What's in a Lexicon ?

Cem Bozşahin

Computer Engineering

Middle East Technical University (METU), Ankara

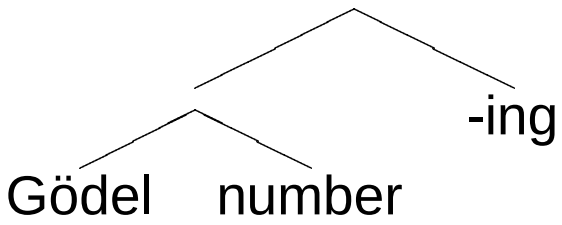
March 17, 2004

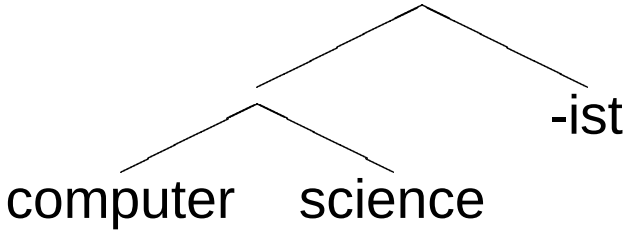
Overview

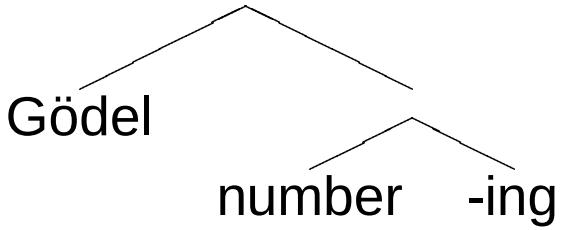
- Examples of bracketing mismatches and phrasal scope of inflections
- Architectures for morphology-syntax-semantics interface
- Morphosyntax: with words or morphemes?
- Morphosyntactic types
- Lexical representation of free/bound morphemes
- Sample derivations of the parser (and performance)

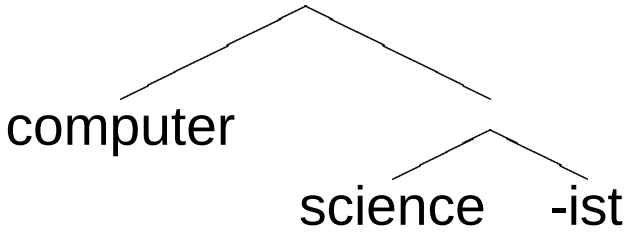
Derivational morphology

- **bracketing mismatches** were first noted in derivational morphology (Williams, 1981)









Verbal inflection

- The problem arises in inflectional morphology as well
- West Greenlandic (Fortescue, 1984)

  Aatsaat tikeraa-nngi-laq
  for.first.time visit-NEG-INDIC/3s
  ‘It is not the first time he has visited.’

- It does not mean ’This is the first time he failed to visit.’

Coordination

- German (Müller, 1999)

  Wenn Ihr Lust und noch nichts anderes vor-habt,
  if you pleasure and yet nothing else intend

  können wir sie ja vom Flughafen abholen
  can we them PARTICLE from.the airport pick up
  'If you feel like it and have nothing else planned, we can pick them up at the airport.'

- semantics:Ihr Lust habt UND noch nichts anderes vorhabt

Subordination

- Turkish

  Mehmet Ayşe'nin [duzenli uyu]-ma-ma-sı-na kız-ıyor
  M.NOM A.-GEN regularly sleep-NEG-INF-AGR-DAT anger-TENSE
  'Mehmet is angry with Ayşe for not sleeping regularly.'
  not 'Mehmet is constantly angry with Ayşe for not sleeping.'

  Mehmet Ayşe'nin kitab-ı oku-ma-sı-nı iste-di
  M.NOM A.-GEN book-ACC read-INF-AGR-ACC want-TENSE
  'Mehmet wanted Ayşe to read the book.'

- semantics: want (read book ayse) mehmet

Relativisation

- Turkish (Bozsahin, 2002)

- Local and non-local morphosyntactic requirements of rel. noun may be different

  Ben Mehmet'in çocuğ-a/*-u ver-diğ-i kitab-ı oku-du-m
  I.NOM M-GEN child-DAT/*ACC give-REL.OP book-ACC read-TENSE-PERS1
  'I read the book that Mehmet gave to the child.'

  Ben Mehmet'in kitab-ı ver-diğ-i çocuğ-u/*-a gör-dü-m
  I.NOM M-GEN book-ACC give-REL.OP child-ACC/*DAT see-TENSE-PERS1
  'I saw the child to whom Mehmet gave the book.'

Lexemic vs. morphemic lexicons

ver-diğ-i :=

$$\left[\begin{array}{l}\text{LOCAL}\left[\begin{array}{ll}\text{CAT} & \left[\begin{array}{l}\text{HEAD}\left[\begin{array}{ll}\text{AGR} & \left[\begin{array}{ll}\text{PERSON} & \textit{third}\\ \text{NUMBER} & \textit{sing}\end{array}\right]\\ \text{CASE} & \textit{dat}\end{array}\right]\\ \text{SUBCAT} \quad \langle \boxed{3}\,\text{NP[gen]}, \boxed{2}\,\text{NP[acc]}, \boxed{1}\,\text{NP[dat]} \rangle\\ \text{MOD} \mid \text{MODSYN} \mid \text{LOCAL} \mid \text{CONT} \mid \text{INDEX} \quad \boxed{1}\end{array}\right]\\ \text{CONTENT} & \left[\begin{array}{ll}\text{RELN} & \textit{give}\\ \text{GIVER} & \boxed{3}\\ \text{GIVEE} & \boxed{1}\\ \text{GIFT} & \boxed{2}\end{array}\right]\end{array}\right]\\ \text{NONLOCAL} \mid \text{TO-BIND} \mid \text{SLASH} \quad \{\boxed{1}\}\end{array}\right]$$

-diğ-i :=

$$\begin{bmatrix} \text{LOCAL} \begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD} & \textit{noun} \left[\text{acc or dat}\right] \\ \text{SUBCAT} & \langle\rangle \end{bmatrix} \\ \text{CONTENT} & \textit{npro} \left[\text{INDEX} \quad \boxed{1}\right] \end{bmatrix} \\ \text{NONLOCAL} \mid \text{INHER} \mid \text{SLASH} \quad \left\{\boxed{1}\right\} \end{bmatrix}$$

Nominal inflection

- Morphological richness of the language does not seem to be the issue
- English (Carpenter, 1997)

  four truck-s

  semantics: four (plu truck)

  alleged thiev-es

  semantics: plu (alleged thief), not alleged (plu thief)

