In Proceedings of NLPRS 2001, pages 525-531

A Modular Approach to Turkish Noun Compounding: The Integration of a Finite-State Model

Aysenur Akyuz Birturk[1] and Sandiway Fong[2]

[1]Department of Computer Engineering
Middle East Technical University, Ankara
birturk@ceng.metu.edu.tr

[2]NEC Research Institute
Princeton, New Jersey
Sandiway@research.nj.nec.com

Abstract

In this paper, we describe the design and integration of a three level cascaded non-deterministic finite state model of Turkish compounding into Turkish PAPPI, a comprehensive syntactic parser in the principles-and-parameters(P&P) framework. Our approach is to handle compounding as an intermediate stage between morphological analysis and syntactic parsing. We discuss how the compounding machine handles bracketing paradoxes and adjectives in compounds and how the non-determinism allows for ambiguity due to different bracketings and Turkish subject/object pro-drop.

1 Introduction

Noun compounding poses a special challenge for natural language processing (NLP) systems because of its productivity and the wide range of possible semantic relations that can be encoded through compounding.

Compounds such as *tea bag, tea leaf, tea garden, tea cake, tea break,* and *tea service* behave like common nouns with respect to syntax but encode different semantic relations between the head and the modifier elements. Furthermore, Downing (1977) claims that there are no linguistic restrictions on the possible relations implicit in nominal compounds, and newly created compounds may be interpreted in many ways with no contextual context, for example, *cousin chair*. Thus compounding is semantically unpredictable.

Semantic compositionality means that the semantics of a complex expression is a function of the semantics of its parts and the mode of combination (Partee, 1984). However, semantic compositionality appears to be violated in many examples:

(1) (a) school girl
girl school
(b) summer school
*school summer

One controversy with respect to compounding is whether they are formed by transformation, e.g. ‘girl friend’ from ‘friend who is a girl’, or by special word formation rules. Important early work on a comprehensive but purely transformational treatment of compounding, encoding a variety of sentence types and grammatical relations, was done by Lees (1966). In the same framework, Botha (1968), arguing from Afrikaans compounding data, showed that phonological considerations must also come into play.

Arguments against the transformational approach include violation of recoverability of deletion and that compounding is acquired before relative clauses formation (see Hoeksema (1985) for further discussion).

Levi (1978) attempted to characterize compounding using a small set of basic semantic relations, e.g. *make*, *for* or *cause*. However, this reductionist approach is of limited practical import given the polysemy inherent in compounding; issues of pragmatics and real-world knowledge must also come into play. Along similar lines, Johnston and Busa (1999) attempted to treat core cases of compounding in the Generative Lexicon framework (Pustejovsky, 1991) by co-opting qualia information present in the lexical entry for the head noun.

Most compounds obey the morphological island constraint of Botha (1980) :

> *The individual constituents of the complex words formed by means of word formation rules lose the ability to interact with inflectional, derivational and syntactic processes.*

For example :

(2)(a) bus ticket
(b) *buses ticket (plural number insertion)

Turkish constitutes a special case for compounding. In particular, the indefinite '*izafet*' construction exhibits many interesting properties as we will see in section 2. However, indefinite '*izafet*' construction in Turkish also obey the morphological island constraint in most cases. This allows us to treat compounding in a modular fashion, as a distinct process separate from the rest of syntax. Our approach is to handle compounding as an intermediate stage between morphological analysis and syntactic parsing. Our goal is to package and identify compounds for syntactic analysis without analyzing the semantic relations between elements of a compound.

The organization of the paper is the following. In Section 2, we give the analysis of noun compounds in Turkish. Problem description is given in section 3. Sections 4 and 5 describe our approach and the design of the compounding machine. We discuss the limitations of the approach in Section 6 and our concluding remarks are in Section 7.

