Probabilistic CCG

From Probabilities to Internal Categories

Cem Bozşahin

bozsahin@metu.edu.tr

joint work with Mark Steedman,
with some help from Luke Zettlemoyer

November 30, 2020

Cognitive Science Department, The Informatics Institute, ODTÜ

The Verb

- Thematic structure
- Argument structure
- Word order
- Tense
- Aspect
- Mood

They have a lot to do with how we report ACTIONs and EVENTs.

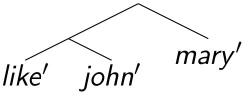
A brief history of Linguistics

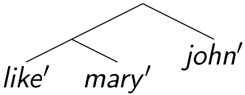
- Thematic structure from argument structure
- Argument structure from thematic structure
- syntax suffices
- semantics suffices
- CCG: it is the correspondence, which seems necessary and sufficient
    - argument from necessity: syntactic identity of semantic differences (e.g. control verbs, ECM, raising)
    - argument from sufficiency: combinators (surface constituency)
- How do we get syntactic forms, logical forms and phonological forms ON THE FLY?
- These "interfaces" must be quite trivial

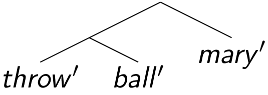
Language of Thought

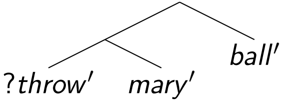
- There seems to be a universal conceptual space (things, actions, events)
- That space seems to be asymmetric
- All human languages are probably homomorphic in that space
- Since languages differ in surface structure, surface structure cannot be homomorphic to that space
- That brings us to WORD ORDER
- How do we capture BOTH aspects if parsing is a reflex? (Garrett)

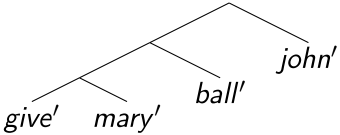
Asymmetry and structure











$(((give'\, mary')\, ball')\, john') =$

$give'\, mary'\, ball'\, john'$

(1) a. The window broke.
$init'(broken'w')$

b. The stone broke the window.
$\lambda z.cause'(init'(broken'w'))st'\,z$

c. The man broke the window with a stone.
$cause'(init'(broken'w'))st'man'$

$init'state'$: an inchoative event culminating in state $state'$

What are these primes and LFs? Hypotheses about LOT.

Surface structure

(2)
a. persuades := $(S\backslash NP_{3s})/VP_{\text{to-inf}}/NP$ — English

b. promises := $(S\backslash NP_{3s})/VP_{\text{to-inf}}/NP$

c. expects := $(S\backslash NP_{3s})/VP_{\text{to-inf}}/NP$

d. broke := $(S\backslash NP)/PP/NP$

e. broke := $(S\backslash NP)/NP$

f. broke := $S\backslash NP$

g. gwelodd (*saw*.3s) := $(S/NP)/NP_{3s}$ — Welsh

h. iniutusan (*order*) := $(S_{\text{DV}}/VP)/NP_{\text{ANG,AGR}}/NP_{\text{NG}}$ — Tagalog

i. savundum (*defended*.1s) := $S\backslash S'_{\text{acc}}$ — Turkish

j. viu (*saw*.3s) := $(S\backslash NP_{3s})/NP$ — Portuguese

Categories: surface and argument structure

(3) a. The window broke.
broke := $S\backslash NP : \lambda x.init'(broken'x)$

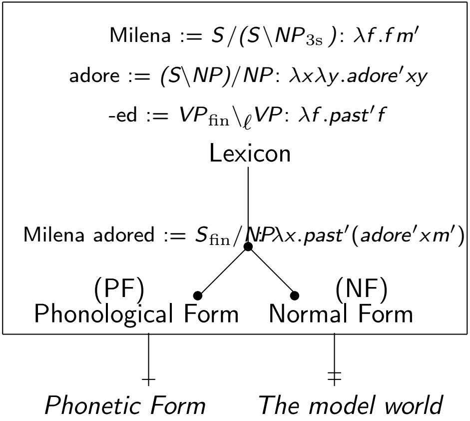
b. The stone broke the window.
broke := $(S\backslash NP)/NP : \lambda x\lambda y\lambda z.cause'(init'(broken'x))y\,z$

c. The man broke the window with a stone.
$(S\backslash NP)/PP/NP : \lambda x\lambda y\lambda z.cause'(init'(broken'x))y\,z$

We get thematic structure as an inference over predicate-argument structure

Question

If surface structure is not homomorphic to universal conceptual structure, how do children learn the word order of their language?