Ki-relativisation

a. araba-da-ki
   car-LOC-REL
   'the one in the car'

b. [[ç ocu**ğ**-un ev-i-nde]-ki-ler-in]-ki
   child-GEN house-POSS-LOC2-REL-PLU-GEN-REL
   lit. 'The one that belongs to the ones that are in the child's house'

c. Ben ev-de-ki-ni hiç kullan-ma-dı -m
   I.NOM house-LOC-REL-ACC2 never use-NEG-TENSE-PERS.1s
   'I never used the one at home.'

d. [k¨uç ¨uk ev-de]-ki hediye
   house-LOC-REL present
   'the present$_i$, the one$_i$ at the little house'
   not 'the little present at the house.'

reduplication

- kırmızı kıpkırmızı yeşil yemyeşil
- yeni yepyeni/yesyeni açık apaçık
- Semantics is uniform, yet there seems to be no morphemic representation.
- It is actually phonology at work on a lexical item (similar processes apply in Tagalog)
- Lexical rules are *unary* rules; they can create new lexical items, and semantics of that can be associated with the lexical rule (specified only once)

- NB: lexical rules refer to *substantive* categories (adjective, adverb etc.), whereas combinatory rules use *formal* categories ($\boldsymbol{X}$, $\boldsymbol{Y}$ etc.)

Resolving the mismatch

- semantics may require affixes to have scope larger than the inflected word
- Alternatives for the morphology-syntax-semantics interface
  - Autonomous levels of morphology, syntax, and semantics (e.g. Sadock, 1991)
  - Morphosyntax-driven semantics (Heylen, 1997; Bozsahin, 2002)
- The lexicon can be morphemic in either case, but it is a combinatory morphemic lexicon in a more lexicalist approach

Inflectional morphology & linguistic theory

- GB (Anderson, 1982) and LFG (Bresnan, 1995) consider inflectional morphology to be part of syntax, (in GB, it is not part of combinatory aspects of grammar)
- MP (Chomsky, 1995) assumes words enter syntax fully inflected (numeration)
- HPSG (Pollard & Sag, 1994) keeps it in the lexicon (lexical rules, or lexical inheritance hierarchy)
- CG work in general (Hoffman, 1995; Heylen, 1997; and others) assumes word-based lexicons, although this is not a theoretical commitment

TLG and inflectional morphology

- Heylen's (1997, 1999) unary modalities. Frau := $\Box_{case}\Box_{fem}\Box_{sg}\Box_{3p}\Box_{decl}\boldsymbol{N}$
- Morphosyntactic type assignment is to inflected forms
- Structural rules regulate scope of inflections, e.g. $\Box_{sg}\Box_{case}\boldsymbol{N}$ can be turned into $\Box_{case}\Box_{sg}\boldsymbol{N}$ by a structural rule
- some iterative morphological processes challenge the lexical rules for word-based type assignment (e.g. -ki in Turkish)
- A more lexical solution is to have morphemic lexicons and morphosyntactic calculus (i.e. -ki as lexical item)

Lexical Syntactic types

- syntactic categories and features

  N, NP, S

  feature-decorations, NP_{acc},S_{fin}

- But features as such are not part of combinatorics,

  unlike e.g. $NP_{case} \rightarrow$ Det N Case

Syntactic calculus

Application (<): $\boldsymbol{Y}:\boldsymbol{a} \quad \boldsymbol{X}\backslash\boldsymbol{Y}:\boldsymbol{f} \quad \Rightarrow \quad \boldsymbol{X}:\boldsymbol{f}\boldsymbol{a}$

Composition (>B): $\boldsymbol{X}/\boldsymbol{Y}:\boldsymbol{f} \quad \boldsymbol{Y}/\boldsymbol{Z}:\boldsymbol{g} \quad \Rightarrow \quad \boldsymbol{X}/\boldsymbol{Z}:\boldsymbol{\lambda x.f}(\boldsymbol{gx})$

Type Raising (>T): $\boldsymbol{X}:\boldsymbol{a} \quad \Rightarrow \quad \boldsymbol{T}/(\boldsymbol{T}\backslash\boldsymbol{X}):\boldsymbol{\lambda f.f}[\boldsymbol{a}]$

Leftward Contraposition (<XP): $\boldsymbol{X}:\boldsymbol{a} \quad \Rightarrow \quad S_{+t}/(\boldsymbol{S}/\boldsymbol{X}):\boldsymbol{\lambda f.f}[\boldsymbol{a}]$
$S_{+t}/(S_{+t}/\boldsymbol{X}):\boldsymbol{\lambda f.f}[\boldsymbol{a}]$

Rightward Contraposition (>XP): $\boldsymbol{X}:\boldsymbol{a} \quad \Rightarrow \quad S_{-t}\backslash(\boldsymbol{S}\backslash\boldsymbol{X}):\boldsymbol{\lambda f.f}[\boldsymbol{a}]$
$S_{-t}\backslash(S_{-t}\backslash\boldsymbol{X}):\boldsymbol{\lambda f.f}[\boldsymbol{a}]$

Lexical Morphosyntactic types

- Two kinds of unary modalities on syntactic types

  $\overset{a}{\triangleleft} X$ (flexible morphosyntactic domain for X : "up to certain inflectional type")

  $\overset{a}{\bowtie} X$ (strict domain : "require certain inflectional type")

- if inflectional paradigm is Stem-Number-Case,

  $\overset{c}{\bowtie} N$ stands for case-marked nouns

  $\overset{c}{\triangleleft} N$ stands for noun stems , number-marked, and case-marked nouns

- Lattice $L = (\mathcal{D}, \leq, =)$
- The set of basic morphosyntactic types: $\mathcal{A}_{ms}$

  $\overset{i}{\triangleleft} X \in \mathcal{A}_{ms}$ and $\overset{i}{\bowtie} X \in \mathcal{A}_{ms}$ if $i \in \mathcal{D}$ and $X \in \mathcal{A}_s$ ($\mathcal{A}_s$: syntactic types)
- The set of complex morphosyntactic types: $\mathcal{B}_{ms}$

  $\mathcal{A}_{ms} \subseteq \mathcal{B}_{ms}$

  If $X \in \mathcal{B}_{ms}$ and $Y \in \mathcal{B}_{ms}$, then $X \backslash Y$ and $X / Y \in \mathcal{B}_{ms}$

Lattice of diacritics (inflectional types)

