

# Chart Parsing

## Lecture #15

### Introduction to Natural Language Processing

### CMPSCI 585, Spring 2004

University of Massachusetts Amherst

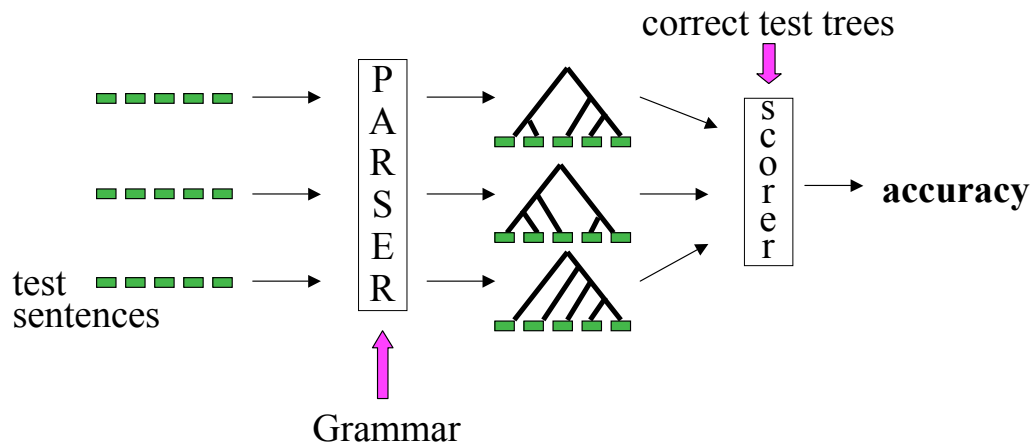


**Andrew McCallum**

*(agglomeration of slides from Jason Eisner)*

Andrew McCallum, UMass Amherst

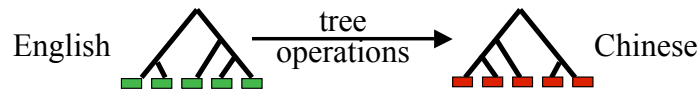
## The parsing problem



Andrew McCallum, UMass Amherst

## Applications of parsing (1/2)

- Machine translation (Alshawi 1996, Wu 1997, ...)



- Speech synthesis from parses (Prevost 1996)

The government plans to raise income tax.

The government plans to raise income tax the imagination.

- Speech recognition using parsing (Chelba et al 1998)

Put the file in the folder.

Put the file and the folder.

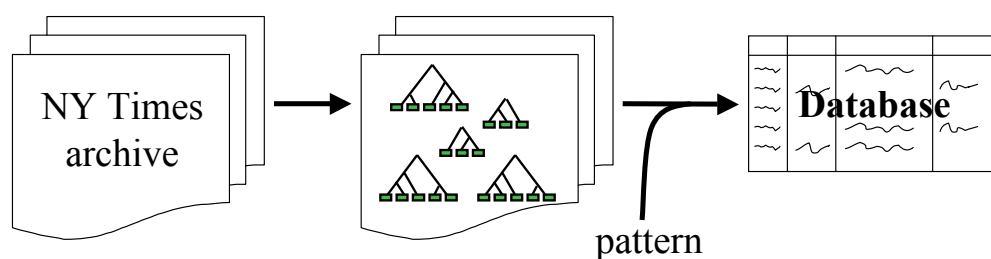
Andrew McCallum, UMass Amherst

## Applications of parsing (2/2)

- Grammar checking (Microsoft)
- Indexing for information retrieval (Woods 1997)

... washing a car with a hose ...      vehicle maintenance

- Information extraction (Hobbs 1996) (Miller et al 2000)



Andrew McCallum, UMass Amherst

## Parsing State of the Art

- Recent parsers quite accurate, e.g.,
  - *A Maximum-Entropy-Inspired Parser*  
Eugene Charniak  
Proceedings of NAACL-2000.
  - *Three Generative, Lexicalised Models for Statistical Parsing*  
Michael Collins  
Proceedings of ACL, 1997.
- Most sentences parsed correctly, or with one error

