlü dile yönelik istatistik tabanlı çözümlemeler kural tabanlı yöntemlere göre daha çok başarım gösterdiğinden, kural tabanlı çözümlemeler konuşma tanımada önemini yitirmektedir (Burch ve Osborne, 2003: 269).

Burch ve Osborne'a göre (2003: 270), istatistiksel DDİ'nin avantajlarını şu biçimde belirtmek mümkündür:

Bu yöntemde hızlı prototipler oluşturulabilir. Elle oluşturulacak bir veri seti zaman alıcıdır. Derleme dayanan veri setleri makine öğrenmesine de uygun olduğundan daha zaman kazandırıcıdır, öyle ki Al-Onazian ve arkadaşları Çince-İngilizce makineli çeviri sistemini 24 saatlik bir sürede tamamlamışlardır. Bununla birlikte bu sistemin

başarımıyla ilgili Burch ve Osborne herhangi bir bilgi vermemişlerdir. İngilizce literatürde oluşturulan uygulamaların sağlamlığı için "robust" sözcüğü sık kullanılmaktadır. Bu sözcükle, sisteme giren veri ne kadar kötü düzenlenirse düzenlensin sistem bir çıktı üretebilmektedir. Burch ve Osborne sağlamlık kavramını anlatmak için metin sınıflama çalışmalarından örnek vererek, metinde geçen sözler daha önceden karşılaşılmamış sözler olsalar bile istatistiksel sınıflayıcının yine de bu metinleri sınıflayabileceğini belirtmektedir.

İstatistiksel yöntemin maliyeti, kural tabanlı yönteme göredüşüktür, çünkü elle hazırlanacak yöntemlere göre daha otomatikleştirilmiş olacak, çalışanların sayısı da düşecektir. Bu yönteme dayanan sistem, bir veri seti üzerinden sayısal yöntemlerle öğreneceğinden, kural tabanlı sisteme göre incelenen dile özgü bilgiye daha az gereksinim duyacaktır.

Kural tabanlı yöntemde olduğu gibi istatistiksel yöntemde de incelenecek dile özgü bir modelin oluşturulması gerekmektedir. Verinin yoğun olarak kullanıldığı DDİ'de modelleme, verinin doğrudan kendisi üzerinden elde edilir, yani bir dilin bir alana bağlı kesiti incelenecekse (özellikle terime duyarlı bilimsel metinlerde) o kesite özgü hazırlanmış derlemler model oluşturmada birincil kaynak durumundadır. Burch ve Osborne (2003: 275), modellemeyi, işlemleri taklit eden makineler kurma çalışmalarının genel adı olarak tanımlayarak istatistiksel model oluşturma aşamalarını şu biçimde ele almışlardır:

Modelin yapısal bileşenleri ortaya çıkarılır. Bu işlem daha çok elle yapılabileceği gibi metin sınıflamada belgelerin benzerliklerine göre kümelendirilmesi gibi otomatik olarak da gerçekleştirilebilir. Burada en önemli nokta ele alınacak dil birimlerinin niteliksel yapılarının sayısallaştırılması aşamasıdır. Nitel verilerin sıklığı, ancak nitel verinin bir sayıya karşılık gelen işaretiyle ölçüleceğinden, girdi olarak mutlaka bir niteliğin sayıya karşılık gelmesi gerekmektedir. Olasılık hesaplarında da tekrarlanan yani sıklık bakımından bir özellik gösteren birimler işlenebilmektedir.

Modeli oluşturan parametre bileşenleri belirlenir. Parametreler istatistikteki değişken terimine karşılık gelmekte, parametreleri alacağı değerler de olasılıklar olarak görülmektedir. Bir dil bilgisi modelinin parametreleri kuralların olasılıklarına denk düşmektedir, dolayısıyla olasılıklar, kuralların bir araya gelme düzenini belirleyecektir.

Son olarak modeller parametrelere atanan değerlerin örnekler aracılığıyla belirlenmesiyle tam olarak oluşturulurlar. Örnekler dil bilgisel etiketlemesi yapılmış

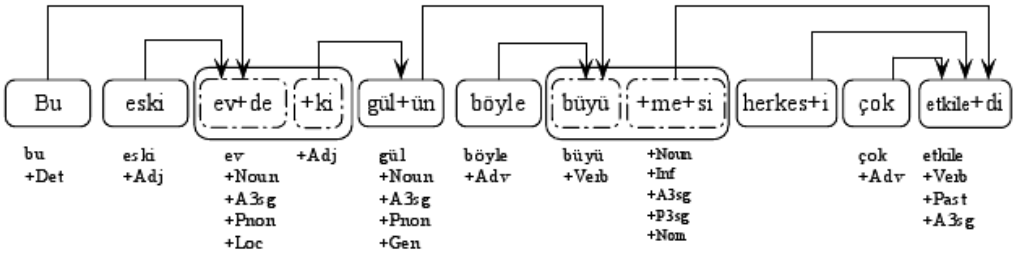
derlemlerden elde edilir. Her bir dil bilgisel yapının olasılığı dikkate alınarak model kurulmuş olur.

Modeller üretici (generative) ve ayırıcı (discriminative) modeller olmak üzere ikiye ayrılırlar. Üretici model cümlelerin üretilmesi anlamına gelmektedir, yani verili bir seçenek olmadan cümlelerin kurulması söz konusudur, bu anlamda dil bilgisine dayanarak cümleler oluşturulabilir. Verili seçenekler (olasılıklar) arasından uygun olanı belirlemek ayırıcı modelin niteliğidir. Ayırıcı model daha çok makineli çeviride tercih edilmektedir, bu anlamda kaynak dilde ele alınan bir cümle için hedef dildeki olasılığı yüksek cümle seçimi yapılır. Bu biçimde aslında kaynak dildeki cümleler modellenmemekte hedef dilin verisindeki olasılıklar dikkate alınmaktadır. Üretici ve ayırıcı modeller olasılık hesabı dikkate alındığında birtakım farklılıklar gösterirler. *Ortak olasılık* (joint probability) terimi iki olayın birlikte gerçekleşme olasılığı hesabına dayanır ve ***P (sözcükler,etiketler)*** biçiminde gösterilir. Burada sözcüklerle etiketler aslında birbirine bağımlıdır, dolayısıyla her bir sözcüğün bir etiketi olmalıdır. Bu modelde makine, birimlerin bütün dağılımlarının çıkarılmasına gerek duyar. ***P (sözcükler|etiketler)*** biçiminde gösterilen yapıyı ayırıcı model kullanır. Burada daha önceden verilen örneklerden yola çıkılarak olası durumlar kestirilmeye çalışılır. Bir sözcüğün sıklığı biliniyorsa, girdi olarak ele alınan metinde, verilen sıklığa dayanılarak çıkarılan formülle sözcüğe atanacak etiketin türü kestirilebilir. Üretici model bağlama bağlıyken, ayırıcı model bağlamdan bağımsızdır (age., 2003: 276, 277).

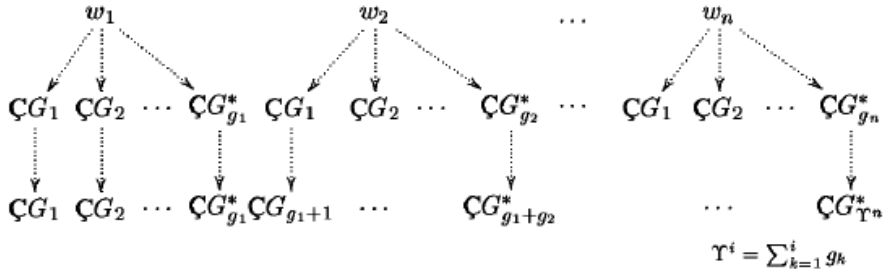
Bilindiği gibi dil birimleri doğrusal olarak sıralanırlar. Bir birimden sonra bir başka birim söylenir ya da yazılır. DDİ'de ***n-gram*** olarak adlandırılan bu birimler ikili, üçlü ve daha çok birimden oluşabilirler. N-gram terimini Baker ve arkadaşları şu biçimde tanımlarlar: "Boşluklar kaldırıldıktan sonra, verili bir söz katarında (string) n tane harfin dizlimi. Örneğin, 'how are you' cümlesinde n=3 için şu biçimler ortaya çıkar: 'how', 'owa', 'war', 'are', 'rey'" (Baker vd.: 2006: 122). Tanımda geçen örnekte, boşluk karakterli dışında kalan harfler 3'lü gruplar halinde bir sonraki sıradan devam etmektedirler. N-gramlar ikili olduklarında ***bigram***, üçlü olduklarında ***trigram*** adını alırlar. N-gramlar istatistiksel çözümlemenin birimleri olarak düşünülebilir. Baker ve arkadaşları n-gramları boşluk dışında kalan harf dizileri olarak değerlendirirken, Jurafsky ve Martin olasılık hesabında boşluk karakterlerini de içerecek biçimde bu tanımı genişletirler. Onlara göre, bir sözcükten sonra gelecek sözcüğün olasılığı bir önceki sözcüğe bağlıdır, dolayısıyla temelde bigram hesabı yapılır bu da koşullu olasılıkta P ($W_n|W_{n-1}$) formülüyle ifade edilir. Görüldüğü gibi Jurafsky ve Martin n-

gramları sözcük düzeyinde düşünerek (formülde W ile temsil edilirler) iki sözcüklük yapıları bigram olarak tanımlarlar. Bir sözcükten önceki sözcüğün sayısal verisine bakılarak elde edilen olasılıksal düzen Markov varsayımı olarak adlandırılmaktadır. Markov zinciri olarak olarak adlandırılan yapı, SDM'de durumların sayısal ağırlıklarına dayanmaktadır. Bir ağırlıklı SDM'de bir sonraki durumun olasılığı önceki durumların ağırlığıyla belirlenmektedir (Jurafsky ve Martin, 2000: 195).

Türkçe için olasılık tabanlı bir çalışma Eryiğit ve arkadaşlarınca gerçekleştirilmiştir. Olasılığa dayalı bağlılık bölümlemesi bir söz dizimsel çözümleme olup, Eryiğit ve arkadaşları ***bağlılık dil bilgisi***nden (dependency grammar) hareketle Türkçe için hazırlanmış sınırlı sayıdaki bir ağaç yapılı derlemdeki (Oflazer vd., 2003) verileri kullanarak istatistiksel parametreleri belirlemişlerdir. Bağlılık çözümlemesinde, söz dizimsel birimlerin, çalışmada iye ve uydu birimler olarak adlandırılan ilişkisel doğası üzerinde durulur. Sözcükler, yüzeysel yapıları dikkate alınmadan çekim kümeleri adı verilen ve ÇK olarak kısaltılan kümelere ayrılırlar. Çekim kümelerinde, ele alınan sözcüğün türetim sınırlarından sonra gelen çekim eklerinin sınırı, iye sözcüğe bağlanma noktası olarak kabul edilir. Bölümleyici, birimler arasındaki olasılıklar hesaplandıktan sonra, bir bağlılık ağacındaki tüm olasılıklar içinde en olası durumu bulmaya çalışır. Aşağıdaki şekilde bir cümlenin bağlılık gösterimi ile W ile temsil edilen sözcükler ve çekim kümeleri görülmektedir.



Şekil 3. Türkçe bir tümcedeki bağlılıklar



Şekil 34. Dil Bilgisel Bağlılık Gösterimi (Eryiğit vd., 2008: 110).

Çalışmada, Türkçe ağaç yapılı derlemdeki bağlılıkların %95'i sağa yönelimli olarak bulunmasından hareket edilerek Sekine ve arkadaşlarının (2000) geliştirdiği geriye doğru demetli arama algoritması kullanılmıştır. Bu arama yönteminde, üzerinde bulunulan birim sağındaki birime bağlanmaya çalışılır. Adım adım ilerleyen yöntemde olasılığı yüksek birimler bir demet içinde tutulur, belirlenen demetten daha yüksek olasılıklı başka bir demet bulunduğunda olasılığı daha yüksek demetler düşük olanların yerini alır. Çekim kümelerinin bulunması için biçimbirimsel çözümlemede gerektiğinden aynı zamanda biçimbirimsel çözümleyici de işlem sırasında kullanılmıştır. Çalışmada çekim kümelerinin sınırları aynı zamanda içinde bulunulan bağlam bilgisini de sağlamaktadır. Sonuçta başarım oranı olarak %73.5'lik bir orana ulaşılmıştır (Eryiğit vd., 2008: 109,110, 116).

1.4.2.3. Melez Yöntem

Melez yöntemler, *konuşma tanıma* (speech recognition) başta olmak üzere DDİ'nin çeşitli çalışma alanlarında kural tabanlı ve istatistiksel yöntemlerin birleştirilmesiyle oluşan yöntemlerdir. Kural tabanlı yöntemlerin oluşturulan tanıma sistemlerinin başarımlarında karşılaşılan eksiklikleri, istatistik yöntemlerle (başta saklı Markov modelleri) giderilmeye çalışılır.

Wermter ve Lehnert (1992: 101-102) son yıllarda *bağlantıcı yöntem*in (connectionist) sembolik yöntemler yerine daha sık kullanılmaya başlandığını belirterek bu yöntemin yapay zeka alanında uygulama yönelimli ve iş yönelimli olmak üzere iki ana yönde geliştiğini belirtirler. Uygulama yönelimli bağlantıcı yöntem, *anlamsal ağ* (semantic network), kural tabanlı sistemler, hiyerarşilerde ve ağaç yapılı uygulamalarda kullanılmıştır. İş yönelimli bağlantıcı yöntem; bölümleme, anlamda belirsizlik çözümü, birleşik yapıların tanınması, cümle üretimi gibi daha özgül alt alanlarda modelleme için kullanılmıştır.

Wermter ve Lehnert, hem bağlantıcı hem de sembolik yöntemi kullanarak teknik İngilizcedeki isim öbeklerini bulmaya yönelik bir çalışma yapmışlardır. Bunun için her ismi 16 özellikli ikili bir vektörle ilişkilendirmişler, bu yapılırken isimler arasındaki anlamsal ilişkileri dağıtık ağ yapısıyla tanımlamışlardır. Anlamsal ağları, söz dizimsel özelliklerle de güçlendirerek kavramların çıkarımı için de çalışmaya sembolik yöntemi eklemişlerdir. Bu çalışma, bir yapay sinir ağı oluşturmaya yönelik olup, isim

öbeklerinin kurucu ögelerinin özellikleri, kavram sözlükleri yardımıyla üst anlam ilişkileri ve genellemelere yönelik düğümleri barındırırken, *yerelci* (localist) bağlantılarla da ismin bulunduğu bağlamın diğer isim ya da edatlarla olan uzaklıkları göz önünde tutulmuştur. Bugün daha çok eşdizimlilikler olarak bilinen bu yaklaşımla makineye kurucu ögelerin bağlam bilgileri sağlanmış olmaktadır. Sembolik yöntem, isim dizilimleriyle isim öbeklerinde yer alan edatların çıkarımında kullanılarak ağ yapısına girdi verilerini sağlamıştır. Sonuç olarak Wermter ve Lehnert, bağlantıcı ve sembolik yöntemin her iki yöntemin tek başına kullanımına göre sistemin başarım oranını yükselttiğini belirtmişlerdir (Wermter ve Lehnert, 1992: 115, 116).

1.4.3. Doğal Dil İşlemede Uygulama Alanları

1.4.3.1. Belge Bölütleme

Var olan basılı metinlerin, belgelerin içeriklerinin değiştirilmeden biçimsel özelliklerinin olduğu gibi korunarak saklanması, istendiğinde sorgulanabilmesi için öncelikle belgenin sayısallaştırılması gerekmektedir. Sayısallaştırılmış bir metin günümüz internet ortamında sunulmak üzere kaydedilebileceği gibi, özelleşmiş *veri madenciliği* (*data mining*) yöntemlerinde kullanılmak üzere hazırlanabilmektedir.

Basılı belgelerin sayısallaştırılmasında en çok kullanılan yöntem *optik karakter tanıma* (*OKT*) (*Optic Charecter Recognition-OCR*) yöntemidir. Bu yöntemde basitçe ev kullanıcıları için hazırlanmış küçük boyutlu *tarayıcılar* (*scanner*) kullanılabildiği gibi, büyük metin işleme projelerine özgü hazırlanan belge tarayıcılar yoğun bir biçimde kullanılmaktadır. Tarayıcıdan salt görüntü olarak makinede kaydedilen belge işlemenin en önemli bölümü olan yazılım aşaması adını verebileceğimiz aşamada belge tanımaya özgül hazırlanmış yazılımlarla yeniden düzenlenebilir biçime (formata) dönüştürülür.

OKT'de, belge sınırlarının belirlenmesinden karakter sınırlarının çizilmesine ve çözümlenmesine kadar bazı işlemler gerçeklenir. Bu aşamalar sırasıyla şunlardır: Arka plan ayrıştırma, yerleşim planı çözümleme, kayıklık açısının düzeltilmesi, satır ayrıştırma, sözcük bölütleme/tanıma, karakter ayrıştırma, karakter tanıma, sözcük tanıma, başarım değerlendirmesi (Yanıkoğlu, 2006:111).

1.4.3.2. Bilgiye Erişim

İnternetle birlikte artan elektronik metin ortamı, çeşitli tür ve boyutta metinden oluşan karmaşık bir metinsel ağ yapısını da ortaya çıkarmıştır. Son kullanıcının bilgiye

ulaşma gereksinimi karşılamak üzere çok sayıda arama motoru bu karmaşık metin yığınları arasından değişik arama ve bulma algoritmaları kullanarak sonuçlar üretmektedirler. ***Bilgiye erişim*** (information retrieval) kavramını Manning ve arkadaşları (2008:1) "Yapılandırılmamış karakterdeki metinlerden, bir bilgi gereksinimini karşılamak amacıyla belge niteliğindeki malzemeyi bilgisayarlarda depolanan büyük koleksiyonlardan bulmak" biçiminde tanımlamışlardır. Bu tanımda geçen koleksiyonlar, sayıları binleri bulan metinlerden oluşan ve değişik sınıflara ait yapılardır. Baker ve arkadaşları (2006:90), bilgiye erişim kavramını "Bilgisayarları kullanarak derlem, veri tabanı ya da internet gibi büyük ölçekte veri barındıran ortamlardan belirli bir enformasyonu ayırıp çıkarmak" olarak tanımlamışlardır. Bu tanımda koleksiyonlar yerine derlem ve veri tabanı ve daha büyük bir ortam olarak internetin veri kaynağı olarak ayrımlandığı görülmektedir. Günümüz internet arama motorlarının yaptığı temel işlemlerde kullanılan yöntemler aslında bir tür bilgiye erişim yöntemidir.

Bilgiye erişim temelde, ilgili belge veya belgeleri bulma teknolojisidir. Bu teknolojinin kullanımı, karmaşık veri yığınları içinden son kullanıcının bilgisayar kullanarak yapamayacağı birçok sorgulamayı ve sorgulama sürecinde harcayacağı zamanı en aza indirmektedir. Doğal dil işleme tekniklerinin kullanılması, bir metinde geçen terimlerden yaralanılarak terime duyarlı belgelere ulaşılmasını ve bu belgelerin sınıflandırılmasını kolaylaştırmaktadır. Türkçe belge koleksiyonları üzerinde Can ve arkadaşlarının (2008) yapmış olduğu çalışmada, ***gövdeleme*** (stemming) tekniği kullanılmış, sözcüklerin çekimli biçimleri ayıklanarak elde edilen dizine ait olabilecek belgeler bir araya getirilmeye çalışılmıştır. 2001-2005 yılları arasında yayımlanan Milliyet gazetesinin haber metinlerinden oluşan 408.305 adet belgeyi içeren 800 MB boyutundaki koleksiyonun sözcük sayısı tekrarlı sözcükler de dahil olmak üzere 95.5 milyondur. Bu sayıya numerik biçimler de dahil edilmiştir.

Çalışmada, sözcüklerin gövdeleme yapılmadan, gövdeleme yapılarak ve bir ***dur sözcükleri listesi***ne (stop-word) başvurularak yapılan yöntemlerin genel olarak istenilen bilgiye erişimde başarım oranları değerlendirilmiştir. Dur sözcükleri listesi, bilgi içeriği düşük ve çok sık kullanılan sözcükleri içerir. Örneğin Türkçede sık kullanılan edatlar (sonra, için, bile, ancak vb.) bu liste içinde yer alırlar. Çalışmada bu listelerin sistemin başarımına etkisinin olmadığı görülmüştür. Sabit önek gövdeleme (fixed prefix stemming) algoritmasıyla madde başı gövdeleme (Lemmatizer-based stemming***)*** kullanılarak gerçekleştirilen çözümlemelerin başarım ölçeğine göre doğru sonuçlar

üretme oranının arttığı bulunmuştur. Sabit önek gövdelemede, sözcüğün başından (soldan sağa) bir doğrusal ilerlemede 3 ile 7 karakter aralığındaki sözcük biçimleri kök kabul edilmiştir. Madde başı gövdelemede ise elde edilen biçim sözlükte bulunacak madde başlarıdır. Bu teknikte sözcüğün ait olduğu söz türü de belirlenmeye çalışılarak olası birden çok örneklerin sözcük türüne bakılarak teke indirilmesi amaçlanır. Çalışmada madde başı gövdeleme algoritmasında birden çok sonuç üreten yapılarla karşılaşılmış, doğru biçimin bulunması için Türkçe için belirlenmiş ortalama kök karakter uzunluğuna sahip olanlar aday biçimler arasından seçilmiş, bu da doğru biçimi bulamadığında sözcük türüne bakılmıştır.

Çalışmada, bir bilgiye erişim sisteminin özellikle tam metin barındıran koleksiyonlar üzerinde daha fazla başarım gösterdiği ayrıca son kullanıcı sorgusunun uzunluğuyla (üç ya da daha çok sözcükten oluşan sorguların) istenilen nitelikteki belgeleri getirmesi bakımından daha seçici sonuçlar ürettiği belirtilmiştir. Ayrıca etkinlik bakımından, madde başı gövdeleme ile derleme dayalı istatistiksel gövdeleme arasında bir fark bulunmamıştır (Can vd., 2008: 410, 411, 418).

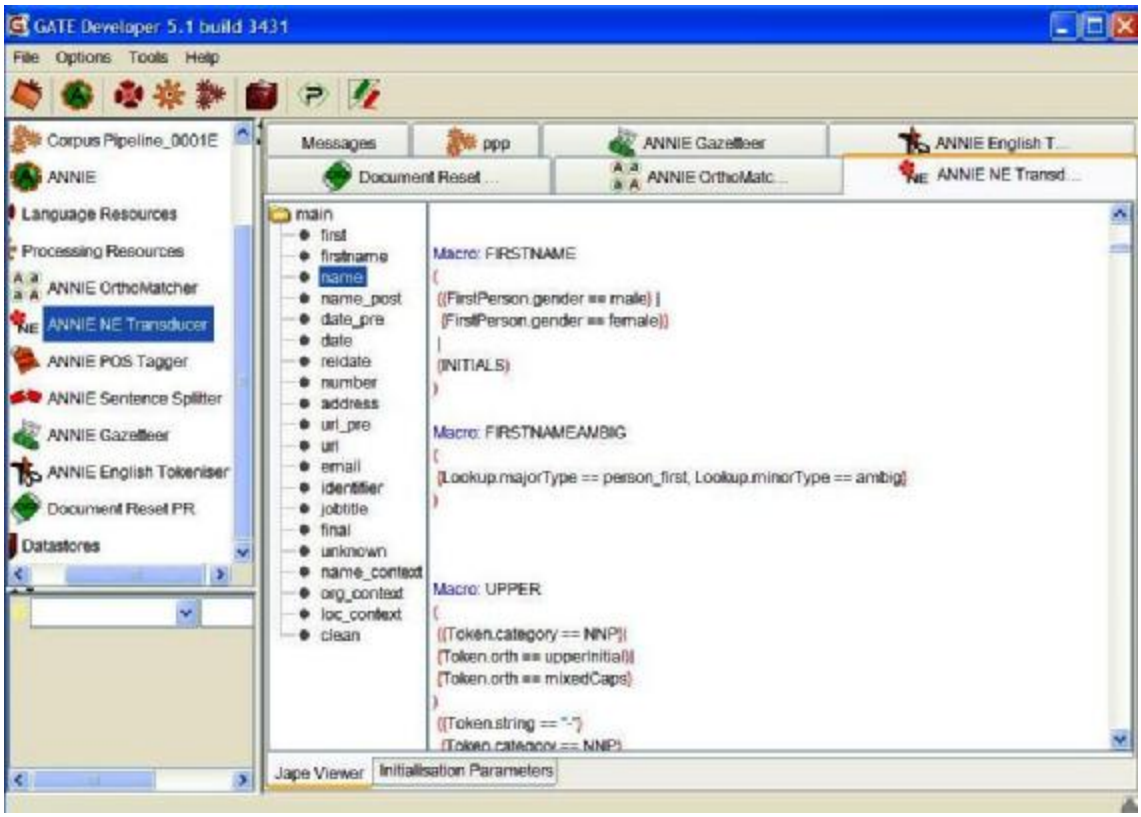
1.4.3.3. Bilgi Çıkarımı

Bilgi çıkarımı, bilgiye erişim kavramıyla yakından ilişkili olmakla birlikte, ondan daha ince ve ayrıntılı teknikleri içerir. Belge koleksiyonlarının belirlenmesi ve istenilen bilgiyle ilgili belgenin ilişkilendirilmesi sürecini kapsayan ve daha genel bir sınıflamaya götürecek teknikleri içeren bilgiye erişimin bir sonraki adımı olarak düşünülebilecek bilgi çıkarımı, metin düzeyinde özel olarak açıkça ifade edilmiş ya da gizli bir örüntü olarak yer almış yapıları bulup sınıflandırma işi olarak tanımlanabilir. Cowie ve Wilks'e göre, bilgi çıkarımında bir ya da birden çok metin üzerinde uygulanan tekniklerin çıktıları bir veri tabanını güncellemek için kullanılır. Çözümleme yapan araştırmacı, daha önceden veri tabanını bilgi çıkarımı için manuel olarak hazırlamış olmalıdır. Metinler üzerinde otomatik bilgi çıkarımı teknikleri, bilgiye erişim tekniklerine göre daha yeni bir çalışma alanıdır. 1979'da Yale Üniversitesinde DeJong'un metinler üzerinden arama işlemi için kurduğu sistemde, metinlerde geçen yapılar script (bugün kullanılan ve küçük program parçası olan betiklerden farklı bir anlamda) olarak tanımlanmış, örneğin metinlerde geçen film yönetmenleri kategorisi bir script olarak ele alınarak bu kategoriye girecek değerler bulunmaya çalışılmıştır. Cowie

ve Wilks, bu scriptleri template (şablon) olarak adlandırmaktadır (Cowie ve Wilks, 2000: 241,244).

Genel olarak bilgi çıkarımı teknikleri, DDİ yaklaşımlarını içeren ve metnin bütün bileşenlerinin dil bilimsel olarak çözümlemesini içermektedir. Bir metinde geçen özel adlar, yer adları ve daha üst düzeyde anlamsal ilişkilerin çözümlenmesi başta bütün metin birimlerinin (sayılar, karakter grupları, noktalama yapısı, paragraflar) ayrımlanması ve ayrımlanan birimlerin sıklıklarının çıkarılması gerekmektedir. Böylelikle metin, birimlerden bütüne doğru giden bir yaklaşımla ele alınmış olur.

Açık kaynak kodlu olarak kullanıma sunulan ve metin madenciliği uygulamalarının geliştirilmesi için grafik arayüzlü bir ortam sunan GATE (General Architecture for Text Engineering), Java diline dayanan modülleri kullanarak basit ve karmaşık DDİ uygulamalarının yazılmasını, test edilmesini amaçlayan bir yazılımdır. Yazılım Sheffield Üniversitesince desteklenmekte ve bu yazılıma dayanan birçok proje hayata geçirilmektedir.



Şekil 35. GATE 5.1 Uygulama Geliştirme Ortamı.

1.4.3.4. Belge Sınıflama

İnternetin ortaya çıkardığı elektronik metinler ve birbiriyle bağlantılı sayfalar ağı her an güncellenmekte ve büyümektedir. Toplam sayfa sayısı 20 milyar 54 milyon sayfa olarak verilmektedir.[14] 2001'de BrightPlanet adlı şirketin yaptığı araştırmaya göre 550 milyar belge bulunmakta bu belgelerden yalnızca 1 milyarı indekslenmiş bulunmaktadır.[15] Arama motorlarının indeksleyemediği sayfalar görünmez ağ sayfaları (invisible web) olarak adlandırılmakta, bu sayfaların sayıca çokluğu karşısında günümüz arama motorlarının yetersizliği vurgulanmaktadır.

Milyarlarca sayıda belgenin konusuna göre sıralanması ve sınıflanması işi belgelerin otomatik sınıflanmasının başlıca konusudur. Sorgusu yapılan terimin konusuyla ilgili bir başka deyişle konu-duyarlı ortamının otomatik yapılması günümüz bilgi teknolojisi çalışmalarının önemli bir alanı durumundadır. Yukarıda verilen sayfa sayılarının elle taranması olanaksız olduğundan, çözüm olarak hızlı ve başarım oranı yüksek uygulamaların geliştirilmesinden yola çıkılarak birçok teknik yaklaşım ortaya sürülmüştür. Belgelerin bir sorguya bağlı olarak bir sonuç kümesine atanması ve gösterilmesi günümüz internet arama motorlarının ana çalışma prensibini de oluşturmaktadır. Arama ve bulma işi için sorgu kutucuğuna yazılan her bir karakter arka planda potansiyel olarak ilişkilendirilmeyi bekleyen devasa bir belge yığınına ulaşmaktadır. Bir sorgunun başarım ölçütü, sorguyla ilişkili terimin eşleştirildiği belgeyle olan konu-duyarlılığıdır. Örneğin bir tıp veri tabanı olan PubMed'de sorgulama yapan bir kullanıcı, dopamin terimi için beyinle ilgili bir belgeye ulaşabiliyorsa ve daha özel olarak beynin çeşitli anatomik yapılarında dopamine özgül sınıflara bu sorgu terimi atanabiliyorsa, sistem oldukça başarılıdır denebilir.

Çevrimiçi belge sınıflamada öncelikle internet sayfalarının barındırdığı üst verileri (meta data) ve sayfalarda yer alan yapı tanımlayıcı birimleri (etiketleri) normal karakterlerden ayırma işlemi uygulanmaktadır. Böylece salt doğal dil görünümü olan sözler ve cümleler nihayet paragraflar ortaya çıkarılarak kabaca metin yapıları işlenmek üzere sisteme girdi olarak hazırlanır. Üst veriler, sayfanın oluşturucusu, kullanılan karakter kodlaması, oluşturulan yazılım ortamı, anahtar sözcükler gibi ana metin gövdesinin dışında, belgenin başlık bölümünde tanımlanan verilerdir. Amaç, veri ortamı için gürültü olarak tanımlanan bu birimlerin ayıklanması ve değerli bilgi içeren metin

[14] http://www.worldwidewebsize.com/
[15] http://www.powerhomebiz.com/vol25/invisible.htm

gövdeleriyle sorgu girdilerinin ilişkilendirilmesidir. Bu işlem yapılmadığında istenilen terim-bilgi çiftinin sonuçları yeni bir gürültü ortamı oluşturmaktadır.

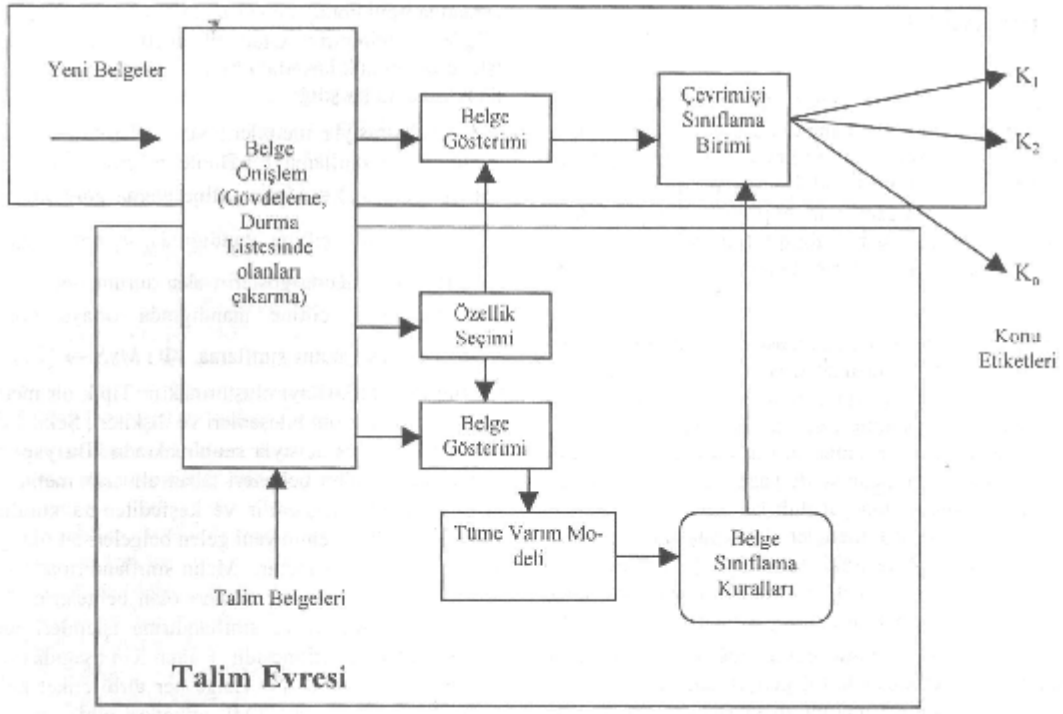
Sever ve Bolat (2006), işlem yapılacak metinleri ve sınıfları M ve S olarak verdikten sonra, boole değeri olarak adlandırılan doğru ve yanlış çiftinden teste tutulan her bir birimin atamasını $<m_i, s_j> \in$ MxS formülüyle gösterdikten sonra, D değerinin m_i nolu metnin s_j sınıfına girdiğini, Y değerininse tersi durumda atandığını belirtirler. Metnin gürültülerden arındırılmasında çok sık geçen ve tek başına anlamı olmayan sözler (edatlar ve bağlaçlar) dur listelerine alınarak işlemede etkileri giderilir, ayrıca çekimli sözlerin yapım ekleri kalacak biçimde çekim eklerinden ayrılarak gövdelemeleri yapılır. Metinde anlamlı birimler "terimler" ya da "özellikler" olarak adlandırılırlar. Sever ve Bolat'a göre metin sınıflamada klasik olarak boole, olasılıksal ve vektör tabanlı yöntemler arasında en çok kullanılanı vektör tabanlı yöntemdir. Bu yöntemde belgede yer alan terimler (özellikler) terimler vektörü olarak gösterilir. Terimin bir derlemde ya da belgelerdeki sıklığı o terimin ağırlığını vermektedir. Terimlerin ağırlık durumları için iki önemli özellik belirtilir:

1- Bir terim bir belgede ne kadar sık geçerse ağırlı o kadar artar.

2- Bir terim farklı belge türlerinde ne kadar çok bulunursa, o terimin anlam ayırt edici özelliği o kadar düşer.

Belgede geçme sıklığına terim sıklığı, farklı tür belgelerdeki terim sıklığına ise ters belge sıklığı adı verilmektedir. Bu durumda bir terimin ağırlığını Sever ve Bolat tf x df olarak gösterirler.

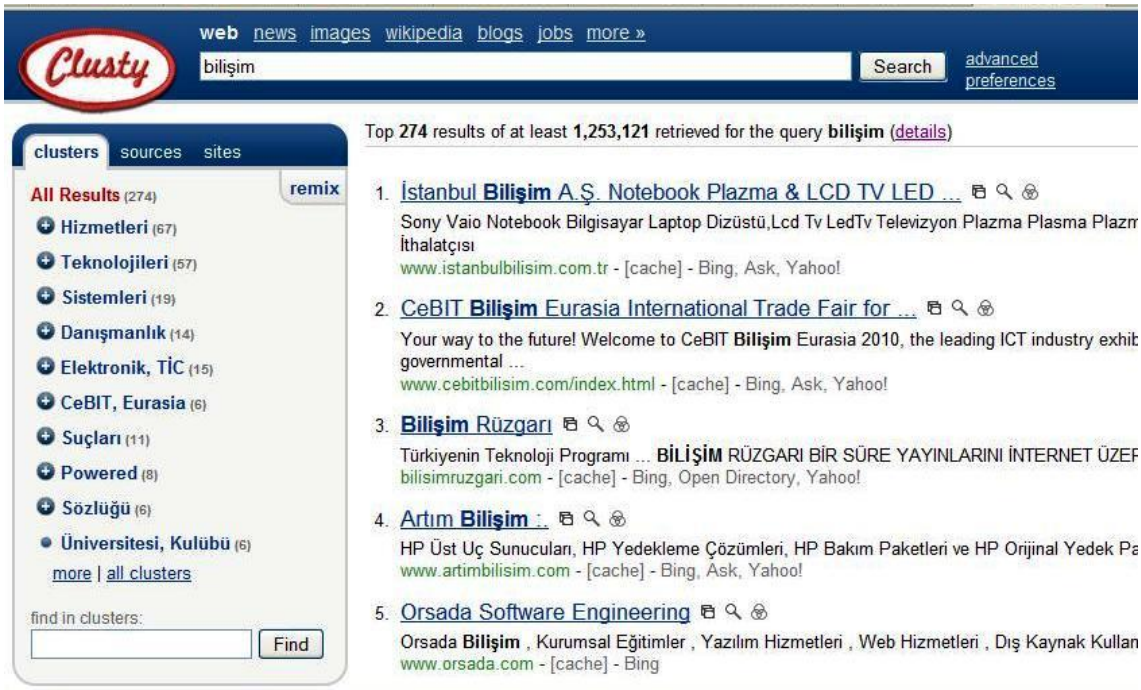
Aşağıdaki şekilde bir belge sınıflama işleminin aşamaları görülmektedir.



Şekil 36. Belge Sınıflama Aşamaları (Sever ve Bolat, 2006: 174).

Şekilde yer alan belge gösterimi aşaması terim-belge çiftinin gösteriminde hangi modellerin kullanılacağının belirlendiği aşamadır. Özellik seçimi belgeyi temsil edecek terimlerin sıklıklarının çıkarılması ve buna dayanan terimler dizgesinin belirlenmesi aşamasıdır. Sonuç olarak da K etiketleriyle gösterilen çıktılar, terimlerin ait olabileceği konuları içermektedir (Sever ve Bolat, 2006: 173, 174).

Belge sınıflama uygulaması olarak internet üzerinde sıkça kullanılan clusty adlı arama motoru, sorguları konularına göre ayırarak ilgili konuya bağlantılar içermektedir. Aşağıdaki şekilde clusty'den alınan "bilişim" sözcüğünün konulara göre sayfalandırma sonuçları yer almaktadır.



Şekil 37. Clusty Arama Motoru Sonuç Sayfası.

Sorgu sonuçları arasında "bilişim" sözcüğü ile sıklık örüntüsü yüksek olduğu anlaşılan "hizmetleri", "teknolojileri" ve "sistemleri" sözleri, bir eşdizimlilik görüntüsü oluşturarak konu sınıflamasında yer almıştır.

0.4.3.5. Varlık Adı Tanıma

Genel olarak varlık adı tanıma olarak çevrilebilecek bu kavram için Diri ve Amasyalı[16] "özel anlamlı kelimeleri belirleme" karşılığını kullanmıştır. İngilizcede *named entity recognition* terimi NER olarak kısaltılmaktadır. Çevrimiçi olarak internet sayfaları ve çevrimdışı ortam olarak bilgisayarda tutulan metinlerde genel olarak özel adların (kişi, yer, kurum) ve özel ad dışında kullanılan sayı adlarının, zaman ifadelerinin ve para birimlerinin otomatik belirlenmesi ve listelenmesi işi olarak tanımlanan *varlık adı tanıma* (VAT) bilgi çıkarımının alt alanı olarak görülmekle birlikte DDİ teknikleriyle de bir kavşak noktası oluşturmaktadır.

VAT sistemleri, üretilen yazılımların değerlendirildiği bilgi çıkarımı yarışması olan Mesaj Anlama Konferanslarında (Message Understanding Conference-MUC) önemli bir değerlendirme alanı olarak görülmektedir. "Named entity" terimi ilk kez 1996'da yapılan 6. Mesaj Anlama Konferansı'nda ortaya atılmıştır.

[16] http://www.ce.yildiz.edu.tr/mygetfile.php?id=2657

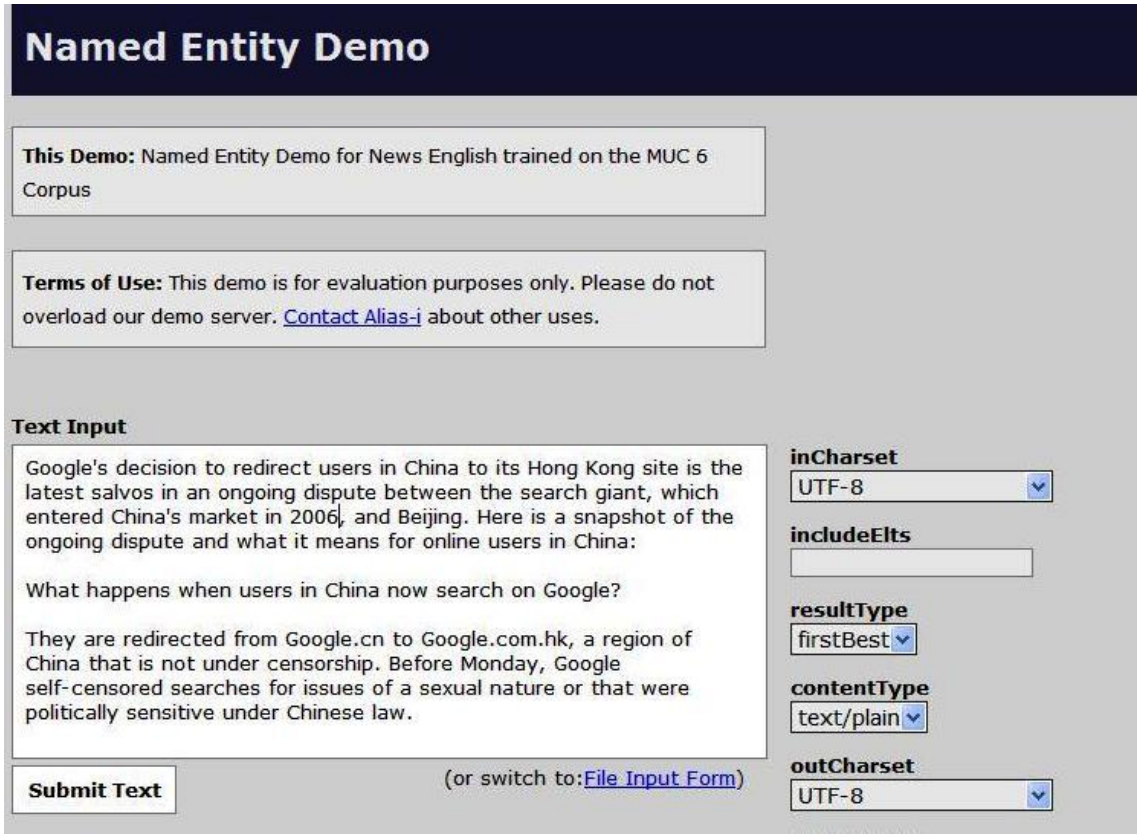
Çeşitli diller için VAT sistemleri çalışması yapılmaktadır. İngilizce için hazırlanan VAT sistemlerinin başarım oranı %90'lara varmakta, Türkçe için bir tam başarımlı VAT sistemi çalışmaları ise henüz yeni başlamaktadır. Türkçe için yapılan istatistiksel cümle bölme, konu bölümleme çalışmalarında istatistiksel yöntem, Türkçenin üretken ek yapısı dolayısıyla doğrudan uygulanamadığı sonucuna varılırken istatistiksel yöntemin bilgi dağınıklığı sorununa yol açtığı vurgulanmaktadır (Küçük ve Yazıcı, 2009: 525)[17].

