


Formal Languages

Context-Sensitive Languages

# Context-Sensitive Grammars:

Productions

$$u \rightarrow v$$


String of variables  
and terminals

String of variables  
and terminals

and:  $|u| \leq |v|$

# Context sensitive grammars

*Context-sensitive production:* any production  $\alpha \rightarrow \beta$  satisfying  $|\alpha| \leq |\beta|$ .

*Context-sensitive grammar:* any generative grammar  $G = \langle \Sigma, \Gamma, \Pi, \sigma \rangle$  such that every production in  $\Pi$  context-sensitive.

No empty productions.

# Context-Sensitive Language

Language  $L$  *context-sensitive* if there exists context-sensitive grammar  $G$  such that either  $L = L(G)$  or  $L = L(G) \cup \{\varepsilon\}$ .

# Context-Free and Context-Sensitive Languages

Any context-free language context-sensitive despite fact that there exist context-free grammars that are not context-sensitive grammars.

A bit of engineering.

# Language generated?

$$S \rightarrow aSBC|aBC$$

$$CB \rightarrow BC$$

$$aB \rightarrow ab$$

$$bB \rightarrow bb$$

$$bC \rightarrow bc$$

$$cC \rightarrow cc$$

# Language generated?

$$S \rightarrow AB$$

$$A \rightarrow aAX/aX$$

$$B \rightarrow bBd/bd$$

$$Xb \rightarrow bX$$

$$Xd \rightarrow cd$$

$$Xc \rightarrow cc$$

Generates  $\{a^i b^j c^i d^j \mid i, j \geq 1\}$

$$S \rightarrow AB$$

$$A \rightarrow aAX/aX$$

$$B \rightarrow bBd/bd$$

$$Xb \rightarrow bX$$

$$Xd \rightarrow cd$$



# Language generated?

$$S \rightarrow aS'bX|abX$$

$$S' \rightarrow aS'bC/S'bC/S'C/bC/C$$

$$Cb \rightarrow bC$$

$$CX \rightarrow Xc$$

$$X \rightarrow c$$

Generates  $\{a^i b^j c^k \mid 1 \leq i \leq j \leq k\}$

$$S \rightarrow aS'bX \mid abX$$

$$S' \rightarrow aS'bC \mid S'bC \mid S'C \mid bC \mid C$$

$$Cb \rightarrow bC$$

$$CX \rightarrow Xc$$

$$X \rightarrow c$$

Grammar for  $\{a^n b^j a^n b^k a^n \mid 1 \leq j, k, n\}$

# Grammar for $\{a^n b^j a^n b^k a^n \mid 1 \leq j, k, n\}$

$S \rightarrow X_1 Y Z A_1 T A_2 X_6$

$T \rightarrow A_1 T A_2 A_3$

$T \rightarrow A_1 A_2$

$A_3 A_2 \rightarrow A_2 A_3$

$Z A_1 \rightarrow A_1 Z$

$Z A_2 \rightarrow A_2 Z$

$Y A_1 \rightarrow A_1 Y$

$A_1 Y A_2 \rightarrow A_1 C A_2$

$A_2 Z A_3 \rightarrow A_2 D A_3$

$C \rightarrow C B$

$C \rightarrow B$

$D \rightarrow D B$

$D \rightarrow B$

$X_1 A_1 \rightarrow a X_1$

$X_1 B \rightarrow b X_2$

$X_2 B \rightarrow b X_2$

$X_2 A_2 \rightarrow a X_3$

$X_3 A_2 \rightarrow a X_3$

$X_3 B \rightarrow b X_4$

$X_4 B \rightarrow b X_4$

$X_4 A_3 \rightarrow a X_5$

$X_5 A_3 \rightarrow a X_5$

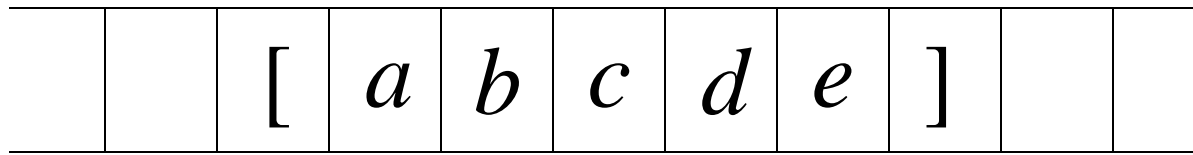
$X_5 X_6 \rightarrow a a$

Linear Bounded Automata (LBAs)  
are the same as Turing Machines  
with one difference:

The input string tape space  
is the only tape space the machine is  
allowed to use

# Linear Bounded Automaton (LBA)

Input string



Working space  
on tape

Left-end  
marker

Right-end  
marker

All computation is done between end markers

Example languages accepted by LBAs:

$$L = \{a^n b^n c^n\}$$

$$L = \{a^{n!}\}$$

LBA's have more power than NPDA's

LBA's have also less power  
than Turing Machines

We define LBA's as NonDeterministic

## Open Problem:

Do NonDeterministic LBA's  
have the same power as  
Deterministic LBA's ?



Theorem:

A language  $L$  is context sensitive  
if and only if

$L$  is accepted by a Linear-Bounded automaton

# The Chomsky Hierarchy

# Unrestricted Grammars:

Productions

$$u \rightarrow v$$

String of variables  
and terminals

String of variables  
and terminals

Example for an unrestricted grammar:

$$S \rightarrow aBc$$

$$aB \rightarrow cA$$

$$Ac \rightarrow d$$

Why not context-sensitive?

$$S \rightarrow aBc$$

$$aB \rightarrow cA$$

$$Ac \rightarrow d$$

## Theorem:

A language  $L$  is recursively enumerable  
if and only if  $L$  is generated by an  
unrestricted grammar

# The Chomsky Hierarchy

All formal languages

Recursively-enumerable

Recursive

Context-sensitive

Context-free

Regular

# Formal Languages and Natural Language Processing



# Relevance of Formal Languages

## Regular languages

One of the key formalisms in NLP  
(morphology, shallow parsing, text processing)

## Context-free languages

Key formalism for parsing in NLP

## Computability

What can be computed? What is an algorithm?