2 Compound Nouns in Turkish

Compound nouns in Turkish are divided into two types (Spencer, 1991; Lewis, 1967; Lapointe, Brentari and Farrell, 1998):

a. Single-word Construction: These are phonologically single words and have idiosyncratic (i.e. non-compositional) meaning. The types of single-word compounds are Noun+Noun, Adj+Noun, Noun+Adj, Verb+Verb and Noun+Verb. The construction is not productive. An example from Spencer(1991) is:

(3) başbakan
head + minister
'prime minister'

b. '*izafet*' Construction: There are two types of '*izafet*' constructions[1].

1. Definite (possessive construction): It takes the form 'Noun-GEN Noun-POSS' and generally corresponds to the English 'Noun's Noun' or 'Noun of the Noun' type syntactic phrases.

(4) bahçenin kapısı
garden-GEN gate-POSS
'the gate of garden'

2. Indefinite: It takes the form 'Noun Noun-POSS' and corresponds to the English 'Noun Noun' compounds.

(5) bahçe kapısı
garden gate-POSS
'garden gate'

Two types of branching are possible in indefinite constructions and branching is explicitly signaled by overt morphology:

1. Right branching ([N [N … [N N-POSS]…]])

(6) Türk Dil Kurumu
Turk Language Organization-POSS
'Turkish Language Council'

2. Left branching ([…[[N N-POSS] N-POSS]…])

(7) Dil Kurumu Sözlüğü
Language Organization-POSS dictionary-POSS
'Language Council Dictionary'

Both type of branching may occur in one construction:

(8) Türk Dil Kurumu Sözlüğü
Turk Language Organization-POSS dictionary-POSS
'Turkish Language Council Dictionary'

In general, the head, i.e. the last noun in the compound, cannot be modified directly and non-heads lose referential and other syntactic properties (see section 5.1.2 for a counterexample):

(9)(a) bahçe kapısı
garden gate-POSS
'garden gate'

(b) *bahçe-ler kapısı
garden-PLR gate-POSS

(c) *bahçe yeni kapısı
garden new gate-POSS

3 Problem Description

As we have described in the previous section, Turkish noun compounding consist of single-word, and indefinite 'izafet' constructions. Our goal is to extend an existing, comprehensive principles-and-parameters (P&P) parser for Turkish, Turkish PAPPI (Birturk, 1998), to accept and parse Turkish noun compounds.

Single-word compounds are lexicalized and thus are inserted whole in the lexicon. The definite 'izafet' construction can be shown to be a syntactic construction and this is already handled by the current syntactic engine employed by Turkish PAPPI. Our main concern in this study is therefore the analysis of indefinite 'izafet' constructions that are also very productive. Hereafter, in this paper, we will restrict our attention to cases of indefinite compound constructions only.

[1] In the following constructions, GEN stands for the genitive suffix, POSS stands for the possessive suffix, and PLR stands for the plural suffix.

4 Pre-parsing Compounds

Our approach is to handle compounding before syntactic parsing begins but after initial morphological analysis (Fig.1). Thus the compounding component serves as a preprocessing phase in Turkish PAPPI. Since compounds are opaque to syntactic processes, i.e. they function as a single syntactic unit once the head has been identified, we can simply treat them as specially-marked nouns inheriting the features of the head noun. We simply 'encapsulate' them as nouns, marked with a special feature 'compound'. In other words, compounds are seen and treated by syntactic components as ordinary nouns. For example, they can take part in syntactic constructions such as possessive constructions, -ki constructions or adjective phrases:

(11) (a) evin [bahçe kapısı]
house-GEN garden gate-POSS
'the garden gate of house'

(b) tahta [bahçe kapısı]
wooden garden gate-POSS
'wooden garden gate'

The task of initial morphological analysis is to expand or decompose input words into their constituent morphemes or tokens in the lexicon. Such a stage of analysis is especially important for agglutinating languages like Turkish where it is impractical to store all forms of a word.