Türkçe haber metinleri üzerinde, Küçük ve Yazıcı kural tabanlı bir yaklaşımla VAT uygulaması geliştirmişlerdir. Bu uygulamanın kaynakları olarak elle derlenmiş söz listeleri ve kural yapıları kullanılmıştır. Haber metinleri dışında, hazırlanan uygulama çocuk hikâyeleri, tarihsel metinler ve haber videolarının metinleri üzerinde de denenmiştir. Böylece uygulamanın değişik metin türleri üzerindeki başarım oranı belirlenerek, bir alan için uygulanan kuralların diğer alanlar için de uygulanıp uygulanmayacağı denenmiş olmaktadır. Uygulamanın yazarları kural tabanlı yaklaşımından elde edilecek çıkarımların makine öğrenmesi gibi daha karmaşık yöntemlere girdi sağlayacağını da belirtmişlerdir. Makine öğrenmesi yönteminin, önceden hazırlanmış bir işaretlenmiş veri setine gereksinim duyması ve işaretlemenin zaman alması bu yöntemin en büyük dezavantajı olarak görülmektedir. Uygulama için 8300 kişi adını içeren bir kişi adları sözlüğü, adları iyi bilenen ünlü kişiler listesi, en çok bilinen yer adları ve en çok bilinen kurum ve kuruluş adları listesi oluşturularak sistemin bu adlara göre belirli bir yapı içinde aranmaları amaçlanmıştır. Kural yapısında X yolu/Kulesi/Stadyumu…gibi bir arada bulunabilecek kalıplar kural olarak belirlenmiş, bu yapıda X yerine geçecek olanlar metinden sözlük listelerine bakılarak belirlenmeye çalışılmıştır. Burada X için belirlenecek olan metin birimi daha önceden belirlenen listeden bir birimse X ve yapıda X'ten sonra yer alan liste elemanı bir yer adı birlikteliği oluşturmakta, liste dışı bir birim atandığında da X silinip yerine yeni bir birliktelik oluşturularak yer adı tanımlanmaktadır. Bu çözümlemede birden çok birimli sokak adlarında Türkçede sıkça kullanılan Abdullah Cevdet örneği verilerek, bu birimde X'in yerine geçecek olan birim birden çok sözcükten oluşmamışsa doğrudan atanarak X'i silme, birden çok sözcükten oluşmamışsa X ile birlikte bir birliktelik oluşturma yolu seçilmiştir. Sonuç olarak kural tabanlı yöntemin çalışmada genel başarım oranı haber

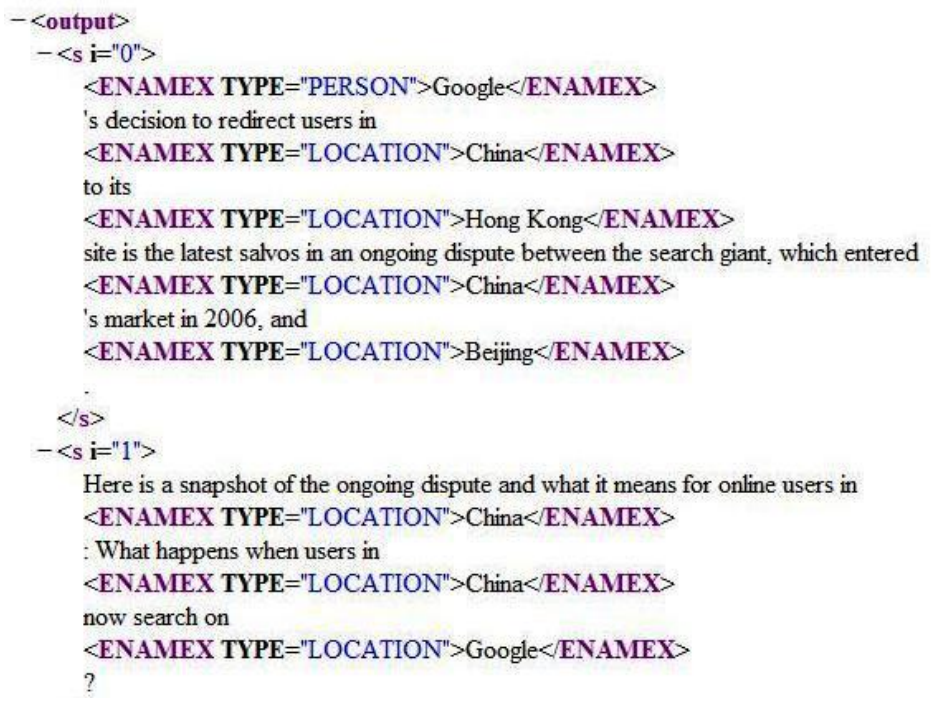
[17] http://www.springerlink.com/content/l831084713048233/fulltext.pdf

metinleri için %75, çocuk hikayeleri için % 70 ve tarihsel metinler için yaklaşık %55 olarak ölçülmüştür (Küçük ve Yazıcı, 2009: 526, 527,530,532).

İngilizce için hazırlanmış Java diliyle yazılmış bir çevrimiçi uygulama olarak Lingpipe sistemi genel olarak DDİ araçları sunmakla birlikte, VAT için de ayrı bir modüle sahiptir. İnternet üzerinden girilen herhangi bir metnin özel adları ile ilgili çözümlemesini XML biçimli çıktı olarak vermektedir. Aşağıdaki şekillerde Lingpipie'ın metin giriş ve örnek metin çözümlemesi görülmektedir.



Şekil 38. Lingpipe Metin Giriş Formu (http://lingpipe-demos.com:8080/lingpipe-demos/ne_en_news_muc6/textInput.html).



Şekil 39. Lingpipe VAT İçin XML Çıktısı (http://lingpipe-demos.com:8080/lingpipe-demos/ne_en_news_muc6/response.xml).

Yukarıdaki şekilde, s olarak etiketlenen birimler cümleler olup, yer adları otomatik olarak işaretlenmiştir.

1.4.3.6. Özetleme

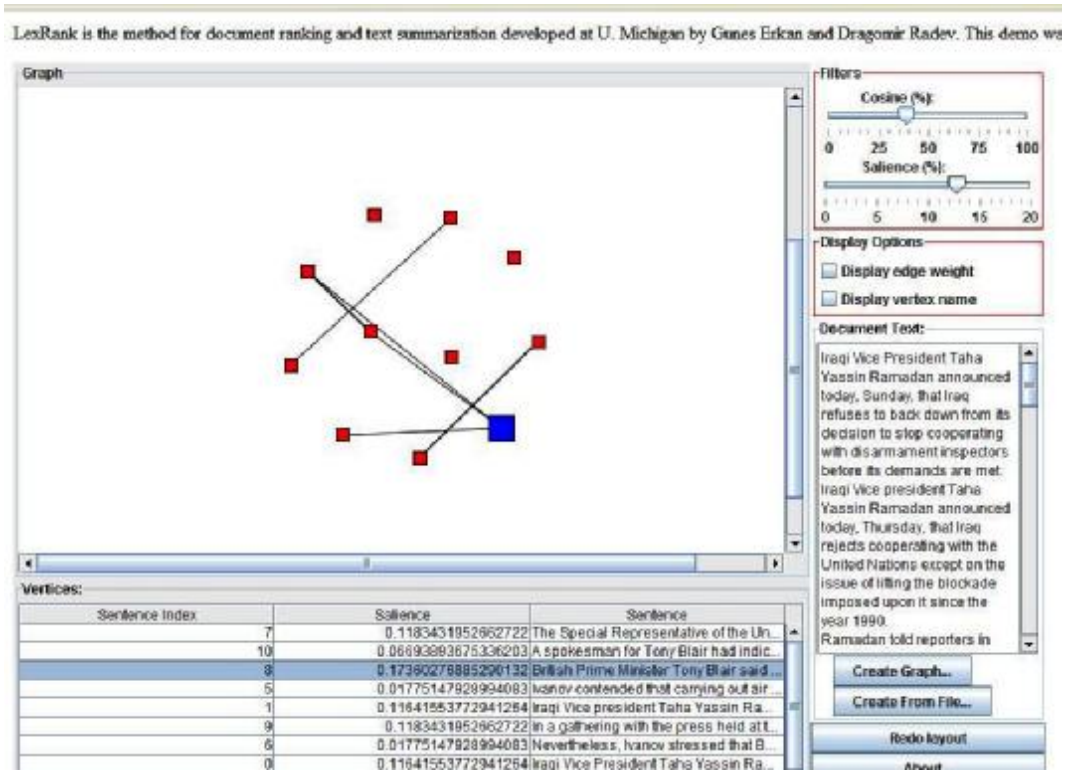
Özetleme, genel olarak bir belgenin ya da belgeler topluluğunun ana düşünceleri sabit tutularak ve ona bağlı olarak daha kısa bileşenleri ile temsil edilmesi biçiminde tanımlanabilir. Elektronik metinlerin ve bunlardan oluşan belge koleksiyonlarının sayısının üssel biçimde artması beraberinde bilgiye ve içeriğe kısa yoldan ulaşma gereksinimlerini de ortaya çıkarmıştır. Bir arama motorunda gereksinim duyduğu bilgiye ulaşmaya çalışan kullanıcı, sorgusuyla ilgili belgeleri istediği hızda listeler biçiminde karşısında bulurken aynı zamanda araştırdığı konuyla örtüşen belgeyi kısa bir sürede okuyup değerlendirmek de istemektedir. Bu noktada durağan (statik) metinler yerine internet metinlerinin sorguyla olan ilgisinin zayıflatılmadan hızlı bir biçimde ana hatlarının çıkarılması, ana düşüncesiyle olan ilgisinin azaltılmadan cümleler ya da daha büyük birimlerle (paragraflar) temsil edilmesi gerekmektedir. Günümüz teknolojileri, metinlerin yorumlanması açısından, en kısa sürede temsil gücü yüksek güçlü ana düşüncelerin ortaya çıkarılması yönünde işlemeye başlamışlardır.

1950'li yıllardan bu yana özetlemeyle ilgili bir çok yöntem geliştirilmekle birlikte temelde iki yöntem öne çıkmaktadır: *Cümle çıkarımı* (sentence extraction) ve *soyutlama* (abstraction). Cümle çıkarımı, metinde geçen sözlerin sayısını ve bunların örüntülerini esas alırken, soyutlama işlemi daha karmaşık olarak cümlelerin ayrıştırılmasından sonra cümleler arası ilişkilere dayanan yeni bir yapının kurgulanmasını gerektirir. Öğrenme temelli özetleme tekniği yanında graflara dayanan özetleme, olasılıksal hesaplamalara dayanan özetleme, söz zincirleri ve ilişkili anlamlarına dayalı özetleme olarak adlandırılan teknikler kullanılmaktadır. Makine öğrenmesine dayanan yöntemde denetimli ve denetimsiz öğrenme olarak ikiye ayrılan teknikler kullanılır. Araştırmacının önceden hazırladığı alıştırma veri setleri makineye girdi olarak verilerek çıktılar bir değerlendirme ölçeğine göre sıralanırlar. Bir metin özetinde özetleme yapılabileceği gibi bir konuyla ilgili birçok belge kümesi de özetlemenin içinde yer alabilir (Kutlu vd., 2010: 2).

Kutlu ve arkadaşları (2010: 1-2), *anahtar ifade* (key phrase) olarak adlandırılan birimleri belgelerin yoğunlaştırılmış ve kısaltılmış biçimleri olarak tanımlamışlar, anahtar ifadelerin bir metni anlamsal olarak temsil eden ifade setleri olarak nitelendirmişlerdir. Türkçe için geliştirdikleri özetleme uygulamasında anahtar ifadeleri, cümlelerin içinde yer alan ve onların özelliklerini (features) gösteren birimler biçiminde ele alarak bu ifadeleri sayısal ağırlıklarına göre sınıflandırmışlardır. Anahtar ifadeler dışında, cümlelerin özellikleri kapsamında değerlendirilen terim sıklıkları, yüzeysel düzey cümle özellikleri (surface-level sentence features) biçiminde adlandırılan tekniğin birimleri olarak atanmışlardır. Uygulamanın asıl amacı, metinde yer alan cümlelerin, makine öğrenmesiyle beslenen bir skor işlevi yardımıyla atanan sayısal değerlere göre en yüksek değerli cümleden başlayarak sıralanması olarak belirtilmiştir. Türkçe metin özetleme çalışması olarak yapılan bu uygulama ilk olma özelliğiyle birtakım zorlukları da beraberinde getirmiştir. Genel olarak DDİ için gerekli olan ve makinece işlenebilir nitelikteki doğal dil kaynaklarının (sözlük veri tabanı, ayrımlanmış birimler, işaretlenmiş metinler vb.) yeterince olmaması, veriye dayanan sistemler için oldukça sıkıntılı durumlar oluşturmaktadır. Bu çalışma için araştırmacıların kendilerinin derlediği 120 adet haber metni kullanılmış, başarım ölçütü olarak Recall-oriented understudy for Gisting evaluatin (ROUGE) adı verilen ölçüt kullanılarak sonuçlar insan merkezli özetlemeye bakılarak değerlendirilmiştir.

Çevrimiçi otomatik metin özetlemede LexRank adlı sistem, verilen metindeki cümleleri benzerliklerine göre çözümleyerek sonuç üretmektedir. Sonuçlar graflarlala

gösterilirken metinde geçen her bir cümle bir diğeriyle bağlantılı olmakta, uygulamada filtreleme aracıyla özetlemenin duyarlılığı artırılarak metni ifade edecek en belirgin cümleler gösterilmektedir. Çıkarımsal (extractive) olarak adlandırılan yöntemde, metin birimi olarak cümleler temel alınmakta, değeri yüksek cümleler seçilmektedir. En önemli cümleler taşıdıkları terimler bakımından bir merkezi (centroid) yalancı cümleyle (pseudo-sentence) olan ilgisine göre belirlenmektedir. Erkan ve Radev (2004), matematikte kullanılan ve "eigenvector" adı verilen vektörel tabanlı bir matris sistemiyle cümlelerin benzerliklerini bulmaya çalışmışlardır. Erkan ve Radev, bu çalışmadan önce yapılan çalışmalarda genel olarak cümlelerde geçen sözlerin sıklıkları ve *ters belge sıklığı* (inverse document frequency) terimiyle açıklanan, bir sözün çeşitli türdeki belge koleksiyonlarında bulunma sıklığına bakılarak yapılan benzerlik sınıflamasından söz etmektedirler. Ayrıca, cümlede geçen anahtar sözlerin sıklığı da bu tür özetleme çalışmalarının ana çıkış noktası olarak belirtilmektedir[18].



Şekil 40. LexRank Uygulaması [19]

[18] http://www.jair.org/media/1523/live-1523-2354-jair.pdf
[19] http://clair.si.umich.edu/clair/lexrank/

0.4.3.7. Soru Yanıtlama

Bilgisayar sistemlerinin doğal dilde üretilmiş bir soruyu istenilen doğrulukta yanıtlaması, bu çalışma alanının temel hedefidir. Günümüz internet arama motorlarının çeşitli bilgi çıkarımı tekniklerini kullanarak dizinledikleri veri tabanlarında sorguya karşılık gelen birim ya da ilişkili birimleri çıkarmaları temel çalışma prensipleridir. Oysa "Bugün hava nasıl olacak?" ya da "Dünyanın en büyük binası nerededir?" gibi doğrudan istenilen bilgiye yönelik, hedefleyici soruları yanıtlayacak bir sistemin geliştirilmesi, internet gibi karmaşık bir bilgi ağında aranılanın bulunması açısından giderek daha önemli bir konuma gelmektedir.

Doğal dil sorusunun girdi olarak uygulamaya alınmasından sorunun çıktısına karşılık gelen hedef bilginin bulunması aşaması neredeyse DDİ'nin bütün alanlarının devreye girmesini gerektirmektedir. Öncelikli olarak sorunun temel düzeyde sözcüklerine ayrılması, her bir sözcüğün diğer sözcükle kurduğu söz türüne dayalı ilişkilerin belirlenmesi ve bu ilişkilerin oluşturduğu cümle tiplerinin hangi anlamlarda kullanıldığı, hangi paragraf tiplerine gönderme yaptığı ve sonuçta bu parça birimlerin hangi alanla ilgili metinler kategorisinde değerlendirileceği saptanmalıdır. Bilgi çıkarımı ve bilgiye erişim yöntemlerinin belgelerin sınıflanmasında kullanılması ve günümüzde anlam bilimsel bir ağ oluşturmak için kullanımı giderek artan ontolojilerin hazırlanması, soru yanıtlamada başarımı ve etkinliği artıran çözümlerdir.

Sayıt ve Karaoğlan (2007), soru yanıtlama sistemlerinde yanıtlanacak soru tiplerinin gerçek tabanlı sorular (factoid questions) ve listeleme soruları (list questions) olarak iki grupta ele alınabileceğini belirterek gerçek tabanlı sorulara "Türkiye'nin başkenti neresidir?", listeleme sorularına da "Türkiye'nin bölgelerini listeleyin" örneklerini vermişlerdir. TREC (Text REtrieval Conference) adlı konfreranslarda soru yanıtlama sistemleri değerlendirilip, sistemlerin başarımlarını ölçecek standartlar belirlenmeye çalışılmaktadır. TREC'te yukarıda verilen soru tipleri dışında yer alan sorular "diğer" sorular sınıfına alınmaktadır. Bu sorular, sistemlerin geliştirilmesi sırasında yanıtları alınamayan sorular olup yanıtla ilgili ek bilgiler içerirler[20].

Zheng (2002), otomatik soru yanıtlama sistemlerinin oluşturulmasıyla ilgili çalışmaların 1960'lı yıllara kadar gidebileceğini belirtirken, çalışmaların genellikle uzman sistemler (expert systems) alanında konuya özgü metinler üzerinde yapıldığını dile getirmektedir. 1960'lı yıllardan günümüze değişen metin ortamlarıyla birlikte

[20] http://ab.org.tr/ab07/bildiri/88.pdf

elektronik belge yayıncılığıyla geniş belge koleksiyonlarına dayanan soru yanıtlama sistemleri oluşturulmaya başlanmıştır. Bu sistemler arasında AnswerBus, AQUA, Webclopedia, PİQASso, NSIR, START, LCC, QuASM ve IONAUT öne çıkanlarıdır. Zheng bütün bu sistemlerin dezavantajları olarak soru yanıtlama sürelerinin uzun olmasını göstermektedir. AnswerBus sistemi alana bağlı olmayan (open-domain), İngilizce, Almanca, Fransızca, İspanyolca, İtalyanca ve Portekizce soruları çeşitli arama motorlarına bağlı olarak yanıtlamaya bağlı bir sistemdir. 200 soruluk bir sette başarım oranı %70.5 olarak verilmektedir. Temel olarak İngilizce çalışma ara dili olarak kullanılmaktadır. Sorunun öncelikle dil tanıma modülüyle hangi dile ait olduğu belirlenir, sonra Altavista arama motorunun BabelFish adlı aracı kullanılarak İngilizce olmayan sorular tekrar İngilizceye çevrilerek işlenmeye başlanır. Arama motorlarına yönlendirilen sorular, arama motorlarından gelen veriler etrafında soruyla ilgili cümlelere çıkarılarak ve bir bağlam bilgisi eklenerek kullanıcıya sunulur. Bu yapılırken arama motorunda ilgi düzeyi yüksek bağlantılar tercih edilmektedir. Arama motoruna yönlendirilecek sorunun her bir sözcüğü için, motordan dönen sonuçlar sayısal olarak gruplandırılırlar. Sorular, içlerindeki sözcüklerin sıklığına göre bir soru cümleleri dizininde tutularak olası yeni sorular için veri tabanı oluşturulmuş olur[21].



Şekil 41. AnswerBus Çevrimiçi Soru Yanıtlama Sistemi Ana sayfası (http://www.answerbus.com/index.shtml)

[21] http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=79B7429763EAD8645508905697892991?doi=10.1.1.11.9666&rep=rep1&type=pdf

1.5. Sözlük Biliminde Bilgisayarlı Yaklaşımlar

1.5.1. Bilgisayarlı Sözlük Bilimi ve Bilgisayar Destekli Sözlük Bilimi Terimleri

DDİ'nin bilgisayarlı dil bilimiyle ortak çalışma yönlerinin olmasının yanında iki alan arasında mühendislik yaklaşımı ve dil bilimi yaklaşımı ayrımının olduğu belirtilmişti. Aynı biçimde bilgisayarlı sözlük bilimi ve bilgisayar destekli sözlük bilimi arasındaki farkı da belirtmek gerekmektedir

Yüksekkaya (2001: 218) bilgisayar destekli sözlükçülük ve bilgisayar sözlükçülüğü terimlerini kullanarak bu iki terimin içeriğindeki farka dikkat çeker. Bilgisayar destekli sözlükçülükteki temel amacın, sözlük hazırlamada bilgisayardan yararlanmak olduğunu belirten Yüksekkaya, bilgisayarın, sözlük yazım aşamalarının hızlandırılmasında kolaylık sağlandığını belirtir. CD-ROM formatına getirilen sözlükler kullanıcıya *çevrim dışı* (off-line) olarak sunulur. Bu çalışmalarda sözlük verisinin bilgisayarca okunur biçime dönüştürülmesi de amaçlanır.

Ooi, Lenders'in (1991:48) sözlüğü dilin sözvarlığı bilgisini taşıyan ya da belirli bir alanın terimlerini içeren "uzman sistem" olarak ele aldığını belirterek DDİ sistemlerinin bu uzman sistemlere sıklıkla başvurduğundan söz eder. Temelde üç amaç bilgisayarlı sözlük bilimi çalışmalarını özetler: 1- Sözlük oluşturumunu otomatikleştirmek, bunun içinde çeşitli sözlük yazılımları geliştirmek. 2- Bilgisayarın kullanabileceği sözlükleri hazırlamak. 3- İnsanların kullanacağı sözlükleri hazırlamak (Ooi, 1998: 30).

1.5.2. Sözlük Birimi Elde Etme

1.5.2.1. Klasik Yöntemle Elde Etme

Ooi'ye göre (a.g.e.:47), sözlüğün hazırlanmasında ve sözlüksel birimlerin elde edilmesinde üç yöntem kullanılır. Yazarın sözcüksel iç gözlem (lexical introspection) adını verdiği yöntemde, sözlük bilimcinin sezgisine dayalı olarak veriler süzülür, derlenir ve sunulur. Bu yöntemin güvenirliği her zaman tartışmalıdır. Ooi bu yönteme masa başı sözlükçülük (armchair lexicography) adını da vermektedir. İç gözlem verileri durağan verilerdir, değiştirilmesi ve güncellenmesi için çok yoğun bir çalışmayı gerektirir. İkinci yöntem gündelik alıntılama (casual citation) yöntemidir. Burada, bir kişinin içinde bulunduğu ortamdan yaptığı rast gele sözcük alıntılamaları söz konudur.

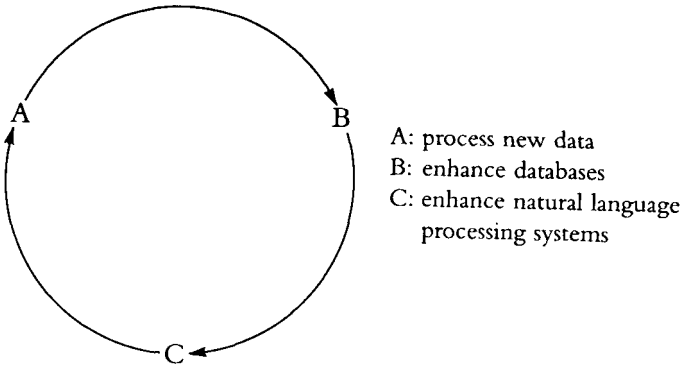
Alıntılanan sözlüksel birimlerin hangi bağlamdan nasıl ne kadar güvenirlikte elde edildiği oldukça tartışmalıdır. Üçüncü yöntem tamamen gözleme dayanan derlem tabanlı yöntemdir. Gündelik alıntılama da gözleme dayanmakla birlikte derleme dayalı yöntem özel olarak seçilmiş ve tutarlı biçimde bir araya getirilmiş verilerden oluştuğundan güvenirliği yüksektir. Ooi, bu üç yöntemin bir arada kullanılmasının kaliteli ve güvenilir sözlüklerin oluşumu için kullanılması gerektiğini savunmuştur.

Sözlüklerin yapılandırılmasında, madde başı birimlerin kaydedilmesinde kullanılan ilk yöntem elle hazırlama (manuel) yöntemidir. Emek yoğun bir işlem olması ve çalışmanın uzun sürmesi; günümüzde gelişen teknolojilerle birlikte bu yöntem, yerini veri tabanı yazılımlarına, çevrim içi sözlükçülüğe bırakmaya başlamıştır. Elle hazırlama yöntemi bilgisayarlara girdi sağlamada 1980'li yılların sonunda kullanılmıştır. Carnegie-Mellon Üniversitesindeki makineli çeviri biriminin hazırladığı ve üç modülden oluşan TRANSLATOR, ONTOS ve DIONYSUS adlı sisteme girdiler otomatik olmayan el yöntemiyle girilmiştir. Çok dilli sözlüklerin, ontolojilerin oluşturulduğu sistem aynı zamanda dünya modelleme (world-modelling) denilen bir yapılandırmayı da içerir. Bu yapılandırmada üç ayrı sözlük türü kullanılır. Dünya kavram sözlüğü (world concept lexicon), insanın dünya bilgisiyle ilgili verileri tutar; çözümleme sözlüğü (analysis lexicon) doğal dilde yer alan sözcükleri, söz öbeklerini ve bunların dünya kavram bilgisiyle olan bağlantılarını içerir. Üretim sözlüğü (generation lexicon), doğal dille bağlantılı olan kavram sözlüğünün dizininin yer aldığı bölümdür. Bu sistemde ONTOS yapısı bir tür kavramlar sözlüğüdür. Bir kısıtlı dil gibi dünyayla ilgili genel kavramları içerir. Dünya modeli, çerçeveler (frames) denilen alt yapılardan oluşur. Çerçeveler kavramları temsil eden yuvaları (slots) içerir. Bu sistem günümüzde çerçeve modeliyle oluşturulmuş kavramsal sözlüklerin temelini atmıştır (age., 1998: 74).

0.5.2.2. Derleme Dayalı Yöntemle Elde Etme

Derlemden elde edilecek dilsel birimler *sözcüksel veri tabanı* (lexical database) denilen bir yapıda tutulurlar. Bu yapıda biçim birimsel, söz dizimsel, anlam bilimsel ve kullanım bilimsel ayrımlamalar makinece okunur biçimde tutulurlar. Bu sözcüksel veri tabanı Sinclair'in monitör derlem adını verdiği derlem türü için güncellemeye uygun olması bakımından özellikle değerlidir (age., 1998: 66).

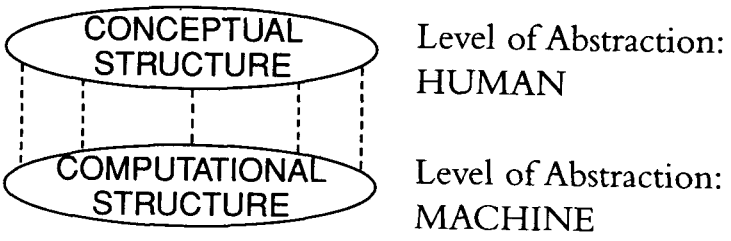
Derleme bağlı sözlüksel birimler çeşitli araçlarla bir dış sözcüksel veri tabanına bağlanabilirler. Bu veri tabanında yukarıda sözü edilen düzeylerde sözcüklerin dil bilgisel özellikleri yer alabilir. Veri tabanı derlem değiştikçe güncellenebilir. Böylelikle sözcük-derlem ilişkisi bir monitör derlem gibi işlev görebilir. Leech, derlem ile derleme dışarıdan eklenecek veri arasındaki ilişkiyi bir döngü biçiminde aşağıdaki şekilde olduğu gibi göstermiştir.



Şekil 42. Derlem, Derleme Eklenen Dış Veri ve DDİ Sistemi İlişkisi (Ooi, 1998: 68).

Şekilde A, derlem için yeni verinin işlenmesini, B var olan veri tabanının yeni verilerle güçlendirilmesini, C ise verilerle DDİ araçlarına daha iyi işlem yapabilecek düzeyde işlenmiş veri akışının sağlanmasını göstermektedir (age., 1998: 68).

Söz dağarcığı (lexicon)[22] adı verilen yapı, makinece okunur formatta tutulan dil bilgisel bilgidir. Ayrıca halihazırda bulunan makinece okunur sözlükler de bir sözlük veritabanı olarak kullanılabilmektedir. Sözlük veri tabanı bir dilbilgisi gösterim konusu olmasının yanında ***bilgi gösterimi*** (knowledge represantation) gibi bilişim alanında kullanılan bir terimdir. Bilgi gösterimi dil bilgisel modellemede de kullanılabilir. Bu gösterim aşağıdaki şekilde verilmiştir.

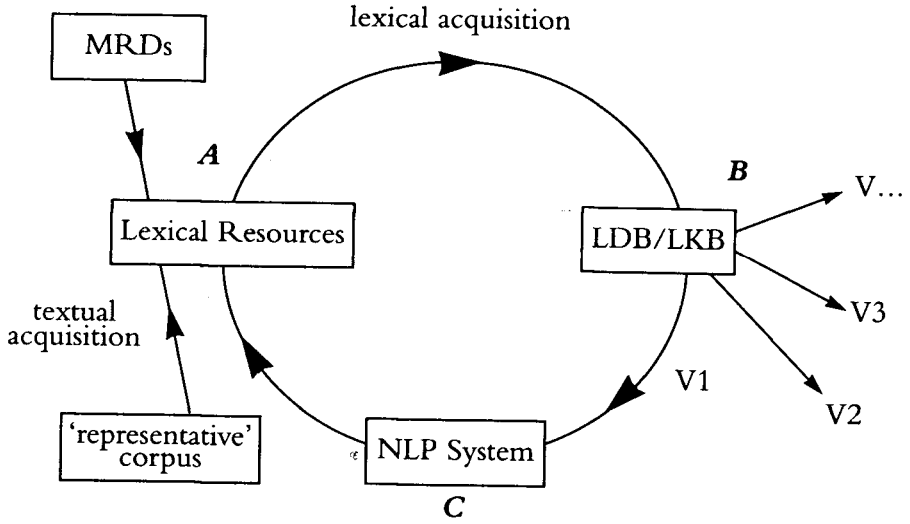


Şekil 43. Soyutlama Düzeyi Gösterimi (age., 1998: 70).

[22] İngilizce lexicon sözcüğü için Yöndem (2005:43) *söz dağarcığı* karşılığını kullanmıştır.

Şekle göre insan ve makine soyutlama düzeylerinde bilgisayar yapısı alt düzeyde bu düzeye bağlı üst düzeyde ise insan tarafından anlaşılabilir yapı kavramsal düzeyde gösterilmiştir (age., 1998: 70).

Ooi'nin sözcüksel veritabanı oluşturma ve diğer kaynaklarla paylaşımı için önerdiği döngü gösterimi de aşağıdaki biçimdedir.



Şekil 44. Sözlük Birimi Elde Etme Sistemi (age., 1998: 72).

Yukarıdaki döngüde A, B ve C ile gösterilen üç ana bölüm vardır. A bölümü sözlüksel veri ya da bilgi adı verilen B bölümüne girdi sağlar. A bölümünde makinece okunur sözlükler (MRD) ve yine makinece okunur metinler (Machine readable texts) yer alır. Bilgisayarca okunur metinler işaretleniş, etiketlenmiş derlemlerdir. Bu derlemin temsil gücünün yüksek olması gerekir. Sözlük birimi elde etme işlemi döngüde yatay okla gösterilmiştir. Elle, yarı otomatik ve otomatik biçimde sözlük birimi elde etme işlemi Ooi'nin sözlüksel taban (lexical data knowledge base) olarak andığı ve V ile gösterdiği sanal sözlükleri oluşturur. Bu sözlükler en sonunda DDİ işlemine girdi sağlayarak sürekli güncellenebilen bir sistem oluştururlar. Döngüde DDİ araçlarının geliştirilmesi için etiketlenmiş derlem ne kadar önemliyse DDİ araçlarının az hata payıyla işlem yapabilme kapasitesinde olması da sözlük birimi elde etme sürecini o kadar otomatik ve kolay gerçekleştirilebilir duruma getirmektedir (age., 1998: 72).

TRANSLATOR, ONTOS ve DIONYSUS gibi üçlü modülden oluşan sözcük işleme sistemi, başarımı yüksek DDİ uygulamaları için sadece sözlük veri tabanları ve ve derlemlerin tek başına yeterli olmayacağını, aynı zamanda etkileşimli bir kavramsal sözlük yapısının da içinde bulunduğu yüksek işlem kapasiteli karmaşık sistemlerin

gerekliliğini ortaya koymuştur. Sözü edilen üç modüllü sistemi hazırlayan Nirenburg ve Raskin, Sözlüksel Yönetim Sistemi (lexical Management System) adını verdikleri içinde çeviri modülünün de yer bir aldığı yapıyı oluşturmuşlardır (age., 1998: 75).

Sözlük birimlerin otomatik olarak elde edilmesi çalışmalarında eşdizimli yapıların çıkarımı önemlidir. Eşdizimli yapıların çıkarımının, hem dil bilimsel incelemelerde, hem de bilgisayar destekli dil bilgisel sistemlerin çalışmasında bilinen bir öge durumuna getirilmesi açısından da önemlidir. Eşdizimsel sözlük birimlerin çıkarımında Smadja'nın (1993: 150) geliştirdiği XTRACT adlı istatistiksel eşdizimli birim çıkarımı yazılımı ilk yazılan uygulamalardandır. Yazılımın İngilizce metinlerden eşdizimli yapıları hatasız olarak çıkarma oranı %80 olarak verilmektedir. XTRACT'in çıkarım algoritmasında span adı verilen tek cümlelik uzamda sözcüklerin +5 ve -5 bir arada bulunabilirlik sınırlarına bakılmıştır. Sözcük çiftlerinin en çok bu 5'lik sağ ve sol mesafe içinde bulunduğu belirtilmiştir. Sistemde eşdizimli yapılar, bir arada bulunabilirlikleri şans eseri olmayan ve sözdizimsel ilişkiler nedeniyle katı (rigid) biçimde bir araya gelmiş birimler arasından seçilmiştir.

Church ve arkadaşları, otomatik sözlük birimi elde etmenin insan kontrolüne gereksinim duyacağını belirterek, kendi kendine öğrenen sistemlerin tam anlamıyla işlevsel olabileceği bir zaman için henüz erken olduğunu dile getirmişlerdir. Kullanılan bütün yöntemleri prematüre olarak değerlendirmektedirler (age., 1998: 81).

Ooi, derlem dil bilimi araştırmalarında Çerçeve Çözümlemesi (Lexical Frame Analysis) adını verdiği bir çözümleme yönergesi hazırlamıştır. Bu yönergeye göre kısaca şunlar yapılmalıdır:

- Sözlükler dilin odak noktasıdır. Dil bilgisel ve diğer sözcüğe ait bilgiler sözlüklerle bağlantılı biçimde bir arada tutulmalıdır.

- Üzerinde uzlaşılmış dil bilgisel kavramlar sözlük biliminde de kullanılmalıdır.

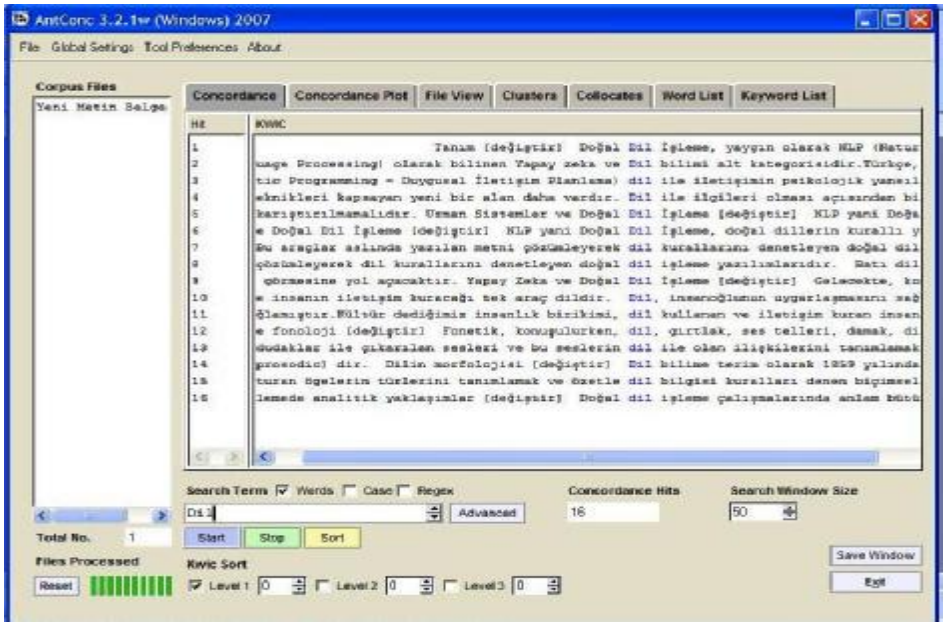
- Sözcük ve gramer bilgisi aynı zamanda istatistiksel bilgi olarak ele alınmalıdır. Sıklık ve eşdizimlilik bilgileri, Z skoru ve diğer istatistik olasılık teknikleri sözlükte yer almalıdır.

- Sözlük bilgisi gözleme dayalı olarak derlemden elde edilmelidir. Makinece okunur sözlükler de bu bilgiyi desteklemelidir (age., 1998: 87).

0.5.3. Bağlamlı Dizinler

"Sözlük bilgisinde, bağlamıyla birlikte sunulan sözcüklerin oluşturduğu abecesel dizge." (Vardar vd., 1998: 35) olarak tanımlanan ***bağlamlı dizin*** (concordance) teriminin Baker ve arkadaşları *key word in context* (KWIC) olarak da adlandırıldığını belirterek, bağlamlı dizninin temelde, bir derlemde aranan sözün sağ ve sol satır bağlamına göre dizim gösterimi olduğunu dile getirmişlerdir. Bu dizinde sözün bütün biçimleri biçim birimsel çözümleme yapılmadan ortaya konur. Bağlamlı dizin oluşturmak için kullanılan yazılımlara genel olarak *concordancer*[23] adı verilmektedir (Baker vd., 2006: 42-44). Bu yazılımlar bir DDİ sistemince anında otomatik işlem yapılacak ve yorumlanacak çıktı üretmezler.

Bağlamlı dizinler paket programlar (AntConc, Concordance, ConcApp, PhraseContext, WordSmith vb.) biçiminde bulunacağı gibi internet tabanlı olarak da hazırlanabilirler.



Şekil 45. AntConc Yazılımı Bağlamlı Dizin Görünümü.

Dizinleme yapacak yazılımların hemen hepsinde desteklenen dosya formatı salt metin (text) olan ve yapısal biçimleme barındırmayan dosyalardır. Etiketlenmesi ve işaretlenmesi henüz yapılmamış bir derlemin salt metinler biçiminde saklanan yapısını, sözcük sayısı ve sıklığının dökümü ile eşdizimlilik görünümlerinin elde edilmesi amacıyla kullanımları yanında, bazı yazılımlar, düzenli ifadeleri destekleyerek

[23] Concordancer terimi için henüz Türkçe karşılık bulunmamaktadır. Terim için "bağlamlayıcı" sözü önerilebilir.

karmaşık ve ileri düzey metin aramalarına olanak sağlarlar. Bir derlem üzerinde var olan söz listeleriyle derlemde kullanılan sözlerin karşılaştırılması amacıyla da kullanılabilecek hazır listeler girdi olarak yazılıma tanıtılabilmektedir. Sıklığı çok yüksek olarak bilinen sözlerin sıklık listeleri yazılıma önceden girdi olarak verilerek bu sözler dışında derlemde geçecek sözlerin listeleri de üretilebilmektedir.

1.5.4. Sıklık Kavramı

Vardar ve arkadaşları ***sıklık*** (frequency) terimini "belli uzunlukta bir konuşma ya da yazıda aynı dilsel olgu ya da birimin gerçekleşme sayısı" biçiminde tanımlamışlardır (Vardar vd., 1998: 183). Sıklık dil bilgisi incelemelerinde değişik amaçlarla kullanılmaktadır. Dil öğretiminde bir derlemden yararlanılarak sıklığı en yüksek sözcükler öğretim stratejisinde ilk olarak öğretilecek ögeler arasında yer alırlar. Günümüzde, dil öğretimi için hazırlanan yabancı dildeki sözlüklerde madde başlarının seçiminde sıklık önemli bir ölçüt olarak karşımıza çıkmaktadır. Sıklığı yüksek birimler, sözlüğün bağlı olduğu derleme göre yanlarında özel işaretlerle belirtilerek öğrencinin bu birimlere daha çok dikkat etmesi gerektiği vurgulanmaktadır.

Bod ve arkadaşları, günümüzde hakim olan dil bilimsel görüşün kategorik dil anlayışı olduğunu, dilin kategorik özellikli yapısının ön plana çıkarıldığını belirterek, dildeki sayıların dolayısıyla sıklık kavramının ve olasılığa veri sağlayan sayısal birimlerin göz ardı edilmiş olduğunu vurgulamışlardır. Üretici dönüşümsel dil bilgisine göre dildeki sayısal gerçekleşmeler kullanımdan kaynaklanan yapay bir durumdur. Dil yetisi, kategoriktir ve düzenli, belirli dil bilgisel sınıflardan oluşmaktadır. Dil yetisinin, kategorik özelliğine, Bod ve arkadaşları son yapılan çalışma sonuçlarının meydan okuduğunu da belirterek ***olasılıklı*** (probabilistic) dil bilimi çalışmalarında ve özellikle ruh dil bilimi araştırmalarında bireyin sözcük ve cümleler üzerindeki yetisinin, bu cümle ve sözcüklerin olasılıklarının bir araya gelmesiyle öngörülebilir bir özellik taşıdığını söylemektedirler. Dolayısıyla dil, olasılıksal bir görünüm de sunmaktadır. Bu açıdan, dildeki ses birimler ve biçim birimler sınıflayıcı bakış açısıyla ele alınırlarken aynı zamanda bu birimler belirli bir devamlılık durumu ve değişim hızına da sahip olabilmektedirler. Olasılıklı dil bilimi, var olan kategorik anlayışa ters düşmekle birlikte, aynı zamanda dile ait bilgimizin de sınırlarını genişletici ve zenginleştirici çalışmalar ortaya koyabilmektedir (Bod vd., 2003: 1).

Yukarıda sözü edilen olasılıklı dil bilimi açısından sıklığı yüksek birimler seyrek birimlere göre daha hızlı bir biçimde tanınabilmekte, birimlerin sadece sayısal özellikleri yanında çok anlamlı birimlerin en sık geçen anlamları da daha hızlı olarak işlenebilmektedir. Sık kullanılan birimlerin diğer özelliği değişime açık olmaları ve özellikle konuşma dilinde kısaltmalara daha çok eğilimli olmalarıdır. Ses bilimsel ve biçim bilimsel açıdan sık kullanılan birimler kendi içinde bir dil bilgisi örüntüsü oluşturabilmektedir. Olasılık kuramında, sıklığı yüksek karmaşık birimler, kendisini oluşturan parçaların sıklıklarının sağladığı veriyle cümle düzeyinde bir bir araya gelme olasılığı oluştururlar. Oluşan bu yüksek birim, tekrar alt birimlerine olasılık kuramına göre bölünebilirler (age. vd., 2003: 3).