- combinatory projection to constituents of *string* := *syn*:*sem*
- serialization of feature geometry from *string* and *syn*
- normalization from *syn* and *sem*
- realization and intake
- inference and valuation

Conjecture

- Syntax is a HIDDEN variable for the child (NB. necessity of correspondence).
- She has to infer a CATEGORY from form-meaning pairs only (NB. sufficiency of correspondence).
- These pairs are subject to universal combinatorics. (PARSING)
- These inferences set up a MENTAL GRAMMAR.
- Such grammars are necessarily PROBABILISTIC.

- Change the nature of categories, and we might end up with PLANS and SCRIPTS.

Now with full categories

(4) a. persuades :=
$(S\backslash NP_{\mathrm{3s}})/VP_{\mathrm{to\text{-}inf}}/NP$: $\lambda x\lambda p\lambda y.persuade'(px)xy$

b. promises :=
$(S\backslash NP_{\mathrm{3s}})/VP_{\mathrm{to\text{-}inf}}/NP$: $\lambda x\lambda p\lambda y.promise'(py)xy$

c. expects := $(S\backslash NP_{\mathrm{3s}})/VP_{\mathrm{to\text{-}inf}}/NP$: $\lambda x\lambda p\lambda y.expect'(px)y$

d. broke := $(S\backslash NP)/PP/NP$: $\lambda x\lambda y\lambda z.cause'(init'\,broken'\,x)yz$

e. broke := $(S\backslash NP)/NP$: $\lambda x\lambda y\lambda z.cause'(init'\,broken'\,x)yz$

f. broke := $S\backslash NP$: $\lambda x.init'(broken'\,x)$

g. gwelodd (*saw*.3s) := $(S/NP)/NP_{\mathrm{3s}}$: $\lambda x\lambda y.saw'\,yx$

h. iniutusan (*order*) :=
$(S_{\mathrm{DV}}/VP)/NP_{\mathrm{ANG,AGR}}/NP_{\mathrm{NG}}$: $\lambda x\lambda y\lambda p.order'(py)yx$

i. savundum (*defended*.1s) :=
$S_{\mathrm{pd}}\backslash S'_{\mathrm{acc}}$: $\lambda p.defend'\,pi' \wedge topic'\,i'$

j. viu (*saw*.3s) := $(S\backslash NP_{\mathrm{3s}})/NP_{\mathrm{rex}}$: $\lambda x\lambda y.saw'\,xy$

(5) a. $(S\backslash NP)/NP$: $\lambda x\lambda y.verb'xy$ (SVO)

b. $(S/NP)\backslash NP$: $\lambda x\lambda y.verb'yx$ (SVO′)

c. $(S/NP)\backslash NP$: $\lambda x\lambda y.verb'xy$ (OVS)

d. $(S\backslash NP)/NP$: $\lambda x\lambda y.verb'yx$ (OVS′)

e. $(S\backslash NP)\backslash NP$: $\lambda x\lambda y.verb'xy$ (SOV)

f. $(S\backslash NP)\backslash NP$: $\lambda x\lambda y.verb'yx$ (OSV)

g. $(S/NP)/NP$: $\lambda x\lambda y.verb'xy$ (VOS)

h. $(S/NP)/NP$: $\lambda x\lambda y.verb'yx$ (VSO)

Categorial Word order: not a metrical concept

(6)	a.	$(S\backslash NP)/NP: \lambda x \lambda y.verb' xy$	(SVO)	English
	b.	$(S/NP)\backslash NP: \lambda x \lambda y.verb' yx$	(SVO′)	Huastec
	c.	$(S/NP)\backslash NP: \lambda x \lambda y.verb' xy$	(OVS)	Hixkaryana
	d.	$(S\backslash NP)/NP: \lambda x \lambda y.verb' yx$	(OVS′)	Päri
	e.	$(S\backslash NP)\backslash NP: \lambda x \lambda y.verb' xy$	(SOV)	Turkish
	f.	$(S\backslash NP)\backslash NP: \lambda x \lambda y.verb' yx$	(OSV)	Dyirbal
	g.	$(S/NP)/NP: \lambda x \lambda y.verb' xy$	(VOS)	Tagalog
	h.	$(S/NP)/NP: \lambda x \lambda y.verb' yx$	(VSO)	Welsh

A Micro Word-order Thought Experiment

- Syntax is a HIDDEN variable for the child.
- She has to infer a CATEGORY from form-meaning pairs only.
- These pairs are subject to universal combinatorics. (PARSING)
- These inferences set up a MENTAL GRAMMAR.
- Such grammars are necessarily PROBABILISTIC.
- All word orders are available to the child IN THE BEGINNING.

