DESIGN AND IMPLEMENTATION OF THE SEMANTIC TURKISH LANGUAGE AND DIALECTS DICTIONARY

by

Pınar ÖNDER

July 2009

DESIGN AND IMPLEMENTATION OF THE SEMANTIC TURKISH LANGUAGE AND DIALECTS DICTIONARY

by

Pınar ÖNDER

A thesis submitted to

The Graduate Institute of Sciences and Engineering

of

Fatih University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Engineering

July 2009
Istanbul, Turkey

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

____

Assist. Prof. Tuğrul YANIK
Head of Department

This is to certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

____

Assist. Prof. Zeynep ORHAN
Supervisor

Examining Committee Members

Assist. Prof. Zeynep ORHAN ________________________

Assist. Prof. Tuğrul YANIK ________________________

Assist. Prof. Özgür ÖZDEMİR ________________________

It is approved that this thesis has been written in compliance with the formatting rules laid down by the Graduate Institute of Sciences and Engineering.

____

Assoc. Prof. Nurullah ARSLAN
Director

DESIGN AND IMPLEMENTATION OF THE SEMANTIC TURKISH LANGUAGE AND DIALECTS DICTIONARY

Pınar ÖNDER

M. S. Thesis - Computer Engineering
July 2009

Supervisor: Assist. Prof. Zeynep ORHAN

ABSTRACT

Traditional dictionaries provide the word, definition and sometimes example sentences. However, most of the important features, information and relationships for the words are not represented. While it is possible to find applications that have some specific features and relationships of the words for English, it is not possible to see these applications for Turkish language. Therefore, the main idea of this study is to represent the semantic relationships between Turkish words.

In this study, a framework to facilitate comparison among the words and access to these words, semantic information is extracted from the word definitions in a way to render implicit information explicitly. In order to transform this implicit information to an explicit representation, the interactions of word definitions via significant relations have been studied and association of words by these predefined relations, automatic inferencing of new relationships by considering the interaction of the relations are provided.

Keywords: Knowledge Base, WordNet, Turkish and Turkic languages(dialects) dictionary, semantic and structural relationships, natural language processing(NLP), etymology, computational linguistics

ANLAMSAL TÜRKÇE VE LEHÇELERİ SÖZLÜĞÜ TASARIMI VE UYGULAMASI

Pınar ÖNDER

Yüksek Lisans Tezi – Bilgisayar Mühendisliği
Temmuz 2009

Tez Yöneticisi Yrd. Doç. Dr. Zeynep ORHAN

ÖZ

Geleneksel sözlükler kelime, tanım ve bazen de örnek cümleler sağlarlar. Bununla birlikte, önemli özelliklerin bir çoğu, kelimeler hakkında bilgi ve aralarındaki ilişkiler gösterilmez. İngilizce için bazı özelliklere ve kelimeler arası ilişkilere sahip uygulamalar bulmak muhtemelken, bu tür uygulamaları Türkçe için görmek mümkün değildir. Bu yüzden bu çalışmanın ana fikri Türkçe kelimeler arasındaki anlamsal ilişkileri göstermektir.

Bu çalışmada, kelimeler arası karşılaştırmayı ve bu kelimelere erişimi kolaylaştırmak için bir yapı oluşturulmuş, kelime tanımlarından anlamsal bilgiler çıkarılarak bir bakıma örtülü bilgi açık hale getirilmiştir. Örtülü haldeki bu bilgiyi açık bir gösterime çevirmek için, belirgin ilişkiler üzerinden kelime tanımları arasındaki etkileşimler ve bu önceden tanımlı ilişkiler ile kelimeler arasındaki birliktelikler çalışılmış, ilişkiler arasındaki etkileşim göz önünde bulundurularak yeni ilişkilerin otomatik çıkarımı sağlanmıştır.

Anahtar Kelimeler: Bilgi tabanı, WordNet, Türkçe ve Lehçeleri sözlüğü, anlamsal ve yapısal ilişkiler, doğal dil işleme(DDİ), etimoloji, hesaplamalı dilbilim.

.

DEDICATION

To my family

ACKNOWLEDGEMENT

I am heartily thankful to my supervisor, Asist. Prof. Zeynep Orhan, whose encouragement, guidance and support from the initial to the final level enabled me to develop and implement this project. The preparation of this thesis study would not have been possible without the support, hard work and endless efforts of my advisor. I appreciate her patience, understanding and encouragement throughout my thesis study.

I would like to give my special thanks to my husband Turan, friends and whole family whose patient love and endless help enabled me to complete this work

Additionally I would like to thank my friend Volkan USLAN for the support and help he has given during my thesis.

Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of this study.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

LIST OF SYMBOLS AND ABBREVIATIONS

SYMBOL/ABBREVIATION

AC	Accuracy
AHFD	American Heritage First Dictionary
CG	Conceptual Graph
CL	Computational Linguistics
NLP	Natural Language Processing

CHAPTER 1

INTRODUCTION

In recent years, the developments in the technology has led to the concept of information age by putting the knowledge forward. Nowadays, especially on the internet, it has become very easy to access and to communicate any information. In parallel to this, the need for the language processing software has emerged for facilitating the knowledge and the technology sharing between different individuals and communities. There will be many important changes in the business world, and also in the international relationships depending on the communication, which will be easier and faster via computational technology. However, man-machine interaction is an important obstacle at this point. Designs/implementations which can provide direct communication in natural languages with the computers will be the key solutions to solve these problems.

Turkish is a language that has been widely used and has an important role among the world languages. Today, it has been spoken with different accents and dialects in more than 20 different geographical areas over the world. Turkic languages are spoken by some 180 million people as a native language; and the total number of Turkic speakers is about 200 million, including speakers as a second language. Despite of the interdisciplinary applications, such as computational linguistics (CL), natural language processing (NLP), artificial intelligence, etc. that have gained increasing attention in the world and its common usage, Turkish is a lesser studied language in these fields.

Today in countries like Europe, America and Japan which attach importance to information processing and communication, great investments for NLP made and as a result of this, softwares and computer systems that provide advantages to the users are developed. Because the most acceptable language in such countries is English, it can be seen that the studies in this scope is in that language. Although Turkish is a widespread language with millions of speakers, it is in the scope of less examined languages. Therefore it can be said that the studies of applied linguistic studies in Turkish are creeping and not enough studies are completed in this area.

However many other technological developments can be copied and used with small changes in different parts of the world or in different cultures, the studies on the field of NLP can not be shared so easily. It is impossible for the rules and algorithms defined for English or any other language to be used in the same way or without any modification for Turkish or any other language duet to the different structures of the languages. The adaptation of the existing systems for the specific language can be achieved only as a result of long and time consuming work. Furthermore the adaptation of these is not often possible and the re-construction obligatory of many systems peculiar to language appears. Also; the works on this field requires proficiency especially on computer science and linguistics. Therefore, scientific works on a language can be carried out by the linguists of that language and computer science experts, and also the scientists who have a wide acquaintance with the language.

Because the valid language in the countries where the researches related to NLP mostly done and carried out in English, it is observed that the studies are largely on this language. NLP technologies which are already obvious to be building stones of feature’s world have a different aspect from other technologic improvements. It is important that these studies are done by the native speakers of the studied language. In other words, the systems of Turkish language which are achieved by native speakers of Turkish will be more productive. Considering that Turkish is being used by millions of people and the dialects of Turkish, the outcomes and gainings of the work is obvious to be exciting.

In order to model the knowledge acquisition, processing, usage and communication abilities of humans in computational domain to some extent, the simulation should be started from the smallest units of human learning mechanisms. The applications mentioned below motivated this study about Turkish and it is planned to study in the word level in the context of this project. Therefore, the main goal of this study is to represent the semantic relationships between Turkish words.

Words are the fundamental building blocks of the communication, thinking, and decision making cognitive processes. While the learning process of words takes place, most of the information related to these words is also kept in the background. Although, the most commonly used dictionaries have been transferred to the electronic environment and have been utilized by information technologies in the last decade, they generally provide only the words and their definitions. However, various useful information and features about the words and relationships among them can not be represented. Therefore, the valuable data can not be facilitated by many other applications. Storing the words along with their various features and relationships in a knowledge base, implementation of WordNet that allows demonstration of wide variety of relationships between words is aimed to put together in the context of this study (Bariere, 1997).

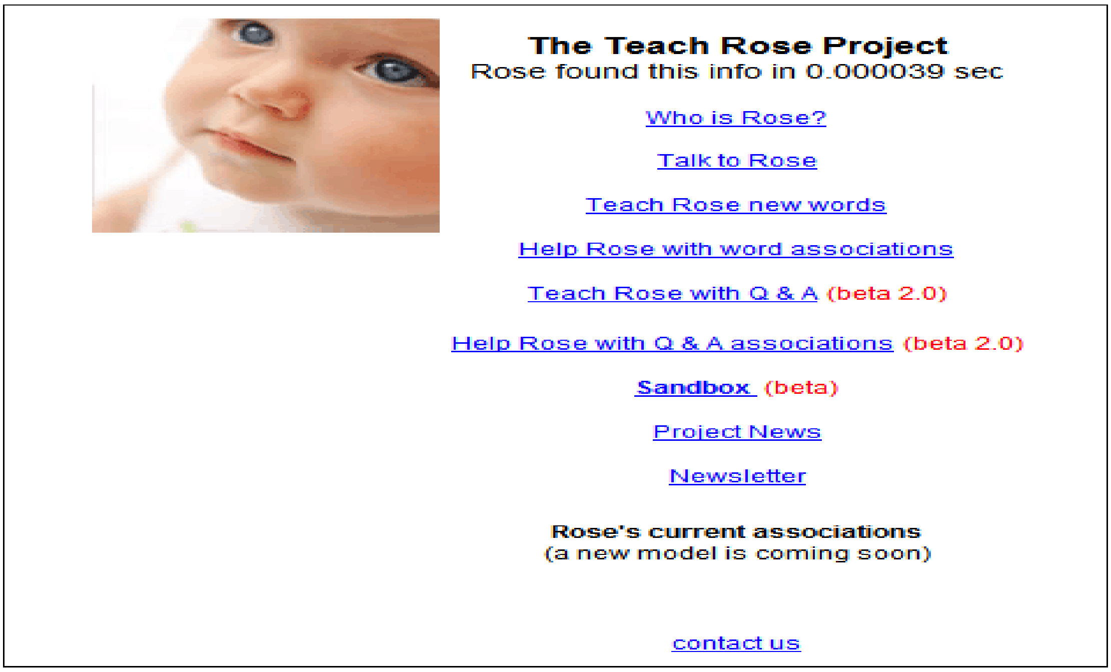
Traditional dictionaries have some fundamental features and generally in various dictionaries word and its definition is the most commonly shared feature. In the context of this study, all useful features that are provided in traditional dictionaries will be brought together, and additionally, insertion of new words and definitions, description of different relationships between words and association of words by these predefined relations, automatic inferencing of new relationships by considering the interaction of the relations will be provided as the fundamental utilities. In the meanwhile, the semantic annotations will be protected by keeping the link between the words and their various senses. An interface will be formed that simulates human language acquisition process and collects the information via internet by the contribution of many people. However, the data formed in this environment will be controlled by experts before the direct transfer to the knowledge base and only the approved ones will be allowed to permanently effect for further processing steps.

1.1 RELATED WORK

While it is possible to find applications that have some specific features and relationships of the words for English such as WordNet (Fellbaum, 1998) and other languages, it is not possible to see these applications for Turkish language. It will be explained detailed in the next section.

1.1.1 The Teach Rose Project

The Teach Rose Project[1] that has been started in the first quarter of 2007 for English has a close relationship with this study. It is simulating the learning mechanism of a child named Rose by an approach called *Hive Mind*. *Hive Mind* uses the theory that if everyone contributes a tiny bit, much likes bees in a bee hive; a massive bee hive can be built. Rose simulates human intelligence by participating in dialogue with site visitors, building vocabulary, building associations, and asking questions.



The Teach Rose Project
Rose found this info in 0.000039 sec

Who is Rose?

Talk to Rose

Teach Rose new words

Help Rose with word associations

Teach Rose with Q & A (beta 2.0)

Help Rose with Q & A associations (beta 2.0)

Sandbox (beta)

Project News

Newsletter

Rose's current associations
(a new model is coming soon)

contact us

Figure 1.1*:* The Information of the Words in the Teach Rose Project

[1] The teach Rose Project: http://teachrose.com/index.php

1.1.2 Lexical Knowledge Base of Conceptual Graphs

The study of Bariere (Bariere, 1997) aims at building Lexical Knowledge Base by extracting information from a machine readable dictionary American Heritage First Dictionary (AHFD) designed for children. The data extracted from the dictionary is represented as a conceptual graph (CG) presenting the explicit relations and information about the words.

