A Categorial Framework for Composition in Multiple Linguistic Domains

H. Cem Bozşahin and **Elvan Göçmen**
Laboratory for the Computational Studies of Language
Computer Engineering Department
Middle East Technical University
06531 Ankara, Turkey
{bozsahin,elvan}@lcsl.metu.edu.tr

Abstract

We describe a computational framework for a grammar architecture in which different linguistic domains such as morphology, syntax, and semantics are treated not as separate components but compositional domains. Word and phrase formation are modeled as uniform processes contributing to the derivation of the semantic form. The morpheme, as well as the lexeme, has lexical representation in the form of semantic content, tactical constraints, and phonological realization. The model is based on Combinatory Categorial Grammars.

1 Introduction

The division of morphology and syntax in agglutinative languages is difficult compared to relatively more isolating languages. For instance, in Turkish, there is a significant amount of interaction between morphology and syntax. Typical examples are: causative suffixes change the valence of the verb, and the reciprocal suffix subcategorize the verb for a noun phrase marked with the comitative case. Moreover, the head that a bound morpheme modifies may be not its stem but a compound head crossing over the word boundaries, e.g.,

(1) *iyi oku-muş çocuk*
well read-REL child
'well-educated child'

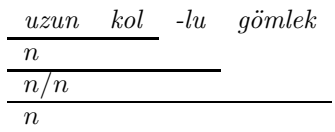
In (1), the relative suffix *-muş* (in past form of subject participle) modifies [*iyi oku*] to give the scope [[[*iyi oku*]*muş*] *çocuk*]. If syntactic composition is performed after morphological composition, we would get compositions such as [*iyi* [*okumuş çocuk*]] or [[*iyi okumuş*] *çocuk*], which yield ill-formed semantics for this utterance.

As pointed out by Oehrle [5, 6], there is no reason to assume a layered grammatical architecture which has linguistic division of labor into components acting on one domain at a time. As a computational counterpart of this idea, rather than treating morphology, syntax and semantics in a cascaded manner, we integrate the process models of morphology and syntax, providing

[1]This research is supported in part by grants from Scientific and Technical Research Council of Turkey (contract no. EEEAG-90), NATO Science for Stability Programme (contract name TU-LANGUAGE), and METU Graduate School of Applied Sciences.

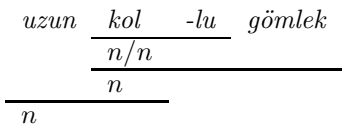
lexical entry	syntactic category	semantic category
uzun	n/n	$\lambda p.long(p(z))$
kol	n	$\lambda x.sleeve(x)$
-lu	$(n/n) \setminus n$	$\lambda q.\lambda r.r(y, has(q))$
gömlek	n	$\lambda w.shirt(w)$

(1a)



$shirt(y, has(long(sleeve(z)))) =$ 'a shirt with long sleeves'

(1b)



$long(shirt(y, has(sleeve(z)))) =$ 'a long shirt with sleeves'

Figure 1: Scope ambiguity of a nominal bound morpheme

semantic composition in parallel. The model, which is based on Combinatory Categorial Grammars (CCG) [1, 9], uses the morpheme as the building block of composition at all three linguistic domains.

2 Morpheme-based Compositions

When the morpheme is given the same status as the lexeme in terms of its lexical, syntactic, and semantic contribution, the distinction between the process models of morphotactics and syntax disappears. In this case, new scoping problems arise in word and phrase formation.

CG accounts of scoping problems concentrate on syntactic and semantic issues such as quantifier scoping [6, 7]. In word formation, morphological bracketing paradoxes are introduced by lexicalized composite affixes which require mixed compositions [4]. However, the scoping problems in morphosyntax go beyond bracketing paradoxes as they may also produce different semantic forms. Consider the example in (2):

(2) *uzun kol-lu* *gömlek*
long sleeve-ADJ shirt

Two different compositions[2] in CCG formalism are given in Figure 1. Both interpretations are plausible, with (1a) being the most likely in the absence of a long pause after the first adjective. To account for both cases, the suffix *-lu*

