

CSE5026: Foundation and Frontiers of Cognitive Science

Language and Syntax

Yang Xu 徐炆

Department of Computer Science and Engineering
Southern University of Science and Technology

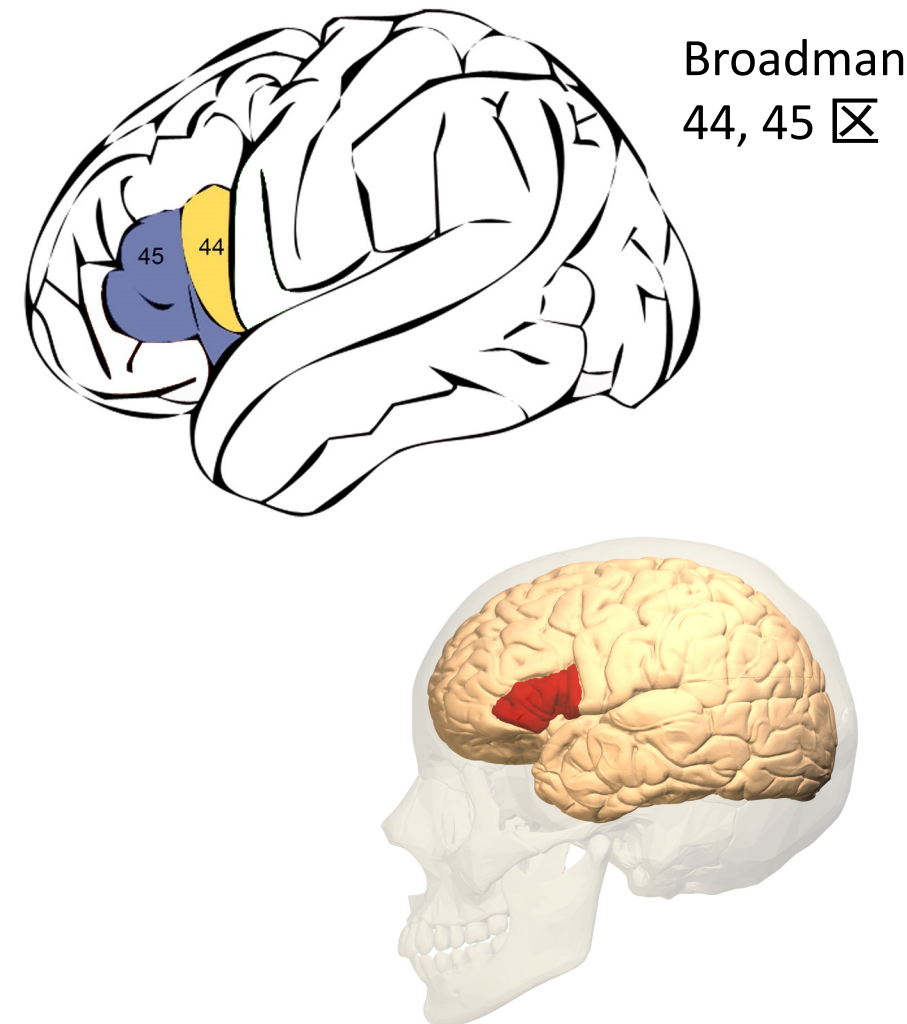
Overview

- Neural Basis of Language Processing 语言加工的神经基础
- Part-of-Speech Tagging 词性解析
- Syntactic Parsing 句法分析
- Cognitive Processing 认知加工

Source: https://en.wikipedia.org/wiki/Broca%27s_area

Neural Basis of Language

- 法国神经学家 Broca 通过研究失语症 (aphasia) 病人，认为产生口语的区域在左半球额叶下回，被称为 **Broca 区**。
- 主要功能：计划和执行说话 (production)。
- 该区病变损伤会导致运动性失语 (expressive aphasia):
 - 无法讲话，但可以理解语言内容。
 - 说话缓慢且不连贯。



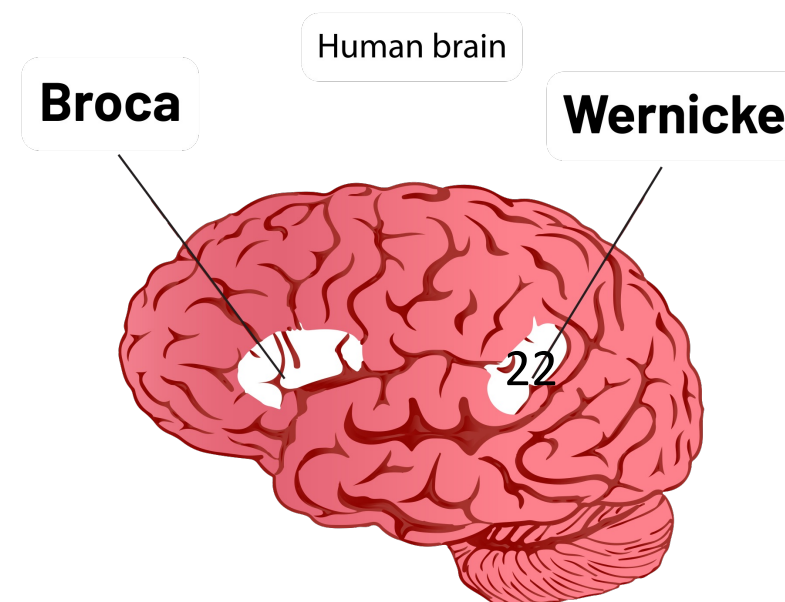
Expressive Aphasia

- Broca's patients:
 - LVL (Tan): He was able to repetitively produce only the word temps (French word for "time").
 - Lelong: He could only say five words, 'yes', 'no', 'three', 'always', and 'lelo' (a mispronunciation of his own name).
- Example:

是...阿...星期一...阿...父親和Peter H(他的姓名)...,
及父親....阿...醫院...及阿...星期三...星期三九點...以及
, 喔...星期二...十點,
阿, 醫生...兩個...醫生...及阿...牙齒...對的。
- Sarah Scott's case: [Aphasia caused by stroke](#) and [recovery after five years](#).

