A LEXICON FOR IDIOMATIC COMPOUNDS IN TURKISH

KADRİYE ELİF EYIGÖZ

BOĞAZİÇİ UNIVERSITY

SEPTEMBER 2007

A Lexicon for Idiomatic Compounds in Turkish

Thesis submitted to the Institute of Social Sciences
in partial satisfaction of the requirements for the degree of

Master of Arts
in
Cognitive Science

by
Elif Eyigöz

Boğaziçi University
September 2007

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

Assist. Prof. Dr. Mine Nakipoğlu Demiralp
Co-advisor

Prof. Dr. Cem Say
Co-advisor

Assist. Prof. Dr. Tunga Günör

Assist. Prof. Dr. Meltem Kelepir

Prof. Dr. A. Sumru Özsoy

Abstract

A Lexicon for Idiomatic Compounds in Turkish

Elif Eyigöz

This work presents and comprises a constraint-based case-frame lexicon for idiomatic compounds headed by verbs in Turkish. The lexicon covers ten Turkish verbs with the highest number of senses to be used in natural language processing applications for representing and resolving senses of idiomatic compounds. This thesis gives detailed instructions, suggests conventions and describes a structure for organizing the data in the lexicon. It also provides a sample lexicon for ten verbs organized according to the structure proposed, in order to form a guideline for future lexicographic work based on this study.

Tez Özeti

Türkçe Deyimsel Tamlamalar Sözlüğü

Elif Eyigöz

Bu çalışma, doğal dil işleme uygulamalarında, fiil başlı deyimsel tamlamaların anlamlarını belirlemek ve temsil etmek amacıyla kullanılmak üzere sınırlama-tabanlı ve isimlerin hallerine dayalı bir sözlük tasarımı önermektedir. Önerilen tasarım, Türkçe'de en çok anlamı olan ilk on fiil için hazırlanmış bir sözluk ile örneklenmiştir. Bu teze dayanarak yapılacak sözlük derleme çalışmalarında izlenmek üzere, verinin düzenlenmesi için sözlüğun yapısı ve kodlama kuralları belirlemiş, detaylı direktifler verilmiştir.

Acknowledgements

I would like to express my gratitude to my thesis supervisor Ass. Prof. Mine Nakipoğlu Demiralp for her guidance, patience, and contribution. I am also deeply thankful to Prof. Sumru Özsoy for sharing her time and ideas to improve this study. It would be impossible to finish this work without their support.

Table of Contents

List of Tables

CHAPTER 1

INTRODUCTION

1.1 Aim

This work presents and comprises a constraint-based case-frame lexicon for idiomatic compounds headed by verbs in Turkish. The lexicon covers ten Turkish verbs with the highest number of senses (meanings) and it will be used in natural language processing applications for representing and resolving senses of idiomatic compounds. Idioms pose a problem for natural language processing as their meanings cannot be predicted from the meanings of their parts. Thus, preparation of lexicons that include particular meanings associated with idioms is necessary.

The main proposal for a constraint-based lexicon for idiomatic compounds has been made by Oflazer and Yılmaz (1995), and possessive markers on the non-heads and control properties of idiomatic compounds has been first studied by Kartal (1995). This thesis not only adopts the constraint-based structure offered by Oflazer and Yılmaz but also extends this structure through Kartal's observations on syntactic, morphological and control properties of Turkish idiomatic compounds. This thesis gives detailed instructions, suggests conventions and describes a structure for organizing the data in the lexicon. It also provides a sample lexicon for ten verbs organized according to the structure proposed, in order to form a guideline for future lexicographic work based on this study.

1.2 Overview of the Thesis

The thesis is structured as follows: The selection of the data in the sample lexicon is justified in Section 1.3. This selection is independent of the structure of the lexicon, as

the same structure can be used for compiling non-idiomatic noun-verb pairs as well. Thus, this selection by no means has to be followed in future lexicographic work.

An introduction to Oflazer and Yılmaz's proposal for a constraint based case-frame lexicon is given in Section 1.4. The detailed examples related to this section are postponed until Chapter 4, where a description of the structure of the lexicon is given based on Oflazer and Yılmaz's proposals and Kartal's observations. The structure of the lexicon is anticipated to have a spreadsheet file format (e.g. Microsoft Office Excel file) that allows the users to enter data into a table with rows and columns. Examples, instructions, tests and conventions for organizing the data in such a spreadsheet file are given in Chapter 4.

To compile the semantic data, I have resorted to the use of certain lexical resources. These resources are described in Chapter 2. Chapter 3 summarizes Kartal's observations of syntactic, morphological and control properties of Turkish idiomatic compounds. Systematic tests for determining the control properties of the compounds are developed in this chapter. After presenting the structure of the lexicon in Chapter 4, in Chapter 6 I will summarize the results, observations and make suggestions for lexicographers. Finally, an electronic database presents the sample lexicon for ten Turkish verbs.

1.3 Data

The lexicon covers the idiomatic compounds headed by ten Turkish verbs with the highest number of senses. Table 1.1 shows the verbs, the number of senses associated with these verbs in the Turkish-Turkish dictionary of *Türk Dil Kurumu* The Turkish Language Foundation, the number of idiomatic compounds headed by these verbs in the lexicon offered in this study, and the English translation of the most frequently used sense. The number of senses has been used as an heuristic in choosing the verbs, in order to compile a significant number of idiomatic compounds. However, as Table 1.1 shows, the relationship between the number of senses and the number of idiomatic compounds is not proportional for the first ten verbs. The verb *geç* has 38 senses but has only 10 idiomatic compounds associated with it in the lexicon. Moreover, the verb with the highest number of senses

Table 1.1: The verbs and the number of senses

Verbs	Senses	Compounds	English
çık	57	32	leave
tut	50	33	hold
çek	46	28	pull
al	35	51	take
gel	38	50	come
geç	38	10	pass
at	37	34	throw
düş	32	21	fall
aç	28	23	open
vur	28	16	hit
Total	297	389	

cık has 57 senses and is associated with 32 idiomatic compounds in the lexicon, but the verb *gel* which also has 38 senses like *geç* is associated with 50 idiomatic compounds. Nevertheless, one of the important uses of this lexicon is word-sense disambiguation; and choosing the verbs according to the number of senses serves this purpose well.

It is interesting to note that all of the verbs in Table 1.1 are motion verbs, which are known to have creative usage in metaphors and have been widely studied by psychologists, linguists and computer scientists (Tenny, 1995).

This study is only concerned with a description of idiomatic compounds formed with these verbs, hence excludes noun-verb pairs which appear to be non-idiomatic. For example, some noun-verb pairs formed by these verbs are non-idiomatic, yet they have one word English translations. For example, one sense of the verb *tutmak* is (to form in), as exemplified in *buz tutmak* (to form ice in), *pas tutmak* (to form rust in), and *küf tutmak* (to form mold in), which is translated as *to freeze*, *to rust* and *to mold* respectively. As such pairs are not idiomatic, they are excluded from the lexicon offered in this study. If such pairs which are more predictable as far as their senses go were included in the lexicon, the number of entries would increase enormously and would make this study go beyond what it aims for.

Another type of non-idiomatic pair that has not been included in this lexicon is the one that requires a pair specific translation. For example, *çıkmak* should be translated as 'to renounce' in *dinden çıkmak* (to renounce religion) only with the complement *religion*. This sense of *renounce* is usually associated with *vazgeçmek* in Turkish (to give up), but in Turkish *dinden çıkmak* is used instead of *dinden vazgeçmek*.

As will be laid out in the ensuing sections, the two types of non-idiomatic pairs exhibit similar syntactic and morphological restrictions when compared with the idiomatic ones. Thus in future work, the structure of the lexicon described here can be extended to cover non-idiomatic compounds as well.

1.4 A Constraint-based Case-frame Lexicon

Verbs convey several senses when they are used in certain syntactic, morphological, lexical, and semantic frames. Accordingly, the basic idea of frame semantics is that one cannot understand the meaning of a verb without access to all essential knowledge related to that verb (Fillmore, 1968; Fillmore, Baker 1998). For example, one would not be able to understand the senses of the verb "sell" in the context of a commercial transfer, without knowing anything about the situation of commercial transfer, which involves a seller, a buyer, goods, money, the relation between the money and the goods, the relations between the seller and the goods and the money, and so on. Thus, the meaning of a verb is jointly determined by the meanings of and the relations between its arguments and adjuncts. Therefore, semantic properties of the arguments and adjuncts of verbs, i.e. frame elements must be included in lexicons as well.

Studies in English frame semantics is conducted by the Berkeley FrameNet. Moreover, similar studies have been conducted by German, Japanese and Spanish FrameNet projects. As for Turkish, one study which aims to capture the insights of frame semantics is offered by Oflazer and Yılmaz (1995). Oflazer and Yılmaz present a constraint-based case-frame lexicon architecture for Turkish. The crucial aspect of the case-frame is that argument and adjuncts in a sentence are denoted by the case markers they carry. Oflazer and Yılmaz list the following categories of constraints in the case-frame of a verb:

(1.1)
1. Semantic constraints on the arguments (e.g. thematic roles, frame roles).
2. Constraints of syntactic and morphological features on the arguments and adjuncts (e.g. case, possessive markers, agreement, phrase type, verb form).
3. Constraints on argument co-occurrence (e.g. obligatory, optional, impossible).
4. Constraints on verb features (e.g. voice, agreement).