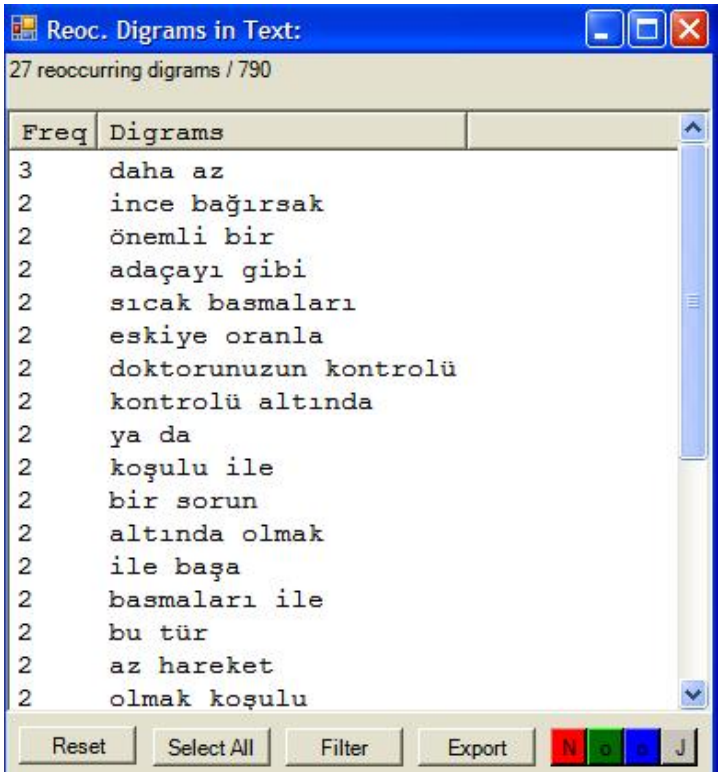
Sıklık, istatistik temelli DDİ uygulamalarının önemli bir veri sağlayıcı ölçütüdür. Olasılık çözümlemelerinde özellikle ikili birimlerin elde edilmesinde, aynı anda bulunabilirlik ölçütünün hesaplanmasında her bir birimin çalışılan derlemdeki tekli sıklık sayıları öncelikli olarak hesaplanıp ikili sıklıkları en yüksek birimler bir araya getirilmektedir. Sıklık tabloları otomatik olarak çıkarılmaktadır. Türkçe için madde başı birimlerin sıklığının hesaplanması, çekimli sözcüklerin çekim eklerinden sonra gövde biçimine getirilmesiyle en sağlıklı sonucu vermesi beklenirken hali hazırda genel bir uygulama aracı olarak bu işlemi yapacak bir uygulama bulunmamaktadır. Şu aşamada çekimli birimlerin sıklığı yazılı metinlerde, sözcük sınırı olarak boşluk karakteri kabul edildiğinden, sıklık bulucu uygulamalar örneğin “ağaç” sözcüğünü ayrı, “ağaçları” sözcüğünün sıklığını ayrı olarak hesaplayabilmektedir. Sözcüğün aldığı her bir ekin biçim birimsel sıklık dağılımıyla birlikte “ağaç” sözcüğüne eklenmesinin sağlanması gerekmektedir. Bu açıdan, özellikle biçem (üslup) çalışmalarında gramatikal dizin oluşturma uygulamalarında Türkçe için bir madde başı oluşturucunun hazırlanması gerekmektedir. Aşağıdaki şekilde bir metinden elde edilen basit bir sıklık tablosu görülmektedir. Şekilde sıklığı belirlenen metnin 606 adet farklı sözcükten oluştuğu görülmekte, sözcüklerin sıklığına göre sıralanmaları çoktan aza, azdan çoğa doğru seçenekli olarak sıralanabilmektedir. Şekilde de görüleceği gibi “çocukların”, “olanlarda”, “basmaları” gibi çekimli biçimler ayrı sayılmış, herhangi bir madde başı oluşturmak için biçimbirimsel çözümleme yapılamamıştır.

Sıklık, otomatik kavram hiyerarşisi oluşturmada ve günümüzde özellikle internet üzerinde eğilim çözümlemelerinin yapılmasında da önde gelen bir ölçüttür. Google’ın trend çözümleyicisi de “en çok” neyin arandığını, istemcilerden gelen verilerin sıklığını hesaplayarak yapmaktadır.



Şekil 46. NooJ Sistemi Tek Birimli Sözcük Sıklığı Tablosu.

Bir metni oluşturan birimlerin tek tek ya da eşdizimlilik örüntüleri gibi çeşitli sıklık görünümleri hesaplanabilmektedir. Kullanılan yazılımın "yeteneğine" göre sıklık aralıkları, ikili üçlü ya da daha çok birimleri kapsayacak biçimde filtreleme özelliği kullanılarak belirlenebilmektedir. Aşağıdaki şekilde, örnek bir eşdizimlilik sıklığı tablosu gösterilmiştir.



Şekil 47. NooJ Sistemi İki Birimli Sözcük Sıklığı Tablosu.

1.5.5. Otomatik Terim Çıkarımı

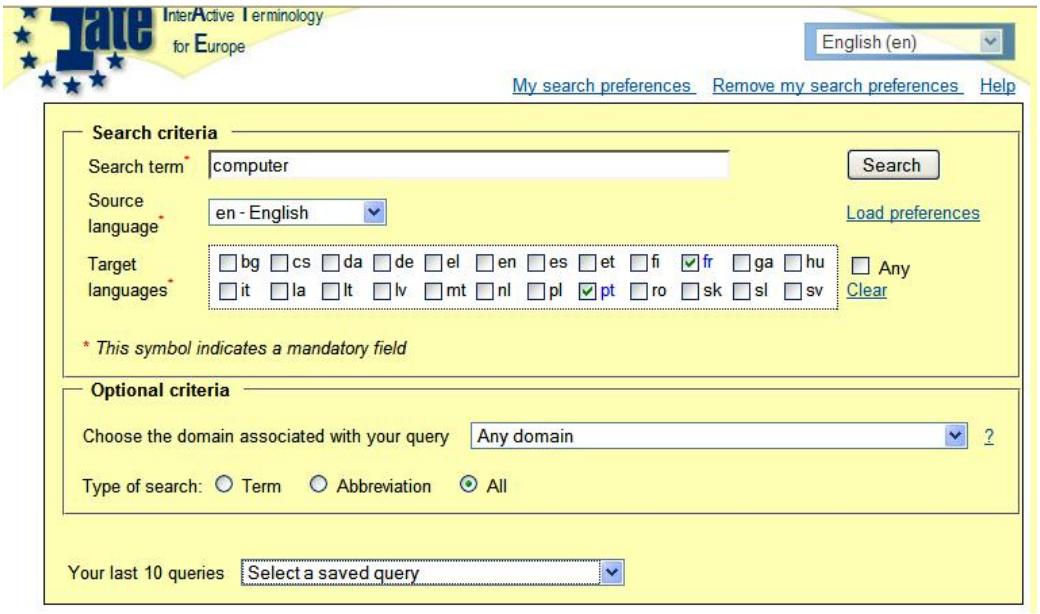
Günümüzde DDİ çalışmalarının ayrı bir alt alanı olarak gelişimini sürdüren *terim çıkarımı*nı (terminology extraction) Baker ve arkadaşları şu biçimde açıklamıştır: Terminoloji elde etme (terminology acquisition) olarak da bilinen terminoloji çıkarımı bir tür bilgi çıkarımı işlemidir. Terim çıkarımı için geliştirilen yazılımlar bir metinden ya da teknik terimleri içeren derlemlerden terim listelerini çıkarırlar. Bir terim bankası oluşturmak için de bu yazılımlar kullanılmaktadır (Baker vd., 2006: 156).

Bir bilim dalının kendine özgü kavramlarını yansıtan terimlerin neden otomatik olarak çıkarımının gerektiği sorusu sorulabilir. Bugün bilinmektedir ki, artan bilimsel yayın sayısı beraberinde bu yayınlara dayalı terim sözlüklerinin de hazırlanmasını gerektirmektedir. Modern sözlük biliminde, sözlük birimlerin belirlenmesi işi artık otomatik yöntemlerle yapılır duruma gelirken, alana özgü (domain spesific) yayınlar olan bilimsel yayınların sözlüklerinin hazırlanması da teknik yöntemlere dayanmaktadır.

Bilimsel belge koleksiyonlarının gerek işaretlenmiş bir derleme dönüştürülmesi gerekse işaretlenmemiş, yapısal olarak karmaşık bir örüntü sunan koleksiyonların ham derlem olarak bilgisayar ortamında bir *terim bankası* (term bank) yapısı oluşturmak

amacıyla tutulmaları gittikçe yaygınlaşmaktadır. Terim bankası, özel bir tip sözlüktür. İçerdiği sözcüklerin ilgili alanıyla ilişkili olduğu, tek dilli ya da çok dilli özellik göstererek otomatik teknik çeviri yazılımlarının temel veri sağlayıcısı olmaktadırlar (Baker vd.:2006: 156).

Terim bankalarının önde gelenlerinden biri Eurodicautom'dur. Bu çevrimiçi terim bankası şu anda 24 faklı dilden çeşitli alanlara ait terimleri barındıran ve Avrupa Birliği'nin desteklediği bir veri tabanıdır. Veri tabanı sorgu arayüzünde sorgusu yapılan sözcük sadece o dile ait terimleri içerebileceği gibi, sorgu sonuçlarının çoklu seçme seçeneğiyle farklı dillerdeki karşılıkları da bilim alanına duyarlı olarak gösterilebilmektedir. Aşağıdaki şekilde Eurodicautom'un sorgu arayüzü görülmektedir:



Şekil 48. Eurodicautom Terim Bankası Sistemi Sorgu Arayüzü.

Cabré, bilgisayar bilimlerinin terminoloji araştırmalarına sağladığı katkıları belirten, bir terimin anlaşılması için onun görsel bağlamının yani terimle birlikte bir görüntüsünün de modern terim bankalarına aktarılmasının, terim araştırmacısına terimi daha iyi çözümleme ve tanımlama fırsatı vereceğini belirtmektedir. Cabré, Sager adlı araştırmacıdan yaptığı alıntıda, terim biliminin bir sanat olmaktan çıkıp, bilgisayarlar aracılığıyla bir tür tekniğe dönüştüğünü dile getirmektedir. Aslında DDİ çalışmalarıyla birlikte dil biliminin de klasik dil bilgisi modelinden gittikçe daha teknik bir modele kaydığı söylenebilir. Daha önce, sıklık kavramının anlatıldığı bölümde de belirtildiği gibi, olasılık çözümlemelerinin dil bilgisel kategoriler üzerine yeniden düşünülmesi

gerektiği, özellikle yazılı metinlere ait otomatik olasılıksal çözümlemenin dayandığı yöntem ve modellerin yeni bir dil bilgisel kategorik sistem ortaya çıkarabileceği bile düşünülebilir. Cabré, bilgisayar aracılığıyla terimlerin belirleneceği belgelerin seçimi, bunların depolanması, sözcük listelerinin çıkarılması, madde başlarının yazımı, madde başlarının anlamlarının kontrolü gibi uygulamaların kolaylıkla yapılabilir duruma geldiğini de belirtmiştir (Cabré, 1999:164,165).

Terimlerin diğer dil bilgisel birimler gibi çeşitli kategorik özelliklerinin bulunması bir yana, alan duyarlı yapıları terimleri bağlamsal olarak daha sınırlı bir kullanımla kısıtlamaktadır. Castellvi ve arkadaşları terimleri bir bilgi alanının merkezi birimleri olarak görerek bunların çıkarımlarında kullanılan uygulamaların da özelleşmiş yeteneklerinin olması gerektiğini belirtmişlerdir. Son yıllarda terimlerin belirlenmesi çalışmalarında derleme bağlı olarak terimlerin anlamsal ilişkilerinin bulunması da amaçlanmaya başlamıştır. Anlamsal ilişki çıkarımı için gerekli olan zengin bir yapısal sözcük ağının varlığı, son yıllarda araştırmacıları ve özellikle terim bilimcileri anlamsal ağ ilişkileri yapılarını modellemeye yöneltmekte, sadece terimlerin birim olarak çıkarımı ikinci plana düşmektedir. Castellvi ve arkadaşları bu noktada, gelinen teknolojinin ve halihazırdaki otomatik terim çıkarıcıların yeterli bir olgunluğa ulaştıklarını ancak anlamsal olarak ilişki modellerinin yapılanmasının hala bir sorun olarak karşımıza çıktığını dile getirmişlerdir. *Ontoloji* (ontology) adı verilen anlamsal ilişkiler ağı, çok anlamlı birimlerde belirsizliklerin giderimi için kullanılırken,, terimlerin belirlenmesi için de gerekli bir araç durumuna gelmiştir. Bir uzmanlık alanına yönelik yayın sayısı arttıkça içerdiği terim yapısı da değişmekte, böylece buna bağlı ontolojilerin de güncellenmesi zorunlu olmaktadır (Castellvi vd., 2007: 1-2,4).

Bilgisayar ortamında geliştirilmiş bir Türkçe terim bankası henüz bulunmamaktadır. TDK'nin geliştirdiği Bilim ve Sanat Terimleri Ana Sözlüğü veri tabanı, TDK'nin yayımladığı çeşitli bilim alanlarındaki terim sözlüklerini içermekte ve 170. 196 terimden oluşmaktadır. Çok dilli madde başlarının sorgularının da yapılabildiği veri tabanı, bütün terimleri ilgili alanla ilişkili olarak tek bir arayüzde sorgulanabilir duruma getirmiştir. Bu durum, örneğin "bilgisayar" sözcüğünün terim duyarlı yapısını alanlara göre çözümlenmesini de olanaklı kılmaktadır. Bu çalışma alana özgü DDİ uygulamalarının geliştirilmesinde bir bilgi tabanı olarak kullanılabileceği gibi, daha özgül sözlüklerin hazırlanmasını da kolaylaştırıcı rol üstlenecektir.

1.5.6. Çok Anlamlılık Çözümlemesi

Anlam, dildeki somut birimlerin dizge ya da sistem düzeyinde bir ara gelmesiyle oluşan ve dilsel mesajın iletiminde gerçekleşen en üst aşama olarak ele alınabilir. Sözün bağlamından bağımsız olarak bir anlamından söz edilip edilmeyeceği konusu, hem felsefi boyutu hem de sözlük bilimsel boyutu olan oldukça karmaşık bir konudur. İnsanlar arası iletişimde olduğu gibi bir DDİ sisteminde de anlam, ulaşılması gereken en son nokta olarak hedeflenir. Makinenin bir dili anlaması denildiğinde elbette insanın dilsel iletişimde gerçekleştirdiği anlama düzeyi kastedilmekle birlikte, beyindeki sinir hücresi düzeyinden başlayarak insanın öğrenmesinin ve kavramlaştırmasının ne kadar karmaşık olduğu göz önüne alındığında, makinenin ürettiği sonuçların bugün için çoğu zaman eksik ve hatalı olduğunu görmekteyiz.

DDİ'de *çok anlamlı sözcüklerin çözümlenmesi* (word-sense disambiguation) verilen bağlamdaki sözcüğün anlamlarını açıkça ortaya koymaktır. Burada bir sözcüğün sözlükte numara ile belirtilen anlamlarından hangisinin bağlamda kastedildiği belirlenmeye çalışıldığı gibi özellikle eş sesli ama anlamları farklı sözcüklerin neye gönderme yaptığı da bulunmaya çalışılabilir. Örneğin Güncel Türkçe Sözlük'te *sevmek* sözcüğü için 5 farklı anlam verilmiştir. Bu sözcüğün geçtiği bağlamda, sözcüğün "hangi anlamı"nın kullanıldığı belirlenmeye çalışılabilir ya da *koyun* gibi bir sözcüğün geçtiği bağlamda; hayvan adı olan *koyun* mu, kucak anlamındaki *koyun* mu, emir 2. tekil kişi olarak kullanılan *koyun* mu, yoksa denizin ya da gölün girinti yaptığı yerin adı olan *koy* sözcüğünün tamlayan durum ekli kullanımının mı kısacası "hangi *koyun*"dan söz edildiği gibi insan için değil ama makine için karmaşık bir anlamsal durum bulunmaktadır.

Çok anlamlılık çözümlemesinin tarihi, makineli çeviri sistemlerinin tasarlandığı 1950'li yıllara kadar gitmektedir. Makineli çeviri sistemlerinde bir dilden diğerine yapılan çözümlemede herhangi bir sözcüğün birden fazla anlama sahip olması anlam bulanıklığına yol açmakta, hedef dilde kabul edilmeyecek yanlış çevirilere neden olmaktadır. Bu yüzden Wilks ve Stevenson'a göre, anlam çözümlemesi özellikle makineli çeviri için yaşamasal olmakla birlikte bilgi çıkarımı (information retrieval) çalışması için de değerlidir (Stevenson ve Wilks, 2003: 251).

Çok anlamlı sözcüklerin bağlamdaki yerlerine göre anlamlarının indirgenmesi için ilk çalışmalardan birini Wilks 1972'de anlam bilimsel formülleştirmelere gittiği ve *seçim anlam bilimi* (preference semantics) adı verilen bir kuramsal çerçeve ile

yapmıştır. Buna göre sözlükte yer alan her bir anlam bir formülle ilişkilidir. Wilks'in geliştirdiği sistem 80 anlam bilimsel özellikle yüklü ve birbiriyle eş anlamlılık ilişkisi olan "insan", "istemek", "soyut" ve "hareket" kavramlarını içine almaktadır. Bu sistemde çok anlamlılık, bir sözcüğün anlamı birden fazla olası formülle temsil edilir. Yazarın düzenli çok anlamlılık adını verdiği bir anlam düzeneği daha vardır. Burada sözcüklerle ilgili formüllerin yeni bağlamlarda da geçerli olması söz konusudur. Sonuçta bu yaklaşımın amacı metnin anlam bilimsel gösterimini çıkarmak olmakla birlikte çok anlamlılığın indirgenmesi bütün metin anlam bilimi çözümleme işleminin bir parçası niteliğindedir. Bu sistem 1973'te İngilizceden Fransızcaya küçük bir çeviri uygulamasında kullanılmıştır.

1986'da Lesk ilk kez bir sözlüğü çok anlamlılık çözümlemelerinde kullanmıştır. Kullandığı sözlük Oxford Advanced Learner's Dictionary adlı sözlüğün makinece okunur biçimidir. Lesk, sözlükteki tanımlardan yola çıkarak bir sözcüğün anlamının çevresindeki diğer sözcükle ilişkisine dayandığını belirlemeye çalışmıştır. Stevenson ve Wilks, DDİ'de çok sık kullanılan Longman'in makinece okunur sözlüğündeki dil bilgisel ipuçlarından yararlanarak bir ***anlam etiketleyicisi*** (sense tagger) geliştirmeye çalışmışlardır. Diğer yaklaşımların anlam çözümlemesinde tek başlarına yetersiz kaldıkları belirtilerek etiketleyiciyle birlikte makine öğrenmesi yaklaşımını kullanarak 40.000 sözcüklük bir anlam bulanıklığı bulunan derlemde yaptıkları testlerde yaklaşık %90'lara varan başarım elde etmişlerdir.

Yapay sinir ağlarını 1990'da Ide ve Véronis bir ağ yaklaşımı olarak kullanmışlardır. Bu sistemde sözcük bilgi tabanı olarak Collins English Dictionary temel alınmış, sistem sözcük düğümleri ve anlam düğümleri olmak üzere iki yapıdan oluşmuştur. Sözcük düğümü, her bir sözcüğün olası farklı anlamlarıyla pozitif olarak ilişkiliyken anlam düğümleri de sözlükte tanımı verilen sözcük düğümüyle bağlı bulunduğu sözcükle pozitif ilişkili, bulunmadığı tanımla yani diğer sözcük düğümleriyle negatif ilişkilidir. Bu sistem bu ağ yapısıyla aynı zamanda geri bildirim sağlayan bir sitemdir. Çözümleme süreci cümle içinde kullanılan sözcüğün bir sözcük düğümünü harekete geçirmesiyle başlayıp stabil duruma gelene kadar olan sürede çeşitli geribildirimlerle sonuçları üreterek sonlanmaktadır. Sistemin başarımı %85 olarak verilmiştir.

Sözcüklerin çok anlamlı yapısıyla ilgili olarak Yarowsky, Gale ve Churhc iki farklı anlam bilimsel yaklaşım geliştirmişlerdir. Birinci yaklaşıma söylem başına bir anlam (one sense per discourse) adını vermişler, buna göre bir sözcük çeşitli

kullanımlarıyla birlikte iyi düzenlenmiş bir söylemde % 98 oranında aynı anlama gelmektedir. İkinci yaklaşım sadece Yarowsky'nin kendi çalışmalarına dayanır. Buna göre önceki yaklaşımın daha dar bir yapıya indirgenmesi söz konusudur. Çok anlamlı sözcüklerin farklı anlamlarla aynı eşdizimde bulunması çok düşük bir olasılıktır. Wilks ve Stevenson'a göre bu son iki bakış açısından özellikle eşdizimli yapılardaki sözcüklerin anlamsal farklılıklarından yola çıkarak bütün bir metinde çok anlamlılığın yarattığı belirsizlikler büyük ölçüde giderilebilir.

Çok anlamlılık çözümleme sistemlerinin ortaya çıkardığı ürünlerin değerlendirilmesi için 1998'de SENSEVAL adlı bir proje oluşturulmuştur. İlk değerlendirme toplantısında 17 sistem değerlendirmeye girmiştir. Değerlendirmede Oxford Üniversitesi'nin geliştirdiği HECTOR adlı özel yapılı bir sözlük kullanılmıştır. Sözlük bir anlam seti sunmaktadır, böylece değerlendirmeye giren sistemler bu sete göre ürettikleri çıktıları karşılaştırmakta başarım düzeylerini ortaya koymaktadırlar. Bu sözlükle, sistemlere ilk önce HECTOR'da etiketlenmiş anlama göre sözcük bağlamı örnekleri verilmekte sonra etiketlenmemiş bağlamlar üzerinden sistemler denenmektedir (Stevenson ve Wilks, 2003: 252-259).

Altan ve Orhan anlam çözümlemede en etkili yöntem olarak makine öğrenme yönteminde kullanılan algoritmaları dile getirmişler, *ön görmeli* (supervised) ve *ön görmesiz* (unsupervised) olarak adlandırılan öğrenme türleri vurgulamışlardır. Öngörmeli ya da denetimli tipte ele alınan verinin bütün ayrıntıları gösterilmişken öngörmesiz ya da denetimsiz tipte çalışılan veride sınıflandırma içeriği bilinmemektedir. Bu tür öğrenmelerden öngörmeli öğrenmenin başarım açısından daha olumlu sonuçlar verdiği görülmüştür. Burada en önemli nokta öngörmeli öğrenim için etiketlenmiş derlemin ayrıntısıdır. Ses bilimsel, biçim bilimsel ve özellikle söz dizimsel etiketleme ayrıntılarının anlam çözümünde başarım oranını artırdığı bilinmektedir (http://turkoloji.cu.edu.tr/DILBILIM/anlam_belirsizligi.pdf)

DDİ çalışmalarının Türkçe için çok eski olmayan tarihi içinde özellikle anlam bilimsel çözümleme için bir çalışma henüz tam olarak yapılmamış olmakla birlikte ODTÜ Türkçe Derlemi'nde söz dizimsel ve anlam bilimsel kimi etiketlemelere gidilmiştir. Aşağıdaki listede "git-" fiili için bir etiketleme görülmektedir:

No Tümce

1 ülkeyi keşfetmek için yazar da [birlikte] (DuZ) $gidiyor$ {1} ve karada kalıyor

2 gemi, [Suratya] (YeZ) $gidiyordu$ {1}

3 Umut Burnuna kadar [rüzgâr] (ÖZ) [çok iyi] (DuZ) $gitti$ {3}
4 Biraz daha kuzeye dönerek Tataristanın kuzeybatısına ve [Buzdenizine] (YeZ) $gitmek$ {1} [olasılığı karşısında] (ZaZ), bulunduğumuz rotayı izlemenin daha iyi olacağını düşündük
5 [merakımı] (KEyl) $giderecek$ {5} bir şey de göremediğimden
6 [olanca hızımla] (DuZ) [o önce] (ZaZ) $gittiğim$ {1} [yana] (YeZ) koşmuştum
7 [sesim ve işaretlerim] (ÖZ) [hoşuna] (KEyl) $gitmiş$ {4} [gibiydi] (Eyl)
8 Fakat nasıl davrandığımı, kocasının işaretlerine göre ne kadar iyi davrandığımı görünce bana alıştı ve $gitgide$ {12} [artan bir sevgi] (DoT) beslemeye başladı
9 Bu yaptığım [pek] (MiZ) [hoşlarına] (KEyl) $gitmişti$ {4}
10 Ben de $gittim$ {1}, elini öptüm

Listede git- filinin çekimleri iki $ işaretiyle etiketlenmiş zarflar parantez işareti içine etiketlenmiş aynı zamanda iki köşeli ayraçla gösterilmiştir. git- fiilinin kaçıncı anlamda kullanıldığı ise ({}) işaretiyle belirtilmiştir,[24]

ODTÜ-Sabancı ağaç yapılı derleminin de kullanıldığı bir küçük derlem üzerinde Altan ve Orhan öngörmeli öğrenme yöntemini denemişler ve örnek tabanlı algoritmaların istatistiksel bir yaklaşım olarak adlandırılan Naïve Bayes yaklaşımına göre başarımı daha yüksek sonuçlar ürettiğini belirtmişlerdir.

1.6. Derlem Dil Bilimi

1.6.1. Derlem Kavramı

Son yıllara değin Türkçe yazında ***bütünce*** olarak karşılanan ***derlem*** (***corpus***) sözcüğü Güncel Türkçe Sözlük'te iki anlamlı olarak verilmiş, birinci anlamında "koleksiyon" sözcüğüne karşılık olarak, ikinci anlamında da terim anlamı olarak "Bir incelemede kullanılmak üzere bir araya getirilmiş metinlerin bütünü." biçiminde tanımlanmıştır.[25]

Sözcük, Merriam-Webster çevrimiçi sözlükte üç anlamlı olarak verilmiştir. Sözcüğün İngilizcedeki temel anlamı "hayvan ya da insan bedeni" anlamında olup

[24] http://turkoloji.cu.edu.tr/DILBILIM/anlam_belirsizligi.pdf
[25] http://www.tdk.gov.tr/TR/SozBul.aspx?F6E10F8892433CFFAAF6AA849816B2EF4376734BED947CDE&Kelime=derlem

özellikle de ölü bedeni anlamında kullanılmış, günümüzde terim anlamı yaygınlık kazanmıştır.[26]

Derlem, terim olarak özellikle dil araştırmalarına temel olmak üzere belirli ilkeler, kurallar çerçevesinde bir araya getirilmiş, özel ya da genel çeşitli türlerde hazırlanmış, elektronik ortamda tutulan ve bilgisayarca okunur biçime getirilmiş sözlü ya da yazılı bütündür (Say, 2005, 81).

Derlem araştırmalarının deyiş biliminden (stilistik), bütün dil bilgisi araştırmalarına, DDİ çalışmalarındaki dil modellemesine kadar geniş bir kullanım alanı vardır. Son 20 yılda gelişen bilgi teknolojilerine bağlı olarak derlem oluşturma ve sorgulama biçim ve hızında büyük ilerlemeler yaşanmıştır.

İngilizce için derlem oluşturma çalışmalarıyla tanınan John Sinclair, derlemin temsil gücünün önemine değinerek, derlemin rasgele bir araya getirilmiş metinler yığını olmadığını özellikle vurgulamıştır. Belirli dilsel ölçütlere göre bir araya getirilen metinlerin derlem oluşturabileceğini ve derlemi geliştirme amacının birincil ölçütlerden olduğunu belirtmiştir. Böylelikle derlem kavramı makinece okunur, örneklemesi belirli ölçütler çerçevesinde konuşma dilinden ya da yazılı metinlerden yapılmış ve dil bilimsel işaretlemesi hazırlanmış bir koleksiyon olduğu modern dil bilimde kabul edilen tanımdır (McEnery vd. 2006:4).

Derlem tabanlı çalışmaların ya da derlem dil biliminin bir yöntem mi olduğu yoksa söz dizimi, biçim bilimi, ses bilimi gibi dil biliminin bir alt alanı mı olduğu sorusu tartışmalı konular arasındadır. McEnery ve Wilson (1998:2) derlem tabanlı çalışmaların dil biliminin ayrı bir disiplini olarak kabul edilemeyeceğini, diğer dil bilimi disiplinlerinin derlemi bir araç olarak kullanabileceğini belirtmektedirler. Bununla birlikte, Tognini ve Bonelli'nin 2001 yılındaki çalışmalarında derlem dil bilimini ayrı bir dil bilimi disiplini olarak kabul ettiklerini belirtmektedirler (McEnery vd. 2006: 7).

Derlem dil biliminin önemli araştırmacılarından Geoffrey Leech bir derlemin özelliğini şu biçimde dile getirmiştir: "Derlemin (bilgisayar destekli derlem) nadiren gelişigüzel toplanan bir bütün olduğu düşünülmelidir. Derlemler genellikle belirli amaçlar için dilin bir bölümünü ya da kimi metin türlerini ***temsil etmek*** için oluşturulurlar" (age., 2006:4).

Derlem oluşturma yöntemlerinin kuramsal çerçevesi ve uygulamalarına yönelik araştırmalar günümüzde hız kazanmış, İngilizce başta olmak üzere diğer dillere yönelik

[26] http://www.m-w.com/cgi-bin/dictionary?book=Dictionary&va=corpus

derleme dayalı yöntemler dil bilgisel modellerin oluşturulması ya da desteklenmesi amacıyla daha çok kullanılmaya başlamıştır. İşin kuramsal yönünde doğrudan derlemden çıkan sonuçlar üzerinden dil bilgisel modellemelerle, var olan kuramsal dil bilgisi çerçevesinin derlemle desteklenmesi konusunda da bir ayrıma gidildiği görülmektedir. Bu konuda Ooi (1998: 51) derleme dayalı bu iki bakış açısı arasındaki farkları şöyle belirlemiştir:

Derlem Tabanlı Yöntem	Derlem Yönelimli Yöntem
- Dil bilimci var olan kuramı kanıtlama için derlemi kullanır. - Dil bilimci dayandığı kuramın sağlamlığından kuşku duymaz dolayısıyla derlemden çıkan sonucu tehdit olarak görmez.	- Gözlemlenen veriden yeni kuramsal çıkarımlar yapılabilir. - Dil bilimci derlem verisinin kuramda değişiklikler yapmasını bekleyebilir.

1.6.2. Derlem Araştırmalarının Tarihçesi

Derlem dil bilimi terimi, 1980'lerin başlarında ortaya çıksa da, derlem dil bilimi yöntemini Chomsky öncesi dil bilimcilerinden Boast ve yapısalcı dil bilimi akımının Amerika'daki temsilcilerinde Sapir, Mewman, Bloomfield ve Pike kullanmışlardır.

1950'li yılların sonlarında derlem çalışmalarına, araştırmacıların bireysel çabalarının yetersiz kalmasının ve teknolojinin ilk derlem çalışmaları denebilecek çalışmalar için bulunmamasının yol açtığı kimi olumsuzluklar özellikle Chomsky'nin sert eleştirilerine hedef olmuştur. İlk derlem tabanlı çalışmaların ses bilimsel ayırt edici özellikler üzerine kurulan küçük boyutlu derlemler olduğu görülür. Jesperson ve Fries ses bilimi dışında genel dil bilgisine yönelik derlem tabanlı çalışmalar yapmışlardır.

Teknolojinin gelişimiyle küçük, elle toplanan verilerden bilgisayar verisi olarak derlenen ve depolanan derlemler oluşturulmaya başlanmıştır. 1960'lı yıllarda Brown Üniversitesi'nde hazırlanan Brown Derlemi, bilgisayar destekli olarak hazırlanan ilk derlem olarak derlem teknolojisinin başlamasını sağlamıştır. 1980'li yıllardan

günümüze devam eden teknolojik gelişmelere bağlı olarak bugün birçok derlem çeşitli amaçlar için hazırlanmaktadır (McEnery vd., 2006:3-4).

1.6.3. Derlem Türleri

Derlemler genel ve özel derlemler olarak iki gruba ayrılmaktadır. Genel derlemler bir dildeki bütün alt çeşitliliği barındıran derlemlerdir. British National Corpus (BNC), 100 milyon işaretlenmiş sözcükten oluşan ve derlem dil bilimi kaynaklarında sözü edilen ilk büyük derlemdir. Bu derlemin modern İngilizceyi temsil ettiği belirtilmektedir.

Özel derlemler alan ve türe özgü derlemlerdir. Hukuk, tıp, mühendislik gibi bilim alanlarına göre ayrılabildikleri gibi, yine kendi içinde gazete, roman, dergi, akademik yazılar gibi türlerden de oluşabilmektedir. Genel ve özel derlemler temsil güçlerine göre farklılıklar gösterirler. Genel derlemlerin içlerinde barındırdıkları türe bağlı olarak temsil güçleri artar ya da eksilir. Özel derlemler de barındırdıkları sözcük düzeyi kadar temsil gücüne sahiptirler (Mc Enery vd. 2006:15).

Bugün derlem dil bilimi çalışmalarının en çok gerçekleştirildiği ve buna bağlı uygulama yazılımlarının en çok yazıldığı dil İngilizcedir. İngilizce için geliştirilen genel ve özel amaçlı derlemlerin yanında özellikle bir derlem türü öne çıkmakta, bu derlem türünün temsil niteliği ve kullanım alanıyla ilgili tartışmalar sürmektedir. *Monitör derlem* (monitor corpus) adı verilen bu derlem türünde sürekli gelişen ve genişleyen bir yapı söz konusudur. Malzemesi örneklemeyle elde edilen durağan (statik) derlemler yerine bu türde, belirlenen amaç için çeşitli konularda genişleyen boyutlarda metin yığınları güncel olarak bir araya getirilmektedir. Aylık, yıllık ve hatta günlük olarak konuyla ilgili metinler değerlendirilmek, ön işlemlerden geçmek üzere ham olarak toplanır. Bank of English 524 milyon sözcüklük derlem büyüklüğüyle bugün için monitör derlemin en önemli temsilcisi olarak öne çıkmakta, 1980'den günümüze sürekli geliştirilmektedir. Global English Monitor Corpus adlı derlem de 2001 yılından günümüze var olan metinlerden oluşturulmakta, 1 milyar sözcüklük bir büyüklüğe ulaşacağı öngörülmektedir. Bu derlemin oluşturulma amacı; İngiltere, ABD, Avustralya, Pakistan ve Güney Afrika'da kullanılan İngilizcenin farklılıklarını ve değişimini izlemektir. Bir başka monitör derlem örneği Birmigham Üniversitesinde geliştirilen AVIATOR (Analysis of Verbal Interaction and Automated Text Retrieval) adlı derlem yapılı sistemdir. Sistemin kurulma amacı, İngilizce metinleri sürekli ve

güncel olarak izleyerek dilsel değişimleri ve özellikle yeni ortaya çıkan ögeleri (neologism) belirlemektir. Bu tür derlemde bir sınırlama getirilmediğinden artan ya da büyüyen bir arşiv olarak da değerlendirilmektedir (age., 2006: 67).

1.6.4. Başlıca Derlemler

Derlem araştırmalarıyla ilgili yayınlarda, tarihsel olarak ilk makinece okunur biçimli derlemin Brown derlemi olduğu belirtilmektedir. 1964'te bir el kitapçığıyla birlikte tamamlanıp sunulan Brown derlemini Nelson Francis ve Henry Kučera hazırlamıştır. Derlem Brown Corpus olarak kısaltılarak kullanılmakla birlikte orijinal adı Brown University Standard Corpus of Present-Day American English'tir. Chomsky'nin 1957'de yayımladığı Sözdizimsel Yapılar (Syntactic Structures) adlı yapıtının dil biliminde üretici-dönüşümsel gramere dayalı bakış açısının ortaya çıkmasını sağladığı ve bu kurama dayalı çalışmaların arttığı bu yıllarda Brown derleminin ortaya konması, dil biliminde ürün olarak ortaya çıkmış dilsel ifadelerin değil de ideal konuşucu ve dinleyicinin üretebileceği zihinsel yapılara dayalı bir gramer anlayışı karşısında oldukça anlamlıdır. Chomsky'nin derlem çalışmalarının hiçbir zaman yeterince kapsayıcı olmayacağı düşüncesinde olduğu bilinmektedir.

Brown derlemi 1961 yılına dayanan Amerikan İngilizcesinin eşzamanlı bir derlemidir. Her biri yaklaşık 2000 sözcük içeren 500 örnekten toplam 1.014.300 sözcük o zamanın bilgisayar teknolojisine göre öncelikle delikli kartlara sonra da manyetik teyplere kaydedilmiştir. Çalışmanın amacı varsayılan bir biçemsel kalitenin temsili amacıyla değil de o güne ait İngilizcenin karşılaştırmalı çalışmalar ve çözümlemeler için ölçünleştirilmiş bir görünümünü sunmaktır. Derlemin malzemesi yerel kütüphanelerden elde edilen metinlerden oluşmaktadır. Sözcüklerin türlerini belirlemek için 80'in üzerinde etiket belirlenmiştir (Kennedy, 1998: 23, 26, 27).

1970-1978 yılları arasında Brown derlemine koşut olarak Oslo ve Lancaster Üniversitelerinin ortak olarak hazırladığı Lancaster-Oslo/Bergen derlemi (LOB), Brown derleminin oluşturulma teknikleri örnek alınarak oluşturulmuştur. Bu derlemle birlikte derlem dil biliminin hem kuramsal hem de yapısal birtakım sorunları gün ışığına çıkmaya başlamıştır. Özellikle bir derlemin temsil gücünün metinlere oranlanması konuları tartışılmaya başlamıştır. Bu bağlamda derlem ekibinin başkanı Johansson, LOB'un metin seçiminin rastlantısallıktan uzak olarak amaçlı bir sınıflamaya dayandığını belirtmiştir. Brown derlemine göre LOB'un metin türlerinin

sınıflandırılması daha tutarlı görünmektedir. 70'li yıllardaki bilgisayar teknolojisinde kaydedilen ilerlemelere paralel olarak derlem, Machintos, MS-DOS ve Unix platformlarına taşınmak üzere kaydedilmiştir. Ayrıca bu derlemle birlikte cümle sınırlarının ve kısaltmaların işaretlenmesi için de ayrıca bir kodlama sistemi geliştirilmiştir. Bağlamlı dizin oluşturucu bir sistemde sorgulanan sözcüğün ortada gösterimine dayanan bağlamlı satırların ortaya konulması yöntemi de bu derlemde uygulanmaya başlamıştır. İngiliz İngilizcesini temel alan bu derlemle Amerikan İngilizcesini yansıtan Brown derlemindeki sözcük sıklıkları bilgisi bu iki İngilizcenin karşılaştırılmalı çalışmalarına da katkı sağlamıştır (age., 1998: 28).

Konuşma diline yönelik olarak hazırlanan ilk derlem London-Lund derlemidir (LLC). Bu derlem Survey of English Usage (SEU) olarak bilinen derlemin çeşitli ses bilimsel ölçütler içerinde işaretlenmesini içermektedir. 87 adet metin ve 435.000 sözcüğün işaretlendiği derlem 1990'lı yıllara kadar kullanılan tek sözlü dile dayanan derlemdir. Bu derlem sözlü dilde kullanılan sözcükleri ve cümle yapılarını prozodik açıdan betimlediğinden özellikle söylem çözümlemesi çalışmalarında veri sağlayıcı olarak kullanılmıştır.

1984-1987 yılları arasında 11 kategoride yetişkinlerle yapılan görüşmelerden alınan ve 52.600 sözcüklük Lancaster/IBM Spoken English Corpus (SEC) 1992'de hazırlanmıştır. Bu çalışmada da vurgu, tonlama gibi konuşmaya bağlı özellikler işaretlenip, noktalama ve dil bilgisel sınıflandırma da yapılarak yoğun disk (YD) biçiminde kullanıma sunulmuştur. 900 konuşmacının 80 saatlik çeviriyazısı yapılmış konuşmalarından elde edilen ve yaklaşık 1 milyon sözcüklük bir toplama ulaşan Corpus of Spoken Amerian English (CSAE) de bir diğer önemli konuşma diline dayanan derlemdir. Bu derlem son zamanlarda yapılmış ve bilgisayarca okunur özellikli en büyük sözlü dil derlemidir. Yüz yüze görüşme, kendiliğinden (spontan) konuşma ortamı, iş yaşamı konuşmaları, tartışmalar, haber yayınları, seromoniler ve radyodaki telefon konuşmaları gibi değişik türlerdeki ortamlar kaydedilmiştir. Bu özelliğiyle özellikle ağız çalışmaları için değerli veriler sağlamaktadır.

8. yüzyıl ve sonrası eski İngilizceyi derlemek üzere Toronto Üniversitesinde hazırlanan ve 1981 yılında yayınlanan Complete Corpus of Old English adlı derlem, 3.022 adet metnin bir araya getirilmesinden oluşmuştur. Elektronik olarak hazırlanan ilk tarihsel derlem Helsinki Üniversitesi'nde Helsinki Corpus of English Texts: Diachronic Part adlı derlemdir. Çalışma 1984-1991 yılları arasında tamamlanmış olup eski

İngilizceden modern İngilizcenin ilk dönemlerini kapsamakta 400 örnekten ve 1.5 milyon sözcükten oluşmaktadır.

Birmingham Üniversitesi'nde hazırlanan ve Cobuild olarak bilinen Collins Birmingham University International Language Databese adlı derlem, 1960 ve sonrası dönemi kapsamaktadır. Bu derlemin oluşturulması sırasında gözetilen hedef kitle başta sözlük gereksinimi olan öğrenciler ve eğitmenlerdir. Ağız özelliklerinin dışında ölçünlü İngiliz İngilizcesinden %70, Amerikan İngilizcesinden %20 oranında metinler derlenmiştir. Derlemin % 25'lik bölümü sözlü dile ayrılmıştır. Metinler rastgele bir seçimle elde edilmemiş, daha çok popüler ve çok okunan eserlerden oluşmuştur. Üniversite arşivinde bulunan görüşme kayıtları ve ders konuşmalarıyla radyo yayınlarından elde edilen kayıtlar sözlü bölümün malzemesini oluşturur. 1982'de derlemin sözcük sayısı 7.3 milyondur. Bu rakam 1987'de 13 milyona ulaşmıştır. 1990'da derlem projesinin yöneticisi John Sinclair, Cobuild'in sözcük sayısı bakımından daha kapsamlı bir derleme dönüşmesini sağlamak üzere The Bank of English adlı projeyi duyurmuştur. 1997'de 300 milyonu aşan derlemin günümüzde 500 milyon sözcüğü içerdiği bilinmektedir. Bu derlem bir monitör derlem özelliği taşımakta, yani sürekli artan bir içerik örüntüsü göstermektedir.