`github.com/bozsahin/ccglab`

Probabilistic CCG

L: Logical Form S: Sentence D: Derivation

- Hidden variable problem: θ are weights of lexical assumptions

$$\arg\max_{L} P(L \mid S; \bar{\theta}) = \arg\max_{L} \sum_{D} P(L, D \mid S; \bar{\theta}) \qquad (1)$$

- Ds are inferred, not observed.
- Relating probabilities and weights (hence categories):

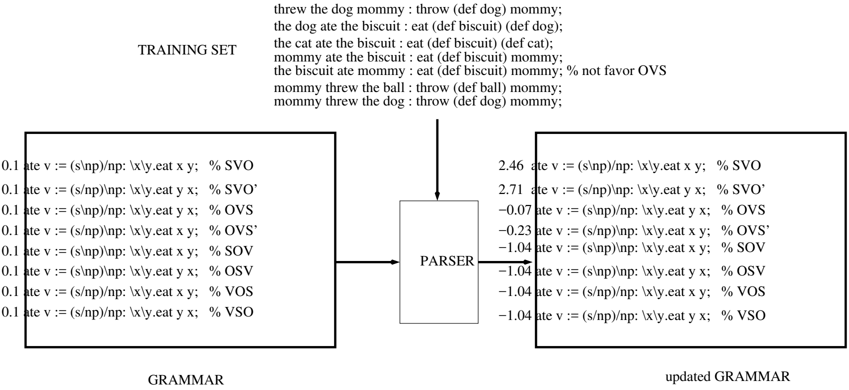
$$P(L, D \mid S; \bar{\theta}) = \frac{e^{\bar{f}(L,D,S)\cdot\bar{\theta}}}{\sum_{L}\sum_{D} e^{\bar{f}(L,D,S)\cdot\bar{\theta}}} \qquad (2)$$

- $f(L, D, S)$: some local measure function (use counts etc.)

Parameter estimation: weight update

- Assume a mental grammar.
- Assume some data to be true (training set)
- What is assumed is form-meaning correspondence
- Parse-to-learn to re-estimate the parameters

Learning workflow



Log-likelihood of training data S_i, L_i

Log-linear model of structure prediction

$$O(\bar{\theta}) = \sum_{i=1}^{n} log\, P(L_i \mid S_i; \bar{\theta}) = \sum_{i=1}^{n} log(\sum_{D} P(L_i, D \mid S_i; \bar{\theta})) \qquad (3)$$

Derivative: How well each parameter serves training

$$\frac{\partial O}{\partial \theta_j} = E_{f_j(L_i,D,S_i)P(D|S_i,L_i;\bar{\theta})} - E_{f_j(L,D,S_i)P(L,D|S_i;\bar{\theta})} \tag{4}$$

Which is

$$\frac{\partial O}{\partial \theta_j} = \sum_{i=1}^{n} \sum_{D} f_j(L_i, D, S_i) P(D \mid S_i, L_i; \bar{\theta}) - \sum_{i=1}^{n} \sum_{L} \sum_{D} f_j(L, D, S_i) P(L, D \mid S_i; \bar{\theta}) \tag{5}$$

Roughly, the expected value of

item j's contribution to correct parses - contribution to all parses

NB. Both are expected values, not counts.

Too many items: we apply Stochastic Gradient to re-estimate

$$
\begin{aligned}
&\text{Initialize } \bar{\theta} \text{ to some value.} \qquad (6)\\
&\text{for } k = 0 \cdots N-1\\
&\quad \text{for } i = 1 \cdots n\\
&\qquad \bar{\theta} = \bar{\theta} + \frac{\alpha_0}{1 + c(i + kn)} \frac{\partial log\, P(L_i|S_i;\bar{\theta})}{\partial \bar{\theta}}
\end{aligned}
$$

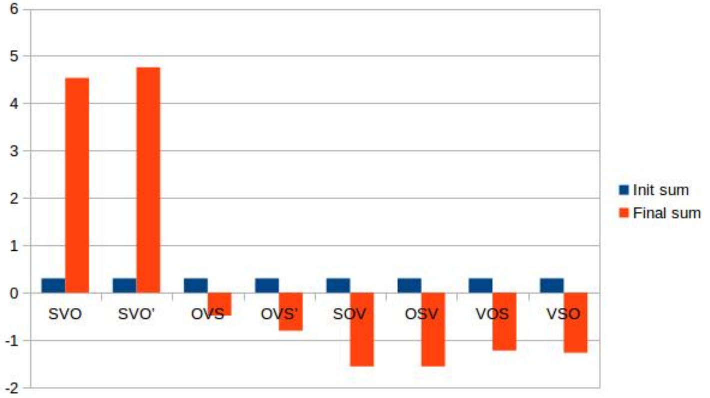
Micro-world experiment

- All possibilities for all verbs are entered in initial mental grammar.
- Give a set of utterances whose meanings are assumed to be known.
- Train the categories on probabilities coming from the parses.
- Some will increase, some decrease, some stay the same.
- How big is the difference from initial grammar?