As an example, 'elmalarımı' is expanded into 'elma PLR POSS1SG ACC'. Here, PLR, POSS1SG, and ACC are abstract morphemes (PLR : plural, POSS1SG : 1.singular possessive, ACC : accusative case) that are implemented as 'markers' in the lexicon (Birturk, 1998). Markers are a special class of morphemes that do not project structure like regular heads or other morphemes such as verbal causatives or passives. Instead, markers in the PAPPI system are realizations of feature elements that are attached to regular categories (Fong, 1998), in the sense of the Case insertion mechanism described in Chomsky (1996).

Hence, markers are applied before syntactic parsing; in other words, they're resolved as features or modifications on existing features of heads. In our example, PLR, POSS1SG and ACC simply instantiate the number, possessive and case features of the head noun 'elma'.

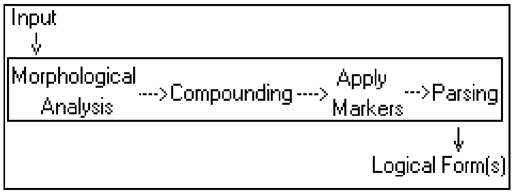

Figure 1. Levels of Linguistic Analysis.

5 A Finite State Machine for Compounding

Chomsky (1957) demonstrated that finite state methods are incapable of representing the full richness of natural languages. However, there are many subsets of natural languages that are adequately covered by finite state methods (Roche and Schabes 1997; Kornai, 1999; Karttunen and Oflazer, 2000). In this section, we assume familiarity with finite state automata, if not the reader is referred to the introduction in Lewis and Papadimitriou (1998).

We demonstrate a successful application of finite state methods in Turkish noun compounding. A cascaded nondeterministic finite state machine (FSM) is designed for compounding. The compounding machine permits left and right branching in compounds, and handles bracketing paradoxes and adjectives in the compounds. The non-deterministic design of the machine allows for ambiguous parses due to subject/object pro-drop in Turkish and different bracketings due to syntax.

The cascaded finite state machine has three levels:

1. Simple Compounding
2. Nested Compounding
3. Title-ProperNoun Compounding

We describe each of the three levels in the following sub-sections.

5.1 Simple Compounding

The initial FSM designed for compounding is shown in Fig.2. Note that this machine permits left and right branching in compounds. Whenever a compound is parsed, the features of the head, i.e. the last noun in the compound, are copied and instantiate the compound's features as a whole. As a consequence, markers (described in section 4) that modify the head noun apply also to the entire compound.

As a rule, we do not employ a compound marker in the lexicon; the POSS marker is used to serve for both compounding and agreement in possessive constructions. This also prevents double marking for compound possessees in possessive constructions:

(12) (a) evin bahçe kapısı
house-GEN garden gate-POSS
'the garden gate of house'

(b) *evin bahçe kapısısı
house-GEN garden gate-POSS-POSS

In order to save branching information, we employ two features, namely RBRANCH and LBRANCH for the compound noun. They are initially set to 0. RBRANCH is incremented each time state 0 or 1 is visited and LBRANCH is incremented each time state 2 is visited.

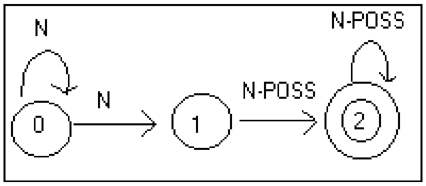

Figure 2. The initial FSM for compounding.

5.1.1 Handling Bracketing Paradoxes

Compounds in Turkish may involve bracketing paradoxes that cannot be resolved by lexical rules (Sehitoglu and Bozsahin, 1999). In the following example, taken from Goksel (1993), the plural marker has semantic scope over the nominal compound marker; however, the opposite is true with respect to the morphological bracketing:

(13) otobüs bilet-ler-i
bus ticket-PLR-POSS
'bus tickets'

We solved this bracketing paradox by considering –leri as a composite plural compound marker and modifying N-POSS labels by N-{PL}-POSS, i.e. N-POSS or N-PLR-POSS.

