An Electronic Lexicon for Turkish Idiomatic Compounds Headed by Verbs

Elif Eyigoz
University of California

Turkish is a very creative language in terms of idiomatic compounds headed by verbs. Although traditional dictionaries include such compounds, syntactic and morphological properties of compounds are left unrepresented. Moreover, although essential elements of idiomatic compounds can be represented in subcategorization frames that refer to the argument positions of the verbs, it has been observed that subcategorization frames are impractical and even inadequate for representing the argument structure of idiomatic compounds headed by verbs in Turkish. This paper presents a design for representing properties of Turkish idiomatic compounds in a machine readable dictionary, which has been showcased in a sample dictionary for 322 idiomatic compounds.

1. Introduction

Turkish is a very creative language in terms of idiomatic compounds headed by verbs (e.g. *to give word to someone* (to promise someone), *to give hand to someone* (to support someone), *to climb to the head of someone* (to abuse someone)). Although traditional dictionaries include such compounds, syntactic and morphological properties of compounds are left unrepresented. This paper presents a design for representing properties of Turkish idiomatic compounds in a machine readable dictionary. The design is implemented as a part of a Turkish-English subcategorization lexicon for Turkish verbs (Eyigoz 2007), which is mapped to WordNet (Miller 1990), FrameNet (Baker et al. 1998) and VerbNet (Kipper et al. 2000). The proposed design has been showcased in a sample dictionary for 322 idiomatic compounds (Eyigoz 2007).

2. Data

The sample dictionary covers the idiomatic compounds headed by ten Turkish verbs with the highest number of senses. Table 1 shows the verbs, the number of senses associated with these verbs in the Turkish-Turkish dictionary *Güncel Türkçe Sözlük* (Contemporary Turkish Dictionary) which was created and is maintained by *Turk Dil Kurumu* (TDK), the number of idiomatic compounds headed by these verbs in the lexicon, and the English translation of the most frequently used sense.

Turkish	English	Number of Senses (TDK)	Number of Compounds in the Lexicon
çık	leave	57	37
tut	hold	50	30
çek	pull	46	33
al	take	35	56
gel	come	38	52
geç	pass	38	8
at	throw	37	41
düş	fall	32	24
aç	open	28	25
vur	hit	28	16
Total		**297**	**322**

Table 1: The verbs and the number of senses

The number of senses has been used as an heuristics in choosing the verbs in order to compile a significant number of idiomatic compounds. However, as Table 1 shows, the relationship between the number of senses and the number of idiomatic compounds is not proportional for the first ten verbs. Please note that the number of senses is less than the compounds associated with the verbs.

3. Properties of idiomatic compounds in Turkish

Non-heads in idiomatic compounds can be noun phrases (NPs) or adverbs. Turkish has five cases: An NP non-head can be a *non-case marked* direct object (categorical direct object), or an *accusative* case marked direct object (definite direct object). Alternatively, it can be marked with *dative* or *ablative* cases. Case-marked and non-case marked non-heads are exemplified in (1).

(1) a. birin-e fırça at-mak
someone-DAT brush throw-INF
"to throw brush at someone" = "to scold someone"
b. turna-yı göz-ün-den vur-mak
crane-ACC eye-POSS3sg-ABL hit-INF
"to hit the crane at the eye" = "to be lucky"
c. küçük düş-mek
little fell-INF
"to fall small" = "to be humiliated"

Non-heads in instrumental/comitative case and subject non-heads are very rare. Likewise, compounds with more than one non-head, as exemplified in (1b) are also not very common. Finally, adverbials can be non-heads in idiomatic compounds, as in (1c). The adverbials in idiomatic compounds are mostly adjectives used adverbially.

3.1. *Possessive marking on the non-heads*

On third of the compounds in the sample lexicon bear a possessive marker on their non-heads. In fact, the possessive marker is sometimes necessary for the idiomatic reading. The possessive marker on the non-head can be coindexed with the matrix subject, as in (2a), or the direct object as in (2b), or with the other arguments of the verb.

(2) a. Emek$_i$ sıkıntı-sın-ı hep iç-i –ne$_i$ at-ar.
Emek distress-POSS3sg-ACC always inside-POSS3sg-DAT throw-AOR
"Emek$_i$ always throws her$_i$ stress her$_i$ inside" = "Emek always suppresses her distress"
b. O ben-i$_i$ sırt-ım-dan$_i$ vur-du.
She I-ACC back-POSS1sg-ABL shoot-PAST
"She stabbed me in my back' = 'She betrayed me."

Alternatively, the possessive marker can be coindexed with the *possessor* NP, which is a genitive case marked modifier, as in (3). 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), the possessor *Elif* corresponds to the subject of the English translation. It has the semantic role *Experiencer*.