Neural Basis of Language

- 德国神经学家、医生 Wernicke 提出 **Wernicke 区**，病人的该脑区受损后仍能流利说话，但说出的是无意义的声音、词和句子。
 - 位于颞上回的 Broadman 22 区。
 - 主要功能：理解口语和书面语言 (comprehension)。
 - 该区域损伤会导致流利性失语症 fluent aphasia。语言理解能力严重受损，但言语仍保持自然的节奏和相对正常的句法



Fluent Aphasia

- A patient with Wernicke's aphasia was asked what brought him to the hospital. His response was:

Is this some of the work that we work as we did before? ... All right ... From when wine [why] I'm here. What's wrong with me because I ... was myself until the taenz took something about the time between me and my regular time in that time and they took the time in that time here and that's when the time took around here and saw me around in it's started with me no time and I bekan [began] work of nothing else that's the way the doctor find me that way...

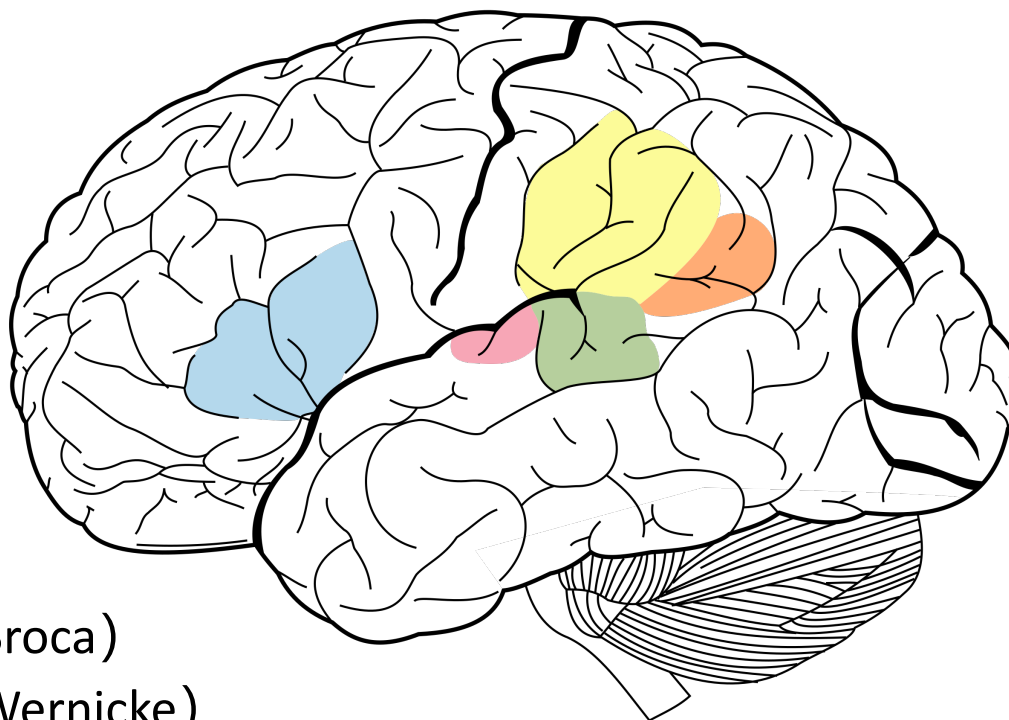
这是我们以前做过的工作吗？ 好吧.....自从我喝了酒之后，我就一直在这里。我怎么了？因为我.....一直都是我自己，直到他们改变了我和我正常时间之间的某些东西，他们改变了我在这里的时间，然后时间就在这里流逝，他们看到我四处游荡，一切都从我没有时间开始，我开始做其他任何事，医生就是这样发现我的.....

- 患者通常意识不到自己的言语能力受损，因为他们对自己言语的理解能力发生了改变。

Source: https://en.wikipedia.org/wiki/Language_center

Language Center

- 大脑中负责语音、语言处理的区域的集合。



- 蓝色: Broca 区
- 绿色: Wernicke 区
- 橙色: 角回 angular gyrus
- 粉色: 初级听觉皮层
- 黄色: 缘上回

主要功能:

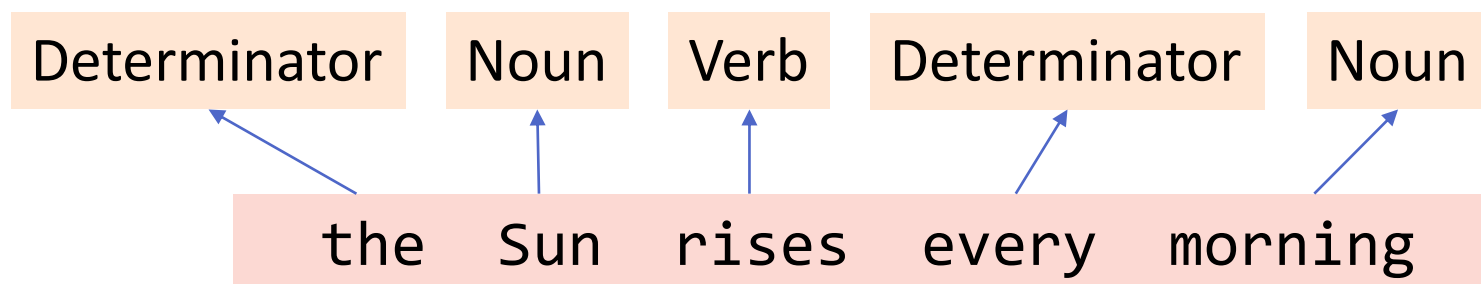
- **运动性**语言中枢 (Broca)
- **听觉性**语言中枢 (Wernicke)
- **视觉性**语言中枢 (Wernicke、角回)

Overview

- Neural Basis of Language Processing 语言加工的神经基础
- **Part-of-Speech Tagging 词性解析**
- Syntactic Parsing 句法分析
- Cognitive Processing 认知加工

Part of Speech

- Words can be classified into grammatical categories



determinator: *n.* 限定词

"n" for Noun 😊

Results returned from:
<https://corenlp.run/>

- These categories are known as **part of speech (POS, POS tags)**, word classes, or simply grammatical categories.