[2]derived and basic categories in the examples are in fact feature structures; see section 4. We use $\frac{x \quad y}{z}$ to denote the combination of categories x and y giving the result z

lexical entry	syntactic category	semantic category
kadın	n	$\lambda x_1.female(x_1)$
-a	$(s \setminus n)/(s \setminus n) \setminus n$	$\lambda p_3.\lambda p_2.p_2(to(p_3(m)))$
dön	$(s \setminus n) \setminus n$	$\lambda x_4.\lambda x_5.turn(x_5, x_4)$
-erek	$(s \setminus n)/(s \setminus n) \setminus ((s \setminus n) \setminus n)$	$\lambda p_6.\lambda p_7.p_7(\lambda x_6.(x_6, by(p_6(x_6))))$
konuş	$s \setminus n$	$\lambda x_8.speak(x_8)$
-tu	$(s \setminus n) \setminus (s \setminus n)$	$\lambda p_9.p_9(z, past)$

kadın	*-a*	*dön*	*-erek*	*konuş*	*-tu*
$(s \setminus n)/(s \setminus n)$					
$(s \setminus n) \setminus n$					
$(s \setminus n)/(s \setminus n)$					
$s \setminus n$					
$s \setminus n$					

$\lambda x_6.speak(x_6, by(turn(x_6, to(female(m))))$ = 'facing the lady, (he/she) talked'

Figure 2: Composition with a verbal bound morpheme

must be allowed to modify the head it is attached to (e.g., 1b in Figure 1), or a compound head encompassing the word boundaries (e.g., 1a in Figure 1).

Example (3) shows a composition with a verbal head. Figure 2 depicts the CCG treatment of this example. The verb *konuş* does not subcategorize for a dative noun phrase (cf. example 3b); *kadına* is the argument of *dön*. In this case, the adverbial suffix *-erek* must modify [*kadına dön*] to obtain the correct reading.

(3) a. *kadın-a* *dön-erek* *konuş-tu*
woman-DATIVE turn-ADV talk-TENSE
'Facing the lady, (he/she) talked.'

b. * *kadına konuştu*

3 Multi-domain Combination Operator

Oehrle [5] describes a model of multi-dimensional composition in which every domain D_i has an algebra with a finite set of primitive operations F_i. As indicated by Turkish data in sections 1 and 2, F_i may in fact have a domain larger than—but compatible with—D_i.

In order to perform morphological and syntactic compositions in a unified (monostratal) framework, the slash operators of categorial grammar must be enriched with the knowledge about the type of process and the type of morpheme. We adopt a representation similar to Hoeksema and Janda's [2] notation for the operator. The 3-tuple (*direction, morpheme type, process type*) indicates direction[3] (left, right, unspecified), morpheme type (free, bound), and the type

[3] we have not yet incorporated into our model the word-order variation in syntax. See [3] for a CCG based approach to this phenomenon.

of morphological or syntactic attachment (e.g., affixation, syntactic concatenation, reduplication[4], clitic). Examples of different operator combinations are as follows:

Operator	Morpheme	Example
< \, bound, clitic>	*de*	*Ben de yaz-ar-ım* I too write-TENSE-PERS 'I write too.'
< \, bound, affix>	*-de*	*Ben-de kalem var* I-LOCATIVE pen exist 'I have a pen.'
< /, bound, redup>	*ap-*	*ap-açık durum* INT-clear situation 'Very clear situation'
< /, free, concat>	*uzun*	*uzun yol* long road 'long road'
< \, free, concat>	*başka*	*bu-ndan başka* this-ABLATIVE other 'other than this'
<|, free, concat>	*oku*	*adam kitab-ı oku-du* man book-ACC read-TENSE or *adam okudu kitabı* 'The man read the book'

4 Information Structure and Tactical Constraints

Entries in the categorial lexicon have tactical constraints, grammatical and semantic features, and phonological representation. Similar to HPSG [8], every entry is a signed attribute-value matrix.