## Efficiency

What can be computed efficiently?

## Linguistics

What kind of formal language is natural language?  
Where is it in the Chomsky hierarchy?

# Relevance of Formal Languages

## Regular languages

One of the key formalisms in NLP  
(morphology, shallow parsing, text processing)

## Context-free languages

Key formalism for parsing in NLP

## Computability

What can be computed? What is an algorithm?

## Efficiency

What can be computed efficiently?

## Linguistics

What kind of formal language is natural language? Where is it in the Chomsky hierarchy?

Most of  
this  
class

# Relevance of Formal Languages

## Regular languages

One of the key formalisms in NLP  
(morphology, shallow parsing, text processing)

## Context-free languages

Key formalism for parsing in NLP

## Computability

What can be computed? What is an algorithm?

## Efficiency

What can be computed efficiently?

## Linguistics

What kind of formal language is natural language?  
Where is it in the Chomsky hierarchy?

Next  
part

# Context-free grammars

(may) have rules like

$NP \rightarrow \text{Det } N$

$PP \rightarrow \text{Prep } NP$

cannot have rules like

$NP \ PP \rightarrow PP \ NP$

*ADV anfangen*  $\rightarrow$  *fangen* *ADV an*

This restriction has implications for the processing resources and speed.

# Issues

Why do computational linguists use formal grammars for describing natural languages?

Are natural languages context-free languages?

# The goal of Natural Language Processing (NLP)

Given a natural language utterance (written or spoken):

Determine: who did what to whom, when, where, how, why (for what reasons, for what purpose)?

Towards this goal: Determine the syntactic structure of an utterance.

# Steps to syntax analysis

For every word in the input string determine its **word class**.

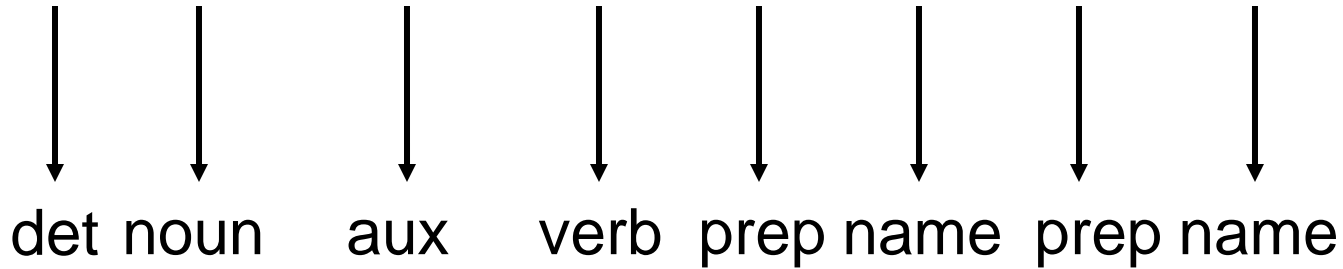
Group all words into **constituents**.

Determine the **linguistic functions** (subject, object, etc.) of the constituents.

Determine the **logical functions** (agent, recipient, transferred-object, place, time ...)