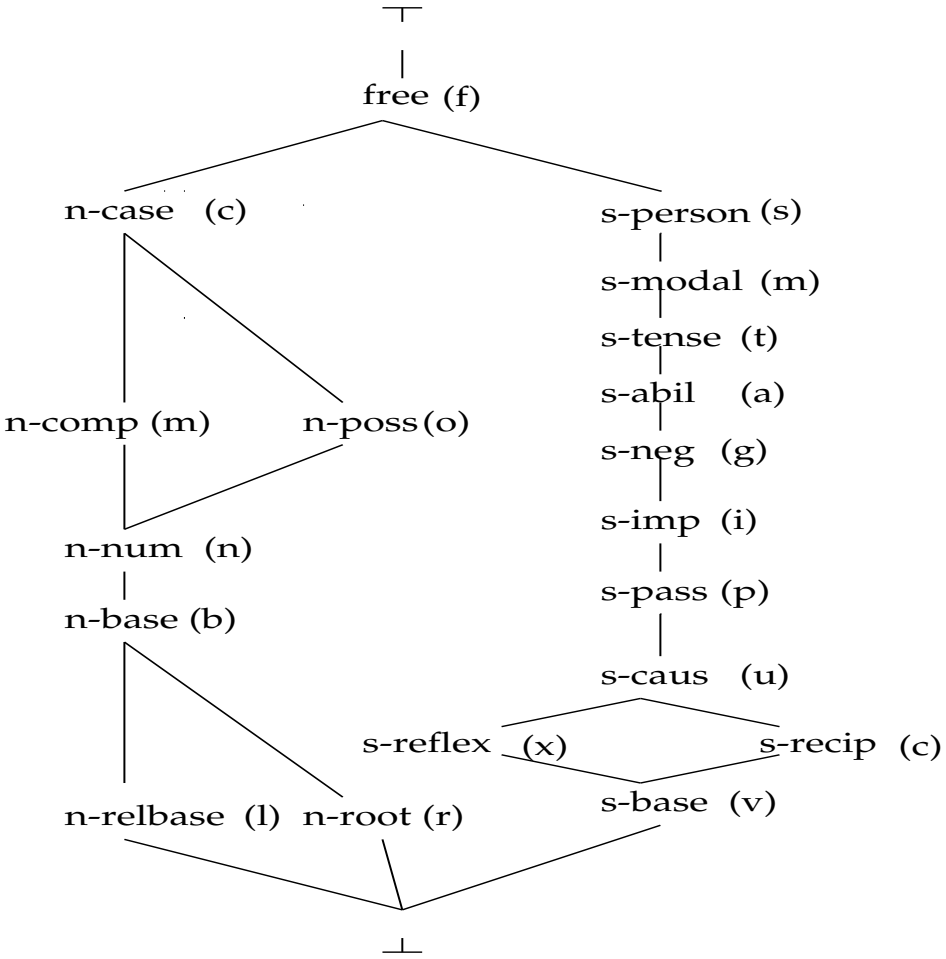
- Inclusion of domains is specified in a language-particular lattice

  This comes in handy for specifying morphotactics as well

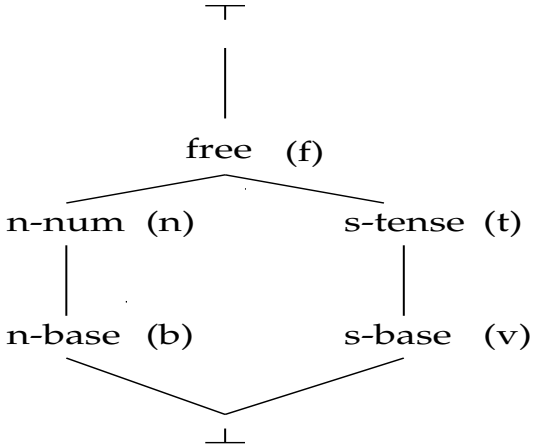
- More importantly, it allows morphosyntactic types to pick semantic domains independent of surface attachment

- All of this is specified in the lexical entry

  attachment type, morphosyntactic type, diacritic, semantic type


free (f)
n-case (c)
s-person (s)
s-modal (m)
s-tense (t)
s-abil (a)
n-comp (m)
n-poss (o)
s-neg (g)
s-imp (i)
n-num (n)
s-pass (p)
n-base (b)
s-caus (u)
s-reflex (x)
s-recip (c)
s-base (v)
n-relbase (l)
n-root (r)

(a)


free (f)
n-num (n)
s-tense (t)
n-base (b)
s-base (v)

(b)

Morphosyntactic lexicon & grammar

- -PLU := $\overset{a}{\circ}$ s – $\overset{n}{\triangleleft}$ *N*\ $\overset{b}{\triangleleft}$ *N*: $\lambda x.\mathrm{plu}\, x$

- Forward Application (>):

$$\frac{\overset{i}{\circ}\, s_1 - X / \overset{\alpha_1}{\square_1}\, Y : f \qquad \overset{j}{\circ}\, s_2 - \overset{\alpha_2}{\square_2}\, Y : a}{\overset{k}{\circ}\, (s_1 \overset{k}{\bullet}\, s_2) - X : fa}\ >$$

if $\alpha_2 \square_1 \alpha_1$ in lattice L, for: $\square_1, \square_2 \in \{\bowtie, \triangleleft\}$, $\alpha_1, \alpha_2 \in \mathcal{D}$ in L, $i, j, k \in \{a, s, c\}$, $\overset{i}{\circ}\ \overset{j}{\circ} \vdash_a \overset{k}{\circ}$