Derlem dil bilimi araştırmalarında adından sıkça söz edilen British National Corpus (BNC), 1991-1995 yılları arasında ve İngiliz hükümetinin de proje maliyetinin yarısını desteklemesiyle tamamlanmıştır. 100 milyon sözcük içeren BNC, 1960-1975 arası ölçünlü İngiliz İngilizcesini temel almakta, %90'ı yazılı ve %10'u sözlü dilden derlenen malzemeyi içermektedir. Malzemenin %60'ı kitaplardan, %25'i dergilerden, %5'i broşürlerden ve %5'i de yayımlanmamış mektuplardan oluşturulmuştur. Sözlü malzeme çok çeşitli toplumsal katmanlardan derlenmiştir. Bunlar arasında ders konuşmaları, sunumlar, görüşmeler, seromoniler, politik konuşmalar, spor yorumları, söyleşi gösterileri sayılabilir. 15 ile 60 yaş arası farklı sosyo-ekonomik özellikteki 124 kişiyle yapılan 2000 saatlik görüşme kayıtları da derlemin sözlü dile dayanan içeriğinde yer almaktadır. Görüşmeciler yapılan görüşmenin kaydedildiğinden haberdar edilmiş, her bir görüşmeciyle yapılan kaydın ayrıntılı kayıt ortamları (zaman, yer, görüşmecinin mesleği, yaşı vb.) tutulmuştur. Derlemde yer alan metin örnekleri katmanlı örnekleme tekniğiyle elde edilmiş (yer, zaman, ortam ve düzey) mümkün olduğunca temsil gücü yüksek bir yapı elde etmek amacıyla kullanılmıştır. Burada amaç, yöntemin daha çok metnin kendi doğasına ve özelliklerine dayanmasıdır. SGML işaretleme diliyle

işaretlenen derleme, internet üzerinde sorgu sistemiyle ulaşılabilmektedir (age., 1998: 23-54).

1.6.5. Derlem Oluşturma Yöntemleri

Amacı belirlenmiş ve dilsel alt grupları iyi seçilmiş, temsil niteliği yeterli olan bir derlemin oluşturulmasında hangi örnekleme tekniklerinin kullanılacağı tartışmalı bir konudur. Derlemin;

a) Makinece okunur olması,
b) Otantik özellikli metinlerden oluşması,
c) Temsil niteliğinin iyi olması

bir derlem için gerekli niteliklerdir.

Bilgisayarca okunur dilsel verinin sağladığı yararlar arasında şunlar sıralanabilir:

1- Verinin hızlı bir biçimde sıralanması, sorgulanması ve biçimlenmesi.
2- İnsan sezgisine ve önyargısına bağlı yanlışlıkların en aza indirilmesi.
3- Meta veri adı verilen verinin tanımını içeren işaretlemelerin çok daha hızlı ve otomatik yapılması.

Bilgisayarın derlem çalışmalarına sağladığı katkıyı Tognini-Bonelli dil bilim incelemesinde bir yöntemsel yön gösterici ve belirleyici olarak kabul etmektedir. Bilgisayarın belirleyici özelliğinden dolayı "bilgisayar destekli derlem dil bilimi" terimi de önerilmiştir (McEnery vd. 2006:6).

Derlem çalışmalarında dil bilimcinin sezgisine dayanan incelemeler yerine gerçek, otantik, kullanımı temel alan dil verisinin tercih edildiği görülmektedir. Sezgiye dayalı yöntemde, araştırmacının kendi bireysel dil kullanımı, dış etkilenmeleri, önyargıları gibi incelemeyi etkileyecek değişkenler derlem tabanlı çalışmalarda sorun olmaktan çıkar. Araştırmacının sezgisine dayanan incelemede dil verisinin niceliksel yönlerinin çıkarımı da olanaklı değildir. Derlem tabanlı yöntemle veride yer alan farklılıklar hem niteliksel hem de niceliksel yönden kolay bir biçimde ortaya konulabilir. Derlem çalışmalarında geleneksel sezgiye dayalı yöntem "deneysel" bir yönteme dönüşmektedir (age., 2006:7).

Derlemde ***denge*** (balance) konusu da derlemin sağlamlığı ve güvenirliği açısından önemlidir. Derlemin dengeli olması içerdiği metin sınıflarına (kategorilerine) ve derlemin kullanım amacına bağlıdır. BNC dengeli derlem örneği içinde en önemlilerindendir. Sadece sözlü dilden örnekler içeren CANCODE, alana özgü olan

HKUST (bilgisayar bilimi derlemi) ve yalnızca yazılı dile özgü derlem olan Brown derlemi de diğer dengeli derlemler arasındadır. Bir derlemin hangi ölçütlere göre dengeli sayılacağı konusu günümüzde henüz tartışmalı konular arasındadır. BNC; American Ulusal Derlemi, Kore Ulusal Derlemi, Polonya Ulusal Derlemi ve Rusya Referans Derlemi gibi diğer ülke derlemlerinin hazırlanmasında kaynaklık etmiştir. BNC'nin dengeli bir derlem olarak kabul edilmesi onun yapısal olarak geniş bir dil malzemesinden hazırlanması nedeniyledir. 100 milyon sözcüklük derlemin %90'ı yazılı dilden, %10'u çeviriyazısı yapılmış sözlü dilden oluşmuştur.

Bir dilin bütün yönlerini ayrıntılarıyla temsil edecek bir derlemin oluşturulmasının zorluğu, ***örnekleme*** (sampling) konusunu derlem dil biliminin önemli bir sorunu durumuna getirmiştir. Örneklenecek malzemenin seçimi ve örnekleme yönteminin sağlamlığı tartışmalı konular arasındadır. Derlem ister durağan, ister sürekli geliştirilen bir yapıda olsun aslında genişçe bir dil nüfusundan elde edilmiş, seçilmiş örneklerdir. Temsil özelliği yeterli olan bir derlem için yapılması gereken, öncelikle dil ürününün sınırlarının belirlenmesi ve bu üründen ***örnek birimi***nin (sampling unit) belirlenmesidir. Örnek birimdeki ögeler ***örneklemenin çerçevesi***ni (sampling frame) oluşturmaktadır. Örneklem birimi; kitap, dergi ya da gazete gibi metin türlerinden oluşur. Örneklem çerçevesi için örneğin Brown derlemini oluşturan Brown Üniversitesinde Kütüphanesinden yer alan kitap ve dergi koleksiyonlarının listesi verilebilir (age., 2006:16-19).

Derlem tasarımı ve örneklemesi hazırlandıktan sonra bilgisayar ortamında tutulan metin verisinin yapılandırılmamış, ham durumuyla bulunması, incelenen dil hakkında elde edilecek niteliksel ve niceliksel bilginin sağlamlığı ve güvenirliği konusunda kuşkulara yol açmaktadır. Derlemdeki dilsel verinin sağlam bir bilgiye dönüşebilmesi için bilgisayar ortamında tutulan verinin işaretlenmesi gerekmektedir. ***İşaretleme*** (annotation), "elektronik form biçiminde depolanan belgenin içine, belgenin kendisiyle ilgili bilgilerin standart bir kodlamayla girilmesidir." (age., 2006: 22). Derlemin işaretlemesi, derlemin işlenmesinde ilk aşamadır. Derlemdeki belgelerin sınırlarının belirlenmesi belge üzerinde daha sonra gerçekleştirilecek işlemler için kolaylık sağlamakla birlikte anlamlı bilgilerin elde edilmesi için gereklidir. Metinlerin işaretlenmesinde, belgede geçen tablolar, paragraflar varsa yabancı dilden yapılan alıntılar yerleşim yerlerine göre işaretlenir. Belge üzerinde editör yorumları da bu süreç içinde ayrıca yer alabilir. Sözlü dilden örneklenen metinlerin çeviriyazısı yapılırken de

konuşmada geçen durak yerleri, bürünsel özellikler de ayrıca işaretlenir (age., 2006: 23).

Derlem işlemede tekli veri akışı ve çoklu veri akışı olarak ikiye ayrılabilecek bir yöntemden söz edilebilir. Tekli veri akışında, derlem metinleri tek tek ele alınır ve metinde geçen özellikler belirlenir. Bu yöntem bilgisayarın görece zayıf donanıma sahip olduğu zamanlardan kalmadır. Doğrusal metin girdisinin her bir biriminin tek bir zaman biriminde işlendiği bir yöntemdir. Çoklu veri akışı için verilen örnek The Bank of English adlı derlemde kullanılan yöntemdir. Bu derlemde geçen metinlerdeki her bir birimin (sözcük, noktalama işareti vb.) bir numarası ve bu numaranın ilişkili olduğu bir metni bulunmaktadır, dolayısıyla, derlemde arama yapacak sözlük bilimci bulduğu birimin metnine daha kolay ve hızlı bir biçimde ulaşmaktadır. Elde edilen sonuç SGML gibi işaretleme diline çevrilebilmektedir (Sinclair, 2003: 179-190).

1981'de Avrupa Bilim Kurumu bilgisayarların sözlük yayımında kullanımı ve sınırlıklarıyla ilgili bir çalıştay düzenlemiştir. Bu çalıştayda ele alınan konulardan biri çok işlevli sözlükler konusudur. Çok işlevli sözlük konusunun gündeme gelmesinde daha önce de değinildiği gibi artan bilginin disiplinler arası nitelik kazanmış olmasıdır. Kruyt, makinece okunur sözlük ve derlemleri dil kaynakları (language resources) adı verilen bir yapının içinde ele alır. Dil kaynakları içindeki dilsel veri tabanlarını da derlemler, sözlükler ve kavram sözlükleri olarak düşünen Kruyt'a göre, dilsel veri tabanları yapılandırılmış bir veri düzeni değil bir çok ilşevli kullanılabilen dil kaynaklarıdır (Kruyt, 2003: 194-195).

Sözlükler, dil hakkındaki bilginin düzenli bir biçimde tutulduğu rezervuarlardır. Makinece okunur sözlükler, insanın kullanımı için elverişliyken bir bilgisayar destekli sözlük sistemi için yeterince yapılandırılmış değildir. Bilgisayarca değişik uygulamalarda insan dilinin anlaşılabilmesi için çok daha açık ve algoritmik, daha yapısal bir sözlük yapısı olmalı ve bir makinece okunur sözlüğün (MOS) bu duruma getirilmesi de uzun ve zor bir çalışma sürecini gerektirmektedir.

MOS'ların bilgisayar uygulamalarında kullanılabilmesi için bu sözlüklerden bilgi çıkarımıyla ilgili bir yçalışmayı 1989'da Boguraev ve Briscoe yapmıştır. Yöntem için temel aldıkları sözlük Longman Dictionary of Contemporary English olup kısaltması LDOCE'dir. Ayrıca tek bir MOS'tan bilgi çıkarımı yerine çoklu MOS'tan çıkarım için Byrd ve arkadaşları, yine Boguraev ve Briscoe ve Atkins çalışmışlardır. Bu çalışmaların tarihleri 1989-91 yılları arası olup, çalışmalardan büyük boyutlu derlemlerin önemi o zamanlar için daha iyi anlaşılmıştır. Church ve Mercer bilgisayar

uygulamalarında kullanılacak sözlük yapılarının veri yönelimli (data-driven) olması gerektiğini belirterek kural tabanlı (rule-based) yöntemlerin, gelişen depolama ortamları ve bilgi işlem gücünün artmasıyla daha da önem kazandığını savunmuşlardır (age., 2003: 196).

LDOCE'den bilgi çıkarımı ve kullanımı amacıyla Alshawi 1989'da bir tanım çözümleyicisi geliştirmiştir. Kyurt'a göre, bu sistemin başarım oranı % 88'dir. Uygulama, sözlükteki tanımları anlamsal baş (semantic head) adı verilen bir yapıya dönüştürmektedir. Yani madde başı tanımları bir anlamsal düğüm olarak yorumlanmaktadır. Alshawi, bu uygulamada üstten-inme (top-down) adı verilen bir bölümleme yöntemi kullanmıştır. Bu sözlük için bir başka uygulama örneği Pentherovdakie ve Vanderwande'nin 1993'te geliştirdiği uygulamadır. Uygulama, otomatik olarak sözlükte yer alan biçimbilgisel sözcüklerle anlamları arasındaki ilişki kurmaktadır. Türemiş madde başlarıyla bir ya da daha çok anlam arasındaki bağlantıyı ortaya çıkarmak amaçlanmıştır. Başarım oranı % 90 olarak verilmiştir.

Anlamsal ağ sözlükleri örnekleri olan WordNet ve EuroWordnet de otomatik bilgi çıkarımı çalışmalarına konu olmuştur. WordNet İngilizce için hazırlanan tek dilli, her bir sözcüğün bir kavram alanı ile ilişkilendirildiği yapısal bir sözlüktür. Psiko dil bilimi esaslarına göre hazırlanmıştır. WordNet üzerinde 2000 yılında Chai çalışmış, bu sözlüğü makine öğrenmesi yöntemiyle çözümlemeye çalışmıştır. Chai, çalışması sırasında önceden elle hazırlanmış bir etiketlenmiş eğitim setinin gerekliliğini bu çalışma için vurgulamıştır. EuroWordnet için benze bir çalışma diller arası bilgi çıkarımı (cross-language information retrieval-CLIR) amacılya Verdejo ve arkadaşları gerçekleştirmiştir. EuroWordnet İngilizce, İspanyolca gibi dillerin de bulunduğu diiler arası indeks sistemi kullanılan bir yapı özelliği gösterir (Kyurt, 2003: 197-198).

Sözlük bilgisi ya da sözlüğe dayanan bilgi çıkarımı bu açıklamalarla birlikte sadece sözlük biliminin konmusu olmaktan çıkmaya başlamıştır. Sözlük, yeni yapılandırma anlayışıyla çeviriden, dil öğretimine ve derlem oluşturmaya kadar uzanan geniş bir sözlük tabanlı ortamın ana veri sağlayıcı nesnesi durumuna gelmektedir.

1.6.6. Derlem ve İstatistiksel Çıkarım

Metinsel çeşitliliği ve dağılımı yapılandırılmış, temsil niteliği yeterli bir derlemden istatistiksel tekniklerle dil bilimsel çözümlemelere kaynaklık edecek listeler, dağılımlar ve tablolar elde edilmektedir. Betimleyici ve yorumlayıcı istatistiksel

yöntemler olarak başlıca iki başlık altında verilen istatistik yöntemlerde ortalama, standart sapma, mod, medyan gibi merkezi dağılım eğilimleri var olan bütünün ya da veri setinde yer alan birimlerin genel olarak dağılımı hakkında bilgiler verir. Bir derlemden sadece sıklık listelerine dayalı çıkarımlar yapılması eksik sonuçlara neden olmaktadır. Birçok derlemin karşılaştırmalı bilgisine dayalı listeleri, sıklık bilgisine dayanan istatistik teknikler için girdi olabilmektedir. İstatistik çıkarımlar için hazırlanmış SPSS gibi profesyonel çözümler için hazırlanmış yazılımlar hem betimleyici hem de yorumlayıcı çalışmalar için raporlar üretebilmektedir (McEnery vd. 2006: 52).

Dilsel hipotezlerin test edilmesi için istatistiksel anlamlılık değerleri ölçüt olarak alınmaktadır. Elde edilen sonucun şansa dayalı olup olmadığı P <0.05 gibi bir değerle test edilmektedir. Burada P olasılığı gösterir. Gözlemlenen değişkenlerdeki farklılıkların derecesi P <0.05 ya da kısaca 0.05'ten küçükse sonuçlar anlamlı kabul edilmekte, verideki farklılıkların şansa bağlı olmadığı belirlenmektedir. Derlem dil biliminde verinin test edilmesinde en çok kullanılan testlerden biri ki-kare testidir. Bu testte gözlemlenen değerle beklenen değer arasındaki fark ele alınır. Beklenen değerle gözlemlenen değer arasındaki fark ne kadar büyükse, farkın şans eseri olma olasılığı düşmektedir. Bu durumun tersinde de aradaki farkın şansa bağlı olduğu sonucuna ulaşılır. Sıkça kullanılan bir başka test log-likelihood testidir. Bu test, dağılımı normal olan veri üzerinde uygulanır. P değeri 0'a yaklaştıkça testin güvenirliği artmakta, böylelikle test edilen değişkenin şansa bağlı olmadığı anlaşılmaktadır (age., 2006: 55).

Eşdizimli yapılar gibi sıklıkla bir arada bulunan yapıların istatistiksel otomatik yöntemlerle çıkarımı günümüzde oldukça önem kazanmıştır. Bu yapıların çıkarımında en çok tercih edilen test tekniği *mutual information* (MI) tekniğidir. Bu tekniğin dışında T testi ve Z testleri de özellikle T testi sıklıkla kullanılmaktadır.

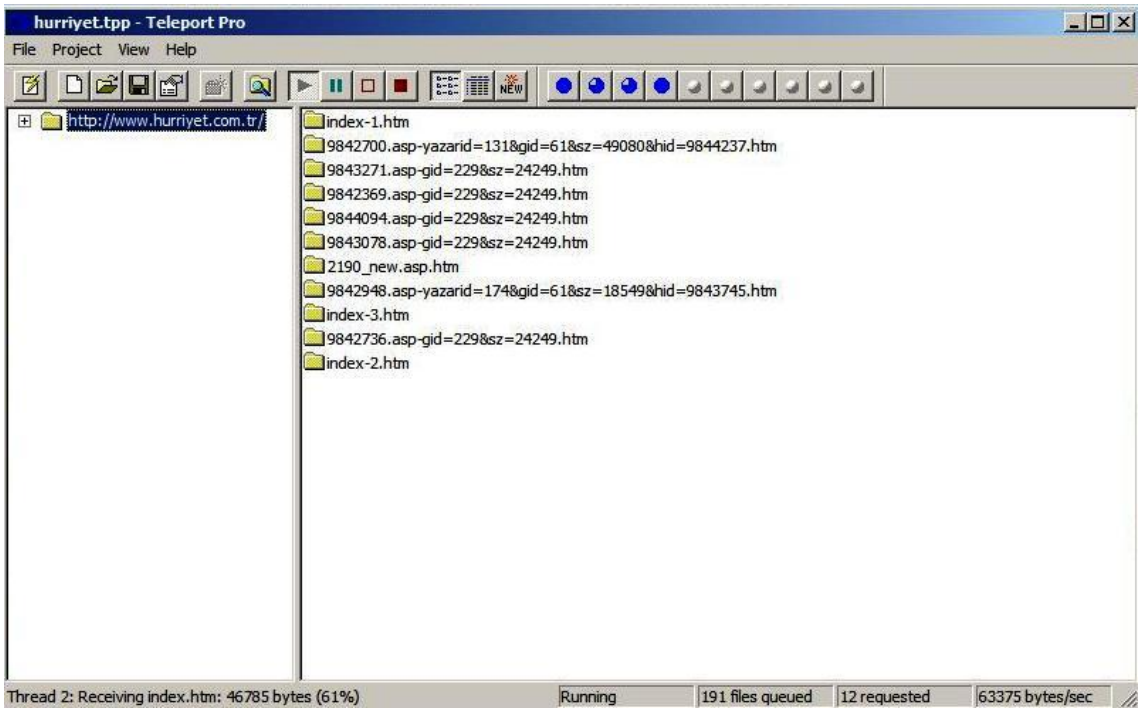
MI tekniği bilgi teorisinden alınmıştır. Bu teknikte, sağında ve solunda birkaç birim bulunan merkez birimin (düğüm-node) dört birim sağı ve dört birim solundaki pozisyona bakılır. Buna *4:4 penceresi* (4:4 window) adı verilir. MI bir eşdizimlilik ölçüsü durumundadır. MI skoru yükseldikçe birimler arasındaki bağın gücü anlamlı kabul edilir ve iki birimin birlikteliğinin (co-occurence) rastgele olmadığı sonucuna ulaşılır. Hunston, MI skorunun 3 ve daha üzeri olmasının iki birimin eşdizimli kabul edilmesi için yeterli olacağını belirtmektedir (age., 2006: 56).

1.6.7. İnternet Üzerinde Derlem Çalışmaları

Tüm dünya üzerinde haberleşmeden basit veri iletimine kadar birçok alanı kapsayan internet, dil çalışmalarının sayısallaştırılması çalışmalarının sonuçlarının da anında kullanıcısına sunulduğu bir araştırma-geliştirme ortamı özelliğini son beş yılda kazanmaya başlamıştır.

Derlemlerin basılı metinlerden elde edilmesinin getirdiği güçlükler karşısında internette yayımlanan özellikle HTML sayfaları, internet tarayıcısının (browser) kopyalama özelliği ile kolayca bilgisayarda saklanabilen veriler olarak kullanılmıştır. Son yıllarda giderek artan oranda makale, kitap formu standardı PDF başta olmak üzere MS Word ve diğer belge formları da derlemlere ham veriler sağlamaktadırlar. Günlük gazete ve dergilerin veri tabanına dayalı yapıları internet tarayıcısında çoğunlukla kopyalanabilir özellikte görüntülenmektedir. Bu nitelik, derlem araştırmacısı için sayfaların tek tek kopyalayabilmesini sağlamakla birlikte geçmişe dönük gazete metinlerinin toplanması ve büyük boyutlu verinin toplanmasının amaç edildiği projelerde kullanışsız duruma gelmektedir. HTML olmayan diğer belge biçimlerinin derlenmesi içinse Google gibi arama motorlarının gelişmiş arama özelliklerinin kullanıcı tarafından bilinmesi gerekmektedir. Google'da " aranacak ifade filetype:pdf", aranacak ifade "filetype:doc" gibi bazı arama motoruna özel deyimlerin girilmesi aranan belgeyi bulabilmektedir. İnternet ortamında bilimsel makalelerin yayımlanmasında PDF belgeleri tercih edilmektedir. Bu durumun yaygınlaşmasında karakter sorunlarının daha az yaşanması, belge biçiminin korunması ve işletim sisteminden bağımsız bir yapısının olması rol oynamaktadır. Derlemin bu türden bilimsel belgeye dayanan verilerden oluşturulması planlanmışsa tam metin yayın yapan özgül bilimsel siteler seçilmektedir.

Derlem verisinin internet üzerinden derlenmesinde kullanılan kopyalama ve belge indirme biçimlerinde kolaylıklar sağlayan yazılımlar bulunmaktadır. Toplu sayfa indirme, sayfadaki belgeleri sitede yer alan bağlantı ağına göre site dizinine göre kullanıcının bilgisayarına indiren yazılımlardan biri Teleport'tur. Bu yazılım aynı anda çok sayıda adresten istenilen bağlantı derinliğinde sayfa ya da belgeyi indirip depolayabilmektedir. Yazılım salt metin verisinden, diğer dosya uzantılarını (.htm, .html, .shtml vb.) desteklemektedir. Araştırmacılar için internetten metin verisi elde etmek, özellikle geniş bant bağlantılı bilgisayarlar ve depolama hacmi yüksek donanımlarla gittikçe kolaylaşmaktadır.



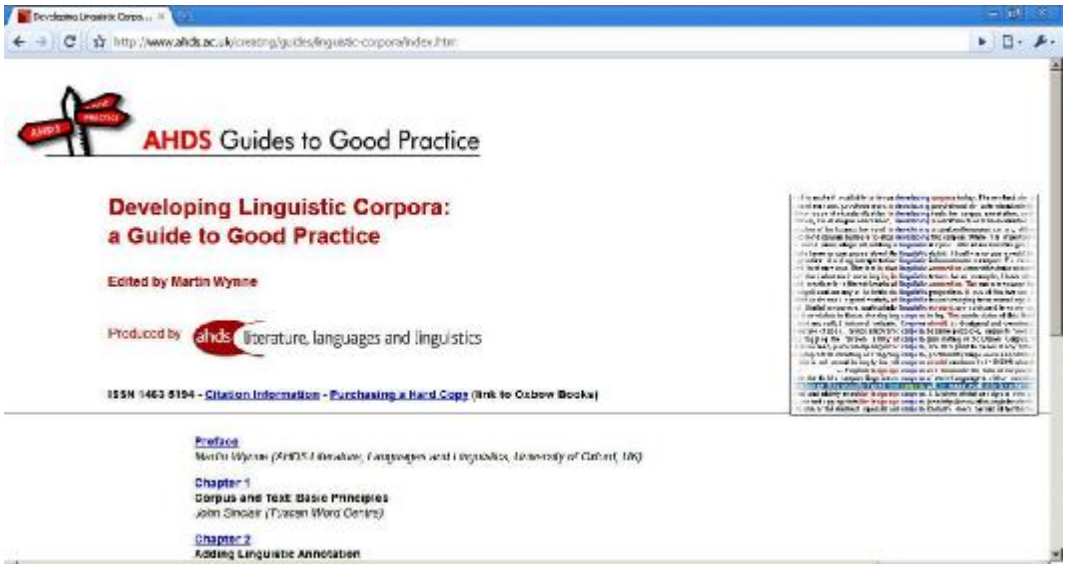
Şekil 50. Teleport Yazılımı Ekran Görüntüsü.

İnternet üzerinde derlem oluşturma yöntemleri ve sorunları üzerine kuramsal bilgi içeren web siteleri yer almaktadır. Bu sitelerde genel olarak derlem tanımının hemen her zaman yer aldığı görülmekle birlikte bazı sitelerde derlem konusunda kullanıcının bilgisinin üst düzeyde olduğu kabul edilerek doğrudan uygulamaya yönelik konular anlatılmaktadır. Hazırlanan sitelerin büyük çoğunluğu İngilizce içerikli ABD ve İngiltere kaynaklı siteler olarak göze çarpmaktadır. Üniversitelere ait dil bilimi ve bilgisayar bilimi bölümleri derlem dil bilimi için ayrı merkezler kurarak bu merkezlerin web sitelerini yayına sokmuşlardır. Üniversite sitelerinde derlem dil bilimi için kurslar ve tanıtımları da bulunmaktadır. Üniversitelerde düzenlenen konuyla ilgili sempozyum duyuruları ve ilgili bağlantılar da güncel olarak sunulmaktadır.

Derlem hakkında kuramsal bilgiye dayanan site bölümlerinde derlemden nicelik ve nitelik araştırmalarının nasıl yapılacağına dair bilgiler özellikle İngilizce metinler üzerinden örneklerle açıklanmaktadır. Bu sitelerin hemen hepsinde derlem için kullanılan yazılım sayfalarına bağlantılar yer almaktadır.

Çevrim içi metin arşivlerinin ve koleksiyonlarının oluşturulması amacıyla İngiltere'de kurulan AHDS (Arts and Humanities Data Sevices) adlı kuruluş kendi bünyelerinde tutulacak derlemlerin oluşturulmasıyla ilgili 6 bölümlük bir kılavuz hazırlamıştır. Derlemde kullanılacak işaretleme ve etiketleme standartları, sözlü dil için

malzeme toplanması ve derlemin kullanıcılar için dağıtımıyla ilgili konular bu kılavuzda ayrıntılı bir biçimde yer alıp kuruluşun sitesi üzerinden yayımlanmaktadır.



Şekil 51. AHDS Derlem Geliştirme Kılavuzu
(http://ads.ahds.ac.uk/project/goodguides/apandrs/)

Cobuild Concordance and Collocations Sampler adlı çevrimiçi yazılım, İngilizce hazırlanan ve 56 milyon etiketlenmiş sözcüğün standardı kendi içinde oluşturulmuş etiketleme sistemiyle sorgulanmasına olanak vermektedir. Sitede verilen örnekte örneğin fool|fools|fooling|fooled/VERB gibi bir sorguda fool sözcüğünün sözcükbiçimleri aranırken "|" karakteriyle bu biçimlerden herhangi birinin sonuç olarak döndürülmesi ve fiil olarak etiketlenenler arasından seçilmesi istenmektedir. Fool sözcüğünün fiil olarak kullanılan bütün biçimleri içinse (fool|fools|fooling|fooled)/VERB. Deyimi kullanılmaktadır[27]. Bu çevrimiçi yazılımın collocation sampler bölümü istatistiksel eşdizimlilik çıkarımında çok kullanılan T testi ve mutual information yöntemi kullanılarak hazırlanmış, iki yöntemin seçenekli sorgu olarak kullanıcıya sunulduğu bir formdan oluşmaktadır. Aşağıdaki örnekte İngilizce "computer" sözcüğünün T testi ile hesaplanan eşdizimlilik sonucundan bir kesit yer almaktadır.

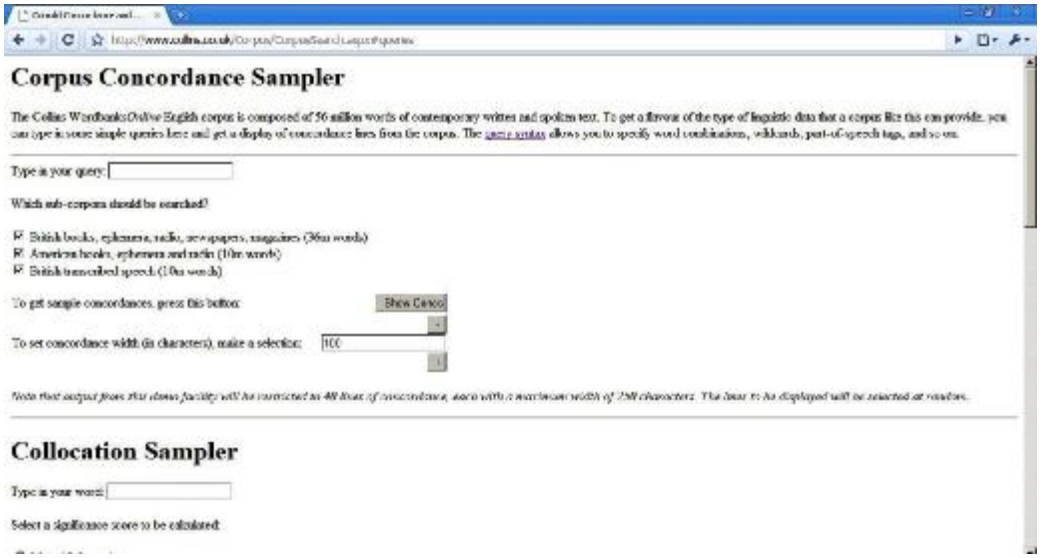
[27] http://www.collins.co.uk/Corpus/CorpusSearch.aspx#queries

Tablo 1. Cobuild Concordance and Collocations Sampler Çevrim içi Eşdizim Çözümlemesi

Collocate	Corpus Freq	Joint Freq	Significance
a	973489	1417	17.967513
system	12514	205	13.652854
systems	3402	177	13.109586
science	4206	127	10.985474
on	307147	459	10.516900
software	1216	103	10.057733
games	4946	106	9.930135
computer	4399	88	9.024057
personal	7230	82	8.447933

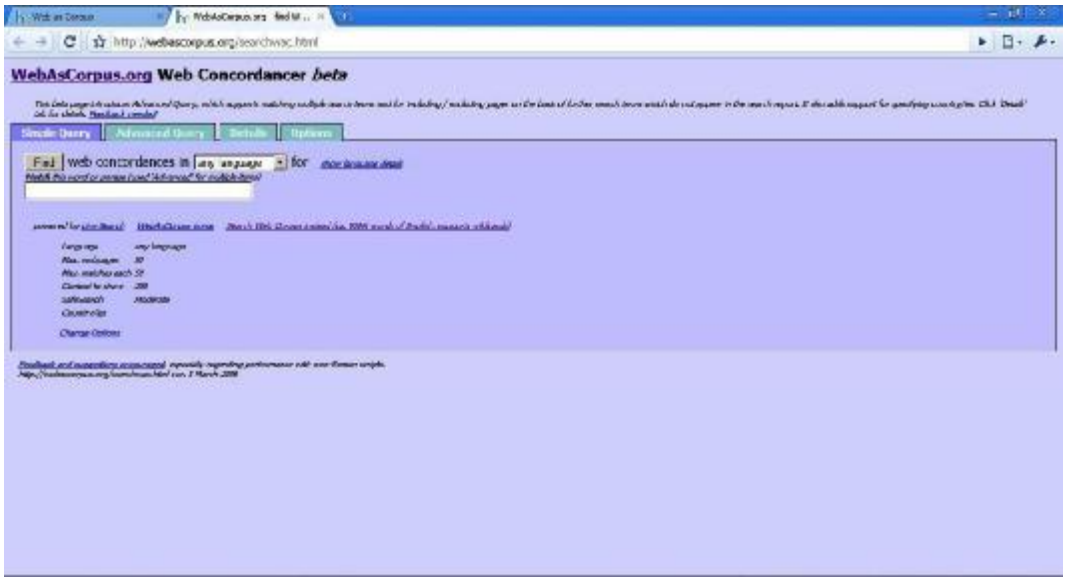
(http://www.collins.co.uk/Corpus/CorpusSearch.aspx#queries)

Yukarıdaki tabloda “computer” sözcüğüyle birlikte sıklığı yüksek sözcükler derlemdeki sıklıkları da dikkate alınarak hesaplanmış ve 56 milyon sözcüklük bu derlemden T testi skoru en yüksek “a” tanımlığının sözcükle eşdizimlilik örüntüsü oluşturduğu anlaşılmıştır.



Şekil 52. Cobuild Concordance and Collocations Sampler Çevrimiçi Derlem Sorgulayıcısı (http://www.collins.co.uk/corpus/CorpusSearch.aspx)

Google arama motoru gibi çalışan ancak interneti başlı başına bir derlem gibi kullanmayı amaçlayan WebAsCorpus, sorgulaması yapılan sözcüklerin geçtiği bağlamları dizinlemekte, aralarında Türkçenin de bulunduğu 34 farklı dildeki web sayfalarını işlemektedir. Yazılımla; mantıksal arama (or, and operatörleri), n-gram sorgulama, sorgulanan sözcüğün değişik arama motorlarında kaç kez geçtiği gibi oldukça gelişmiş sorgulama seçenekleri sunulmaktadır. Arama sonuçları sınırlanmıştır, en fazla 500 sonuç gösterilebilmektedir. Gelişmiş arama seçeneklerinde aranan sözcüğün bağlamında bulunması istenen sözcük, sonuçlarda görüntülenmesi istenmeyen sözcükler, işlenecek en fazla web sayfası gibi yazılımın işleme yeteneğini özgülleştiren yapılar bulunmaktadır. Çeşitli veri tabanlarından terim arama motoru da yazılıma eklenmiştir.



Şekil 53. WebAsCorpus Bağlamlı Dizin Oluşturucusu
(http://webascorpus.org/searchwac.html)

Türkçede derlem kavramı ve derlem dil bilimine yönelik araştırmaların çok yeni olması, bu alanın dil bilimcileri de kapsayan geniş bir araştırma alanı olmasının dünyadaki diğer örnekleri göz önüne alındığında zaman alacağı görülmektedir. İnternet üzerinde derlem ve derlem dil bilimine yönelik Türkçe kaynaklar oldukça sınırlı kalmaktadır.

1.6.8. Türkçe için Derlem Çalışmaları

Türkçenin bilgisayarla işlenmesi üzerine ilk çalışma olarak 1976'da Aydın Köksal'ın otomatik biçim birimsel çözümleme konusunda hazırladığı "Automatic Morphological Analysis of Turkish" adlı doktora tezi öne çıkmakta, 1981'de Sagay'ın İngilizce-Türkçe arasında, 1987'de de Stoop'un Almanca ve Türkçe arasında otomatik çeviri sistemi hazırlama çalışmaları ilk çalışmalar olarak görülmektedir. 1990'lı yıllarla birlikte bilgisayar destekli dil bilimi çalışmalarının tüm dünyada internet'in gelişmesiyle Türkçe bilgisayar destekli çalışmalarının da proje düzeyinde yapılmaya başlandığı görülür. Bu çerçevede 1994'te NATO projesi olarak Oflazer ve Bozşahin'in Türkçe doğal dil işleme inisiyatifiyle ilgili projeleriyle hızlanan çalışmalar, 2000 yılında Tür'ün Türkçe için ilk istatistiksel çıkarım yöntemine dayalı doktora çalışmasıyla bir başka açıdan ele alınmıştır. Biçim birimsel ve çeviri ağırlıklı çalışmaların yanında 2001 yılında Yöndem'in hazırladığı Türkçe için söylem düzeyinde bölümleme konusundaki doktora çalışmasıyla doğal dil işleme ve derlem dil bilimi alanı, tüm dünyada olduğu gibi Türkiye'de de öncelikle bilgisayar bilimciler ve dil bilimciler arasında tanınmaya başlamıştır. Yukarıda sözü edilen çalışmaların büyük bölümünde kural tabanlı yöntemler kullanılmıştır (Yöndem, 2006: 45).

Geleneksel dil çalışmalarında metne dayalı çalışmalar öteden beri yapılmakla birlikte, Türkiye'de, tek tek metinlerden değil de metin bütünlerinden topluca yararlanmak, bütünler arası karşılaştırma yapmak gibi derlem çalışmalarında ayrı ilgi ve çalışma alanı konuları üzerine kapsamlı olarak çalışmalar yeni başlamıştır. Say (2006: 82), dil bilimciler arasında yapılan bir ankette, araştırmacıların %41'inin çalışmalarının derleme dayalı olduğunun ortaya çıktığını belirtmektedir. Derlem olarak nitelenen bu çalışmaların büyük çoğunluğu, kişisel çabalarla elektronik formlar dışında toplanmış ve paylaşıma açık olmayan yapıdadır.

Günümüzde derlem kullanım amaçlarından biri doğal dil işleme uygulamalarına gereksinim duyduğu veri tabanını sağlamaktır. Etiketlemesi gerçekleştirilmiş bir derlem DDİ uygulamalarının hata oranını düşürmektedir. Türkçe DDİ uygulama çalışmalarına kaynaklık edebilecek dengeli ve temsil gücü hesaplanmış ilk derlem Orta Doğu Teknik Üniversitesinde (ODTÜ) hazırlanan ve 1999-2003 yılları arasında tamamlanan ODTÜ Türkçe Derlemi'dir. Derlem, 1990 sonrasında toplanan yazılı kaynaklara dayalı metinlerden oluşmaktadır. Derlem genel amaçlı bir derlem olarak nitelenmiş, sözlü dile dayanan örnekler bir araya getirilmemiştir. Derlem dil bilimi literatüründe denge ve

temsil gücüne yönelik hazırlama ölçütleri de bu derlemin hazırlanmasında göz önüne alınmış, böylece 14 metin türü her biri 2000 sözcüklük 999 örneklem (201 kitap, 87 dergi 3 günlük gazete) bir araya getirilerek toplam 2 milyon etiketlenmiş sözcüklük derlem oluşturulmuştur.

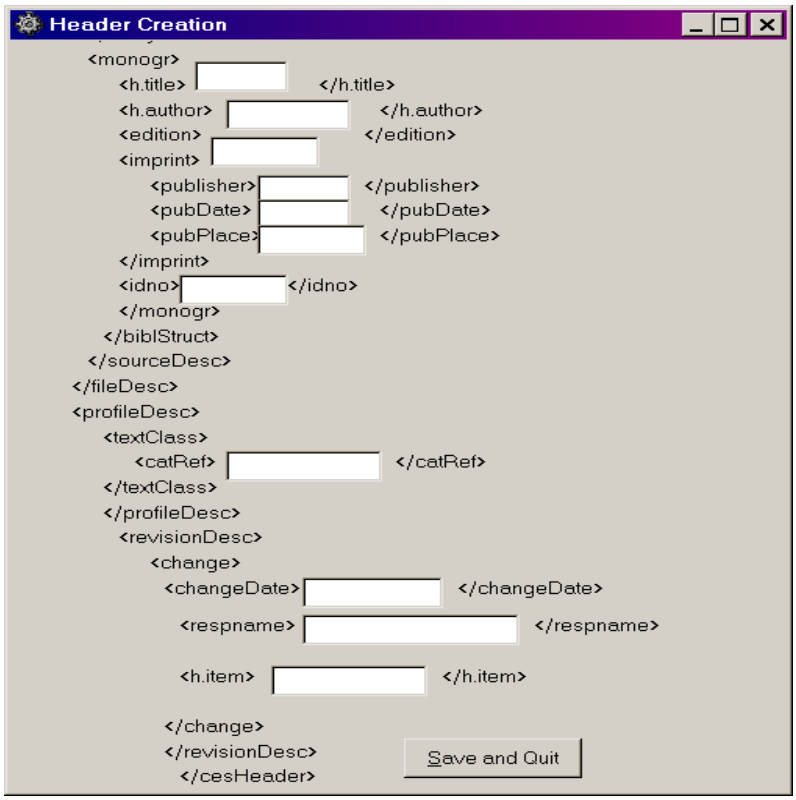
Tablo 2. ODTÜ Türkçe Derlemi Metin Türleri Dağılımı (Say, 2006: 83)

Köşe Yazısı	8%
Haber	42%
Roman	13%
Öykü	11%
Makale	8%
Deneme	7%
Araştırma	5%
Gezi	2%
Söyleşi	1%
Diğer	3%
	100%

Derlem için metin toplama, işaretleme ve sorgulamaya yönelik yazılımın geliştirilmesi çalışma planı izlenmiştir. Metinler Unicode salt metin olarak kaydedilmiş, derlemde yer alan sözcükler için Oflazer'in 1994'te hazırladığı biçim birimsel çözümleyici kullanılmıştır. XEROX Araştırma Merkezi Sonlu Durum Araçları kural tabanlı çözümleyici olarak biçim birimsel çözümlemede kullanılmıştır. Derlem daha sonra Sabancı Üniversitesi'yle ortaklaşa yürütülen bir proje kapsamında ODTÜ-Sabancı Ağaç Yapılı Derlemi adını almıştır. Bu ortak proje kapsamında 7300 cümle ve 65.000 sözcük etiketlenmiştir.

Biçimbirimsel çözümlemede birden çok sonuç veren yapılar için insan kontrolüne dayalı bir etiketleme standardı oluşturulan işaretleme yazılımında yer alarak belirlenen XCES XML tabanlı derlem etiketleme standardı yönergesi izlenerek etiketlemenin son olarak varsa hatalarının düzeltilmesi etiketleyici kişiye bırakılmıştır. Dil bilimciler için sorgulama yazılımında düzenli ifadeler ve mantıksal sorgulama (ve, veya) seçenekleri oluşturulmuştur. Biçim birimsel çözümleme dışında cümle düzeyinde otomatik çözümleme için de bağlılık grameri adı verilen ve Türkçe için çekim öbeklerinin belirlendiği bir form göz önüne alınmıştır (Say, 2006: 85). Derlem ve

yazılımının internet ortamında araştırmacılara ücretsiz olarak paylaşıma açık biçimi bulunmaktadır.



Şekil 49. ODTÜ Türkçe Derlemi İşaretleme Yazılımı İşaretleme Formu (http://www.tdk.gov.tr/TR/dosyagoster.aspx?DIL=1&BELGEANAH=2908&DOSYAISIM=sunum06.ppt)

Türkçe için derlem geliştirme çalışmalarında Çebi ve Varlıkların (2006) oluşturduğu ve TurCo adı verilen işaretlenmemiş derlemdir. TurCo'nun internet üzerinden toplanan bölümü 10 farklı web sitesinin internet tarayıcı yazılımının salt metin kaydetme özelliği kullanılarak oluşturulmuştur. Bu işlemde en büyük sorun olarak site sayfalarında geçen bağlantı bilgileri (hyperlink), tekrarlanan menü adları gösterilmiştir (Çebi ve Varlıklar, 2006: 60).

Derlemdeki toplam sözcük sayısı 50.111.828 olarak verilmiştir. Farklı sözcük sayısı 686.804 olarak çıkarılmıştır. Derlemin %90,40'ı internette yayımlanan gazete, dergi gibi güncel yayınlar ve Türkiye Büyük Millet Meclisi'nin internet sitesinde yayımlanmış metinlerinden ve Devlet İstatistik Enstitüsü'nün web sayfaları taranarak elde edilmiştir. Derlemin bilgisayar ortamındaki toplam büyüklüğü 362 MB'dir. TBMM tutanaklarının yazılı ve sözlü Türkçeyi temsil ettiği derlem bilgisi olarak belirtilmektedir.