These constraints are employed in the design of the lexicon; detailed examples are given in Chapter 4 where the structure of the lexicon is explained. In general, constraints 2, 3 and 4 i.e. syntactic and morphological constraints are language internal. They are described and determined by the lexicographer's linguistic knowledge of Turkish. On the contrary, semantic constraints are language independent. They are compiled from the semantic resources utilized in this study, which will be introduced in the next section.

CHAPTER 2

SEMANTIC RESOURCES

The senses (meanings) in the lexicon are mapped to WordNet definitions. Semantic frame elements are compiled from VerbNet and FrameNet. Mapping of semantic elements to English sources renders the lexicon as a bilingual dictionary to be used in machine translation applications. Moreover, as WordNet, VerbNet and FrameNet databases have been lingua franca for many languages such as German, Japanese, Spanish, and Balkan languages, mapping Turkish to them connects Turkish to these languages as well. With respect to semantic constraints, the lexicon relies heavily on these lexical resources and needs to be constantly updated as these resources are updated.

2.1 WordNet

WordNet is a semantic lexicon for English. It was created and is being maintained at the Cognitive Science laboratory of Princeton University. WordNet covers the vast majority of nouns, verbs, adjectives, and adverbs of English (Miller, 1990). The words in WordNet are organized as synonym sets, called synsets. (2.1) shows the list of senses associated with the verb *handle*. There are six senses, each of which is associated with a synset given above the description of the sense.

(2.1) 1. {handle#1, manage#2, deal#7, care#4}
be in charge of, act on, or dispose of

2. {handle#2, treat#1, do by#1}
interact in a certain way

3. {handle#3, cover#5, treat#4, plow#2, deal#1, address#8}

act on verbally or in some form of artistic expression

4. {handle#4, palm#1}
   touch, lift, or hold with the hands
5. {handle#5, wield#2, manage#7}
   handle effectively
6. {handle#6}
   show and train

Every synset includes an instance of *handle* followed by the sense number and other verbs that share the same synset. The synset {handle#1, manage#2, deal#7, care#4} is associated with the definition *be in charge of, act on, or dispose of*. This definition holds for the first sense of the verb *handle*, the second sense of the verb *manage*, the seventh sense of the verb *deal* and the fourth sense of the verb *care*. The order of senses reflects the frequency of usage.

Each synset thus represents a sense and provides a short, general definition, and records the various semantic relations between these senses including hypernymy/ hyponymy (IS-A), meronymy/ holonymy (HAS-A), antonyms, entailment etc. As of 2006, the database contains 150,000 words organized in over 115.000 synsets for a total of 207.000 word-sense pairs.
(http://WordNet.Princeton.edu/man/wnstats.7WN)

2.2 FrameNet

Based on frame semantics, the Berkeley FrameNet Project has been creating an online lexical resource for English, which is supported by corpus evidence. It documents semantic and syntactic combinatory possibilities of each word in each of its senses.(Baker et al., 2003) It is based on a theory of frame semantics and sentence level ontology.

In frame semantics, a *frame* corresponds to a scenario that involves an *interaction* and its *participants* with certain roles. A frame has a name, which is used to identify the semantic relation that groups together the semantic roles. For example, the verb *buy*

evokes the frame *commerce_ buy*, the elements of which are given in (2.2). The non-core elements are shared by other frames and the full list of non-core elements is not given here. The italic words refer to the elements in the frame.

Commerce_ buy

(2.2)

- **Core Elements**

  Buyer The *Buyer* wants the *Goods* and offers *Money* to a *Seller* in exchange for them.

  Ex: *Jess* bought a coat.

  Goods The frame element *Goods* is anything (including labor or time, for example) which is exchanged for *Money* in a transaction.

  Ex: Only one winner purchased *the paintings*.

- **Some non-core elements**

  Seller *The Seller* has possession of the *Goods* and exchanges them for *Money* from a *Buyer*.

  Ex: Most of my audio equipment, I purchased *from a department store near my apartment*.

  Money *Money* is the thing given in exchange for *Goods* in a transaction.

  Ex: Sam bought the car for *$12,000*.

  Means The means by which a commercial transaction occurs.

  Ex: Will they allow you to purchase *by check*?

  Purpose of goods *The Buyer's* intended purpose for the *Goods*.

  Ex: I purchased the calculator *for easier calculation of my debts*.

Currently, FrameNet defines about 3040 verbs attached to 310 different frames. The knowledge provided by FrameNet is used to identify the frames and semantic roles in the lexicon presented here. The incompleteness of sources like FrameNet calls for the need to update the lexicon presented here.

2.3 VerbNet

VerbNet is a verb lexicon compatible with WordNet. It was created at the University of Pennsylvania. It has explicitly stated syntactic and semantic information based on Levin's (1993) verbs classes. VerbNet associates the semantics of a verb with its syntactic frames: it combines traditional lexical information such as thematic roles with syntactic selectional restrictions (Kipper et al., 2000). (2.3) shows the entry for the verb *handle*. The entry for *handle* lists only the frame for the fourth sense of *handle*: touch, lift, or hold with the hands (handle, palm). It lists one class (hold-15.1) and three frames for the fourth sense of *handle*.

(2.3)
- **Verb class:** hold-15.1 Basic Transitive

  She *held* *the rail.*
  Agent V Theme
- **Verb class:** hold-15.1 Body-Part Possessor Ascension Alternation

  She *held* *him* *by* *the arm.*
  Agent V Theme Prep(by) Oblique
- **Verb class:** hold-15.1-1 Transitive (Body-part Object)

  She *held* *his arm.*
  Agent V Theme

The information from VerbNet will be relevant mostly in terms of semantics, namely in compiling traditional thematic roles and verb classes. Just as FrameNet, VerbNet is also incomplete. This calls for the constant update of the lexicon as VerbNet is updated.

2.4 TDK, Zargan and Ekşisözlük

Turkish resources are used mainly for compiling the idiomatic compounds and secondarily for inquiring for the meanings of idiomatic compounds. Senses have been compiled from the online Turkish-Turkish dictionary *Güncel Türkçe Sözlük* (Contemporary Turkish Dictionary), which was created and is maintained by the *Türk Dil Kurumu* The Turkish Language Foundation, and the Turkish-English online dictionary www.Zargan.com. Idiomatic compounds have been compiled from the *Güncel Türkçe Sözlük*, and online in-

formal dictionary of Ekşisözlük (www.sourtimes.org). Searching with wildcards such as "*" has been useful in accessing the compounds associated with the verbs.

CHAPTER 3

PROPERTIES OF IDIOMATIC COMPOUNDS IN TURKISH

Turkish idiomatic compounds headed by verbs participate in various syntactic and morphological constructions. The following section presents the syntactic, morphological and control properties of idiomatic compounds that are pertinent to the practical aims of this study.

3.1 Some Background on Turkish

Göksel and Kerslake (2005) categorize Turkish verbs into transitive and intransitive verbs. The verbs which take a direct object complement are transitives, the verbs which take dative, ablative or instrumental/ comitative marked complements and the verbs which do not take complements at all are intransitives.

Direct objects in Turkish are usually non-case marked if they are indefinite or categorial, they are marked by accusative if they are definite. (3.1a) exemplifies an indefinite direct object *mektup*, the determiner position of which is filled by the indefinite article *bir*. (3.1b) exemplifies a categorial direct object *film*, the determiner position of which is empty.[1] Finally, the definite direct object *parça* in (3.1c) bears an accusative case marker.

(3.1) a. *bir mektup yaz-dı-m*
a letter write-PAST-1SG
'I wrote a letter'

b. *Onlar film seyred-er-di.*
they film watch-AOR-PAST
'They used to watch movies'

[1]Categorial direct objects cannot be modified by any determiners. They cannot be plural-marked. When functioning as subjects they have to be in the immediately preverbal position. When functioning as direct objects, they do not receive accusative case unless topicalized (Göksel and Kerslake, p. 377).

c. *Bu parça-yı çok iyi bir piyanist çal-abil-ir.*
This piece very good a pianist play-ABIL-AOR
'A very good pianist can play this piece.'

Just as the direct object, there is no one-to-one relation between the subject and the case markers it may carry. Subjects are marked with nominative case in finite sentences and usually marked with genitive case in nominalized and relative clauses. Because of the special status of the direct object and the subject with respect to case marking properties, Oflazer and Yılmaz include them in the case-frame as *subject* and *direct object*. The dative, ablative and instrumental/ comitative case marked objects however have not been described according to their grammatical functions. The whole list of the case-frame elements are shown in Table 3.1.

Table 3.1: Case-Frame Elements

Subject	Sub
Direct Object	Dir
Dative	DAT
Ablative	ABL
Instrumental/ comitative	INS

The dative, ablative and instrumental/ comitative case marked complements are categorized as *oblique arguments* by Göksel and Kerslake. However, the case-frame structure offered by Oflazer and Yılmaz refers to these elements only by their case markers, it does not make use of the more general categorization of *oblique argument*. As dative, ablative and instrumental/ comitative cases may also mark adjuncts, the case marker on the element itself is not sufficient to classify the element as an argument or an adjunct. The argument-adjunct status of an element is represented in the case-frame by the cooccurence constraints *optional* and *obligatory*: Arguments are coded as obligatory and adjuncts are coded as optional elements of the case-frame. Therefore, although Göksel and Kerslake's conception of the *oblique argument* as 'a dative, ablative or instrumental/ comitative case marked argument' is not represented in the lexicon as such, it is derivable from the lexicon. Likewise, the categorization of verbs in Göksel and Kerslake's sense as transitives

and intransitives, with respect to their ability to take a direct object complement, is also derivable from this lexicon.[2] The lexicon includes the properties of the verbs and their arguments in a relatively theory neutral way due to its practical aims. Moreover, this renders higher order linguistic concepts such as *oblique argument* derivable from this lexicon as essential parts of their definitions are included in the lexicon. Therefore, such higher order classifications are omitted, unless they prove themselves to be necessary. They are necessary only if they are required to explain empirical facts that are not rule governed or exceptions to the rules, both of which must be represented in the lexicon.

