ISTANBUL UNIVERSITY –
JOURNAL OF ELECTRICAL & ELECTRONICS ENGINEERING

YEAR : 2004 (1017-1023)
VOLUME : 4
NUMBER : 1

BUILDING A CORPUS BASED ADJECTIVE LEXICON FOR TURKISH

Yaşar ERENLER

Istanbul Technical University, Electrical-Electronics Faculty
Computer Engineering Department, 80626 Maslak, İstanbul

E-mail: erenler@cs.itu.edu.tr

ABSTRACT

This paper describes the design and construction of a lexical database for Turkish adjectives. We used a textual corpus of about one million running words that we collected from on-line newspapers and magazines available on the Internet. The lexicon contains syntactic category, semantic category, gradability, and thesaurus information about adjectives as well as selectional restrictions. It supports Natural Language Processing (NLP) applications such as parsing, text generation, natural language understanding, and information retrieval. It has been implemented as a relational database. The process of building the lexicon from the textual corpus has been performed semi-automatically using a series of extraction programs. We also implemented a Graphical User Interface to the lexicon.

Keywords: *Natural Language Processing, Corpus Linguistics, Computational Lexicography.*

1. INTRODUCTION

Observations [1] indicated that there is immediate need for Turkish Natural Language Processing (NLP) applications in diverse areas such as Machine Translation, Intelligent Tutoring Systems, Word Processing, Parsing, and Natural Language Interfaces to databases and operating systems. The lexicon is a central component in any NLP system and building a lexicon becomes a very important task. Besides several approaches such as using a machine-readable dictionary, or manual compilation, corpus based methods are also becoming very popular in recent years. Therefore, we decided to follow a corpus-based approach to build a lexicon to support Turkish NLP systems. Our lexicon is limited to the adjectives for now. This paper is based on a Ph.D. research [2] that was completed at the Illinois Institute of Technology.

Received Date : 06.06.2002
Accepted Date: 20.12.2003

1.1 Related Works

There has been extensive research in computational lexicography for the English language. One such project, The WordNet project [3,4], led by Miller et al. at Princeton University, has been a major lexicon project. WordNet is an on-line lexical database system in which nouns, verbs, and adjectives are organized into synonym sets, each representing one underlying lexical concept. Lexical-semantic relations such as synonymy, part-whole, and is-a relations have been the basis of this massive and comprehensive project.

Most previous work on the lexicon in both English and Turkish has focused on nouns and verbs. In the Turkish lexicon field, Yılmaz [5] worked on a verb lexicon and verb sense disambiguation of Turkish. Yorulmaz [6] has worked on a general purpose lexicon architecture for Turkish based on feature-based representation.

The scope of our work is Turkish adjectives. Our approach has been a corpus-based approach.

Extracting lexical information from a textual corpus and building a lexicon in this way has been our goal.

1.2 Types of Lexical Information

A basic lexicon would typically include information about morphology and syntax such as the complement structures of each word. Several researchers [7, 8, 9] have specified the information that a computational lexicon should contain. Their common items are: entry word, part of speech, definition, synonyms including cross-references, and examples.

2. CONTENT OF THE LEXICON

Our adjective lexicon contains syntactic and semantic information, selectional restrictions, and gradability information for each sense of the adjectives. It also includes synonymy and antonymy relations between the adjectives. Different NLP tasks emphasize different information in the lexicon. Our lexicon supports NLP tasks such as parsing, text generation, and information retrieval. Determining the structure and the kind of information to be stored is therefore a critical step.

2.1 Adjective Classifications

There have been several classifications of adjectives in Turkish, defined by different linguists. The common approach [10,11] is to classify the adjectives syntactically and also semantically. Table 1 and Table 2 show the semantic and syntactic classifications of adjectives in Turkish.