Tablo 3. TurCo'da Sözcüklerin Dağılım Oranları (Çebi ve Varlıklar, 2006: 59).

Bölüm No	**Farklı Kelime Sayıları**	**Farklı Kelime Oranı (%)**	**Kelime Sayıları**	**Kelime Sayıları Oranı (%)**	**Farklı Kelime Kullanım Oranı (%)**
1	342.544	27,74	23.396.817	46,69	1,46
2	255.024	20,65	9.746.093	19,45	2,62
3	99.432	8,05	9.415.716	18,79	1,06
4	309.030	25,02	4.668.306	9,32	6,62
5	20.760	1,68	948.116	1,89	2,19
6	42.208	3,42	753.571	1,50	5,60
7	46.743	3,78	527.757	1,05	8,86
8	29.228	2,37	203.620	0,41	14,35
9	13.103	1,06	160.562	0,32	8,16
10	37.057	3,00	135.519	0,27	27,34
11	25.294	2,05	96.857	0,19	26,11
12	14.633	1,18	58.894	0,12	24,85
Toplam	**1.235.056**	**100,00**		**100,00**	
TurCo	**686.804**		**50.111.828**		**1,37**

Derlemde kirlilik yaratacak unsurların yer alması, derlem ekibini ilk olarak metinlerde cümle sonu bulma algoritmalarını yazmaya yöneltmiştir. Bunun dışında sözcük kök ve gövdelerinin bulunması, sözcük türlerinin belirlenmesi ve veri tabanlarının oluşturulması çalışmalarının derlem için düşünüldüğü de belirtilmektedir. Derlem için veri tabanı yönetim sistemi seçiminde MS-SQL ve MySQL gibi popüler yazılımların, doğal dil işleme çalışmalarının kendi özel sorunları ve çözüm gereksinimleri gereğince yetersiz kaldıkları, XML gibi işaretleme standartlarının bu nedenle TurCO'nun işaretlenmesinde kullanılacağı vurgulanmıştır (Çebi ve Varlıklar, 2006: 67).

Derlem geliştirme sürecinde karşılaşılan güçlüklere karşı TurCo ekibinin önerdiği çözümlerin başında dil bilimcilerle bilişim araştırmacılarının iş birliğinin sağlanması gerektiğidir. Bu konuyla ilgili olarak Türk Dil Kurumu'nun 2005 yılında Türkiye'de ilk olarak bilişim bilimcilerle dil bilimcileri bir araya getiren Bilgisayar Destekli Dil Bilimi Çalıştayı düzenlenmiştir. Çalıştay süresince Türkçe üzerine bilgisayarda geliştirilen derlemlerle, genel olarak DDİ uygulama alanları tartışılmıştır.

1.7. Veri Tabanı

1.7.1. Veri, Enformasyon, Bilgi ve Veri Tabanı Kavramı

Günümüz bilgi ve iletişim teknolojilerinin üretilmesinde ve geliştirilmesinde *veri* (data), verinin elde edilmesi, saklanması ve işlenmesinde artık veri tabanı sistemlerinin vazgeçilmez bir yeri vardır. Bilgisayar sistemlerinin gelişmeye başladığı 1960'lı yıllardan bu yana çeşitli veri tabanı sistemleri geliştirilmiş, bunlardan bugün en çok kullanılan ve tercih edilen veri tabanı modeli "ilişkisel veri tabanı" modeli olmuştur.

Veri tabanı (database) kavramının anlaşılması için öncelikle veri, enformasyon ve bilgi sözlerinin anlaşılması gerekmektedir. Bu kavramlardan öncelikli olarak verinin, üzerinde işlem yapılan ve bir çözümleme sürecinden sonra elde edilen bilginin temel dayanağı olduğu anlaşılmalıdır.

GTS'de "veri" sözünün çok anlamlı olarak verilen tanımında, temel anlam olarak verilen "Bir araştırmanın, bir tartışmanın, bir muhakemenin temeli olan ana öge, muta, done." tanımı, verinin veri tabanı sistemlerinin ana ögesi olduğunu ortaya koymaktadır. Aynı sözlükte bilişim terimi [28]olarak verilen " Olgu, kavram veya komutların, iletişim, yorum ve işlem için elverişli biçimli gösterimi" tanım ise daha çok verinin modellenmesi sürecini gösteren bir tanımdır. Çağıltay ve Tokdemir (2010), veriyi "olaylar ve varlıklar ile ilgili kaydedilebilir, dolaylı olarak bir anlam ifade eden ve işlenmemiş (ham) gerçekler ve ölçümler" olarak ele almaktadır. Bu bağlamda veri, kaydedilebilir nitelikte olup işlenmediği sürece bilgi gibi açıkça bir anlam ifade etmeyen salt verili gerçekler olarak görünmektedir (Çağıltay ve Tokdemir, 2010: 8:)

Enformasyon sözü için GTS'de verilen "Danışma, tanıtma." ve "Haber alma, haber verme, haberleşme" tanımlarının, veri tabanı sistemleri için kullanılan terim nitelikli anlamı dışında olduğu görülmektedir. Enformasyonun bilgiye ulaşım sürecinde Çağıltay ve Tokdemir'in ele aldığı biçimde "karar vermek için bir değeri olan ve organize edilmiş verilen özetlenmesi ile edle edilen gerçeklerdir." tanımının veri tabanı sistemlerinin bir parçası olduğu görülmektedir. Bilginin çıkarımında bir ara kesit niteliğindeki enformasyonun verinin çözümlenmeye başladığı bölüm olarak ele alınması enformasyonu oluşturmaktadır (Çağıltay ve Tokdemir, 2010: 8:)

---

[28]http://tdk.gov.tr/TR/Genel/SozBul.aspx?F6E10F8892433CFFAAF6AA849816B2EF4376734BED947CDE&Kelime=veri

Bilgi (knowledge) sözü için GTS'de[29] terim anlamı olarak "Kurallardan yararlanarak kişinin veriye yönelttiği anlam." tanımı verilmiştir. Veri-enformasyon ikilisinin son bileşeni olarak bilgi için, Çağıltay ve Tokdemir'de "enformasyon haline dönüştürülmüş gerçeklerin analiz eilmesi ve sentezlenmesi sonucu karak vermeye yönelik olarak elde edilen daha üst seviyeli gerçekleri içerir" açıklaması getirilmiştir. Bilgi bu bağlamda elde edilen ve eyleme yönelik son çıktı, ürün olarak gerçeğin yorumu olarak ele alınabilir (Çağıltay ve Tokdemir, 2010: 9).

Veri, enformasyon ve bilgiyle ilgili hiyerarşik bir üçgen aşağıdaki şekilde gösterilmiştir. Yukarıdaki tanımlar uyarınca bu üçgende Çağıltay ve Tokdemir üçgenin solunda, bilgiye ulaşım eylemlerini sıralamıştır.



Şekil 54. Bilginin Elde Edilmesi Süreçleri (Çağıltay ve Tokdemir, 2010: 7).

İnsanoğlunun salt dünya gerçekliğini düzensiz bir ortamdan gruplayarak soyut ve kendi içinde karmaşık bir bilgi ağına dönüştürme çabaları yazı sistemlerinin geliştirilmesinden günümüz bilgi destekli karar verme sistemlerine ve yapay zeka uygulamalarına taşınmıştır. Veri tabanları yapay zeka uygulamalarının yoğun olarak kullanıldığı sektörlerin yanında, bankacılıktan alışverişe, coğrafi bilgi sistemlerine kadar verinin değerli olduğu tüm iş kollarına destek sağlamaktadır.

Veri tabanı birçok biçimde tanımlanmaktadır. GTS'de[30] "Bilgisayar kullanımında çözüme erişmek için işlenebilir duruma getirilmiş bilgi ortamı." olarak tanımlanan bu kavramı Yarımağan, "Bir ya da birkaç uygulamada kullanılmak için, gereksiz yinelemelerden arınmış olarak, düzenli biçimde bilgisayar belleklerinde

[29] http://tdk.gov.tr/TR/Genel/SozBul.aspx?F6E10F8892433CFFAAF6AA849816B2EF4376734BED947CDE&Kelime=bilgi
[30] http://tdk.gov.tr/TR/Genel/SozBul.aspx?F6E10F8892433CFFAAF6AA849816B2EF4376734BED947CDE&Kelime=veri%20taban%C4%B1

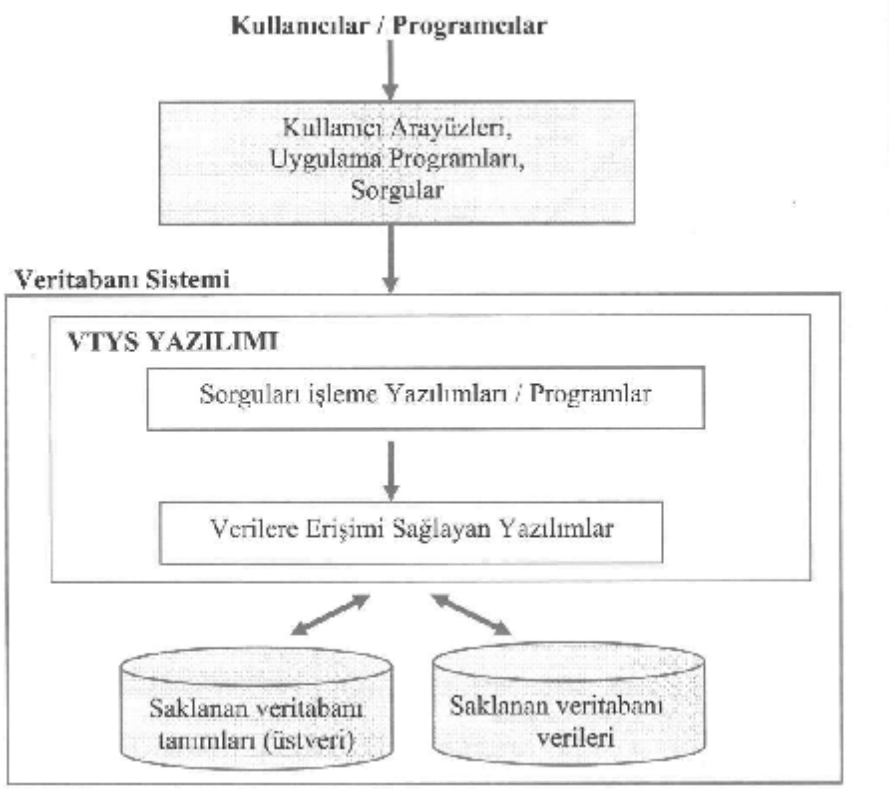
saklanan birbiriyle ilişkili veriler topluluğu" olarak tanımladıktan sonra bu genel tanımın eksik olduğunu veri tabanı sistemlerinin değişen uygulamalara bağlı olarak çeşitlilik gösterdiğini belirtmiştir (Yarımağan, 2000: 1).

Çağıltay ve Tokdemir, veri tabanını "verilere kolaylıkla erişilmesi ve organize edilmesini sağlayan yapılar" olarak tanımladıktan sonra veri tabanının gerçek dünyadaki olayların ve ilişkilerinin modellenmesini sağladığını belirtir. Veri tabanında benzer özelliklere sahip veriler karmaşık bir dağınık yapıdan aynı niteliklere sahip kümeler biçiminde tablolar içinde saklanırlar. Bu anlamda gruplanarak oluşturulan bu yapı bir gerçek dünya modelinin tablolar biçimde saklanması olarak da düşünülebilir, ancak küçük dünya (mini-world) terimiyle veri tabanında gerçek dünya modelinin yer alması sağlanırken dünyadaki bütün olayların ve niteliklerin bu yapı içinde ele alınamayacağı da belirtilmektedir (Çağıltay ve Tokdemir, 2010: 14, 17). Veri tabanı, ilişkili verilerin oluşturduğu bir koleksiyon olarak da tanımlanmaktadır. Gerçek dünya verilerinin yanında veri tabanının kendi iç dinamiğine ait ***üst verilerin*** (metadata) de içinde barındırıldığı yapıda verilerin sıradüzeni, ilişkilerin tutulma biçimleri de yer alır (Kaya ve Tekin, 2007: 16-17).

1960'lı yıllardan başlayarak veri tabanı mimarilerinin gelişiminde ilk olarak ağ modeli (network model) olarak adlandırılan model geliştirilmiştir. Bu model verinin kümeler ve her kümenin bir üst kayıt adıyla birlikte temsil edildiği kümelerde yer alan üyelerin birden fazla küme içinde bulunabildiği bir modeldir. Model Unisys, Hp gibi değişik platformlarda kullanılmıştır. IBM firmasının geliştirdiği hiyerarşik ya da sıradüzen modelinde (Hierarchical model) veriler ağaç yapısı biçiminde tutulmaktadır. Yapıda üst (parent) ve ast (child) veriler bulunur. Ebeveyn-çocuk sıradüzeni olarak XML teknolojisinde olduğu gibi ast ve üst birimler arasında bağıntılar kurulur. 1970'li yıllarda bugün en çok kullanılan ve E. F. Codd'un geliştirdiği ilişkisel modelde (relational model) veriler tablolar biçimde saklanmakta, her tablonun bir birincil anahtarıyla (primary key) veriler bir diğerinden benzersiz biçimde ayrılarak tablolar arası ilişkiler yabancı anahtar (foreign key) ile sağlanmaktadır. Önceleri özellikle uygulama geliştirme sürecinde kullanılan klasik dosyalara kaydedilen ve veri tekrarı ile düzensiz bir veri yığını biçimindeki veriler ilişkisel modelle gereksiz tekrarlardan arındırılıp benzer nitelikli gruplamalarla daha kolay ve hızlı bir erişime olanak sağlamıştır (Çağıltay ve Tokdemir, 2010: 26).

Veri tabanı üzerinde verilerin barındırılacağı tabloların oluşturulması, veri tiplerinin belirlenmesi, silme, ekleme ve değiştirme gibi işlemler için veri tabanı

yönetim sistemi (VTYS) adı verilen sistemler kullanılır. Bu sistemlerde veri üzerinde gerçekleştirilecek işlemler bir arayüz üzerinde veri tabanı yöneticisinin kontrolünde yapılır. Kullanıcı arayüzlerinin oluşturulması, raporların hazırlanması için içlerinde çeşitli uygulama geliştirme araçları barındırırlar. Günümüzde en çok kullanılan VTYS'ler SQL Server, MS Access, MySql, Oracle, Sybase gibi kullanıcı ve yöneticiler için hem çevrimdışı hem de çevrimiçi hizmetlerde kullanılan yazılımlardır. Aşağıdaki şekilde bir VTYS'nin yapısı görülmektedir.



Şekil 55. VTYS ve Veri Tabanı (Çağıltay ve Tokdemir, 2010: 16)

Bir veri tabanı sisteminin temel özellikleri Çağıltay ve Tokdemir 2010'da 12 başlık altında ele alınmıştır. Buna göre;

1- Bir veri tabanı sistemi kendi içeriğine sahiptir. Veri tabanındaki verilerin bilgisini barındıran katalog bilgisi ile üstveri (metadata) bilgisi VTYS'ler aracılığıyla kolayca paylaşılır. Farklı VTYS'ler arasında bilgi alışverişi platform bağımsız olarak taşınabilir.

2- VTYS içinde programlar ile veriler birbirinden ayrıştılır. Son kullanıcı için geliştirilen uygulamalar ile kullanıcıya ulaşmaması gereken ve güvenlik amacıyla tutulan yapılar birbirinden bağımsız olarak saklanabilir.

3- Veriler arasındaki bağıntılar özet veri (data abstraction) özelliğiyle uzun uğraşlara gerek kalmadan görülebilir.

4- Kullanıcılar için verilere farklı bakış açıları sağlanabilir. VTYS'lerde bulunan ve görünüm (view) özelliğiyle eklenecek işlemler sonucu farklı kullanıcılar veri tabanına kendi gereksinimlerine göre ulaşabilirler.

5- Tekrar edilen verilerin kontrolü sağlanmış olur. Böylece veri tutarsızlığının önüne geçilir. VTYS'ler oluşturulacak kurallar sayesinde veri tekrarının önüne geçilmesini sağlayabilirler.

6- Veriler çok sayıda kullanıcı tarafından ulaşılabilir duruma getirilebilir. VTYS'ler aynı anda çok sayıda kullanıcının erişimine izin verilmesini ve kullanıcı kontrolünün sağlanmasını sağlarlar.

7- VTYS'ler erişim kontrolünü sağlayarak yetkisiz kullanıcı girişlerini önceleyebilir.

8- Veriler arasındaki karmaşık bağıntıların yönetilebilmesi yüklü miktarda veri taşıyan sistemler için sorun olmakla birlikte VTYS'lerde bulunan çeşitli gösterim araçları ile bu karmaşanın akış modelleri ortaya konularak karmaşanın kontrolü sağlanabilir.

9- Verilerin yedeklenmesi ve veri kaybının önlenebilmesi için VTYS'ler kurtarma ve yedekleme araçları barındırırlar.

10- VTYS'ler değişik platformlarda çalışan sistemleri belirli bir strandart kapsamında birleştirilebilir.

11- Sorgulamaların hızlı ve etkin bir biçimde sağlanması için VTYS'ler eniyileme (optmizasyon) araçları barındırabilirler. Böylece son kullanıcılar açısından verinin çağrılma zamanında iyileşme sağlanabilir.

12- Programcılar için küçük program parçaları ve nesneler VTYS'lerde tutulabilir. Böylece programcı, veri tabanıyla ilgili yazılım geliştirme sürecini daha etkin biçimde yönetebilir (Çağıltay ve Tokdemir, 2010: 17, 20).

1.7.2. Dilsel Veri Tabanları

Verinin değişik biçimlerde temsil edilmesi ve gösterilmesi, bilginin insan için hem ulaşılması hem de işlenmesi açısından önem taşımakla birlikte, birçok veri türünün bir yerde bulunması, yer alması kısaca depolanması da verinin bilgiye dönüşmesi bakımından önem taşımaktadır.

Bilişim uygulamalarında kullanılan ya da genel ağ (internet) ortamında yer alan verinin temel olarak metin odaklı, yazıya dayalı olduğu görülmektedir. Genel ağdaki metinlerin birimleri doğal olarak dil birimleridir. Kullanılan her söz, sayısallaşmış bu metinlerin işlenebilir ana birimleridir. Günümüz genel ağındaki görsel ögelerin başta resim biçimi olmak üzere video biçiminde de gittikçe artması, yazının bu ögelerle birlikte kullanımını da artırmaktadır. Bu bakımdan internet, metin türü veri tabanları için her zamankinden daha önemli bir dil araştırması kaynağı durumuna gelmiştir.

Dil çalışmalarında kullanılmak üzere oluşturulan özel veri tabanları bulunmaktadır. Bunlardan ses bilimi çalışmaları için UCLA Phonetics Laboratory 1996 en çok bilinen veri tabanları arasında yer almaktadır. Dil biliminin yalnızca biçim bilimi, ses bilimi, söz dizimi çalışmaları yanında tarihsel dil bilimi araştırmalarında değil, dil biliminin uygulamalı alanı sözlük bilimi için de veri tabanları hazırlanmıştır (Nerbonne, 1998: 2)

Dil araştırmalarının amaçlandığı uygulamalarda veri tabanı terimi, dil verisinin elektronik kayıtlarından oluşan koleksiyon anlamına gelmektedir. Cümlelerin bulunduğu bir dosya ya da dosyalar seti basitçe bir veri tabanı olabilmekle birlikte, sonradan işlenmek, anlamlı sonuçlar çıkarmak üzere cümlelerin ya da sözcüklerin ayrıldığı bildirimli (declarative) yapılar da veri tabanı olabilmektedir (age., 1998: 2)

Nerbonne (1998), DDİ'de kullanılan cümle ayırıcı (ing. parser) ve çeşitli dilsel yapıları otomatik üreten araçların aynı veriyi kullanmalarına karşın bir veri tabanı sayılamayacağını belirtmektedir. Oluşturulacak veri tabanında kullanılacak verinin ne kadarının otantik ya da doğal olması gerektiği, ne kadarının sezgisel olarak yer alması konusunda da tartışmalar sürmektedir.

Simons, dilsel veri tabalarını; özel amaçlı dilsel veri tabanı yazılımları, genel amaçlı ilişkisel veri tabanları ve genel amaçlı *nesne yönelimli* (object-oriented) veri tabanı sistemleri olarak üçe ayırır. Özel amaçlı veri tabanı yazılımları içinde, bütün dil araştırmacılarına yönelik kullanımı kolay yazılımları örnek verir. IT (Interlinear Text processor), Shoebox, LinguaLinks Simons'ın sözünü ettiği yazılımlardan olup, satır arası metin çözümlemelerinde ve sözlüksel bilginin depolanıp sorgulanmasında kullanılabilmeleri açısından çıkmaktadırlar (http://www.sil.org/computing/routledge/simons/databases.html).

Günümüzde en çok tercih edilen veri tabanı modeli ilişkisel veri tabanı modelidir. Dil verisinin çok boyutluluğu ve karmaşıklığı ilişkisel model açısından avantaj sağlamasına karşın, verinin emek yoğun bir ön işlemeden geçmesinin gerekliliği

bu modellerin dezavantajı olarak görülebilir. İlişkisel model ile Waterloo Üniversitesi'nde bulunan New Oxford English Dictionary and Text Research adlı merkez Text/Relational Database Management System Project başlıklı bir proje çalışması yürütmektedir. İlişkisel veritabanı yönetim sistemleri için Visual FoxPro, Oracle, Sybase, Informix, Paradox, MS SQl, MySql, PostgreSql örnek olarak verilebilir.

Nesne yönelimli (object oriented) veri tabanı sistemlerinde, günümüz programlama mantığında önemli bir yeri olan nesne ve sınıf kavramları kullanılır. Nesne olarak adlandırılan veri tipi, programlama sırasında çağrılıp işleme sokulan sınıfların bir üyesi olarak görev alır. Nesneler ilişkisel modeldeki alanlar olarak düşünülebilir. Veriler nesnenin içine alınarak bir şekilde "kapsüllen"miş olmaktadır. Bu nesnelere çeşitli methodlar (yöntem) kullanılarak ulaşılabilmektedir. Bu modele örnek olarak CELLAR (Computing Environment for Linguistic, Literary, and Anthropological Research) adlı çok dilli olarak oluşturulan sistem verilebilir.[31]

Veri tabanına dayalı çıkarımlardan çoğunlukla DDİ'de yararlanılmaktadır. Verilerin düzenli biçimde yapılandırılmasının zorunluluğu, sağlam (robust) uygulamaların çok titiz ve sağlam veriler üzerine kurulmasının gerekliliği DDİ'deki veri tabanlarının titizlikle hazırlanmasını açıklamaktadır. Dilsel yapıların *işaretlenmesi* (annotation), tutarlılıklarının sağlanmış olması, çok dilli olarak oluşturulmaları da DDİ'de tercih edilmelerini sağlamaktadır (Oepen vd., 1998: 13)

Oepen ve arkadaşları TSNLP (Test Suites for Natural Language Processing) adlı dilsel veri tabanı sistemi geliştirmişlerdir. Bu sistemin geliştirilmesindeki amaçlar şu biçimde özetlenmiştir:

- Dil veri tabanlarının oluşturulmasında tutarlı ve sistematik koleksiyonların elde edilmesine kılavuzluk etmek.
- Değişik dil kuramlarını tarafsız bir biçimde göz önünde tutarak *işaretleme şeması* (annotation schema) oluşturmak.
- Değişik DDİ uygulamaları için (cümle ayırıcıları, dil bilgisi denetimi, dil tanımlayıcıları) İngilizce, Fransızca ve Almanca temel test amaçlı parçaları (ögeleri) oluşturmak.
- Test verisini depolamak, bu veriye ulaşmak ve değiştirmek genişletilebilir bir dilsel veri tabanı tasarımı sağlamak.

[31] http://www.sil.org/computing/routledge/simons/databases.html

- yarı otomatik uygulamalar için değişik yöntemleri ve araçları gözden geçirmek (Oepen vd., 1998: 14).

Deutsch ve arkadaşları (1998), genel amaçlı veri tabanları dışında, ses örneklerinin depolanıp işlenebildiği The S_Tools Database-Management System (STDBMS) adlı bir sistem geliştirmek için çalışmışlardır. Sistem bir *yerel ağ* (LAN) ve Metropolitan ağı (Metropilitan Area Network) (LAN/MAN) adı verilen ağa bağlı olarak *gerçek zamanlı* (real-time) veri girişine olanak sağlamaktadır. Sistemde yerel ağa bağlı olan *istemci*lerdeki (client) çevirici cihazlar yardımıyla çevreden kaynaklanan gürültünün en az indirilmesi amaçlanmış, çeşitli noktalardan aynı anda birçok kişinin ses kaydı yapabilmesi için sunucu kurulmuştur. Gerçek zamanlı ses verisi kaydı yapmanın işlem gücü ve bellek açısından güçlü bir sunucu bilgisayar gerektirmesi, bu gibi *çoklu ortam* (multi media) sunucu mimarilerinin, sadece metin depolanan genel amaçlı *veri tabanı yönetim sistemleri*nin (VTYS) çalıştırıldığı sunuculardan farklı yanlarını ortaya koymaktadır. Sistemde orta düzeyli 50 ile 200 saat arası bir kayıt için 100 terabayt ve daha fazlası depolama alanının gerektiği belirtilmiştir.

Sistemin, ses kaydı girişinin tamamlanmasının ardından ses verisinin bölümlenmesi aşaması için editör arayüzü de hazırlanmıştır. Bu arayüzde her bir bürünsel özellik için başlangıç ve bitiş noktalarının işaretlenip, arka planda çalışan ses dosyası bölümüne kaydı sağlanmaktadır. Sistem üzerinden otomatik içerik tanımlama çalışmasının yapılabilmesi için üç düzeyli bir düzeneğin gerekli olduğu belirtilmektedir. Bunlar sırayla, belgenin hiyerarşik tipi, kavramsal yapısı ve belge içeriğinin gösterimidir. Belgeye dayalı çözümlemede olduğu gibi sesbilimsel çözümleme de kendi birimleri sesliler, sessizler, prozodik nitelikler gibi bölümler (segment) için aday sınırlar olabilmelidir. Bu ses niteliklerine dayalı sınırların hatasız olarak belirlenmesi için kaliteli, gürültü düzeyi son derece düşük kayıtların yapılmış olması gerekmektedir. Sinyal işleme teknolojilerindeki gelişmeler, eskiden karmaşık ve zaman alan algoritmaların bugün daha hızlı olarak işlenmesine olanak vermektedir (Deutsch vd.,1998: 80).

STDBMS'nin oluşturulmasında iki farklı derlem kullanılmıştır. Ölçünlü Avusturya-Alman Derlemi (Standard Austrian German Corpus), ölçünlü cümle listelerinin ana dili Almanca olan Avusturyalı kişilere okutturulmasıyla hazırlanmıştır. 5 ile 10 arasında değişen sayıda sözcükten oluşan vurgu, ton, diftong gibi ses bilimsel birimler dikkate alınarak özenle seçilmiş cümleler, konuşucuların doğal ortamıyla, ses yalıtımlı özel odalarda kaydedilmiştir. Diğer derlem, Adli Söylem Veri Derlemi

(Forensic Speech Data Corpus), telefon konuşmalarının çeşitli suç ortamlarından elde edilmiştir. Toplanan telefon konuşmaları sonradan yüksek kalitede çözümleme yapan laboratuarlarda karşılaştırma amacıyla tekrar işlenmiştir. Derlem, Alman vatandaşı Arnavurt, Sırp, Hırvat, Rus, Polnyalı ve İtalyanlar arasından seçilerek oluşturulmuştur. Toplam 220 kişi, 36.000 ses sinyali bölümü (segment) kayıt sağlamış, 10 ile 400 arasında söz türü belirlenerek bu kayıtlarda geçen ses birimler bölümlenerek ele alınmıştır (age., 1998: 84).

Verilerin yapılandırılışı açısından STDBMS şu alanlardan oluşmuştur: Kayıt oturum kimliği, ana kimlik, veri tabanı kimliği ve oturum tipi, çalışma istasyonu kimliği, derlem kimliği, derlem tipi kimliği, dil kimliği, çeviriyazı, sözcük listesi, sesbirimler listesi, ses bölümü (segment) kimliği, ses dosyası kimliği, dosya tipi kimliği. Bu alanlara ayrılan veri tabanına erişim ***hiper metin*** (hypertext) desteği de olan AskSam adlı ticari bir yazılımla sağlanmış, AskSam'in xBase veri tabanı destekli yapısıyla taban üzerinde tam metin sorgulama işlemi sağlanmıştır. AskSam'de ses segementi sınırları cümle/sözcük/sesbirim olarak başlıca üçe ayrılan veri setleri içinde aşağıda gösterilen yapıda saklanmıştır. Veri tabanının MS Access 2.0 sürümünde de modellemesi yapılmış, bu da aşağıdaki tabloda gösterilmiştir.

Sonuç olarak STDBMS sistemine DOS ve Windows desteğinin güçlü olması, oluşturulacak diğer ses verisi derlemlerinin ya da veri tabanlarının hazırlanmasında kolaylıklar sağlayacaktır. Ses işleme için çeşitli uygulama yazılımlarının geliştirilmesi, sesle birlikte videoların da eş zamanlı olarak anlamsal işlemeye de elverişli bir biçime getirilmesi araştırma konuları arasındadır. Büyük boyutlu metin verilerinin de otomatik olarak elde edilmesi ve sayısallaştırılması hazırlanacak veri tabanlarına ön işlem açısından büyük kolaylık sağlayacaktır (age., 1998: 91).

WORDNET adıyla bilinen sözcük veri tabanı yapısı, birçok DDİ çalışması için veri kaynağı olarak kullanılmaktadır. Ad ve sıfat türü sözcüklerin alt anlamlılık-üst anlamlılık, eş anlamlılık, parça-bütün hiyerarşisi içinde yapılandırıldığı bu sistemi, Chirist (1998) çevrim dışı olarak hazırladığı IMS derlem sorgu sistemi (IMS Corpus Query System) için bir çevrim içi yöntem geliştirerek kullanmıştır. Dinamik *öznitelikler* (attributes) adını verdiği sistemde cümle sınırları, söz türleri (part-of-speech) WORNET'e bağlanılarak belirlenmekte, örneğin kill "öldürmek" fiilinin adlarla kullanıldığı birlikteliklerinin sorgusu çağrıldığında WORDNET yapısında human (insan) olarak işaretlenen veriler yardımıyla "killed a client", "kill Mr.", killed the woman" gibi yapılar, çevrimdışı sistemin bağlamlı dizin penceresinde

görüntülenmektedir (Christ, 1998: 192). Aşağıdaki şekillerde bu sistemin sorgu yapısı ve sonuçlarından örnekler bulunmaktadır.

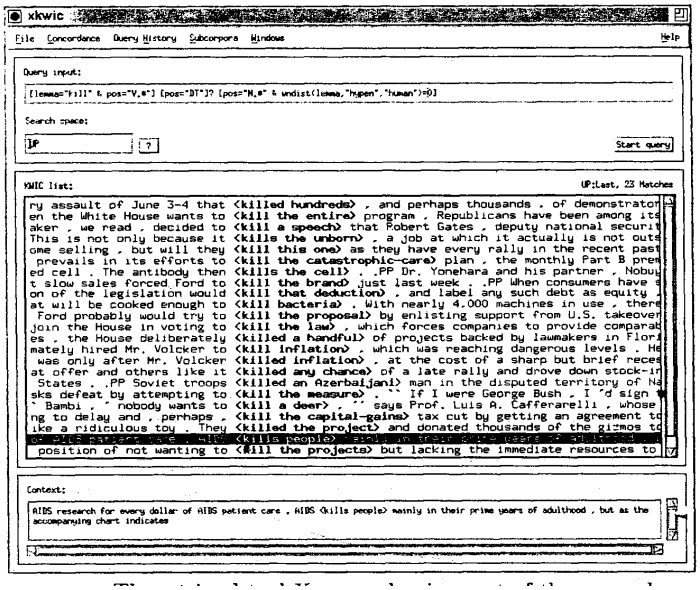
```
[lemma="kill" & pos="V.*"] [pos="DT"]?
  [pos="N.*" & wndist(lemma,"hypen","human")>0]
```

Şekil. IMS Sorgu Sistemi WORDNET Sorgulama Söz Dizimi (Christ, 1998: 193)

Yukarıdaki şekilde madde başı kill sözcüğünün öncelikle konumu (pozisyon) belirlenmekte, bu konuma göre fill+tanımlık+ad yapısı ve (&) mantık operatörüyle birleştirilmekte, wndist (WORDNET'teki sözlerin yerleşim hiyerarşisinin çağrılması) fonksiyonu ile üst anlamlılık (hypen) ilişkisi "human" olan ve sonucu 0'dan büyük bütün sonuçların listelenmesi sağlanmaktadır.

```
in one botched job <killed a client> . Her remorse was sh
lem group vowed to <kill Americans> if the U.S. implement
ld single-handedly <kill the BART> funds unless the congr
 sugar , which can <kill diabetics> . .PP The usual warni
 of threatening to <kill Mr.> Laff . .PP Mr. Laff 's atto
 way the U.S. will <kill a madman> is by making sure we t
 used primarily to <kill worms> in the intestinal tracts
o blame because he <killed the woman> without `` will or
```

Şekil 56. IMS Corpus Query System Sorgu Sonuçları (Christ, 1998: 193)



Şekil 57. IMS Corpus Query System Bağlamlı Dizini (Christ, 1998: 194).

Çok dilli bir dünyada, DDİ çalışmalarının, çeşitli dil çözümlemesi gerektiren otomatik işlemlerin de çok dilli olarak hazırlanmaları gerekir. Bilgisayar bilimciler dil çözümlemesi yapan yazılımlar için çok dilliğin neden olduğu karakter karmaşasına

çözüm getirmek için çalışmışlardır. Görsel ara birimlerin tasarlanması aşamasında karakter sorunlarının çözümü önemini artırmış, yazılım şirketleri çok dilliliği göz önüne alarak yazılımlarını uluslararası biçime, birçok dile çevrilerek hazırlamaya başlamışlardır. Gerçek dünya çok dillidir, yazılımların ve bilgisayarların da gerçek dünya sorunlarını çözmek için çok dilli modüllerle yapılandırılmaları gerekir. Bundan hareketle, The Summer Institute of Linguistics adlı kuruluş CELLAR (Computing Environment for Linguistic, Literary and Anthropological Researc) adındaki veri tabanını hazırlamıştır (Simons ve Thomson, 1998: 204).

CELLAR projesi, bilgisayarlı dil araştırmalarının sağlam ve olabildiğince az hatalı yapılabilmesi için her şeyden önce veri tabanı kavramından yararlanmanın gerekliliğinden hareketle veri tabanı çok dilli olacaktır. Her dilin kendi verileri kendi yapısı doğrultusunda yapılandırılmış, her dilin kendi verileri kendi yapısı doğrultusunda depolanmıştır. Hiyerarşik ve oldukça ilişkili bir düzende kurulan sistem nesne yönelimli model ile oluşturulmuştur. Simons ve Thomson (1998), gerçek dünyadaki nesnelerin kavramsal modellemeye el verişli yapısıyla nesne yönelimli modeli CELLAR'da kullanmayı tercih etmişler, böylece nesne yönelimli modelin saydam yapısı, her nesnenin kendi özniteliklerinin ve davranışlarının (metodlar) kurularak kolayca ulaşılabilecek yapıda kendi sınıfı içinde kullanmışlardır. Dünya nesnelerinin de bire bir ve saydam yapılı olduklarını belirten yazarlar, ilişkisel modelin bir varlık için bir çok ilişkili tablo gerektirmesinin ve kayıt kontrolünün emek yoğun olması nedeniyle dezavantajları olduğunu vurgulamaktadır (age., 1998: 208).

Princeton Üniversitesinde geliştirilen WORDNET İngilizce için hazırlanmış; adlar, sıfatlar, fiiller ve zarfların kavramsal hiyerarşisinden oluşan bir sözcük veri tabanıdır. Proje George A. Miller'ın başkanlığında 1993 yılında başlamıştır. Geliştirilmesi halen süren veri tabanının son olarak 3.0 sürümü yayınlanmıştır. Eşanlamlılık setlerine (synsets) göre hazırlanan veri tabanı için birçok araştırmacı zaman içinde katkıda bulunmuş, bilgisayar destekli dil bilimi ve DDİ araştırmalarının anlamsal inceleme gerektiren bölümleri için önemli bir kaynak durumuna gelmiştir. WORDNET'in çevrimiçi sürümü, genel sözcük sorgulaması yanında veri tabanında sözcüklerin kaydediliş bilgileri ayrıntılarına göre de sorgulamaya olanak vermektedir. Çevrimdışı sürümü de hem UNIX hem de Windows sürümü olarak genel ağdan indirilip

çalıştırılabilmektedir ticari kullanımlar için lisans sözleşmesi gerektiren yazılım akademik kullanımlar için ücretsizdir.[32]

WORDNET istatistiklerine bakıldığında, adların tekil olarak 117798, anlamlarıyla birlikte 146312 adet; fiillerin 11529 tekil, anlamlarıyla birlikte 25047 adet; sıfatların 21479 tekil, anlamlarıyla birlikte 30002 adet olmak üzere toplam 155287 tekil, anlamlarıyla birlikte 206941 adet söz varlığının bulunduğu görülmektedir. Veri tabanının çok anlamlılık istatistikleri de yayımlanmıştır, buna göre çok anlamlı sözcük türü sayısı şu şekildedir: Adlar: tek anlamlı 101863, çok anlamlı: 15935; fiiller: 6277 tek anlamlı, 5252 çok anlamlı; sıfatlar: 16503 tek anlamlı, 4976 çok anlamlı; zarflar: 3748 tek anlamlı, 733 çok anlamlı. Söz türlerine göre ortalama çok anlamlılık sayıları göz önüne alındığında fiillerin sözcük başına 2.17'lik istatistik ile diğer türlerden ayrıldığı görülmektedir.[33]

Projenin geliştirilmesi sürdüğünden, veri tabanı yazılımıyla ilgili kod ortamının çeşitli programlama dillerinde kütüphaneleri yayınlanmıştır. Aşağıdaki resimde görüleceği gibi sorgulanan "book" sözcüğünün ada ve fiil olarak kullanımına ait anlamları ve tanıkları listelenmiştir.



Şekil 58. WORDNET'te Sorgulama Sonuçları (http://wordnetweb.princeton.edu/)

Geniş bir kullanım kılavuzu bulunan WORDNET'in kılavuz sayfalarında veri tabanında kullanılan dosya formatları, giriş bilgileri, komut dosyası bilgileri yanında

[32] http://wordnet.princeton.edu/
[33] http://wordnet.princeton.edu/man/wnstats.7WN

özellikle sözlükbilimciler için hazırlanan sözcüklerin anlam ilişkileriyle ilgili bilgiler de bulunmaktadır. Buna göre, örneğin sıfatların ilişkisellikleri (pertainyms) ile zarflar ve adlar ayrı dosyalarda saklanmaktadır[34].

[34] http://wordnet.princeton.edu/man/lexnames.5WN

İKİNCİ BÖLÜM

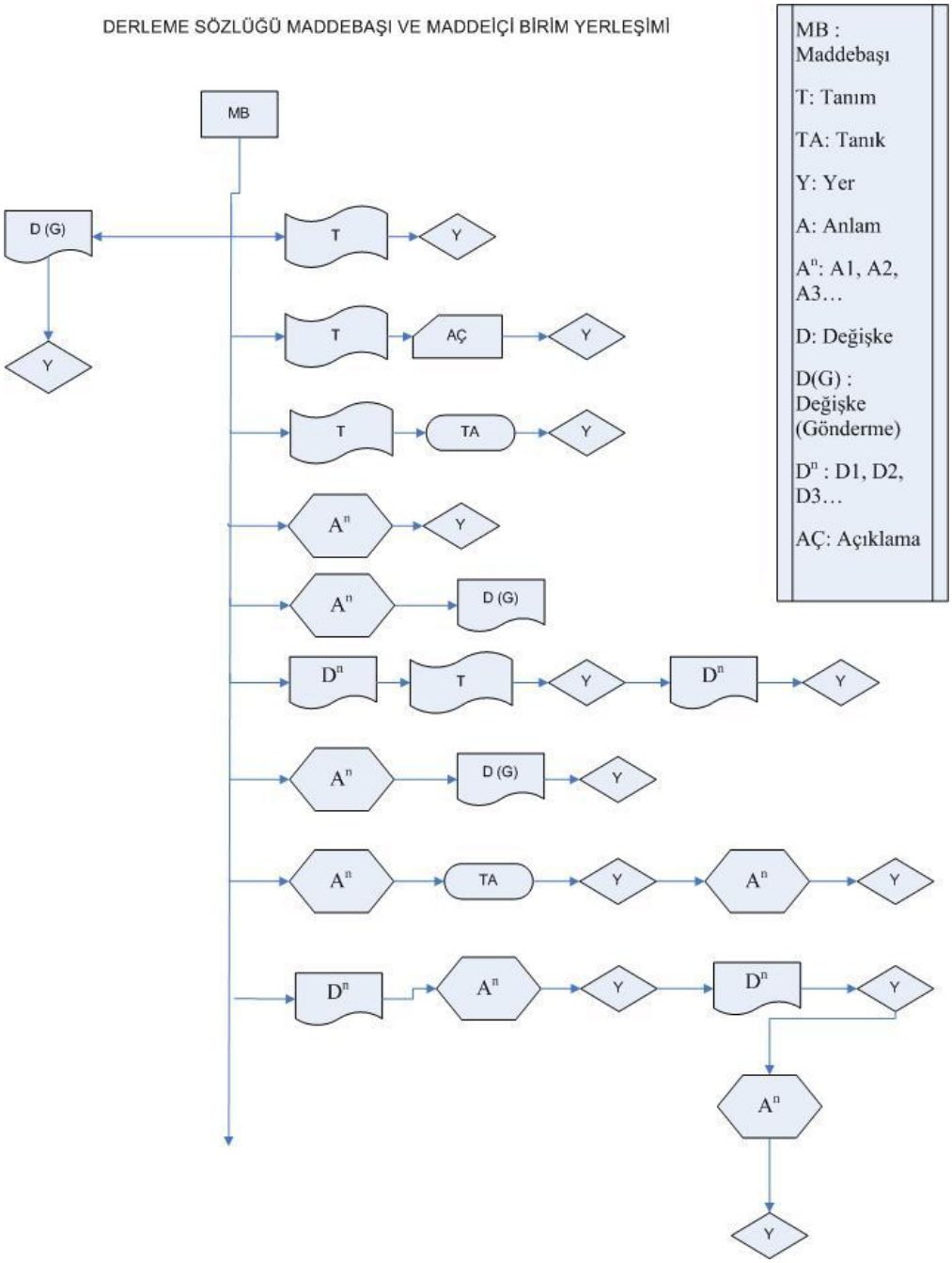
DERLEME SÖZLÜĞÜ VERİ TABANI MODELİ

2.1. Hazırlama Yöntemi ve Kullanılan Teknoloji

Derleme Sözlüğü sonradan hazırlanan kavramlar dizini dışında toplam 12 ciltten oluşmaktadır. Sözlüğün tamamı OKT kullanılarak word belgesi biçimine dönüştürülmüştür. 28 MB büyüklüğündeki belge MySQL açık kaynak kodlu veri tabanı yönetim yazılımı kullanılarak ilişkisel veri tabanı biçimini almıştır. MySQL standart SQL (Structured Query Language) dilini destekleyen, bugün milyonlarca ağ kümesinin (web sitesi) veri tabanının tutulduğu geniş kullanımlı ücretsiz bir yazılımdır. OKT'nin yol açtığı karakter sapmaları ilk aşamada göz önüne alınmamış, sözlükte yer alan il kısaltmaları sorgulama sırasında kısaltma biçimlerinden açık biçimlerine dönüştürülmüştür.

