What do we parse when we parse?

Cem Bozşahin

Computer Engineering & Cognitive Science
Middle East Technical University (METU), Ankara
bozsahin@metu.edu.tr

April 3, 2008, Boğaziçi University

contributors:
Berfin Aktaş, Jason Baldridge, Orkan Bayer, Ruken Çakıcı, Onur Çobanoğlu, Çağrı Çöltekin, Güneş Erkan, Burcu Karagöl-Ayan, Aysun Kunduracı, Kürşad Kurt, Mark McConville, Umut Özge, Müge Sevinç, Mark Steedman, Michael White, Murat Yasavul

(1)a. (Maymun) (elmayı yedi).

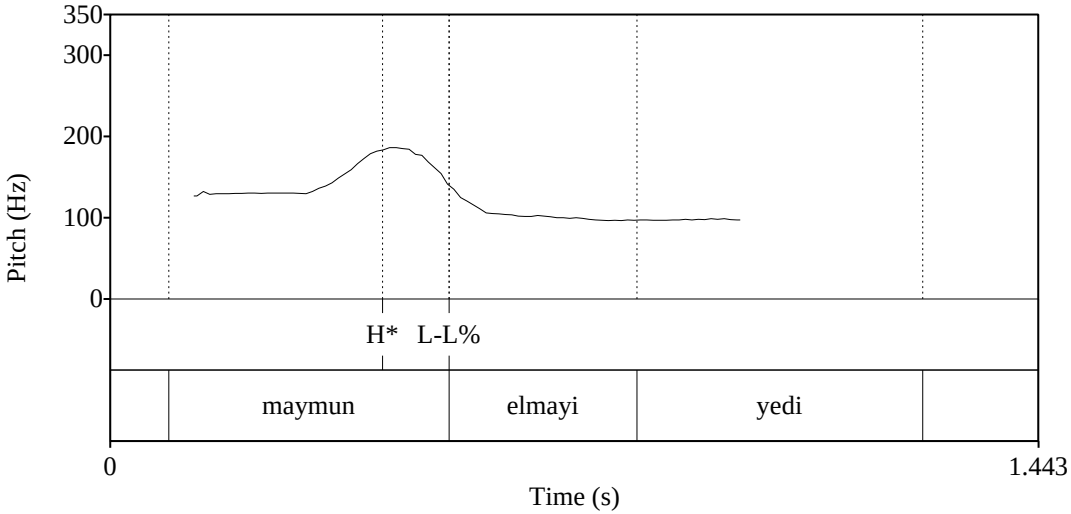
monkey.NOM apple-ACC ate

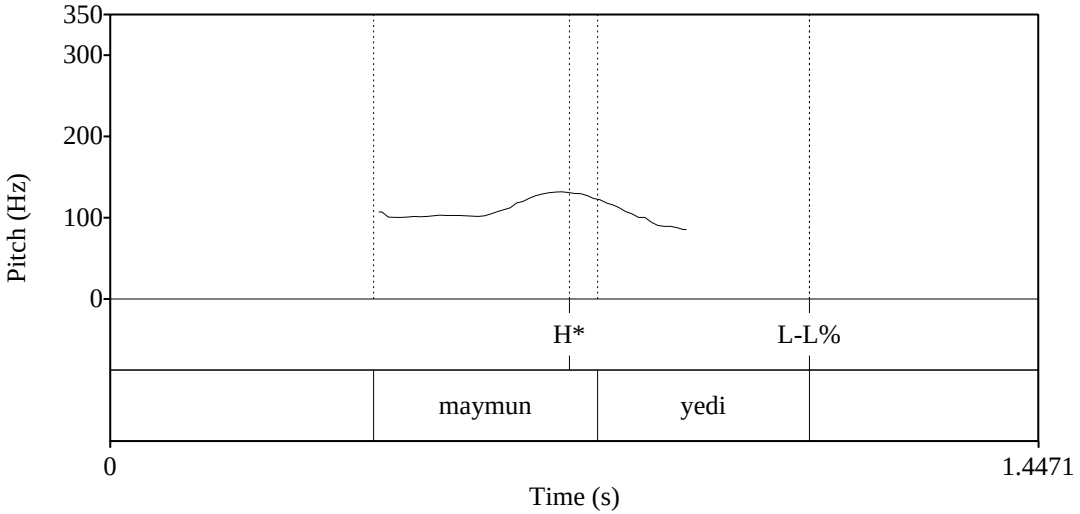
H* L-L%

‘The monkey ate the apple.’

b. (Maymun yedi). ‘The monkey ate (it).’