Table 1: Semantic Subcategories

Level 1	**Level 2**	**Level 3**
1. Qualificative	1. State/Status	
	2. Shape	
	3. Size	
	4. Color	
	5. Other	
2.Determinative	1. Demonstrative	
	2. Interrogative	
	3. Numeral	1. Cardinal
		2. Distributive
		3. Ordinal
		4. Fractional
	4. Indefinite	

Table 2: Syntactic Subcategories

Level 1	**Level 2**	**Level 3**
1. Simple		
2. Compound	1. Semantically Fused	
	2. Regulated	1. Suffixed
		2. Intensifiers
		3. Other
3. Derived	1. Noun Derived	
	2. Verb Derived	

2.2 Gradability

Gradability is a semantic feature that indicates whether an adjective can take comparative and superlative forms. Gradable adjectives can be used in comparative and superlative forms, as in the following examples:

küçük	*daha küçük*	*en küçük*
(small),	*(smaller),*	*(smallest);*
Mavi	*daha mavi*	*en mavi*
(blue),	*(bluer),*	*(bluest);*

However, not all adjectives are gradable. For example, adjectives denoting provenance such as ***Türk*** *kahvesi (**Turkish** coffee)* cannot be graded. Since gradability is useful information especially in text generation systems, we store this information in our adjective lexicon.

2.3 Selectional Restrictions

Some adjectives have selectional restrictions on what kind of nouns they can modify. The nouns might be Concrete or Abstract. Concrete nouns may be Human, Animate, or Inanimate nouns. Following are examples of nouns in these categories:

human: *şarkıcı (singer)*
animate: *aslan (lion)* inanimate: *masa (table)*
concrete: *ağaç (tree)* abstract: *fikir (idea)*

The adjective *yeşil (green)*, for example, can modify only a concrete noun. Therefore, the phrase ***yeşil*** *fikir (**green** idea)* does not make sense.

Conlon et al. [12] pointed out that selectional restrictions can be useful in disambiguation in natural language understanding. This information is of great importance also for both parsing and text generation systems in order to be able to parse and generate sentences that are semantically correct.

2.4 Lexical-Semantic Relations

Most dictionaries contain only synonyms and antonyms, so we consider these only. This kind of information would be very useful in information retrieval systems, or text generation systems.

Synonymy gives information about the words with similar meaning. This relation is observed generally in qualificative adjectives. Synonymy is a symmetric relation; that is, if A is a synonym of B, then B is a synonym of A. Examples of synonyms are:

tatlı (sweet) and *şekerli (with sugar)*;
eğlenceli (amusing) and *gülünç (funny).*

Antonymy gives information about the words with opposite meaning. It is the most widely encountered lexical-semantic relation for adjectives. Almost all qualificative adjectives have an antonym. It is also a symmetric relation; that is, if A is the antonym of B, then B is the antonym of A. Examples of antonyms are:

yeni (new) and *eski (old);*
uzun (tall) and *kısa (short).*

Taxonomy is a relation between two words that are in the same part of speech category, and gives information regarding the taxonomical relation of those two words. For example, it can be used between two nouns: *a horse ISA animal*; or between two verbs: *to walk ISA to move*. Since nouns are most often defined in terms of other nouns and verbs in terms of other verbs, it makes taxonomy a natural relation to use. Adjectives, on the other hand, are rarely defined in terms of other adjectives, so the taxonomy relation is not used for adjectives.

Similarly, Part-Whole, Attribute, Collocation, and Grading relations also can be omitted for adjectives.

3. IMPLEMENTATION AND BUILDING PROCESS

We implemented our adjective lexicon as a relational database. A relational database has several advantages because the data model is independent of how data is stored and accessed. So, we decided to use the Microsoft Access relational database system. The tables in the database and their fields are as follows. Keyfields are underscored.