Andrew McCallum, UMass Amherst

## Last class...

- We defined a CFG,  
where it sits in the Chomsky hierarchy
- Talked about parsing as **search**...  
...through an exponential number of possible trees
- Gave examples of bottom-up and top-down search.
- Discussed problems:
  - Infinite loop with left-recursive rules
  - Much duplicated work in exponential space...  
backtracking

Andrew McCallum, UMass Amherst

## Earley Parser (1970)

- Nice combination of
  - dynamic programming
  - incremental interpretation
  - avoids infinite loops
  - no restrictions on the form of the context-free grammar.
    - $A \rightarrow B C \text{ the } D \text{ of}$**  causes no problems
  - $O(n^3)$  worst case, but faster for many grammars
  - Uses left context and optionally right context to constrain search.

Andrew McCallum, UMass Amherst

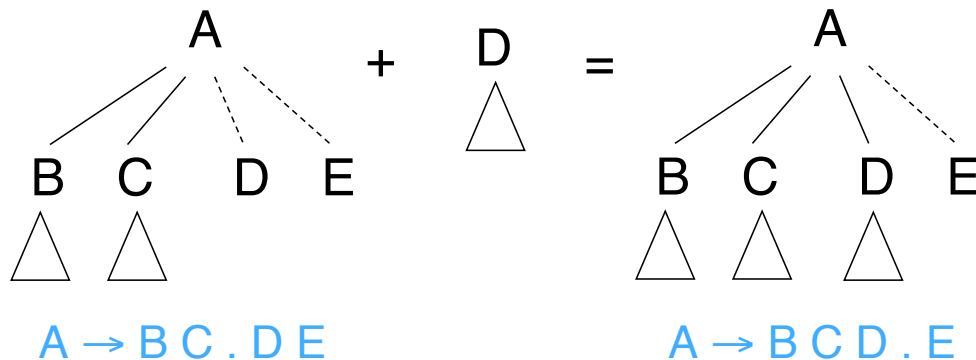
## Earley Parser

- Input: String of words and grammar
- Output: yes/no  
(i.e. recognizer, but can turn into a parser)
- Data Structure:
  - columns 0 through n,  
corresponding to the gaps between words
  - column j is a list of entries like  
 **$(i, A \rightarrow X Y . Z W)$**   
meaning there could be an A starting at i, and we  
have found the X Y part of it from i to j.

Andrew McCallum, UMass Amherst

## Overview of the Algorithm

- Finds constituents and partial constituents in input
  - $A \rightarrow BC.DE$  is partial: only the first half of the  $A$



Andrew McCallum, UMass Amherst

## Overview of the Algorithm

- Proceeds incrementally left-to-right
  - Before it reads word 5, it has already built all hypotheses that are consistent with first 4 words
  - Reads word 5 & attaches it to immediately preceding hypotheses. Might yield new constituents that are then attached to hypotheses immediately preceding *them* ...
  - E.g., attaching  $D$  to  $A \rightarrow BC.DE$  gives  $A \rightarrow BCD.E$
  - Attaching  $E$  to that gives  $A \rightarrow BCDE.$
  - Now we have a complete  $A$  that we can attach to hypotheses immediately preceding the  $A$ , etc.