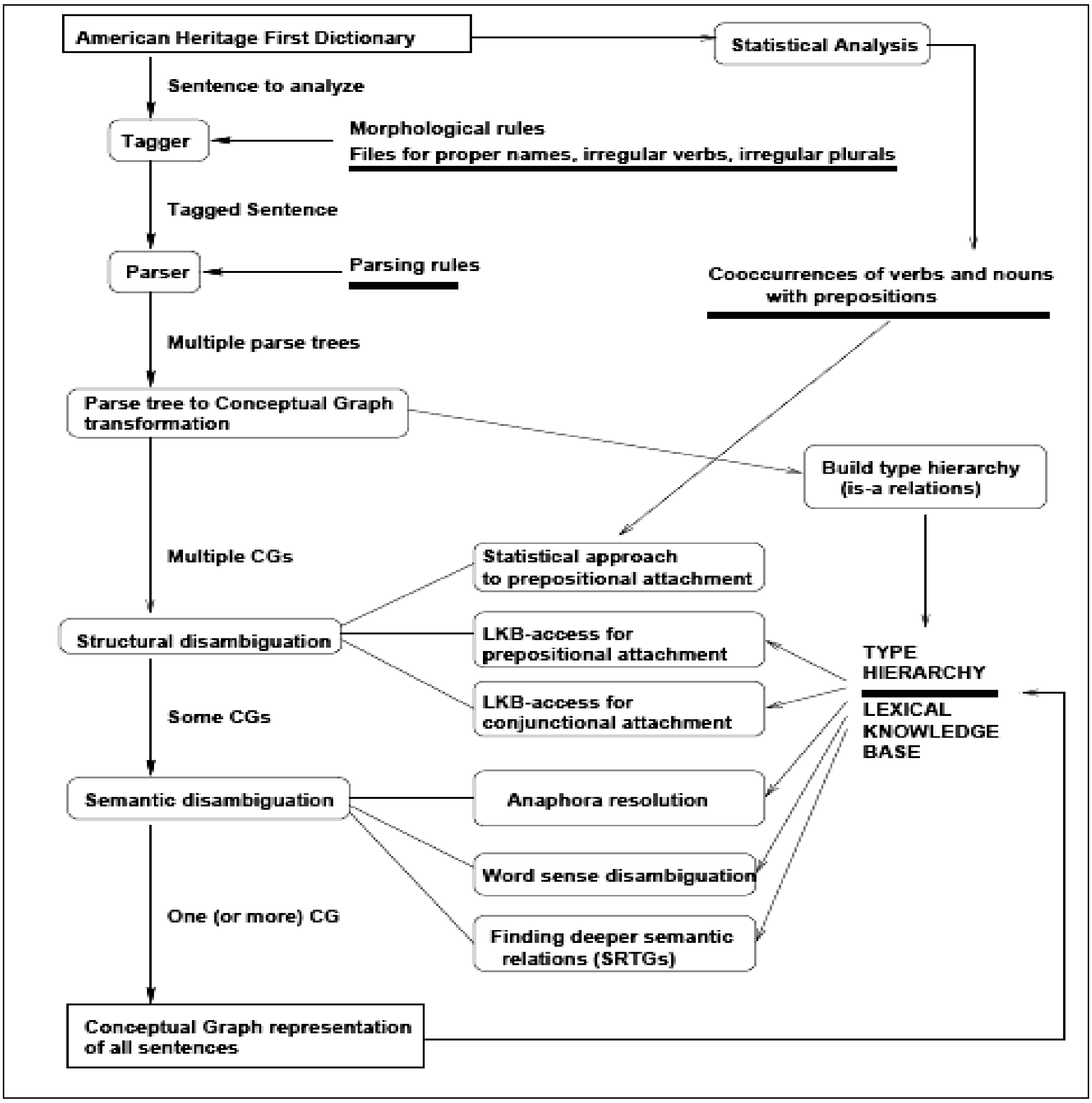

American Heritage First Dictionary
Statistical Analysis
Sentence to analyze
Tagger
Morphological rules
Files for proper names, irregular verbs, irregular plurals
Tagged Sentence
Parser
Parsing rules
Cooccurrences of verbs and nouns with prepositions
Multiple parse trees
Parse tree to Conceptual Graph transformation
Build type hierarchy (is-a relations)
Multiple CGs
Statistical approach to prepositional attachment
Structural disambiguation
LKB-access for prepositional attachment
TYPE HIERARCHY
LEXICAL KNOWLEDGE BASE
LKB-access for conjunctional attachment
Some CGs
Semantic disambiguation
Anaphora resolution
Word sense disambiguation
One (or more) CG
Finding deeper semantic relations (SRTGs)
Conceptual Graph representation of all sentences

Figure 1.2 All Steps from a Sentence to Conceptual Graphs

The type hierarchy, extracted automatically from the definitions, groups all the nouns and verbs in the dictionary into taxonomy. The relation hierarchy is built manually which groups into subclasses/superclasses the relations used in CG representation of definitions. Its graph representation is joined to the graph representations of other words in the dictionary that are related to it. The set of related words form a concept cluster and their graph representation, showing all the relations between them and other related words, is a concept clustering knowledge graph as shown in Figure 1.2 All Steps from a Sentence to

One important aspect of this study is the underlying thread of finding similarity through concept as a general way of processing information.

An important study on creating an information base from the words and their meanings and creating their graphics is implemented using AHFD (Bariere, 1997). Conclusions are done according to meaningful sentences and the relations between the words are shown by graphics.

1.1.3 Sesli Sözlük

Sesli Sözlük[2] which is developed for some languages including Turkish is a good study in this area. It is improving with the contributions of the users. The information entered by the users is added to the system by voting method. Also the pronunciation of the words can be listened, translations can be found and it can be used with mobile devices shows finding a word in Sesli Sözlük as shown in Figure 1.4.

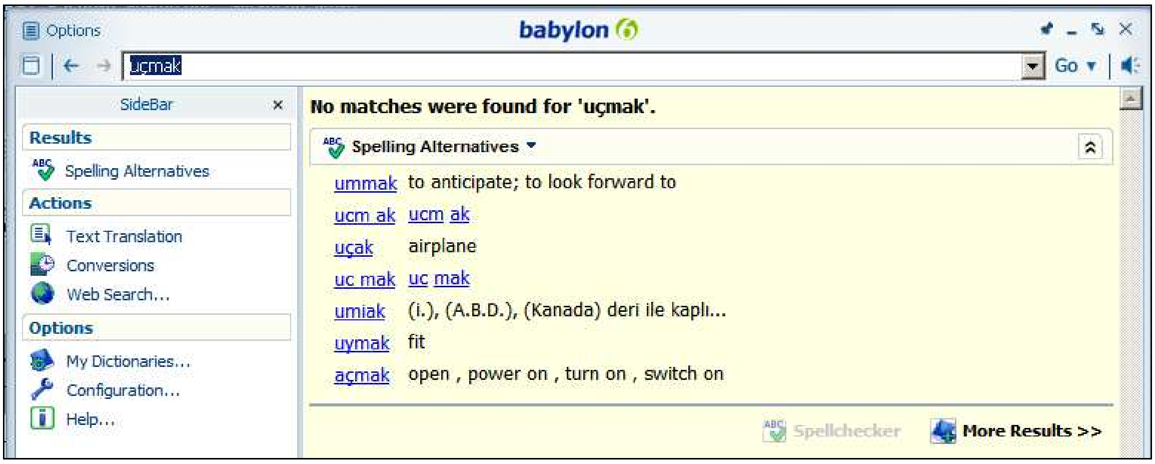
1.1.4 Babylon

Another major study in this area is Babylon[3] which developed by Babylon. It is founded in 1997. It translates and gives information of the words which are clicked on as shown in Figure 1.3. In its 7th version it translates in 17 languages and gives Wikipedia

[2] Sesli Sözlük: http://www.seslisozluk.com

[3] Babylon: http://www.babylon.com/

information in 13 languages. It uses 1300 database in 75 languages. It is used by 35 million computers in 168 countries.



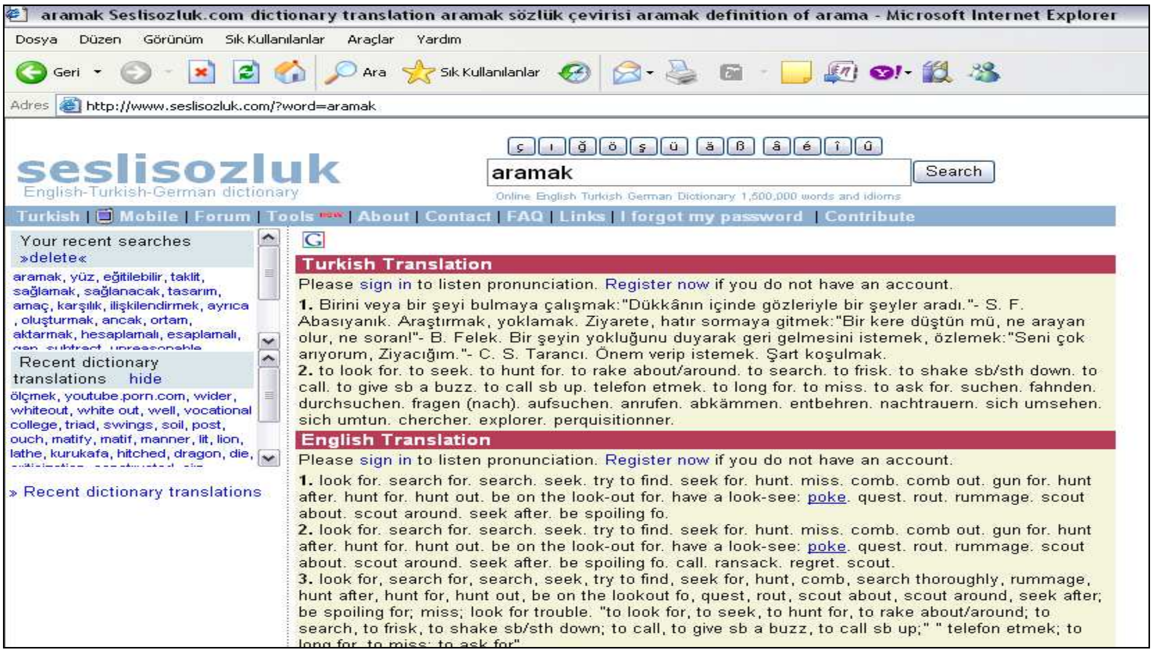
No matches were found for 'uçmak'.

Spelling Alternatives

ummak	to anticipate; to look forward to
ucm ak	ucm ak
uçak	airplane
uc mak	uc mak
umiak	(i.), (A.B.D.), (Kanada) deri ile kaplı...
uymak	fit
açmak	open , power on , turn on , switch on

Figure 1.3 Information of Words in Babylon

Also the pronunciations of the words are added. It gives Turkish results but it needs to be improved.


aramak Seslisozluk.com dictionary translation aramak sözlük çevirisi aramak definition of arama - Microsoft Internet Explorer
Dosya Düzen Görünüm Sık Kullanılanlar Araçlar Yardım
Geri
Ara
Sık Kullanılanlar
Adres http://www.seslisozluk.com/?word=aramak
seslisozluk
English-Turkish-German dictionary
ç ı ğ ö ş ü ä ß â é î û
aramak
Search
Online English Turkish German Dictionary 1,500,000 words and idioms
Turkish | Mobile | Forum | Tools new | About | Contact | FAQ | Links | I forgot my password | Contribute
Your recent searches
»delete«
aramak, yüz, eğitilebilir, taklit, sağlamak, sağlanacak, tasarım, amaç, karşılık, ilişkilendirmek, ayrıca, oluşturmak, ancak, ortam, aktarmak, hesaplamalı, esaplamalı,
Recent dictionary translations hide
ölçmek, youtube.porn.com, wider, whiteout, white out, well, vocational college, triad, swings, soil, post, ouch, matify, matif, manner, lit, lion, lathe, kurukafa, hitched, dragon, die,
» Recent dictionary translations

Turkish Translation

Please sign in to listen pronunciation. Register now if you do not have an account.

1. Birini veya bir şeyi bulmaya çalışmak:"Dükkânın içinde gözleriyle bir şeyler aradı."- S. F. Abasıyanık. Araştırmak, yoklamak. Ziyarete, hatır sormaya gitmek:"Bir kere düştün mü, ne arayan olur, ne soran!"- B. Felek. Bir şeyin yokluğunu duyarak geri gelmesini istemek, özlemek:"Seni çok arıyorum, Ziyacığım."- C. S. Tarancı. Önem verip istemek. Şart koşulmak.
2. to look for. to seek. to hunt for. to rake about/around. to search. to frisk. to shake sb/sth down. to call. to give sb a buzz. to call sb up. telefon etmek. to long for. to miss. to ask for. suchen. fahnden. durchsuchen. fragen (nach). aufsuchen. anrufen. abkämmen. entbehren. nachtrauern. sich umsehen. sich umtun. chercher. explorer. perquisitionner.

English Translation

Please sign in to listen pronunciation. Register now if you do not have an account.

1. look for. search for. search. seek. try to find. seek for. hunt. miss. comb. comb out. gun for. hunt after. hunt for. hunt out. be on the look-out for. have a look-see: poke. quest. rout. rummage. scout about. scout around. seek after. be spoiling fo.
2. look for. search for. search. seek. try to find. seek for. hunt. miss. comb. comb out. gun for. hunt after. hunt for. hunt out. be on the look-out for. have a look-see: poke. quest. rout. rummage. scout about. scout around. seek after. be spoiling fo. call. ransack. regret. scout.
3. look for, search for, search, seek, try to find, seek for, hunt, comb, search thoroughly, rummage, hunt after, hunt for, hunt out, be on the lookout fo, quest, rout, scout about, scout around, seek after; be spoiling for; miss; look for trouble. "to look for, to seek, to hunt for, to rake about/around; to search, to frisk, to shake sb/sth down; to call, to give sb a buzz, to call sb up;" " telefon etmek; to long for, to miss; to ask for"

Figure 1.4 The Information of Words in Seslisozluk

1.1.5 Thinkmap Visual Thesaraus

ThinkMap Visual Thesaurus[4] is an interactive dictionary and thesaurus which creates word maps that blossom with meanings and branch to related words. Its innovative display encourages exploration and learning. The word relations are represented by visual interactive components as shown in Figure 1.5.