Syntactic and semantic information are of grammatical (g) sign and semantic (s) sign, respectively. These properties include agreement features such as

[4] intensifiers such as *ap-* and *bes-* in *ap-açık* and *bes-belli* may appear as prefixes but they are in fact reduplicated from the first syllable of the stem

person, number, and possessive, and selectional restrictions:

$$g\begin{bmatrix}\text{CAT} & \\ & \text{NPROP}\begin{bmatrix}\text{PERSON}\\ \text{NUMBER}\\ \text{POSS}\\ \text{CASE}\\ \text{RELATIVE}\\ \text{FORM}\end{bmatrix}\\ & \text{VPROP}\begin{bmatrix}\text{REFLEXIVE}\\ \text{RECIPROCAL}\\ \text{CAUSATIVE}\\ \text{PASSIVE}\\ \text{TENSE}\\ \text{MODAL}\\ \text{ASPECT}\\ \text{PERSON}\\ \text{FORM}\end{bmatrix}\\ \text{RESTR} & \text{<list of conditions>}\end{bmatrix} \qquad s\begin{bmatrix}\text{TYPE} & \\ \text{FORM} & \\ \text{RESTR} & \text{<list of conditions>}\end{bmatrix}$$

Basic and derived categories of CG are of p (property) or f (function) sign, respectively.

$$p\begin{bmatrix}\text{SYN}\\ \text{SEM}\end{bmatrix} \qquad f\begin{bmatrix}\text{RES}\\ \text{OP}\\ \text{ARG}\end{bmatrix}$$

RES-OP-ARG is the categorial notation for the element. Every RES and ARG feature has an f or p sign.

Lexical and phrasal elements have functional representation (f or p sign) and the PHON feature. PHON represents the phonological string. Lexical elements may have (a) phonemes, (b) meta-phonemes such as H for high vowel, and D for a dental stop whose voicing is not yet determined, and (c) optional segments, e.g., *-(y)lA*, to model vowel/consonant drops, in the PHON feature. During composition, the surface forms of composed elements are mapped and saved in PHON. PHON also allows efficient lexicon search. For instance, the causative suffix *-DHr* has eight different realizations but only one lexical entry.

A special feature value called `none` is used for imposing certain morphotactic constraints. For instance, most of the inflectional morphemes of Turkish have the category X\X where X is the category of the stem. `none` is used to make sure that the stem is not inflected with the same feature more than once; it also ensures, through SYN constraints, that inflections are marked in the right order. A sample lexicon entry for a derivational suffix is given in Figure 3. For composition, we use a generalized LR parser [11] in which CCG rules are encoded as recursive rewrite rules with equational constraints.

5 Conclusion

Turkish is a language in which grammatical functions can be marked morphologically (e.g., case), or syntactically (e.g., indirect objects). Semantic composition is also affected by the interplay of morphology and syntax, for instance the change in the scope of modifiers and genitive suffixes, or valency and thematic role change in causatives. To model interactions between domains, we propose a categorial approach in which composition in all domains proceed in parallel. In the domain of phonology, there are categorial accounts of prosody [10] and voice assimilation [12]. Our treatment of phonology is not yet integrated into the uniform grammar architecture. Morphophonemic processes such as vowel harmony and devoicing are modeled as mappings from the operator and the phonological strings to surface forms. Integrating categorial phonology into the architecture will help restore the modularity of processing at all domains.

References

[1] A. E. Ades and M. Steedman. 1982. On the order of words. *Linguistics and Philosophy*, 4:517–558.

[2] Jack Hoeksema and Richard D. Janda. 1988. Implications of process-morphology for categorial grammar. In R. T. Oehrle, E. Bach, and D. Wheeler, editors, *Categorial Grammars and Natural Language Structures*, D. Reidel, Dordrecht, 1988.

[3] Beryl Hoffman. 1992. A CCG approach to free word order languages. In *Proceedings of the 30th Annual Meeting of the ACL*, Student Session, 1992.

[4] Michael Moortgart. 1988. Mixed Composition and Discontinuous Dependencies. In R. T. Oehrle, E. Bach, and D. Wheeler, editors, *Categorial Grammars and Natural Language Structures*, D. Reidel, Dordrecht, 1988.