Andrew McCallum, UMass Amherst

## The Parse Table

- Columns 0 through n corresponding to the gaps between words
- Entries in column 5 look like  $(3, S \rightarrow NP . VP)$ 
  - (but we'll omit the  $\rightarrow$  etc. to save space)
  - Built while processing word 5
  - Means that the input substring from 3 to 5 matches the initial NP portion of a  $S \rightarrow NP VP$  rule
  - Dot shows how much we've matched as of column 5
  - Perfectly fine to have entries like  $(3, S \rightarrow is\ it .\ true\ that\ S)$

Andrew McCallum, UMass Amherst

## The Parse Table

- Entries in column 5 look like  $(3, S \rightarrow NP . VP)$
- What will it mean that we have this entry?
  - *Unknown right context: Doesn't* mean we'll necessarily be able to find a VP starting at column 5 to complete the S.
  - *Known left context: Does* mean that some dotted rule back in column 3 is looking for an S that starts at 3.
    - So if we actually do find a VP starting at column 5, allowing us to complete the S, then we'll be able to attach the S to something.
    - And when that something is complete, it too will have a customer to *its* left ...
    - In short, a top-down (i.e., goal-directed) parser: it chooses to start building a constituent not because of the input but because that's what the left context needs. In **the spoon**, won't build **spoon** as a verb because there's no way to use a verb there.
    - So any hypothesis in column 5 *could* get used in the correct parse, if words 1-5 are continued in just the right way by words 6-n.

Andrew McCallum, UMass Amherst

## Earley's Algorithm, recognizer version

- Add **ROOT**  $\rightarrow$  **. S** to column 0.
- For each  $j$  from 0 to  $n$ :
  - For each dotted rule in column  $j$ , (including those we add as we go!) look at what's after the dot:
    - If it's a word  $w$ , **SCAN**:
      - If  $w$  matches the input word between  $j$  and  $j+1$ , advance the dot and add the resulting rule to column  $j+1$
    - If it's a non-terminal  $X$ , **PREDICT**:
      - Add all rules for  $X$  to the bottom of column  $j$ , with the dot at the start: e.g. **X**  $\rightarrow$  **. Y Z**
    - If there's nothing after the dot, **COMPLETE**:
      - We've finished some constituent,  $A$ , that started in column  $l < j$ . So for each rule in column  $j$  that has  $A$  after the dot: Advance the dot and add the result to the bottom of column  $j$ .
- Output “yes” just if last column has **ROOT**  $\rightarrow$  **S .**
- **NOTE: Don't add an entry to a column if it's already there!**

Andrew McCallum, UMass Amherst

## Summary of the Algorithm

- Process all hypotheses one at a time in order. (Current hypothesis is shown in blue.)
- This may add **new hypotheses** to the end of the to-do list, or try to add **old hypotheses** again.
- Process a hypothesis according to what follows the dot:
  - If a word, **scan** input and see if it matches
  - If a nonterminal, **predict** ways to match it
    - (we'll predict blindly, but could reduce # of predictions by *looking ahead*  $k$  symbols in the input and only making predictions that are compatible with this limited *right context*)
  - If nothing, then we have a complete constituent, so **attach** it to all its customers

Andrew McCallum, UMass Amherst

## A Grammar

S	→	NP VP	NP	→	Papa
NP	→	Det N	N	→	caviar
NP	→	NP PP	N	→	spoon
VP	→	V NP	V	→	ate
VP	→	VP PP	P	→	with
PP	→	P NP	Det	→	the
			Det	→	a

## An Input Sentence

*Papa ate the caviar with a spoon.*

Andrew McCallum, UMass Amherst



initialize

*Remember this stands for (0, ROOT → . S)*





0
0 ROOT . S
0 S . NP VP
0 NP . Det N
0 NP . NP PP
0 NP . Papa
0 Det . the
0 Det . a

predict the kind of Det we are looking for (2 kinds)

0
0 ROOT . S
0 S . NP VP
0 NP . Det N
0 NP . NP PP
0 NP . Papa
0 Det . the
0 Det . a

predict the kind of NP we're looking for  
*but we were already looking for these so  
 don't add duplicate goals! Note that this happened  
 when we were processing a left-recursive rule.*

0	Papa	1
0 ROOT . S	0 NP Papa .	
0 S . NP VP		
0 NP . Det N		
0 NP . NP PP		
0 NP . Papa		
0 Det . the		
0 Det . a		

scan: the desired word is in the input!