ev -de -ki

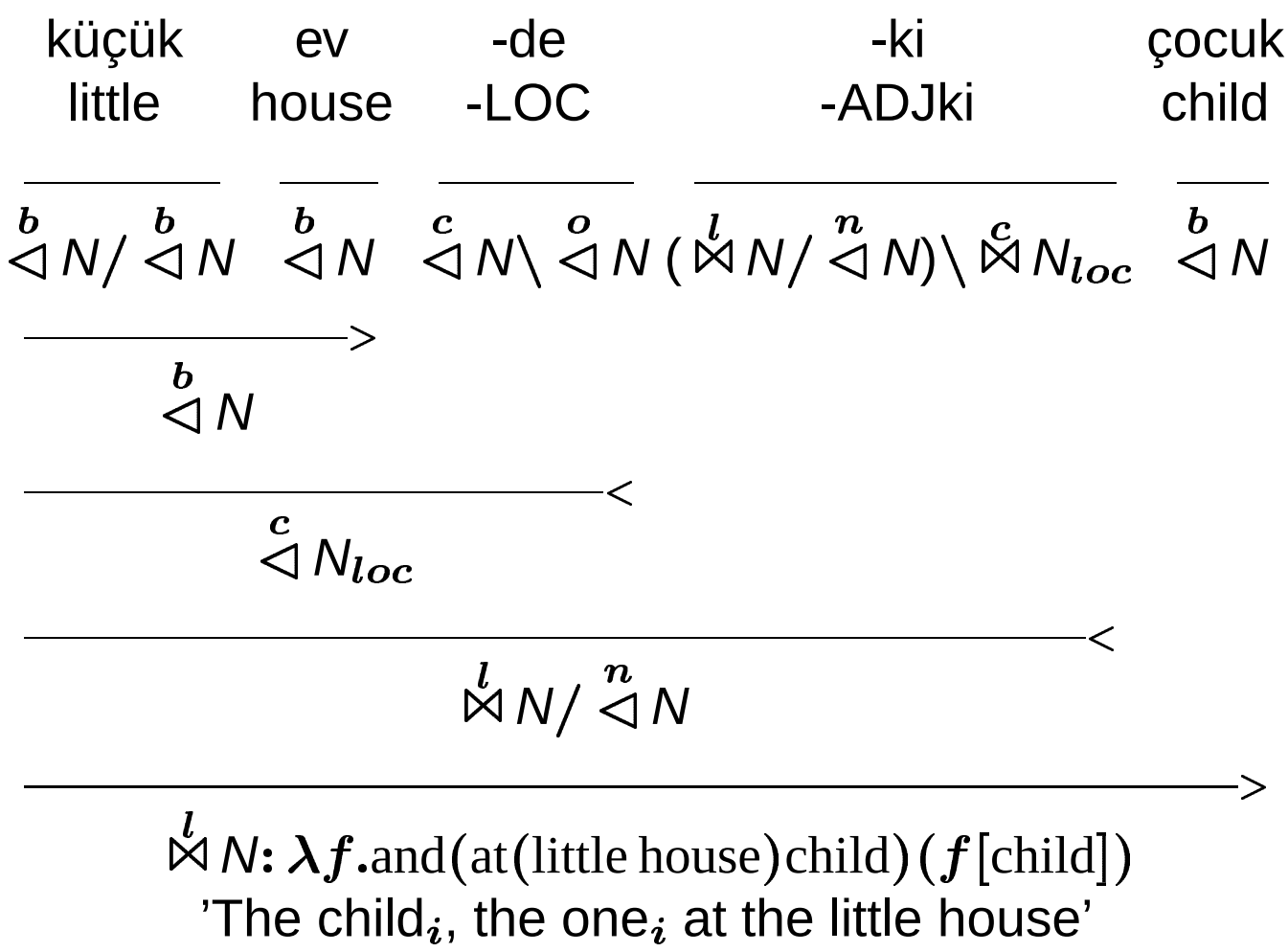
house -LOC -PROki

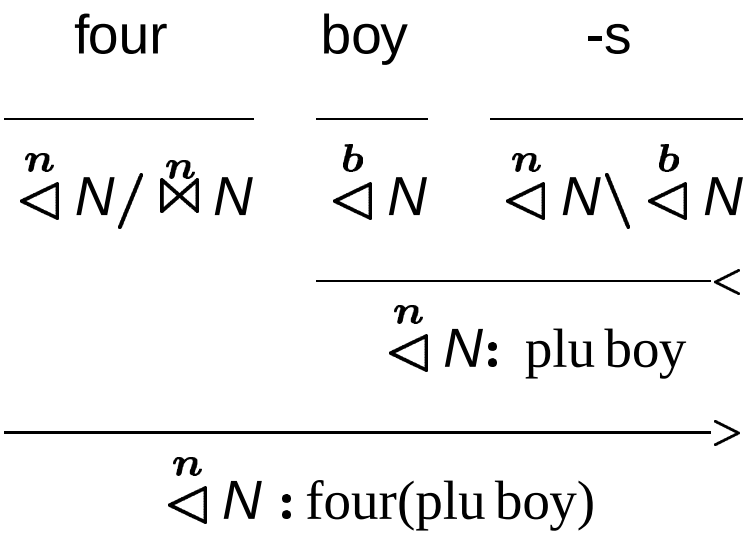
$\overset{b}{\triangleleft} N \quad \overset{c}{\triangleleft} N\backslash \overset{o}{\triangleleft} N \quad \overset{l}{\bowtie} N\backslash \overset{c}{\bowtie} N_{loc}$

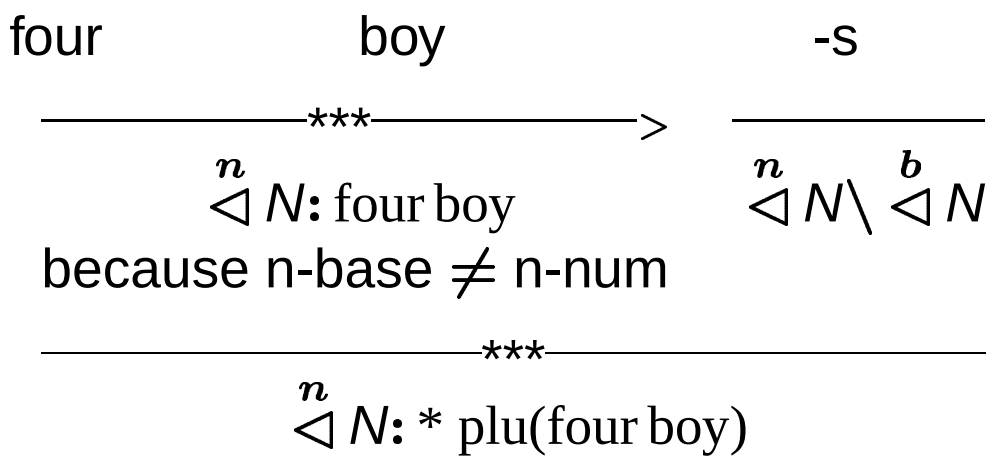
$\overset{c}{\triangleleft} N_{loc}$ <

$\overset{l}{\bowtie} N: \boldsymbol{\lambda f}.\text{and}(\text{at PRO house})(\boldsymbol{f}[\text{PRO}])$ <