POS in Chinese Language

Not a complete list

实词 Notional Words **or content words**

名词 Nouns	动词 Verbs	助动词 Auxilliary Verbs	形容词 Adjectives	数词 Numerals	量词 Measure Words	人称代词 Personal	指示代词 Demonstrative	疑问代词 Interrogative
国家、妹 妹、玫瑰、 颜色、月亮	工作、学 习、走、 吃、打	应该、可 以、能、 会、想	漂亮、诚实、 慢、坏、红	一、二、 百、万、亿	名量词 Nominal 个、条、 张、公分、 只	动量词 Verbal 次、遍、 回、趟、 轮		
						我们、你、 他、她、咱 们	这、那、各、 每、该	什么、啥、 哪、谁、怎么

unique in Chinese

虚词表 Functional Words

副词 Adverbs	介词 Prepositions	连词 Conjunctions	助词 Particles			叹词 Interjections	象声词 Onomatope
			结构助词 Structural	动态助词 Aspectual	语气助词 Modal		
很、都、就、 也、已经	从、向、在、 被、把	跟、但是、或者、并 且、因为	的、地、得	了、者、过	吗、吧、 呢、了	喂、哎呀、嗯、 哦	哗哗、乒乓

source: https://chinesenotes.com/grammar_intro.html

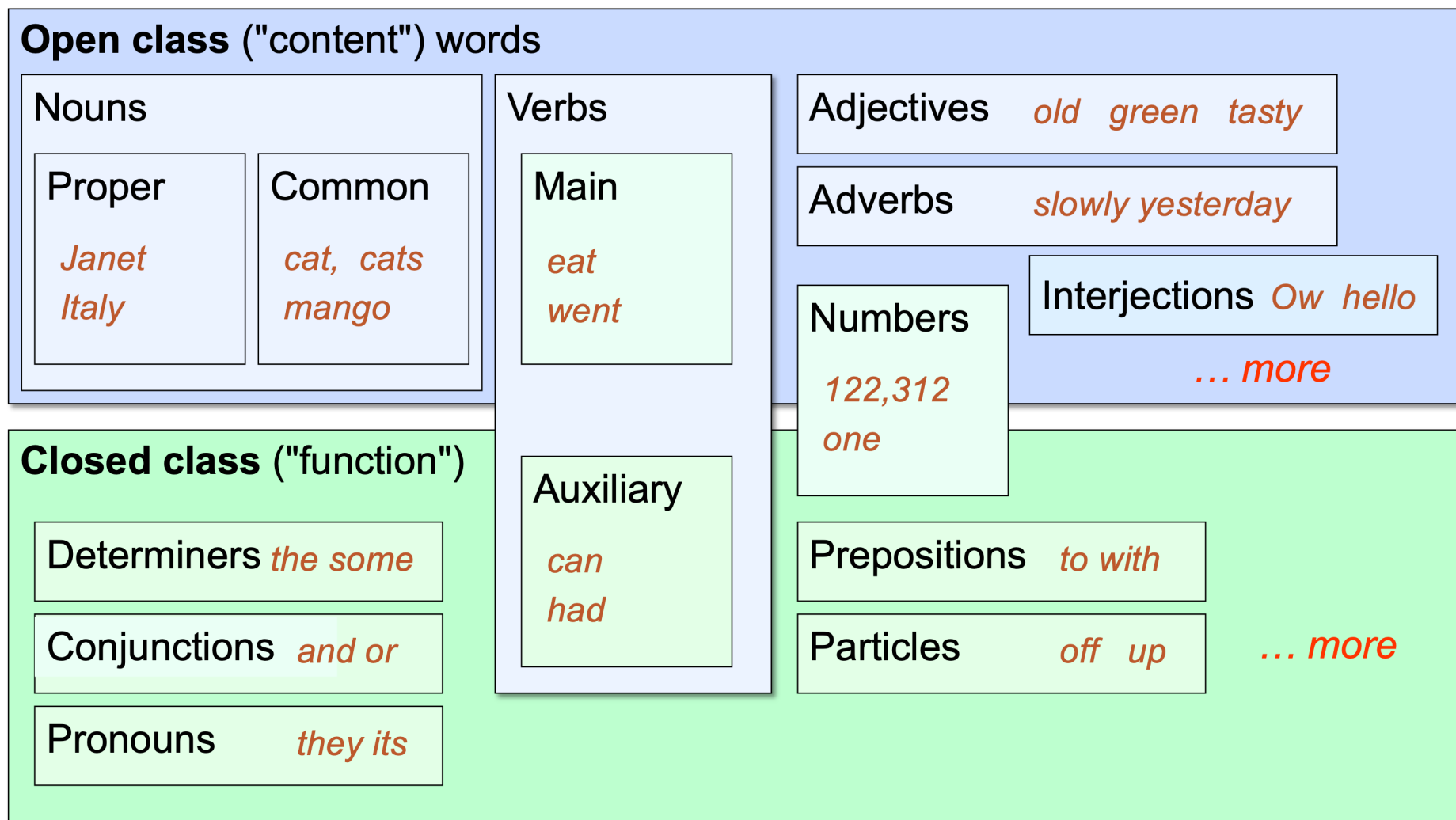
Two classes of words: Open vs. Closed

- Open class
 - Usually **content** words: **Nouns, Verbs, Adjectives, Adverbs**
 - New nouns and verbs like *AI* or *GPT* ... continuously being created/borrowed
- Closed class
 - Relatively fixed membership
 - Usually **function** words: short, frequent words with grammatical function
 - **determiners** (限定词): *a, an, the*
 - **pronouns** (人称代词): *she, he, I*
 - **prepositions** (介词): *on, under, over, near, by, ...*