Veri tabanındaki bilgiye ulaşım ve sorgulama ara yüzü PHP (PHP: PHP: Hypertext Preprocessor) sunucu taraflı betik (script) dili ile hazırlanmıştır. PHP de MySQL gibi açık kaynak kodlu ve ücretsizdir. PHP'yi 1993 yılında Rasmus Lerdof geliştirmiş, günümüzde PHP ve MySQL hem ağ kümesi sunumunda hem de metin işlemede Perl dili gibi sunduğu geniş kütüphanesiyle çokça tercih edilen ikili yazılımlar durumuna gelmiştir.

Sözler veri tabanına kaydedilmeden önce sözlüğün yapısal özelliği aşağıdaki şekilde görüldüğü gibi çıkarılmıştır.



Şekil 59. Derleme Sözlüğü Birim Yerleşimi

Yukarıdaki şekilde birimler değişik geometrik çizimle temsil edilmiştir. Bu çizimdeki yapıların çalışmamızda seçilen sorgulama biçiminde düzenli ifadeler (regular expression) ile kodlanması gerekmektedir. Aşağıda bu kodlamadan örnekler verilmiştir.

MB \T \Y

uydül Ödül. (-Bo.)

MB \D(G)

uyga [-> uyaroğlu]

MB \T \AÇ \Y

uykucu Uyumayı seven, çok uyuyan (kimse). (*Bor -Nğ.)

MB \T \TA \Y

uyku durak Dinlenme olanağı: Hiç uyku durak yok! (*Bor -Nğ.)

MB \A^n \Y

uykuluk (II) 1. Yeni doğan çocukların başındaki kepek. (Peşman *Daday, *Taşköprü -Ks.; -Sn.) 2. Bebeklerin parmak aralarında biriken kir. (-Sm.; *Bor -Nğ.) 3. Göğüsle karın boşluklarını ayıran zar, perde. (*Güdül -Ank.)

MB \A^n \D(G)

vağıldamak 1. Coşkun su çağıldamak. (*Gelendost -Isp.; *Afşin -Mr.; -Yz.) **2.** [-> vadırdamak]

MB \$D^n$ \T \Y \$D^n$ \Y

utancak [udlu, utangan, utangan, utangeç, utansak, utlu, uyalçak, uyalğan, uyalûvçu] Utangaç. (-Es. köyleri; *Kurşunlu -Çkr.; *İskilip -Çr.;-Sm.; *Taşova köyleri, *Merzifon -Ama.; Bayadı -Or.; Yavuz, Salcı *Şavşat, *Ardanuç köyleri -Ar.; *Erciş, - Vn.; -Dy.; -Sv. ilçe ve köyleri; Çanıllı *Ayaş -Ank.; Bahçeli, *Bor, -Nğ.; *Ermenek - Kn.; Güzelsu *Akseki-Ant.; *Mut köyleri -İç.; -E d.; *Lüleburgaz köyleri, -Krk.) [udlu]: (*Eğridir köyleri -Isp.) [utangan]: (*Arapkir -Ml.) [utangan]: (*Düzce -Bo.) [utangeç]: (Bereketli *Tavas -Dz.) [utansak]: (*Eğridir köyleri -Isp.; -Dz.; *Alaşehir - Mn.; Yeniköy -Ba.; *Düzce -Bo.; Nefsiköseli *Görele -Gr.; -İç. köyleri; Pmar-lıbelen *Bodrum, *Milâs -Mğ.) [utlu]: (Cebel -Isp.; Tokat -Es.; -Ks.; *Kurşunlu -Çkr.; -Sv.; *Talaş, Afşarlar *Pınarbaşı -Ky.; *Bor -Nğ.; *Beyşehir, *Bayındır -Kn.) [uyalçak]: (Karaçay Aşireti, Baş-höyük *Kadınhanı -Kn.) [uyalğan]: (Karaçay Aşireti, Baş-höyük *Kadınhanı -Kn.) [uyalüvçu]: (Karaçay Aşireti, Baş-höyük *Kadınhanı -Kn.)

MB \A^n \D(G) \Y

uvmaç 1. [-> umaç -2] 2. Hamur çorbası. (-Zn.)

MB \ A^n \TA \Y \ A^n \Y

yatım (I) 1. Durum, tutum: Elin oğlu onun yatımından anlar. (*Yalvaç -Isp.; -Dz.; Tokat -Es.; -Tr.) 2. Yol, yöntem. (*Bozdoğan -Ay.)

MB \$D^n$ \ A^n \Y \$D^n$ \Y \$A^n$ \Y

üçayak [üçayaklı] 1. Sacayağı. (Senir *Keçiborlu -Isp.; Honaz -Dz.; -Tr. dağ köyleri; - Ar.; -Bt.; -Dy.; *Varto -Mş.; *Nizip -Gaz.; -Mr.; Amik Ovası

Türkmenleri *Reyhanlı - Hat.; -Nğ.) [üçayaklı]: (Bahçeli *Bor -Nğ.) 2. Sehpa. (*Akyazı çevresi -Sk.; *Nizip - Gaz.; *Mut köyleri -İç.; -Tk.) 3. Üç ayaklı merdiven. (*Milâs -Mğ.) 4. Bir çeşit horon. (-Rz.; Hacıilyas *Koyulhisar -Sv.) 5. Bir yerel oyun. (-Ur.)

Düzenli İfadeler ile Gösterim:

Maddebaşı:

1) [a-z\x00E7|011F|0131|015F|00FC|00E2|00EE|00FB|014B|00F1]+ (boş karakter)

.+ (boşluk yok)

azsınmag

2) [a-z\x00E7|011F|0131|015F|00FC|00E2|00EE|00FB|014B|00F1]+ (boş karakter)

[a-z\x00E7|011F|0131|015F|00FC|00E2|00EE|00FB|014B|00F1]+

.+\s.+

azne etmek

3) [a-z\x00E7|011F|0131|015F|00FC|00E2|00EE|00FB|014B|00F1]+ (boş karakter)

[a-z\x00E7|011F|0131|015F|00FC|00E2|00EE|00FB|014B|00F1]+ (boş karakter)

[a-z\x00E7|011F|0131|015F|00FC|00E2|00EE|00FB|014B|00F1]+

.+\s.+\s.+

ayoç boşoç yatmak

4) [a-z\x00E7|011F|0131|015F|00FC|00E2|00EE|00FB|014B|00F1]+ (boş karakter)

[a-z\x00E7|011F|0131|015F|00FC|00E2|00EE|00FB|014B|00F1]+ (boş karakter)

[a-z\x00E7|011F|0131|015F|00FC|00E2|00EE|00FB|014B|00F1]+ (boş karakter)

[a-z\x00E7|011F|0131|015F|00FC|00E2|00EE|00FB|014B|00F1]+

.+\s.+\s.+\s.+

ayın oyun iş değil

5) [a-z\x00E7|011F|0131|015F|00FC|00E2|00EE|00FB|014B|00F1]+, (boş karakter)

[a-z\x00E7|011F|0131|015F|00FC|00E2|00EE|00FB|014B|00F1]+

.+\,\s.+

aylakcı, aylakçı

6) [a-z\x00E7|011F|0131|015F|00FC|00E2|00EE|00FB|014B|00F1]+ (boş karakter)

[a-z\x00E7|011F|0131|015F|00FC|00E2|00EE|00FB|014B|00F1]+, (boş karakter)

[a-z\x00E7|011F|0131|015F|00FC|00E2|00EE|00FB|014B|00F1]+ (boş karakter)

[a-z\x00E7|011F|0131|015F|00FC|00E2|00EE|00FB|014B|00F1]+

.+\s.+\,\s.+\s.+

aynak gazzak, aynak kazak

Madde içi

1) \T \Y

uydül Ödül. (-Bo.)

.+\s\(\-.+\.\)

2) **uyga** [-> uyaroğlu]

\[.+\s.+\]

3) **uykucu** Uyumayı seven, çok uyuyan (kimse). (*Bor -Nğ.)

.+\.\s\(*.+\s\-.+\.\)

4) **uyku durak** Dinlenme olanağı: Hiç uyku durak yok! (*Bor -Nğ.)

.+\:\s.+\s\(*.+\s\-.+\.\)

2.1.2. Veri Tabanı Yöntemi

Sözlüğün veri tabanı *ilişkisel veri tabanı* (relational database) modeliyle söz ve madde tablolarından oluşan iki *varlık* (entity) alanı düşünülerek oluşturulmuştur. Veri tabanlarında tablo alanlarının atomik yapılandırılmasının başarım düzeyini artırdığı bilinmekle birlikte, DS veri tabanında madde başı sözleri içeren "söz tablosu", bütün anlamları yani maddebaşlarının tanımını içeren "anlam tablosu", anlamlarda örnek cümlelerin verildiği "tanık tablosu", madde başı sözlerin varyasyonlarının gösterildiği "değişke tablosu" ve derleme yeri bilgilerinin tutulduğu "yer tablosu" olmak üzere 5 tablo yer almaktadır. Aşağıdaki şekilde bu 5 tablonun ilişkili biçimleri görülmektedir.



Şekil 60. Derleme Sözlüğü Veri Tabanı İlişkiler Tablosu

Söz tablosunun birincil anahtarı diğer 4 tabloya yabancı anahtar olarak atanmış ve tablolar arasındaki bağlantılar bu yabancı anahtar ile sağlanmıştır.

Veri tabanının hazırlanmasından önceki aşamada, OKT (OCR) işlemi yapılan basılı metin Word belgesi biçiminde kaydedilmiştir. Basılı metindeki madde başı sınırlarının OKT işleminden sonra kaybolması ve her birimin iç içe girmiş olması,

maddebaşı ve maddeiçi sınırlarının ön işlem olarak iki bölüme (madde başı ve madde içi) ayrılmasını gerektirmiştir. İki bölüme ayrılan sözlük verisi, MySql veritabanı yönetim sistemi için yazılan PhpMyAdmin sistemi kullanılarak başta soz ve madde tablosu olarak birim sıra sayısı alanında ilişkilendirilerek tutulmuştur. Maddeiçi birimlerin yapısının önceki bölümlerde açıklandığı gibi düzenli ifadeler ile örüntüleri tanımlanmış, ve PHP betik dilindeki preg_match_all fonksiyonu ile tanımlanan bu birimlerden yer adları kabaca elde edilmiştir. Aşağıdaki ifadede sözlükte ( ) içine alınan yer adlarının elde edilmesi amaçlanmıştır.

(preg_match_all ("/\(.+\)/",$sira["madde"], $bulunan)

Bu fonksiyonun döndürdüğü sonuç şu biçimde olmuştur:

---->(Çayağzı *Şavşat -Ar.; Yeniköy *Düzce -Bo.; Soğukpınar *Kangal -Sv.)
---->(*Alaşehir -Mn.; -Ba.; -Gaz.; Kayalıpmar *Yıldızeli -Sv.)2. Usta. (Karahisar *Tavas -Dz.)3. Efendi. (Karahisar Tavas -Dz.)4. Büyük erkek kardeş, ağabey. (*Akçaşehir -Bo.; Mahzemin -Ky.)5. Baba: A bana kitap al! (Çalılı -Yz.; Yarasa *Haymana -Ank.)
--->(I)] Hayır, yok, olmaz, istemiyorum anlamına gelir: Karnın aç mı? Aa! (Karatekeli aşireti -Üş.; -Brd.; Bekilli *Çal, *Sarayköy köyleri, İğdir *Çivril, Beylerli-Dz.; Tepeköy *Torbalı -İz.; -Mn.; *Edremit, Demirkapı, *Susurluk -Ba.; Evreşe *Gelibolu, Küçükkuyu, *Ayvacık -Çkl.; Eğirler *Simav, *Emet -Kü.; -Es.; Baklalı, *Çatalca -İst.; -Gr.; Kalafka -Tr.; -Gm.; Hamidiye *Pazar -Rz.; -Nğ.; Yelten *Korkuteli -Ant.; Orhaniye *Marmaris, *Bodrum, Yerkesik -Mğ.; *Meriç, Kayapa, -Ed.; *Lüleburgaz ve köyleri -Krk.; Büyükmanika *Saray -Tk.) [aah (I)]: (*Eşme köyleri, -Üş.; Ekse *Çal, *Buldan köyleri, *Sarayköy köyleri -Dz.; Hamzabali, Eymir *Bozdoğan, Beyköy -Ay.; *Bergama -İz.; *Akhisar köyleri -

Bu sonuçlarda da görüldüğü gibi, OKT hatalarından kaynaklanan karakter sapmaları nedeniyle her zaman beklenmesi gereken örüntüyü fonksiyon bulamayıp, istenen örüntünün dışında da gürültü olarak adlandırabilinecek fazla veri de elde edilebilmektedir. Bu örnekte yer adları dışında [ ] işareti içinde bulunan değişke ve rakamla başlayan anlamlar da bulunmaktadır.

Örüntü dışındaki verilerin çok fazla belirsizliğe yol açması, soz ve madde tablolarından elde edilen tüm veriyi ortaya koyan aşağıdaki fonksiyon yardımıyla bütün veriler elde edilmiştir.

(preg_match_all ("/.+/",$sira["madde"], $bulunan))

Elde edilen sonuç şu biçimdedir:

11--->**aba (III)**--->[-> abo, abö (I) -1]
12--->**aba (IV)**--->[-> abe (I)]
13--->**aba (V)**--->Yaba. (*Hozat -Tn.)
14--->**aba atmak**--->1. Kendisini kurtarmak için suçu başkasına yüklemek. (*Ünye -Or.) 2. Birine yük olmak: Bizim Hasan dayı aba atacak bir yer arıyor. (Nudra *Şarki Karaağaç -Isp.)
15--->**ababa (I)**--->1. Baş ve kuyruk tarafları kara, gövdesi yeşil bir kuş, incir kuşu. (Buluklu *Mersin -İç.) 2. Akbaba. (-Çr.)
16--->**ababa (II)** --->Ağababa, büyükbaba, dede. (*Akyazı çevresi -Sk.)
17--->**âbabaççа**--->Papatya. (Çalış *Haymana -Ank.)
18--->**ababak**--->[-> abulabut -1]

Bu ikinci yöntemle bütün veri maddebaşları ve sıra numaraları koyu karakterde olmak üzere elde edildikten sonra, RegexBuddy adlı yazılımın tam işlevli tanıtım sürümüyle bölümlenmeye başlanmıştır.

İç içe girmiş sözlük verisindeki örüntüleri tanımlamak için şu düzenli ifadeler tanımlanmıştır:

Tek anlam:

```
>.+?--|>[^\d].+?\.\s\(|[^\d].+?\:
```

Çok Anlam:

```
>.+?---
|>(\u015E).+?[.:]|>(\u00C7).+?[.:]|>(\u0049).+?[.:]|>(\u00D6).+?[.:]|>(\u00DC).
+?[.:]|>(\u0130).+?[.:]|>(\u0134).+?[.:]|>[A-Z].+?[.:]
```

Değişke:

```
>.+?--|\d\..+?[.:]
```

Tanık:

```
>.+?---|\[.+?\]
```

Yerler

Köy:

```
>.+?--|\:\s.+?[.!]\s|\:\s.+?[?]\s\(
```

İlçe:

```
.+?--->|\s.+? köyleri|\s.+? ve köyleri|\(.+?\s\*|\(.+?\s\-|\;\s.+?\*
\s\*.+?\s|\-.{2,3}\.
```

İl:

```
.+?---\(|.+?---|\*.+?\s|ve köyleri|\*.+?\,.+?\-
```

```
.+?--->|\[.+?\]|-.{2,3}\.|\s.{2,3}\.
```

Yurtdışı:

\d+--->.+?---

>|[Kk]ıbrıs|[Ll]imasol|[Kk]erkük|[Mm]agosa|[Ll]arnaka|[Ll]efkoş[ae]|[Kk]ırcaal i|[Ss]elanik|[Yy]ugoslavya|[Mm]anastır

Yukarıdaki örüntü gösterimleri RegexBuddy adlı yazılımın tam işlevli deneme sürümü kullanılarak istenen veriler elde edildikten sonra aşağıdaki ekran görüntüsündeki gibi tablo ve alanlara ayrılmıştır. Tablolama sırasında sözlüğün basılı biçiminde yer alan il adı kısaltmaları normal uzun biçimlerine dönüştürülmüştür.



Şekil 61. Derleme Sözlüğü Veri Tabanı Tablolarının VTYS'de görünümü.

Yukarıdaki şekilde MySQL veri tabanı yönetim sistemi olarak kullanılan phpMyAdmin adlı yazılımın toplam 19 adet tabloyu barındırdığı görülmektedir. Ana sorgulama sayfasında bu tablolardan 5'i kullanılırken anlamlar ve bu anlamların il, ilçe ve köylere göre groplamasını içeren tablolar da ana sorgu sayfasına bağlantılı olarak görüntülenebilmektedir.

Aşağıdaki tablo anlam tablosundan ilk 30 satırın verisini göstermektedir. Tabloda yer alan ve köşeli parantezle verilen değişkeler, madde başlarından anlamına bakılması için diğer bir madde başına olan göndermelerdir.

Tablo 4. 30 Satırlık Anlam Tablosu İçeriği

id	anlam	soz_id
1	İşte:	1
2	1. Ağa. 2. Usta. 3. Efendi. 4. Büyük erkek kardeş,...	2
3	Hayır, yok, olmaz, istemiyorum anlamına gelir: Kar...	3
4	Burada:	4
5	Çocuk dilinde büyük abdest.	5
6	[-> aa I)]	6
7	1. Pişmanlık ünlemi. 2. Acıma ünlemi. 3. İstek, di...	7
8	Ağırşak.	8
9	1. Abla, büyük kız kardeş. 2. Anne. 3. Üvey anne, ...	9
10	1. Çoban, deveci ve göçebelerin giydiği uzun, yaka...	10
11	[-> abo, abö I) -1]	11
12	[-> abe I)]	12
13	Yaba.	13
14	1. Kendisini kurtarmak için suçu başkasına yükleme...	14
15	1. Baş ve kuyruk tarafları kara, gövdesi yeşil bir...	15
16	Ağababa, büyükbaba, dede.	16
17	Papatya.	17
18	[-> abulabut -1]	18
19	Kebab kestane.	19
20	[-> abbak I) -1]	20
21	1. Kırlangıç. 2. Kırlangıç yavrusu.3. Tip ve büyük...	21
22	Umut vererek oyalamak:	22
23	[-> abâbül -3]	23
24	1. [-> ababara]2. Kestane. 3. Kavun.	24
25	[-> ababil -3]	25
26	[-> ababil -1]	26

id	anlam	soz_id
27	Çocuk dilinde bir korku ünlemi:	27
28	1. Atıcı, yalan yanlış söyliyen: Onun sözüne inanı...	28
29	[-> abacık]	29
30	Uzak yakın akraba, tanıdık, ahbap:	30

ÜÇÜNCÜ BÖLÜM

DERLEME SÖZLÜĞÜ VERİ TABANI SORGULAMA YÖNTEMİ

Derleme Sözlüğü'nün (DS) veritabanı oluşturulduktan sonra arayüzü ve sorgulama siteminde PHPRunner 5.2 adlı yazılımın tam işlevli deneme sürümü kullanılmıştır.

Ana sorgu tablosu dışında, anlam tablosundan elde edilen anlam söz listesi, anlamlarla birlikte il, ilçe ve köylerin gruplandırıldığı anlam-il, anlam-ilçe ve anlam-köy gruplama tabloları, anlam tablosundaki 2'li ve 3'lü eşdizimlilik tabloları ve derleme yeri sıklık tablosu olmak üzere 9 sorgulanabilir tablo yapısının arayüzü hazırlanmıştır.

3.1. Madde Başında Sorgulama

DS madde başı sorgulaması, veri tabanında "soz" olarak tutulan tablodaki sözlerin sorgusunu ilişkili oldukları diğer tablolarla beraber sonuç sayfası olarak üretmektedir. Aşağıdaki şekilde, DS veri tabanı sorgulamasının gerçekleştirilebildiği ana sayfası görülmektedir.



Şekil 62. Derleme Sözlüğü Sorgulama Ana sayfası.

Veri tabanında "no", "maddebaşı", "değişke", "anlam", "tanık" ve "derleme yeri" olmak üzere altı alan bulunmakta, her bir alan madde başında yapılan sorgulamaya göre ilişkili olduğu madde başıyla birlikte satır ve sütünlar biçiminde görüntülenmektedir. Her bir madde başının numarası da yanında gösterilerek toplam 124.692 adet sözün (madde başı) verisine ulaşılabilmektedir.

Ana sorgulama sayfasının sol üst köşesinde yer alan arama kutucuğu herhangi bir alanda sorgulama yapılmasını olanaklı kılmaktadır. Sayfanın solunda yer alan seçenekli arama bölümünde istenirse iki ya da daha fazla alanda seçenekler eklenerek sorgulama yapılabilmekte, eklenebilen seçenekler arasında içerir, eşit, ile başlayan,ile başlamayan, boş, boş değil, içermez, eşit/eş olmayan gibi metne duyarlı sorgular oluşturulabilmektedir. Hem madde başlarında hem de diğer tabloların içeriğinde sayısal değerlerin az olması; daha büyük, daha küçük, arasında, küçük olmayan, büyük olmayan gibi sayısal değerlere duyarlı alanlardaki sorgulamayı kısıtlamaktadır.



Şekil 63. Madde Başı “eşit” Sorgusu.

Yukarıdaki ekran görüntüsünde “abablandırmak” madde başı “eşit”tir sorgusuyla elde edilmiştir. Eşit sorgusunda aranan madde başının kesinlikle aynı söz olması gerekmektedir, yani “abablandırmak” dışında başka bir söz sorgu sonucu olarak getirilmez. Bunun tersine “içerir” sorgusu örneğin “aba” sözü sorgulandığında aşağıdaki sorgu sonuçlarını vermektedir:

aba (I), aba (II), aba (III), aba (IV), aba (V), aba atmak, ababa (I), ababa (II), âbabaççа, ababak, ababara…

Görüldüğü gibi, “aba” sözcünün içinde geçtiği bütün madde başları numara sırasına göre sıralanmaktadır. DS’de “aba” sözünü içeren madde başı sayısı 1204 adettir.

Ana sorgulama sayfasının sağ üst köşesinde yer alan “Gelişmiş Arama” özelliği aslında ana sayfada yer alan seçenekli sorgulamayla bağlantılıdır.



Şekil 64. Gelişmiş Arama Sorgu Arayüzü.

Gelişmiş arama sayfasında iki temel radyo düğmesi bulunmaktadır: Tüm Şartlar ve Herhangi Biri. Kullanıcı, altı sorgu kutusuna tüm şartlar etkin durumda iken örneğin No'sunda "2" içeren "aba" sözünü aramak istediğinde bu iki şartın sorgu sonuçlarında kesinlikle bulunması gerektiğini belirtmiş olmaktadır. Gelişmiş arama sorgusunda "DEĞİL"leme yapılabilecek seçenekli kutular da bulunmakta, böylece "tüm şartlar" ya da "herhangi biri" şartında girilen sorgunun "olmadığı" durumları sorgulayabilmektedir. Numarasında 2, madde başında "aba" sözcüğünü içeren toplam 553 adet madde başı bulunmaktadır. Madde başında "aba" sözüyle başlayan sözleri bulmak için gelişmiş aramada, madde başı alanının açılır kutu seçeneğinde "ile başlayan" sorgusu seçildiğinde toplam 176 sözün sonuçları şu biçimde başlayan sözleri getirmektedir:

abacıh, abacı kebeci, abacık, abacur, abaç, abacık, abadan (I), abadan (II), abadan (III), abadan (IV), aba gibi atmak, abağına çalışmak, abah, abalıam, abalımı, abahandan ,abahon, abajor…

Madde başlarında dil bilgisel biçimlerin sorgulamasını yapmak da DS'nin şu anda hazırlanan veri tabanı biçimiyle kısıtlı da olsa yapılabilmektedir. Örneğin "-mak" isim fiil ekiyle yapılmış sözleri bulmak için madde başı alanının "içerir" sorgusuyla yapılan "mak" sorgusu 13938 adet madde başının "mak" isim fiil ekini çierdiğini göstermektedir. Bunlardan bazıları şu biçimdedir:

aba atmak, abablandırmak, aba gibi atmak, abağına çalışmak, abak atmak, abalamak (I), abalamak (II), abalamak (III), abanmak, aballamak, abamak (I), abamak (II), abarot satmak, abartıyı basmak, abartmak (I), abatmak, abaz abaz olmak, abazımak, abcalamak, …. yıramak, yırlamak, yolaklamak, yonmak (II), yormak, yorulup arılmak, zıypıtmak, zifiri azmak, ziğ olmak…

İçerisinde hem “-mak” hem de “-mek” isim fiil eki geçen madde başları da sorgu ana sayfasında yer alan sorgulama kutusunun “alan ekle” özelliğiyle bulunabilmektedir. Alan ekle seçeneği ile madde başı kutusuna istenildiği kadar “içerir” sorgusu yüklenerek sonuçlar elde edilebilir. “mak” ve “mek” sorgusunun sonucu olarak 18 adet madde başının olduğu görülmektedir:

abıçmak, ebiçmek; ağzını silmek, ağzını yıkamak; annaçlamak, anneçlemek; basıla gelmek, basıla kalmak; başına çevirmek, başına sallamak; başına ekmek çağırmak, bılk bılk etmek, oynamak; bıllık bıllık etmek, oynamak; bıllım bıllım etmek, oynamak; boğum bağlamak, boğum vermek; cırmak çekişmek, çitmek yapmak, damak değiştirmek, ekmek karıştırmak, tuğmek olmak, tuz ekmek doymamak, ayağını mercimek kütüğüne dayamak, ekmek yamamak.



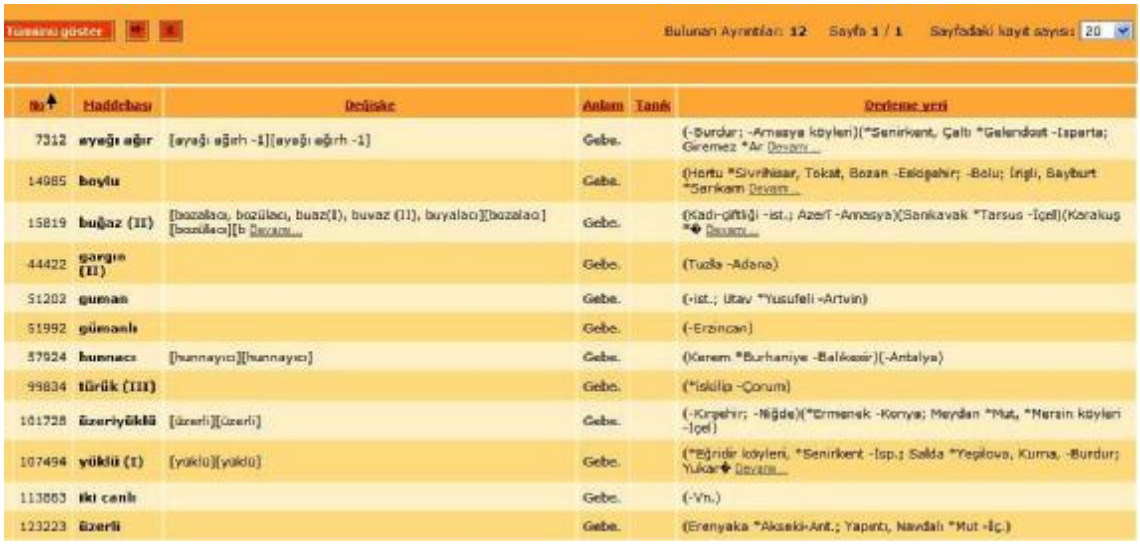
Şekil 65. Madde Başı Sorgusu Gelişmiş Özellikler.

Madde başı sorgulama kullanıcının değişik sorgu türlerine yanıt verecek düzeyde olup, gerek dil bilgisel gerek genel sorgulamalar için kullanılabilmektedir. Burada sorgu yapan kullanıcının ne sorgulamak istediği ve sorguyu genişletebilme düşüncesi önemli yer tutmaktadır. DS’nin şu andaki sorgulama dili klasik SQL veri tabanı sorgulama diliyle yapıldığından daha karmaşık duyarlı sorgular da temel sorgulama tiplerine eklenip genişletilebilir.

3.2. Anlam İçinde Sorgulama

Bu sorgu biçimi, “anlam” tablosunun içinde yer alan ve DS’de “degiske” alanı olarak belirlenmiş alanla “tanik” alanından önce gelen bölümü içermektedir. Veri tabanı oluşturma süreci içinde OKT’si yapılmış metnin kimi OKT hatalarının yol açtığı karakter değişiklikleri de şu anda bu bölümde izlenebilmekle birlikte, hataların büyük çoğunluğu giderilmiştir. Bu alanın sorgu sonuçları da madde başı sorguslarında olduğu gibi ilişkili olduğu diğer tablolardan sorguyla bağlantılı alanları getirmektedir. Gelişmiş arama ve sorgu ana sayfasında sözlükte anlamla ilgili karakter düzeyinden cümle düzeyine kadar sorgular yapılabilmektedir. Örneğin “Çocuk dilinde bir korku ünlemi” gibi bir sözlük cümlesi sorgusu ilişkili olduğu bütün alanları getirebilmektedir.

Anlam sorgusunun kullanıcı sorgularının en serbest olduğu alan olması olasıdır, çünkü akla gelebilecek her terim ya da kavram için sözlükte bir karşılığın olma olasılığı yüksektir. Anlamında tam olarak “gebe” sözü geçen bir sorgulama yapıldığında, sözlükte anlam alanında sadece gebe olarak girilmiş, “engebe” gibi bir sözün içinde geçmeyen sonuçlar gelmektedir. Bununla birlikte anlam alanındaki metinde “gebe” sözü için içerir sorgusu daha geniş ve anlamlı sonuçlar getirerek, metnin herhangi bir yerinde geçen “gebe” sözü gebelikle ilgili madde başlarına erişimi sağlamaktadır.


Tümünü göster

Bulunan Ayrıntılar: 12 Sayfa 1 / 1 Sayfadaki kayıt sayısı: 20

No	Maddebaşı	Değişke	Anlam	Tanık	Değişke yeni
7312	ayağı ağır	[ayağı ağırh -1][ayağı ağırh -1]	Gebe.		(-Burdur; -Arnavya köyleri)(*Senirkent, Çallı *Gelendost -Isparta; Giremez *Ar Devamı...
14985	baylu		Gebe.		(Hartu *Sivrihisar, Tokat, Bozan -Eskişehir; -Bolu; İnigl, Bayburt *Serikam Devamı...
15819	buğaz (II)	[bozalacı, bozülacı, buaz(I), buvaz (II), buyalacı][bozalacı][bozülacı][b Devamı...	Gebe.		(Kadı-gittiği -ist.; Azer'i -Amasya)(Sankavak *Tarsus -İçel)(Karakuş *Devamı...
44422	gargın (II)		Gebe.		(Tuzla -Adana)
51202	guman		Gebe.		(-ist.; Utav *Yusufeli -Artvin)
51992	gümanlı		Gebe.		(-Erzincan)
57924	hunnacı	[hunnayıcı][hunnayıcı]	Gebe.		(Karem *Burhaniye -Balıkesir)(-Antalya)
99834	türük (III)		Gebe.		(*Iskilip -Çorum)
101728	üzeriyüklü	[üzerli][üzerli]	Gebe.		(-Kırşehir; -Niğde)(*Ermenek -Konya; Meydan *Mut, *Mersin köyleri -İçel)
107494	yüklü (I)	[yuklu][yuklu]	Gebe.		(*Eğridir köyleri, *Senirkent -Isp.; Salda *Yeşilova, Kuma, -Burdur; Yukar Devamı...
113863	iki canlı		Gebe.		(-Vn.)
123223	üzerli		Gebe.		(Eranyaka *Akseki-Ant.; Yapıntı, Navdalı *Mut -İç.)

Şekil 66. Anlam Sorgusu Sonuç Sayfası.

DS’de Anlam sorgulaması yapılırken, sözlüğün madde başalrına karşılık gelen anlamlarının belirli bir örüntü içinde verildiği dikkati çekmektedir. Bütün anlamlar büyük harflerle başlarken, çok anlamlı madde başlarının anlamları 1., 2., gibi noktalı

rakamlar kullanılarak verilmiştir. Her bir anlamdan sonra, anlam "." ile bitirilirken, tanıklardan önce ":" noktalama işareti kullanılmıştır. Ayrıca anlama açıklama eklerken açıklamalar (kimse, bitki, argo vb.) parantezler içinde verilmiştir. Anlama yönelik bu tür işaretlemeler, anlam sorgulamalarında ipucu niteliğinde ayrıntıya inen sonuçlar için sorgu oluşturmaya yardımcı olabilmektedir.

Yukarıda açıklana ipuçlarına dayanarak DS'de içerir sorgusu için "1." kullanıldığında, yani çok anlamlı sözler numaralandırılarak verildiğinden her bir çok anlamlı söz 1 numaradan başladığından, sorgu sonucu olarak toplam 9620 adet sözün çok anlamlı olduğu görülmektedir. Sözlükte anlam olarak tanımlanan madde başları için kimi zaman göndermeler kullanılmıştır. Örneğin **abahon** *1. [-> abulabut -1] 2. Düşüncesiz. 3. Tembel, kaba adam.* abahon sözü için 1. anlam abulabut sözüne göndermedir. Sorgu sırasında çok anlamlı sözlerin tanımları içerip, göndermeleri içermemesi sağlanmıştır. Aşağıda çok anlamlı sözlerin sorgu sonucundan bir bölüm tablo olarak verilmiştir.

Tablo 5. Çok Anlamlı Sözler Sonuç Tablosu Örneği.

No	Maddebaşı	Anlam
2	â	1. Ağa. 2. Usta. 3. Efendi. 4. Büyük erkek kardeş, ağabey.
7	ahh (II)	1. Pişmanlık ünlemi. 2. Acıma ünlemi. 3. İstek, dilek ünlemi. 4. Kızgınlık ünlemi.
10	aba (II)	1. Çoban, deveci ve göçebelerin giydiği uzun, yakasız üstlük, kepenek. 2. Beyaz yünden dokunan bir çeşit kilim. 3. Yün terlik.
14	aba atmak	1. Kendisini kurtarmak için suçu başkasına yüklemek. 2. Birine yük olmak: Bizim Hasan dayı aba atacak bir yer arıyor.
15	ababa (I)	1. Baş ve kuyruk tarafları kara, gövdesi yeşil bir kuş, incir kuşu. 2. Akbaba.
28	abacı	1. Atıcı, yalan yanlış söyliyen: Onun sözüne inanılmaz, abacıdır. 2. Bedavadan

		yiyip, içen.
35	abadan (I)	1. Sofra örtüsü. 2. Aşağı yukarı dört metre kare genişliğinde, üzeri iki ayrı renkte küçük karelerle süslü olan; yatak, yorgan sarmakta örtü olarak veya kışın omuz atkısı olarak kullanılan yün dokuma, ince battaniye. 3. Kalın kumaştan yapılmış, işlemeli, cepkene benzer bir çeşit ceket.
39	aba gibi atmak	1. Büyük söz söylemek, mübalâğalı konuşmak. 2. Yüksekten atmak, palavra savurmak. 3. Öğünmek. 4. Kar lapa lapa, iri parçalar halinde yağmak.
44	abalımı	1. Tembel, uyuşuk. 2. Beceriksiz. 3. Şaşkın.
60	abalamak (I)	1. Çocuk emek lemek, emekliyerek yürümeğe başlamak. 2. Yerde sürünmek, sürünerek yürümek, yere eğilip emekler gibi yürümek: Avcı abalıyarak ayının inine girdi. 3. Çabalamak: Şapkam kayboldu abaladım abaladım bulamadım. 4. Sendelemek.
73	abamak (I)	1. Çirkin bir kızı veya kötü bir malı kandırma yoliyle birine yamamak. 2. Bir şeyi birisine zorla vermek. 3. Yüklemek: Suçu benim üzerime aba-dılar. 4. İftira etmek.
74	abamak (II)	1. Menetmek, alıkoymak: Bu adamı işinden abadım. 2. Muhalefet etmek, dayatmak.
77	abana (II)	1. Ana su. 2. Suyu bol memleket.
83	abanges (II)	1. Aklı eksik, kıt akıllı: Osmanın abanges olduğu yürümesinden belli. 2. Her

		hareketinde beceriksizlik sezilen adam.
97	abanuz [abanus]	1. Abanoz.
102	abara (I)	1. Su değirmenlerinde suyun basıncını çoğaltmak için yapılan, büyük bir huni şeklindeki hazne. 2. Tarlalarda bir taraftan bir tarafa su geçirmekte kullanılan tahta oluk. 3. Çift demiri ve pullukla açılan su yolu: Tarlaya abara çektim. 4. Çift demirin açtığı çizgi, saban izi. 5. Su oluğunun iki başından üstüne oturduğu duvar. 6. Köy evlerindeki tavanlarda iki direk arasındaki boşluk. 7. Tünel.
103	abara (II)	1. Buğday ambarı. 2. Hayvan yemliği.
114	abarmak (II)	1. Almak, alıvermek. 2. Alıp götürmek.
119	abartmak (I)	1. Olduğundan fazla göstermek, büyütmek, mübalâğa etmek. 2. Yalan söylemek, uydurmak. 3. Pohpohlamak, öğmek. 4. Nazlandırmak, şı martmak. 5. Kışkırtmak, teşvik etmek.
129	abaskal	1. Bedava yaptırılan iş: Ne abaskalın var. 2. Yarıda bırakılan iş.

Çok anlamlı sözlerin sorguları özelleşitirilebilmektedir. Hem çok anlamlı olup hem de insan için kullanılan sözleri bulanilmek için sözlükte (kimse) olarak işaretlenmiş anlam özelliğinden yararlanılabilir. Bunun için ana sorgulama sayfasının anlam kutucuğu sayısı ikiye çıkarılarak içerir sorgusu olarak “1.” ve “kimse” eklenir. Anlam içindeki göndermelerin elenmesi için içermez sorgusuna “[“ karakteri de eklenerek sorgu tamamlanır. Bu biçimde hazırlanan sorgu sonucunda 510 adet sözün insan için çok anlamlı kullanıldığı görülmektedir.



Şekil 67. Çok Anlamlı Sözler Ayrıntılı Sorgu Formu.

Tablo 6. Çok Anlamlı Sözler Ayrıntılı Sorgu Sonuç Örneği.

No	Maddebaşı	Anlam
914	aç kabadayı	1. Kabadayılık yapan yoksul ***kimse***. 2. - Yoksul olduğu halde şuna buna yardım eden.
1276	ağa (II)	1. Şeyh. 2. Seyit, peygamber sülâlesine mensup ***kimse***.
1501	ağır (I)	1. Olgun, terbiyeli, oturaklı, aklı başında. 2. İtibarlı, hatırı sayılır ***kimse***). 3. Kıymetli eşya, para: Evde ağırın, dağda davarın olsun. 4. Beyaz, mor veya başka renk çizgili kumaştan yapılan kadın elbisesi. 5. Kâbus: Bu gece beni ağır bastı.
1540	ağıt (I)	1. Ölen bir ***kimse***nin iyi hallerini ve ölmesinden duyulan acıları sayıp dökmek üzere söylenen ezgi, ezgi ile mersiye söyliyerek ağlama, 2. Gelin olan bir kızın arkasından meziyetlerini sayıp dökerek ağlama.
1543	ağıtçı	1. Ağıt söyliyen, ağıt söyliyerek cenaze sahiplerini ağlatan. 2. Gelin olan kızın etrafındaki topluluğa söyle diği ağıtlarla

		hüzün veren, onları ağlatan ***kimse***).
1563	ağız bağı (II)	1. Yalancı şahitlik yapacak ***kimse***ye verilen rüşvet. 2. İstenilen bir kızın verilmesine razı olmıyan, zorluk gösteren akraba ve dostlarına erkek tarafından verilen hediye. 3. Düğün gecesi gelini konuşturmak için damadın verdiği hediye.
1675	ağmak (I)	1. Aşağı inmek, yük veya terazide denge bozularak bir yanı ağır gelmek. 2. Aşmak: Ben tüfeği ateşleyene kadar tepeyi ağdı. 3. *Yağmur*inmek, düşmek. 4. Yıldız, yılan v.b. kaymak: Gökten bir yıldız ağdı. 5. Toplanmak, üşüşmek: Yemeğe karıncalar ağmış. 6. Güneş batmak: Gün ağdı. 7. Çömelmek, otur mak: Yanına ağıver. 8. Yatıp yuvarlanmak, debelenmek: Bu gece yatakta ağdım döndüm. 9. Bir ***kimse***nin üzerine yüklenmek, çullanmak, asıl mak: Üzerime fazla ağma. 10. Yükselmek, yukarı çıkmak: Gönülsüz namaz, göklere ağmaz. Sürü yamaca ağdı. 11. Hayvanlar çiftleşmek.
1735	ağrek	1. Koyun, keçi ve sığırların yaylımda dinlendikleri yer. 2. İşsiz, avare ***kimse***lerin toplandığı yer, delikanlılar meclisi.
1744	ağrıcaklı	1. Hastalıktan kurtujamıyan ***kimse***, hastalıklı, ağrılı sızılı. 2. Hastalıklı hayvan. 3. Gözü hastalıklı ***kimse***.

2071	ahırlamak	1. Her hangi bir hizmet hayvanını kış süresince ahırda beslemek. 2. Hayvan ahıra alışmak. 3. Hayvan uzun zaman ahırda kalarak hamlaşmak: Hayvana ara sıra bin ki ahırlamasın. 4. Uzun zaman ahırda kalan hayvan zayıflamak: Ahırlayan hayvanları uzun zaman kırlarda otlatmah. 5. Uzun zaman ahırda kalan hayvan yabanileşmek, hırçınlaşmak: Bizim at ahırlamış, ***kimse***yi yanına yaklaştırmıyor. 6. Uzun zaman ahırda kalmış bir hayvan, havalar ısınmadan otlağa çıkarıldığında hastalanmak.
2111	ahlı	1. Duygulu, içli ***kimse***. 2. Ah almış, ilençîi, beddualı ***kimse***.
2190	ahu (II)	1. Ak sakallı, saygı değer yaşlı adam. 2. Sevgili, sevimli ***kimse***.
2259	akarcalı	1. Başkasının malına zarar veren ***kimse***. 2. Bünyece zayıf, hastalıklı ***kimse***.
2792	akyavaş (I)	1. Sinsice hareket eden insan veya hayvan. 2. Ağır kanlı, tembel ***kimse***. 3. Bit.
2806	al (I)	1. Hile, tuzak: Haydin arkadaşlar al oldu bize, Sılanın dikenleri gül oldu bize. 2. Lohusa kadınların üstüne çökerek onları boğduğu sanılan görüntü. 3. Kadınlardan başka, bir ***kimse***nin üstüne çökerek onları boğduğu sanılan görüntü.
3085	alak (II)	1. Oyunlarda aynı taraftaki eş, bir guruba katılan ***kimse***: Bizim alak nereye gitti.