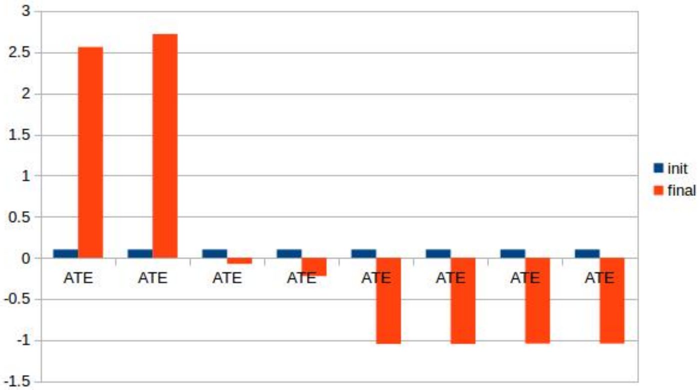
The training set

```
the dog ate the biscuit :  eat (def biscuit) (def dog);
the cat ate the biscuit :  eat (def biscuit) (def cat);
mommy ate the biscuit :  eat (def biscuit) mommy;
the biscuit ate mommy :  eat (def biscuit) mommy;
mommy threw the ball :  throw (def ball) mommy;
mommy threw the dog :  throw (def dog) mommy;
threw the dog mommy :  throw (def dog) mommy;
the dog saw mommy :  see mommy (def dog) ;
mommy saw the dog :  see (def dog) mommy;
the dog walked :  walk (def dog);
the cat slept :  sleep (def cat);
the cat walked :  walk (def cat);
the dog :  def dog;
the cat :  def cat;
mommy :  mommy;
```

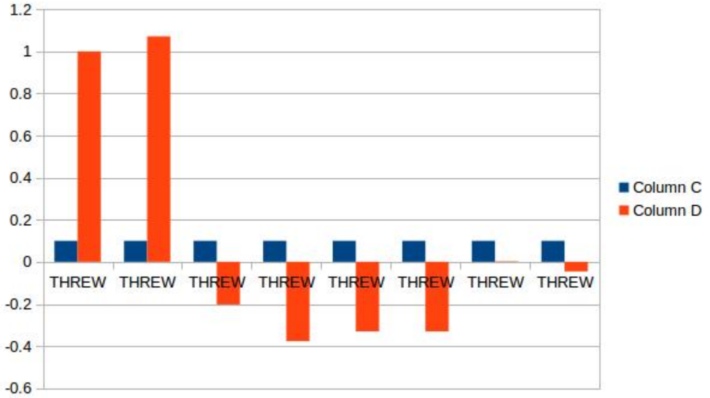
Verb categories update: 3 verbs total $\times$ 8 categories



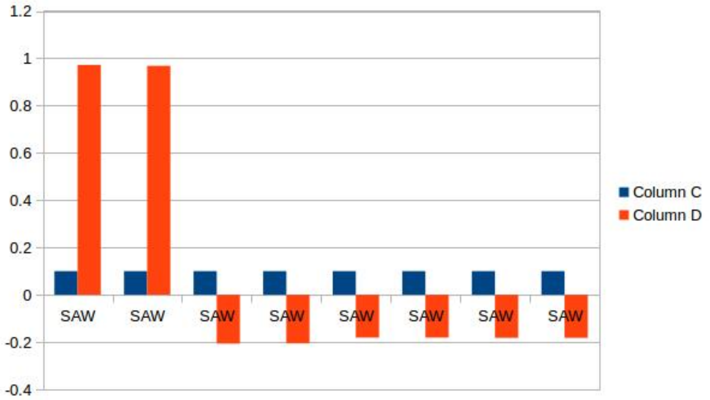
Ate parameters



Threw parameters



Saw parameters



Summary

- These conditions are quite realistic for the child.
- A computationally efficient grammatical theory can attack the hidden-variable problem in syntax
  - to fast convergence without brain switches,
  - to meet the timeframe of language acquisition and the size of grammars children learn.
- Having a correspondence is the prerequisite (necessity)
- Having a combinatory correspondence is sufficient