To do semantic inference, in addition to other resources, a database that includes the relationships between words and terms in the language is needed. There are various studies to create such databases in the literature.

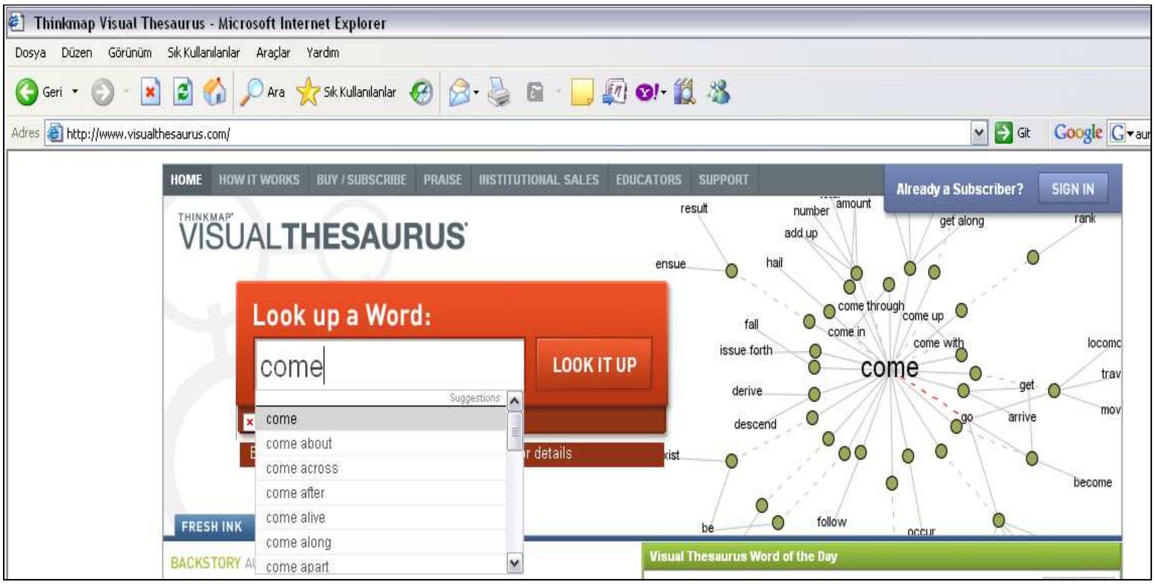


Figure 1.5 The Word Relations in ThinkMap Visual Thesaurus

1.1.6 WordNet

The most common one in these studies is WordNet[5] which includes synonym sets for nouns, verbs and adjectives and some semantic relations between them. WordNet first appeared after five years of study with a great labor and taken up a lot of time and includes

[4] ThinkMap: http://www.visualthesaurus.com

[5] WordNet: http://wordnet.princeton.edu/man/wngloss.7WN.html

150.000 word formats consist of one or more words and 115.000 synonym sets. WordNet uses a hierarchic structure that includes hypernym and hyponym relations. Hypernyms are extracted from descriptions, and then this process is used to obtain new hypernyms by using new inferences as shown in Figure 1.6.

Information in WordNet is organized around logical groupings called synsets. Each synset consists of a list of synonymous words or collocations (eg. "fountain pen" , "take in"), and pointers that describe the relations between this synset and other synsets. A word or collocation may appear in more than one synset, and in more than one part of speech. The words in a synset are grouped such that they are interchangeable in some context (Miller et al., 2005).

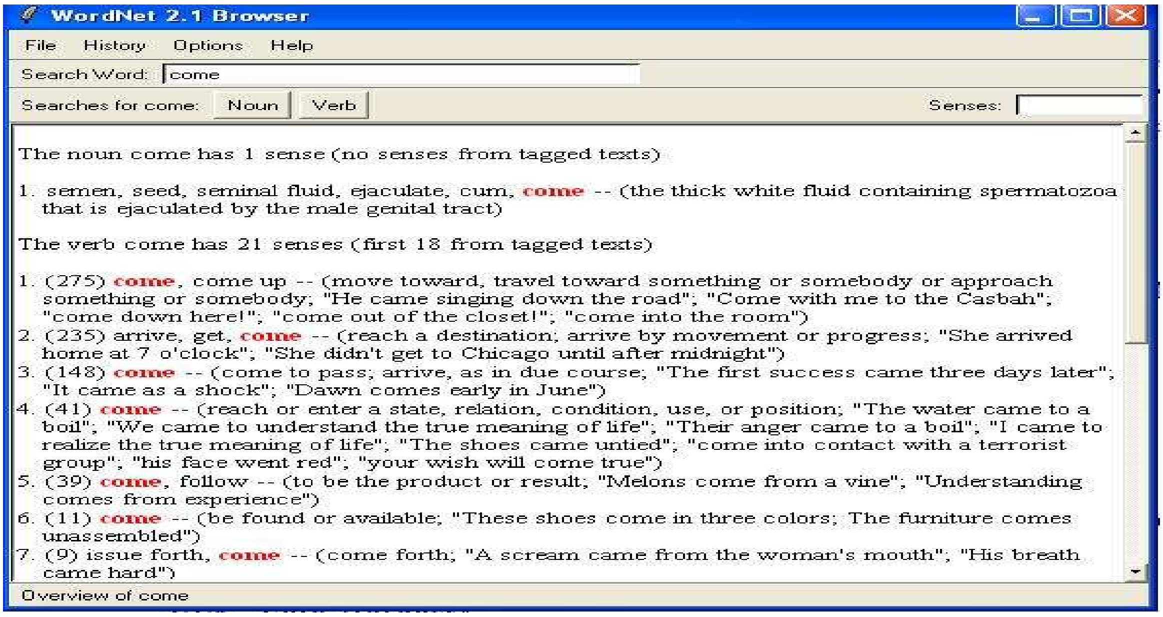
1.1.7 Euro WordNet

EuroWordNet[6] was a European resources and development project supported by the Human Language Technology sector of the Telematics Applications Programme. EuroWordNet is a multilingual database with wordnets for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). The wordnets are structured in the same way as the American WordNet for English (Princeton WordNet, Miller et al 1990) in terms of synsets (sets of synonymous words) with basic semantic relations between them. Each wordnet represents a unique language-internal system of lexicalizations as shown in Figure 1.7.

In addition, the wordnets are linked to an Inter-Lingual-Index, based on the Princeton WordNet. Via this index, the languages are interconnected so that it is possible to go from the words in one language to similar words in any other language. The index also gives access to a shared top-ontology of 63 semantic distinctions. This top-ontology provides a common semantic framework for all the languages, while language specific properties are maintained in the individual wordnets. The database can be used**,** among others, for

[6] EuroWordNet: http://www.illc.uva.nl/EuroWordNet/

monolingual and cross-lingual information retrieval, which was demonstrated by the users in the project (Peters et al., 1998).



Search Word: come

Searches for come:

Senses:

```
The noun come has 1 sense (no senses from tagged texts)

1. semen, seed, seminal fluid, ejaculate, cum, come -- (the thick white fluid containing spermatozoa
   that is ejaculated by the male genital tract)

The verb come has 21 senses (first 18 from tagged texts)

1. (275) come, come up -- (move toward, travel toward something or somebody or approach
   something or somebody; "He came singing down the road"; "Come with me to the Casbah";
   "come down here!"; "come out of the closet!"; "come into the room")
2. (235) arrive, get, come -- (reach a destination; arrive by movement or progress; "She arrived
   home at 7 o'clock"; "She didn't get to Chicago until after midnight")
3. (148) come -- (come to pass; arrive, as in due course; "The first success came three days later";
   "It came as a shock"; "Dawn comes early in June")
4. (41) come -- (reach or enter a state, relation, condition, use, or position; "The water came to a
   boil"; "We came to understand the true meaning of life"; "Their anger came to a boil"; "I came to
   realize the true meaning of life"; "The shoes came untied"; "come into contact with a terrorist
   group"; "his face went red"; "your wish will come true")
5. (39) come, follow -- (to be the product or result; "Melons come from a vine"; "Understanding
   comes from experience")
6. (11) come -- (be found or available; "These shoes come in three colors; The furniture comes
   unassembled")
7. (9) issue forth, come -- (come forth; "A scream came from the woman's mouth"; "His breath
   came hard")
```

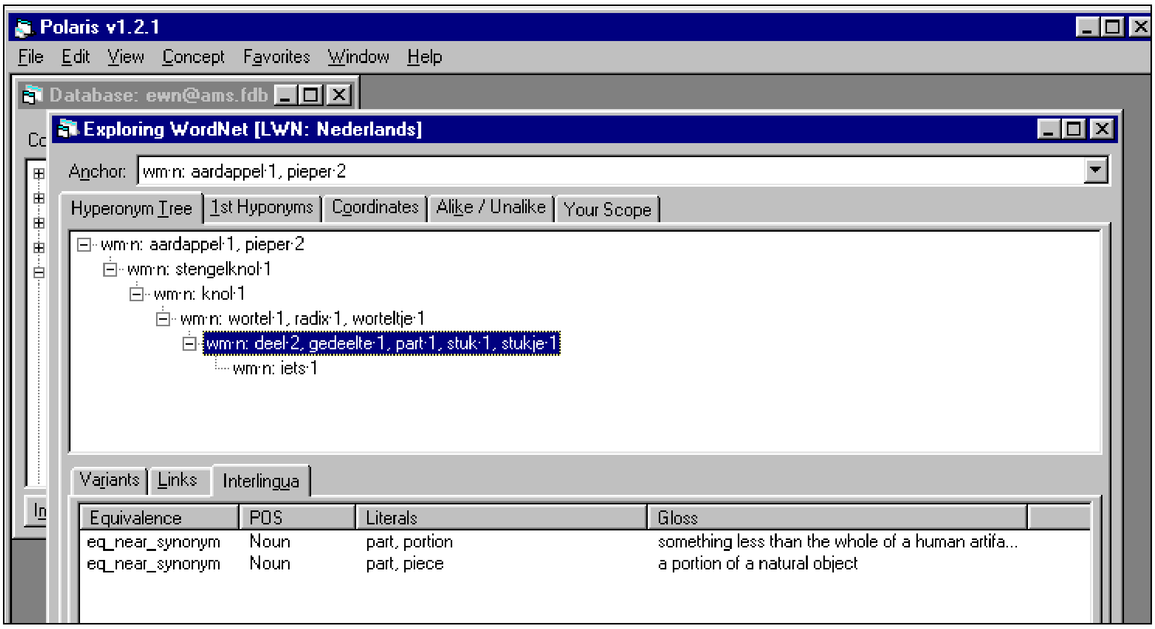
Overview of come

Figure 1.6 The Information of Words in WordNet

The cooperative framework of EuroWordNet is continued through the Global WordNet Association as shown in Figure 1.7. This is a free and public association that builds on EuroWordNet and Princeton WordNet The aim is to stimulate further building of wordnets, further standardization and interlinking and the development of tools, dissemination of information.

1.1.8 BalkaNet Project

Sabancı University Turkish Lexical Database Project is a part of BalkaNet project which aims to design and develop a multilingual lexical database by using the own cognitive dictionaries (-wordnet- electronic dictionaries according to the meaning of the words instead of their structures) of Turkish, Greek, Bulgarian, Czech, Romanian and Serbian languages. Cognitive dictionaries are attemped to associate with cosets (synonym set) and the Interlingual Index which attaches the cognitive dictionaries to each other like Euro WordNet (Bilgin at el, 2004).



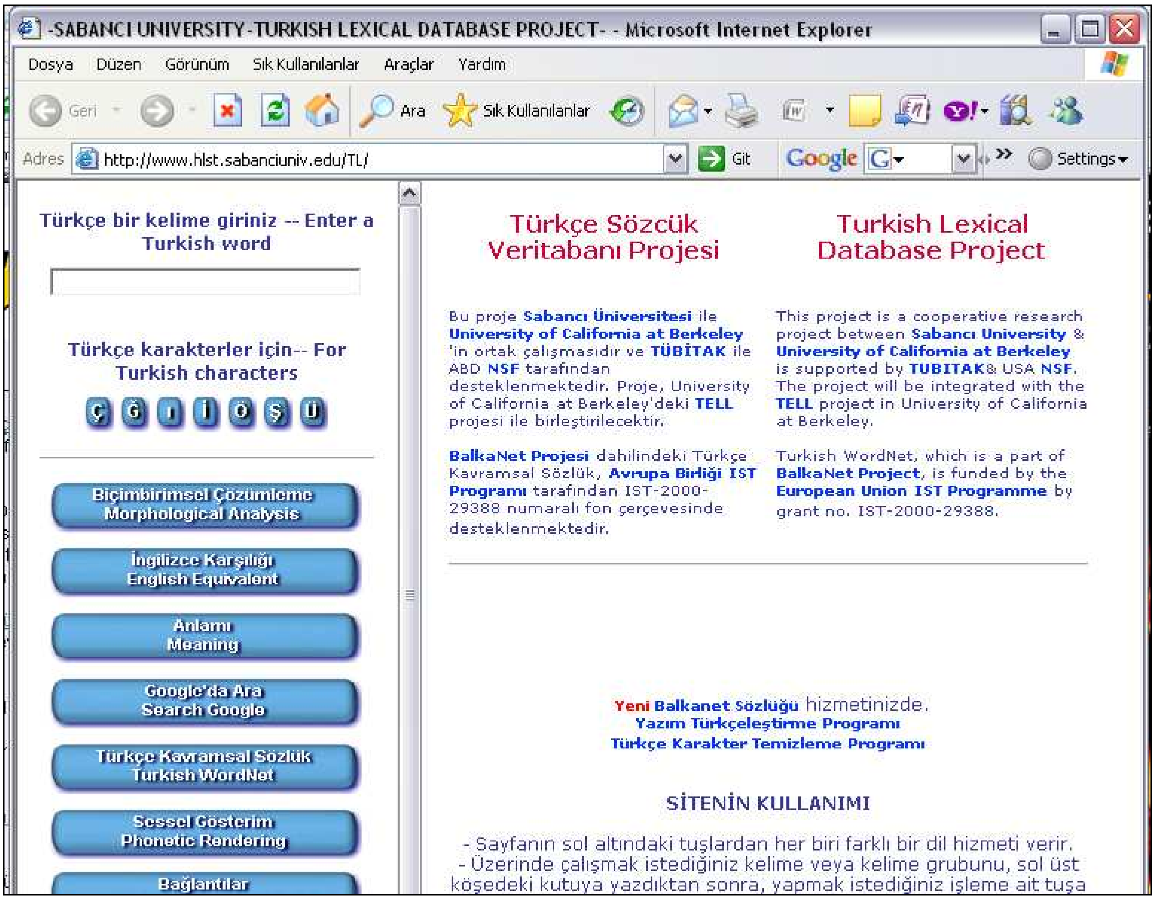
Anchor: wm·n: aardappel·1, pieper·2

```
⊟ wm·n: aardappel·1, pieper·2
   ⊟ wm·n: stengelknol·1
      ⊟ wm·n: knol·1
         ⊟ wm·n: wortel·1, radix·1, worteltje·1
            ⊟ wm·n: deel·2, gedeelte·1, part·1, stuk·1, stukje·1
                  wm·n: iets·1
```