H* L-L%


350
300
200
100
0
Pitch (Hz)
H*
L-L%
maymun
elmayi
yedi
0
1.443
Time (s)


350
300
200
100
0
Pitch (Hz)
H*
L-L%
maymun
yedi
0
1.4471
Time (s)

Pitch accents and boundary tones are distinct prosodic events (Pierrehumbert, 1980)

H* : local maximum on stressed syllable

L- : Intermediate phrase boundary

L% : Intonational phrase boundary

Turkish has a pitch accent system:

(Demircan, 1996; Ergenç, 1989; Van Der Hulst and Van De Weijer, 1991; Johanson, 1998; Kornfilt, 1997; Levi, 2005; Lewis, 2000; Nash, 1973; Özsoy, 2004; Selen, 1973; Üçok, 1951; Underhill, 1976)

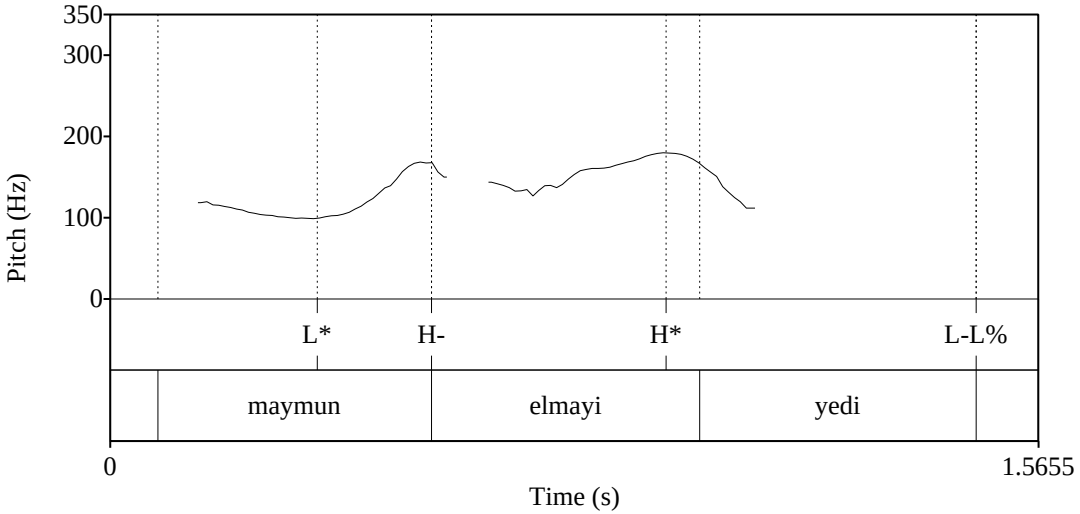
Pitch accent is different than stress: It is **movement** of pitch associated with stressed syllable.

H*, L*, L+H*, H*+L, L*+H

L+H*: *Marcel proved COMPLETENESS*

Word stress without high pitch accent:

(2) (Maymun) (elma-YI ye-di.)
L* H- H* L-L%


350
300
200
100
0
Pitch (Hz)
L*
H-
H*
L-L%
maymun
elmayi
yedi
0
1.5655
Time (s)

H*+L

(3)a. Siz-in-ki-ler nerede?

you-GEN-REL-PLU where

‘Where are your parents?’

b. *Theme* *Rheme*

(Annem) (ARABAYA biniyor.)

kontrast

H*+L H- H* L-L%

‘My mother is getting on the car.’

L+H*

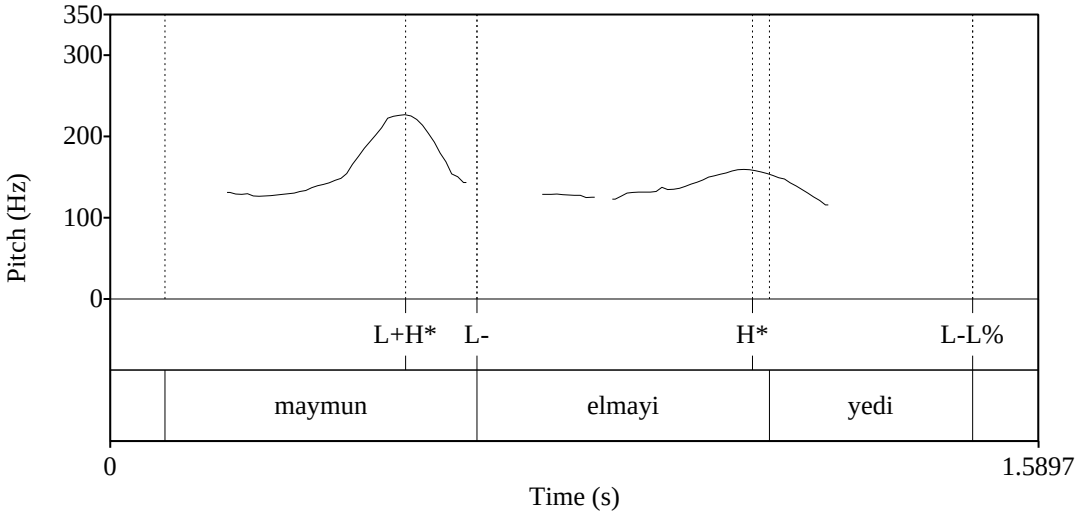
(4)a. (mayMUN) (elma-YI ye-di.)