MAIN
(WORDKEY, WORD, SENSE-NUMBER, ENGLISH_DESCRIPTION)

SELECT_REST_GRAD
(WORD, SENSE-NUMBER, GRADABLE, HUMAN, ANIMATE, INANIMATE, ABSTRACT, CONCRETE)

SYNTACTIC_CATEGORY
(WORD, SENSE-NUMBER, SYN_SUB1, SYN_SUB2, and SYN_SUB3)

SEMANTIC_CATEGORY
(WORD, SENSE-NUMBER, SEM_SUB1, SEM_SUB2, SEM_SUB3)

SYNONYMY
(WORD, SENSE-NUMBER, SYN_WORD, SYN_SENSE_NUMBER)

ANTONYMY
(WORD, SENSE-NUMBER, ANT_WORD, ANT_SENSE_NUMBER)

3.1 Building the Adjective Lexicon

Our approach to build the lexicon was extraction of lexical information from widely available textual corpora, building the lexicon semi-automatically. We have collected a corpus of one million running words from on-line resources such as daily newspapers and magazines available on the Internet. Certain kinds of information about adjectives can be found better in text. Since adjectives are used in a specific context within a text, we can get an idea of different sense uses. Information such as gradability, selectional restrictions, and antonymy can usually be found more easily in a text. For similar reasons, corpus based computational linguistics research has gained a lot of attention in recent years.

3.2 Textual Corpus as a Resource

In order to collect Turkish texts, we captured daily news from on-line newspapers, and weekly news from on-line magazines and saved all text files on a Personal Computer. All of the texts are in ISO-8859-9 encoding. The newspapers have similar sections such as headlines, politics, economy, world news, art, sports, and commentaries.

After collecting a large textual corpus, we electronically sent some of our text to the director of the TU-LANG Project at the Bilkent University, and they very kindly put part-of-speech (POS) tags on the adjectives for us. Then we extracted the words that were marked as the adjectives and imported them to the lexicon database. Figure 1 shows a sample of the POS tagged text.

We extracted approximately 140,000 adjectives from the parsed files, which have one million running words in total. In this corpus, the total number of unique adjectives is about 2,209. In other words, there were 140,000 adjective tokens and 2,209 adjective types.

```
yapmış
1. yap+Verb+Pos+Narr+A3sg
2. yap+Verb+Pos+Narr^DB+Adj+Zero

olduğu
1. ol+Verb+Pos^DB+Adj+PastPart+P3sg
2. ol+Verb+Pos^DB+Noun+PastPart+A3sg
   +P3sg+Nom

bu
1. bu+Det
2. bu+Pron+DemonsP+A3sg+Pnon+Nom

açıklamalarla
1. açıkla+Verb+Pos^DB+Noun+Inf+A3pl
   +Pnon+Ins

birlikte
1. bir+Num+Card^DB+Noun+Ness+A3sg
   +Pnon+Loc
2. birlik+Noun+A3sg+Pnon+Loc
3. birlikte+Adv
4. birlikte+Postp+PCIns
```

Figure 1 : Sample POS Tagged Text

3.3 KWIC Indexing

We created a KWIC (Key Word In Context) index for our corpus. It provides us with a list of all words and the sentences in which they appeared. KWIC indexing is a useful tool to analyze word frequencies, and word usage in different contexts. It permits efficient word search for human users. By accessing different usages of a given word via this index, one can analyze and determine part of the speech semi-automatically. A KWIC index is also useful for determining selectional restrictions of an adjective.

While we examine an adjective in context using the KWIC index, there is a high chance that an antonym of that adjective will be in the same sentence or possibly in the next one. Justeson and Katz [13] studied the co-occurrences of antonymous adjectives and their contexts for the English language. They did an empirical study of the Brown Corpus, which has one million words of English text, and found that antonymous adjectives exist in the same sentence very frequently. For example, *far* and *near*, *cold* and *hot* often appeared in the same sentence to create contrast. There is a similar trend in Turkish texts too. We make use of this observation to discover antonymy relationships automatically, rather than sorting out all adjectives and finding all antonyms manually.