Adapted from: https://web.stanford.edu/~jurafsky/slp3/slides/8_POSNER_intro_May_6_2021.pptx

Common POS tags in English

Adapted from: https://web.stanford.edu/~jurafsky/slp3/slides/8_POSNER_intro_May_6_2021.pptx



How difficult is POS tagging in English?

- Roughly 15% of word types are ambiguous
- Hence 85% of word types are unambiguous
- E.g., *Janet* is always PROP, *hesitantly* is always ADV
- But those 15% tend to be very common. So ~60% of word tokens are ambiguous

Some ambiguous examples:

Love loves to love love. -- *Ulysses*
NP VP VP NP JAMES JOYCE

Source: https://web.stanford.edu/~jurafsky/slp3/slides/8_POSNER_intro_May_6_2021.pptx

Difficulty of Automatic POS Tagging (in English)

- How many tags are correct? (Tag accuracy)
- About **97%**
 - Hasn't changed in the last 10+ years; **HMMs**, **CRFs**, BERT perform similarly
 - Human accuracy about the same
- But baseline is **92%**!
 - Baseline: stupidest possible method -- "Most frequent class baseline"
 - Tag every word with its most frequent tag (and unknown words as *nouns*)

Source: https://web.stanford.edu/~jurafsky/slp3/slides/8_POSNER_intro_May_6_2021.pptx

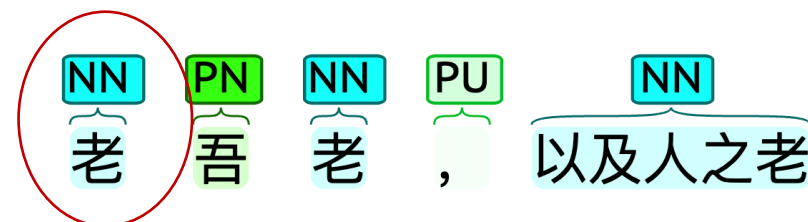
POS Tagging in Chinese

- Similar performance as English

System	F1 score
Tian et. al. (2020)	96.92
Meng et. al. (2019) (Glyce + BERT)	96.61
Meng et. al. (2019) (BERT)	96.06
Shao et. al. 2017	94.38

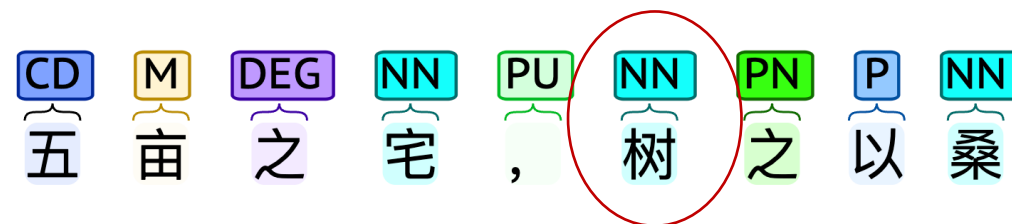
Data source: https://chinesenlp.xyz/docs/pos_tagging.html

Maybe true for modern Chinese (with better labeled data), but not necessarily true for all cases:



Example:

- 快速的棕色狐狸跳过了懒惰的狗
- [快速] VA [的] DEC [棕色] NN [狐狸] NN [跳过] VV [了] AS [懒惰] VA [的] DEC [狗] NN



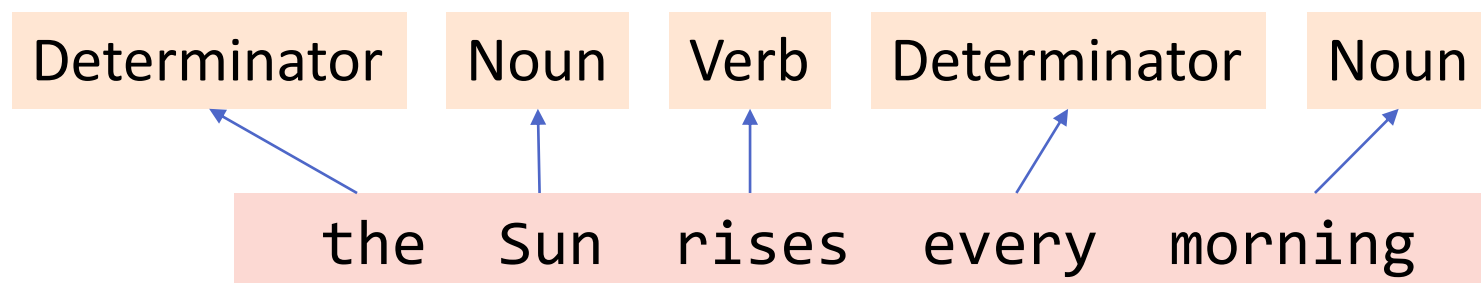
Noun or Verb?