3250	alaşa (II)	1. Kötü kadın, , orospu, oynak, cilveli. 2. Çok süslü, allı pullu. 3. Herkesçe beğenilen, hoş görülen, yakışıklı ***kimse***. 4. İkiyüzlü, ara bozucu, yaltaklık eden. 5. Alçak, engin, basık: Duvar da çok alasaymış. 6. Çok aceleci, her işte acele eden. 7. Yaramaz, hırçın, yaygaracı.
3407	al basma (I)	1. Loğusa kadınlarda görülen bir hastalık, loğusa humması: Aman loğusayı yalnız bırakmayın al basar. 2. Loğusa olmıyaıı ***kimse***lerde al denilen görüntünün uyku arasında verdiği boğucu u k.ntı, kâbus. 3. Üzüntü ve kederden bayılma. 4. Sara hasta lığı. 5. Kızıla benzer bir çeşit hastalık.
4187	anadan cavlak	1. Doğuştan kel olan ***kimse***. 2. Doğuştan deli olan ***kimse***.
4231	analık (I)	1. Üvey anne. 2. Kaynana. 3. Düğünde sağdıçlık eden ***kimse***nin karısı, annesi. 4. Gelini düğüne hazırlıyan kadın. 5. Kardeşlik olan kadın ve kızlara şerbeti içiren kadın.

Anlam tablosunun içeriğindeki sorgulamanın daha özelleştirilmesi için anlam-il, anlam-ilçe ve anlam-köy gruplamaları yapılmıştır. Bu yapılırken her bir gruba ait tablolar oluşturularak her bir grubun sorgu sayfasında, gruba ait arama seçenekleri de eklenmiştir. Böylece, kullanıcı hangi anlamın hangi il, ilçe ya da köyde kullanıldığını görebilmektedir. Aşağıdaki şekilde bu gruplamalardan anlam-il gruplama tablosunun sorgu sayfası görülmektedir.



Şekil 68. Anlam-İl Gruplama Tablosu Sorgu Sayfası

3.3. Kavram Sorgulaması

İnsan, hayvan, bitki, madde, yiyecek, araç vb. genel nitelikli kavramların sorgulaması da anlam tablosu üzerinden yapılabilir. Bu noktada sözlükte anlam bölümünde verilen “madde”, “kimse”, “insan”, “bitki”, “bir çeşit”, “bir tür” gibi açıklayıcı nitelikteki işaretlemelerden yararlanılması gerekmektedir. Sözlükte açıkça belirtilmemiş anlam nitelikleri için, kullanıcının bu türden genel kavramları bulabilmesi için sözlüğün söz listesinden yararlanması gerekmektedir. Bu amaçla DS’de anlam bölümünde geçen bütün sözlerin sıklıkları çıkarılmış ve bunlar derleme listesi tablosuna kayedeilmiştir. Bu tabloda en sık geçen sözlerin listesi ve yüzdeleri sıklığı en çok olandan en aza doğru sıralanmıştır. Aşağıdaki listede bu tablonun çıktısından bir bölüm yer almaktadır. Tabloda yer alan “bir-Bir”, “Çok-çok”, “Küçük-küçük” gibi hem büyük hem küçük karakterli sözler, tanımlarda geçen bu sözlerin cümle başında ve içinde kullanılmasına göre değişkenlik gösterdiğinden listede hem büyük hem de küçük karakterli biçimleriyle yer almışlardır. Bu durum bize, “Bir” sözünün tanıma başlarken, “bir” sözünün ise tanım içinde kullanıldığını göstermektedir.

Tablo 7. Anlam Tablosu Söz Sıklığı Örneği.

No	Söz	Sıklık	Yüzde
1	bir	7449	1,80
2	ve	5488	1,33
3	çeşit	4171	1,01
4	için	4061	0,98
5	Bir	3123	0,76
6	yapılan	2512	0,61
7	da	2101	0,51
8	kimse	2069	0,50
9	ya	1911	0,46
10	olan	1593	0,39
11	yer	1582	0,38
12	kullanılan	1458	0,35
13	ağaç	1456	0,35
14	Çok	1311	0,32
15	gibi	1174	0,28
16	vb	1170	0,28
17	etmek	1154	0,28
18	Küçük	1150	0,28
19	küçük	1105	0,27
20	ot	1011	0,24
21	su	1009	0,24
22	olmak	986	0,24
23	yarayan	974	0,24
24	çok	930	0,23
25	hayvan	872	0,21

26	toprak	842	0,20
27	çocuk	826	0,20
28	yemek	787	0,19
29	ile	763	0,18
30	yapılmış	744	0,18
31	yapmak	721	0,17
32	ince	676	0,16
33	büyük	659	0,16
34	iki	659	0,16
35	konulan	643	0,16
36	bitki	615	0,15
37	araç	584	0,14
38	Büyük	573	0,14
39	olarak	559	0,14
40	sonra	552	0,13
41	Su	549	0,13
42	renkli	546	0,13
43	kadın	539	0,13
44	yavrusu	534	0,13
45	uzun	527	0,13
46	üzüm	518	0,13
47	veya	515	0,12
48	taş	502	0,12
49	tahta	496	0,12
50	gelmek	462	0,11
51	iş	454	0,11
52	şey	453	0,11

53	ses	452	0,11
54	buğday	442	0,11
55	demir	442	0,11
56	yetişen	437	0,11
57	ekmek	430	0,10
58	yere	424	0,10
59	parçası	417	0,10
60	kuş	414	0,10
61	takılan	414	0,10
62	keçi	409	0,10
63	vermek	408	0,10
64	oyunu	401	0,10
65	İki	396	0,10
66	iyi	394	0,10
67	anlamında	392	0,09
68	çıkan	388	0,09
69	ünlemi	384	0,09
70	tarla	382	0,09
71	Hayvan	380	0,09
72	olmayan	379	0,09
73	gelen	377	0,09
74	yeri	375	0,09
75	söz	374	0,09
76	adam	370	0,09
77	erkek	370	0,09
78	şeyi	370	0,09
79	çıkarmak	365	0,09

80	kadar	354	0,09
81	koyun	354	0,09
82	kullanılır	350	0,08
83	kalan	349	0,08
84	bulunan	345	0,08
85	hastalığı	343	0,08
86	az	341	0,08
87	üstüne	337	0,08
88	Çocuk	331	0,08
89	parça	331	0,08
90	denilen	330	0,08
91	İnce	329	0,08
92	beyaz	328	0,08
93	meyve	327	0,08
94	zaman	326	0,08
95	hastalık	324	0,08
96	üzerine	323	0,08
97	biçiminde	322	0,08
98	Bu	322	0,08
99	Mısır	313	0,08
100	benzer	311	0,08
101	tatlı	306	0,07
102	verilen	304	0,07
103	iri	302	0,07
104	işi	302	0,07
105	almak	298	0,07
106	ağacı	295	0,07

107	Ağaç	294	0,07
108	ist	294	0,07
109	Yeni	291	0,07
110	at	290	0,07
111	insan	287	0,07
112	ip	286	0,07
113	Her	282	0,07
114	kalın	277	0,07
115	sert	275	0,07
116	Az	274	0,07
117	yuvarlak	270	0,07
118	yapan	265	0,06
119	anlatır	263	0,06
120	kız	262	0,06
121	Kısa	260	0,06
122	ucu	260	0,06
123	gitmek	259	0,06
124	yol	259	0,06
125	kalmak	258	0,06
126	mısır	257	0,06
127	oyun	254	0,06
128	ekin	253	0,06
129	konuşmak	251	0,06
130	tutmak	251	0,06
131	benzeyen	248	0,06
132	içine	248	0,06
133	Toprak	248	0,06

134	içinde	246	0,06
135	yüz	245	0,06
136	kısmı	243	0,06
137	pişirilen	240	0,06
138	boylu	238	0,06
139	konuşan	238	0,06
140	açılan	237	0,06
141	At	236	0,06
142	Kötü	236	0,06
143	tahıl	235	0,06
144	Uzun	235	0,06
145	birbirine	232	0,06
146	kap	230	0,06
147	para	229	0,06
148	her	227	0,05
149	İri	227	0,05
150	kırmızı	227	0,05

Derleme Sözlüğü veri tabanından yorumlanması amacıyla sözlükteki ipuçları yardımıyla bilgi çıkarımı yapmak mümkündür. Bunun için sözlüğün yazım düzenini, üslup özelliklerini bilmek önemlidir. Tanımlar, cümle sonu örüntüleri adını verebileceğimiz yapılara sahiptir. “insan”, “nesne”, “madde”, “şey”, “kişi”, “renk” gibi sözler, kullanılan cümlenin sonunda “.” ya da “:” ile bitirilmektedir. Şu cümlelerde bu sözlerin kullanımı görülmektedir: “Yumuşak, sözünü geçiremeyen **kişi.**”, “Zayıf, çarpık **insan.**”, “Yapışkan bir **madde.**”, “Yıkılmak ya da devrilmek üzere olan **şey.**” **“**Siyahla kırmızı arası bir **renk.**” “[… **aranılan söz.]**” örüntüsüyle çıkarımı yapılacak genel kavramların sayısı artırılabilir.

Yukarıdaki cümlelerdeki yapıları veri tabanından çıkarmak için anlam kutucuğuna “insan.”, “kişi.”, “madde.”, “şey.”, “renk.” sorgu ifadeleri girilerek Anadolu ağızlarında geçen insan, renk ve madde adlarına ulaşılabilir. Bu çalışmada “insan”,

“kişi”, “madde” ve “renk” kavramları ile derlendikleri iller arasındaki ilişkiyi belirlemek için kavramların geçtiği iller bölgelere göre atanmış ve sıklıkları çıkarılmıştır. Aşağıdaki tablo ve grafiklerde bölge ve illere göre dağılımlar verilmiştir.

Tablo 8. “insan.” Sorgusunun Bölge ve İllere Göre Dağılımı.

Kavram	**Bölge**	**İl**	**Sıklık**
“insan”	Ege	Muğla	4
		Aydın	3
		Denizli	3
		Afyon	1
		İzmir	1
		Manisa	1
		Toplam	*13*
	Akdeniz	Burdur	4
		Isparta	4
		İçel	3
		Adana	1
		Antalya	1
		Toplam	*13*
	Karadeniz	Artvin	4
		Trabzon	4
		Bolu	3
		Giresun	3
		Ordu	3
		Gümüşhane	2
		Kastamonu	2
		Çorum	2
		Amasya	1
		Rize	1

		Samsun	1
		Tokat	1
		Toplam	*27*
	Marmara	Balıkesir	2
		Kocaeli	1
		Çanakkale	1
		toplam	4
	İç Anadolu	Sivas	7
		Eskişehir	4
		Kırşehir	2
		Ankara	1
		Niğde	1
		Çankırı	1
		Toplam	*16*
	Doğu Anadolu	Erzincan	2
		Van	2
		Ağrı	1
		Bitlis	1
		Elazığ	1
		Toplam	*7*
	Güneydoğu Anadolu		0

Yukarıdaki tabloda da görüldüğü gibi Karadeniz bölgesi için “insan” ile biten tanımlar çoğunluktadır. “insan” kavramının alt düzey kavramı olarak düşünülebilecek “kişi” kavramı içinde aşağıdaki sonuçlar elde edilmiştir.

Tablo 9. “kişi.” Sorgusunun Bölge ve İllere Göre Dağılımı

Kavram	**Bölge**	**İl**	**Sıklık**
“Kişi”	Ege	Denizli	2
		Afyonkarahisar	1
		İzmir	1
		Kütahya	1
		Toplam	*5*
	Akdeniz	Adana	2
		Isparta	2
		Antalya	1
		Burdur	1
		İçel	1
		Kırşehir	1
		Toplam	*8*
	Karadeniz	Trabzon	3
		Çorum	3
		Amasya	2
		Giresun	2
		Samsun	2
		Artvin	1
		Bolu	1
		Toplam	*14*
	Marmara	Kocaeli	1
	İç Anadolu	Niğde	14
		Sivas	4
		Ankara	3
		Kayseri	3
		Konya	3

		Yozgat	1
		Toplam	*28*
	Doğu Anadolu	Kars	1
		Toplam	*1*
	Güneydoğu Anadolu	Gaziantep	1
		Toplam	*1*

"insan" ve "kişi" kavramları arasında Karadeniz ve İç Anadolu bölgesi arasında bir farklılık görülmektedir. Tanımlarda "insan" için Karadeniz, "kişi" için İç Andolu bölgesine bir eğilim görülmektedir. "madde" kavramı için elde edilen sonuçlar da aşağıdaki gibidir:

Tablo 10. "madde." Sorgusunun Bölge ve İllere Göre Dağılımı

Kavram	**Bölge**	**İl**	**Sıklık**
"madde"	Ege	Denizli	2
		Kocaeli	2
		Kütahya	1
		Toplam	*5*
	Akdeniz	İçel	3
		Burdur	2
		Isparta	2
		Adana	1
		Toplam	*8*
	Karadeniz	Amasya	2
		Giresun	2
		Samsun	2
		Sinop	2

		Tokat	2
		Artvin	1
		Gümüşhane	1
		Kastamonu	1
		Ordu	1
		Trabzon	1
		Çorum	1
		Toplam	*16*
	Marmara	Çanakkale	1
		Toplam	1
	İç Anadolu	Ankara	2
		Sivas	2
		Kayseri	1
		Konya	1
		Yozgat	1
		Toplam	*7*
	Doğu Anadolu	Erzincan	3
		Erzurum	1
		Malatya	1
		Muş	1
		Toplam	*6*
	Güneydoğu Anadolu	Gaziantep	1
		Toplam	*1*

“renk” kavramı için elde edilen tabloya bakıldığında Karadeniz bölgesi illerinin sayısı ve sıklığı dikkat çekmektedir.

Tablo 11. “renk.” Sorgusunun Bölge ve İllere Göre Dağılımı

Kavram	**Bölge**	**İl**	**Sıklık**
"renk"	Ege	İzmir	1
		Manisa	1
		Muğla	1
		Toplam	*3*
	Akdeniz	Adana	4
		Kahramanmaraş	5
		İçel	3
		Isparta	1
		Toplam	*13*
	Karadeniz	Kastamonu	3
		Ordu	2
		Çorum	2
		Amasya	1
		Bolu	1
		Gümüşhane	1
		Trabzon	1
		Toplam	*11*
	Marmara	İstanbul	2
		Tekirdağ	1
		Çanakkale	1
		Toplam	*4*
	İç Anadolu	Kırşehir	2
		Niğde	2
		Sivas	2
		Konya	1
		Yozgat	1

		Toplam	*8*
	Doğu Anadolu	Kars	5
		Elazığ	2
		Erzurum	1
		Toplam	*8*
	Güneydoğu Anadolu	Gaziantep	2
		Şanlıurfa	1
		Toplam	*3*

Yukarıda verilen “madde”, “insan”, “kişi” ve “renk” kavramlarıyla ilgili sorgu sonuçlarında Karadeniz bölgesinden sözlüğe giren verinin daha çok olduğu görülmektedir. Doğu ve Güneydoğu Anadolu bölgelerinden gelen verinin görece az olduğu dikkati çekmekte, “kişi” sorgusu için Niğde ilinden gelen verinin yüksek sıklığı öne çıkmaktadır.

3.4. Derleme Yeri Sorgulaması

DS’de derleme yeri sorgulaması için sadece derleme yerinde geçebilecek il, ilçe ve köy adları sorgulanabilirken, sorgulama ana sayfasında derleme yeriyle birlikte anlam, değişke, madde başı ve tanık sorgulaması da yapılabilir. Derleme yerinde sözlükte kullanılan “*”, “-“ işaretleri sırasıyla ilçeleri ve illeri göstermektedir. Bu işaretlerden yararlanılarak örneğin anlamında üzüm geçen ve derleme yerinde de bütün iller listelenmek istendiğinde, derleme yeri kutusuna “-“ işaretini koymak yeterlidir. Aşağıdaki tablo yukarıda belirtilen üzüm-il sorgusunadan elde edilen sonuçlardan örnekleri göstermektedir.

Tablo 12. Anlam-Derleme Yeri Sorgu Sonucu Örneği

No	Anlam	Derleme yeri
335	1. Ceviz ağacının mobilya yapmaya	(-Zonguldak; Hamidiye, *Pazar, *Fındıklı çevresi,

	yarıyan iç kısmı. 2. Ağaçların dayanıklı iç kısmı. 3. Yaban armudu, ahlat. 4. İyice kararmamış, olgunlaşmamış ***üzüm***.	Karadere, -Rize; *Hopa -Artvin)(Kalafka -Trabzon; Ardeşen *Pazar -Rize)(Körküler *Yalvaç -Isparta)(*Kilis -Gaziantep)
371	1. Çilli ve çopur yüzlü, sarı saçlı, açık renk gözlü adam. 2. Yüzü, vücudu alaca benekli, lekeli hayvan veya adam. 3. Alnındaki beyazlık alt dudağı na kadar inen -at, inek, manda, köpek b.v.-hayvan. 4. Doru at. 5. Hay vanların kıç taraflarındaki leke. 6. Alaca bulaca, karışık renk li: Aldığın yün abraş çıktı.*Alaşehir -Mn.; -Gaz. köyleri; Kızıllar *Karaman -Kn.; *Silifke -İç.) 7. İki tarafı denk gelmemiş halı. 8. Halıdaki alacalık, renk bozuk luğu. 9. Bicimsiz, çirkin. 10.	(*Keçiborlu, -Isparta; *Karşıyaka, Seyrekköy *Menemen -İzmir; *Gelibolu -Çanakkale; Bozan -Eskişehir; *Kurşunlu, Genek -Çankırı; -Sinop; Meydancık bucağı köyleri *Şavşat -Artvin; *Kilis, -Gaziantep; -Niğde; Ortaca, *Akşehir,-Konya; *Mut ve köyleri -İçel)(*Bergama -İzmir; Aşağı-kayı *Tosya -Kastamonu; -Amasya; -Niğde; *Karaman -Konya; *Mut -İçel)(Bor nova -İzmir; -Bursa; Baklalı *Çatalca -İstanbul; Eraslan -Amasya; *Maçka köy leri, Kalafka-Trabzon; *Bozukkale-Rize -Siirt; -Gaziantep; *Antakya -Hatay; -Adana Ortakçı *Lalapaşa -Edirne; -Kırklareli)(*Safranbolu -Zonguldak)(*Dikili -İzmir;

	Çarpık, eğri. 11. Şaşı: Şu çocuk abraş bakıyor. 12. Sert huylu, ters, kaba, muaşeret bilmez. 13. Patavatsız, sözü hoşa gitmeyen. 14. Tedirgin edici, obur. 15. Tembel, uyuşuk, ağır canlı. 16. Pinti, hasis. 17. El ve yüzde, beyaz lekeler halinde olan bir çeşit deri hastalığı. 18. Çil. 19. Yarı olmuş, iyice kararmamış ***üzüm***. 20. Benekli, iyi kurumamış tütün yaprağı. 21. Düzgün olmı-yan, pütürlü, pürüzlü deri.	Yeniköy -Balıkesir; *Ezine köyleri, Fili *Biga - Çanakkale; *Mudanya - Bursa; *Kandıra -Kocaeli; *Çarşamba -Samsun; - Amasya ilce ve köyleri; *Zile, Çilehane *Reşadiye - Tokat; Zile *Mesudiye, Kuzköy *Akkuş -Ordu; *Kangal -Sivas; *Mut köyleri -İçel; *Gazipaşa köyleri -Antalya; Kavaklı, *Meriç -Edirne; Çavuşköy *Babaeski - Kırklareli)(Keçek *01tu - Erzurum)(Varay-Amasya)(- Kayseri)(*Alaşehir - Manisa; -Gaziantep köyleri; Kızıllar *Karaman -Konya; *Silifke -İçel)(Dallıca *Nazilli -Aydın; Bahçeli *Bor -Niğde)(*Bünyan- Kayseri; Bahçeli*Bor - Niğde)(Karahisar, *İncesu - Kayseri; *Kayalı, - Konya)(Yeniköy *?Or hangazi-Bursa; Bozan - Eskişehir; *Mesudiye - Ordu; *Maçka köyleri - Trabzon; -Gümüşhane; Önsen, Hop uru, - Kahramanmaraş; Hacıilyas *Koyulhisar -Sivas; Divanlı

		-Yozgat; -Niğde)(Hacıilyas *Koyulhisar -Sivas; -Niğde; -Konya)(Bozan -Eskişehir; Uzunmusa -Ordu; Kayadibi-Giresun)(*Çal -Denizli; Mencilis *Keşap -Giresun)(*Kargı -Kastamonu; Güneyce, *İkizdere -Rize; Eldelek *Elbistan -Kahramanmaraş)(Bozan -Eskişehir; Eldelek *Elbistan, -Kahramanmaraş)(-Çorum)(Karadere -Rize)(-Çorum)(*?Antakya köyleri -Hatay; -Niğde)(Bozan -Eskişehir; *Kilis -Gaziantep)(Lohan, -Gaziantep)(*Alaçam -Samsun)(-Samsun)
637	1. Amerikan bağ çubuğu. 2. ***Üzüm***ü acı ve küçük olan aşısız bağ çubuğu.	(Ekse *Çal -Denizli; Beki *Karahallı -Uşak)(Irgıllı *Çivril -Denizli)(Ekse *Çal -Denizli)
696	1. Sık, yuvarlak ve küçük taneli bir çeşit ekşi ***üzüm***. 2. Bir çeşit elma.	(Kozluca *Keçiborlu, Akça-şar, *Eğridir, Atabey -Isparta; Bunak, Karamanlı, *Tefenni, Akça-köy *Yeşilova, Çerçin -Burdur; Mo-raca *Çal, Çöplü *Çivril, Babadağ, Yeniköy,

		Ahallı *Sarayköy, Kösten - Denizli; Çomaklı *Korkuteli, Kışla, *Elmalı - Antalya; Genek *Yatağan - Muğla)(İlyas *Keçiborlu,- Isparta; Ürkütlü *Bucak, - Burdur; İğdir *Çiv-ril, Moran -Denizli)(Söğüt *Tefenni -Burdur)
717	***Üzüm***.	(Mençek *Ermenek - Konya)
782	Bahçe kenarlarında yetişen ve ***üzüm***den küçük, siyah meyvası olan bodur bir ağaç.	(-Van)
1023	Bir cins siyah ***üzüm*** ve bu ***üzüm***ün ekşi pekmezi.	(*Ünye-Ordu)
1068	İnce kabuklu, güzel kokulu bir çeşit beyaz ***üzüm***.	(*Bor-Niğde)
1163	***Üzüm*** suyu çıkarılan yer, şırahane.	(Navlu *Yeşilova -Burdur)
1332	***Üzüm*** suyunun daha çabuk ve daha kolay durulması için, içine konulan beyaz toprak veya ufaltılmış taş kırıntıları.	(-Şanlıurfa)

1377	Bir çeşit ***üzüm***.	(Kesirik -EL)
1410	1. ***Üzüm*** sıkılan ve pekmez yapılan yer, şırahane. 2. Pekmez yapmaktan başka işe yaramıyan ***üzüm***.	(*Elmalı -Antalya)(*Elmalı -Antalya)
1437	Çok tatlı, kokulu, iri sal-kımlı bir çeşit ***üzüm***.	(-Niğde)
1665	Bir çeşit ***üzüm***.	(-Tokat köyleri)
1692	Bir çeşit ***üzüm***.	(*Çemişkezek -Tunceli; Kesirik -Elazığ)
1797	1. [-> ağrışak]2. Her hangi bir cismin dönmesi için takılan ağırca demir parçası: Bu taş dönmüyor, ağşak getirelim. 3. Hayvan nallarını delmekte kullanılan çelik parça. 4. [-> ağırşak -2]5. Zımba ile beraber kullanılan ve çivi delmeye yarıyan bir kalaycı aleti. 6. Koyunların kuyruk altında tüylerine yapışarak kuruyan pislik parçaları. 7. Tavı geçmiş tarladaki	(Vasıldan *Divriği-Sivas)(Güdül *Ayaş -Ankara)(-Ankara)(Bahçeli *Bor -Niğde)(Bahçeli *Bor -Niğde)(*Bor -Niğde)(Ekse *Çal-Denizli)

	kesekler: Ağşakları saban da ezmiyor. 8. Çıbanın altındaki sertlik. 9. Sarı renkte, iri taneli bir çeşit ***üzüm***. 10. [-> ağırşak -5]	
1816	1. Badana yapmakta kireç yerine kullanılan bir çeşit beyaz toprak. 2. Pekmez yapılırken ***üzüm***ün ekşitsini almak için kullanılan toprak.	(*Bayburt köyleri - Gümüşhane köyleri; - Gaziantep; Cihanpaşa - Yozgat; Mahzemin - Kayseri)(Kızılçakçak *Arpaçay -Kars)(- Gaziantep)(-Gaziantep)(- Gaziantep)
1851	1. Beyaz renkte bir çeşit ***üzüm***. 2. Yeşilimsi beyaz renkte bir çeşit ***üzüm***.	(Kuşbaba *Bucak -Burdur; İsabey *Çal - Denizli)(Bereketli *Tavas - Denizli; *Erciş -Van)
1971	Bir çeşit kara ***üzüm***.	(Er-kinis *Yusufeli -Artvin)
2058	1. [-> ahar I) -1]2. [-> ahar I) -3]3. [-> ahar I) -4]4. [-> ahar I) -5]5. [-> ahar II) -5]6. İçinde ***üzüm*** çiğnenen ağaç tekne, şırahane.	(Çerçin -Burdur; Karadere *Gündoğmuş -Antalya)

Derleme yeri tablosunda sadece iller listelenmek istendiğinde derleme yeri kutucuğu alan ekle özelliğiyle çoğaltıldıktan sonra içerir sorgusuna "-" işareti, içermez sorgusuna da "*" işareti konularak ilçeler elenmiş olmaktadır. Bu sorgu sonucuna göre

yaklaşık 23.268 adet kaydın içinde il geçmektedir. Hem iller hem de ilçelerin toplamı da iki derleme yeri kutucuğuna içerir sorgusuyla “*”, “-“ eklendiğinde 68.336 sonucunu vermektedir.

Derleme yeri tablosundaki bilgilerden yararlanarak en çok hangi il, ilçe ya da köyden derlemenin yapıldığını belirlemek için derleme yeri tablosunun bir sıklık listesi çıkarılmıştır. Aşağıdaki sıklık listesi tablosunda da yer aldığı gibi en çok köylerden derleme yapıldığı görülmektedir. İl olarak da Niğde dikkati çekmektedir. Bu sonuçlarda hangi faktörlerin etkili olduğunu anlamak güçtür. Bölgede derleme yapan kişilerin yoğun olarak çalıştığı mı, bölge halkının ilgisinin mi etken olduğunu belirlemek gerekmektedir.

Tablo 13. Derleme Yeri Sıklık Listesi Örneği

No	Derleme Yeri	Sıklık	Yüzde
1	köyleri	15939	2,33
2	ve	14585	2,13
3	Niğde	11012	1,61
4	Konya	10362	1,51
5	Denizli	9769	1,43
6	Isparta	9582	1,40
7	Sivas	9536	1,39
8	Bor	8650	1,26
9	Ankara	6949	1,01
10	İçel	6821	1,00
11	Tokat	6721	0,98
12	Muğla	6416	0,94
13	Burdur	6119	0,89
14	Bolu	5622	0,82
15	Amasya	5442	0,79
16	Kayseri	5434	0,79
17	Adana	5422	0,79

18	Samsun	5257	0,77
19	Çorum	5249	0,77
20	Kars	5133	0,75
21	çevresi	4975	0,73
22	Ordu	4963	0,72
23	Eskişehir	4771	0,70
24	Kastamonu	4761	0,70
25	Kahramanmaraş	4708	0,69
26	Gaziantep	4637	0,68
27	İzmir	4512	0,66
28	Antalya	4367	0,64
29	Artvin	4069	0,59
30	Balıkesir	4010	0,59
31	Afyonkarahisar	4009	0,59
32	Gümüşhane	3977	0,58
33	Aydın	3788	0,55
34	Ermenek	3775	0,55
35	Trabzon	3730	0,54
36	Giresun	3711	0,54
37	Merzifon	3445	0,50
38	Kütahya	3404	0,50
39	Hatay	3261	0,48
40	Manisa	3257	0,48
41	Çankırı	3164	0,46
42	Malatya	3145	0,46
43	Erzincan	3124	0,46
44	Çal	3083	0,45

45	Çanakkale	3068	0,45
46	Silifke	2997	0,44
47	Bursa	2975	0,43
48	Eğridir	2942	0,43
49	Yozgat	2883	0,42
50	Mut	2705	0,39

DÖRDÜNCÜ BÖLÜM

DERLEME SÖZLÜĞÜ'NÜN SÖZLÜKSEL GÖRÜNÜMLERİ

4.1. Eşdizimlilik Görünümleri

DDİ'de n-gram terimiyle bir derlemde geçen her bir birim n-gram olarak düşünülebileceği gibi sözcük düzeyinde de n-gramlar ele alınabilmektedir. Sözcük düzeyinde n-gramlarda, birimlerin sıklığı en önemli ölçütlerden biridir ve bu sıklık, sıklığı hesaplanan birimin ardından gelebilecek birimin sıklığıyla yakından ilişkilidir. Örneğin bir derlemde ***koyun*** sözcüğü bir sözcük n-gram olarak düşünüldüğünde, koyun'un sıklığı ile ***sürüsü*** sözünün sıklığı bu iki n-gramın dil bilimsel olarak bir eşdizim kurma olasılığını artırmaktadır. İçerik çözümlemesi yapabilen yazılımlarda eşdizimlilikleri ortaya çıkarmak için her bir birimin sıklığı hesaplanıp, sonra bu sıklıkların olasılıksal anlamlılıklarıyla birlikte sözcüklerin eşdizim yapıları belirlenebilmektedir.

Belirlenecek n-gramların boyutu ya da uzamı genellikle 2 ve 3 birimlik aralıklarda hesaplanmaktadır. 2'li birimlerin çıkarımında örneğin ***kapıyı açık tutmak*** deyiminde olduğu gibi uzamın dışında kalabilecek sözlerin olabileceği de göz önüne alınarak n-gram uzamı 3 birim olarak tutulur, böylece ***kapıyı açık tutmak*** gibi 3 birimli bir eşdizim elde edilmiş olur. İstatistiksel DDİ'de öğrenen sistemlerin eğitiminde, n-gramlar ve her bir n-gramın sıklık bilgisi eğitim setinin en önemli parçalarından biridir.

Derleme Sözlüğü'nün eşdizimlilik yapısının ortaya çıkarılmasında n-gram uzamı 2 ve 3 birimlilerden oluşacak şekilde hesaplanmıştır. N-gram çıkarımı için sözlüğün asıl metin bölümü olan ve söz sıklığı hacmi yüksek olması yanında madde başı birimlerinin elde edilebileceği alan olması bakımından anlam tablosu esas alınmıştır. N-gram çıkarımı için derlem dil bilimi çevrelerinde sıkça kullanılan AntConc adlı yazılım kullanılmıştır. Aşağıdaki tabloda n-gram uzamı 2 olan birimlerden bir bölüm verilmiştir.

Tablo 14. Derleme Sözlüğü Anlam Tablosu 2 Birimli n-gramlar

Sıklık	**Olasılık**	**N-gram(2)**
2934	0,398	bir çeşit
1882	0,983	ya da
1166	0,375	Bir çeşit
578	0,234	yapılan bir
446	0,151	için kullanılan
363	0,088	çeşit ot
312	0,042	bir ot
279	0,194	kullanılan bir
275	0,037	bir bitki
224	0,153	olan bir
200	0,048	çeşit bitki
191	0,061	Bir şeyi
181	0,510	anlamında kullanılır
179	0,957	çağırma ünlemi
170	0,467	ses çıkarmak
163	0,055	için yapılan
159	0,038	çeşit yemek
159	0,038	çeşit üzüm
154	0,922	hangi bir
149	0,484	benzer bir
146	0,689	oynanan bir
144	0,272	renkli bir
143	0,019	bir araç
137	0,033	çeşit kuş
134	0,043	Bir şeyin

132	0,018	bir kuş
127	0,518	Kısa boylu
123	0,436	Her hangi
122	0,017	bir şey
122	0,230	çocuk oyunu
120	0,125	yarayan bir
118	0,016	bir şeyi
115	0,016	bir hastalık
112	0,640	meydana gelen
108	0,044	yapılan yemek
107	0,073	olan kimse
102	0,637	elde edilen
102	0,320	yapmak için
100	0,032	Bir işi
98	0,398	benzeyen bir
98	0,191	olarak kullanılan
97	0,023	çeşit tatlı
94	0,013	bir cins
90	0,012	bir ağaç
89	0,012	bir yere
89	0,093	yarayan araç
89	0,022	çeşit ağaç
88	0,203	yetişen bir
86	0,262	denilen bir
85	0,012	bir böcek
84	0,884	bildirir ünlem
83	0,020	çeşit oyun

81	0,039	da denilen
79	0,444	duruma gelmek
78	0,494	kokulu bir
78	0,019	çeşit hastalık
76	0,102	ile yapılan
75	0,025	için açılan
74	0,363	yapraklı bir
74	0,030	yapılan ve
73	0,018	çeşit buğday
72	0,035	da hayvan
72	0,017	çeşit çocuk
71	0,066	küçük bir
71	0,013	ve köyleri
71	0,703	Yavaş yavaş
71	0,317	yüz tutmak
69	0,423	dikenli bir
68	0,861	yapmakta kullanılan
67	0,123	hayvan için
67	0,151	vb şeyler
66	0,688	Alay etmek
66	0,308	altına konulan
66	0,327	Ufak tefek
65	0,670	çıkardığı ses
64	0,009	bir şeyin
64	0,496	yağan yağmur
63	0,306	insan ya
63	0,926	Mudurnu Bo

63	0,071	su kabı
63	0,778	ununda n yapılan
63	0,015	çeşit balık
62	0,008	bir oyun
61	0,953	iğneciler Mudurnu
61	0,473	yenilen bir
61	0,015	çeşit armut
59	0,277	Koyun keçi
58	0,892	Bayburt Sarıkamış
58	0,052	gibi olmak
58	0,040	olan ve
57	0,500	ağaçtan yapılmış
57	0,740	boynuna takılan
57	0,320	görülen bir
57	0,261	işe yaramaz
57	0,254	yüz tutmuş
57	0,014	çeşit mantar
56	0,008	bir çocuk
56	0,163	demir araç
56	0,050	Küçük çocuk
56	0,144	tahta araç
55	0,063	ağaç parçası
55	0,018	Bir cins
55	0,407	esen yel
55	0,019	için kullandıkları
55	0,833	koymaya yarayan
55	0,519	yenen bir

54	0,667	Parça parça
54	0,261	taneli bir
53	0,018	için kullanılır
53	0,083	konulan ağaç
53	0,283	yavaş yavaş
53	0,174	Çocuk dilinde
52	0,754	korumak için
52	0,010	ve uzun
52	0,825	îrişli Bayburt
51	0,007	bir araya
51	0,016	Bir şey
51	0,981	Bor Nğ
51	0,927	kovalama ünlemi
51	0,239	Koyun ve
51	0,053	yarayan tahta
51	0,440	çiçek açan
50	0,272	güzel kokulu
50	0,068	yapılmış bir
49	0,034	kullanılan ağaç
49	0,262	Köpek yavrusu
49	0,700	Sarıkamış Kr
49	0,278	Zayıf cılız
49	0,012	çeşit araç
48	0,449	Ağaçtan yapılmış
48	0,015	Bir yere
48	0,696	Domuz yavrusu
48	0,203	pişirilen bir

48	0,247	tahıl ölçeği
48	0,111	yetişen ve
48	0,449	çiçekli bir
48	0,090	çocuk dilinde
47	0,006	bir yemek
47	0,164	Mısır koçanı
47	0,758	Silifke İç
47	0,671	tahtadan yapılmış
46	0,301	almak için
46	0,189	koyun keçi
46	0,162	Yeni doğmuş
45	0,127	anlamında ilenç
45	0,692	karıştırılarak yapılan
45	0,542	Kendi kendine
45	0,207	yapan kimse
44	0,014	Bir kimseyi
44	0,200	Kadınların başlarına

Yukarıdaki tablonun sıklık sütunu en sık bulunan birimlerden en aza doğru sıralanmıştır. Olasılık sütununda ilk birimin diğer birimle olan ilişkisi görülmektedir. Bu ilşkinin anlamlılık derecesi olasılıksal olarak 1'e yakınlığına göre verilmiştir. *tahtadan yapılmış* n-gram yapısında *tahtadan* sözünün yapılmış ile birlikte bulunma olasılığı anlam tablosunda 0,671 olarak hesaplanmıştır. Bu değer 1'e ne kadar yakınsa anlamlılık değeri de o kadar artmaktadır. 1'e yakın olan birimler arasında anlamlı bir ilişkinin olduğundan söz edilebilir.

DS'de anlam tablosunda belirlenen 2'li n-gram toplam sayısı 210.758'dir. Bu sayının yüksek olması, birimlerin tekrarından kaynaklanmakta, tekrarlı birimler çıkarıldığında toplam 198.211 adet 2'li n-gram olduğu görülmektedir.

Anlam tablosunun 2'li n-gramlarına içerik olarak bakıldığında sıklığı en yüksek birimin sıklık sayısı 2934 olan *bir çeşit* sözü olduğu görülür. Bu sözün sıklığının yüksek

olmasında tanımlama anlatımının rolü olduğu düşünülebilir. Tanımı yapılan sözlerin daha çok nesnelerden oluştuğu kanısına da buradan ulaşılabilir. Sıklık sıralamasında ikinci sırada yer alan *ya da*; tanımlarda "olabilirliğin", bir diğer seçeneğin bulunduğu sözlerin derlendiğinin de göstergesi olarak yorumlanabilir. Bunun yanında, DS'de tanımlamalar büyük harfle başladığından, tanımlamalarda hem *bir çeşit* (2934), hem de *Bir çeşit* (1166) gibi iki farklı birimin hesaplanması, birinin tanıma başlarken diğerinin de tanım içinde kullanıldığını göstermektedir. Sıklığı yüksek ilk 150 2'li n-grama bakıldığında; *için kullanılan* (446), *kullanılan bir* (279), *olarak kullanılan* (98), *yapmakta kullanılan* (69), *için kullandıkları* (55), *için kullanılır* (53), *kullanılan ağaç* (49) "kullanmak" kavramıyla ilgili sözlerin tanımlarda önemli yer tuttuğu görülür.

N-gram uzamı 3 birim olanlar da çalışma kapsamında çıkarılmış ve bunların bir bölümü de aşağıdaki tabloda gösterilmiştir.

Tablo 15. Derleme Sözlüğü Anlam Tablosu 3 Birimli n-gramlar

Sıklık	Olasılık	N-gram(3)
450	0,267	yapılan bir çeşit
315	0,054	bir çeşit ot
173	0,030	bir çeşit bitki
148	0,025	bir çeşit yemek
131	0,022	bir çeşit üzüm
129	0,112	kullanılan bir çeşit
115	0,431	Her hangi bir
107	0,018	bir çeşit kuş
102	0,100	olan bir çeşit
82	0,014	bir çeşit tatlı
80	0,276	benzer bir çeşit
78	0,013	bir çeşit ağaç
78	0,174	renkli bir çeşit
72	0,038	ya da hayvan
71	0,012	bir çeşit hastalık

70	0,012	bir çeşit oyun
67	0,166	yetişen bir çeşit
64	0,023	için kullanılan bir
63	0,377	insan ya da
62	0,084	yarayan bir çeşit
61	0,330	oynanan bir çeşit
58	0,906	iğneciler Mudurnu Bo
55	0,009	bir çeşit buğday
54	0,033	çeşit çocuk oyunu
51	0,212	benzeyen bir çeşit
49	0,008	bir çeşit araç
49	0,778	îrişli Bayburt Sarıkamış
46	0,015	Bir çeşit ot
46	0,269	taneli bir çeşit
44	0,154	denilen bir çeşit
43	0,219	pişirilen bir çeşit
42	0,007	bir çeşit yiyecek
41	0,007	bir çeşit armut
41	0,013	Bir çeşit çocuk
41	0,222	oynanan bir çocuk
40	0,007	bir çeşit balık
40	0,007	bir çocuk oyunu
39	0,007	bir çeşit mantar
39	0,021	da denilen bir
38	0,200	her hangi bir
37	0,006	bir çeşit böcek
36	0,563	Bayburt Sarıkamış Kr

36	0,243	kokulu bir çeşit
36	0,188	yapraklı bir çeşit
36	0,391	yemeği yapılan bir
34	0,309	yenilen bir çeşit
33	0,075	olarak kullanılan bir
33	0,020	yapılan bir yemek
32	0,005	bir çeşit elma
32	0,011	Bir çeşit kilim
32	0,337	yenen bir çeşit
31	0,027	kullanılan bir araç
31	0,157	Mısır unundan yapılan
30	0,062	ağaç ya da
29	0,005	bir çeşit börek
29	0,005	bir çeşit çocuk
29	0,044	ile yapılan bir
29	0,165	tatlı bir çeşit
29	0,015	ya da ağaç
29	0,056	yapılmış bir çeşit
28	0,009	Bir çeşit üzüm
28	0,176	görülen bir çeşit
28	0,583	Şaşma bildirir ünlem
27	0,009	Bir çeşit kuş
27	0,184	güzel kokulu bir
26	0,295	kabuklu bir çeşit
26	0,040	küçük bir çeşit
26	0,148	tatlı su balığı
26	0,014	ya da bir

26	0,035	yarayan bir araç
25	0,004	bir çeşit yaban
25	0,025	olan bir hastalık
25	0,439	oynadıkları bir çeşit
25	0,121	taş ya da
25	0,013	ya da keçi
24	0,004	bir çeşit ekmek
24	0,130	oynanan bir oyun
23	0,004	bir çeşit sinek
23	0,004	bir çeşit çorba
23	0,085	bulunan bir çeşit
23	0,141	de denilen bir
23	0,008	için kullanılan ucu
23	0,489	Çelik çomak oyununda
22	0,007	Bir çeşit balık
22	0,007	Bir çeşit bitki
22	0,004	bir çeşit çiçek
22	0,138	görülen bir hastalık
22	0,116	koyun ya da
22	0,012	ya da başka
22	0,030	yarayan tahta araç
21	0,004	bir çeşit yabanıl
21	0,162	dikenli bir ot
21	0,162	dikenli bir çeşit
21	0,042	sonra geriye kalan
21	0,280	yapmakta kullanılan bir
21	0,012	yapılan bir bitki

21	0,226	çiçek açan bir
20	0,007	Bir çeşit armut
20	0,092	biten bir çeşit
20	0,370	boz renkli bir
20	0,120	hangi bir şeyi
20	0,120	hangi bir şeyin
20	0,105	Koyun keçi vb
20	0,247	ununda n yapılan bir
20	0,011	ya da demir
20	0,073	Yeni yürümeye başlayan
20	0,206	çiçekli bir çeşit
20	0,800	încekum Silifke İç
19	0,297	Bayburt Sarıkamış Selim
19	0,003	bir çeşit erik
19	0,114	insan ve hayvan
19	0,007	için açılan çukur
19	0,050	uzun bir çeşit
19	0,184	veren bir çeşit
19	0,012	çeşit tatlı su
18	0,006	Bir çeşit mantar
18	0,006	için kullanılan ağaç
18	0,006	için kullanılan ünlem
18	0,038	konulan bir çeşit
18	0,164	Köpek çağırma ünlemi
18	0,360	zarar veren bir
18	0,065	çıkan bir çeşit
17	0,347	Alay etmek eğlenmek

17	0,003	bir çeşit beyaz
17	0,006	Bir çeşit buğday
17	0,003	bir çeşit kavun
17	0,006	Bir çocuk oyunu
17	0,810	birleşmelere aracılık eden
17	0,041	büyük bir çeşit
17	0,112	büyüklüğünde bir çeşit
17	0,266	dökülerek yapılan bir
17	0,026	küçük bir kuş
17	0,680	Sarıkamış Selim Kr
17	0,089	yapraklı bir ot
17	0,680	Yolsuz birleşmelere aracılık
17	0,018	Çok konuşan geveze
17	0,304	çıkardığı ses için
16	0,168	ağacı ve meyvesi
16	0,320	beliren bir hastalık
16	0,003	bir halk oyunu
16	0,005	Bir çeşit küçük
16	0,003	bir çeşit küçük
16	0,096	demir ya da
16	0,017	gibi ses çıkarmak
16	0,043	hayvan ya da
16	0,006	için kullanılan tahta
16	0,078	Kısa boylu kimse
16	0,016	olan bir bitki
16	0,036	renkli bir kuş
16	0,410	Sulak yerlerde yetişen

DS anlam tablosunda toplam 218.333 adet 3 birimli n-gram yer almış, bu n-gramların tekil olarak sayısı da 215.055 olarak hesaplanmıştır. 3'lü birimlerde sıklığı en yüksek birim 450 sıklık sayısıyla *yapılan bir çeşit* birimidir. 2'li birimlerde geçen ***bir çeşit*** sözü, *yapılan* sözüyle eşdizimlilik kurarak tanımlarda anlatılan nesnenin yapmak kavramıyla ilişkili olduğunu göstermektedir. bir çeşit ot (315), bir çeşit bitki 173), *bir çeşit yemek* (148), *bir çeşit üzüm* (131), *kullanılan bir çeşit* (129) gibi içinde *bir çeşit* geçen ve 2'li n-gramlarda görülmeyen diğer birimler bu tabloda ortaya çıkarak, *bir çeşit*'in aslında *yapılan* bir tür nesne için kullanıldığını ortaya koymaktadır. Anlam tablosunun anahtar söz çözümünde görülebileceği gibi 3'li birimlerde sık geçen *bir çeşit ot, bir çeşit bitki* eşdizimlilikleri, DS'de bitki kavramıyla ilgili sözlerin daha çok derlendiğini ortaya koymaktadır.