0	Papa	1
0 ROOT . S	0 NP Papa .	
0 S . NP VP		
0 NP . Det N		
0 NP . NP PP		
0 NP . Papa		
0 Det . the		
0 Det . a		

scan: failure

0	Papa	1
0 ROOT . S		0 NP Papa .
0 S . NP VP		
0 NP . Det N		
0 NP . NP PP		
0 NP . Papa		
0 Det . the		
0 Det . a		

scan: failure

0	Papa	1
0 ROOT . S		0 NP Papa .
0 S . NP VP		0 S NP . VP
0 NP . Det N		0 NP NP . PP
0 NP . NP PP		
0 NP . Papa		
0 Det . the		
0 Det . a		

attach the newly created NP  
 (which starts at 0) to its customers  
 (incomplete constituents that *end* at 0  
 and have NP after the dot)

0	Papa	1
0 ROOT . S		0 NP Papa .
0 S . NP VP		0 S NP . VP
0 NP . Det N		0 NP NP . PP
0 NP . NP PP		1 VP . V NP
0 NP . Papa		1 VP . VP PP
0 Det . the		
0 Det . a		

predict

0	Papa	1
0 ROOT . S		0 NP Papa .
0 S . NP VP		0 S NP . VP
0 NP . Det N		0 NP NP . PP
0 NP . NP PP		1 VP . V NP
0 NP . Papa		1 VP . VP PP
0 Det . the		1 PP . P NP
0 Det . a		

predict

0	Papa	1
0 ROOT . S		0 NP Papa .
0 S . NP VP		0 S NP . VP
0 NP . Det N		0 NP NP . PP
0 NP . NP PP		1 VP . V NP
0 NP . Papa		1 VP . VP PP
0 Det . the		1 PP . P NP
0 Det . a		1 V . ate

predict

0	Papa	1
0 ROOT . S		0 NP Papa .
0 S . NP VP		0 S NP . VP
0 NP . Det N		0 NP NP . PP
0 NP . NP PP		1 VP . V NP
0 NP . Papa		1 VP . VP PP
0 Det . the		1 PP . P NP
0 Det . a		1 V . ate

predict







0	Papa	1	ate	2
0 ROOT . S	0 NP Papa .	1 V ate .		
0 S . NP VP	0 S NP . VP	1 VP V . NP		
0 NP . Det N	0 NP NP . PP	2 NP . Det N		
0 NP . NP PP	1 VP . V NP	2 NP . NP PP		
0 NP . Papa	1 VP . VP PP	2 NP . Papa		
0 Det . the	1 PP . P NP			
0 Det . a	1 V . ate			
	1 P . with			

predict

0	Papa	1	ate	2
0 ROOT . S	0 NP Papa .	1 V ate .		
0 S . NP VP	0 S NP . VP	1 VP V . NP		
0 NP . Det N	0 NP NP . PP	2 NP . Det N		
0 NP . NP PP	1 VP . V NP	2 NP . NP PP		
0 NP . Papa	1 VP . VP PP	2 NP . Papa		
0 Det . the	1 PP . P NP	2 Det . the		
0 Det . a	1 V . ate	2 Det . a		
	1 P . with			

predict (these next few steps should look familiar)

0	Papa	1	ate	2
0 ROOT . S	0 NP Papa .	1 V ate .		
0 S . NP VP	0 S NP . VP	1 VP V . NP		
0 NP . Det N	0 NP NP . PP	2 NP . Det N		
0 NP . NP PP	1 VP . V NP	2 NP . NP PP		
0 NP . Papa	1 VP . VP PP	2 NP . Papa		
0 Det . the	1 PP . P NP	2 Det . the		
0 Det . a	1 V . ate	2 Det . a		
	1 P . with			

predict

0	Papa	1	ate	2
0 ROOT . S	0 NP Papa .	1 V ate .		
0 S . NP VP	0 S NP . VP	1 VP V . NP		
0 NP . Det N	0 NP NP . PP	2 NP . Det N		
0 NP . NP PP	1 VP . V NP	2 NP . NP PP		
0 NP . Papa	1 VP . VP PP	2 NP . Papa		
0 Det . the	1 PP . P NP	2 Det . the		
0 Det . a	1 V . ate	2 Det . a		
	1 P . with			

scan (this time we fail since Papa is not the next word)