Overview

- Neural Basis of Language Processing 语言加工的神经基础
- Part-of-Speech Tagging 词性解析
- **Syntactic Parsing 句法分析**
- Cognitive Processing 认知加工

Grammar $\approx$ Syntax

- Greek: *syntaxis*, meaning “setting out together”
- The way words are arranged together
- Syntax: *n.* 句法, 句法规则;
- Part-of-speech is already a kind of syntactic notions



Constituency 成分

- Groups of words can behave as **single units**, or **constituents** (*n.* 成分)
- Ex. noun phrases (**NP**, sequence of words surrounding at least one noun)

Harry the Horse
the Broadway coppers
they

a high-class spot such as Mindy's
the reason he comes into the Hot Box
three parties from Brooklyn

- Why they form constituents? (the following NPs all appear before verbs)

three parties from Brooklyn *arrive*...
a high-class spot such as Mindy's *attracts*...
the Broadway coppers *love*...
they *sit*

Invalid sentences:

*from <i>arrive</i> ...	*as <i>attracts</i> ...
*the <i>is</i> ...	*spot <i>sat</i> ...

Evidence for constituency (more examples)

- Preposed (前置) or postposed (后置)

“On February 2nd” is a
prepositional phrase (**PP**)

On February 2nd, I’d like to fly from Shanghai to Osaka.

I’d like to fly on February 2nd from Shanghai to Osaka.

I’d like to fly from Shanghai to Osaka on February 2nd.

* On February, I’d like to fly 2nd from Shanghai to Osaka.

* I’d like to fly on February 2nd from Shanghai to Osaka.

I’d like to fly on February from Shanghai to Osaka 2nd.

Terminology

- Phrase: 词组
- Noun Phrase: 名词词组; Prepositional Phrase: 介词词组
- Proper Noun: 专有名词
- Nominal: 名词性的词

How do we use constituency to describe grammar?

- How do we describe rules such as “*Noun Phrases can occur before verbs*”, “*Prepositional Phrases can occur after verbs*” and so on?
- Need a more formal way!
- -- Context-Free Grammar (CFG)

Context-Free Grammars (CFG)

- Also called Phrase-Structure Grammar
- Wilhelm Wundt (1900), Chomsky (1956), Backus (1959), Naur (1963)
- A set of **rules** (规则) or **productions** (产出) that express the ways symbols can be grouped and ordered together.

Ex. NP $\rightarrow$ Det Nominal
 NP $\rightarrow$ ProperNoun
 Nominal $\rightarrow$ Noun | Nominal Noun

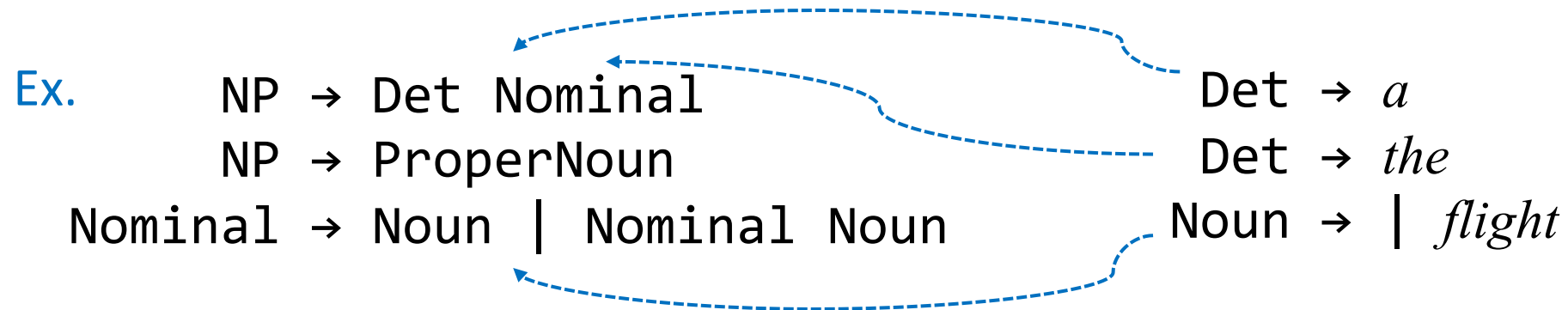
Wundt, German psychologist
Chomsky, American linguist
Backus & Naur, American
computer scientists (FORTRAN,
ALGO)

A **NP (noun phrase)** can be composed of

- either a *ProperNoun* (专有名词)
- or a determiner (*Det* 冠词)
 followed by a *Nominal* (名词性词)

A *Nominal* can be one or more *Nouns* (名词)

Rules can be hierarchically embedded



Two classes of symbols: **terminal** and **non-terminal**

- **terminal**: symbols that are actual words in language (“a”, “the”, ...)
- **non-terminal**: symbols that express abstractions over terminals

CFG as a device for generating sentences (生成)

Rules:

- ① $NP \rightarrow Det\ Nominal$
- ② $NP \rightarrow ProperNoun$
- ③ $Nominal \rightarrow Noun \mid Nominal\ Noun$
- ④ $Det \rightarrow a$
- ⑤ $Det \rightarrow the$
- ⑥ $Noun \rightarrow \mid flight$

Start from *NP*

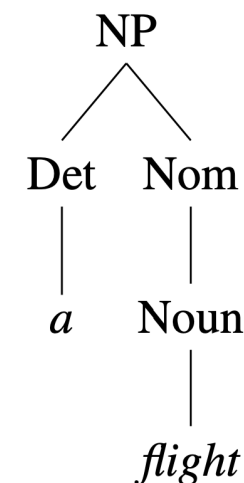
Use rule ① to rewrite *NP* as: *Det Nominal*

Use rule ③ to rewrite *Nominal* as: *Noun*

Use rule ④ to rewrite *Det* as: *a*

Use rule ⑥ to rewrite *Noun* as: *flight*

Result: **parse tree** 

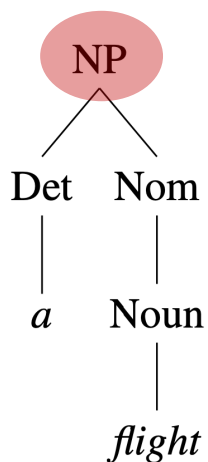


CFG: More terms and rules

A grammar must have one designated **start symbol**, often called **S**.