3.4 Automating the Extraction of Adjectives

The process of extracting the adjectives from the text is automated to a certain degree. Syntactically, adjectives in Turkish are classified in three subgroups: Simple, Compound, and Derived. All of the derived adjectives are created by adding certain suffixes to nouns or to verbs. It is, therefore, possible to identify derived adjectives by doing a morphological analysis of given words. For example, in the word *mavimsi (blueish),* the suffix **-imsi (-ish)** indicates that this is a derived adjective. But, it is not possible to identify the word *mavi (blue)* as a color adjective without human knowledge, so this kind of non-derived adjective had to be identified manually.

The intensifier adjectives are also discovered automatically by looking for repeat patterns (syllable) on both sides of an intensifier letter (M, P, R, or S). For example the adjective *çabuk (quick)* is intensified with the letter R as in the adjective *çarçabuk (quickly).*

The suffixes shown in Table 3 are used to determine and extract the adjectives that are derived from a noun or a verb. For the automatic extraction of adjectives that are identified by the suffixes, we wrote a C program . This program takes a source text file as input, examines each word by trying a match with the suffixes, and if successful, writes those matched words to an output file. Besides the adjective word itself, the program also outputs the syntactic classification information, telling whether the adjective is simple, derived, or compound. All variants of vowel harmony in suffixes are included in the suffix arrays.

Table 3: Adjective Suffixes for Automatic Extraction

NOUN DERIVED	
Suffix	**Example**
ce	*dostça (friendly)*
ci	*kitapçı (bookseller)*
cik	*küçücük (tiny)*
cil	*bencil (egoist)*
(in)de	*yerinde (proper)*
den	*toptan (wholesale)*
(i)msi	*ekşimsi (sourish)*
(i)mtrak	*sarımtırak (yellowish)*
(de)ki	*dünkü (yesterday's)*
li	*yağmurlu (rainy)*
lik	*kiralık (for rent)*
man	*şişman (fat)*
sel	*bilimsel (scientific)*
si	*çocuksu (childish)*
siz	*şekersiz (sugarless)*
î	*millî (national)*
ik	*ekonomik (economical)*
el	*potensiyel (potential)*
VERB DERIVED	
Suffix	**Example**
ç	*gülünç (laughable)*
dik	*tanıdık (known)*
ecek	*gelecek (upcoming)*
(e)k	*parlak (shining)*
(i)k	*patlak (exploded)*
(y)en	*sevilen (loved)*
gen	*unutkan (forgetful)*
gin	*yorgun (tired)*
ici	*besleyici (nutrious)*
me	*kestirme (shortcut)*
miş	*pişmiş (cooked)*
r	*döner (rotating)*
z	*paslanmaz (stainless)*

4. IMPLEMENTATION OF A GRAPHICAL USER INTERFACE FOR THE LEXICON

We implemented a GUI program for the users of the lexicon. The lexical database and the GUI program can be used as an electronic desktop dictionary / thesaurus for the Turkish adjectives. Of course, the database can also be utilized by any NLP system directly. The program allows a user to view, add, delete, or modify any adjective entries, synonymy and antonymy relations, and all other syntactic and semantic information.

Figure 2 is the main user interface for the lexicon. In Figure 3 and Figure4, snapshots of syntactic and semantic categories for the adjective *"aceleci"* are shown.