# An example

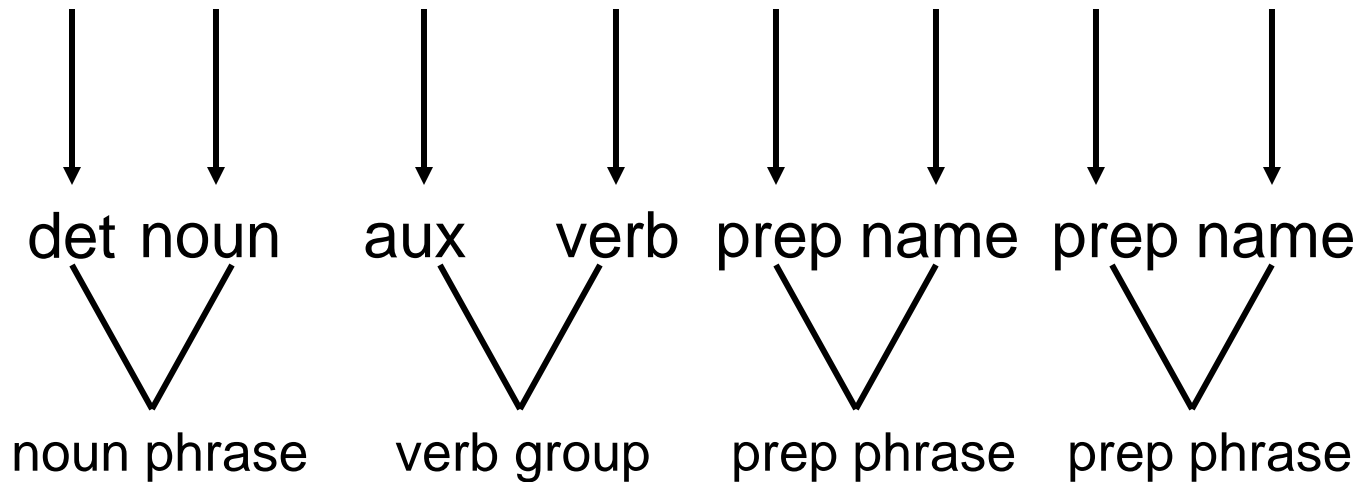
*A book was given to Mary by Peter.*





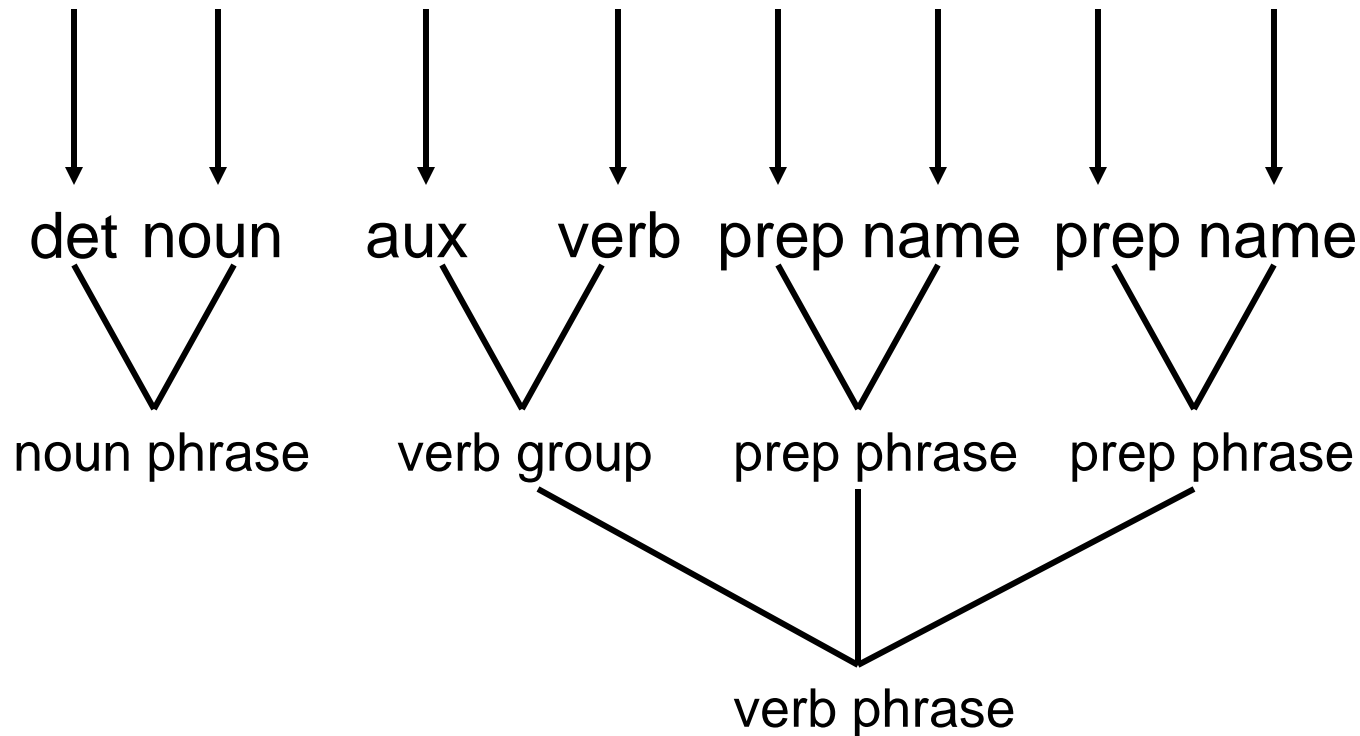
# An example

*A book was given to Mary by Peter.*



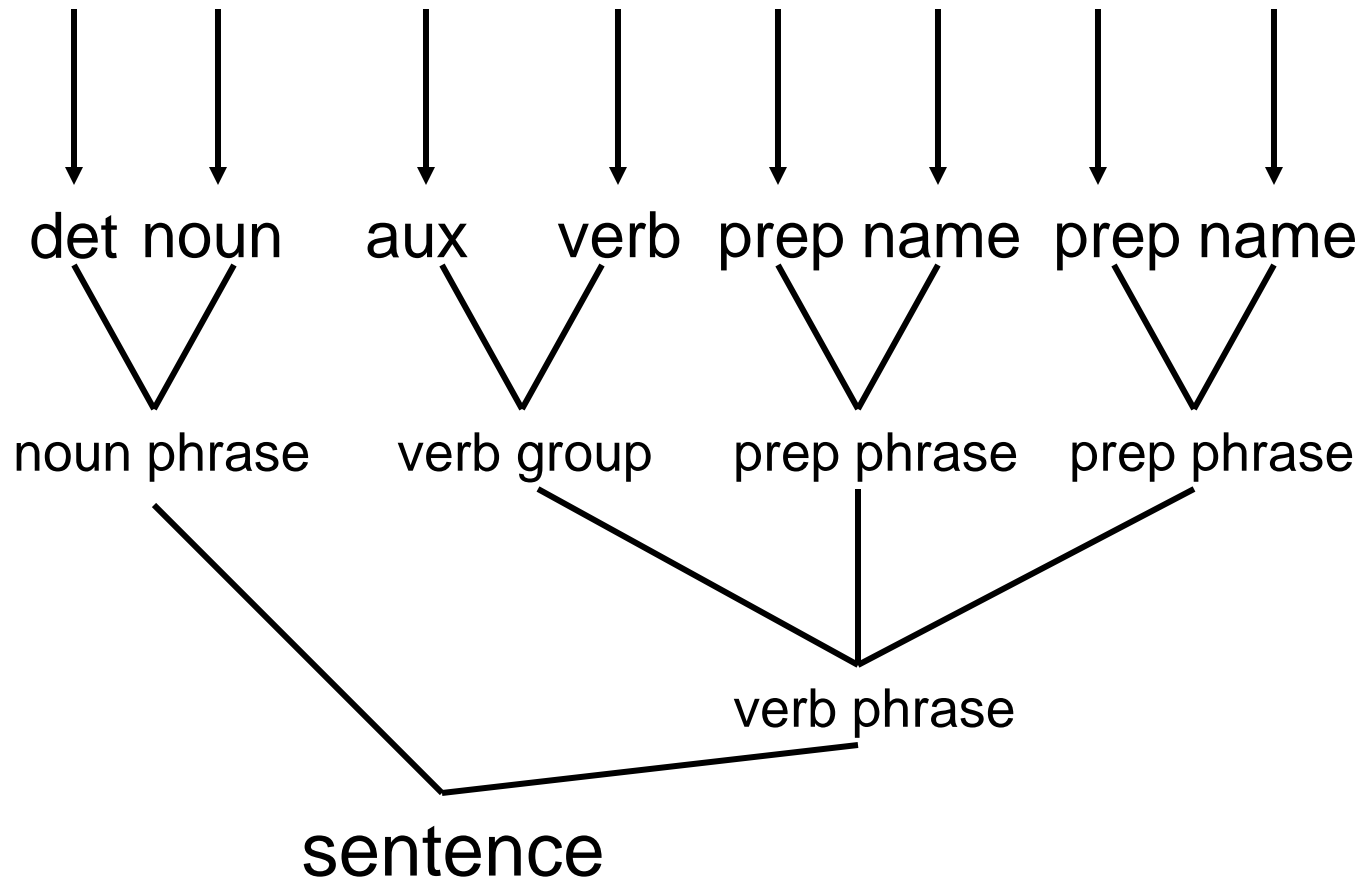
# An example

*A book was given to Mary by Peter.*