(3) a. [Elif-in$_i$ can-ı$_i$] pasta çek-ti.
Elif-GEN spirit-POSS3sg cake pull-PAST
"Elif's soul wants some cake" = "Elif wants some cake."
b. Eylem [bulasık-lar-ın$_i$ kaba-sı-nı$_i$] al-dı
Eylem dish-PL-GEN base-POSS3sg-ACC take-PAST
"Elif took the base of the dishes" = "Elif cleaned the dishes superficially."

By the same token, *bulaşık-lar* (dishes) is the *possessor* of the direct object in (3b), and it corresponds to the object in the English translation, which has the semantic role *Theme*.

3.2. *Bare noun non-heads*

A bare (non-case marked) non-head can be a categorical 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 (4a) and (4b), there is no overt case marking on the non-head. In (4b) the non-head *kan* is the subject, since it is not possible to insert a subject in the sentence, as shown in (5). However, we can insert the subject *Ahmet* in (6), indicating that the non-head *kafa* in (4a) is an object.

(4) a. birin-e kafa tut-mak
someone-DAT head hold-INF
"to hold head to someone" = "to defy someone"
b. birin-i kan tut-mak
someone-ACC blood hold-INF
"to be hold by blood" = "to be irritated by blood"

(5) * Ahmet kan tut-tu.
Ahmet blood hold-PAST
* "Ahmet held blood."

(6) Ahmet Ali-ye kafa tut-tu.
Ahmet Ali-ACC head hold-PAST.
"Ahmet defied Ali."

(6) is an example of *theme incorporation* as the verb *tutmak* (to hold) assigns its direct object the thematic role *Theme*. Kartal (1995) observed that theme incorporation in compounds headed by unaccusative verbs results in a transitive structure. For example in (7), the compound headed by the unaccusative verb *çıkmak* (appear) subcategorizes for a new dative marked object.

(7) biri-ne destek çık-mak
someone-DAT support appear-INF
"to support-appear someone" = "to support someone"

Theme incorporation occurs with almost all unaccusatives unless the non-head bears a possessive marker. When it occurs, the subcategorization frame is not adequate for representing the valency of the verb. The incorporated theme cannot be included in the frame as the subject, because the transitive structure subcategorizes for a new subject instead of the *Theme*. Moreover, the object position is filled by the new object.

A very similar syntactic change is observed with some transitive verbs. The idiomatic compound headed by a transitive verb subcategorizes for a new direct object in addition to the *Theme* non-head. Just as theme incorporation, the non-head cannot be coded in the subcategorization frame, as the frame already has a new direct object. The incorporated non-head in unaccusatives and the exceptional cases of the sort described are listed outside of the subcategorization frame as defined in the following section.

4. Representing properties of idiomatic compounds

The dictionary is in a spreadsheet file format in which every compound is associated with a subcategorization frame and the values described in this section. The restrictions on the compounds listed in this section are exemplified in Table 2 on the idiomatic compound *kan beynine çıkmak* shown in (8), which involves two non-heads: one in subject position *kan* (blood) and one in the dative object *beynine* (brain-POSS3sg-DAT). It also involves an argument at the possessor position of the dative object (someone).

(8) kan beyn-in-e çık-mak
blood brain-POSS3sg-DAT climb-INF
"blood to climb on the brain of someone" = "someone to rage"

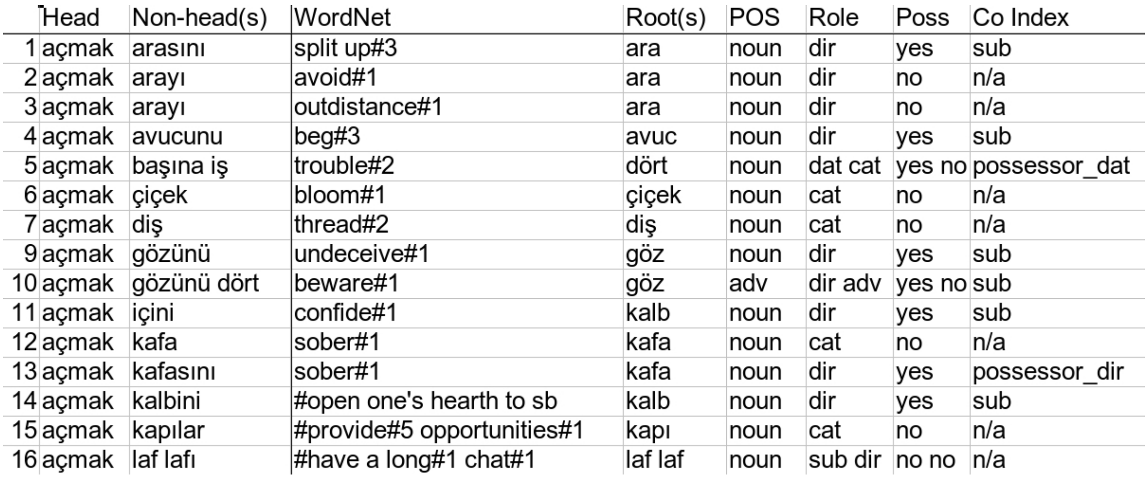
Head	Non-Head(s)	Root	POS	Role	Poss	Co Index
çıkmak	kan beynine	kan beyin	noun noun	sub dat	no yes	possessor_dat