Equivalence	POS	Literals	Gloss
eq_near_synonym	Noun	part, portion	something less than the whole of a human artifa...
eq_near_synonym	Noun	part, piece	a portion of a natural object

Figure 1.7 The Cooperative Framework of EuroWordNet

Sabancı University that carries out the Turkish parts of the BalkaNet Project, created the Turkish Lexical Database which includes topics like development of Turkish synsets and semantic organization, addition of language specific features to the cognitive dictionary and the structural design of the database system which forms the Turkish Lexical Database. But it is needed to extend the scope of the dictionary, increase the relations between words and improve the sample uses of the dictionary.


-SABANCI UNIVERSITY-TURKISH LEXICAL DATABASE PROJECT- - Microsoft Internet Explorer
Dosya Düzen Görünüm Sık Kullanılanlar Araçlar Yardım
Geri
Ara
Sık Kullanılanlar
Adres
Git
Google
Settings
Türkçe bir kelime giriniz -- Enter a Turkish word
Türkçe karakterler için-- For Turkish characters
Ç Ğ ı İ Ö Ş Ü
Biçimbirimsel Çözümleme Morphological Analysis
İngilizce Karşılığı English Equivalent
Anlamı Meaning
Google'da Ara Search Google
Türkçe Kavramsal Sözlük Turkish WordNet
Sessel Gösterim Phonetic Rendering
Bağlantılar

http://www.hlst.sabanciuniv.edu/TL/

Türkçe Sözcük Veritabanı Projesi

Turkish Lexical Database Project

Bu proje Sabancı Üniversitesi ile University of California at Berkeley 'in ortak çalışmasıdır ve TÜBİTAK ile ABD NSF tarafından desteklenmektedir. Proje, University of California at Berkeley'deki TELL projesi ile birleştirilecektir.

BalkaNet Projesi dahilindeki Türkçe Kavramsal Sözlük, Avrupa Birliği IST Programı tarafından IST-2000-29388 numaralı fon çerçevesinde desteklenmektedir.

This project is a cooperative research project between Sabancı University & University of California at Berkeley is supported by TUBITAK& USA NSF. The project will be integrated with the TELL project in University of California at Berkeley.

Turkish WordNet, which is a part of BalkaNet Project, is funded by the European Union IST Programme by grant no. IST-2000-29388.

Yeni Balkanet Sözlüğü hizmetinizde.
Yazım Türkçeleştirme Programı
Türkçe Karakter Temizleme Programı

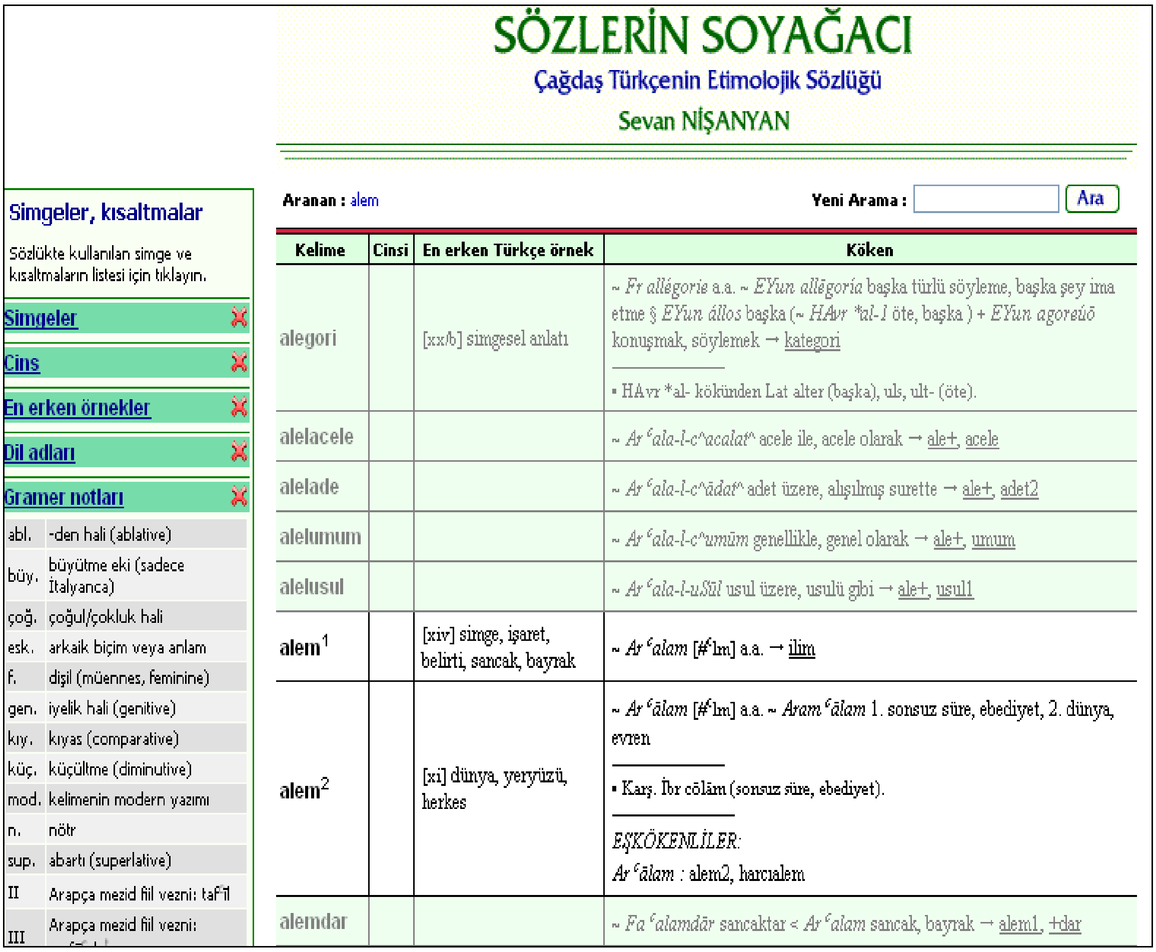
SİTENİN KULLANIMI

- Sayfanın sol altındaki tuşlardan her biri farklı bir dil hizmeti verir.
- Üzerinde çalışmak istediğiniz kelime veya kelime grubunu, sol üst köşedeki kutuya yazdıktan sonra, yapmak istediğiniz işleme ait tuşa

Figure 1.8 Sabancı University Turkish Lexical Database Project

1.1.9 Turkish Etymologic Lexicon Project

Some etymologic dictionaries are created for the languages. The one for Turkish is Nişanyan’s dictionary and includes some etymologic information as shown in Figure 1.9. The dictionary has 12.760 Turkish words. The old words that no longer exist in standard Turkish language, local terms and proper nouns don’t take place in this dictionary. If not necessesary the definiton of the words are not shown and only history and etymologic sources are mentioned (Nişanyan, 2007).



SÖZLERİN SOYAĞACI

Çağdaş Türkçenin Etimolojik Sözlüğü

Sevan NİŞANYAN

Simgeler, kısaltmalar

Sözlükte kullanılan simge ve kısaltmaların listesi için tıklayın.

abl.	-den hali (ablative)
büy.	büyütme eki (sadece İtalyanca)
çoğ.	çoğul/çokluk hali
esk.	arkaik biçim veya anlam
f.	dişil (müennes, feminine)
gen.	iyelik hali (genitive)
kıy.	kıyas (comparative)
küç.	küçültme (diminutive)
mod.	kelimenin modern yazımı
n.	nötr
sup.	abartı (superlative)
II	Arapça mezid fiil vezni: tafᶜīl
III	Arapça mezid fiil vezni:

Aranan : alem

Kelime	Cinsi	En erken Türkçe örnek	Köken
alegori		[xx/b] simgesel anlatı	~ *Fr allégorie* a.a. ~ *EYun allēgoría* başka türlü söyleme, başka şey ima etme § *EYun állos* başka (~ *HAvr *al-1* öte, başka) + *EYun agoreúō* konuşmak, söylemek → kategori ▪ HAvr *al- kökünden Lat alter (başka), uls, ult- (öte).
alelacele			~ *Ar ᶜala-l-c^acalat^* acele ile, acele olarak → ale+, acele
alelade			~ *Ar ᶜala-l-c^ādat^* adet üzere, alışılmış surette → ale+, adet2
alelumum			~ *Ar ᶜala-l-c^umūm* genellikle, genel olarak → ale+, umum
alelusul			~ *Ar ᶜala-l-uSūl* usul üzere, usulü gibi → ale+, usul1
alem^1		[xiv] simge, işaret, belirti, sancak, bayrak	~ *Ar ᶜalam* [#ᶜlm] a.a. → ilim
alem^2		[xi] dünya, yeryüzü, herkes	~ *Ar ᶜālam* [#ᶜlm] a.a. ~ *Aram ᶜālam* 1. sonsuz süre, ebediyet, 2. dünya, evren ▪ Karş. İbr cōlām (sonsuz süre, ebediyet). *EŞKÖKENLİLER:* *Ar ᶜālam* : alem2, harcıalem
alemdar			~ *Fa ᶜalamdār* sancaktar < *Ar ᶜalam* sancak, bayrak → alem1, +dar

Figure 1.9 The Information of the Words in the Turkish Etymologic Lexicon Project

These applications partially similar to some parts of this study are studied and also going on in different languages. As mentioned in the project’s purposes the studies in the literature show the importance and necessity of a study in this area in Turkish.

1.2. THE MOTIVATION

The main purpose of this study can be summarized as:

- Having studies devoted to recover from the language handicaps to facilitate accessing intended knowledge and communication via internet

- Providing contribution to the necessary studies to utilize the interdisciplinary applications and their benefits that keep gaining value all over the world.
- Recovering Turkish from the scope of less examined languages in technologic and linguistic studies in spite of its broad usage.
- Examining humans' ability of obtaining, processing, using information and communication and admitting of modelling these in some degree, obtaining and storing the information kept about words on the background.
- Providing opportunity to natural language processing and computational linguistics studies providing enrichment and expansion for the features of the words in use, obtaining a rich database and a structural and semantic word web,
  - Gathering different meanings of the words,
  - Showing different areas of usage,
  - Obtaining example sentences of the words and having morphologic analysis on these sentences for an open collection text for Turkish language,
  - Giving the pronunciations as voiced or international phonetic spelling,
  - Inserting the pictures if exists,
  - Showing the relationships between words and creating a wordnet,
  - Binding the words with the equivalents in other languages for usage in multilingual platforms,
  - Being scalable, flexible, and trainable for humanbeings
  - Giving many people a chance for contributions using the interfaces which will allow transfer on internet for obtaining and creating information, but also existence of the control mechanism to keep accuracy and data quality high,
  - Immitating the ability of learning and using of humanbeings,
  - Letting the system to update obtaining new automatic inferences extracted from the data,
  - Presenting the obtained results both as visual and as text based,
- Providing the Turkish NLP studies to be faster and center of interest letting the use of the products obtained from the study to the people who want to study on that area.

It is very difficult to find open source studies in Turkish about this in such a wide scope. Therefore this study has the character of being an important study on Turkish natural language processing. The purpose information mentioned above briefly as this study shows this importance.

Finally one of our purposes is, having scientific writings in respectable scientific magazines on such important topics in behalf of our country and taking a step to participate in the European Union projects in this scope and providing Turkish and Turkic languages to be represented in projects, contests and NLP studies.

In this study, latter approach is taken into consideration for representing the meanings of the words. The issue of how one can semi and/or full automatically forms a substantial lexical knowledge base that is useful for natural language processing applications from existing information resource(s) is addressed in this study. Although there are some researches about Turkish language in this area, this study has an importance as being the first one which has the special properties mentioned above for Turkish natural language processing field.