0	Papa	1	ate	2	the	3	caviar	4
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .				
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N	3 N caviar				
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar					
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon					
0 NP . Papa	1 VP . VP PP	2 NP . Papa						
0 Det . the	1 PP . P NP	2 Det . the						
0 Det . a	1 V . ate	2 Det . a						
	1 P . with							

attach

0	Papa	1	ate	2	the	3	caviar	4
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .				
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N	3 N caviar				
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar	1 VP V NP .				
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon	2 NP NP . PP				
0 NP . Papa	1 VP . VP PP	2 NP . Papa						
0 Det . the	1 PP . P NP	2 Det . the						
0 Det . a	1 V . ate	2 Det . a						
	1 P . with							

attach  
(again!)



































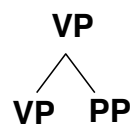


# Left Recursion Kills Pure Top-Down Parsing ...

**VP**

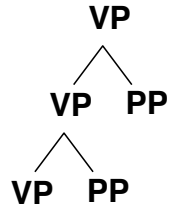
Andrew McCallum, UMass Amherst

# Left Recursion Kills Pure Top-Down Parsing ...



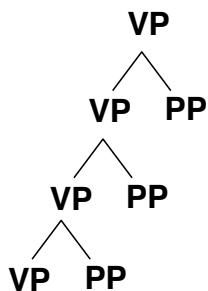
Andrew McCallum, UMass Amherst

## Left Recursion Kills Pure Top-Down Parsing ...



Andrew McCallum, UMass Amherst

## Left Recursion Kills Pure Top-Down Parsing ...



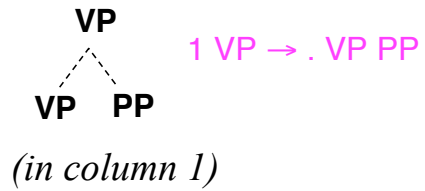
makes new hypotheses  
ad infinitum before we've  
seen the PPs at all

hypotheses try to predict  
in advance how many  
PP's will arrive in input

Andrew McCallum, UMass Amherst

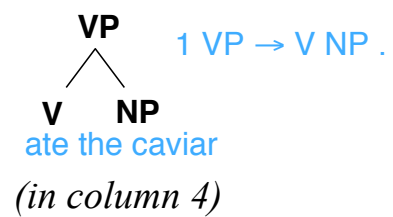
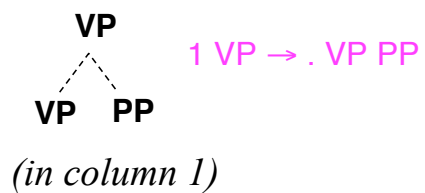