Table 2: Elements of the Compound

- **Head.** The head of the compound is the verb. In case of the compound in Table 2, it is *çıkmak* (climb).
- **Non-head(s).** It is possible for a compound to have more than one non-head, as shown in Table 2.
- **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 reading of the lexicon. As the possessive marker on the non-head should agree with what it is coindexed with, the form of the possessor suffix in the non-head column is not always relevant.

- **Part of Speech (POS).** This section lists the part of speech category of the non-head(s).
- **Syntactic Role.** This column lists the case on the non-head(s) if the non-head is an NP. 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 categorical direct objects. If the non-head occupies the subject position then it is listed as *sub*. Dative, ablative and instrumental/comitative are coded as *dat*, *abl* and *ins* respectively. If the non-head is an adverbial, then it is listed as *adv*. The theme incorporation is coded as *inc*. Finally, exceptional cases described in Section 3.2 are coded as *exc*.
- **Possessive Marker.** This column has the value *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 in Table 2 has “no yes” referring to the order of the words in the non-head column.
- **Co-indexation.** This column lists the constituent 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*. Possessor positions are coded as prefixes on the syntactic roles as specified above.



	Head	Non-head(s)	WordNet	Root(s)	POS	Role	Poss	Co Index
1	açmak	arasını	split up#3	ara	noun	dir	yes	sub
2	açmak	arayı	avoid#1	ara	noun	dir	no	n/a
3	açmak	arayı	outdistance#1	ara	noun	dir	no	n/a
4	açmak	avucunu	beg#3	avuc	noun	dir	yes	sub
5	açmak	başına iş	trouble#2	dört	noun	dat cat	yes no	possessor_dat
6	açmak	çiçek	bloom#1	çiçek	noun	cat	no	n/a
7	açmak	diş	thread#2	diş	noun	cat	no	n/a
9	açmak	gözünü	undeceive#1	göz	noun	dir	yes	sub
10	açmak	gözünü dört	beware#1	göz	adv	dir adv	yes no	sub
11	açmak	içini	confide#1	kalb	noun	dir	yes	sub
12	açmak	kafa	sober#1	kafa	noun	cat	no	n/a
13	açmak	kafasını	sober#1	kafa	noun	dir	yes	possessor_dir
14	açmak	kalbini	#open one's hearth to sb	kalb	noun	dir	yes	sub
15	açmak	kapılar	#provide#5 opportunities#1	kapı	noun	cat	no	n/a
16	açmak	laf lafı	#have a long#1 chat#1	laf laf	noun	sub dir	no no	n/a

Figure 1: An example of the application

Various properties of an idiomatic compound have to be considered together to understand its meaning, which would be difficult and time consuming if these properties would be embedded in the subcategorization frame of the verb. The proposed design exemplified in Figure 1 not only organizes the data in a meaningful way, but also facilitates the retrieval of the statistical information of the compounds in the lexicon.

5. Conclusion

Although the essential elements of an idiomatic compound i.e. the morphological, syntactic, semantic and lexical restrictions on the head and non-heads(s) can be represented in the usual subcategorization schemes that refer to the argument positions of the verbs, it has been observed that these subcategorization schemes are impractical and even inadequate for representing the argument structure of the idiomatic compounds headed by verbs in Turkish. Various properties of an idiomatic compound have to be considered together to understand its meaning, which would be difficult and time consuming if these properties would be embedded in the subcategorization frame of the verb. Therefore, we proposed a design which facilitates the coding and accessing the properties of the compounds in the lexicon. Moreover, the proposed design serves the purpose of representing the non-heads in theme incorporation, and other exceptional cases of syntactic change, which the subcategorization frame of the verb is inadequate to represent.

References

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

Eyigoz, E. (2007). "A Lexicon for Idiomatic Compounds in Turkish". Bogazici University Master's Thesis.

Kartal, G. (1995). "Argument Structure and Idiomatic Compounds in Turkish". Bogazici 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.

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