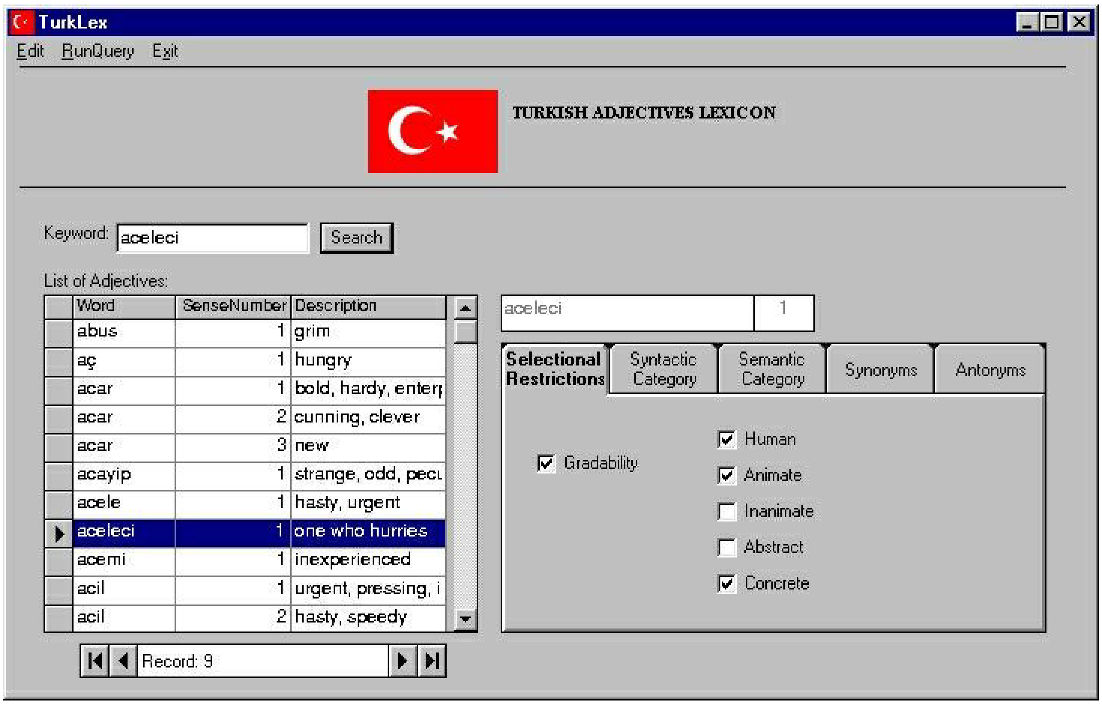
5. CONCLUSION

In this paper, we have described the design and construction of a lexicon for Turkish adjectives. This lexicon is designed to function as a central part of any NLP system, with support for a variety of applications such as parsing, text generation, natural language understanding and information retrieval.

The lexicon is implemented as a relational database using the Microsoft Access database management system. For each of its 2400 adjectives it contains syntactic category information, semantic category information, gradability information and information on selectional restrictions. It also contains thesaurus information about synonyms and antonyms of adjectives. We have implemented a graphical user interface program to allow users to browse, search, extend, and modify the lexical database.

The source of the information included in the lexicon is the corpus that we have collected from the Internet of articles and editorials from on-line newspapers and magazines. Currently, it contains 1,200,000 words of running text.

Once new text was downloaded we added it to our KWIC (Keyword in Context) index of the adjectives in the corpus, which contains each word in the corpus in alphabetical order along with a line of material from each sentence in which that word appears.



Keyword: aceleci

List of Adjectives:

	Word	SenseNumber	Description
	abus	1	grim
	aç	1	hungry
	acar	1	bold, hardy, enter;
	acar	2	cunning, clever
	acar	3	new
	acayip	1	strange, odd, pecu
	acele	1	hasty, urgent
▶	aceleci	1	one who hurries
	acemi	1	inexperienced
	acil	1	urgent, pressing, i
	acil	2	hasty, speedy