## ... but Earley's Alg is Okay!



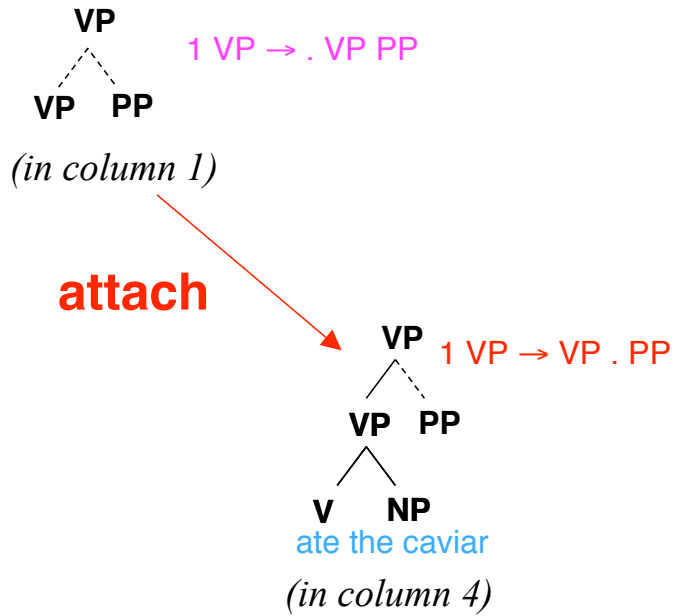
Andrew McCallum, UMass Amherst

## ... but Earley's Alg is Okay!



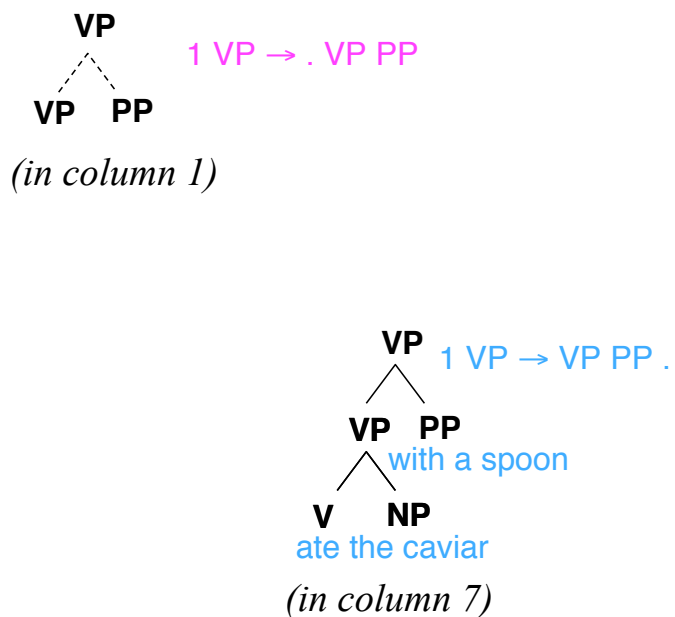
Andrew McCallum, UMass Amherst

## ... but Earley's Alg is Okay!



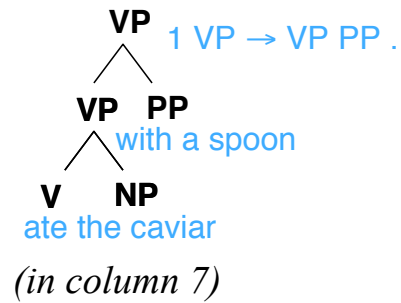
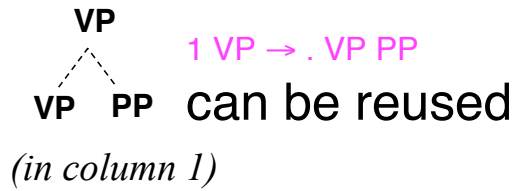
Andrew McCallum, UMass Amherst