'The one that is in the house'


küçük
little
ev
house
-de
-LOC
-ki
-ADJki
çocuk
child
$\overset{b}{\triangleleft} N/\overset{b}{\triangleleft} N$
$\overset{b}{\triangleleft} N$
$\overset{c}{\triangleleft} N\backslash\overset{o}{\triangleleft} N\,(\overset{l}{\bowtie} N/\overset{n}{\triangleleft} N)\backslash\overset{c}{\bowtie} N_{loc}$
$\overset{b}{\triangleleft} N$
>
$\overset{b}{\triangleleft} N$
<
$\overset{c}{\triangleleft} N_{loc}$
<
$\overset{l}{\bowtie} N/\overset{n}{\triangleleft} N$
>
$\overset{l}{\bowtie} N$: $\boldsymbol{\lambda f}.\mathrm{and}(\mathrm{at}(\mathrm{little\ house})\mathrm{child})(\boldsymbol{f}[\mathrm{child}])$
'The child$_i$, the one$_i$ at the little house'


four
boy
-s
$\overset{n}{\triangleleft} N/\overset{n}{\boxtimes} N$
$\overset{b}{\triangleleft} N$
$\overset{n}{\triangleleft} N\backslash \overset{b}{\triangleleft} N$
<
$\overset{n}{\triangleleft} N$: plu boy
>
$\overset{n}{\triangleleft} N$: four(plu boy)


four
boy
-s
***>
$\overset{n}{\triangleleft} N$: four boy
$\overset{n}{\triangleleft} N\backslash \overset{b}{\triangleleft} N$
because n-base ≠ n-num
***
$\overset{n}{\triangleleft} N$: * plu(four boy)

toy gun -s

$\frac{\text{toy} \quad \text{gun}}{\overset{b}{\triangleleft} N/\overset{b}{\triangleleft} N \quad \overset{n}{\triangleleft} N}\,<$: plu gun

$\frac{}{\overset{n}{\triangleleft} N}$: *toy(plu gun)

because n-num $\not\leq$ n-base

toy gun -s

$\frac{\text{toy} \quad \text{gun} \quad \text{-s}}{\overset{b}{\triangleleft} N/\overset{b}{\triangleleft} N \quad \overset{b}{\triangleleft} N \quad \overset{n}{\triangleleft} N\backslash\overset{b}{\triangleleft} N}$

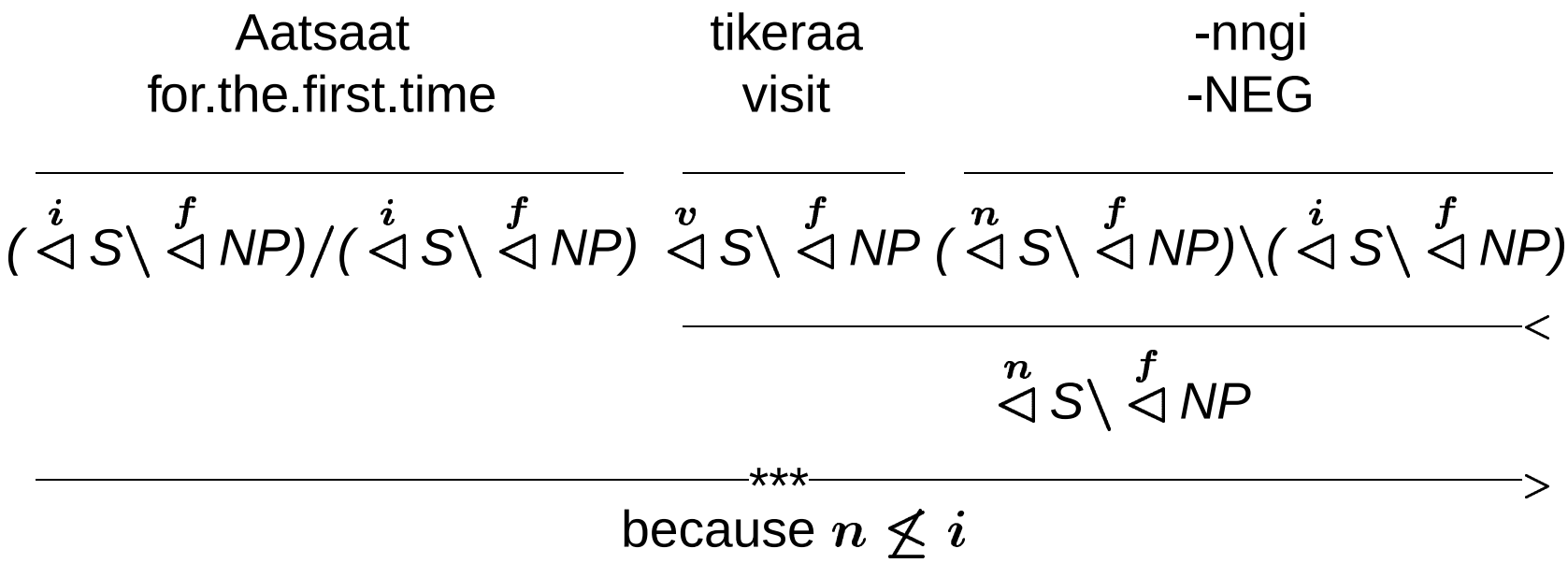
$\frac{}{\overset{b}{\triangleleft} N}>$: toy gun