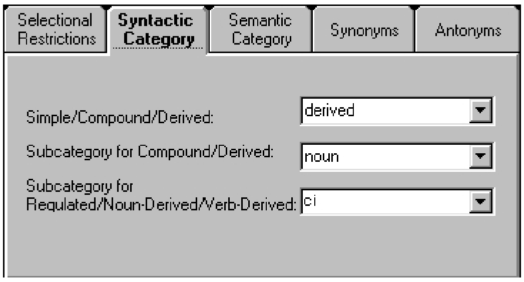
Record: 9

aceleci 1

☑ Gradability

☑ Human
☑ Animate
☐ Inanimate
☐ Abstract
☑ Concrete

Figure 2 : Main User Interface for Lexicon

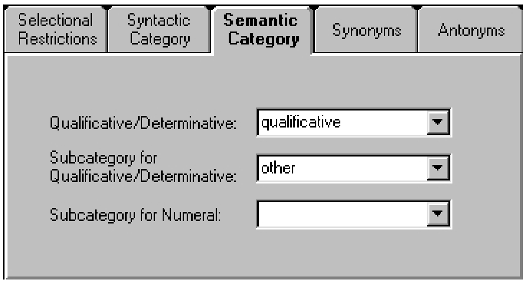


Simple/Compound/Derived: derived

Subcategory for Compound/Derived: noun

Subcategory for Regulated/Noun-Derived/Verb-Derived: ci

Figure 3 : Snapshot of Syntactic Category



Qualificative/Determinative: qualificative

Subcategory for Qualificative/Determinative: other

Subcategory for Numeral:

Figure 4 : Snapshot of Semantic Category

REFERENCES

[1] Oflazer, K., Bozşahin, C. 1994. "Turkish Natural Language Processing Initiative: An Overview", In *Proceedings of the Third Turkish Symposium on Artificial Intelligence and Neural Networks*, Ankara, June 1994.

[2] Erenler, Y. 1999. "Designing and Building an Adjective Lexicon for Turkish Based on a Million Word Corpus", Ph.D. Thesis,. *Illinois Institute of Technology*, Department of Computer Science, Chicago, USA.

[3] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K., 1990. Introduction to WordNet: An On-line Lexical Database, In *Five Papers on WordNet, CSL Report 43*, Princeton University, Princeton, NJ.

[4] Fellbaum, C., (ed.), 1999. WordNet "An Electronic Lexical Database", The MIT Press.

[5] Yılmaz, O. 1994. Design and Implementation of a Verb Lexicon and a Verb Sense Disambiguator for Turkish. M.S. Thesis, *Bilkent University,* Department of Computer Engineering and Information Science, Ankara.

[6] Yorulmaz, A. K. 1997. Design and Implementation of a Computational Lexicon for Turkish. M.S. Thesis, *Bilkent University,* Department of Computer Engineering and Information Science, Ankara.

[7] Apresyan, Yu. D., Mel'cuk, I.A., and Zolkovsky, A. K. 1970. "Semantics and Lexicography: Towards a New Type of Unilingual Dictionary", In F. Kiefer, (ed.) *Studies in Syntax and Semantics*. Reidel, Dordrecht, Holland, 1-33.

[8] Meijs, W. 1992. "Computers and Dictionaries", In Butler, C. S., (ed). *Computers and Written Texts*, Basil Blackwell Ltd., Surrey UK. 141-165.

[9] Smith, R. N. Maxwell, E. 1973. "An English Dictionary for Computerized Syntactic and Semantic Processing", In Zampolli, A. and Calzolari, N., eds. *Computational and Mathematical Linguistics*, Florence, Italy: Olschki, 1977. 303-322.

[10] Ediskun, H. 1996. "Türk Dilbilgisi (Turkish Grammar)", *Remzi Kitabevi*, 5th edition, Istanbul.

[11] Koç, N. 1990. "Yeni Dilbilgisi (New Grammar)", *İnkılap Kitabevi*, Istanbul.

[12] Conlon, S. Pin-Ngern, Evens, M., and Ahlswede, T. 1990. "Generating a Lexical Database for Adverbs", In *Proceedings of the University of Waterloo Centre for the New Oxford English Dictionary*, 28-30, 95-109.

[13] Justeson, J. S., Katz, S. M. 1991. "Co-occurrences of Antonymous Adjectives and Their Contexts", *Journal of Computational Linguistics*, Vol. 17, Num. 1, 1-19.