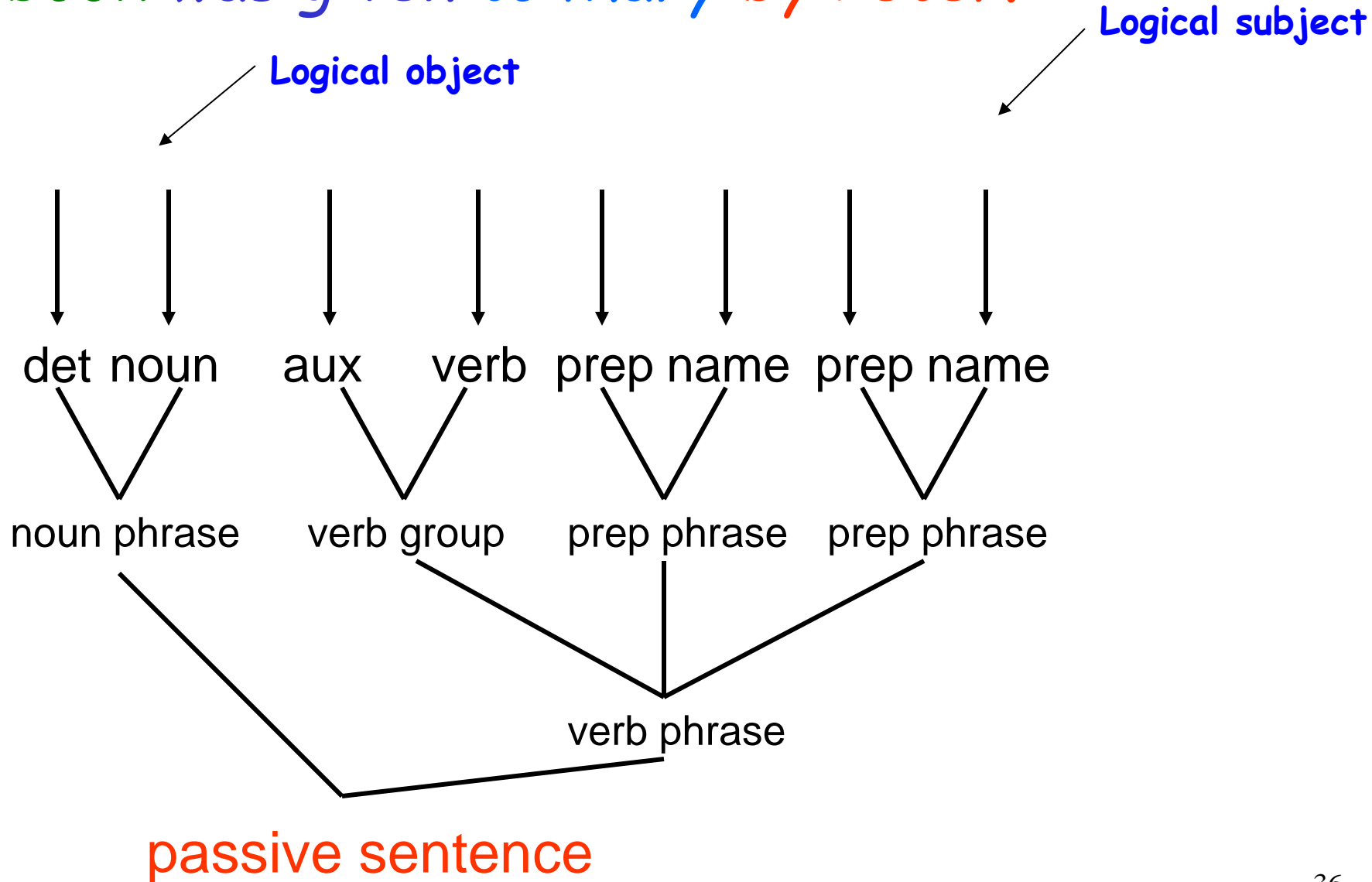
# An example

*A book was given to Mary by Peter.*



# An example

*A book was given to Mary by Peter.*



# Result

Agent (the giver):

Peter

The object:

a book

Recipient:

Mary

Action:

giving

When:

in the past

Via inference

Who has a book now?