Common rules:

NP **dominates** all nodes in the tree



A sentence can consist of a noun phrase followed by a verb phrase

$S \rightarrow NP \ VP$ *I prefer a morning flight*

A VP can consist of a verb followed by assorted things

$VP \rightarrow Verb \ NP$ *prefer a morning flight*

Or VP may have a verb followed by a prepositional phrase

$VP \rightarrow Verb \ PP$ *leaving on Thursday*

A PP generally has a preposition followed by a noun phrase

$PP \rightarrow Preposition \ NP$ *from Los Angeles*

CFG: Examples from ATIS corpus

A sample of grammar rules $\mathcal{L}_0$

Grammar Rules	Examples
$S \rightarrow NP VP$	I + want a morning flight
$NP \rightarrow$	
<i>Pronoun</i>	I
<i>Proper-Noun</i>	Los Angeles
<i>Det Nominal</i>	a + flight
$Nominal \rightarrow$	
<i>Nominal Noun</i>	morning + flight
<i>Noun</i>	flights
$VP \rightarrow$	
<i>Verb</i>	do
<i>Verb NP</i>	want + a flight
<i>Verb NP PP</i>	leave + Boston + in the morning
<i>Verb PP</i>	leaving + on Thursday
$PP \rightarrow$	
<i>Preposition NP</i>	from + Los Angeles

Air Travel Information System (ATIS) pilot corpus <https://aclanthology.org/H90-1021/>

Use ATIS grammar to generate “AIS-language”

$S \rightarrow NP \ VP$

Start with S, expand it with NP VP



Choose a random expansion of NP: *I*

$NP \rightarrow$	<i>Pronoun</i>	<i>I</i>
	<i>Proper-Noun</i>	Los Angeles
	<i>Det Nominal</i>	a + flight



Choose a random expansion of VP: *Verb NP*

$VP \rightarrow$	<i>Verb</i>	do
	<i>Verb NP</i>	want + a flight
	<i>Verb NP PP</i>	leave + Boston + in the morning
	<i>Verb PP</i>	leaving + on Thursday



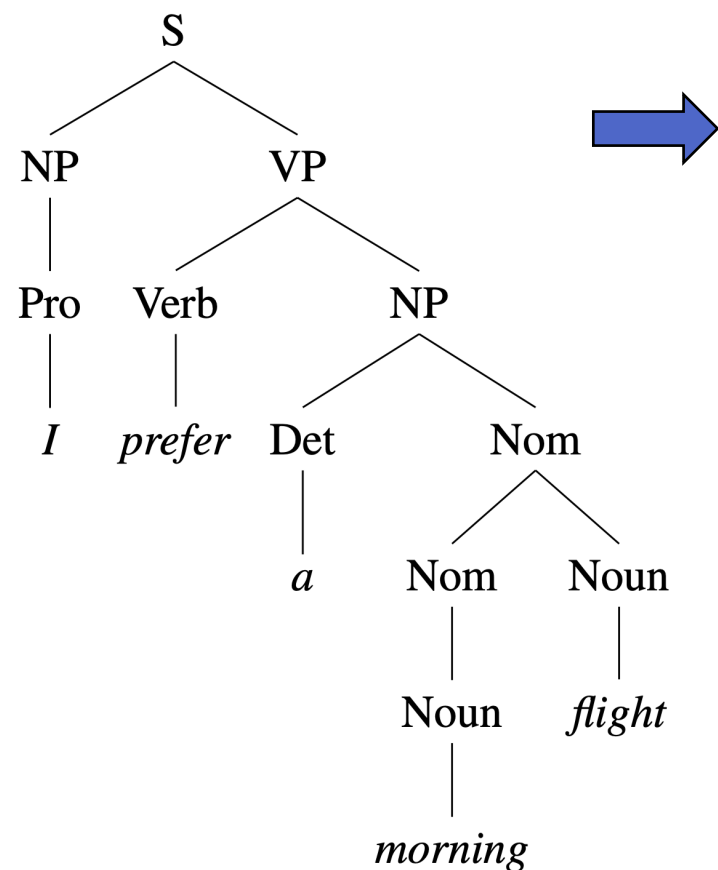
Verb NP

- want + a flight
- want + a morning flight
- prefer + a morning flight*
-



I prefer a morning flight

Syntactic Tree: Bracketed Notation



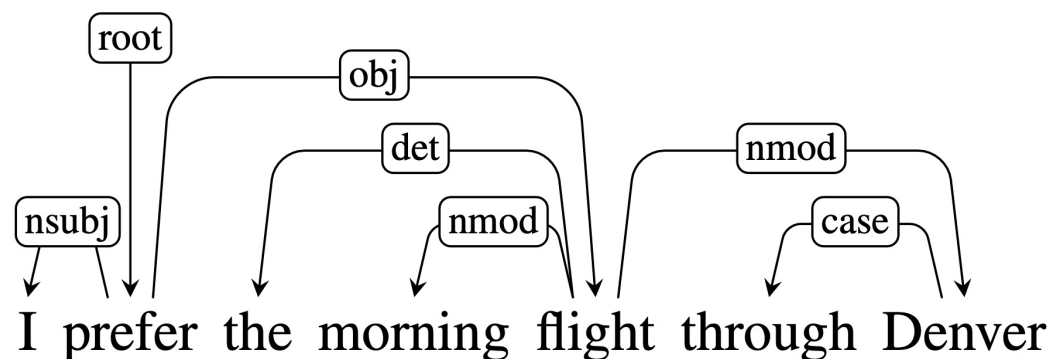
(S (NP (PRO I)
 (VP (V prefer)
 (NP (Det a) (Nom (N morning) (N flight))))))

- Sentences that can be derived by a grammar are called **grammatical**;
- those cannot be derived are referred to as **ungrammatical**.