[5] Richard T. Oehrle. 1988. Multi-dimensional compositional functions as a basis for grammatical analysis. In R. T. Oehrle, E. Bach, and D. Wheeler, editors, *Categorial Grammars and Natural Language Structures*, D. Reidel, Dordrecht, 1988.

[6] Richard T. Oehrle. 1994. Term-labeled Categorial Type Systems. *Linguistics and Philosophy*, 17:633–678.

[7] Fernando C. N. Pereira. 1990. Categorial Semantics and Scoping. *Computational Linguistics*, 16(1):1–10.

[8] C. Pollard and I. A. Sag. 1994. *Head-driven Phrase Structure Grammar.* University of Chicago Press.

[9] Mark Steedman. 1985. Dependencies and coordination in the grammar of Dutch and English. *Language*, 61:523–568.

[10] Mark Steedman. 1991. Structure and intonation. *Language*, 67:260–296.

[11] Masaru Tomita. 1987. An Efficient Augmented-Context-Free Parsing Algorithm. *Computational Linguistics*, 13(1-2):31–46.

[12] Deirdre Wheeler. 1988. Consequences of some categorially-motivated phonological assumptions. In R. T. Oehrle, E. Bach, and D. Wheeler, editors, *Categorial Grammars and Natural Language Structures*, D. Reidel, Dordrecht, 1988.

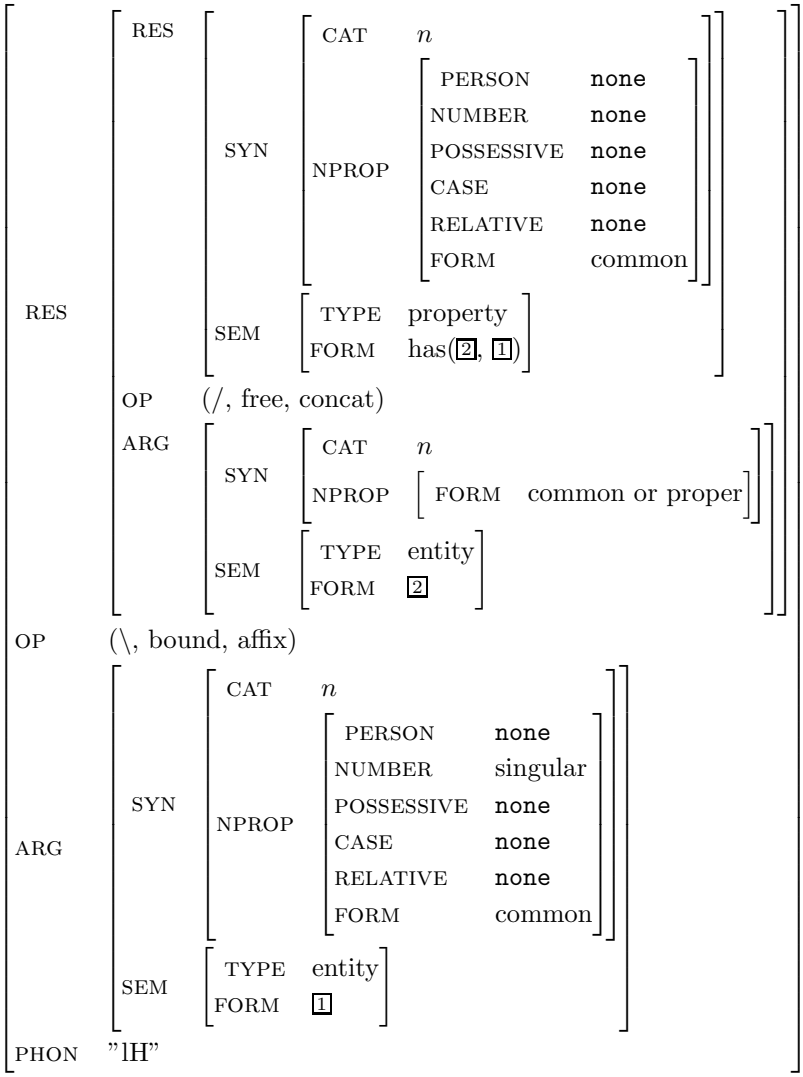


Figure 3: Lexicon entry for *-lH*.