5.1.2 Adjectives in the compound

Adjectives may also be a part of some noun compounds as in the English examples [[*short story*] *competition*], or [[*natural language*] *parser*]. Similar compounds can be found in Turkish:

(14) kısa film yarışması
short movie competition-POSS
'short movie competition'

The adjective 'kısa' modifies 'film', not '[film yarışması]'. Here we put aside the question of internal bracketing for semantics and simply encapsulate the entire sequence as [kısa film yarışması]. We handle this case in the machine by adding a Adj-Noun link between states 0 and 1 in Fig.2.

5.1.3 Allowing Ambiguity

Examples like *white door handle* or *American history teacher* are ambiguous since one may (a) start at N, and get [Adj white][N door handle], i.e. the handle is white; or (b) start at Adj, and get [N white door handle], interpreted as the door is white. Thus compounding may start with the adjective or noun.

The situation in Turkish is further complicated by the possibility of subject and object-drop (indicated by the zero element ø below). For example, there are three possible parses for the sentence:

(15) $_1$NATO $_2$yaz okulu düzenledi.
NATO summer school organize-PAST3SG
i. 'NATO organized [summer school].'
ii. 'ø organized [NATO summer school].'
iii. 'NATO summer school organized ø.'

Thus compounding may start at positions 1 or 2 (given by subscripting) in the sentence above. This is accommodated by adding a nondeterministic component to the finite state machine. The revised FSM given in Fig.3 handles bracketing paradoxes, adjectival insertion in compounds and ambiguity of the aforementioned kind. ***** stands for any terminal symbol, so the machine may stay at initial state or start compounding with a N or Adj. In general, for a sequence of *m* unmarked nouns followed by *n* POSS-marked nouns, *(m*n)* compounds may be generated.

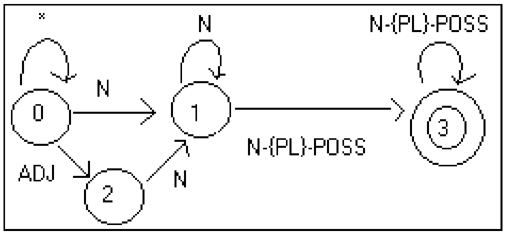

Figure 3. Level 1: Compounding Machine.

As a consequence of the nondeterminism necessary to handle ambiguity, controlling overgeneration becomes an issue. In some cases, illicit compounds can be eliminated by appealing to selectional restrictions. However, in many cases, there are no clear criteria for elimination. Consider, for example:

(16) Zeynep otobüs bileti verdi.
Zeynep bus ticket-POSS give-PAST3SG
Note: *Zeynep* is a personal name in Turkish.
i. 'Zeynep gave bus ticket to ø'
ii. 'ø gave [Zeynep bus ticket]'
iii. '[Zeynep bus ticket] gave ø to ø'

Here, (iii) can be eliminated by selectional restrictions of the verb *give***.** However, (ii) can only be eliminated by appealing to extra-grammatical information such as real-world knowledge or discourse context.

Consider also the following example:

(17) Kennedy havaalanına gitti.
Kennedy airport-POSS-DAT go-PAST3SG
i. 'Kennedy went to airport.'
ii. 'ø went to [Kennedy airport].'

(17ii) is a parallel example to (16ii) above. Hence, in principle, the compounding module is correct in allowing for both possibilities.

5.2 Nested Compounding

Multiple compounding is also a common phenomenon in natural languages. For example:

(18) Middle East Technical University
Computer Engineering Department

In Turkish, nested compounds have the nested structure [[COMPOUND-1] …[COMPOUND-N]].

(19) [[Orta Doğu Teknik Üniversitesi] [Bilgisayar Mühendisliği Bölümü] [yaz okulu kayıtları]] başladı.
[[Middle East Technical University-POSS] [Computer Engineering-POSS Department-POSS] [Summer School registration-PLR-POSS]] start-PAST3SG
'Middle East Technical University Computer Engineering Department Summer School registrations started.'