Mary

# The context-free rules of a natural language grammar

**Noun\_Phrase  $\rightarrow$  Determiner Noun**

*a book*

*the house*

*some houses*

*50 books*

*Peter's house*

# The context-free rules of a natural language grammar

Adjective\_Phrase  $\rightarrow$  Adjective

Adjective\_Phrase  $\rightarrow$  Adverb Adjective

*nice*

*nicest*

*very nice*

*hardly finished*

# The context-free rules of a natural language grammar

**Noun\_Phrase** → **Det** **Adjective\_Phrase** **Noun**

*a nice book*

*the old house*

*some very old houses*

*50 green books*



# The context-free rules of a natural language grammar

Prep\_Phrase → Preposition Noun\_Phrase

*with a nice book*

*through the old house*

*in some very old houses*

*for 50 green books*

# The context-free rules of a natural language grammar

(may) include recursion (direct and indirect)

Examples

$NP \rightarrow NP \ PP$

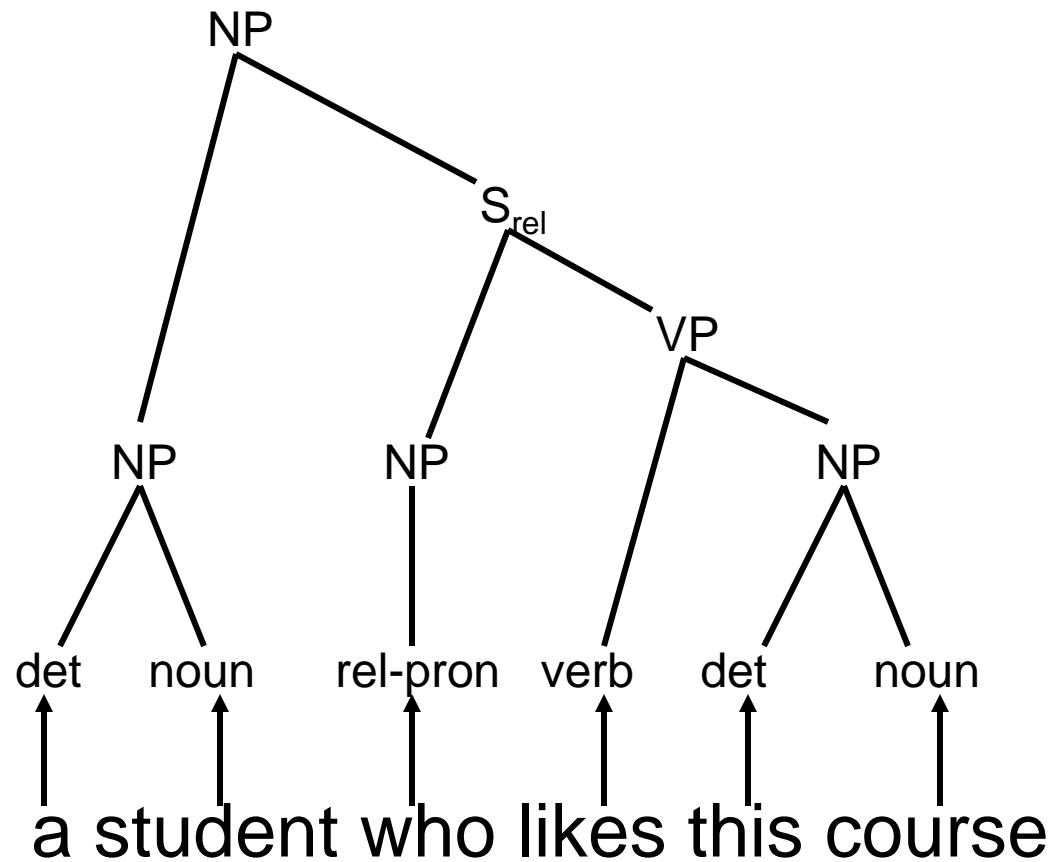
# the bridge over the Nile

$NP \rightarrow NP \ S_{\text{relative}}$

# a student who likes this course

$S_{\text{relative}} \rightarrow NP \ VP$

# who likes this course



# Formal Definition of a Context-free Grammar

A context-free grammar consists of  
a set of non-terminal symbols  $N$

set of terminals  $\Sigma$

a set of productions  $A \rightarrow \alpha$

$A \in N, \alpha\text{-string} \in (\Sigma \cup N)^*$

a designated start symbol (from  $N$ )

# Context-free grammars for natural language

A set of non-terminal symbols  $N$

word class symbols (N, V, Adj, Adv, P)

linguistic constituent symbols (NP, VP, AdjP, AdvP, PP)

A set of terminals  $\Sigma$

all words of the English language

A set of productions  $A \rightarrow \alpha$

the grammar rules (e.g.  $NP \rightarrow Det, AdjP, N$ )

A designated start symbol

a symbol for the complete sentence

NLP: How many (non-)terminals?

# How many ...?

... non-terminals do we need?

word class symbols (N, V, Adj, Adv, P)

usually between 20 and 50

linguistic constituent symbols (NP, VP, ...)

usually between 10 and 20

... terminals do we need?

words of the English language?

Different word stems (see, walk, give)

> 50' 000

Different word forms (see, sees, saw, seen)

> 100' 000

# How many ...?

... grammar rules do we need?

NP → Name # Mary, Peter

NP → Det Noun # a book

PP → Prep NP # to Mary

VP → V NP PP # gave a book to Mary

VP → V NP NP # gave Mary a book

Problem: This grammar will also accept:

\*Peter give Mary a books. # agreement problem

\*Peter sees Mary a book. # complement problem



# Agreement: Why bother?

*\*Peter give Mary a books.*

Consider:

*Peter threw the books into the garbage can that are old and grey.*

*Peter threw the books into the garbage can that is old and grey.*

Agreement can help us determine the intended meaning!

# Agreement: First approach

$NP_{sg} \rightarrow Name_{sg}$   
 $NP_{sg} \rightarrow Det_{sg} Noun_{sg}$   
 $NP_{pl} \rightarrow Det_{pl} Noun_{pl}$   
 $PP \rightarrow Prep NP_{sg}$   
 $PP \rightarrow Prep NP_{pl}$   
 $VP \rightarrow V NP_{sg} NP_{sg}$   
 $VP \rightarrow V NP_{sg} NP_{pl}$   
 $VP \rightarrow V NP_{pl} NP_{sg}$   
 $VP \rightarrow V NP_{pl} NP_{pl}$

# Mary, Peter  
# a book  
# the books  
# to Mary  
# for the books  
# gave Mary a book  
# gave Mary the books  
# gave the kids a book  
# gave the kids the books

Problem?