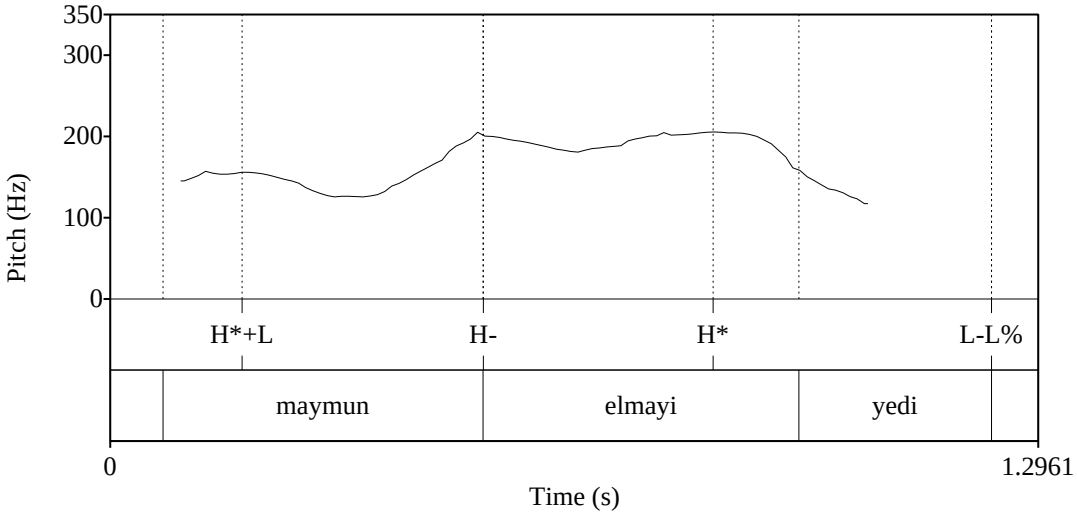
L+H* L- H* L-L%

b. (MAYmun) (elma-YI ye-di.)

H*+L H- H* L-L%

Not truth-conditionally equivalent.


350
300
200
100
0
Pitch (Hz)
L+H*
L-
H*
L-L%
maymun
elmayi
yedi
0
1.5897
Time (s)

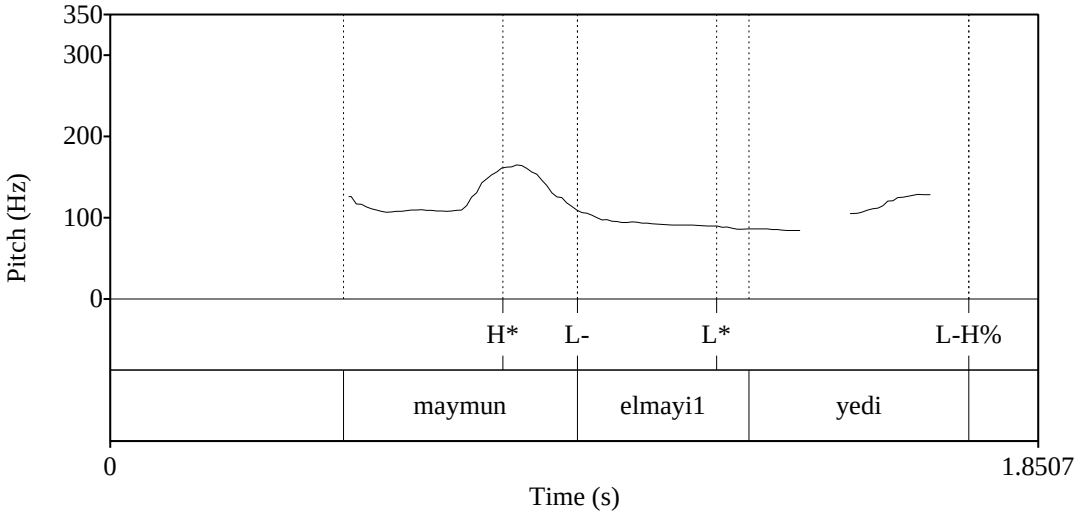

350
300
200
100
0
Pitch (Hz)
H*+L
H-
H*
L-L%
maymun
elmayi
yedi
0
1.2961
Time (s)

When do we complete an utterance?

(5) (Maymun) (elmayı yedi.)

H* L- L* L-H%

Main predicate is *and′*, not *eat′*.


350
300
200
100
0
Pitch (Hz)
H*
L-
L*
L-H%
maymun
elmayi1
yedi
0
1.8507
Time (s)

L-, H-, L% are autonomous strings.

H*, L* are not. They align with syllables of words (segmental).

H*, L* bear on information structure of the words (rheme, theme).

L-, H-, L% are lexical items. They carry theme/rheme to complete intonational phrases.

Words and affixes

(6)a. [*Ziyaretçi ve misafir*]*-ler-de-ki* *hediye-ler*

Visitor and guest-PLU-LOC-REL gift-PLU

lit. The gifts, the ones at the visitors and of the guests
'the gifts at the visitors and at the guests'

b. **Ziyaretçi hediyeler*

c. *Çocuk-lar-ın* *ziyaretçi ve misafir-ler-de-ki* *hediye-ler-i*

child-PLU-GEN.3s visitor and guest-PLU-LOC-REL gift-PLU-COMP.3s

'the gifts of the children at the visitors and the guests'

Clitics

(7)a. [*Sinema-ya* *gid-er*
kitap *oku-y-abil-ir*
ya da *dans* *ed-ebil-ir*]*-mi-y-di-k*?
Cinema-DAT go-AOR
book read-COP-ABIL-AOR
or dance do-ABIL-AOR-QUES-COP-PAS
'Could we go to a movie, read a book or dance?' (Kabak, 2006, ex.20)

Annen kitap, sen de dergi okudun bütün gün.

**Annen kitap, sen dergi de okudun bütün gün.*

**Annen kitap, sen dergi okudun da bütün gün.*

Arabalarla

**Arabalalar*

Evdeler mi? Evde miler?