This is the second step in compounding. After the compounding FSM in Fig.3 has completed the first level of compounding, the newly formed compounds are sent to the FSM in Fig.4 for nested compounds. This machine is also nondeterministic. Thus, we may obtain different bracketings for the following sentence:

(20) Orta Doğu Teknik Üniversitesi Bilgisayar Mühendisliği Bölümü açtı.
Middle East Technical University-POSS Computer Engineering-POSS Department-POSS establish-PAST3SG
i. '[Middle East Technical University Computer Engineering Department] established ø.'
ii. '[Middle East Technical University] established [Computer Engineering Department].'

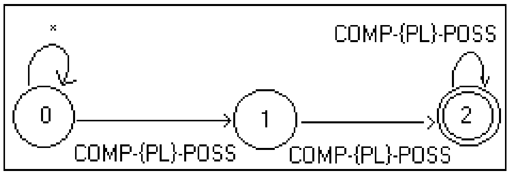

Figure 4. Level 2: Nested Compounding Machine.

5.3 Title-ProperNoun Compounding

Another phenomenon related to compounding is Title-ProperNoun sequences. The title can be a simple noun or a compound as indicated in the following examples:

(21) (a) Başkan Bush
President Bush
'President Bush'

(b) [Turizm Bakanı] Mumcu
Tourism Minister-POSS Mumcu
'Minister of Tourism Mumcu'
Note: *Mumcu* is the surname of the Minister of Tourism.

This is handled by a separate compounding level. After the completion of nested compounding, the nondeterministic FSM given in Fig.5 handles title-properNoun sequences. This machine is also nondeterministic. Thus, we may obtain different parses for the following sentence:

(22) Turizm Bakanı Mumcu'ya bir hediye verdi.
Tourism Minister-POSS Mumcu-DAT a present give-PAST3SG
i.'ø gave a present to Minister of Tourism Mumcu'
ii.'Minister of Tourism gave a present to Mumcu'

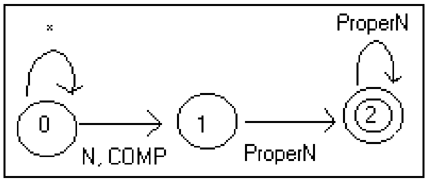

Figure 5. Level 3: Title-ProperN Compounding.

6 Limitations of the Approach

In this section, we discuss limitations of the approach presented so far.

6.1 Conjunction in compounds

Putative compounds like *NY and NJ trains, NY buses and trains, Democracy and Human Rights Report, France and England Territories* involve coordinating conjunctions and ambiguity :

(23) NY ve NJ trenleri
NY and NJ train-PLR-POSS
i.'NY and [NJ buses]'
ii.'[NY and NJ] buses'

Coordination in general is a process at the level of syntax and cannot be handled with our approach at present. It is not even entirely clear whether they should be treated as a compound or handled within a general treatment of conjunction. We leave such cases for future research.

6.2 Other problematic cases

There are some compound like structures which violate the morphological island constraint. For example:

(24) Sokak çocuklarına yardım derneği
Street child-PLR-POSS-DAT help foundation-POSS
'Foundation to help street children'

This is an apparent nested compound in which the first sub-compound is marked by a dative case marker (DAT). The case marker indicates and triggers a syntactic phrase boundary. Thus, instead of modifying the nested compounding machine to accommodate case markers, the example can be shunted aside and handled in the syntactic component after the first level of compounding.

There is a special type of noun compounding in Turkish: [N ADJ-POSS].

(25) dut kurusu
mulberry dry-POSS
'dried mulberry'

The difference in noun-adjective order implies that the level-1 FSM must be revised or perhaps another separate level may be introduced to handle these cases.