Argument Adjunct Distinction

Since the argument-adjunct status of a dative, ablative or instrumental/ comitative case marked element may not always be straightforward, I follow Göksel and Kerslake's adverbial definition for categorizing adjuncts.[3] Göksel and Kerslake suggest that the adjuncts modify the verb by describing:

1. Its destination, or target (indicated by the dative case marker)

   araba-ya (git) '(go) to the car'

2. Its location (indicated by the locative case marker)

   kanepe-de (uyu) '(sleep) on the sofa'

3. The source of the action or the space through which an action takes place (indicated by the ablative case marker)

[2] Another traditional linguistic classification omitted in this lexicon is the *external-internal* argument distinction. *External* argument and *internal* argument are not widely known in the computational linguistics community. The *external* argument is widely refered to as the 'subject'. As for the Turkish internal argument, if we assume it to be the direct object, then it is refered to as "direct object" by Turkish computational linguists. If we assume the 'internal argument' to refer to the arguments excluding the external argument and the adjuncts, then the internal argument may carry dative, ablative and instrumental/ comitative cases in addition to the accusative and nominative cases in Turkish. Despite the theoretical power of this conception of 'internal argument' in explaining rule governed facts in languages, which are of great importance in linguistics, the practical aims of compiling a lexicon requires encoding the arguments in terms of their syntactic, semantic and morphological properties that are more fine grained than the concept of 'internal argument'. Obviously, the concept of internal argument, with any assumed definition can be derived from this lexicon.

[3] Göksel and Kerslake use the term "adverbials" to refer to adjuncts in the usual sense: the elements that are not subcategorized by the verb. I will use "adverbials" to refer to true adverbials (simple adverbs, derived adverbs and adjectives used adverbially) in this study.

büro-dan (kağıtlar-ı al) '(take the papers) from office'

4. The manner in which the action takes place indicated by:

   (a) the instrumental/ comitative marker

   bıçak-la (ekmek kes) '(cut bread) with knife'

   (b) a simple adverb

   yavaş (yürü) '(walk) slowly

   (c) a derived adverb

   yavaş-ca (yürü) '(walk) slowly'

   (d) an adjective used adverbially

   güzel (oku) '(study) well'

3.2 Non-heads in Idiomatic Compounds

The non-head of an idiomatic compound is always an obligatory element of the sentence as it must exist for the idiomatic reading to exist. Therefore, it will never be coded as optional.

Non-head in an idiomatic compound can be a non-case marked direct object (categorial direct object) as in (3.2a), an accusative case marked direct object as in (3.2b). Alternatively, it can be marked with dative, or ablative cases (3.2c, d).

(3.2) a. biri-ne fırça atmak
someone-DAT brush throw
'to scold someone'

b. bir yer-e kapağ-ı atmak
somewhere-DAT lid-ACC throw
'to settle in somewhere'

c. biri-ni/ bir şey-i tehlike-ye atmak
someone/ something-ACC danger-DAT throw
'to risk someone/ something'

d. birşey-i kafa-dan atmak
something-ACC head-ABL throw
'to make up something artificial or untrue'

Non-heads in instrumental/ comitative case (3.3a), and subject non-heads (3.3b) are very rare. Likewise, compounds with more than one non-head, as exemplified in (3.4a) and (3.4b), are not very common.

(3.3) a. bir şey-i ip-le çekmek
something-ACC rope-INST pull
'to anxiously wait for'

b. biri-ni kan tutmak
someone-ACC blood hold
'to be irritated by blood'

(3.4) a. turna-yı gözün-den vurmak
crane-ACC eye-ABL hit
'to be lucky'

b. laf laf-ı açmak
word word-ACC open
'there to be a long chat'

Finally, adverbials can be non-heads in idiomatic compounds. The adverbials in idiomatic compounds are mostly adjectives used as adverbials, as exemplified in (3.5).

(3.5) a. küçük düşmek
little fall
'be humiliated'

b. bitap düşmek
tired fall
'to be exhausted'

.

3.3 Possessive Marking on the Non-heads

A significant number of compounds bear a possessive marker on their non-heads. As a matter of fact, the possessive marker is sometimes be necessary for the idiomatic reading. For example, the idiomatic reading is lost by the omission of the possessive marker in (3.6b). Likewise, different idiomatic readings are introduced by the omission and the inclusion of the possessive marker in (3.7).

(3.6) a. biri-nin tepe-sin-e çıkmak
someone-GEN top-P3sg-DAT climb
'to sauce someone'

b. tepe-ye çıkmak
top-DAT climb
'to climb to the top'

(3.7) a. biri-nin baş-ı-na çıkmak
someone-GEN head-P3sg-DAT climb
'to sauce someone'

b. bir şey-le/ biri-yle baş-a çıkmak
something/ someone-INS head-DAT climb
'to handle something/someone'

The POSS marked NP's in compounds may form a genitive/possessive (GEN/POSS) construction which is coindexed with the matrix subject as in (3.8). The possessive marker can also be coindexed with the direct object as in (3.9).

(3.8) Emek$_i$ sıkıntısını hep iç-i -ne$_i$ at-ar.
Emek distress-P3sg-ACC always inside-P3sg-DAT throw-AOR.
'Emek always suppresses her distress'

(3.9) O ben-i$_i$ sırt-ım-dan$_i$ vur-du.
she I-ACC back-P3sg-ABL shoot-PAST
'She betrayed me.'

Alternatively, the possessive marker can be coindexed with the *possessor* NP, which is a genitive case-marked modifier, as in (3.10). The possessor NP of a non-head is always an argument of the compound, in that it corresponds to an argument in the English translation. For example in (3.10), the possessor *Elif* corresponds to the subject of the English translation. It has the semantic role *Experiencer*.

(3.10) a. [Elif-in$_i$ can-ı$_i$] pasta çek-ti.
Elif-Gen spirit-P3sg cake pull-PAST
'Elif wants some cake.'

b. Eylem [bulaşık-lar-ın$_i$ kaba-sı-nı$_i$] aldı
Eylem dish-PL-GEN base-Poss2sg-ACC take-PAST
'Elif cleaned the dishes superficially.'

By the same token, *bulaşık-lar* (dishes) is the *possessor* of the direct object in (3.10b), and it corresponds to the object in the English translation and has the semantic role *Theme*.

Semantic frames link different syntactic realizations of the same thematic role in English and Turkish. For *canı çekmek* in (3.10a), the VerbNet frame specifies the [Experiencer V Theme] order for the verb *want#3*. On the other hand, our lexicon lists the *Experiencer* under the *possessor* of subject position, and the *Theme* under the direct object position. Therefore, we know that the [Experiencer V Theme] order in English corresponds to an *Experiencer* in the possessor of the subject, and a *Theme* in the direct object position in Turkish.

The Turkish entry for *canı çekmek* also includes the information that the direct object cannot have overt case marking, the verb cannot be in reflexive and reciprocal etc. For a complete description of the structure of lexicon please refer to section 4.

The following example illustrates the need for separate entries in the lexicon for compounds that differ only in values for possessive marking and co-indexation. In (3.11a) there is no possessive marker; in (3.11b) the possessive marker is coindexed with the subject and in (3.11c) the possessive marker is coindexed with the *possessor* of the ablative object. The alternations in meaning necessitate separate lexicon entries for these compounds.

(3.11) a. Erdil$_i$ yol-dan çıktı.
Erdil road-ABL leave-PAST
'Erdil deviated from the true way of life.'

b. Erdil$_i$ yol-u-ndan$_i$ çıktı.
Erdil road-P3sg-ABL leave-PAST
'Erdil deviated from his way of life.'

c. Erdil akıl hoca-sı-nın$_i$ yol-u-ndan$_i$ çıktı.
Erdil mentor-P3sg-GEN road-P3sg-ABL leave-PAST
'Erdil deviated from his mentor's way of life.'

Even in cases where an alternation in meaning cannot be observed, separate lexicon entries are needed, as illustrated in (3.12). The non-head has no overt marking in (3.12a), has an overt accusative marker in (3.12b), and has an additional possessive marker coindexed with the subject in (3.12c). Although they all convey the same sense, such compounds are

listed as separate entries in the lexicon as they have different and contradictory morphological constraints on the non-head.

(3.12) a. kazık atmak
pole throw
'to humbug'

b. kazığ-ı atmak
pole-ACC throw
'to humbug'

c. kazığ-ı-nı atmak
pole-P3sg-ACC throw
'to humbug'

3.4 Bare Noun Non-heads

A bare (non-case marked) non-head can be a categorial direct object or a subject. Kartal (1995) proposes that this can be tested by inserting a subject and an object in the structure and observing whether the resulting structure is ungrammatical. In both (3.13b) and (3.13b), there is no overt case marking on the non-head. In (3.13a) the non-head *kan* is the subject, since it is not possible to insert a subject in the sentence as shown in (3.14). However, we can insert a subject in (3.15) indicating that the non-head *kafa* is an object.