Constituencies

(8)a. Ben, kapı-yı ALİ kır-dı zanned-iyor-du-m.
I door-ACC A break-PAST think-PROG-PAST-1sg
'I thought Ali crashed the door.'

b. Hayır, (PENCERE-Yİ Ali), (kapı-yı) (MEHMET kır-dı.)
No window-ACC A door-ACC M break-PAST
'No, Ali crashed the window, and Mehmet, the door.'

(9)a. Ben, kapı-yı ALİ kır-dı zanned-iyor-du-m.
I door-ACC A break-PAST think-PROG-PAST-1sg
'I thought Ali crashed the door.'

b. Hayır, (PENCERE-Yİ) (Ali kır-dı), (kapı-yı) (MEHMET)
No window-ACC A break-PAST door-ACC M
'No, Ali crashed the window, and Mehmet, the door.'

Interjections

The man, I claim, works for the government.

**The, I can claim with confidence, man works for the government.*

**This man I, claim works for the government.*

Tüm hayallerimizin, biraz düşününce, gerçekleşemeyeceğini anladık.

**Tüm, biraz uyuyup düşününce, hayallerimizin gerçekleşemeyeceğini anladık.*

CCG

Combinatory Categorial Grammar: Steedman (1996, 2000).

Steedman and Baldridge (2007): summary of recent advances.

Hockenmaier and Steedman (2007): wide-coverage parsing.

Çakıcı (2008): Turkish WCP.

All constraints on grammatical meaning composition must bear on a single dynamic aspect of grammar: Its **syntactic types**.

Syntactic derivation is purely syntactic type-driven.

Lexicalizing a Surface Grammar: A Game of Algebra

PSGs encode constituency and order, but some information will be redundantly specified:

S	$\rightarrow$	**NP VP**	
VP	$\rightarrow$	$\mathbf{V}_{\mathrm{iv}}$	
VP	$\rightarrow$	$\mathbf{V}_{\mathrm{tv}}$ **NP**	
$\mathbf{V}_{\mathrm{iv}}$	$\rightarrow$	*slept*	$e \mapsto t$
$\mathbf{V}_{\mathrm{tv}}$	$\rightarrow$	*read*	$e \mapsto (e \mapsto t)$

read's subcategorization for an object **NP** is defined twice: in its lexical category ($\mathbf{V}_{\mathrm{tv}}$) and in the **VP** rule.

NB. Its (normalized) semantic type is non-redundantly specified.

VP is the functor, which, applied to a leftward **NP**, yields an **S**; i.e. **VP = (S\NP)**

$\mathbf{V}_{\mathrm{iv}}$ **= VP = (S\NP)** $\mathbf{V}_{\mathrm{tv}}$ **= VP/NP = (S\NP)/NP**

Therefore, the only ineliminable parts of the grammar above are *slept* and *read*:

slept $\stackrel{\text{def}}{=}$ $\mathbf{V}_{\text{iv}}$ **= (S\NP)**

read $\stackrel{\text{def}}{=}$ $\mathbf{V}_{\text{tv}}$ **= (S\NP)/NP**

Assuming **S** to be of semantic type t, and **NP** to be e, a fully interpretable equivalent of the grammar above is

$$\begin{array}{lcll} slept & := & \mathbf{S\backslash NP} & e \mapsto t \\ read & := & (\mathbf{S\backslash NP})/\mathbf{NP} & e \mapsto (e \mapsto t) \end{array}$$

This much type-dependence can explain Ross's (1967) ATB constraints,

their exceptions,

and exceptions to exceptions (Steedman, 2000, 2008:p.c.).

(10) a. The cat that [John admires]$_{\mathbf{S/NP}}$ and [Mary hates]$_{\mathbf{S/NP}}$

b. *The cat that [John admires]$_{\mathbf{S/NP}}$ and [bites Mary]$_{\mathbf{S\backslash NP}}$

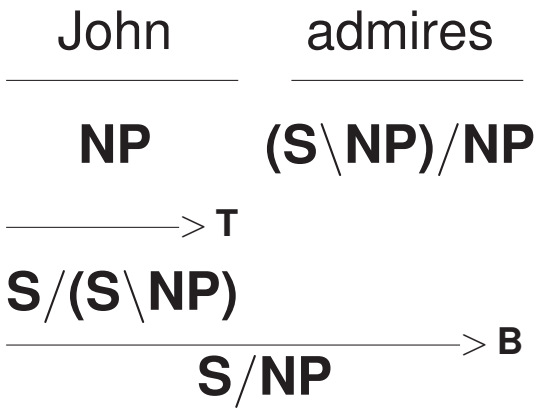
c. *The man that [admires John]$_{\mathbf{S\backslash NP}}$ and [Mary detests]$_{\mathbf{S/NP}}$