# Agreement: First approach

$NP_{sg} \rightarrow Name_{sg}$

$NP_{sg} \rightarrow Det_{sg} Noun_{sg}$

$NP_{pl} \rightarrow Det_{pl} Noun_{pl}$

$PP \rightarrow Prep NP_{sg}$

$PP \rightarrow Prep NP_{pl}$

$VP \rightarrow V NP_{sg} NP_{sg}$

$VP \rightarrow V NP_{sg} NP_{pl}$

$VP \rightarrow V NP_{pl} NP_{sg}$

$VP \rightarrow V NP_{pl} NP_{pl}$

# Mary, Peter

# a book

# the books

# to Mary

# for the books

# gave Mary a book

# gave Mary the books

# gave the kids a book

# gave the kids the books

Combinatorial explosion ... too many rules

# Agreement: Better approach

Variables ensure agreement via feature unification.

NP[Num]  $\rightarrow$  Name[Num]

# Mary, Peter

NP[Num]  $\rightarrow$  Det[Num] Noun[Num]

# a book, the books

PP  $\rightarrow$  Prep NP[X]

# to Mary, for the books

VP[Num]  $\rightarrow$  V[Num] NP[X] NP[Y]

# give Mary a book; gives Mary the books

# Subcategorization

Verbs have preferences for the kinds of constituents they co-occur with.

For example:

VP → Verb                      (*disappear*)

VP → Verb NP                  (*prefer a morning flight*)

VP → Verb NP PP    (*leave Boston in the morning*)

VP → Verb PP                (*leaving on Thursday*)

But not: *\*I disappeared the cat.*

# Why is parsing hard?

## McDonald's CEO steps down to battle cancer

By Neil Buckley in New York

Published: November 23 2004 00:51

Last updated: November 23 2004 00:51

McDonald's said on Monday night Charlie Bell would step down as chief executive to devote his time to battling colorectal cancer, dealing another blow to the world's largest fast food company.

Mr Bell's resignation comes just seven months after James Cantalupo, its former chairman and chief executive, died from a heart attack.

McDonald's moved quickly to close the gap, appointing Jim Skinner, currently vice-chairman, to the chief executive's role.

## McDonald's CEO steps down to battle cancer

By Neil Buckley in New York

Published: November 23 2004 00:51

Last updated: November 23 2004 00:51

McDonald's said on Monday night Charlie Bell would step down as chief executive to devote his time to battling colorectal cancer, dealing another blow to the world's largest fast food company.

Mr Bell's resignation comes just seven months after James Cantalupo, its former chairman and chief executive, died from a heart attack.

McDonald's moved quickly to close the gap, appointing Jim Skinner, currently vice-chairman, to the chief executive's role.



# Problems when parsing natural language sentences

Words that are (perhaps) not in the lexicon.

Proper names

*James Cantalupo, McDonald's, InterContinental, GE*

Compounded words → need to be segmented

*kurskamrater, kurslitteratur, kursavsnitt, kursplaneundersökningarna, kursförluster*

*valutakurs, snabbkurs, säljkurser aktiekurser, valutakursindex*

Foreign language expressions

*Don Kerr är Mellanösternspecialist på The International Institute for Strategic Studies i London , högt ansedd , oberoende thinktank .*

Multiword expressions

Idioms: *to deal another blow*

Metaphors

*to battle cancer*

# How can we obtain statistical preferences?

From a parsed and manually checked corpus  
(= collection of sentences)

Such a corpus is usually a database that  
contains the correct syntax tree with each  
sentence (therefore called a **treebank**).

Building a treebank is very time-consuming.

-> Statistical Methods (Sommersemester)

Can all the syntax of natural language be described with context-free rules?

Are there phenomena in natural language that require context-sensitive rules?

# Limits of Context-free Grammars

It is **not** possible to write a **context-free grammar**

(or to design a Push-Down Automaton (PDA))

for the language  $L = \{a^n b^n a^n \mid n > 0\}$

Why?

Intuitively: The memory component of a PDA works like a stack. One stack! So, it can only be used to count once.

Are natural languages context-free?

Received opinion: generally, yes

But ... there is a famous paper about some constructions in Swiss German of the form

$w a^n b^m x c^n d^m y$

*Jan säit, das mer (em **Hans**) (es **huus**) (**hälfed**) (**aastrüiche**).*

*Jan säit, das mer (d' **chind**)<sup>n</sup> (em **Hans**)<sup>m</sup> (es **huus**) (**haend wele laa**)<sup>n</sup> (**hälfe**)<sup>m</sup> (**aastrüiche**).*

but they are rather strange and rare.

The claim that they are not context-free relies on the assumption that  $n$  and  $m$  are unbounded.

# Relevance of Formal Languages

## Regular languages

One of the key formalisms in NLP  
(morphology, shallow parsing, text processing)

## Context-free languages

Key formalism for parsing in NLP

## Computability

What can be computed? What is an algorithm?

## Efficiency

What can be computed efficiently?

## Linguistics

What kind of formal language is natural language?  
Where is it in the Chomsky hierarchy?

Next  
part

**Chomsky hierarchy:  
Where does natural language fall?**



# The Chomsky Hierarchy

All formal languages

Recursively-enumerable

Recursive

Context-sensitive

Context-free

Regular

# The notion of "context"

We need "context" to understand a natural language utterance!

This notion of "context" is different from the notion of "context" in the name context-free languages.

**Is the set of sentences of a natural language finite?**

# Where Does English Fall - The Finiteness Question

Is the set of English sentences finite?