(3.13) a. biri-ne kafa tutmak
someone-DAT head hold
'to defy someone'

b. biri-ni kan tutmak
someone-ACC blood hold
'to be irritated by blood'

(3.14) * Ahmet kan tuttu.
Ahmet blood hold-PAST.

(3.15) Ahmet bana kafa tut-tu.
Ahmet me head hold-PAST.
'Ahmet defied me'

Öztürk (2004) studies the properties of bare nouns in Turkish with respect to their thematic roles.[4] With respect to the thematic roles, *kan tutmak* in (3.13b) is an example of 'agent incorporation', as the verb *tutmak* (to hold) assigns its subject the thematic role *Agent*.[5] By the same token, (3.15) is an example of 'theme incorporation' as the verb *tutmak* (to hold) assigns its direct object the thematic role *Theme*.

Idiomatic compounds often change the *thematic role assignment* properties of the verb.[6] For example, in *kan tut* (blood hold), which means 'be irritated by blood'

[4]Öztürk shows that just like immediately preverbal bare nouns in non-idiomatic structures, bare nouns in idioms also behave as NP categories syntactically. She provides the following evidence for the NP status of the the bare nouns in idioms: Focus particles -dA and -mI which target XP categories can be inserted between the verb and the bare noun forming the idiom (a). The verb forming the idioms can be elided under identity (b). In idioms bare nouns can be coordinated (c).

a. Ali bu problem-e kafa mı patlat-tı?
   Ali this problem-DAT head Q-particle explode-PAST
   'Did Ali worked hard to solve this problem?'

b. Ali bu problem-e kafa patlat-tı sen-in gibi çene değil.
   Ali this problem-DAT head explode-PAST, you-GEN as chin not
   'Ali worked hard to solve this problem, he didn't just talk a lot about this problem as you did.'

c. Ali hem çile hem acı çek-ti.
   Ali both misery both pain suffer-PAST
   'Ali suffered both from misery and pain.'

[5]It has been claimed that idioms do not involve verbs and their arguments bearing an agent theta role (Marantz 1984). Öztürk argues that Turkish creates an exception for the cross-linguistic restriction that no agentive arguments are allowed in idiomatic constructions.

[6]Traditional linguistics categorizes verbs into groups according to the thematic roles on their arguments (Grimshaw, 1992). The verbs in this lexicon are not categorized as such, as Oflazer and Yılmaz's case-frame structure requires one thematic role to be associated with every argument or adjunct so that such distinctions and others can eventually be derived from this lexicon. This categorization is summarized as follows:

1. Transitive agentive:
   Agent Theme
   Snow White$_{agent}$ discovered *a hut*$_{theme}$.
2. Ditransitive:
   Agent Theme Goal
   The witch$_{agent}$ gave *an apple*$_{theme}$ to *Snow White*$_{goal}$.
3. Psychological agentive:
   Agent Experiencer
   The witch$_{agent}$ frightened *the dwarfs*$_{experiencer}$.
4. Psychological state:
   Experiencer Theme
   The seven dwarfs$_{experiencer}$ adored *Snow White*$_{theme}$.
5. Unergative:
   Agent
   The dwarfs$_{agent}$ cried the whole night when Snow White died.

the idiomatic reading assigns the role *Experiencer* to its direct object although the verb *tutmak* (to hold) assigns the role *Theme* to its direct object.

Theme Incorporation in Unaccusatives

Some cases of theme incorporation not only change the thematic role assignment properties of the verb, but also change the subcategorization properties of the verb. Kartal (1995) observed that theme incorporation in compounds headed by unaccusative verbs (which normally assign the role *Theme* to the subject and do not subcategorize for an object) results in a compound that assigns the role *Agent* to its subject and subcategorizes for an new object with the role *Theme*. In the examples in (3.16), the compound headed by the unaccusative verb *çıkmak* (appear, leave, go out, etc.) subcategorizes for a new dative marked object. Likewise, in (3.17) the compound subcategorizes for a new instrumental/ comitative marked object.[7]

(3.16) a. *biri-ne* *destek* *çıkmak*
someone-DAT support appear
'to support someone'

b. *biri-ne* *sahip* *çıkmak*
someone-DAT owner appear
'to protect someone'

(3.17) *biri-yle* *alay* *geçmek*
someone-INS fun pass
'to gibe with someone'

The examples in (3.18) show that a non-case marked non-head and an unaccusative verb do not always change the subcategorization properties of the verb. When the non-head bears a possessive marker, the possessor position of the non-head is occupied by an argument of the compound, and no new object is introduced.

6. Psychological Causative:
Theme Experiencer
The thunder$_{theme}$ does not frighten *the dwarfs*$_{experiencer}$.

7. Unaccusative:
Theme
Snow White$_{theme}$ died the moment she bit the apple.

[7]Please note that none of the senses of the verb *geçmek* assigns instrumental/ comitative case to its arguments.

(3.18) a. *Adam-ın akl-ı çık-tı.*
Man-GEN mind-POSS leave-PAST
'The man got mad.'

b. *Bebeğ-in kırkı çık-tı.*
Baby-GEN forty-POSS appear-PAST
'The baby is forty days old.'

c. *Ali-nin maske-si düş-tü.*
Ali-GEN mask-POSS fall-PAST
'Ali's mask has fallen.'

The possessor of (3.18a) *adam* (man) is the experiencer, the possessor of (3.18b) *bebek* is the theme, the possessor of (3.18c) *Ali* can also be considered as a theme. This shows that, the theta role assignment properties of verbs do change arbitrarily. Yet, the subcategorization properties of verbs are more rule governed: Theme incorporation in unaccusatives usually introduces a new argument in the structure, unless the non-head bears a possessive marker.

The following examples are exceptions to the regular pattern of theme incorporation in unaccusatives. The verb *düşmek* in (3.19) is an unaccusative verb, it is not transitive, therefore the bare nouns in (3.19) cannot be non-case marked direct objects. Then, they may be subjects with the role *Theme* and these compounds would be cases of theme incorporation just as the previous examples. These are the only two examples in the lexicon in which a bare NP non-head and an unaccusative verb do not change the subcategorization frame of the verb as it was observed in the previous examples. One analysis would be to categorize such compounds as exceptions to the rule that theme incorporation in unaccusatives results in subcategorization of a subject and a new obligatory object. Alternatively, the nouns in (3.16) can be classified as adverbials. Then the bare non-heads would be zero derived adjectives used adverbially. In any case, these are exceptions to the general pattern and will be encoded as exceptions, along with a three letter extention showing what is exceptional about them. *exc-thm* indicates that this is an exception to the theme incorporation rule in unaccusatives.

(3.19) a. şehit düşmek
martyr fall
'to become a martyr'

b. küme düşmek
league fall
'to fall from a league'

Another interesting exceptional case is shown in (3.20). The verb *tutmak* (to hold) normally assigns the role *Theme* to its direct object. The compounds in (3.20) subcategorize for a new direct object, in addition to the bare nouns *mesken* and *mekan*. This is similar to the theme incorporation in unaccusatives, yet there are not enough cases to claim that a similar construction is at work with transitives as well. Alternatively, these may also be classified as zero derived adjectives used adverbially. Anyway, such cases will be coded as exceptions along with a three letter extention showing what is exceptional about them. *exc-new* indicates that this is an exception in that it introduces a new direct object to the sentence.

(3.20) a. *bir yer-i* *mesken* *tutmak*
somewhere-ACC place hold
'to settle in somewhere'

b. *bir yer-i* *mekan* *tutmak*
somewhere-ACC place hold
'to settle in somewhere'

3.5 Control Properties of Idiomatic Compounds

Control Theory accounts for the referential properties of PRO, a silent pronoun which is the subject of infinitival sentences (Chomsky, 1981). For example, the fact that *John* is the understood subject of *to do the dishes* in (3.21a), but not in (3.21b), and *Susan* is the understood subject of *to do the dishes* in (3.21b) is accounted for by control theory. The case in (3.21a) is called 'subject control' and the case in (3.21b) 'object control'.

(3.21) a. John$_i$ promised Susan$_j$ [PRO$_i$ to do the dishes]

b. John$_i$ ordered Susan$_j$ [PRO$_j$ to do the dishes]

Kartal (1995) investigates the control properties of idiomatic compounds headed by verbs in Turkish. This section heavily depends on her observations, yet the model in which

she investigates the behavior of the compounds is not relevant to the lexicographic aims of this study. The properties of idiomatic compounds are presented here are as theory neutral as possible.

Whether a compound participates in a certain control structure is constrained by the syntactic properties of the compound, yet these syntactic constraints do not by themselves determine whether the compound actually takes part in that control structure. The types of control structures that a compound does occur depend on the semantic properties of the compound. Therefore, although some control properties of idiomatic compounds are rule governed by the syntactic constraints, the control structures in which an idiomatic compound actually takes part is a lexical information.