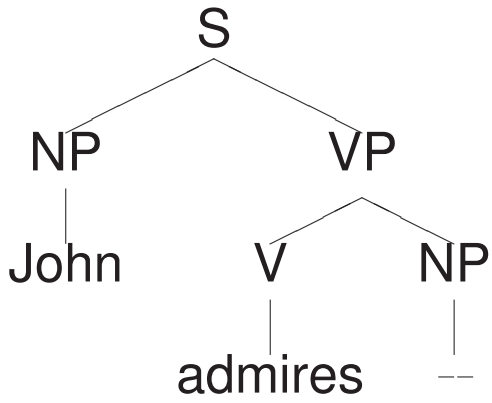
d. The man [that admires John]$_{\mathbf{N\backslash N}}$ and [(that) Mary detests]$_{\mathbf{N\backslash N}}$

e. *The cat that [John admires]$_{\mathbf{S/NP}}$ and [Mary hates it]$_{\mathbf{S}}$

f. *The cat that [John admires it]$_{\mathbf{S}}$ and [Mary hates]$_{\mathbf{S/NP}}$

John admires as a constituent:





CCG's universal syntax

Application

$$\mathbf{X/Y}: f \quad \mathbf{Y}: a \quad \Rightarrow \quad \mathbf{X}: fa \qquad\qquad \mathbf{Y}: a \quad \mathbf{X\backslash Y}: f \quad \Rightarrow \quad \mathbf{X}: fa$$

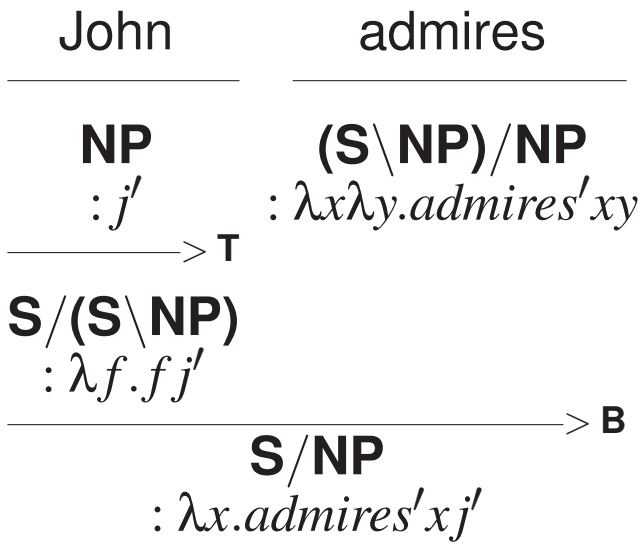
Type Raising

$$\mathbf{A}: a \quad \Rightarrow \quad \mathbf{T/(T\backslash A)}: \lambda f.fa \qquad\qquad \mathbf{A}: a \quad \Rightarrow \quad \mathbf{T\backslash(T/A)}: \lambda f.fa$$

Composition

$$\mathbf{X/Y}: f \quad \mathbf{Y/Z}: g \quad \Rightarrow \quad \mathbf{X/Z}: \lambda x.f(gx) \qquad \mathbf{Y\backslash Z}: g \quad \mathbf{X\backslash Y}: f \quad \Rightarrow \quad \mathbf{X\backslash Z}: \lambda x.f(gx)$$

John likes's compositional semantics is immediate:



Strings and Categories

(11)a. *admires* := $\mathbf{(S\backslash NP_{3s})/NP} : \lambda x \lambda y.admire' \, xy$

b. $\overbrace{\overbrace{admires}^{string} \underbrace{:=}_{\substack{string\\type\\descriptor}} \overbrace{\overbrace{\mathbf{(S\backslash NP_{3s})/NP}}^{syn.type} : \overbrace{\underbrace{\lambda x \lambda y.}_{correspondence} \underbrace{admire' \underbrace{(e,(e,t))}_{sem.type} xy}^{logical\,form}}^{lambda\ term}}^{category}$

Any contiguous string can bear a type. CCG does not commit itself to morphemic morphology or word-based syntax.

Learning which (sub)strings go with which meanings is a learning problem.

(Zettlemoyer and Collins, 2005; Steedman, 2005; Steedman and Hockenmaier, 2007; Coltekin and Bozsahin, 2007)

Constituent structure, intonational structure, information structure

H*L- seems to be the rheme contour in Turkish.

Declarative tune is H*L-L%

(Ali) $\overbrace{\text{(AYNUR-U gördü)}}^{\textit{Rheme}}$ dün gece.

L* H- H* L-L%

‘Ali saw Aynur last night.’

Rheme does not always include the verb, or L%.

(12)a. Ali kim-i gör-dü?
Ali who-ACC see-PAST
‘ Whom did Ali see?’

b. $\overbrace{\text{(AYNUR-U}}^{\textit{Rheme}}\ \overbrace{\text{gördü.)}}^{\textit{Theme}}$
H* L-L%
Aynur-ACC saw

Grammaticizing Turkish tunes and information structure

(13)a. The categories of boundary tones: L- := $\mathbf{S\$}_{\iota}\backslash\mathbf{S\$}_{\rho} : \lambda f.\rho' f$

H- := $\mathbf{S\$}_{\iota}\backslash\mathbf{S\$}_{\theta} : \lambda f.\theta' f$