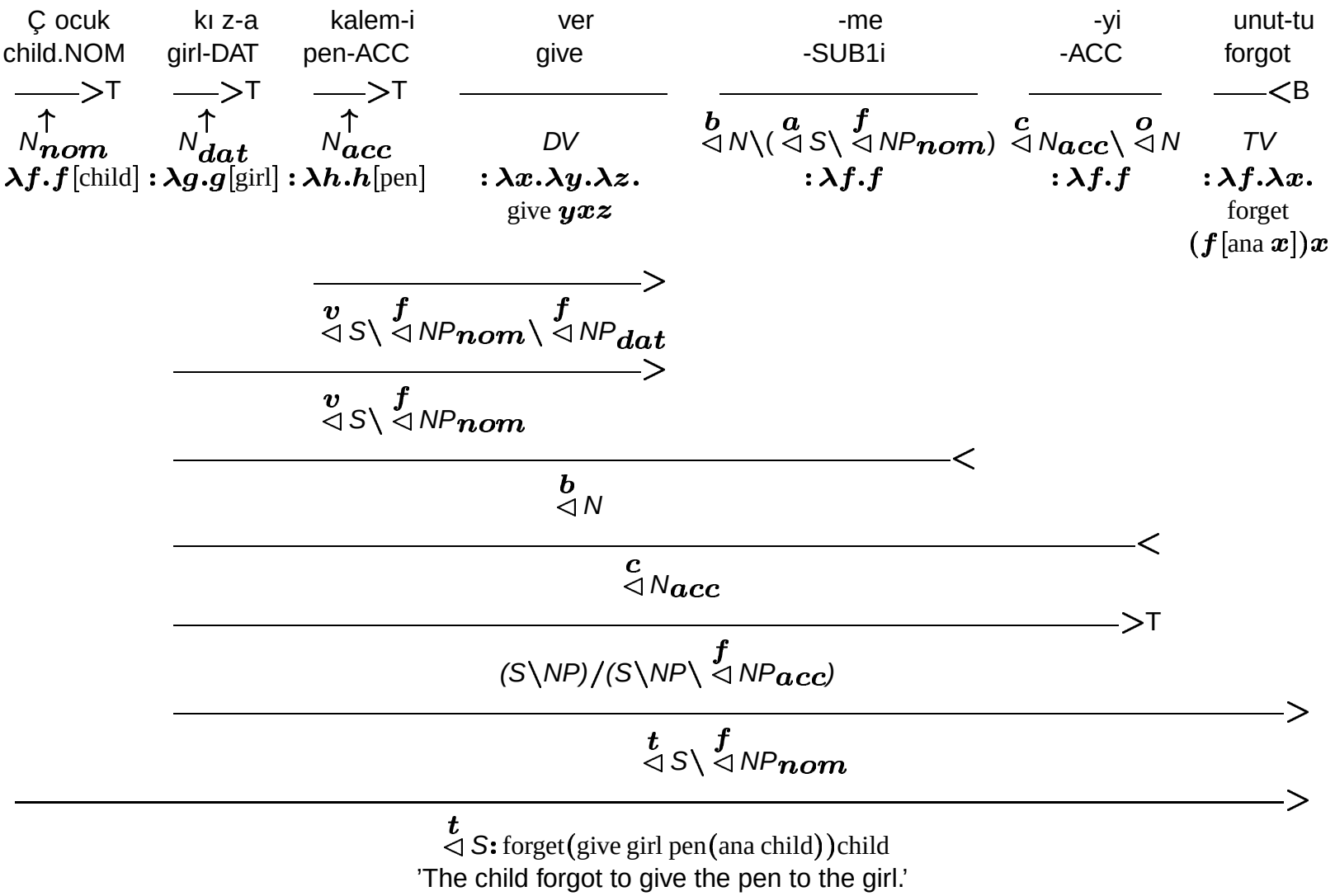
$\frac{}{\overset{n}{\triangleleft} N}<$: plu(toy gun)

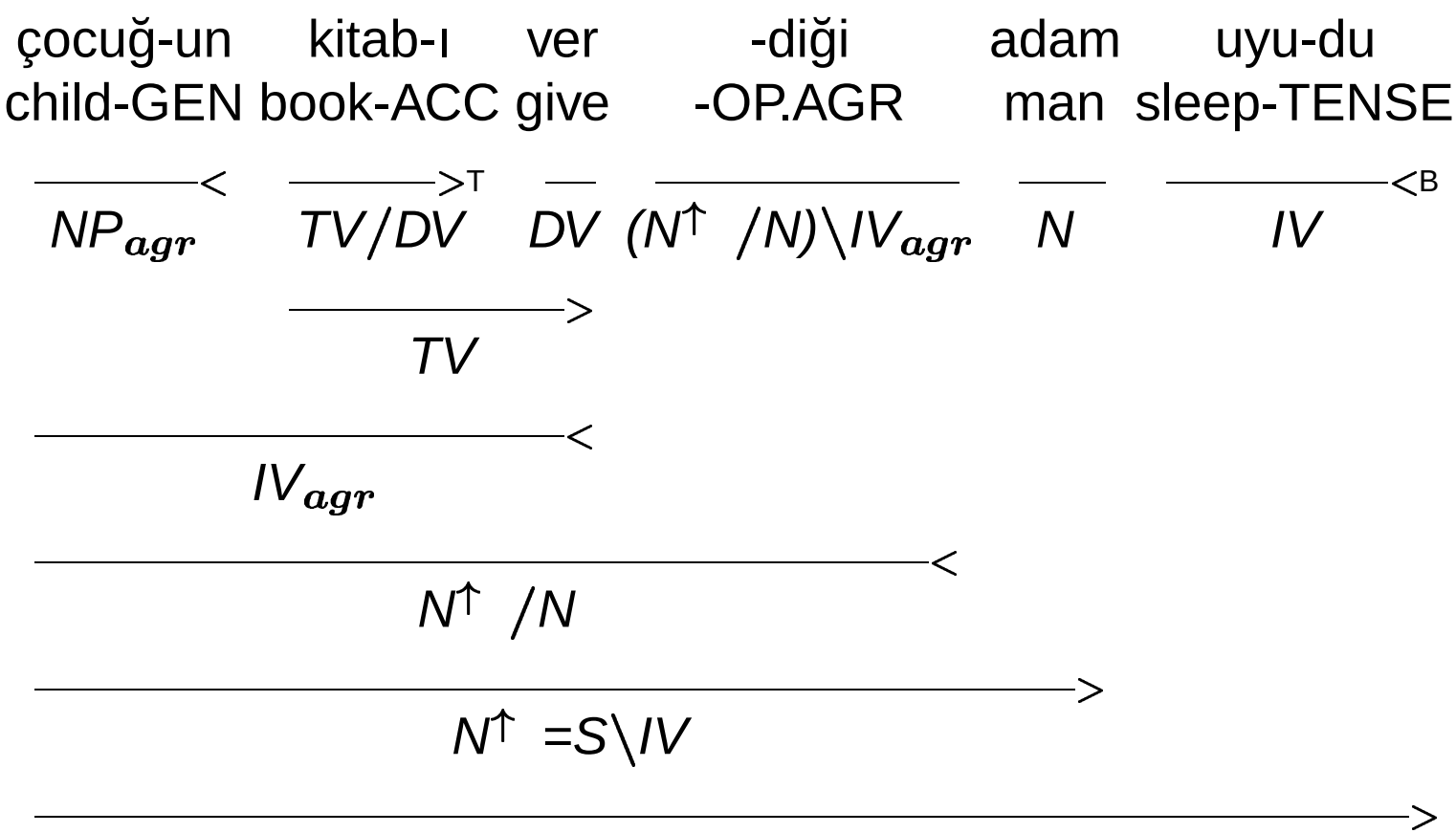
Aatsaat	tikeraa	-nngi	-laq
for.the.first.time	visit	-NEG	-INDIC

$$\frac{(\overset{i}{\triangleleft} S\backslash\overset{f}{\triangleleft} NP)/(\overset{i}{\triangleleft} \backslash\overset{f}{\triangleleft} NP) \quad \overset{v}{\triangleleft} S\backslash\overset{f}{\triangleleft} NP}{\overset{i}{\triangleleft} S\backslash\overset{f}{\triangleleft} NP}\,> \qquad \frac{\overset{i}{\triangleleft} S\backslash\overset{f}{\triangleleft} NP}{\overset{n}{\triangleleft} S\backslash\overset{f}{\triangleleft} NP}\,<$$

