A Computational Framework for Composition in Multiple Linguistic Domains

Elvan Göçmen
Computer Engineering Department
Middle East Technical University
06531, Ankara, Turkey
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. The framework is based on Combinatory Categorial Grammars and it uses the morpheme as the basic building block of the categorial lexicon.

1 Introduction

In this paper, we address the problem of modelling interactions between different levels of language analysis. In agglutinative languages, affixes are attached to stems to form a word that may correspond to an entire phrase in a language like English. For instance, in Turkish, word formation is based on suffixation of derivational and inflectional morphemes. Phrases may be formed in a similar way (1).

(1) *Yoksul-laş-tır-ıl-makta-lar*
poor-V-CAUS-PASS-ADV-PERS
'(They) are being made poor (impoverished)'.

In Turkish, there is a significant amount of interaction between morphology and syntax. For instance, 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.,

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

In (2), 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 (1988), 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 framework, rather than treating morphology, syntax and semantics in a cascaded manner, we propose an integrated model to capture the high level of interaction between the three domains. The model, which is based on Combinatory Categorial Grammars (CCG) (Ades and Steedman, 1982; Steedman, 1985), 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. Consider the example in (3).

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

Two different compositions[1] 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* 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).

3 Multi-domain Combination Operator

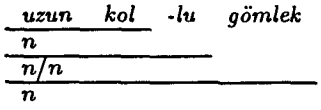
Oehrle (1988) describes a model of multi-dimensional composition in which every domain D_i has an algebra with a finite set of primitive operations

[1] Derived and basic categories in the examples are in fact feature structures; see section 4.

We use $\frac{x}{z}\ \frac{y}{}$ to denote the combination of categories x and y giving the result z.

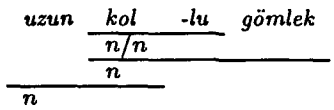
lexical entry	syntactic category	semantic category
uzun	n/n	$\lambda p.long(p(z))$
kol	n	$\lambda x.sleeve(x)$
-lu	$(n/n) \setminus n$	$\lambda q.\lambda r.r(y, has(q))$
gömlek	n	$\lambda w.shirt(w)$

(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

Operator	Morp.	Example
< \, bound, clitic>	*de*	*Ben de git-ti-m* I too go-TENSE-PERS 'I went 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>	*gör*	*kız kedi-yi gör-dü* girl cat-ACC see-TENSE or *kız gördü kediyi* 'The girl saw the cat'

Figure 2: Operators in the proposed model.

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 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 (1988) notation for the operator. The 3-tuple *<direction, morpheme type, process type>* indicates direction[2] (left, right, unspecified), morpheme type (free, bound), and the type of morphological or syntactic attachment (e.g., affix, clitic, syntactic concatenation, reduplication). Examples of different operator combinations are given in Figure 2.

4 Information Structure and Tactical Constraints

Entries in the categorial lexicon have tactical constraints, grammatical and semantic features, and phonological representation. Similar to HPSG (Pollard and Sag, 1994), every entry is a signed attribute-value matrix. Lexical and phrasal elements are of the following f (function) sign:

$$f\begin{bmatrix}\text{res}\\ \text{op}\\ \text{arg}\\ \text{phon}\end{bmatrix}$$

res-op-arg is the categorial notation for the element. phon represents the phonological string. Lexical elements may have (a) phonemes, (b) metaphonemes such as H for high vowel, and D for a dental 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. Every res and arg feature has an f or p (property) sign:

$$p\begin{bmatrix}\text{syn}\\ \text{sem}\end{bmatrix}$$

syn and sem are the sources of grammatical (g sign) and semantic (s sign) properties, respectively. These properties include agreement features such as person, number, and possessive, and selectional re-

[2]We have not yet incorporated into our model the word-order variation in syntax. See (Hoffman, 1992) for a CCG based approach to this phenomenon.

strictions:

$$g\begin{bmatrix}\text{cat} & \\ \text{nprop} & \begin{bmatrix}\text{person}\\ \text{number}\\ \text{poss}\\ \text{case}\\ \text{relative}\\ \text{form}\end{bmatrix} \\ \text{vprop} & \begin{bmatrix}\text{reflexive}\\ \text{reciprocal}\\ \text{causative}\\ \text{passive}\\ \text{tense}\\ \text{modal}\\ \text{aspect}\\ \text{person}\\ \text{form}\end{bmatrix} \\ \text{restr} & \langle\text{cond}\rangle\end{bmatrix} \qquad s\begin{bmatrix}\text{type} & \\ \text{form} & \\ \text{restr} & \langle\text{cond}\rangle\end{bmatrix}$$

A special feature value called none is used for imposing certain morphotactic constraints, and 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 (cf., Figure 3).

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. As an implementation, we have been working on the modelling of Turkish causatives using this framework.

6 Acknowledgements

I would like to thank my advisor Cem Bozsahin for sharing his ideas with me. 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.

$$\begin{bmatrix}\text{res} & \begin{bmatrix}\text{res} & \begin{bmatrix}\text{syn} & \begin{bmatrix}\text{cat} & n \\ \text{nprop} & \begin{bmatrix}\text{person} & \text{none}\\ \text{number} & \text{none}\\ \text{possessive} & \text{none}\\ \text{case} & \text{none}\\ \text{relative} & \text{none}\\ \text{form} & \text{common}\end{bmatrix}\end{bmatrix} \\ \text{sem} & \begin{bmatrix}\text{type} & \text{property}\\ \text{form} & \text{has}(\boxed{2},\boxed{1})\end{bmatrix}\end{bmatrix} \\ \text{op} & \langle /, \text{free}, \text{concat}\rangle \\ \text{arg} & \begin{bmatrix}\text{syn} & \begin{bmatrix}\text{cat} & n \\ \text{nprop} & [\text{form} \quad \text{com. or prop.}]\end{bmatrix} \\ \text{sem} & \begin{bmatrix}\text{type} & \text{entity}\\ \text{form} & \boxed{2}\end{bmatrix}\end{bmatrix} \\ \text{phon} & \end{bmatrix} \\ \text{op} & \langle \backslash, \text{bound}, \text{suffix}\rangle \\ \text{arg} & \begin{bmatrix}\text{syn} & \begin{bmatrix}\text{cat} & n \\ \text{nprop} & \begin{bmatrix}\text{person} & \text{none}\\ \text{number} & \text{singular}\\ \text{possessive} & \text{none}\\ \text{case} & \text{none}\\ \text{relative} & \text{none}\\ \text{form} & \text{common}\end{bmatrix}\end{bmatrix} \\ \text{sem} & \begin{bmatrix}\text{type} & \text{entity}\\ \text{form} & \boxed{1}\end{bmatrix}\end{bmatrix} \\ \text{phon} & \text{"lH"}\end{bmatrix}$$

Figure 3: Lexicon entry for *-lH*.

References

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

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.

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

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.

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

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