## ... but Earley's Alg is Okay!



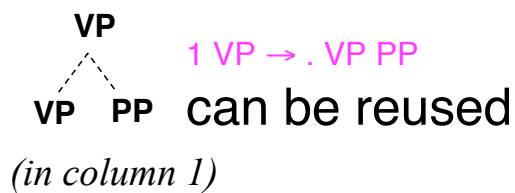
Andrew McCallum, UMass Amherst

## ... but Earley's Alg is Okay!

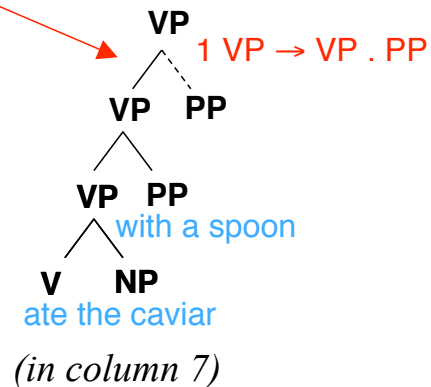


Andrew McCallum, UMass Amherst

## ... but Earley's Alg is Okay!

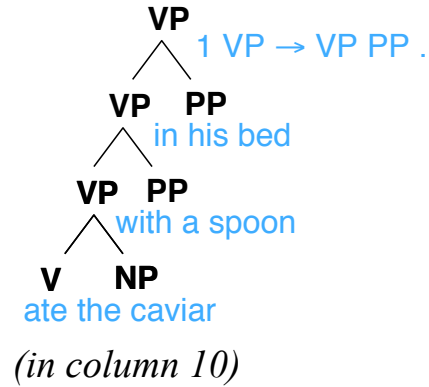
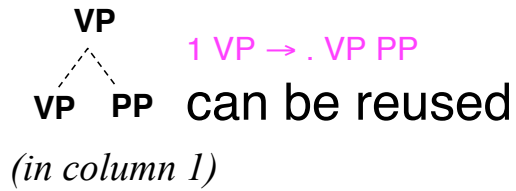


**attach**



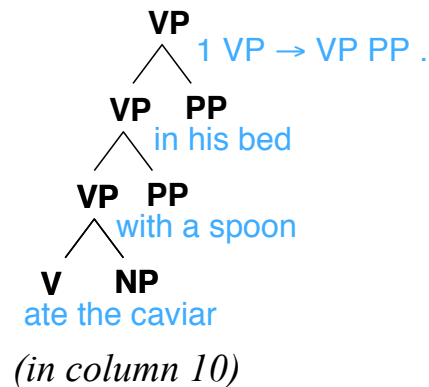
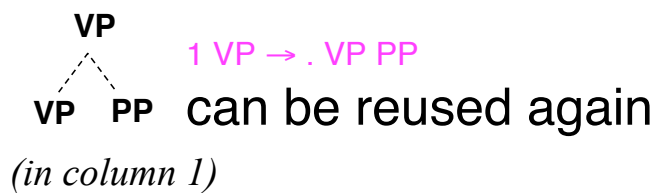
Andrew McCallum, UMass Amherst

## ... but Earley's Alg is Okay!



Andrew McCallum, UMass Amherst

## ... but Earley's Alg is Okay!



Andrew McCallum, UMass Amherst



0	Papa	1	ate	2	the	3	caviar	4	with a spoon	7
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .	...	6 N spoon .				
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N	2 NP Det N .		5 NP Det N .				
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar	1 VP V NP .		4 PP P NP .				
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon	2 NP NP . PP		5 NP NP . PP				
0 NP . Papa	1 VP . VP PP	2 NP . Papa		0 S NP VP .		2 NP NP PP .				
0 Det . the	1 PP . P NP	2 Det . the		1 VP VP . PP		1 VP VP PP .				
0 Det . a	1 V . ate	2 Det . a		4 PP . P NP		7 PP . P NP				
	1 P . with			0 ROOT S .		1 VP V NP .				
				4 P . with		2 NP NP . PP				
						0 S NP VP .				
						1 VP VP . PP				
						7 P . with				
						0 ROOT S .				

completed that VP = VP PP in col 7  
col 1 would let us use *it* in a VP PP structure  
can reuse col 1 as often as we need

## How to change the parser into a recognizer?

# What's the Complexity?

Andrew McCallum, UMass Amherst

# What's the Complexity?

- How many state sets will there be?
  - Length of sentence,  $n$
- How big can the state sets get?
  - Size of grammar,  $G$ , times  $n$
- How long does it take to build a state set?
  - Scan
    - Constant time
  - Predict
    - Need to check for duplicates
  - Complete
    - Search previous state set, also check for duplicates,  $(Gn)^2$
- Total:  $O(n^3)$

Andrew McCallum, UMass Amherst