CHAPTER 2

IMPLEMENTATION

2.1 THE LEXICAL KNOWLEDGE

Lexical knowledge is the complete knowledge expressed by the words. The words can be put together to form an infinite number of sentences, each one expressing a distinct meaning. Furthermore, pragmatics asserts that the meaning can also vary depending on the context of utterance. The study of words aiming to understand the meaning and how they relate to each other is a very wide and complex field in itself. The ultimate goal of rendering this information that can be utilized by the computers presents even a harder problem to tackle. Researchers have tried to constrain this problem in a few ways (Bariere, 1997).

Words can be taken as single entities. Generally, they are examined without a complete investigation of their meanings. The interactions between words are given through statistical measurements at different steps of sentence analysis (Önder at el., 2008).

On the other hand it is possible to work with a sublanguage where the number of words is limited and the sentence structures are more restricted. Investigating a smaller set of words allows researchers to go deeper in their analysis and better understanding of the meanings rather than the words themselves.

First of all, in order to have semantic inference, in addition to other resources, it is needed to create a database that stores the semantic relationships between words and concepts in a language. There are various applications for creating these type of databases in the literature. Among these databases, the most common one is the WordNet which has semantic clusters (synonym set – synset) for nouns, verbs and adjectives (Fellbaum, 1998). WordNet is created with a great effort in 5 years and contains 150.000 word structure which has more than one or more words and 115.000 synonyms. WordNet uses hierarchical structure which contains upper-concept and lower-concept between words.

Since there is no common usage of these kind of applications for Turkish language, an application similar to WordNet is being tried to be implemented. Generally most of the words in the dictionaries have several meanings/senses, therefore the relations of the words are established by linking one sense of the source word to the appropriate sense of the target word rather than the words. Otherwise semantic consistency and integrity will be destroyed when wrong relationship between words exist.

The following example illustrates this situation. The word “yüz” in Turkish has senses like “to swim, a hundred, face, etc.” and whenever a relationship is needed between the “sayı”(number) and “yüz” the sense that is “a hundred” has to be linked and the rest of the senses will be irrelavant.

2.2 THE XML LEXICON

Linguistics is the scientific study of NLP. In linguistic, a corpus (plural corpora) or more specifically text corpus is a large and structured set of texts (now usually electronically stored and processed).

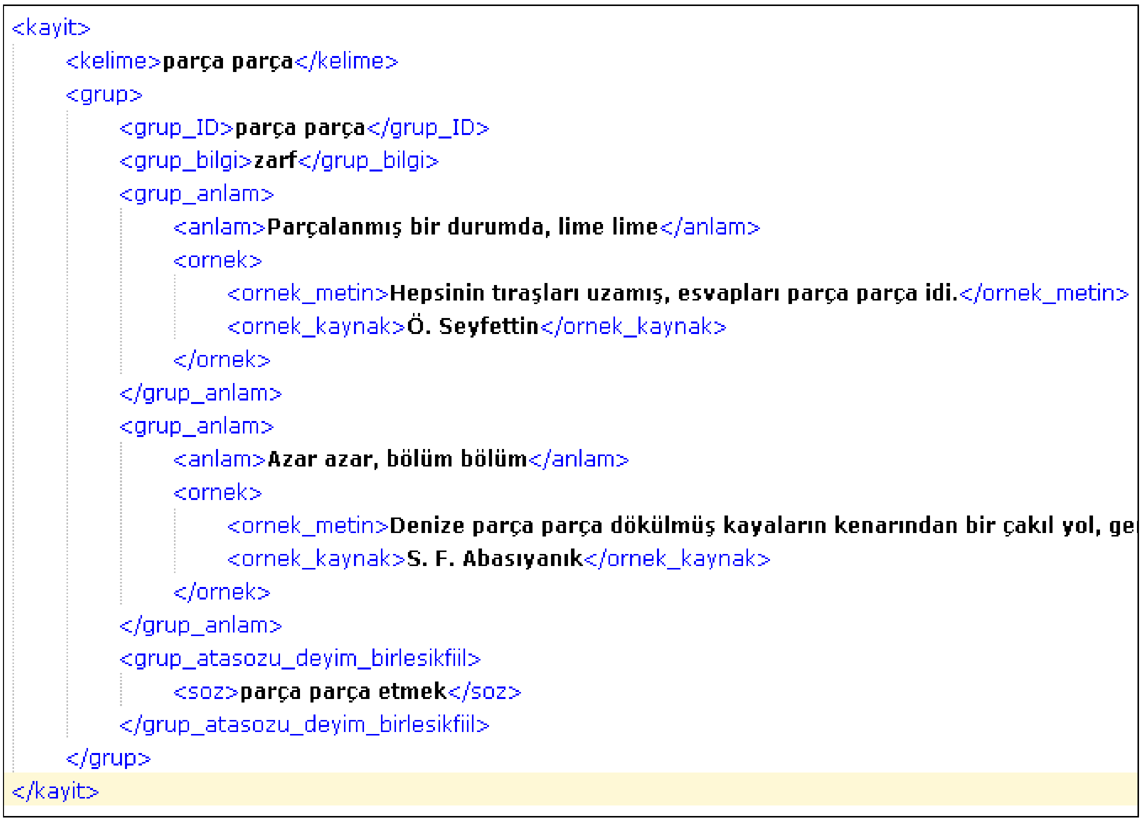
XML is a markup language for documents containing structured information. Structured information contains both content (words, pictures, etc.) and some indication of what role that content. Almost all documents have some structure. A markup language is a mechanism to identify structures in a document. The XML specification defines a standard way to add markup to documents (O’Reilly Media, 2009).

An XML document comprises elements, attributes, processing instructions, comments, and entities.

- **Element**: Text delimited by an opening and a closing /tag/. A tag is a name enclosed within angle brackets.
- **Attribute**: A piece of qualifying information for an element. An attribute consists of a name, an equals sign, and an attribute value delimited by either single-quotes or double-quotes.
- **Processing instruction**: The software that is reading an XML document is referred to as a /processor/. A processing instruction is additional information embedded in the document to inform the processor and possibly change its behaviour.
- **Comment**: An XML comment begins with the characters: less-than, exclamation mark, minus, minus; and ends with the characters: minus, minus, greater-than. Any text within a comment is intended for a human reader and is ignored by the processor.
- **Entity**: An entity is a compact form that represents other text. Entities are used to specify problematic characters and to include slabs of text defined elsewhere. An entity reference consists of an ampersand, a name, and a semi-colon.

XML lexicon is used for Turkish Language Corpus which contains 63K word entries. This corpus consist of an alphabetically series of XML documents. These documents together form the lexicon knowledge base. This lexicon knowledge base is an organized as a description of the lexemes of the language. In each lexeme word senses are described as word-sense entries. For each word sense elements are represented by <kelime>, part of speech is designated by <group_ID>, meaning <anlam> is given and a sample sentence <ornek_metin> is illustrated. The study is implemented by finding relations among words from this lexicon knowledge base.

A typical lexicon entry in the lexicon knowledge base is shown in Figure 2.1.



```
<kayit>
    <kelime>parça parça</kelime>
    <grup>
        <grup_ID>parça parça</grup_ID>
        <grup_bilgi>zarf</grup_bilgi>
        <grup_anlam>
            <anlam>Parçalanmış bir durumda, lime lime</anlam>
            <ornek>
                <ornek_metin>Hepsinin tıraşları uzamış, esvapları parça parça idi.</ornek_metin>
                <ornek_kaynak>Ö. Seyfettin</ornek_kaynak>
            </ornek>
        </grup_anlam>
        <grup_anlam>
            <anlam>Azar azar, bölüm bölüm</anlam>
            <ornek>
                <ornek_metin>Denize parça parça dökülmüş kayaların kenarından bir çakıl yol, ge
                <ornek_kaynak>S. F. Abasıyanık</ornek_kaynak>
            </ornek>
        </grup_anlam>
        <grup_atasozu_deyim_birlesikfiil>
            <soz>parça parça etmek</soz>
        </grup_atasozu_deyim_birlesikfiil>
    </grup>
</kayit>
```

Figure 2.1 A Representation of XML File

2.3 RULE EXTRACTION

The study of words in the goal of understanding their meanings and how they relate to each other is very large and complex field in itself. Aiming to render this information usable by a computer presents an even larger problem. We are interesting in analyzing the definitions given in the Turkish XML lexicon to find the relationships between the words. To do so, we needed to analyze the meaning of the defining sentences from the XML tags <kelime> and <grup_anlam> and in that respect we are interesting in semantic knowledge

Typical relationships and a few examples that can be used in this application are given in Table 2.1.

Table 2.1 Typical RelationShips and Their Examples

RELATION	EXAMPLE
Kind-Of	Fasulye(bean) bitki(plant)
Amount-Of	Hektar(hectare) –ölçü(measurement)
Group-Of	Manga(squad) –asker(soldier)
Member-Of	Burçak(vetch) –baklagil(leguminous)
Synonym	Ak (White) – Beyaz(White)
Antonym	Zor (Hard) – Kolay (Easy)

Much of the work on semantic relations, from a perspective of extraction of information from a dictionary, is done via the analysis of defining formulas. Defining formulas correspond to phrasal patterns that occur often through the dictionary definitions suggesting particular semantic relations.

As it mentioned In PHD Thesis Bariere, the relations presented fall into two main classes: objects or situations. This classification lead us to find new relations.

OBJECTS: The relations found in this group take place in this group, as well as the relation between a unique object and its properties or parts (Bariere, 1997):

- **part/ whole relations:** This class looks at objects that can be segmented into a number of functional parts, or into smaller segments.

Relations: part-of, piece-of, area-of, amount-of, content

- **member/set relations:** This class contains all the relations of quantity of objects, whether we have none, one or many of the same or different types.

  Relations: set-of, element-of

- **human/animal relations**: Some relations only apply to living entities having the capacity for decision, perception and feeling.

  Relations: child-of, possession, home-for

- **comparison relations:** This class contains the relation for comparing the physical properties of two objects.

  Relations: like, more-than, as, less-than

- **spatial relations:** The relations for comparing two objects with regard to their locations.

  Relations: multiple prepositions such as on, in, above, behind

- **word relations**: To some extent, these relations are context independent. They are not part of an object or situation, they are relations on the concepts that words represent instead of the physical entities themselves. We could say they are intensional relations instead of extensional ones.

  Relations: opposite, synonymy, taxonomy

- **description relations**: This class gives the value of different attributes of objects through some formulas that describe the objects.

  Relations: name, attribute, material, function, about

SITUATIONS: This group deals with situations instead of objects; it therefore relates actions to participants, location, and time. As well, it classifies the situations themselves as being states or events depending on the time they take to accomplish in comparison to the surrounding events.

- **action modifiers:** General adverbial modifiers not yet classified as more precise.

  Relations: modify

- **case-role relations:** This is the largest class, it contains all the relations that can be subordinate to a verb.

Relations: instrument, method, manner, agent, experiencer, location, object, recipient, result, cause, transformation, reason, goal, accompaniment, direction, source, destination, path, during, point-in-time, frequency

- **agent involvement**: The agent of the action is a living entity with desires, feelings, goals. The involvement of the agent in a situation is an important factor to its progress.

  Relations: ability, desired-act, intention

- **action relations:** This class contains different types of actions: event, state, process. The three relations form more of a continuum that three independent relations, as the distinction between them is subtle. It involves the ratio of elapsed time between the action itself and the other actions within a situation.

  Relations: act, event, process, sequence.

For example, the relations part-of, made-of can be detected directly via the defining formulas <X1 is a part of X2>, <X1 is made of X2> whenever the definitions contain these patterns. Various rules similar to these have been defined to find the relationships between the words and relationships. Then the frequencies of each rule for the related relations of the words have been calculated. In the meanwhile, transitive or inverse relations have been considered and taken into account. A partial list of rules is provided in Table 2.2.

On the other hand if the relations were too specific, it would be very hard to find formulas for rules from our lexicon that has 63K entries. So the generic rules were defined as shown in Table 2.2 that lists the most frequent defining formulas. The rest of the relations were added by looking through the definition of the words and trying to see which relations would be needed.

Table 2.2 Relationships and the Corresponding Patterns in Turkish

RELATION	**RULES**
Kind-Of	Rule 1: <X: …Y tipi(dir).>
	Rule 2: <X:…. Y çeşidi(dir).>
	Rule 3: <X….: Y türü(dür).>
Amount-Of	Rule 1: <X: ... Y birimi(dir).>
	Rule 2: <X: ….Y miktarı(dır).>
	Rule 3: <X:..... Y ölçüsü(dür).>
Group-Of	Rule 1: <X: …Y topluluğu(dur).>
	Rule 2: <X:… Y kümesi(dir).>
	Rule 3: <X: …Y birliği(dir).>
	Rule 4: <X:… Y(den|dan) oluşan topluluk.>
	Rule 5: <X:…. Y bütünü).>
	Rule 6: <X: ….Y tümü).>
	Rule 7: <X:…. Y sürüsü.>
Member-Of	Rule 1: <X:….. Y'nin üyesi(dir).>
	Rule 2: <X:….. Y+gillerden(dir).>
	Rule 3: <X: …..Ysınıfı.>
	Rule 4: <X: ….Y takımı.>
Synonymy	Rule 1: <X: Y (single word).>
	Rule 2: <X:…..,Y. (after comma,the last word of the sentence)>
Antonymy	Rule 1: <X:…. Y karşıtı.>
	Rule 2: <X:…. Y olmayan.>

2.4 EXTRACTED RELATIONS

In this section from the Object group "synonymy, antonymy, amount-of, member-of" relations have been analyzed in great detail. Additionally the hierarchical relation is shown by the kind-of and a member-of relation extracted from the definitions via defining formulas such as shown in the examples below and followed by illustrative sentences and the predicates that can be derived from them.