This distinguishment is vague, and only a very *simplified* model of how natural language really work.

Dependency Grammars 依赖关系语法

- Different from context-free grammars and constituency-based representations
- **Dependency Grammars** describe syntactic structure of a sentence solely in terms of directed grammatical *relations between words* arc: *n.* 弧线



Labeled arcs from **heads** to **dependents**

$prefer \xrightarrow{\text{nsubj}} I$

prefer is the head of *I*

$prefer \xrightarrow{\text{obj}} flight$

flight is the dependent of *prefer*

$root \rightarrow prefer$

root is the head of *prefer* and also the head of the entire structure (root of the tree)

Overview

- Neural Basis of Language Processing 语言加工的神经基础
- Part-of-Speech Tagging 词性解析
- Syntactic Parsing 句法分析
- **Cognitive Processing 认知加工**

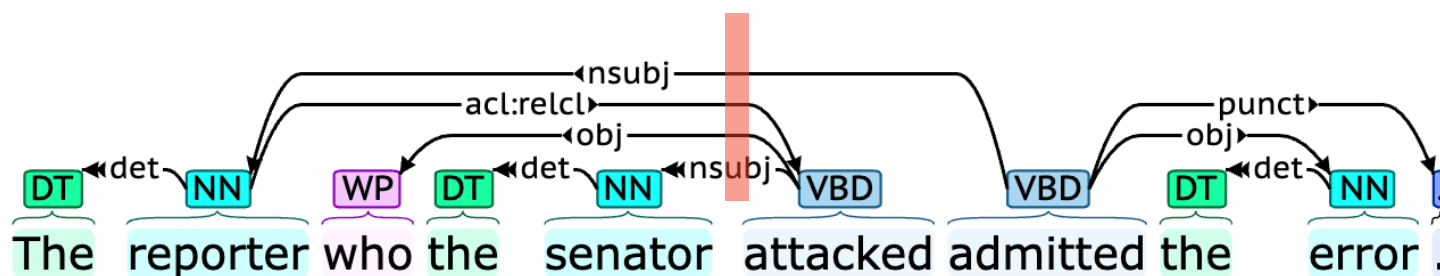
Source: Gibson (2000), [https://doi.org/10.1016/S0010-0277\(98\)00034-1](https://doi.org/10.1016/S0010-0277(98)00034-1)

Cognitive Effort in Processing Language

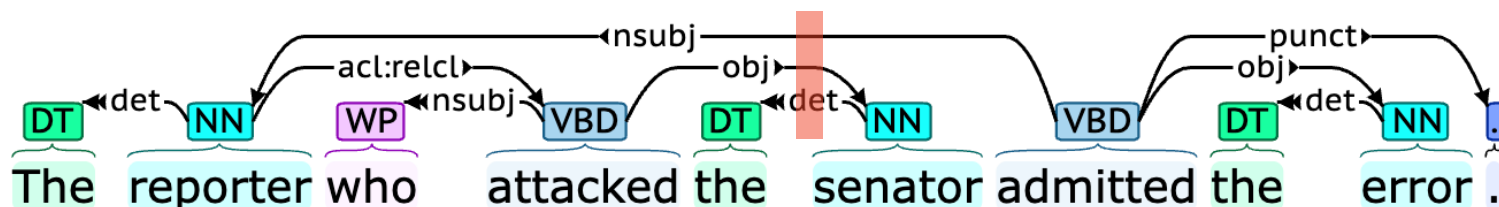
- Theory: **Locality** (局部性) of syntactic dependencies
- 0: The reporter admitted the error.
- 1a: The reporter **who the senator attacked** admitted the error.
- 1b: The reporter **who attacked the senator** admitted the error.
- Both relative clauses (RCs)
 - 1a => “who” is extracted from the **object** position of RC. => Hard
 - 1b => “who” is extracted from the **subject** position of RC. => Easy

Dependency parsed sentences

- 1a: The reporter **who the senator attacked** admitted the error.