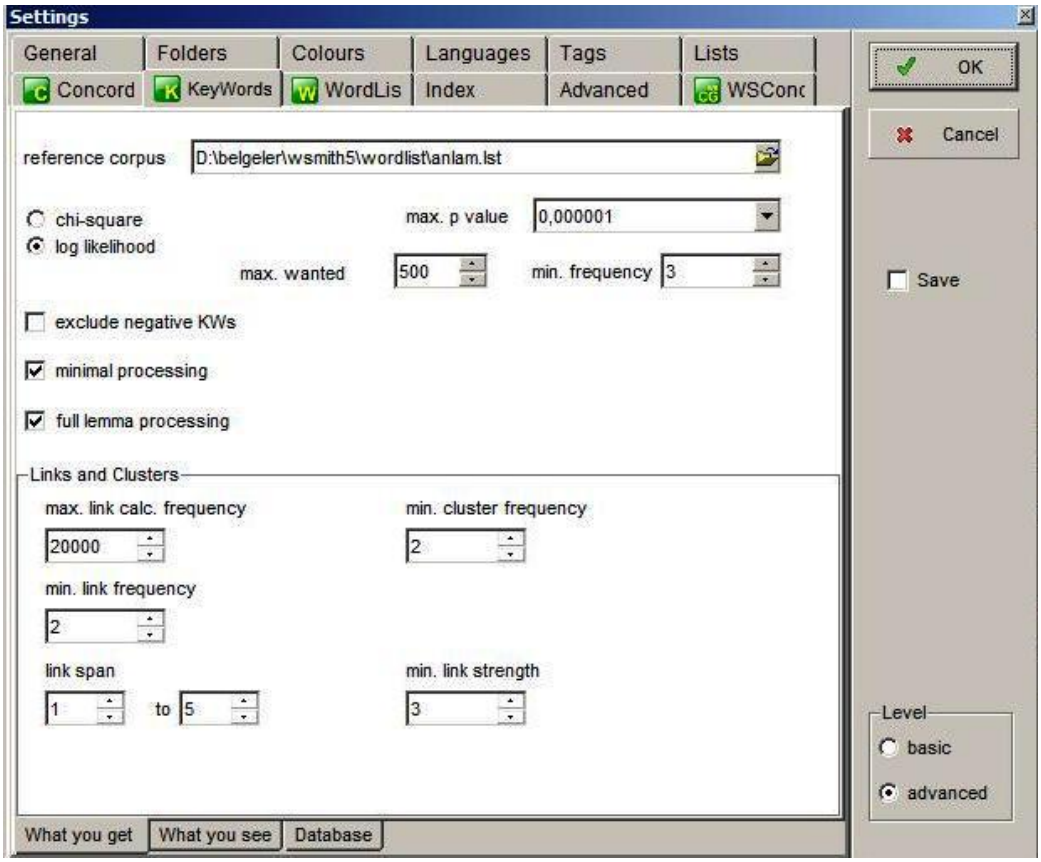
4.2. Anlam Tablosunda Anahtar Sözcük Görünümleri

Anahtar sözcük (keyword) kavramını Vardar ve arkadaşları şu biçimde tanımlamışlardır: "Bir döneme ya da toplumsal çevreye ayırt edici özelliğini veren kavram ya da ülküleri dile getiren sözcüklerden her biri". Bu tanıma dayalı olarak belirli yıllar arasında ortaya çıkan ve gündemi belirleyen olaylara ait sözcükler, olaylarla ilişkili olarak olayların süresine göre ilgili oldukları toplumun düşünüş içeriğini incelemek açısından birtakım çalışmalar yapılabilmektedir. Bu türden çalışmaların günümüz giderek artan internet metinleri aracılığıyla eğilim çıkarma çalışmalarını Google arama motoru Google insight ve Google trends ile belirli zaman aralıkları için yapabilmektedir.

Anahtar sözcük kavramını, genel olarak bir düşünce dünyasını, olaylar yığınını temsil edecek nitelikte ortaya çıkan sözler biçiminde de anlamak mümkündür. Derlem dil biliminde bu kavramı Baker ve arkadaşları istatistik tabanlı tekniklerin ortaya çıkardığı, bir metinde ya da derlemde istatistiksel olarak anlamlı birimler olarak tanımlarlarken, bu tekniğin iki söz listesinden karşılaştırmalı bir yönteme dayandığını belirtmektedirler. Burada kullanılan iki yöntemden log-likelihood, iki veri setinden gözlemlenen ve beklenen değerleri logaritmik bir takım hesaplara dayanarak çıkarırken sıklık değerleri mutlaka göz önünde bulundurulur (Baker vd., 2006: 97, 109).

DS'de anlam tablosunda anahtar sözcüklerin çıkarımı için derlem dil bilimi alanında sıkça kullanılan ve güvenirliği kanıtlanmış bir yazılım olan WordSmith Tool kullanılmıştır. Anhatar sözleri belirlemek için iki söz listesi belirlenmiştir. Bunlardan

birincisi TDK Yazım Kılavuzunda verilen sözler listesi, ikincisi de anlam tablosundan elde edilen listedir. Bu noktada referans derlem olarak anlam tablosu söz listesi belirlenmiş, Yazım Kılavuzunda geçen sözlerin anlam tablosundaki sözlerle karşılaştırılması ve anlam tablosundaki sözlerin sıklıkları ve anahtar olma durumunun anlamlılığını ölçen pozitif ve negatif anahtar olma durumları ortaya çıkarılmıştır. Yazılımda anlamlı olma ölçütü p 0,000001 olarak temel alınmış, sıklığı 3'ten fazla sözler hesaplanmış, teknik olarak da log-likelihood tekniği seçilmiştir.



Şekil 69. WordSmith Tool Anahtar Sözcük Hesaplama Ayarları.

Sonuç olarak, aşağıda bir bölümü verilen 379 adet sözün anahtar sözcük olma durumları hesaplanmıştır.

Tablo 16. Derleme Sözlüğü Anlam Tablosu Anahtar Sözcük Listesi

Id	Anahtar Söz	Referans Liste Sıklığı	Anahtar Olma
1	ve	5488	-1.834,58
2	bir	7449	-1.816,11

3	çeşit	4171	-1.415,85
4	için	4061	-1.377,48
5	ya	1911	-621,93
6	ağaç	1456	-335,34
7	ot	1011	-321,34
8	yer	1582	-238,05
9	hayvan	872	-235,54
10	küçük	1105	-209,8
11	toprak	842	-187,86
12	yemek	787	-179,34
13	çocuk	826	-174,9
14	sonra	552	-165,4
15	araç	584	-163,52
16	renkli	546	-145,44
17	ince	676	-137,55
18	bitki	615	-134,77
19	çok	930	-128,6
20	parçası	417	-108,85
21	tahta	496	-102,69
22	gelen	377	-101,17
23	kadın	539	-97,54
24	yere	424	-96,92
25	büyük	659	-94,93
26	tarla	382	-88,22
27	üzüm	518	-85,34
28	keçi	409	-84,27
29	buğday	442	-83,15

30	Mısır	313	-75,58
31	kadar	354	-75,47
32	iki	659	-71,29
33	hastalık	324	-70,36
34	erkek	370	-69,19
35	parça	331	-68,54
36	koyun	354	-68,02
37	az	341	-67,78
38	iyi	394	-66,28
39	iri	302	-63,69
40	ip	286	-62,83
41	ekin	253	-61,28
42	tahıl	235	-60,39
43	Yeni	291	-60,38
44	uzun	527	-55,31
45	kabuğu	218	-54,94
46	yavaş	221	-51,24
47	yapraklı	205	-50,79
48	kap	230	-49,81
49	hastalığı	343	-49,64
50	sert	275	-48,68
51	ucu	260	-47,7
52	at	290	-46,68
53	ekmek	430	-45,35
54	önce	201	-45,05
55	kuş	414	-43,73
56	mısır	257	-43,63

57	kalın	277	-43,16
58	demir	442	-42,12
59	yuvarlak	270	-41,28
60	bez	213	-40,98
61	çabuk	185	-40,15
62	ufak	208	-39,52
63	armut	181	-38,94
64	oyunu	401	-38,8
65	kötü	226	-38,1
66	insan	287	-37,81
67	yemeği	224	-37,55
68	zayıf	176	-37,42
69	cins	199	-36,92
70	yerde	160	-36,64
71	siyah	186	-36,63
72	boylu	238	-35,54
73	güzel	226	-35,14
74	sık	165	-34,11
75	çivi	162	-33,22
76	beyaz	328	-32,25
77	odun	215	-32,2
78	kabı	145	-32,02
79	taş	502	-31,31
80	iş	454	-31,05
81	her	227	-30,08
82	yoğurt	173	-29,55
83	düzgün	136	-29,28

84	kokulu	158	-28,52
85	işi	302	-28,42
86	tarafı	133	-28,37
87	tatlı	306	-27,43
88	adam	370	-27,07
89	saç	141	-27,02
90	ceviz	139	-26,44
91	pekmez	137	-25,86
92	sulu	170	-25,78
93	birlikte	124	-25,66
94	oyun	254	-25,49
95	kız	262	-25,32
96	daha	157	-25,14
97	hamur	167	-24,98
98	meyve	327	-24,93
99	bölümü	120	-24,46
100	un	153	-24,05
101	ne	142	-24,02
102	köpek	193	-24,02
103	su	1009	-23,98
104	dikenli	163	-23,93
105	sahibi	21	23,95
106	ön	111	24,08
107	anahtarı	7	24,29
108	kitap	7	24,29
109	odası	30	24,4
110	damak	4	24,4

111	dönemi	4	24,4
112	maden	14	24,63
113	sağ	14	24,63
114	sinir	20	24,97
115	kaplama	5	25,23
116	bayram	8	25,37
117	mahalle	8	25,37
118	adalet	0	25,72
119	askerlik	0	25,72
120	astsubay	0	25,72
121	cambazı	0	25,72
122	cari	0	25,72
123	efektif	0	25,72
124	eğrisi	0	25,72
125	elektron	0	25,72
126	fizik	0	25,72
127	geçiş	0	25,72
128	göstergesi	0	25,72
129	iftar	0	25,72
130	kesir	0	25,72
131	Latin	0	25,72
132	lop	0	25,72
133	lüks	0	25,72
134	mahkemesi	0	25,72
135	mektubu	0	25,72
136	muhabbet	0	25,72
137	oyuncusu	0	25,72

138	ötesi	0	25,72
139	özne	0	25,72
140	paralık	0	25,72
141	potasyum	0	25,72
142	savaşı	0	25,72
143	sinema	0	25,72
144	takvimi	0	25,72
145	teminat	0	25,72
146	trafik	0	25,72
147	türeme	0	25,72
148	uzmanı	0	25,72
149	vazife	0	25,72
150	yad	0	25,72

Tabloda referans sıklığı olarak verilen sıklıklar, Yazım Kılavuzundaki sözlerin anlam tablosundaki sıklıklarını göstermektedir. Anahtar olma sütununda yer alan ve (-) işaretiyle verilen negatif işareti, sözün şans eseri olmayacak biçimde anlam tablosunda bulunduğunu göstermektedir. Kısaca (–) işaretliler sözün anahtar olma durumunun anlamlı olduğunu göstermektedir. Bu listeden; “ve”, “bir”, “çeşit”, “için” sözlerinin bir nesneyi ya da durumu tanımlamak için sıkça kullanıldığı sonucuna varılabilir. İçerik olarak yüklü sözlerden “ot”, “ağaç”, “yer”, “hayvan”, “toprak”, “bitki” sözlerinin kırsal yaşamı yansıttığı görülmektedir. “koyun” ve “keçi”nin de hayvan adları arasında anahtar sözcükler olduğu da ortaya çıkmaktadır.

4.3. Anlam Tablosu Söz Sıklıkları

Anlam tablosunda yer alan sözlerin sıklıkları çıkarılmış, toplam kullanılan ayrı söz sayısı 53.758 olarak belirlenmiştir. Aşağıdaki tabloda sıklığı en yüksek 150 söz verilmiştir. Listenin hesaplanmasında sözlükte madde başı olarak bulunabilecek sözler yanında, sözcüklerin çekimli biçimleri olan sözcük biçimler de listede yer almaktadır. Kullanılan yazılım, “Küçük” ve “küçük” örneğinde olduğu gibi büyük-küçük harflere

duyarlı olarak hesaplama yapmıştır Büyük harfli sözler anlam tablosunda tanımlamanın başında kullanılan sözleri göstermektedir.

Tablo 17. Derleme Sözlüğü Anlam Tablosu Sıklık Listesi

No	Söz	Sıklık	Yüzde
1	bir	7449	1,80
2	ve	5488	1,33
3	çeşit	4171	1,01
4	için	4061	0,98
5	Bir	3123	0,76
6	yapılan	2512	0,61
7	da	2101	0,51
8	kimse	2069	0,50
9	ya	1911	0,46
10	olan	1593	0,39
11	yer	1582	0,38
12	kullanılan	1458	0,35
13	ağaç	1456	0,35
14	Çok	1311	0,32
15	gibi	1174	0,28
16	vb	1170	0,28
17	etmek	1154	0,28
18	Küçük	1150	0,28
19	küçük	1105	0,27
20	ot	1011	0,24
21	su	1009	0,24
22	olmak	986	0,24
23	yarayan	974	0,24
24	çok	930	0,23

25	hayvan	872	0,21
26	toprak	842	0,20
27	çocuk	826	0,20
28	yemek	787	0,19
29	ile	763	0,18
30	yapılmış	744	0,18
31	yapmak	721	0,17
32	ince	676	0,16
33	büyük	659	0,16
34	iki	659	0,16
35	konulan	643	0,16
36	bitki	615	0,15
37	araç	584	0,14
38	Büyük	573	0,14
39	olarak	559	0,14
40	sonra	552	0,13
41	Su	549	0,13
42	renkli	546	0,13
43	kadın	539	0,13
44	yavrusu	534	0,13
45	uzun	527	0,13
46	üzüm	518	0,13
47	veya	515	0,12
48	taş	502	0,12
49	tahta	496	0,12
50	gelmek	462	0,11
51	iş	454	0,11

52	şey	453	0,11
53	ses	452	0,11
54	buğday	442	0,11
55	demir	442	0,11
56	yetişen	437	0,11
57	ekmek	430	0,10
58	yere	424	0,10
59	parçası	417	0,10
60	kuş	414	0,10
61	takılan	414	0,10
62	keçi	409	0,10
63	vermek	408	0,10
64	oyunu	401	0,10
65	İki	396	0,10
66	iyi	394	0,10
67	anlamında	392	0,09
68	çıkan	388	0,09
69	ünlemi	384	0,09
70	tarla	382	0,09
71	Hayvan	380	0,09
72	olmayan	379	0,09
73	gelen	377	0,09
74	yeri	375	0,09
75	söz	374	0,09
76	adam	370	0,09
77	erkek	370	0,09
78	şeyi	370	0,09

79	çıkarmak	365	0,09
80	kadar	354	0,09
81	koyun	354	0,09
82	kullanılır	350	0,08
83	kalan	349	0,08
84	bulunan	345	0,08
85	hastalığı	343	0,08
86	az	341	0,08
87	üstüne	337	0,08
88	Çocuk	331	0,08
89	parça	331	0,08
90	denilen	330	0,08
91	İnce	329	0,08
92	beyaz	328	0,08
93	meyve	327	0,08
94	zaman	326	0,08
95	hastalık	324	0,08
96	üzerine	323	0,08
97	biçiminde	322	0,08
98	Bu	322	0,08
99	Mısır	313	0,08
100	benzer	311	0,08
101	tatlı	306	0,07
102	verilen	304	0,07
103	iri	302	0,07
104	işi	302	0,07
105	almak	298	0,07

106	ağacı	295	0,07
107	Ağaç	294	0,07
108	ist	294	0,07
109	Yeni	291	0,07
110	at	290	0,07
111	insan	287	0,07
112	ip	286	0,07
113	Her	282	0,07
114	kalın	277	0,07
115	sert	275	0,07
116	Az	274	0,07
117	yuvarlak	270	0,07
118	yapan	265	0,06
119	anlatır	263	0,06
120	kız	262	0,06
121	Kısa	260	0,06
122	ucu	260	0,06
123	gitmek	259	0,06
124	yol	259	0,06
125	kalmak	258	0,06
126	mısır	257	0,06
127	oyun	254	0,06
128	ekin	253	0,06
129	konuşmak	251	0,06
130	tutmak	251	0,06
131	benzeyen	248	0,06
132	içine	248	0,06

133	Toprak	248	0,06
134	içinde	246	0,06
135	yüz	245	0,06
136	kısmı	243	0,06
137	pişirilen	240	0,06
138	boylu	238	0,06
139	konuşan	238	0,06
140	açılan	237	0,06
141	At	236	0,06
142	Kötü	236	0,06
143	tahıl	235	0,06
144	Uzun	235	0,06
145	birbirine	232	0,06
146	kap	230	0,06
147	para	229	0,06
148	her	227	0,05
149	İri	227	0,05
150	kırmızı	227	0,05

Listede, bitki ve hayvan adları yanında “su”, “toprak”, “çocuk” kavramlarının da ilk 150 söz içinde dikkati çektiği görülmektedir. “Bir”, “bir”, “Çok”, “çok”, “küçük”, “Küçük” gibi sözcüklerin ayrı birimler olarak hesaplanmaları, büyük harflerle başlayan sözcüklerin tanımlamalarda geçen ilk sözcükler olmasından kaynaklanmaktadır.

SONUÇ

Ek ciltleriyle birlikte toplam 12 ciltlik Derleme Sözlüğü'nün 124.692 adet madde başı sözünün veri tabanının hazırlandığı bu çalışma ile Anadolu ağızlarının çeşitli görünümleri elektronik ortamda sorgulanabilir bir biçime dönüştürülmüştür. Çalışma, Anadolu ağızlarının değişke, tanık, madde başı ve özellikle anlam özelliklerindeki bilgiye erişimi kolaylaştırmakla birlikte gelişmiş sorgu formlarıyla daha derin örüntülere ulaşmayı sağlayabilmektedir.

Türk dili ve edebiyatı bölümlerinin Anadolu ve Rumeli ağızlarıyla ilgili çalışmalarına katkı sağlayacağı düşünülen çalışmanın aynı zamanda DDİ araştırmaları için de veri sağlayacağı açıktır. 12 ciltlik, toplam 4402 sayfalık sözlük metninden elde edilen listeler olası bir Anadolu ağızları atlasının görsel olarak hazırlanması çalışmaları için istatistiksel çıkarımlarıyla ve bölgelere özel anlam örüntülerinin, değişkelerinin sorgulanabilirliğiyle çalışmamız hazır veriler sağlamaktadır.

Anlam tablosundan çıkarılan söz listeleri ve özellikle eşdizim örüntüleri Anadolu ağızlarından derlenen malzemenin nesnelere dayandığını ortaya çıkmaktadır. "Bir çeşit", "bir tür", "yapılan bir çeşit" gibi eşdizim örüntüsünü oluşturan birimlerin sıklığının yüksek olması bu birimlerin nesne tanımlama birimleri olarak yorumlanmasını sağlamaktadır.

Dil birimlerinin bir arada bulunabilirliğinin ölçütü olarak sıklığın aynı zamanda bir vurgu ve önem belirtmesi, sıklığın, metin yoğunluklu ve yüz binlerce sözcüklük yığınların yorumlanması aşamalarında da temel alınan bir ölçüt olmasında etkilidir. Sıklık, istatistik ve olasılık kuramlarının basit bir dağılım biriminin adı olması yanında, sıklık ölçütünün olmadığı bir olasılık hesaplamasını düşünmek de zordur. Bu bakımdan bu çalışmada özellikle anlam tablosundan elde edilen veriler sıklık temelli olarak yorumlanmıştır.

Bu çalışmanın, edebi türler üzerine yapılan ve biçem (üslup) çalışmalarından ayrı olarak bir sözlük metninin elektronik ortamda sorgulanabilir ve bağlantılı bir yapıya kavuşturulmasının dışında, sözlük bilimi araştırmalarında ve sözlük yazımında, çevrim içi, paylaşımlı sözlüklerin hazırlanması aşamalarında da veri sağlayabileceği düşünülmektedir.

Çalışmanın kuramsal bölümünü oluşturan derlem dil bilimi, DDİ tanımı ve uygulama alanları, DDİ çalışmalarının tanıtılması bakımından olduğu kadar dil

araştırmalarında bilgisayar ve bilişim teknolojilerinin her geçen gün artan yazılım ve donanım gücüyle yakın gelecekte yorumlamaya giden sürecin en önemli parçaları olacağını göstermesi açısından önemlidir. Sözlük biliminin internet tabanlı otomatik metin işleme yöntemlerinden etkilenmesi ve sözlük yazımından yayınlanmasına kadar olan bütün süreçlerde yeni uygulamalarla zenginleşeceği de açıktır.

KAYNAKÇA

AKALIN, Şükrü Halûk, "Sözcük Bilimi ve Sözlükçülük", *Türk Dili*, C. XCVIII, S.698, TDK Yayını, Şubat, 2010.

AKSAN, Doğan (1998), *Her Yönüyle Dil Ana Çizgileriyle Dilbilim* , TDK Yayını, Ankara

----------(2004), "Türklerde Sözlükçülük, Bugün Türkiye'de Sözlük", *Dilbilim ve Türkçe Yazıları*, Multilingual Yay., Matbaa 70, İstanbul.

BAKER, Paul, Andrew Hardie, Tony McEnery (2006), A Glossary of Corpus Linguistics, Edinburgh University Pres Ltd., Edinburgh.

BAYRAV, Süheylâ (1998), *Filolojinin Oluşumu*, Multilingual Yay., Matbaa 70, İstanbul.

BEESLEY, Kenneth R. (2003), "Computational Morphology and Finite-State Methods", *Language Engineering for Lesser-studied Languages*, (Editör: Kemal Oflazer), NATO Science Series, Series III, Vol.188, IOS Press. Netherlands.

BOD, Rens, Jennifer Hay, Stefanie Jannedy (2003), "Introduction", *Probabilistic Linguistics*, Massachusetts Institute of Technology.

BOESCHOTEN, Hendrik, Hansje Braam (1994), "Bilgisayar ve Sözlükçülük Yöntemleri", *Uygulamalı Dilbilim Açısından Türkçenin Görünümü*, Dil Derneği Yay.:5, s.9

BOOGARDS, Paul (2003), "Uses and Users of Dictionaries", *A Practical Guide to Lexicography*, (Editör: Piet van Sterkenburg), John Benjamins Publishing Company, Amsterdam.

BOWKER, Lynne (2003), "Specialized Lexicography and Specialized Dictionaries", *A Practical Guide to Lexicography*, (Editör: Piet van Sterkenburg), John Benjamins Publishing Company, Amsterdam.

BOZ, Erdoğan (2006), "Sözlük ve Sözlükçülük Sorunu", Türkçenin Çağdaş Sorunları (Editörler: Gürer Gülsevin, Erdoğan Boz), 2. baskı, Gazi Kitabevi, Ankara.

BURCH, Chris Callison, Miles Osborne (2003), "Statistical Natural Language Processing", *Handbook for Language Engineers*, (Editör: Ali Farghaly), CSLI Publications, California.

BUTT, Miriam, Tracy Holloway King (2003), "Grammar Writing, Testing, and Evaluation", *Handbook for Language Engineers*, (Editör: Ali Farghaly), CSLI Publications, California.

CABRÉ, M. Teresa (1999), *Terminology*, John Benjamins Publishing Company.

CAN, Fazlı, Seyit Koçberber, Erman Balçık, Cihan Kaynak, H. Çağdaş Öcalan, Onur M. Vursavaş (2008), "Information Retrieval on Turkish Texts", *Journal Of The American Society For Information Science And Technology*, Volume 59, Issue 3, February.

CASTELLVİ, M. TeresaCabré, Anne Condamines, Fidelia Ibekwe-SanJuan (2007), "Introduction- Application-driven terminology engineering", *Application-Driven Terminology Engineering*, John Benjamins Publishing Company.

CEBİROĞLU, Gülşen, Eşref Adalı (2002), "Sözlüksüz Köke Ulaşma Yöntemi", *19. TBD Bilişim Kurultayı*, http://www3.itu.edu.tr/~gulsenc/papers/TBD_Gulsen_Makale.pdf

ČERMÁK, František (2003), "Source Materials for Dictionaries", *A Practical Guide to Lexicography*, (Editör: Piet van Sterkenburg), John Benjamins Publishing Company, Amsterdam.

CHIRIST, Oliver (1998), "Linking WORDNET to a Corpus Query System", *Linguistic Databases*, Center for the Study of Language and Inf.

COLEMAN, John (2005), Introducing Speech and Language Processing, Cambridge University Press, Cambridge.

COWIE, Jim, Yorick Wilks (2000), "Information Extraction", *Handbook of Natural Language Processing*, (Editörler: Robert Dale, Hermann Moisl, Harold Somers), Marcel Dekker Incorporated.

ÇAYCI, Özgür (2003), *PHP ve MySQL*, Seçkin Yayıncılık, İstanbul.

ÇEBİ, Yalçın- Özlem Varlıklar (2006), "Türkçe Derlem Oluşturmada Karşılaşılan Sorunlar ve Çözüm Önerileri", *Bilgisayar Destekli Dil Bilimi Çalıştayı Bildirileri*, TDK Yay: 868, Ankara

ÇOBANOĞLU, Bülent (2007), *Algoritma Geliştirme ve Veri Yapıları*, Pusula Yayınları, İstanbul.

DAŞDEMİR, yaşar (2004), *Veritabanı Yönetim Sistemleri*, Türkmen Kitabevi, İstanbul.

DERLEME SÖZLÜĞÜ (1982), C. I-XII, TDK Yay.:211/12, Ankara

DEUTSCH Werner A., Ralf Vollmann, Anton Noll, Sylvia Moosmüller (1998), "An Open Systems Approach for an Acoustic-Phonetic Continuous Speech

Database: The S_Tools Database-Management System (STDBMS)", *Linguistic Databases*, Center for the Study of Language and Inf.

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

ERKAN, Güneş, Dragomir R. Radev (2004), "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization", *Journal of Artificial Intelligence Research*, 22, AI Access Foundation.

ERYİĞİT, Gülşen, Eşref Adalı, Kemal Oflazer (2008), "Türkçe'nin Olasılık Tabanlı Bağlılık Ayrıştırması", *İTÜ Dergisi /d Mühendislik*, C.:7, S. 4.

GEERAERTS, Dirk (2003), "Meaning and Definition", *A Practical Guide to Lexicography*, (Editör: Piet van Sterkenburg), John Benjamins Publishing Company.

GILMORE, W. Jason (2006), *Beginning PHP and MySQL 5: From Novice to Professional*, Apress

GUTHRIE, M. Lousie-Brian M. Slator ve Yorick A. Wilks (1996), *Electric Words, Dictionaries, Computers and Meanings*, The MIT Pres, London

GÜRSAKAL, Necmi (2008), *Betimsel İstatistik*, 4. baskı, Dora Yay. No: 5, Bursa.

HANKS, Patrick (2003), "Lexicography", *The Handbook of Computational Lingusitics*, (Editör: Ruslan Mitkov), Oxford University Pres, Oxford.

HARTMANN, R.R.K-Gregory James (1998), *Dictionary of Lexicography*, Routledge, London.

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

HUNDT, Marianne, Nadja Nesselhauf , Carolin Biewer (2006), *Corpus Linguistics and the Web***,** Editions Rodopi BV.

ILSON, Robert F. (2002), "Lexicography", *The Linguistics Encyclopedia*, (Editör: Kirsten Malmkjaer), 2. basım, Routledge, Florence, Ky, USA.

JACKSON, Howard (2002), *Lexicography-An Introduction*, Routledge, USA.

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

KARAOĞLAN, Bahar, Müge Sayıt (2007), "Soru Cevaplama Üzerine Bir Araştırma", Akademik Bilişim 2007, 31 Ocak – 2 Şubat 2007, Kütahya.

KARTTUNEN, Lauri (2003), "Finite-State Technology", *The Oxford Handbook of Computational Linguistics*, (Editör: Ruslan Mitkov), Oxford University Press, USA.

KAYA, Yılmaz, Ramazan Tekin (2007), *Veritabanı ve Uygulamaları*, 1. basım, Papatya Yayınları.

KENNEDY, Graeme (1998), *An Introduction to Corpus Linguistics*, Addison Wesley Longman Limited, New York.

KIRKNESS, Alan (2004), "Lexicography", *The Handbook of Applied Linguistics*, (Editörler: Alan Davies, Catherine Elder), Blackwell Publishing LTD.

KOCAMAN, Ahmet (1998), "Dilbilim, Sözlük, Sözlükçülük", *Kebikeç*, Yıl: 3, S. 6., Ankara.

KRUYT, Truus (2003), "Multifunctional Linguistic Databases: Their Multiple Use", (Editör: Piet van Sterkenburg), *A Practical Guide to Lexicography*, John Benjamins Publishing Company.

KUTLU, Mücahid, Celal Çığır, İlyas Çiçekli (2010), "Generic Text Summarization for Turkish, *The Computer Journal*, January 6, Oxford University Pres.

KÜÇÜK, Dilek, Adnan Yazıcı (2009), "Named Entity Recognition Experiments on Turkish Texts", *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg.

MANNING, Christopher, Hinrich Schütze (1999), *Foundations of Statistical Natural Language Processing*, The MIT Press.

----------Christopher, Prabhakar Raghavan, Hinrich Schütze (2008), *Introduction to Information Retrieval*, Cambridge University Press

MCCARTHY, Diana-Geoffry Sampson (Haz.) (2005), *Corpus Linguistics: Readings in A Widening Discipline*, Continuum, New York.

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

----------(2006), *Corpus-Based Language Studies: An Advanced Resource Book*, Routledge yayınları

NERBONNE, John (1998), *Linguistic Databases*, Center for the Study of Language and Inf.

----------(1998), "Introduction", *Linguistic Databases*, Center for the Study of Language and Inf.

OEPEN, Stephan, Klaus Netter, Judith Klein (1998), "TSNLP Test Suites for Natural Language Processing", *Linguistic Databases*, Center for the Study of Language and Inf.

OFLAZER, Kemal (1994), "Two-level Description of Turkish Morphology", Literary and Linguistic Computing Vol. 9, No:2, Oxford.

----------(2003), *Language Engineering of Lesser-Studied Languages*, (Nato Science Series, Series III : Computer and Systems Science-Vol 188), IOS Press.

----------(2006), "Türkçe Doğal Dil İşleme", Türkiye Bilişim Ansiklopedisi, Papatya Yayıncılık, s. 814.

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

OOI, Vincent (1998), *Computer Corpus Lexicography*, Edinburgh University Pres, Edinburgh.

ÖLMEZ, Mehmet (1994), "Türk Dillerinin Sözlükleri ve Türk Sözlükçülüğü, *Uygulamalı Dilbilim Açısından Türkçenin Görünümü*, Dil Derneği Yay.: 5, 1. baskı, Ankara

ÖZKAN, Yalçın (2006), *Veri Tabanı Sistemleri*, Alfa Yayınları, İstanbul.

ROARK, Brian, Richard Sproat (2007), *Computational Approaches to Morphology and Syntax*, Oxford University Press, New York

SAY, Bilge (2006), "Türkçe İçin Bir Derlem Geliştirme Çalışması", *Bilgisayar Destekli Dil Bilimi Çalıştayı Bildirileri*, TDK Yay: 868, Ankara

SAYIT, Müge, Bahar Karaoğlan (2007), "Soru Cevaplama Üzerine Bir araştırma", Akademik Bilişim 2007, 31 Ocak-2 Şubat 2007, Kütahya, http://ab.org.tr/ab07/bildiri/88.pdf

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

SIMONS, Gary F., John V. Thomson (1998), "Multilingual Data Processing in the CELLAR Environment", *Linguistic Databases*, Center for the Study of Language and Inf.

SINCLAIR, John (2003), "Corpora for Lexicography", *A Practical Guide to Lexicography*, (Editör: Piet van Sterkenburg), John Benjamins Publishing Company, Amsterdam.

SMADJA, Frank (1993), "*Retrieving Collocations from Text: Xtract*", Computational

Linguistics, Vol. 19, Association for Computational Linguistics.

STERKENBURG, Piet van (2003), *A Practical Guide to Lexicography*, (Editör: Piet van Sterkenburg), John Benjamins Publishing Company, Amsterdam.

STEVENSON, Mark, Yorick Wilks (2003), "Word-Sense Disambiguation", *The Handbook of Computational Lingusitics*, (Editör: Ruslan Mitkov), Oxford University Pres, Oxford.

ŞAMLI, Mehmet (2003), *MySQL ve PostgreSQL ile Veritabanı Programlama*, Pusula Yayınları, İstanbul.

----------(2006), *PHP ile Web Programcılığı*, Pusula Yayınları, İstanbul.

TANTUĞ, A. Cüneyd, Eşref Adalı, Kemal Oflazer (2006), "Computer Analysis of the Turkmen Language Morphology", *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg

TUNÇ, Tomris (1995), *Derleme Sözlüğü ve Kavramlar Dizini*, C.I-III, TDK Yay.:627/1-3, Ankara.

YANIKOĞLU, Berrin (2006), "Belge Bölütleme ve Tanıma", *Türkiye Bilişim Ansiklopedisi*, Papatya Yayınları, İstanbul.

YARIMAĞAN, Ünal (2000), *Veri Tabanı Sistemleri*, Akademi Yayın Hizmetleri San. Tic. Ltd. Şti, Ankara.

YILDIRIM, Faruk, Tahir Tahiroğlu (2006), "İnternette Türkçe Kullanımı Sorunu", Türkçenin Çağdaş Sorunları (Editörler: Gürer Gülsevin, Erdoğan Boz), 2. baskı, Gazi Kitabevi, Ankara.

YÖNDEM, Meltem Turhan (2006), "Bilişimsel Dil Bilimin Altyapı Ögeleri", *Bilgisayar Destekli Dil Bilimi Çalıştayı Bildirileri*, TDK Yay.:868, Ankara

YÜKSEKKAYA, Yaşar (2001), "Avrupadaki Modern Sözlük Araştırmaları ve Bu Araştırmaların Türk Metasözlükçülüğüne Yansımaları", *Uluslar arası Sözlükbilim Sempozyumu Bildirileri*, Doğu Akdeniz Üniversitesi Basımevi, Gazimağusa.

UYSAL, Mithat (2001), *SQL Veri Tabanı Sorgulama Dili*, Beta Yayınları, İstanbul.

VARDAR, Berke (1998), *Açıklamalı Dilbilim Terimleri Sözlüğü*, ABC Kitabevi A.Ş.,NovaPrint Basımevi, İstanbul

WERMTER, Stefan, Wendy G. Lehnert (1992), "A Hybrid Symbolic/Connectionist Model for Noun Phrase Understanding", *Connectionist Natural Language Processing*, (Editör: Noel Sharkey), Kluwer Academic Publishers.

WILKS, Yorick A., M. Brian Slator, M. Guthrie Louise (1995), Electric Words, Dictionaris, Computers an Meanings, The MIT Pres, London

ZHENG, Zhiping (2002), AnswerBus Question Answering System, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.11.9666&rep=rep1&type=pdf

Erişim :

http://ab.org.tr/ab07/bildiri/88.pdf

http://acl.ldc.upenn.edu/E/E93/E93-1066.pdf

http://books.google.com/books?id=o86MgEz233cC&printsec=frontcover#v=onepage&q&f=false

http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=79B7429763EAD8645508905697892991?doi=10.1.1.11.9666&rep=rep1&type=pdf

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.11.9666&rep=rep1&type=pdf

http://clair.si.umich.edu/clair/lexrank/

http://clair.si.umich.edu/clair/lexrank/

http://en.wikipedia.org/wiki/Global_Language_Monitor

http://iate.europa.eu/iatediff/SearchByQueryLoad.do?method=load

http://lingpipe-demos.com:8080/lingpipe-demos/ne_en_news_muc6/response.xml

http://lingpipe-demos.com:8080/lingpipe-demos/ne_en_news_muc6/textInput.html

http://sozluk.sourtimes.org/

http://tdk.gov.tr/TR/Genel/SozBul.aspx?F6E10F8892433CFFAAF6AA849816B2EF4376734BED947CDE&Kelime=veri

http://tdk.gov.tr/TR/Genel/SozBul.aspx?F6E10F8892433CFFAAF6AA849816B2EF4376734BED947CDE&Kelime=bilgi

http://tdk.gov.tr/TR/Genel/SozBul.aspx?F6E10F8892433CFFAAF6AA849816B2EF4376734BED947CDE&Kelime=veri%20taban%C4%B1

http://tdkterim.gov.tr/bts/

http://tdkterim.gov.tr/bts/?kategori=verilst&kelime=bilgisayar&ayn=tam

http://tdkterim.gov.tr/bts/?kategori=veritbn&kelimesec=288264

http://tdkterim.gov.tr/bts/?kategori=veritbn&kelimesec=73633

http://turkoloji.cu.edu.tr/DILBILIM/anlam_belirsizligi.pdf

http://turkoloji.cu.edu.tr/DILBILIM/anlam_belirsizligi.pdf

http://webascorpus.org/searchwac.html

http://wordnet.princeton.edu/

http://wordnet.princeton.edu/man/lexnames.5WN

http://wordnet.princeton.edu/man/wnstats.7WN

http://wordnet.princeton.edu/papers

http://wordnet.princeton.edu/man/wnstats.7WN

http://wordreference.com/

http://www.aclweb.org/anthology-new/J/J93/J93-1007.pdf

http://www.answerbus.com/index.shtml

http://www.answerbus.com/index.shtml

http://www.ce.yildiz.edu.tr/mygetfile.php?id=2657

http://www.springerlink.com/content/l831084713048233/fulltext.pdf

http://www.ce.yildiz.edu.tr/mygetfile.php?id=2657

http://www.springerlink.com/content/l831084713048233/fulltext.pdf

http://www.collins.co.uk/corpus/CorpusSearch.aspx

http://www.collins.co.uk/Corpus/CorpusSearch.aspx#queries

http://www.collins.co.uk/Corpus/CorpusSearch.aspx#queries

http://www.corpus.bham.ac.uk/research/index.htm

http://www.dicts.info/ud.php?11=turkish&12=turkish

http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/courses/conferences.html

http://www.google.com/insights/search/#

http://www.google.com/insights/search/#q=bilgisayar&cmpt=q

http://www.google.com/intl/en/press/zeitgeist2008/

http://www.google.com/intl/en/press/zeitgeist2008/mind.html

http://www.jair.org/media/1523/live-1523-2354-jair.pdf

http://www.jair.org/media/1523/live-1523-2354-jair.pdf

http://www.languagemonitor.com/

http://www.ldoceonline.com/

http://www.ldoceonline.com/popup/popupmode.html?search_str=electronic

http://www.ldoceonline.com/search/?q=computer

http://dictionary.cambridge.org/search/british/?q=computer&x=0&y=0

http://www.lexicool.com/

http://www.powerhomebiz.com/vol25/invisible.htm

http://www.powerhomebiz.com/vol25/invisible.htm

http://www.seslisozluk.com/

http://www.sil.org/computing/routledge/simons/databases.html

http://www.sil.org/computing/routledge/simons/databases.html

http://www.springerlink.com/content/7qq2565v22064117/

http://www.tdk.gov.tr/TR/dosyagoster.aspx?DIL=1&BELGEANAH=2908&DOSYAISIM=sunum06.ppt

http://www.tdk.gov.tr/TR/dosyagoster.aspx?DIL=1&BELGEANAH=2908&DOSYAISIM=sunum06.ppt

http://www.tdk.gov.tr/TR/Genel/Ana.aspx?F6E10F8892433CFFAAF6AA849816B2EF4376734BED947CDE

http://www.tdk.org.tr/lehceler/

http://www.worldwidewebsize.com/

http://www.worldwidewebsize.com/

http://www3.itu.edu.tr/~gulsenc/papers/TBD_Gulsen_Makale.pdf

http://www3.itu.edu.tr/~gulsenc/papers/TBD_Gulsen_Makale.pdf

ÖZ GEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı: B. Tahir TAHİROĞLU **Doğum**

Yeri ve Yılı: Kadirli, 01.03.1979

Medenî Hâli: Bekâr

Adres : Çukurova Üniversitesi, Fen-Edebiyat Fakültesi, Türk Dili ve Edebiyatı Bölümü, Balcalı/ADANA

Telefon : 0 555 443 36 36

e-posta : btahir@cu.edu.tr

ÖĞRENİM DURUMU

Yüksek Lisans : 2000-2003, Çukurova Üniversitesi, Sosyal Bilimler Enstitüsü, Türk Dili ve Edebiyatı Anabilim Dalı, Balcalı/ADANA **Lisans** : 1995-1999, Çukurova Üniversitesi, Fen Edebiyat Fakültesi, Türk Dili ve Edebiyatı Bölümü, Balcalı/ADANA

Lise : 1991-1994, Borsa Lisesi / ADANA

Orta Okul: 1988-1991, Taşpınar Lisesi / AKSARAY

İlkokul: 1983-1988, Cumhuriyet İlköğretim Okulu, İslahiye / GAZİ ANTEP

İS HAYATI

2002-ÇÜ Türkoloji Araştırma ve Uygulama Merkezi Uzmanı

YABANCI DİL

İngilizce

BİLGİSAYAR

MS Windows 98, MS Windows 2000, MS Windows 2000 Server, MS Windows XP, Linux Mandrake 9.1, MS Word, Excel, Access, PowerPoint, FrontPage; Macromedia Dereamweaver