2.4.1 Relation: Kind-of

Most of the extraction techniques rely on finding defining formulas within the defining sentences. Rules such as <X: Y tipi(dir). >(X is a type of Y), <X: Y çeşidi(dir) > (X is a kind of Y), <X: Y türü(dür).> (X is a sort of Y).are mentioned as the most common ways to identify a ”Kind-Of” relation between concepts X and Y.

(Rule 1) **X:** W_1 W_2 W_3…………..W_n . *Y* *tipi(dir).*

mavzer:…….orduda kullanılan bir tüfek tipi.

Kind-of *{“mavzer”,”tüfek”}*

mavzer(mauser) is a kind of tufek(rifle)

(Rule 2) **X:** W_1 W_2 …W_n . *Y* *çeşidi(dir).*

defne yaprağı: …………Bir lüfer çeşidi.

Kind-of *{“defne yaprağı”,”lüfer”}*

defneyaprağı(bluefish) is a kind of lüfer(bluefish

(Rule 3) **X:** W_1 W_2 W_3......................W_n . *Y* *türü(dür).*

atari:… basit programlarla düzenlenmiş bir oyun türü.

Kind-of *{“atari”,”oyun”}*

atari(video game systems) is a kind of oyun(game)

2.4.2 Relation: Amount of

(Rule 1) $\underbrace{X:}$ $\underbrace{W_1 W_2 W_3..........................W_n.}$ $\underbrace{Y}$ *birimi(dir)*

Amper:... Elektrik akımında…………… şiddet birimi.

Amount-of *{"şiddet","amper"}*

Amper(ampere) is an amount of şiddet(amplitude)>

(Rule 2) $\underbrace{X:}$ $\underbrace{W_1\ W_2\ \ldots..........................W_n}$ $\underbrace{Y}$ *miktarı(dır).*

kapasite:... ....Bir işletmenin..................üretim miktarı.

Amount-of *{"kapasite","üretim"}*

kapasite(capacity) is an amount of üretim(manufacture)

(Rule 3) $\underbrace{X:}$ $\underbrace{W_1\ W_2\ \ldots.........................W_n}$ $\underbrace{Y}$ *ölçüsü(dür).*

aruz:... ....divan edebiyatı................nazım ölçüsü

Amount-of *{"nazım","aruz"}*

aruz(prosody) is an amount of nazım(poetry)

2.4.3 Relation:Group-of

(Rule 1) $\underbrace{X:}$ $\underbrace{W_1\ W_2\ \ldots..............................W_n}$ $\underbrace{Y}$ *topluluğu(dur)*

cins:... ....Pek çok ortak özellikleri bulunan türler topluluğu.

Group-of *{"cins","tür"}*

cins is a group of tür (species)

(Rule 2) $\underbrace{\boldsymbol{X:}}$ $\underbrace{W_1\ W_2\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots W_n}$ $\underbrace{\boldsymbol{Y}}$ *kümesi(dir).*

skala:... .... Bir bestede kullanılabilecek aynı türden sesler kümesi.

Group-of *{"skala","ses"}*

skala (scale) is a group ses (tone)

(Rule 3) $\underbrace{\boldsymbol{X:}}$ $\underbrace{W_1\ W_2\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots W_n}$ $\underbrace{\boldsymbol{Y}}$ *birliği(dir)*

hece:...........Bir solukta çıkarılan.............ses veya ses birliği, seslem.

Group-of *{"hece","ses"}*

hece(syllable) is a group of ses(tone)

(Rule 4) $\underbrace{X:}$ $\underbrace{W_1\ W_2\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots W_n}$ *oluşan* $\underbrace{\boldsymbol{Y}}$ *topluluğu.*

Grup:……altında birleştirilmesinden .... oluşan kıta topluluğu.

Group-of *{"grup","kıta"}*

grup(group) is a group of kıta(detachment)

(Rule 5) $\underbrace{\boldsymbol{X:}}$ $\underbrace{W_1\ W_2\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots W_n}$ $\underbrace{\boldsymbol{Y}}$ *bütünü*

donanma: Belli bir amaçla kullanılan gemilerin bütünü.

Group-of *{"donanma","gemi"}*

donanma(fleet) is a group of gemi(ship)

(Rule 6) $\underbrace{\boldsymbol{X:}}$ $\underbrace{W_1\ W_2\ \ldots\ldots\ldots\ldots\ldots\ldots W_n}$ $\underbrace{\boldsymbol{Y}}$ *tümü*

bitki örtüsü: Bir bölgede yetişen.......... bitkilerin tümü

Group-of *{"bitki örtüsü","bitki"}*

bitki örtüsü(flora) is a group of bitki(plant).>

(Rule 7) $\underbrace{\textbf{\textit{X:}}}\quad \underbrace{W_1\ W_2\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots W_n}\quad \underbrace{\textbf{\textit{Y}}}$ *sürüsü*

nahır: .... *Sığır sürüsü*

Group-of *{“nahır”,”sığır”}*

nahır(flock of cattle) is a group of sığır(cattle).

2.4.4 Relation: Member-of

(Rule 1) $\underbrace{\textit{X:}}\quad \underbrace{W_1\ W_2\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots W_n}\quad \underbrace{\textbf{\textit{Y}}}$ *üyesi(dir).*

gangster: ....Yasa dışı işler yapan *çete üyesi.*

Member-o f {*“ gangster”,” çete”* ***}***

gangster(gangster) is a member of çete(gang).

(Rule 2) $\underbrace{\textbf{\textit{X:}}}\quad \underbrace{\textbf{\textit{Y}}\textit{+gillerden}}\textit{,}\quad \underbrace{W_1\ W_2\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots W_n}$ *bitki*

ahududu: Gülgillerden, böğürtleni andıran, ... bir bitki

Member-o f{*“ raspberry”,” Rosaceae”****}***

Ahududu (raspberry) is a member of gülgiller (Rosaceae)

(Rule 3) $\underbrace{\textbf{\textit{X:}}}\quad \underbrace{W_1\ W_2\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots W_n}\ \underbrace{\textbf{\textit{Y}}}$ *sınıfı.*

Ilmiye: ....Din işleriyle uğraşan........... hocalar sınıfı

Member-o f {*“ ilmiye”,” hoca”****}***

İlmiye is a member of hoca.

(Rule 4) $\underbrace{\textbf{\textit{X:}}}\quad \underbrace{W_1\ W_2\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots W_n}\ \underbrace{\textbf{\textit{Y}}}$ *takımıı.*

Formül: .... ilkeyi açıklayan .... *simgeler takımı.*

Member-o f {*“ simge”,” formül”****}***

Simge(symbol) is a member of formül(formula)

2.4.5 Relation: Synonymy

(Rule 1) $\underbrace{\textbf{X:}}\qquad \underbrace{\textbf{Y}}$

Bağışlamak: Affetmek

Synonym {*" bağışlamak"," affetmek"****}***

bağışlama (forgiveness) is a synonym of affetme(forgiveness).>

(Rule 2) $\underbrace{\textbf{\textit{X:}}}\qquad \underbrace{W_1\ W_2\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots W_n}\ \underbrace{,\textbf{\textit{Y}}.}$

mazeretli: ....Mazereti olan, mazur.

Synonym {*" mazeretli"," mazur"****}***

mazeretli (excused) is a synonym of mazur(excused).

2.4.6 Relation: Antonymy

(Rule 1) $\underbrace{\textbf{\textit{X:}}}\qquad \underbrace{W_1\ W_2\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots W_n}\ ,\underbrace{\textbf{\textit{Y}}}\quad \textit{karşıtı.}$

aç: ....Yemek yemesi gereken, ... tok karşıtı

Antonym {*" aç"," tok"****}***

aç (hungry) is an antonym of tok(satiated).

(Rule 2) $\underbrace{\textbf{X:}}\qquad \underbrace{W_1\ W_2\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots W_n}\ \underbrace{\textbf{Y}}\quad \text{olmayan.}$

ham: ....Yenecek kadar olgun olmayan (meyve).

Antonym {*" ham"," olgun"****}***

ham(unripe) is an antonym of olgun (ripe).

2.5 DISCOVERY OF NEW RULES

A hypernymy relation denotes that there is a word such that it is more specific than the other word. This relationship is also can be called as an “IS A” relation. On the other hand hyponymy relation denotes that the word is a subclass of a more generic word. Necessary sub-rules for relations to improve the accuracy are applied. Further for finding rules of hypernymy relations through lexicon entries from the lexicon knowledge base are investigated. The table lists below the rules found and implemented in extraction of word relations.

From the definitions via extracted relations such as shown in below:

(Rule 1) ***X:*** *Bu renkte olan .(X is a colour)*

mavi: Bu renkte olan.

In the example above the lexicon entry “mavi”(blue) is related with word “renk”(colour) and in English form the sentence can be expressed as “blue is a colour”.

(Rule 2) ***X:*** *Y (gillerden) ,W_1 W_2 ……….W_n* *bitki.(**X** is a plant which is a member of **Y**)*

ebegümeci: Ebegümecigillerden,………… çiçekli bir bitki.

Ebegümecigiller: Ayrı taç yapraklı iki çeneklilerden, örnek bitkisi ebegümeci olan bir bitki familyası.

In the example above the lexicon entry “ebegümeci” (mallow) is related with word “bitki” (plant) and “ebegümecigiller” (Malvaceae). And entry “ebegümecigiller” (Malvaceae) is related with word “iki çenekliler” (Magnoliopsida) and“bitki”(plant).So in English form the sentence can be expressed as “mallow is a plant which is a member of Malvaceae” and . “Malvaceae is a plant which is a member of Magnoliopsida”. This example is shown in Figure 2.2 as a tree structure.

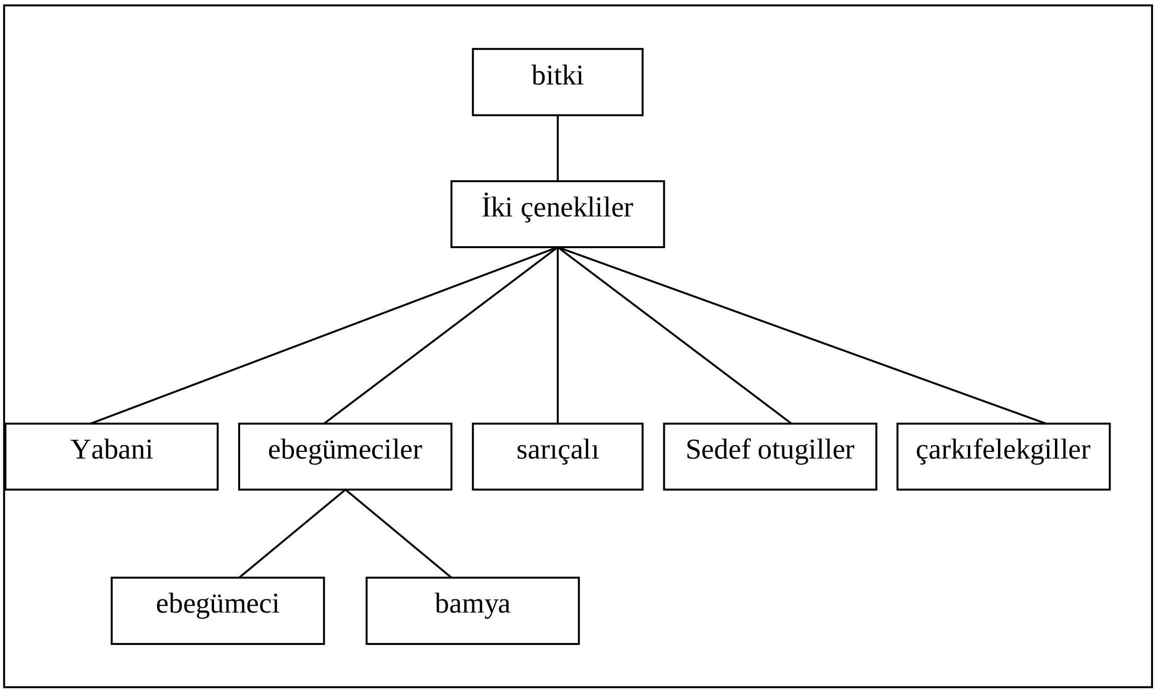


Figure 2.2 Hierarchical Structure Of Plant (Bitki) Class

(Rule 3) ***X:*** ***Y**(gillerden|lerden|lardan),W_1 W_2 …W_n hayvan .(**X** is an animal which is a member of **Y**)*

pars: Kedigillerden, genellikle ……….., etçil, memeli hayvan,.

In the example above the lexicon entry “pars” (leopard) is related with word “hayvan” (animal) and “kedigiller” (Felidae) so in English form the sentence can be expressed as “leopard is an mammal animal which is a member of Felidae”. This example is shown in Figure 2.3.