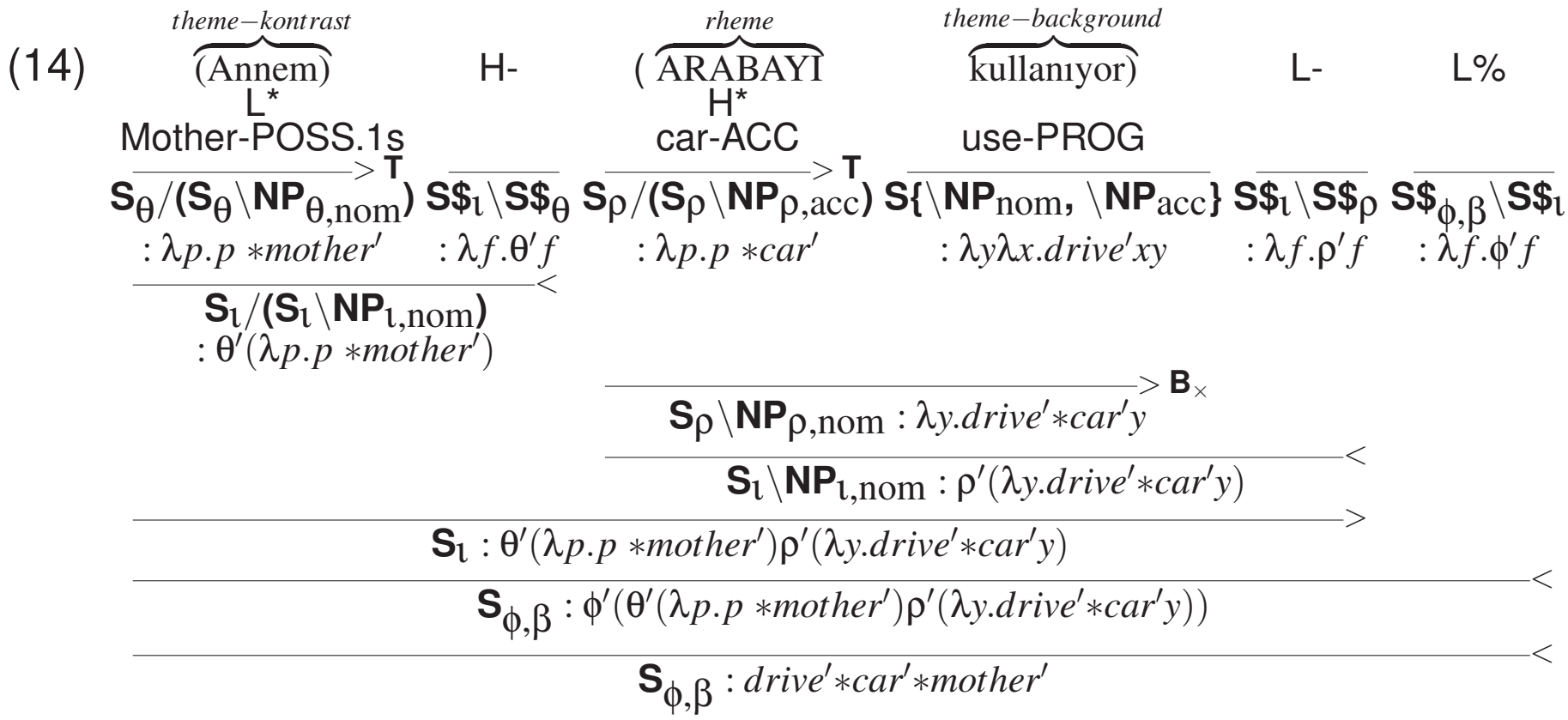
L% := $\mathbf{S\$}_{\phi,\beta}\backslash\mathbf{S\$}_{\iota} : \lambda f.\phi' f$

b. Pitch accents: H* decorates the item in the string with ρ (rheme) feature.

L* decorates the item in the string with θ (theme) feature.

ι: Intermediate phrase. β: background ϕ: Intonational phrase.

(14)



Autosegmental-metrical component aligns pitch accents with syllables (Ladd, 1996; Liberman, 1975; Pierrehumbert, 1980).

Postfocal deaccenting

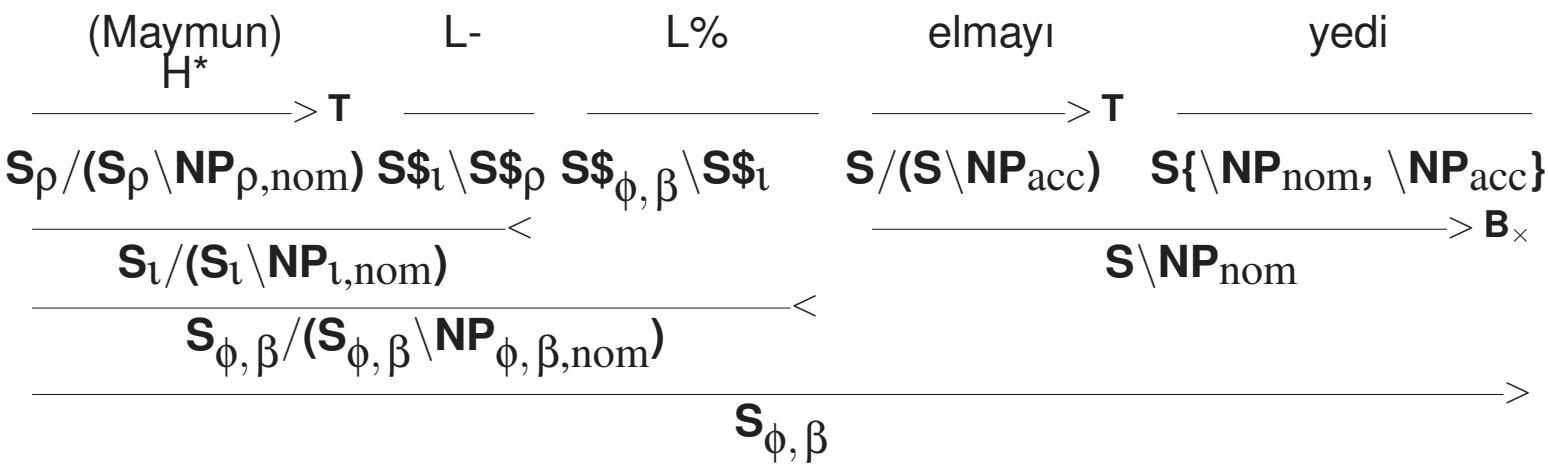
Follows from lexicalized syntactic types.