7 Conclusion

The cascaded finite state machine described in this paper is a descriptively adequate, powerful and simple yet flexible mechanism for handling simple and multiple compounds before real syntactic parsing begins. Compounds constructed by the machine interact with later syntactic processes as common nouns. Ambiguous parses due to different bracketings and subject/object pro-drop in Turkish are obtained through nondeterminism and the cascaded design of the FSM. An important point is that we do not analyze the semantic relations between elements of a compound. As we have pointed out previously, compounding may encode arbitrarily complex and deep semantic relations, requiring both real-world and discourse knowledge. A comprehensive treatment of this topic is beyond the purview of simple finite state machinery (for further discussion on this topic, see for example Johnston and Busa (1999)). The machine has been tested on a variety of different types of compounds found in Turkish and it demonstrably integrates well as a new component to the existing Turkish PAPPI parser. For example, syntactic constraints such as selectional restrictions help filter out unwanted cases of compounding. However, the possibility of overgeneration due to nondeterminism also points to the need for extra-grammatical constraints. The existing machine cannot by itself deal with overgeneration where discourse or real-world knowledge is involved. To handle these cases, the results of compounding must be further filtered after syntactic analysis at the level of syntactic logical form (LF), the interface to semantic interpretation.

Acknowledgements
The authors thanks David Lebeaux and Piroska Csuri for their help in the analysis of Turkish compounds. We would also like to thank Cem Bozsahin and David Lebeaux for valuable comments on this paper. Of course, the responsibility for any errors and omissions lies with us.

References

Birturk A. 1998. *A Computational Analysis of Turkish using the Government-Binding Approach*. Ph.D. thesis, Middle East Technical University, Ankara.

Botha R.P. 1968. *The Function of the Lexicon in Transformational Generative Grammar*. Mouton, The Hague.

Botha R.P. 1980. *Word-based Morphology and Synthetic Compounding*, Stellenbosch Papers in Linguistics 5, Stellenbosch University.

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

Chomsky, N. 1996. *Knowledge as Language*, Prager.

Downing, P. 1977. On the Creation and Use of English Compounds, *Language 53*.

Fong, S. 1998, *The PAPPI Reference Manual*, available at http://www.neci.nj.nec.com/homepages/sandiway/pappi/doc/refman.

Goksel, A. 1993. *Levels of Representation and Argument Structure in Turkish*. Ph.D. Thesis, SOAS.

Hoeksema, J. 1985. *Categorial Morphology*. Garland, New York.

Johnston, M. and Busa F. 1999. Qualia Structure and the Compositional Interpretation of Compounds. In *Breadth and Depth of Semantic Lexicons*, E. Viegas (ed), Kluwer.

Karttunen L. and Oflazer K.(eds.). 2000. *Computational Linguistics: Special Issue on Finite-State Methods in NLP*, Vol.26, No.1., MIT Press.

Kornai, A. (ed). 1999. *Extended Finite State Models of Language*. Cambridge University Press.

Lapointe S.G., Brentari D.K., and Farrell P.M.(eds). 1998. *Morphology and Its Relation to Phonology and Syntax*. CSLI Publications.

Lees, R.B. 1966. *The grammar of English nominalizations*. Bloomington and the Hague.

Levi, J.N. 1978. *The Syntax and Semantics of Complex Nominals.* Academic Press, NY.

Lewis, G.L. 1967. *Turkish Grammar*. Oxford Press.

Lewis H. and Papadimitriou C. 1998. *Elements of the Theory of Computation*. 2.edition. Prentice-Hall.

Partee B. 1984. Compositionality. In *Varieties of Formal Semantics*, Landman F., Veltman F. (eds), Foris Publications.

Pustejovsky, J. 1991. The Generative Lexicon. *Computational Linguistics*, 17.4.

Roche, Emmanuel, and Yves Schabes (eds). 1997. *Finite-State Language Processing*, MIT Press, Cambridge, MA.

Sehitoglu O., Bozsahin C. 1999. Lexical Rules and Lexical Organization. In *Breadth and Depth of Semantic Lexicons*, E. Viegas (ed), Kluwer.

Spencer, A. 1991. *Morphological Theory*, Blackwell Publishers.

Sproat, R. 1992. *Morphology and Computation*. The MIT Press.