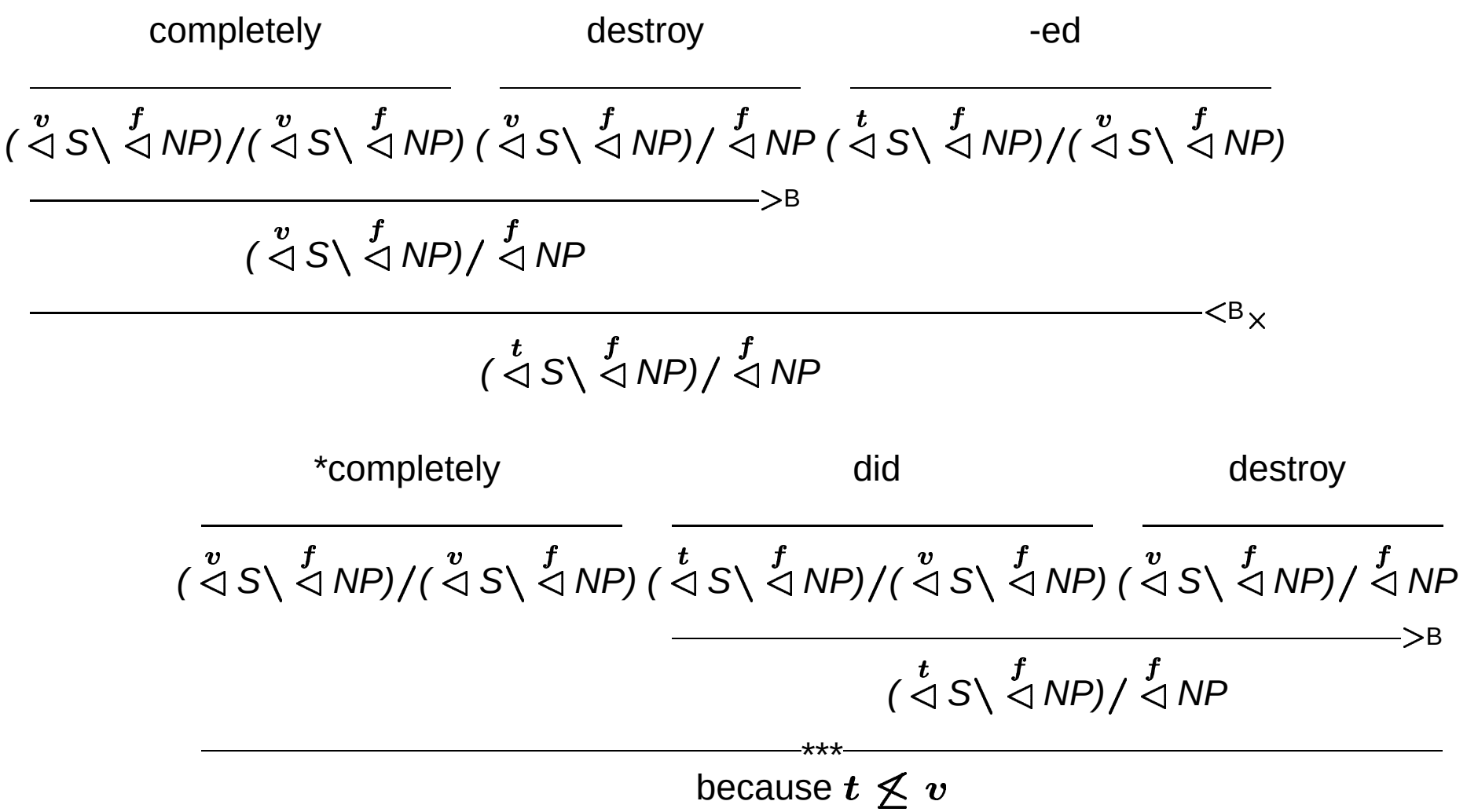
’This is not the first time he visited.’