Subject and Arbitrary Control

Idiomatic compounds allow for two types of control, subject and arbitrary. In subject control structures, the subject of the embedded sentence is filled by a silent PRO which is coindexed with the matrix subject. This type of control is also called obligatory control. To test whether a compound allows subject control, the verb with the infinitive suffix -mAk is embedded as the complement of the verb *iste-* (want). This is exemplified in the sentences in (3.22). In (3.22a) PRO is coindexed with the subject *Cemal*. In (3.22b), PRO is not only coindexed with the subject *Cemal* but also with the possessive marker on the non-head. In (3.22c), PRO is only coindexed with the subject *Cemal* and the possessive marker on the non-head is coindexed with the possessor *Ahmet*.

(3.22) a. Cemal$_i$ [PRO$_i$ bu sorun-a çare ara-mak] isti-yor.
Cemal this problem-DAT solution search-INF want-PROG
'Cemal wants to find a solution to this problem.'

(Kartal, 49)

b. Cemal$_i$ [PRO$_i$ bana iç-i-ni$_i$ aç-mak] iste-me-di.
Cemal me-DAT inside-P3sg-ACC open-INF want-Neg-PAST
'Cemal didn't want to talk about his problems to me'

(Kartal, 49)

c. Cemal$_i$ [PRO$_i$ [Ahmet-i n$_j$ defter- i-ni$_j$] dür-mek] iste-di.
Cemal Ahmet-Gen book-P3sg-ACC roll-INF want-PAST
'Ahmet wanted to kill Cemal.'

(Kartal, 49)

In arbitrary control, the referential properties of PRO are left undetermined. In the sentences in (3.23), a silent PRO$_{arb}$ is the subject of the embedded sentence. To test whether a compound allows arbitrary control, the verb with the infinitive suffix -mAk is embedded as the complement of substantive predicates like *doğru* (right), *gerek* (necessary), *mümkün* (possible), as shown in (3.23). Arbitrary control structures are also called optional control structures.

(3.23) a. [PRO$_{arb}$ onlar-la alışveriş-i kes-mek] gerek.
they-INST shopping-ACC cut-INF necessary
'It is necessary not to shop from them'

(Kartal, 49)

b. [PRO$_{arb}$ o ev-i ucuz-a kapa-mak] mümkün.
that house-ACC cheap-DAT close-INF possible
'It is possible to buy that house at a low price'

(Kartal, 49)

c. [PRO$_{arb}$ herkes-e tepe-den bak-mak] doğru değil.
everyone-Dar top-ABL look-INF right not
'It is not right to look down upon people.'

(Kartal, 49)

d. [PRO$_{arb}$ güven kazan-mak] kolay değil.
trust earn-INF easy not
'It is not easy to make people trust oneself'

(Kartal, 49)

The following are examples of arbitrary control structures with possessive markers on the non-heads. The possessive marker is coindexed with the possessor *bu sorun* in (3.24a) and with the object *biri* in (3.24b). Since the possessive marker is not coindexed with the subject in the following examples, such sentences can have an arbitrary PRO_{arb} as subjects, in other words they allow arbitrary control.

(3.24) a. [PRO_{arb} [bu sorun-un$_i$ kök-ü-nü$_i$] kazı-mak] mümkün.
this problem-Gen root-P3sg-ACC scrape-INF possible
'It is possible to exterminate this problem'

(Kartal, 50)

b. [PRO_{arb} biri-ni$_i$ ekmeğ-i-nden$_i$ et-mek] doğru değil.
someone-ACC bread-Gen-ABL do-INF right not
'It is not right to fire someone'

(Kartal, 50)

When the possessive marker on the non-head is coindexed with the subject, then this structure does not allow arbitrary control. In (3.25a) the non-head *kafasını* is coindexed with the subject *Ahmet*. (3.25b) shows that the same compound in an arbitrary control structure is ungrammatical. However, when the non-head is not marked with possessive as in (3.25b) the compound can occur in an arbitrary control structure.

(3.25) a. Ahmet$_i$ kafasını$_i$ kullan-dı.
Ahmet head-P3sg-ACC use-PAST
'Ahmet used his head.'

b. *[PRO_{arb} kafasını$_i$ kullan-mak] gerek.
head-P3sg-ACC use-INF necessary
'It is necessary that one uses her brain.'

c. [PRO_{arb} kafa-yı kullan-mak] gerek.
head-ACC use-INF necessary
'It is necessary to use brains.'

(Kartal, 51)

This is not the case for all compounds in this structure. (3.25a) and (3.25b) parallel with (3.26a) and (3.26b), but (3.26c) in which the non-head is not marked with a possessive marker is still ungrammatical, unlike (3.25c). Such compounds are lexically marked for allowing or not allowing arbitrary control in these contexts, therefore this information should be included in the lexicon.

(3.26) a. Ahmet$_i$ kalb-i-ni$_i$ aç-tı.
Ahmet hearth-P3sg-ACC open-PAST
'Ahmet shared his thoughts with me'

b. *[PRO$_{arb}$ kalb-i-ni$_i$ aç-mak] gerek.
hearth-P3sg-ACC open-INF necessary
'It is necessary for someone to open her hearth'

c. *[PRO$_{arb}$ kalb-i aç-mak] gerek.
hearth-ACC open-INF necessary
'It is necessary for someone to open the heart'

A final remark about arbitrary control is that a structure in which the possessive marker is coindexed with the subject can allow arbitrary control when the marker itself can be interpreted arbitrary (Kartal, 1995). (3.27a) is ungrammatical but a very similar structure is grammatical in (3.27b) when the marker itself can be interpreted arbitrary. This is also possible in questions. The possessive marker on *peş-i-nden*$_{i\ arb}$ is arbitrary in the question, in that the referent is not a specific person.

(3.27) a. *[PRO$_{arb}$ biri-ni arka-sı-ndan$_i$ sürükle-mek] doğru değil.
someone-ACC behind-P3sg-ABL drag-INF right not
'It is not right to drag someone behind'

b. [PRO$_{arb}$ biri-ni peş-i-nden$_{i\ arb}$ doğru değil.
someone-ACC behind-P3sg-ABL drag-INF right not
'It is not right to drag someone behind'

c. [PRO$_{arb}$ biri-ni peş-i-nden$_{i\ arb}$ sürükle-mek] doğru mu?
someone-ACC behind-P3sg-ABL drag-INF right Ques
'Is it right to drag someone behind?'

(Kartal, 50)

Substantial Predicates

There are two types of structures that do not allow control. First, the compounds in which the non head occupies the subject position does not allow control, as in (3.28). This is impossible by definition as the subject position is filled by the non-head, not with a silent PRO.

(3.28) a. *Akıl bu-nu al-mak iste-mi-yor.
mind this-ACC take want-Neg-PAST
'The mind does not want to grasp such a thing'

b. *Karanlık bas-mak isti-yor
Darkness overwhelm-INF want-PROG
'Darkness wants to overwhelm'

(Kartal, 54)

The compounds in which the non-head and the head form an unaccusative structure or a psychological causative structure do not allow control as well. However, some compounds which do not allow control do allow control when embedded under substantive predicates. The sentences in (3.29) exemplify an unaccusative compound that allows subject control with the substantial predicate *zorunda* (must). However, this is not the case with the unaccusative in (3.30), showing that all verbs do not follow the same pattern. This calls for the need to include this information in the lexicon as a separate value.

(3.29) a. * Öyle bir aksilik$_i$ [PRO$_i$ baş göster-mek] isti-yor.
such a obstacle head show-INF want-PROG
'Such an obstacle wants to arise'

(Kartal, 54)

b. Öyle bir aksilik [PRO$_i$ baş göster-mek] zorundaydı zaten.
such a obstacle head show-INF must-PAST anyway
'Such an obstacle had to arise anyway'

(Kartal, 55)

(3.30) a. *[PRO$_i$ Ekin-ler baş ver-mek] isti-yor.
crop-Pl head give-INF want-Prog
'The crop wants to arise'

(Kartal, 54)

b. *[PRO$_i$ Ekin-ler baş ver-mek] zorunda.
crop-Pl head give-INF must
'The crop must arise'

(Kartal, 54)

Summary

The control properties of idiomatic compounds are investigated and added to the lexicon by employing the following tests. The verb must always be in the infinitival form, i.e. suffixed by -mAk. Test 1, 3 and 4 should not be applied if the non-head(s) is in the subject position. Test 4 should be applied only if the non-head bears a possessive marker. If the value for Test 1 is *yes* then the value for Test 2 is also *yes*.

(3.31)
1. Embedding the compound under the verb *iste-* (want) to test whether the compound allows subject (obligatory) control.
2. If the compound does not allow control, test the subject control by embedding it under the substantive predicate *zorunda* (must).
3. Embedding the compound under *doğru* (right), *gerek* (necessary), *mümkün* (possible) to test whether the compound allows arbitrary control.
4. If there is a possessive marker on the non-head, strip off the possessive marker and test whether it allows arbitrary control in this structure.

CHAPTER 4

STRUCTURE OF THE LEXICON

The lexicon is kept in a spreadsheet file, a copy of which can be found in the electronic database. Every restriction listed in this section corresponds to a column header in this file. Lexicon entries correspond to rows and have certain values for each column listed here.

4.1 Elements of the Compound

The restrictions described in this section can take slightly different values for compounds that have embedded sentences as subparts. They make up less than two percent of the lexicon, thus the structure proposed for such compounds is an extension of the one described here for all other compounds. They are covered in Section (4.6) for completeness.