Issues:

- Size of vocabulary
- Length of sentences

*I know that "1" isn't the largest number and I know that "2" isn't the largest number (...)*

If the set of English sentences is finite, then a regular grammar has enough weak generative capacity.

# Chomsky hierarchy: Where does natural language fall?

We need to refine the question:

The *weak generative capacity* of a grammar is the set of strings that the grammar generates.

The *strong generative capacity* of a grammar is the set of *structures* that the grammar generates.

Note that strong generative capacity mirrors linguistic and psychological reality much better than weak generative capacity does.

# Is This English?

**In the event that the Purchaser defaults in the payment of any instalment of purchase price, taxes, insurance, interest, or the annual charge described elsewhere herein, or shall default in the performance of any other obligations set forth in this Contract, the Seller may: at his option: (a) Declare immediately due and payable the entire unpaid balance of purchase price, with accrued interest, taxes, and annual charge, and demand full payment thereof, and enforce conveyance of the land by termination of the contract or according to the terms hereof, in which case the Purchaser shall also be liable to the Seller for reasonable attorney's fees for services rendered by any attorney on behalf of the Seller; or (b) sell said land and premises or any part thereof at public auction, in such manner, at such time and place, upon such terms and conditions, and upon such public notice as the Seller may deem best for the interest of all concerned, consisting of advertisement in a newspaper of general circulation in the county or city in which the security property is located at least once a week for Three (3) successive weeks or for such period as applicable law may require and, in case of default of any purchaser, to re-sell with such postponement of sale or resale and upon such public notice thereof as the Seller may determine, and upon compliance by the Purchaser with the terms of sale, and upon judicial approval as may be required by law, convey said land and premises in fee simple to and at the cost of the Purchaser, who shall not be liable to see to the application of the purchase money; and from the proceeds of the sale: First to pay all proper costs and charges, including but not limited to court costs, advertising expenses, auctioneer's allowance, the expenses, if any required to correct any irregularity in the title, premium for Seller's bond, auditor's fee, attorney's fee, and all other expenses of sale occurred in and about the protection and execution of this contract, and all moneys advanced for taxes, assessments, insurance, and with interest thereon as provided herein, and all taxes due upon said land and premises at time of sale, and to retain as compensation a commission of five percent (5%) on the amount of said sale or sales; SECOND, to pay the whole amount then remaining unpaid of the principal of said contract, and interest thereon to date of payment, whether the same shall be due or not, it being understood and agreed that upon such sale before maturity of the contract the balance thereof shall be immediately due and payable; THIRD, to pay liens of record against the security property according to their priority of lien and to the extent that funds remaining in the hands of the Seller are available; and LAST, to pay the remainder of said proceeds, if any, to the vendor, his heirs, personals representatives, successors or assigns upon the delivery and surrender to the vendee of possession of the land and premises, less costs and excess of obtaining possession.**

## Assume not finite

(Either because it really isn't finite or because we care about strong generative capacity.)

- English isn't regular.
- English can't be characterized with a context-free grammar without sacrificing simplicity and elegance.
- Some natural languages may not be context free at all.

# Complexity of natural language

There have been many arguments about the complexity of natural language, but all of them have the following form:

Find a particular construction type  $C$  (e.g. center embedding) in a natural language  $L$  (e.g. English)

Assume that the construction type in question is theoretically unbounded: i.e., in theory speakers could go on producing ever longer instances of the construction.

The fact that in real life there is a limit to the length of instances of  $C$  that people can process is held to be irrelevant.

This hinges crucially on the *competence-performance* distinction.

Reduce  $C$  via a homomorphism to a formal expression of known complexity.

Argue thereby that  $L$  cannot be of lesser complexity than  $C$ .

Extrapolate from this one instance to natural language in general. I.e.: if there's this one construction in this one language that has this complexity then the human language faculty must allow this in general.

Caveat: in order for this argument to be correct it is not sufficient in general to show that  $L$  contains  $C$  to draw conclusions about  $L$  from  $C$ .



# NL is not regular: Chomsky's original argument

If  $S_1$ , then  $S_2$

Either  $S_3$ , or  $S_4$

The man who said  $S_5$  is arriving today

If either the man who said  $S_5$  is arriving today, or the man who said  $S_5$  is arriving tomorrow, then the man who said  $S_6$  is arriving the day after . . .

Assume the following homomorphism:

if	→	a
then	→	a
either	→	b
or	→	b
everything else	→	$\epsilon$

Then this reduces to the reversal language  $ww^R$ , which is non-regular (but is context free).

Thus Chomsky concluded:

"English is not a finite state language."

N. Chomsky. 1957. *Syntactic Structures*. Mouton, The Hague. Page 21.

Note that the way Chomsky's original argument was cast was to say that because English *contains* these constructions, which are not regular, then English is not regular. **As stated, the argument is fallacious.**

(Also he means "regular language", but we won't quibble about terminology.)

# Problem with Chomsky's argument?

# Why Chomsky's argument is fallacious

Consider the language  $(a|b)^*$ :

- This is (obviously) regular
- Yet it *contains* the language  $ww^R$  ( $w$  consisting of  $a$  or  $b$ ).
- So English could be regular and still *contain* instances of the reversal language.

# How to state the observation correctly

- $(a|b)^*$  is a regular language.
- Assume English is a regular language.
- Only those instances of  $(a|b)^*$  where the string matches  $ww^R$  are in English.  
Or in other words:  $\text{English} \cap (a|b)^* = ww^R$  (over  $a, b$ ).
- Regular languages are closed under intersection.
- We know (from the *Pumping Lemma*) that  $ww^R$  is not a regular language.
- So the finger of doubt points at English. English cannot be a regular language.