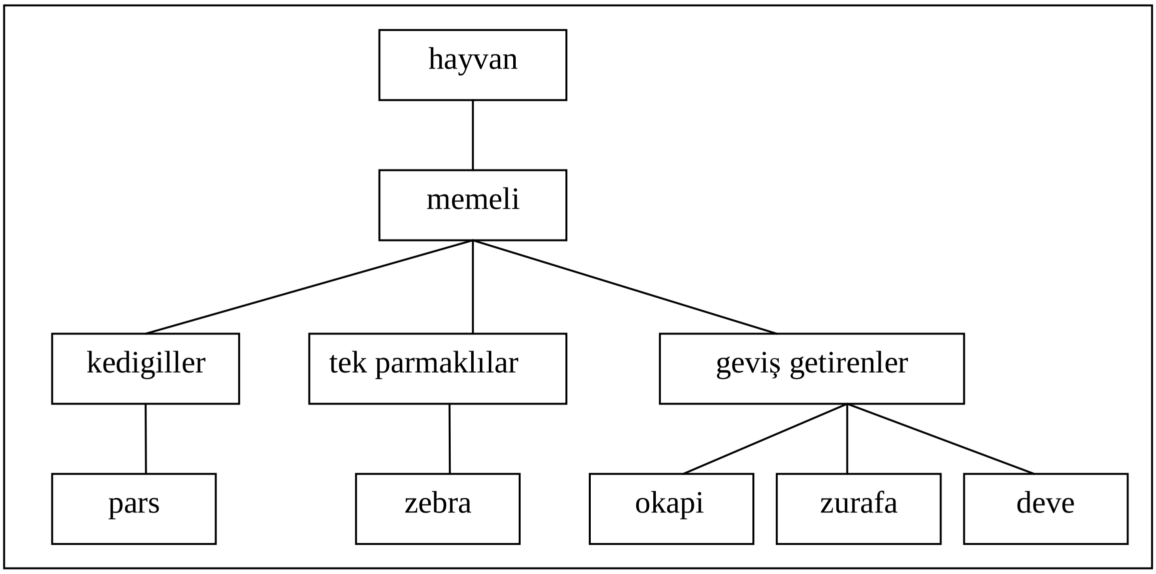


Figure 2.3 Hierarchical Structure of Animal (Hayvan) Class

(Rule 4) ***X:*** W_1 W_2 .... **…**W_n … *bir element.(X is an element)*

kalay: Atom numarası 50, **…………,** *yumuşak bir element.*

In the example above the lexicon entry “kalay” (tin) is related with word “element”(element) and in English form the sentence can be expressed as “tin is an element”.

(Rule 5) ***X:*** W_1 W_2 W_n araç. (X is a tool)

fırın: Bir maddenin fiziksel ı araç.

In the example above the lexicon entry “fırın” (oven) is related with word “araç”(tool) and in English form the sentence can be expressed as “fırın is a tool”.

(Rule 6) ***X:*** W_1 W_2 ………..…W_n .yer. (X is a place)

cephe: Yerde veya daha yükseklerde………………, karşılaştıkları yer.

In the example above the lexicon entry “cephe”(exposure) is related with word “yer” (place) and in English form the sentence can be expressed as “cephe is a place”.

(Rule 7) $\underbrace{W_1}\quad \underbrace{W_2}\ \dots\ \underbrace{W_n}$ ***vb.*** $\underbrace{W_1 \quad W_2 \dots\dots\dots\dots\dots\dots\ W_n}$ *(such as)*

Ölüm, yangın, deprem vb. olayların yarattığı üzüntü, keder, elem

In the example above in lexicon meaning “ölüm”(death), “yangın”(fire), “deprem”(earthquake) is related with word “olay”(event) and in English form the sentence can be expressed as “{death,fire, earthquake} are events”.

(Rule 8) $\underbrace{\boldsymbol{X}\textbf{:}}\quad \underbrace{W_1\ W_2\ \dots W_n}$ *kimse . (X is a person)*

dondurmacı: Dondurma yapan veya satan kimse

In the example above in lexicon entry “dondurmacı”(iceman) is related with word “meslek”(occupation) and in English form the sentence can be expressed as “dondurmacı is an occupation”. This example is shown in Figure 2.4 as a tree structure.

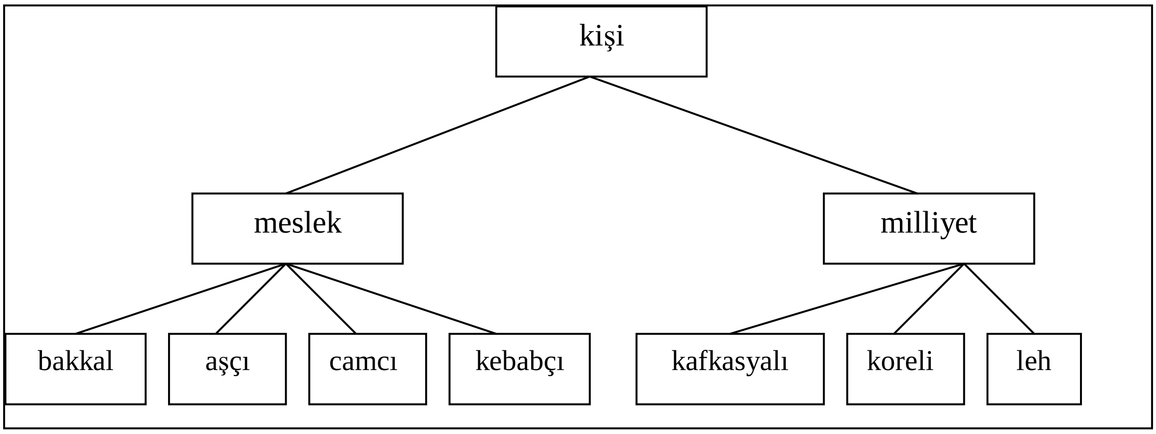


Figure 2.4 Hierarchical Structure of a Person (Kişi) Class

2.6 MORPHOLOGICAL ANALYSIS

Turkish is an agglutinative language and frequently uses affixes, and specifically suffixes, or endings (Lewis, 2001). One word can have many affixes and these can also be used to create new words, such as creating a verb from a noun, or a noun from a verbal root. Most affixes indicate the grammatical function of the word (Lewis, 2001). The only native prefixes are alliterative intensifying syllables used with adjectives or adverbs.

The extensive use of affixes can give rise too long words. To give an example, a morphological structure of a word in a Turkish language is given in the following example (Jurafsky and Martin, 2006):

Turkish: uygarlaştıramadıklarımızdanmışsınızcasına

English: (behaving) as if you are among those whom we could not civilize/cause to become civilized

uygar	+laş	+tır	+ama	+dık	+lar
civilized	+become	+causative	+not able	+participle	+pl
+ımız	+dan	+mış	+sınız	+casına	
+person1pl	+ablative	+past	+2pl	+as if	

Therefore all words that are acquired from the patterns have to be morphologically parsed to obtain the word stems.

Turkish extensively uses agglutination to form new words from nouns and verbal stems. The majority of Turkish words originate from the application of derivative suffixes to a relatively small set of core vocabulary.

An example set of words derived from a substantive root is shown in Table 2.3 and an example of starting from a verbal root is shown in Table 2.4.

Table 2.3 An Example Set of Words Derived From a Substantive Root

Turkish	**Components**	**English**	**Word class**
göz	*göz*	eye	Noun
gözlük	*göz* + *-lük*	eyeglasses	Noun
gözlükçü	*göz* + *-lük* + *-çü*	optician	Noun
gözlükçülük	*göz* + *-lük* + *-çü* + *-lük*	optician's trade	Noun
gözlem	*göz* + *-lem*	observation	Noun
gözlemci	*göz* + *-lem* + *-ci*	observer	Noun
gözle	*göz* + *-le*	observe	Verb (order)
gözlemek	*göz* + *-le* + *-mek*	to observe	Verb (infinitive

The main problem in our application is stemming the words. Stemming is the process for reducing inflectional or derived words in a language to a reduced form that may or may not be the morphological root of the words. It is not necessary that the stemmed words should give the morphological root of the word. It is sufficient that similar words math to similar stem. Eg. the words “call”, ”caller”, ”calls” should match to same stem ”call” (Sanyal, 2006).

Table 2.4 An Example Starting From a Verbal Root

Turkish	**Components**	**English**	**Word class**
yat-	*yat-*	lie down	Verb (order)
yatmak	*yat-mak*	to lie down	Verb (infinitive)
yatık	*yat- + -(ı)k*	leaning	Adjective
yatak	*yat- + -ak*	bed, place to sleep	Noun
yatay	*yat- + -ay*	horizontal	Adjective
yatkın	*yat- + -gın*	inclined to; stale (from lying too long)	Adjective
yatır-	*yat- + -(ı)r-*	lay down	Verb (order)
yatırmak	*yat- + -(ı)r-mak*	to lay down	Verb (infinitive)
yatırım	*yat- + -(ı)r- + -(ı)m*	laying down; deposit, investment	Noun
yatırımcı	*yat- + -(ı)r- + -(ı)m + -cı*	depositor, investor	Noun

Following example is detected according to one of the rules of hypernymy relation

“Ölüm, yangın, deprem vb. olayların yarattığı üzüntü, keder, elem”

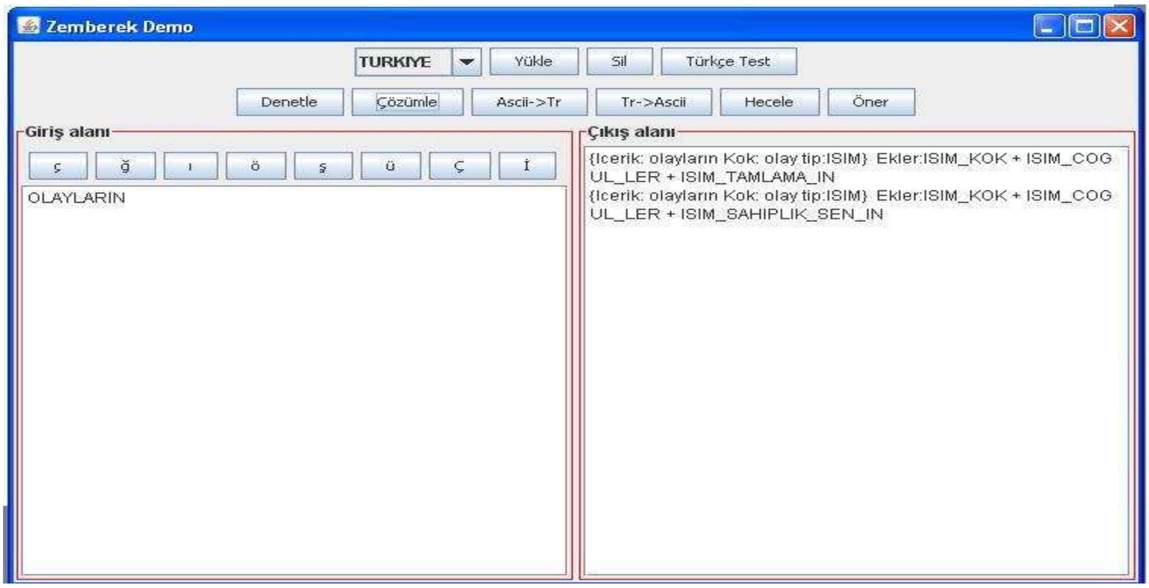
The hypernymy relation is found between the word pairs:

Hypernymy{“ölüm(death)”,”olayların(events’)”}

Hypernymy{“ yangın(fire)”,“olayların(events’)”}

Hypernymy{“ deprem(earthquake)”,“olayların(events’)”}

One of the related word “olayların (events’)” has some suffixes. Morphological analysis is needed to have the stem of the word. To achieve this process an open source, platform independent, general purpose Natural Language Processing library and toolset designed for Turkic languages Zemberek is used as shown in Figure 2.5.



```
OLAYLARIN
```

```
{Icerik: olayların Kok: olay tip:ISIM}  Ekler:ISIM_KOK + ISIM_COG
UL_LER + ISIM_TAMLAMA_IN
{Icerik: olayların Kok: olay tip:ISIM}  Ekler:ISIM_KOK + ISIM_COG
UL_LER + ISIM_SAHIPLIK_SEN_IN
```

Figure 2.5 Morphological Analysis of a Word

Zemberek has the ability to construct words. This is a simpler operation then Morphological parsing, it requires a root word object and a list of suffix objects. The system basically creates the suffixes and appends it after the formed word is shown in Table 2.5.

Table 2.5 Root and the Suffix List in Zemberek

1. {Icerik: olayların Kok: olay tip:ISIM} Ekler:ISIM_KOK + ISIM_COGUL_LER + ISIM_TAMLAMA_IN
2. {Icerik: olayların Kok: olay tip:ISIM} Ekler:ISIM_KOK + ISIM_COGUL_LER + ISIM_SAHIPLIK_SEN_IN

This list main contain many different roots, so it will be impossible to find the true root. Therefore the root of the beginning element of the list (Kok: olay) is accepted as a default root of the word. After this operation the new related word pairs are:

Hypernymy{“ölüm(death)”,”olay (event)”}

Hypernymy{“ yangın(fire)”,“olay (event)”}

Hypernymy{“ deprem(earthquake)”,“olay (event)”}

CHAPTER 3

RESULTS AND COMPARISON

This chapter will show the accuracy results of the automatic detection of word relations. The results in the tables below indicate that some relations are hard to be detected automatically from the definitions. Alternatively, one can also infer that the rules employed are not sufficient and some other rules are necessary for these types of relations. Additionally the accuracy of the results can be improved and the necessary rules can be easily obtained by increasing the rules of the relations. On the other hand, some relations can be completely or at least generally detected without further modifications and this is promising for some other types of relations.

3.1. PERFORMANCE MEASUREMENT

Performance of such systems is commonly evaluated using the data in the matrix. Confusion Matrix is a table with the true class in rows and the predicted class in columns (Kohavi and Provost, 1998). The diagonal elements represent correctly classified compounds while the cross-diagonal elements represent misclassified compounds. The Table 3.1 also shows the accuracy of the classifier as the percentage of correctly classified compounds in a given class divided by the total number of compounds in that class. The overall (average) accuracy of the classifier is also depicted (Hamilton, 2007).