The restrictions on the compounds listed in this section are exemplified in Table 4.1 on the idiomatic compound *kan beynine çıkmak* (to rage).

Lexicon Entry Number

Every lexicon entry must have a number.

Head

The head of the compound is the verb. In case of the compound in Table 4.1, it is *çıkmak* (climb).

Table 4.1: Elements of the Compound

xNo	Head	Non-Head	Root	Role	Poss	Index
1	çıkmak	kan beynine	kan beyin	sub dat	no yes	possessor_dat

Table 4.2: Verb Features and Control

No	Head	Non-Head	Ref	Rec	Cau	Pas	Verb	T1	T2	T3	T4
1	çıkmak	kan beynine	no	no	yes	no		no	yes	no	no

Table 4.3: Semantic Mapping

No	Head	Non-Head	Sense	VerbNet	FrameNet
1	çıkmak	kan beynine	rage#3	marvel-31.3	#experience_obj

Non-head(s)

It is possible for a compound to have more than one non-head. Table 4.1 shows a compound with two non-heads *kan* (blood) and *beynine* (brain-P3sg-DAT).

Root(s)

This section lists the non-heads in their root form. The non-head column lists them in their possessive third person singular suffixed form for easy scanning of lexicon by the lexicographers. As the possessive marker on the non-head should agree with what it is coindexed with, the form in the non-head column is not always relevant.

Syntactic Role

This column lists the case on the non-head(s) as in Table 4.1. The accusative marked direct object is coded as *dir*. If the case on the direct object is not overtly marked, then it is listed as *cat* as they are categorial direct objects. If the non-head occupies the subject position then it is listed as *sub*. If the non-head is an adverbial, then it is listed as *adv*. The theme incorporation in unaccusatives are coded as *thm*. Finally exceptional cases are coded as *exc*. Dative, ablative and instrumental/ comitative are coded as *dat*, *abl* and *ins* respectively. In the above example, the non-heads have the roles *sub* and *dat*.

Possessive Marker

This column has either *yes* or *no* depending on whether the non-head has a possessive marker. In the example, *kan* does not have a possessive marker, and *beynine* has. So this column has *no yes* referring to the order of the words in the non-head column.

Co-indexation

This column lists the constituent in the structure which is coindexed with the possessive marker on the non-head. In the example, this is the *possessor* of the dative object, abbreviated as *possessor_dat*.

4.2 Restrictions on the Verb

Restrictions on the verb include any morphological feature necessary for the idiomatic reading. In Table 4.2, the columns *ref* (reflexive), *rec* (reciprocal), *cau* (causative) and *pass* (passive) have the values *yes* or *no*.[1]This indicates whether the verb can be in reflex-

[1]**Reflexive**
A verb is rendered reflexive by suffixation of **-(I)n-** to the verb stem. It adds the meaning that the action is done to oneself. Moreover, it adds the meaning that the action is done for oneself (Lewis, 1967).

(4.1)	bul-	to find	bul(un)-	to find oneself
	döv-	to beat	döv(ün)-	to beat one's breast
	giy-	to wear	giy(in)-	to dress oneself

Reciprocal
Reciprocal or the co-operative verb is derived by suffixation of **-(I)ş-** to the verb stem. It adds the meaning that the action is done by more than one subject one with another or one to another (Lewis, 1967).

(4.2)	anla-	to understand	anla(ş)-	to understand one another
	döv-	to beat	döv(üş)-	to fight one another
	sev-	to love	sev(iş)-	to make love

Causative
Causative is formed by adding one of the suffixes listed bellow to the stem. (Lewis, 1967) **-dir-** This is the most common causative suffix, but not used with polysyllabic stems ending in a vowel or **l** or **r**.

(4.3)	inan-	to believe	inan(dır)-	to to persuade
	öl-	to die	öl(dür)-	to kill

-ir- This is used with some twenty monosyllables.

(4.4)	doğ-	to be born	doğ(ur)-	to give birth to
	kaç-	to escape	kaç(ır)-	to kidnap, to let escape

-t- This is used with polysyllabic stems ending in a vowel or **l** or **r**.

(4.5)	anla-	to understand	anla(t)-	to explain
	bekle-	to wait	bekle(t)-	to keep waiting

-It- This suffix is used after a few monosyllabic stems, mostly ending in **-k**.

(4.6)	ak-	to flow	ak(ıt)-	to shed
	kork-	to fear	kork(ut)-	to frighten

-Ar- This suffix occurs in only in the following words:

(4.7)	çık-	to go out	çık(ar)-	to remove
	çök-	to collapse	çöker-	to cause to collapse
	git-	to go	gid(er)-	to remove
	kop-	to break off	kop(ar)-	to causue to break off
	on-	to prosper	on(ar)-	to repair

Irregulars The irregular causative forms are given in (4.8).

(4.8)	gel-	to come	getir-	to bring
	gör-	to see	göster-	to show
	em-	to suck	emzir-	to suckle
	kalk-	to rise	kaldır-	to raise, remove

Passive Passive is formed by adding **-il-** after all consonants except **/-l/**.

(4.9)	sev-	to love	sev(il)-	to be loved
	gör-	to see	gör(ül)-	to be seen

ive, reciprocal, causative, and passive voices without losing the idiomatic reading.

The restrictions on tense and agreement markers on the verb are given in a separate column *verb* in Table 4.2. The occurrence of a morphological feature on the verb is coded for example as (morph, fut), which means that the verb must be in future tense in order to preserve the idiomatic reading. Moreover, *not* is used as a prefix on the pairs. For example, [not[morph,fut]] means that that the verb should not be marked with future tense in order to preserve the idiomatic reading. The values for *morph* follow the conventions used in tagging the Turkish Treebank (Oflazer et al., 2003).

4.3 Control Properties

Control properties of compounds are accessed by employing the tests developed in section (3.5). They are summarized here for convenience:

Test 1

Embed the compound under the verb *iste-* (want) to test whether the compound allows subject (obligatory) control. For *kan beynine çıkmak*, the value for this test is *no*, because we cannot embed it under *iste-*, as shown in (4.11). This is expected as *kan* occupies the subject position.

(4.11)

*Kan	beyn-in-e	çık-mak	ist-iyor.
blood	brain-P3sg-DAT	go up-INF	want-PROG

Test 2

If the compound does not allow control, test subject control by embedding it under the substantive predicate *zorunda* (must). (4.12) shows that *kan beynine çıkmak* has the value *yes* for this test.

(4.12)

Kan	beyn-i-ne	çık-mak	zorunda.
blood	brain-P3sg-DAT	go up-INF	must

Stems ending in **-l** or a vowel form their passive by suffixation of **-(I)n-**, which is identical to the reflexive form.

(4.10)

al-	to take	al(ın)-	to be taken
oku-	to read	oku(n)-	to be read

'He must rage.'

Test 3

Embed the compound under *doğru* (right), *gerek* (necessary), *mümkün* (possible) to test whether the compound allows arbitrary control. *kan beynine çıkmak* has the value *no* for this test as expected, as *kan* occupies the subject position, which thus cannot be occupied by an arbitrary PRO.

(4.13) *Kan beyn-i-ne çık-mak mümkün.
blood brain-P3sg-DAT go up-INF possible

Test 4

If there is a possessive marker on the non-head, strip off the possessive marker and test whether it allows arbitrary control in this structure. Although there is no need to employ this test for *kan beynine çıkmak* because *kan* is the subject of the sentence, the test is shown in (4.14) for completeness.

(4.14) *Kan beyin-e çıkmak mümkün.
blood brain-DAT go up-INF possible

Test 1, 3 and 4 should not be applied if the non-head(s) is in the subject position. Test 4 should be applied only if the non-head bears a possessive marker. When a test is not applicable, the value is coded as *n/a*. If the value for Test 1 is *yes* then the value for Test 2 is also *yes*.

Table 4.2 shows the values for the restrictions described above. The reader can easily verify that this compound can take only the causative voice. Moreover, there are no restrictions on tense and agreement markers.

Please be reminded that the Tables 4.1 and 4.2 correspond to columns and rows in the spreadsheet file, thus the order of the columns are flexible. They need not always be in the order given here.

Sense

This is the corresponding WordNet sense, or a short description if a mapping cannot be found in WordNet. If the value is a WordNet sense, then it is a word followed by the # sign and a number. For example, it is *rage#3* for *kan beynine çıkmak*. The word with the sense number corresponds to a synset in WordNet, as described in (2.1). Else, if a mapping cannot be found in WordNet, a short description following a # sign is given. For example, *#be forty days old* is listed for the compound *kırkı çıkmak*. All words in the description then can be mapped to WordNet definitions if possible, for example *#be #1 forty#1 days#1 old#*.

VerbNet

This is the name of the frame that the sense falls under in VerbNet. For *kırkı çıkmak*, *rage#3* falls under the frame *marvel-31.3*. It is not always possible to find a frame for a given WordNet sense in VerbNet. In that case, the value is either left blank, or a frame which is close to the meaning is used. This approximate frame is input to the lexicon with the prefix #, e.g. *#marvel-31.3*. The approximate frames should be revised in every update of the lexicon. It should be checked whether the frame corresponding to the relevant WordNet sense has been entered in VerbNet.