## English isn't regular - An example

Examples:     *The boy she saw yesterday was crying.*

*The boy she saw coming down the road was crying.*

Grammar:

$S \rightarrow NP \ VP$  (not allowed)

So we have to write something like:

$S \rightarrow \text{the } X$

$X \rightarrow \text{boy } Y$

$Y \rightarrow \text{she } Z$

$Z \rightarrow \text{saw } Q$

## English isn't regular - The proof

If  $S_1$  then  $S_2$

Either  $S_3$  or  $S_4$

The man who said  $S_5$  is arriving today.

**If** *either* the man who said *either* quit *or* stay is arriving today *or* the man who said  $S_5$  is arriving tomorrow, **then** the man who said  $S_5$  is arriving the day after tomorrow.

Let:	if	$\rightarrow$	a	then	$\rightarrow$	a
	either	$\rightarrow$	b	or	$\rightarrow$	b
	others	$\rightarrow$	$\epsilon$			

Then this sentence is of the form abbbba, which is an instance of  $x x^R$ .

# English isn't regular - Another proof

The cat the dog the rat the elephant admired bit chased likes tuna fish.

Form: (the noun)<sup>n</sup> (transitive verb)<sup>n-1</sup> likes tuna fish.

# Similar point about center-embedding

The cat smelled

The cat the dog chased smelled

The cat the dog the rat bit chased smelled

The cat the dog the rat the elephant admired bit chased smelled

Assume the following homomorphism:

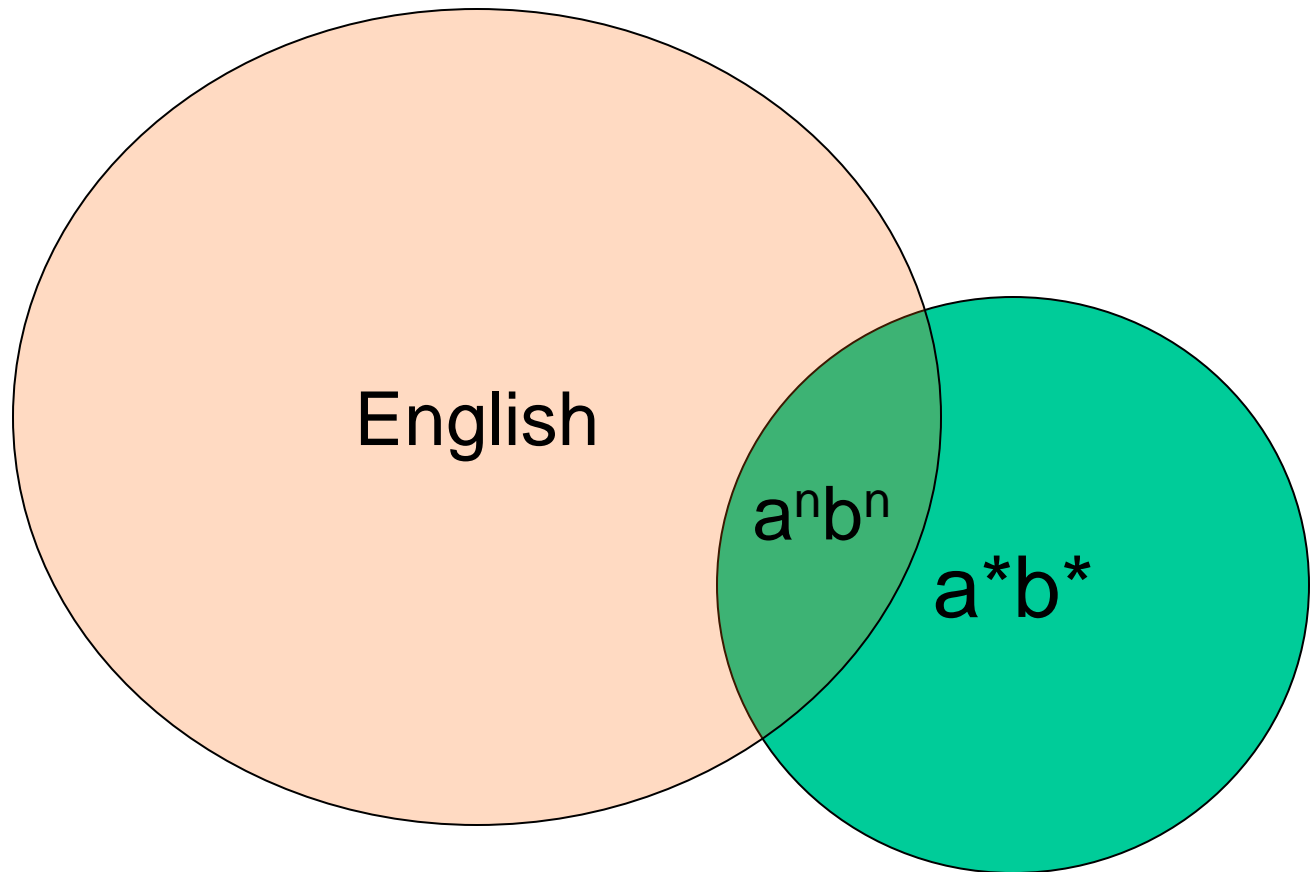
$a = \{\text{the cat, the dog, the rat, the elephant, } \dots\}$

$b = \{\text{chased, bit, admired, ate, smelled, } \dots\}$

Then this is an instance of  $a^n b^n$



# Center embedding



Can we go higher on the hierarchy?

In morphology: reduplication in Bambara  
(Culy, 1985)

In syntax: "cross-serial" dependencies in  
Swiss German (Shieber, 1985)

# Swiss German is not context-free

The nested structures that we've just seen can easily be described with a context-free grammar. But what about sentences of the form  $w w$ :

$x_1 x_2 x_3 x_4 x_5 \dots y_1 y_2 y_3 y_4 y_5 \dots$  (we call these cross serial dependencies)

In Swiss German:

Jan säit das mer em Hans es huus hälfed aastriiche.

Jan says that we Hans/DAT the house/ACC helped paint