(15) (Annem kullanıyor) L- L% arabayı.
H*

$$\frac{\dfrac{\dfrac{\dfrac{\dfrac{\mathbf{S}_{\rho}/(\mathbf{S}_{\rho}\backslash\mathbf{NP}_{\rho,nom})\quad \mathbf{S}\{\backslash\mathbf{NP}_{nom},\ \backslash\mathbf{NP}_{acc}\}}{\mathbf{S}_{\rho}\backslash\mathbf{NP}_{\rho,acc}}\,>\mathbf{B}_{\times}\quad \mathbf{S\$}_{\iota}\backslash\mathbf{S\$}_{\rho}}{\mathbf{S}_{\iota}\backslash\mathbf{NP}_{\iota,acc}}<\quad \mathbf{S\$}_{\phi,\beta}\backslash\mathbf{S\$}_{\iota}}{\mathbf{S}_{\phi,\beta}\backslash\mathbf{NP}_{\phi,\beta,acc}}<\quad \mathbf{S}_{\beta}\backslash(\mathbf{S}_{\beta}\backslash\mathbf{NP}_{\beta,acc})\,>\mathbf{T}_{\times}}{\mathbf{S}_{\phi,\beta}}<$$

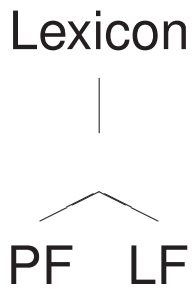
'My mother drives the car.'

It's not always the verb that does this. **Always** the rheme!



'The monkey ate the apple.'

Training the CCG lexical types



No intermediaries between Lexicon and interfaces.

No multiple structures.

Everything projects from the lexicon, led by syntactic types.

Learning algorithms easily apply to annotated data.

Conclusion

Information structure and intonation structure has compositional semantics.

Just like constituent structure.

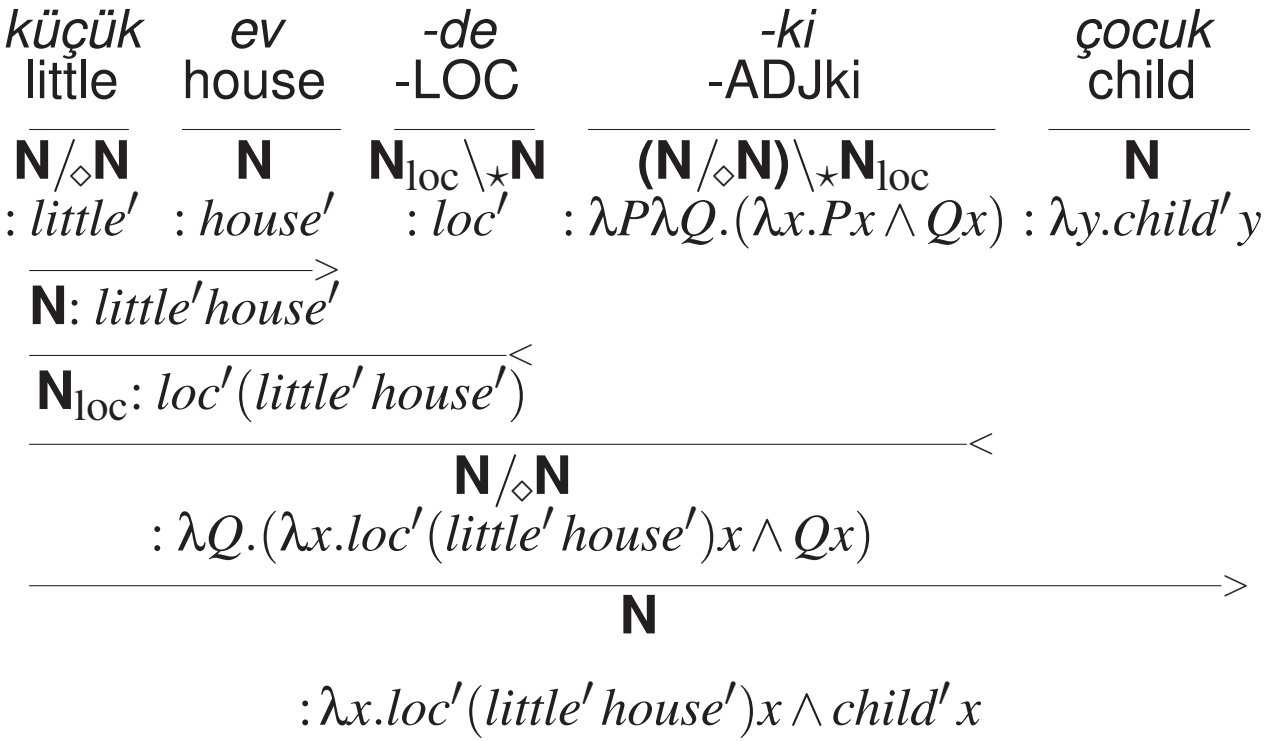
One syntax is good enough for all this

If syntactic type assignments are liberated from words.