FrameNet

This is the name of the frame that the sense falls under in FrameNet. *rage#3* falls under the frame *experience_obj* in FrameNet. It is not always possible to find a specific enough frame for a given sense in FrameNet. The frame *experience_obj* is too general for the meaning of *rage#3*. In that case, the value is either left blank, or a frame which covers the meaning is used. Then this frame is input to the lexicon with the prefix #, e.g. *#experience_obj*.

4.5 Case-Frame

The case-frame includes the slots for the arguments (denoted by their cases) and their *possessor* positions, and adjuncts such as propositional phrases, adverbials etc. The *possessor* is included as a separate slot because this position may be filled by an argument of the compound, which corresponds to an object position in English.

Unlike the restrictions above, which can have only a couple of values, the slots in the case-frame include a list of attribute value pairs. These attribute value pairs are coded in "[]" brackets, with a "," comma between them. The case-frame for *kan beynine çıkmak* (rage#3) is given in Table 4.4. If the slot for a frame element has *no* in it, this means that the compound does not subcategorize for that element. For instance, all intransitive verbs have *no* in their direct object positions. If the list of attribute value pairs is preceded by an *OP* then this is an optional element of the case-frame. For example, *sinir* (nervousness) in *abl* is an optional element in Table 4.4. Otherwise, it is an obligatory element of the case-frame. The explanations of the attributes are:

Lexical

Lexical constraints determine which string to be used in this position. E.g. [lexical, beyin]

Morph

This requires the occurrence of a morphological feature on the constituent. For example, [morph, poss] requires a possessive marker on the nominal form. Moreover, *not* is used as a prefix on the pairs. For example, [not[morph,poss]] means that the nominal should not be marked with possessive.

Index

This assigns an index number to the constituent in order to co-index it with another constituent with the same index number. E.g: [index,1]

Table 4.4: The Case-Frame of *kan beynine çıkmak*

sub	**p_sub**	**dir**	**p_dir**	**dat**	**p_dat**	**ins**	**p_ins**	**abl**	**p_abl**	**adj1**
[lexical, kan]	no	no	no	[lexical,beyin], [index,1]	[marvel-31.3, experiencer], [experience_obj, experiencer], [index,1]			OP [lexical,sinir]		

Table 4.5: The lexicon entry for *ağzından girip burnundan çıkmak*

verb	**sub**	**p_sub**	**dir**	**p_dir**	**dat**	**p_dat**	**ins**	**p_ins**	**abl**	**p_abl**	**adj1**
[lexical,çık]	[amuse-31.1, cause], [suasion,speaker]		no	no	no	no			[lexical,burun], [index,1], [morph, poss]	no	[sentence,1]
[lexical, gir], [morph, afterdoingso]	no	no	no	no	no	no	no	no	[lexical,ağiz], [index,1]	[index,1], [amuse-31.1, experiencer], [suasion,addressee]	

Frame

frame is the name of the VerbNet or the FrameNet frame. Its value is the semantic role of the constituent in that frame. In Table 4.4, VerbNet and FrameNet have the same name for the role of poss_dat. This may not always be the case.

Is-a

This is a hyponymy relation. For example, [is-a, solider#1] is used on the subject of *çürüğe çıkmak* (to be disqualified as a soldier). Using this relation has been very useful for representing senses that are not described specific enough by WordNet definitions. The value for this attribute *soldier#1* is again a WordNet sense.

Has-a

This is a meronymy relation. E.g [has-a, car#1]

Additional attributes are used for compounds that subcategorize for embedded sentences in their non-heads. The values for *morph* and for additional attributes in the following section follow the conventions used in tagging the Turkish Treebank. Please note that the *poss* [morph, poss] is a variable over the set of possessive markers, not a tag used in the Turkish Treebank. Moreover, there is no limit on the number of adjunct slots. They can be added as needed, the number suffixed to the *adj* is increased e.g. *adj1*, *adj2*... Finally, there is a column for the new arguments introduced by theme incorporation and other exceptional new arguments.

4.6 Compounds with Embedded Sentences

An embedded sentence in a compound can be a nominalized sentence in an argument position, or a derived adverb in an adjunct position, or a relative clause inside a noun phrase, or a complement of a postposition.

Embedded sentences are represented in the frames of matrix sentences with the

pair [sentence,1]. The frames of the embedded sentences are listed on the rows immediately following the frame of the matrix sentence, as shown in Tables 4.5, 4.6 and 4.7.

The frames for embedded sentences with increasing numbers are given in consecutive rows, yet finding compounds with more than one embedded sentence is not very probable. Nevertheless, the value of the number in [sentence,1] refers to the order of the row of the embedded sentence.

In Table 4.5, *ağzından girip burnundan çıkmak* (to persuade) has an embedded sentence in a derived adverb, which occupies the *adj1* position. In Table 4.6, *-esi tutmak* (to feel like doing) has a nominalized sentence in the subject position. They are both represented as [sentence,1].

In Table 4.7, *kazdığı kuyuya düşmek* (to be trapped), the embedded sentence is a relative clause inside the noun phrase. Unlike the first two, the embedded sentence is represented as [modifier,[sentence,1]].

The case-frame does not have slots to refer to the internal structure of NPs or PPs. Since these positions are only relevant for very exceptional cases such as idiomatic compounds with embedded sentences, the structure of the lexicon is not complicated to include slots that refer to deeper embeddings inside NPs and PPs. We can nevertheless refer to them by using the following categories, all of which are in accordance with the conventions used in Turkish Treebank.

Determiner

Quantifiers, articles.

e.g. [determiner, [lexical,bazı]]

Modifier

Adjectives, relative clauses, relativized nouns, unit nouns.

e.g. [modifier, [sentence,1], [occurrence,1]], [modifier, [lexical,sarı], [occurrence,2]]

Table 4.6: The lexicon entry for *-esi tutmak*

verb	**sub**	**p_sub**	**dir**	**p_dir**	**dat**	**p_dat**	**ins**	**p_ins**	**abl**	**p_abl**	**adj1**
[lexical, tut], [not [morph,FUT]], [not[morph,Prog1]]	[sentence,1], [want-32.1,Theme], [desire,event]	no	no	no	no	no					
[morph , opt+a3sg+poss],	[want-32.1,experiencer], [morph,gen], [desire, experiencer]										

Table 4.7: The lexicon entry for *kazdığı kuyuya düşmek*

verb	**sub**	**p_sub**	**dir**	**p_dir**	**dat**	**p_dat**	**ins**	**p_ins**	**abl**	**p_abl**	**adj1**
[lexical, düş]	[index,1], [manipulate_into_doing, Victim]		no	no	[modifier, [sentence,1]], [lexical,kuyu]	[manipulate_into_doing, manipulator]					
[lexical,kaz], [morph,PastPart]		no	no								

Classifier

A nominal modifier in the nominative case. For example: [classifier, [lexical, syntax]] in *söz dizim kitabı* (syntax book).

Possessor

A nominal modifier in the possessive case. For example: [possessor,[lexical,Elif],[morph, prop]]. Internal structure of a noun phrase is represented by lists, the head of which is the name of the category and the tail is again a list of attribute value pairs. Since there can be more than one modifier in a noun phrase, [occurrence,1] is included in the list [modifier,[sentence,1],[occurrence,1]].

Object

Object is used to refer to an embedded sentence inside a postpositional phrase. For example, the postposition *için* takes a nominalized sentence as a complement. This postpositional phrase is included in an adjunct slot in the following way:

(4.15) [object,[sentence,5]],[postp, için]

The lexicographer

The head of the list also can be the name of the lexicographer, as she may enter a value on her discretion by using her name in the code. For example, if the lexicographer tentatively decides on a value X because she thinks it is controversial, she should pair X with her name as follows:

(4.16) a. [elifeyigoz, no]

b. [elifeyigoz, [classifier, [lexical, syntax]], [not[morph,plur]]]

c. [elifeyigoz, [not[morph, poss]]]

Please observe that, the frames for embedded sentences include lists, the head of which is a constant e.g. [modifier, [sentence,1]] or [object,[sentence,5]] or [elifeyigoz, no]. In the

above examples, the head of the list is the name of the lexicographer, and the tail is a list of values as described above. Furthermore, the lexicographer may use the question mark *?* whenever she feels indecisive.

CHAPTER 5

NLP APPLICATIONS

5.1 Word Sense Disambiguation

Idioms pose serious difficulties for many NLP applications as they contradict the principle of compositionality: their meanings cannot be deduced from the literal definitions and the arrangement of their parts, but refer instead to a meaning that is known only through common use. Idiomatic rules not only introduce new semantic content, but they also allow for creation of new idiom-specific subcategorization frames i.e. new selectional restrictions. This section discusses the implications of the new semantic content and the selectional restrictions with respect to word sense disambiguation applications.

Word sense disambiguation is the problem of determining which sense of a word is being used in a given context. There are two approaches to handling the lexical ambiguity problem: the constraint-based approach, and the stand-alone approach (Jurafsky and Martin,1999). In the constraint-based approach, selectional restrictions are the primary knowledge sources used to perform disambiguation. They are used to rule out the inappropriate senses and reduce the ambiguity in the analysis. The selection of the correct word senses occurs during semantic analysis as a side-effect of the elimination of ill-formed semantic representations.

For example, the new selectional restrictions introduced by the idiom *birinin tepesine cıkmak* (to sauce someone) in (5.1) evokes the frame *Abusing* in FrameNet, which assigns the role *Abuser* to the subject and *Victim* to the possessor of the dative object. The literal reading *birşeyin tepesine çıkmak* (to climb on top of something) is the only possible analysis in (5.1a) as the possessor of dative *dağ* (mountain) cannot bear the role *Victim*.