The entries in the confusion matrix have the following meaning in the context of our study:

- A is the number of **correct** predictions that an instance is **negative**,
- B is the number of **incorrect** predictions that an instance is **positive**,

- C is the number of **incorrect** of predictions that an instance **negative**, and
- D is the number of **correct** predictions that an instance is **positive**.

Table 3.1 Confusion Matrix

		Predicted	
		Negative	Positive
Actual	Negative	**A**	**B**
	Positive	**C**	**D**

- The *Accuracy* (*AC*) is the proportion of the total number of predictions that were correct. It is determined using the equation:

$$AC = \frac{A + D}{A + B + C + D} \tag{3.1}$$

3.2. ACCURACY OF THE RESULTS

The accuracy of a measurement process is used to evaluate the performance of system.

Table 3.2 demonstrates that the total number of outputs that is obtained from our implementation by using extraction algorithms for the relations and accuracy of this implementation.

Table 3.3 results indicate that some rules are hard to be detected automatically. On the other hand, some rules can be completely or at least generally detected without further modifications and this is promising for some other types of generations

Table 3.2 Accuracy Results for Automatic Detection of Word Relations

Relation	**Total**	**Correct**	**Incorrect**	**AC(%)**
Antonymy	1962	1687	275	84
Synonymy	22124	21510	614	97
Kind Of	630	567	63	90
Amount Of	254	218	36	86
Group Of	421	303	118	72
Member Of	1026	831	195	81

Table 3.3 Number of Relationships Obtained According to Each Rule

Relation	**Rule1**	**Rule2**	**Rule3**	**Rule4**	**Rule5**	**Rule6**	**Rule7**
Antonymy	367	1595	-	-	-	-	-
Synonymy	6757	15367	-	-	-	-	-
Kind Of	12	32	586	-	-	-	-
Amount Of	167	45	42	-	-	-	-
Group Of	129	14	61	66	124	16	80
Member Of	37	805	66	118	-	-	-

The first column of Table 3.4 indicates the rules of the Hypernymy Relation. The second column points the total number of extracted relations from that rules. The columns named total and correct are used to calculate accuracy of each rule for the hypernymy relation.

For example, calculation of one of the rule “like” is :

$$AC = \frac{Correct}{Total} = \frac{544}{581} = 0{,}94 \qquad (3.2)$$

3.3. ERROR SOURCES

Experimental results show that automatic relation extraction of words in Turkish language is really difficult to be accomplished with high accuracy. Some of the sources of incorrect results are explained below.

3.3.1 Subordinative Conjuctions

Two nouns, or groups of nouns, may be joined to form subordinative conjuctions. In our relation extraction algorithm subordinative conjuctions are not considered while finding related words.

Table 3.4 Accuracy Results for Hypernymy Relation

Rule	**Total**	**Correct**	**Error**	**AC%**
Term	7115	7115	0	100
Person	1939	1939	0	100
Action	5453	5453	0	100
Science	58	52	6	90
Animal-Plant	72	64	8	89
Category	141	141	0	100
Colour	68	68	0	100
Element	38	33	5	87
Place	303	303	0	100
Equipment	49	48	1	98
Tool	70	70	0	100
Job	413	413	0	100
Nationality	125	124	1	99
Such as	3119	1560	1559	50
Like	581	544	37	94

In the following example according to Rule 3 of the Kind-of Relation the correct related word with “bal arısı” should be “eklem bacaklı”. Because of the difficulty of detection of the subordinative conjuctions in Turkish Language,they are not considered.

bal arısı: Zar kanatlılardan, bal yapan eklem bacaklı türü (Apis mellifica).

Kind-Of {“bal arısı (honeybee)”,” bacaklı (having legs)”}

3.3.2 Morphological Analysis of Zemberek

Some of the morphological analyses provided by Zemberek are detected as incorrect. There is an example below that shows this situation.

“Bir önceki cümleyle bağlantı kuran yani, demek ki, öyle ki vb. bağlayıcılarla başlayan, söz konusu duygu veya düşünceyi bütünleyen cümle.”

Hypernymy{“demek ki (scil)”,”bağlayıcılarla(with the connectives)”}
Hypernymy{“ yani (I mean)”,“ bağlayıcılarla(with the connectives)”}
Hypernymy{“ öyle ki (such that)”,“bağlayıcılarla(with the connectives)”}

The hypernymy relations show that the morphologic analysis is needed for the second related word “bağlayıcılarla(with the connectives)” as shown in Table 3.5 Morphological Analysis of Word “bağlayıcılarla”.

Table 3.5 Morphological Analysis of Word “bağlayıcılarla”

3. *{ Icerik: bağlayıcılarla Kok: bağla tip:FIIL} Ekler:FIIL_KOK + FIIL_TANIMLAMA_ICI + ISIM_COGUL_LER + ISIM_BIRLIKTELIK_LE*
4. *{ Icerik: bağlayıcılarla Kok: bağ tip:ISIM} Ekler:ISIM_KOK + ISIM_DONUSUM_LE + FIIL_TANIMLAMA_ICI + ISIM_COGUL_LER + ISIM_BIRLIKTELIK_LE*

The correct root of the word *“bağlayıcılarla”* should be *“bağlayıcı”*. After the morphologic analyse of Zemberek it is found as “bağla”.These incorrect relations can be corrected only manually by the experts.

Hypernymy{“demek ki (scil)”,”bağla (fixate)”}
Hypernymy{“ yani (I mean)”,“ bağla(fixate)”}
Hypernymy{“ öyle ki (such that)”,“bağla(fixate)”}

These incorrect relations can be corrected only manually by the experts as:
Hypernymy{“demek ki (scil)”,”bağlayıcı(conjuction)”}
Hypernymy{“ yani (I mean)”,“ bağlayıcı(conjuction)”}
Hypernymy{“ öyle ki (such that)”,“bağlayıcı(conjuction)”}

3.3.3 Conjunctions in Zemberek

Zemberek analyses some of the conjunctions like “ve, birçok, veya,ki….” as they are nouns. According to our kind-of algorithm the relation can be found only between the words has same genus. If the conjunctions are defined as nouns this becomes a problem while finding the related noun pairs

ağ mantarlar: İnsan ve hayvanlarda hastalığa yol açan ve birçok türü içine alan ilkel bitkiler topluluğu.

Kind-of{“ağ mantarlar”,”birçok”} is not correct.

CHAPTER 4

CONCLUSION

4.1 EVALUATION OF RESULTS

In order to model the knowledge acquisition, processing, usage and communication abilities of humans in computational domain to some extent, the simulation should be started from the smallest units of human learning mechanisms. Therefore, it is planned to study in the word level in the context of this project. Words are the fundamental building blocks of the communication, thinking, and decision making cognitive processes. While the learning process of words takes place, most of the information related to these words is also kept in the background.

Although, most commonly used dictionaries are transferred to the electronic environment and are utilized by information technologies in the last decade, they generally provide only the words and their definitions. However, various useful information and features about the words and relationships among them can not be represented and these can not be facilitated by many other applications. Storing the words along with their various features and relationships in a knowledge base, formation of a WordNet that allows demonstration of wide variety of relationships between words, and also to associate the words with their equivalent translations in the other languages for applications of multilingual environments are among the major goals of this study.

The design is implemented in such a way that it is flexible, scalable and trainable by humans and it is possible to imitate the dynamic learning and processing mechanism of human being in this manner.

In our application some formulas are defined for relating the words by using dictionary definitions as the starting point. These formulas are applied to the meaning of the words by using a computer program. All the related words and their relations that are handled from the program which we have done are stored in the files. The results indicate that some relations are hard to be detected automatically from the definitions.

On the other hand, some relations can be completely or at least generally detected without further modifications and this is promising for some other types of relations.

4.2 FUTURE IMPROVEMENTS

In the future work example sentences, pronunciation (as sound files or texts as international phonetic spelling), pictures, etc may be kept in our database. All useful features that are provided in traditional dictionaries will be brought together, and additionally, insertion of new words and definitions, description of different relationships between words and association of words by these predefined relations, automatic inference of new relationships by considering the interaction of the relations and demonstration of words structural changes in different time periods and geographical locations will be provided.

As the knowledge base enriched either by direct input or automatic detection new inferences or corrections on the automatic inferences will take place leading to an exponential growth in the data. Sample sentences will be kept as the source of a corpus, the words in this corpus will be morphologically analyzed and the sentences will be parsed, in the meanwhile, the semantic annotations will be protected by keeping the link between the words/their senses and the example sentences. The example sentences will form a corpus and provide an invaluable scalable resource that is needed but missing for the Turkish semantic applications.

An interface will be formed that provides acquisition and inputting of the information via internet and contribution of many people over it, however, the given information will be controlled before direct transfer to the knowledge base and only the approved data will be allowed to be processed. If there are new automatic inferences that can be acquired from the given information, these will also be added to it. The outcome can be presented to users in a visual and also text format.

As a result, although there are some researches about Turkish language in this area, this project has an importance as being the first one which has the special properties mentioned above for Turkish natural language processing field.

REFERENCES

Arnold, D., Balkan, L., Meijer, S., Humphreys, R.L., Sadler, L., Machine Translation: An Introductory Guide, Essex University, 1993.

Bariere, C., *From A Children's First Dictionary To A Lexical Knowledge Base of Conceptual Graphs*, Phd Thesis, Simon Fraser University, Canada, 1997.

Bilgin, O., Çetinoğlu, Ö., Oflazer, K., Morphosemantic Relations In and Across Wordnets: *A Preliminary Study Based on Turkish, in Proceedings of the Global WordNet Conference*, Masaryk, Czech Republic, January 2004, http://people.sabanciuniv.edu/~oflazer/balkanet/twn_tr.htm

Chomsky, N., Syntactic Structures, Mouton, The Hague, 1957.

Copeland, C., Durand, J., Krauwer, S., Maegaard, B., The Eurotra Formal Specifications, volume 2 of *Studies in Machine Translation and Natural Language Processing*. Office for Official Publications of the Commission of the European Community, Luxembourg, 1991.

Fellbaum, C., WordNet: An Electronic Lexical Database, The MIT Press, 1998.

Gordon, Raymond G., Jr. (ed.), Ethnologue: Languages of the World, Fifteenth edition. Language Family Trees - Altaic (HTML), 2005.

Hamilton H.J, “Confusion Matrix”, 2007, http://www2.cs.uregina.ca/~hamilton/courses/831/notes/confusion_matrix/confusion_matrix.html

Hutchins, J., The History Of Machine Translation In A Nutshell, IMP Magazine, 2005.

Isabelle, P., Machine Translation at the TAUM group, In M. King, editor, *Machine Translation Today: The State of the Art, Proceedings of the Third Lugano Tutorial,* Edinburgh University Press, Edinburgh,, 1987.

Jurafsky, D., Martin, J.H., *Speech and Language Processing*, New Jersey, 2006.

Kohavi R., Provost F., "Special Issue on Applications of Machine Learning and the Knowledge Discovery Process", *Machine Learning*, Vol.30, No.2-3, 1998.

Lewis, G, *Turkish Grammar*. Oxford University Press. ISBN 0-19-870036-9, 1996.

Maas, H.D., The MT system SUSY, In Margaret King, editor, *Machine Translation Today: The State of the Art*, *Proceedings of the Third Lugano Tutorial*, Edinburgh University Press, Edinburgh, 1987.

Mcdonald, D. D., *Natural Language Production as a Process of Decision Making*, Ph.D. thesis, MIT, Cambridge, MA, 1980.

Mckeown, K. R., Text Generation, Cambridge University Press, Cambridge, 1985.

Melby, A. K., A Suggestion Box Translator Aid, in: Proceedings of the annual symposium of the Deseret Language and Linguistic Society, Brigham Young University, Prove, Utah, 1981.

Miller, G., *WordNet*, 2005,
http://wordnet.princeton.edu/man/wngloss.7WN.html

Miller, G., EuroWordNet, 1999,
http://www.illc.uva.nl/EuroWordNet/

Nişanyan, S., "Etymological dictionary of Turkish Nişanyansözlük", 2007,
http://www.nisanyansozluk.com/

Orhan, Z., Önder, P., "Türkçe İçin İlişkili Bilgi Tabanı Ve Sözcük Ağı Geliştirilmesi", *VI. Uluslararası Türk Dili Kurultayı* , VI. Uluslararası Türk Dili Kurultayı Bildiri Kitabı, Ankara/Türkiye, 2008.

O'Reilly Media, What is Xml, 2009, http://www.xml.com/pub/a/98/10/guide0.html?page=2

Önder, P., Özen, N., Unlu, S., Orhan, Z, "A Framework for Building a Turkish Lexicon and Knowledge Base", *The 2008 International Conference on Information and Knowledge Engineering, Proceedings of IKE'08*, Monte Carlo Resort, Las Vegas, Nevada, USA , 2008.

Peters, W., Vossen, P., Díez-Orzas, P., Andriaens G., "Cross-linguistic Alignment of Wordnets with an Inter-Lingual-Index", *Computers and the Humanities*, Springer, 1998.

Pierce, J. R., Carroll, J. B., *Language and Machines - Computers in Translation and Linguistics (ALPAC Report)*. ALPAC, Washington D.C, 1966.

Sanyal R., "Unsupervised Machine Learning Approach to Word Stemming", *Project Guide in Indian Institute of Information Technology*, Allahabad, 2006.