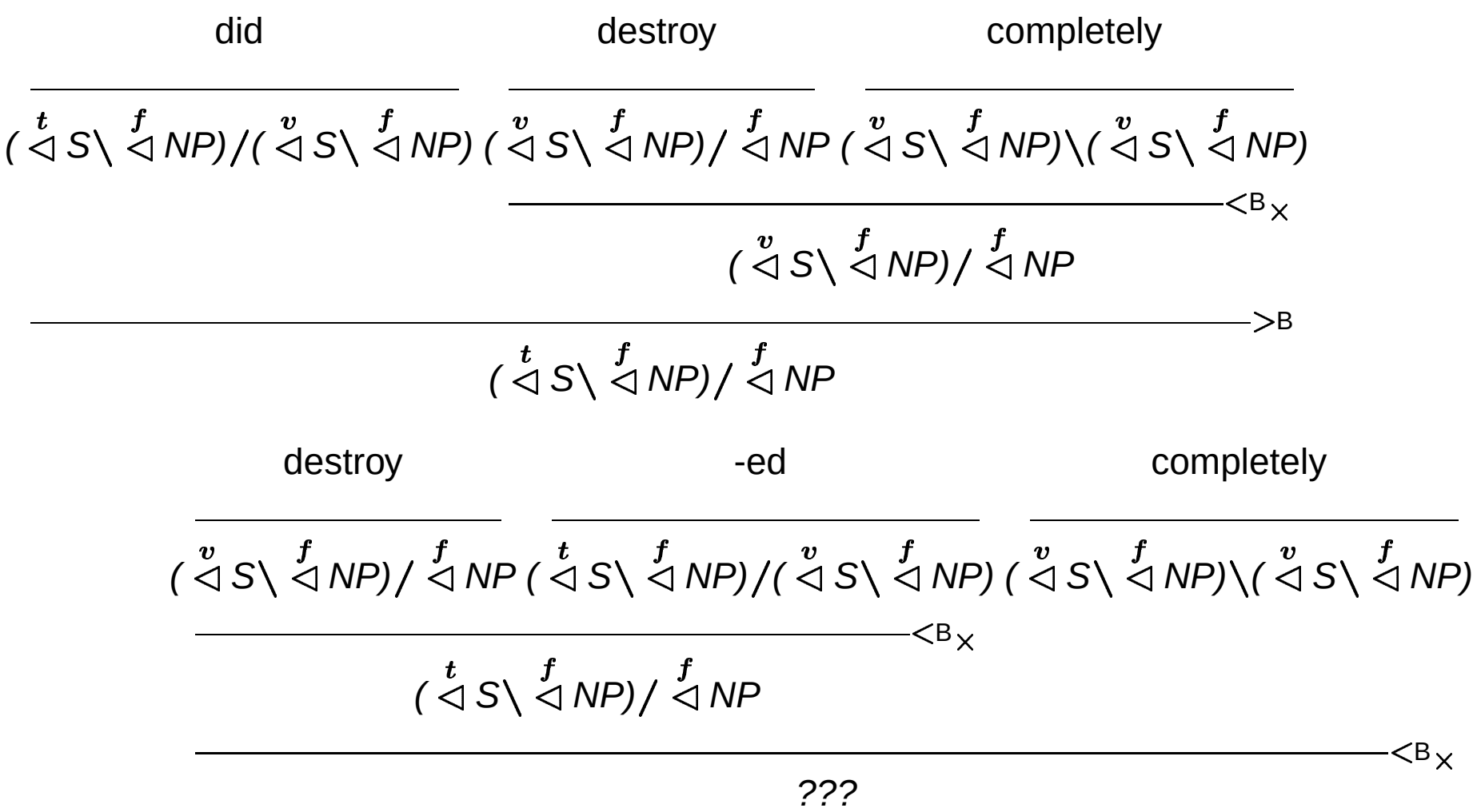

Aatsaat
for.the.first.time
tikeraa
visit
-nngi
-NEG
$(\overset{i}{\triangleleft} S\backslash \overset{f}{\triangleleft} NP)/(\overset{i}{\triangleleft} S\backslash \overset{f}{\triangleleft} NP)$
$\overset{v}{\triangleleft} S\backslash \overset{f}{\triangleleft} NP$
$(\overset{n}{\triangleleft} S\backslash \overset{f}{\triangleleft} NP)\backslash(\overset{i}{\triangleleft} S\backslash \overset{f}{\triangleleft} NP)$
<
$\overset{n}{\triangleleft} S\backslash \overset{f}{\triangleleft} NP$
***
>
because $n \not\leq i$


Ç ocuk
child.NOM
kı z-a
girl-DAT
kalem-i
pen-ACC
ver
give
-me
-SUB1i
-yi
-ACC
unut-tu
forgot
>T
>T
>T
<B
N_nom
N_dat
N_acc
DV
b ◁ N\(a ◁ S\f ◁ NP_nom)
c ◁ N_acc\o ◁ N
TV
: λf.f [child]
: λg.g [girl]
: λh.h [pen]
: λx.λy.λz.
give yxz
: λf.f
: λf.f
: λf.λx.
forget
(f [ana x])x
>
v ◁ S\f ◁ NP_nom\f ◁ NP_dat
>
v ◁ S\f ◁ NP_nom
<
b ◁ N
<
c ◁ N_acc
>T
(S\NP)/(S\NP\f ◁ NP_acc)
>
t ◁ S\f ◁ NP_nom
>
t ◁ S: forget(give girl pen(ana child))child
'The child forgot to give the pen to the girl.'



'The man to whom the child gave the book slept.'


completely
destroy
-ed
$(\overset{v}{\triangleleft} S\backslash\overset{f}{\triangleleft} NP)/(\overset{v}{\triangleleft} S\backslash\overset{f}{\triangleleft} NP)$
$(\overset{v}{\triangleleft} S\backslash\overset{f}{\triangleleft} NP)/\overset{f}{\triangleleft} NP$
$(\overset{t}{\triangleleft} S\backslash\overset{f}{\triangleleft} NP)/(\overset{v}{\triangleleft} S\backslash\overset{f}{\triangleleft} NP)$
>B
$(\overset{v}{\triangleleft} S\backslash\overset{f}{\triangleleft} NP)/\overset{f}{\triangleleft} NP$
<B×
$(\overset{t}{\triangleleft} S\backslash\overset{f}{\triangleleft} NP)/\overset{f}{\triangleleft} NP$
*completely
did
destroy
$(\overset{v}{\triangleleft} S\backslash\overset{f}{\triangleleft} NP)/(\overset{v}{\triangleleft} S\backslash\overset{f}{\triangleleft} NP)$
$(\overset{t}{\triangleleft} S\backslash\overset{f}{\triangleleft} NP)/(\overset{v}{\triangleleft} S\backslash\overset{f}{\triangleleft} NP)$
$(\overset{v}{\triangleleft} S\backslash\overset{f}{\triangleleft} NP)/\overset{f}{\triangleleft} NP$
>B
$(\overset{t}{\triangleleft} S\backslash\overset{f}{\triangleleft} NP)/\overset{f}{\triangleleft} NP$
***
because $t \not\leq v$


did
destroy
completely
(t\f)/(v\f) (v\f)/f (v\f)\(v\f)
<Bx
(v\f)/f
>B
(t\f)/f
destroy
-ed
completely
(v\f)/f (t\f)/(v\f) (v\f)\(v\f)
<Bx
(t\f)/f
<Bx
???

Experiments with the CKY parser

- a 21-morpheme sentence (12 words) parsed in 2.9 seconds

  37-morphemes (20 words) in 40 seconds

- Güngördü & Oflazer's LFG parser takes 10 seconds/sentence with 24,000 word lexicon

- separate morphological analyzers deliver 2 to 5 analyses/second (Oflazer, 1996; Komagata, 1997)

- 2.8 morphemes/word on the average including derivations (Turkish)

  less than 2 inflections/word (Oflazer et. al, 2001)

Sample text type	Number of items in text			Avg. number of parses/gram. input		Avg. CPU time per test (milliseconds)		
	tests	words	morphs	PAS check	NF parse	Unrestr.	PAS check	NF parse
Word order & case	58	216	384	1.26	3.68	39	39	30
Subordination	14	70	137	3.00	5.09	267	270	180
Relativisation	23	130	232	2.04	2.32	796	783	266
Control verbs	33	147	291	1.42	3.34	166	163	137
Possessives & compounds	26	109	200	1.23	2.47	137	135	98
Adjuncts	14	57	100	1.12	4.87	89	88	72
-ki relatives	24	66	179	1.07	1.54	36	36	35

Conclusion

- The key to integration of inflectional morphology and syntax is granting representational status to morphemes
- Morphosyntactic mismatches do not necessitate multi-tiered grammars
- Lexical items can be smaller or larger than words, and project their own semantic domains and attachment characteristics
- Loss of efficiency is tolerable up to medium-length sentences
- Modular grammar-lexicon (in fact, just the lexicon!)