This restriction is satisfied in (5.1b) so the idiomatic reading is available as well as the literal one.

(5.1) a. *çocuk* *dağın* *tepesine* *çıktı.*
child mountain-GEN top-POSS-DAT climb
'The child climbed up the mountain.'

b. *çocuk* *adamın* *tepesine* *çıktı.*
child man-GEN top-POSS-DAT climb
'The child sauced the man.'
'The child climbed up to the man's head'

In the stand-alone approach, sense disambiguation is performed independent of and prior to compositional semantic analysis. In analysing idiomatic compounds, the stand-alone approach may select the correct sense for the non-head, for example the non-head *kırk* (forty) in (5.2a), and *nazar* in (5.2b) do not loose their literal meanings in the idiomatic reading. However, the stand-alone approach can never select the correct sense for the verb if the compound is to be considered idiomatic. For example, *çiçek açmak* 'to open flower' (to bloom) is included in the lexicon although its idiomatic status is debatable, because none of the dictionary definitions of *açmak* in TDK dictionary can convey the meaning of *to bloom*. Therefore, the stand-alone approach always fails in determining the senses of verbs and rarely selects the correct senses of non-heads in idiomatic compounds.

(5.2) a. *kırk-ı* *çıkmak*
forty-POSS leave
'to be forty days old'

b. *nazar-a* *gelmek*
evil eye-DAT come
'to be effected by evil eye.'

The lexicon will be used in conjunction with other word-sense ambiguation applications, and the data to be analysed is going to be compared with the information in the lexicon in addition to other attempts to resolve the meaning. If evidence in favor of a literal reading exists, then the idiomatic sense will be included among the possible senses of the pair. Since idioms are rarely used literally, the idiomatic reading should be assigned higher probability than the literal senses.

5.2 Coding Conventions

The sample lexicon is a Turkish-English machine readable dictionary, which can be loaded in a database and can be queried via various applications. The selectional restrictions on the idiomatic compounds are compiled in a spread sheet file, the structure of which can also be used for representing verb senses and other co-locations such as non-idiomatic noun-verb pairs. This section summarizes the conventions used in coding the selectional restrictions in the dictionary and proposes a guideline to extend the coding conventions for future work. Any future extension of the design should follow these guidelines.

The attribute-value pairs and lists are designed in the Prolog style: with brackets and separated with commas. An attribute value pairs is a list of length two, the head of which is always a constant. It can be one of the constants listed in Table 5.1 as in (5.3a), a name of a VerbNet or FrameNet frame as in (5.3b), or a category in Turkish Treebank as in (5.3c).

(5.3) a. [lexical, nazar]

b. [abusing, victim]

c. [prop, için]

d. [elifeyigoz, [classifier, [lexical, syntax]], [not[morph,plur]]]

e. [modifier, [sentence,1], [occurrence,1]]

Table 5.1: The Constants

lexical	possessor
morph	classifier
is-a	determiner
has-a	modifier
index	occurrence
sentence	

Lists of lists are also used as in (5.3d) and (5.3e). The heads of lists are again constants. In

additon to the constants in Table 5.1 the head of a list can be the name of the lexicographer as in (5.3d). The tails of lists are also lists of attribute value pairs or lists.

There are also two prefixes: not and op (optional) as in (5.4). Please note that the prefixes on the lists do not follow the Prolog conventions.

(5.4) a. not[morph, poss]

b. op[lexical, sinir]

Finally, if an exact macth cannot be found in the semantic resources, the partial macthes will be coded with a "#" sign preceding the value (e. g. #rage#3, #abusing). Moreover, if a mapping to a WordNet sense cannot be found then a definiton will be given by the lexicographer. If possible the words in the definition are also mapped to WordNet definitons (e. g. #be forty#1 days#1 old#1).

5.3 Control Phenomena

The depth of representations in the lexicon are increased by including the control properties of idiomatic compounds in addition to the selectional restrictions. It has been common practice to include control properties in NLP lexicons, control properties are expressed in lexicons such as Genelex, Acquilex, PLNLP, ILCLEX, LDOCE and Comlex.

Control properties of Turkish idiomatic compounds are also lexically determined. The grammaticality of an idiomatic compound in a certain control structure cannot be determined without consulting a lexicon which explicitly lists the acceptable structures. For example, in (5.5) and (5.6) the heads *gelmek* are the same, the non-head positions are both occupied by adverbs and they both subcategorize for a dative object. Still, *ters gelmek* cannot occur in the control structures that *üstün gelmek* can occur. This difference in control properties lie only in the semantic properties of the compounds, no syntactic restriction can account for this pattern.

(5.5) a. * *birine* *ters* *gelmek* *istiyor.*
someone reverse come-INF want-PROG

b. * *birine* *ters* *gelmek* *zorunda.*
someone reverse come-INF must

c. * *birine* *ters* *gelmek* *mümkün.*
someone reverse come-INF possible

(5.6) a. *birine/ bir şeye* *üstün* *gelmek* *istiyor.*
something come-INF want-PROG
'He wants to defeat someone.'

b. *birine/ bir şeye* *üstün* *gelmek* *zorunda.*
something come-INF must
'He must defreat someone.'

c. *birine/ bir şeye* *üstün* *gelmek* *mümkün.*
something come-INF possible
'It is possible to defeat someone'

Control properties will be used as restrictions in language generation and semantic analysis applications for narrowing down the set of well-formed sentences. The applications which can make use of these properties must have access to a classification of substantive predicates that allow subject control (e. g. zorunda), that allow arbitray control (e.g. mümkün), and be able to recognize verbs that subcategorize for infinitival embedded sentences that allow subject control (e.g. iste).

CHAPTER 6

CONCLUSION

The lexicons differ as to the number of entries they contain and to the amount of linguistic information each entry is provided with. Some lexicons give more importance to coverage, others concentrate on depth of lexical representations. The lexicon design proposed here is of the second type. Furthermore, this lexicon has another level of coverage, namely the possibility of including the non-idiomatic noun-verb pairs. Future work may decide to keep the lexicon idiomatic and increase the number of verbs. Alternatively, future work may increase the number of noun-verb pairs per verb, in order to cover non-idiomatic noun-verb pairs as well. At the most extreme, the lexicon design can be used for representing senses of only the verbs without the nouns. This can be accomplished by omitting the properties of the non-head, and just coding the subcategorization properties of the verb in the case-frame.

In terms of semantics, the lexicon relies heavily on VerbNet, FrameNet and WordNet and their future updates. As for syntax, the lexicon is designed as theory neutral as possible. It mainly focuses on the empirical facts such as case markers on the non-heads. As higher order linguistic concepts are derivable from this lexicon, future work will most probably focus on inferring statistical information with respect to these higher order classifications.

The lexicon also assumes a theory of control. The justification for including the control properties in the lexicon has been that control properties of idiomatic compounds are not rule governed, so they must be included in the lexicon. Nevertheless, future work may also include rule governed properties of compounds in order to find and list the exceptions to these rules, as exceptions must be in the lexicon as well.

REFERENCES

Baker, C., Fillmore, C., & Lowe J. (1998). The Berkeley Framenet Project. *Coling-Acl 98: Proceedings of the Conference*. 86-90.

Baker, C., Fillmore, C. & Cronin, B. (2003). The Structure of the Framenet Database. *International Journal of Lexicography. Vol. 16.3.* 281-296.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris, 1981.

Fillmore, C. (1968). The Case for Case. In Bach and Harms (Eds.), *Universals in Linguistic Theory*. New York: Holt, Rinehart, and Winston

Grimshaw, J. (1992). *Argument Structure*. Cambridge: MIT Press.

Göksel, A., & Kerslake, C. (2005). *Turkish: A Comprehensive Grammar*. London & New York: Routledge.

Jurafsky, D., & Miller, H. J. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall.

Kartal, G. (1995). Argument Structure and Idiomatic Compounds in Turkish. Boğaziçi University Master's Thesis.

Kipper, K., Hoa T. D., & Palmer, M. (2000). Class-Based Construction of a Verb Lexicon. *AAAI- Seventeenth National Conference on Artificial Intelligence.* 691 - 696.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press.

Lewis, G. (1969). *Turkish Grammar*. Oxford: Clarendon.

Marantz, A. (1984). *On the Nature of Grammatical Relations*. Cambridge, MA: MIT Press.

Miller, G. (1990). WordNet: An On-Line Lexical Database. *International Journal of Lexicography. Vol. 3* 235-312.

Oflazer, K., and Yılmaz, O. (1995). A Constraint-based Case-frame Lexicon Architecture. *Proceedings of the Workshop on the Computational Lexicon. European Summer School on Logic, Language and Information. Barcelona, Spain*.

Oflazer, K., Say, B., Hakkani Tür, D., & Tür, G. (2003). Building a Turkish Treebank. In Abeille, Anne (Ed.), *Treebanks: Building and Using Parsed Corpora*. Dordrecht/Boston/London: Kluwers.

Öztürk, B. (2004). Complex Predicate Formation in Turkish. *Harvard University Working Papers in Linguistics: Proceedings of Workshop on Light Verbs*. Harvard University.

Tenny, C. (1995). How Motion Verbs are Special. *Pragmatics and Cognition. Vol. 3.1*. 31-73.