Constraint interaction shapes the lexical syntactic types.

Grammar development is setting the constraints right.

The rest is history for the parser; it does not care about the nature of the string, and LF and IS is immediate.



'The child$_i$, the one$_i$ in the little house'

*References

Çakıcı, Ruken. 2008. *Turkish Wide-coverage Parsing with CCG*. Ph.D. thesis, University of Edinburgh. Forthcoming.

Coltekin, Çağrı, and Cem Bozsahin. 2007. "Syllable-based and Morpheme-based Models of Bayesian Word Grammar Learning from CHILDES Database." In *Proc. of the 29th Annual Meeting of Cognitive Science Society*. Nashville, TN.

Demircan, Ömer. 1996. *Türkçe'nin Sesdizimi [Phonology of Turkish]*. İstanbul: Der Yayın.

Ergenç, Iclâl. 1989. *Turkiye Türkçesi'nin Görevsel Sesbilimi [Phonology of Turkey Turkish]*. Ankara: Engin Yayın.

Hockenmaier, Julia, and Mark Steedman. 2007. "CCGBank." *Computational Linguistics*, 33, 3, 356–396.

Johanson, Lars. 1998. "The Structure of Turkic." In Lars Johanson and Éva Á. Csató, eds., *The Turkic Languages*, 30–66. London and New York: Routledge.

Kabak, Barış. 2006. "Turkish Suspended Affixation." Ms., University of Konstanz.

Kornfilt, Jaklin. 1997. *Turkish*. London: Routledge.

Ladd, D. Robert. 1996. *Intonational Phonology*. Cambridge University Press.

Levi, Susannah V. 2005. “Acoustic Correlates of Lexical Accent in Turkish.” *Journal of the International Phonetic Association*, 35, 1, 73–97.

Lewis, Geoffrey. 2000. *Turkish Grammar*. Oxford, New York: Oxford University Press, second edn. First Published in 1967.

Liberman, Mark. 1975. *The Intonational System of English*. Ph.D. thesis, MIT. Published by Garland Press, New York, 1979.

Nash, Rose. 1973. *Turkish Intonation: An Instrumental Study*. The Hague: Mouton.

Özsoy, Sumru. 2004. *Türkçe’nin Yapısı I, Sesbilim [The Structure of Turkish: Phonology]*. Boğaziçi Üniv.

Pierrehumbert, Janet, and Julia Hirschberg. 1990. “The Meaning of Intonational Contours in the Interpretation of Discourse.” In Philip Cohen, Jerry Morgan, and Martha Pollack, eds., *Intentions in Communication*, chap. 14, 271–312. Cambridge, Mass.: MIT Press.

Pierrehumbert, Janet B. 1980. *The Phonology and Phonetics of English Intonation*. Ph.D. thesis, MIT, Cambridge, MA.

Ross, John Robert. 1967. *Constraints on Variables in Syntax*. Ph.D. thesis, MIT. Published as *Infinite Syntax!*, Ablex, Norton, NJ, 1986.

Selen, N. 1973. *Entonasyon Analizleri [Analyses of Intonation]*. Ankara: DTCF Yayın.

Steedman, Mark. 1996. *Surface Structure and Interpretation*. Cambridge, MA: MIT Press.

Steedman, Mark. 2000. *The Syntactic Process*. Cambridge, MA: MIT Press.

Steedman, Mark. 2005. "Grammar Acquisition in Child and Machine." In *Proc. of the 9th Conf. on Computational Natural Language Learning*. Ann Arbor, MI.

Steedman, Mark, and Jason Baldridge. 2007. "Combinatory Categorial Grammar." In Kirsti Börjars, ed., *Non-transformational syntax: a guide to current models*. Blackwell. To appear.

Steedman, Mark, and Julia Hockenmaier. 2007. "The Computational Problem of Natural Language Acquisition." Ms., University of Edinburgh.

Üçok, N. 1951. *Genel Fonetik [General Phonetics]*. Ankara: Ankara Üniv. Yayın.

Underhill, Robert. 1976. *Turkish Grammar*. Cambridge, MA: MIT Press.

Van Der Hulst, Harry, and Jeroen Van De Weijer. 1991. "Topics in Turkish Phonology." In Hendrik Boeschoten and Ludo Verhoeven, eds., *Turkish Linguistics Today*, 11–59. Leiden, The Netherlands: E. J. Brill.

Zettlemoyer, Luke S., and Michael Collins. 2005. “Learning to Map Sentences to Logical Form: Structured Classification with Probabilistic Categorial Grammars.” In *Proc. of the 21st Conf. on Uncertainty in Artificial Intelligence*. Edinburgh.