more stacked arrows

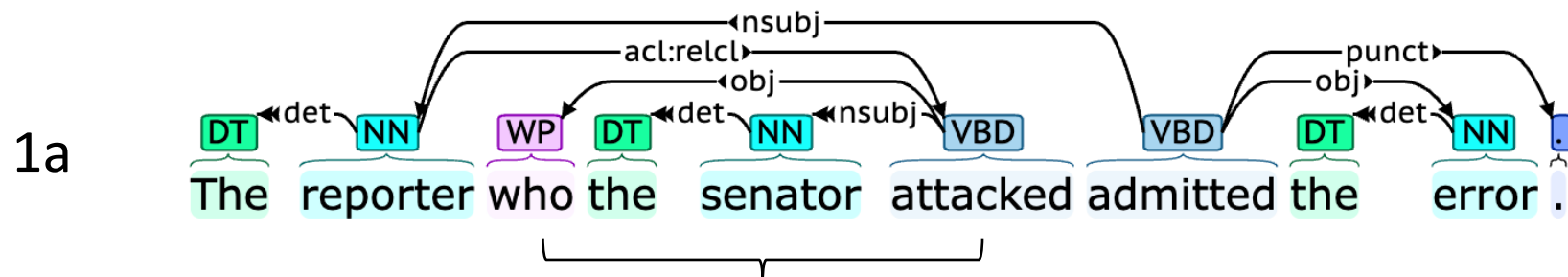


fewer stacked arrows

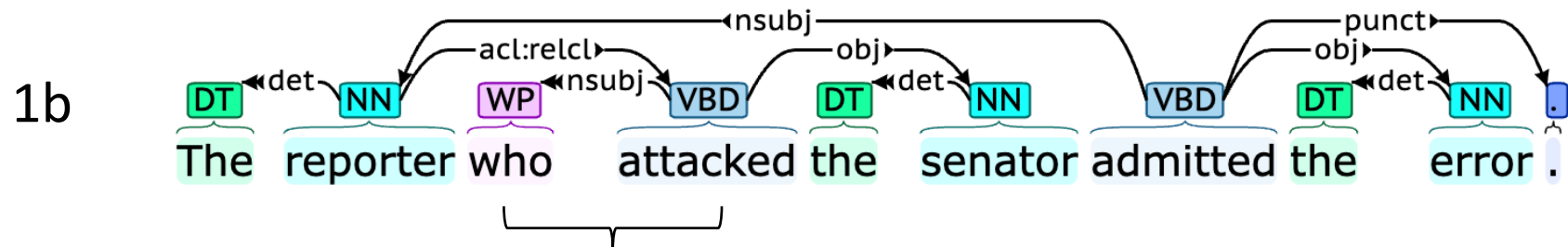
- 1b: The reporter **who attacked the senator** admitted the error.

The memory costs are different

- One explanation: the distance between “who” and “attacked” determines the memory cost to resolve their dependency relation -- locality.



less “local”, more memory cost to wait for “the senator” to be parsed.



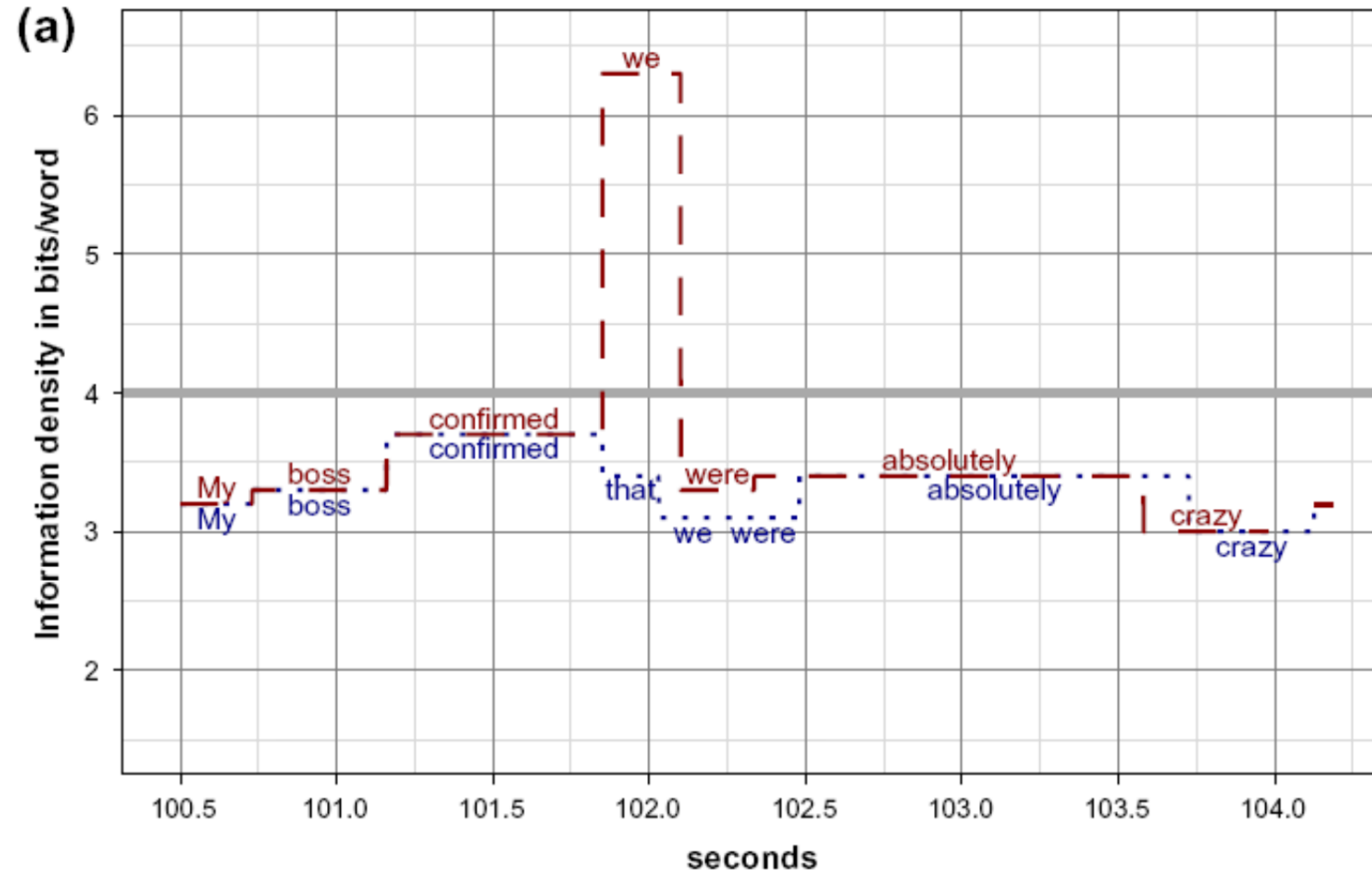
more “local”, less memory cost

Cognitive Effort in Processing Language

- Theory: Uniform information density (**UID**).
 - Human speakers tend to distribute the amount of information uniformly in natural language.
- 1a: My boss confirmed we were absolutely crazy.
- 1b: My boss confirmed **that** we were absolutely crazy.
- Which one reads more smoothly?

Source: Jaeger 2010, [doi:10.1016/j.cogpsych.2010.02.002](https://doi.org/10.1016/j.cogpsych.2010.02.002)

Uniform Information Density



Human speakers tend to avoid the sudden changes in information.