2.1. 自然语言的特点

+ 有限规则集（具递归性）

This is the cat.

This is the cat that caught the rat.

This is the cat that caught the rat that ate the cheese. ……

- This is the house
- This is the house that Jack built
- This is the grain that lay in the house that Jack built
- This is the rat that ate the grain that lay in the house that Jack built
- This is the cat that killed the rat that ate the grain that lay in the house that Jack built
- This is the dog that chased the cat that killed the rat that ate the grain that lay in the house that Jack built

2.1. 自然语言的特点

Recursive Structures

NP → NP PP The flight to Boston

VP → VP PP departed Miami at noon

Flights to Miami

Flights to Miami from Boston

Flights to Miami from Boston in April

Flights to Miami from Boston in April on Friday

Flights to Miami from Boston in April on Friday under $300.

Flights to Miami from Boston in April on Friday under $300 with lunch.

Conjunctions

S → S and S

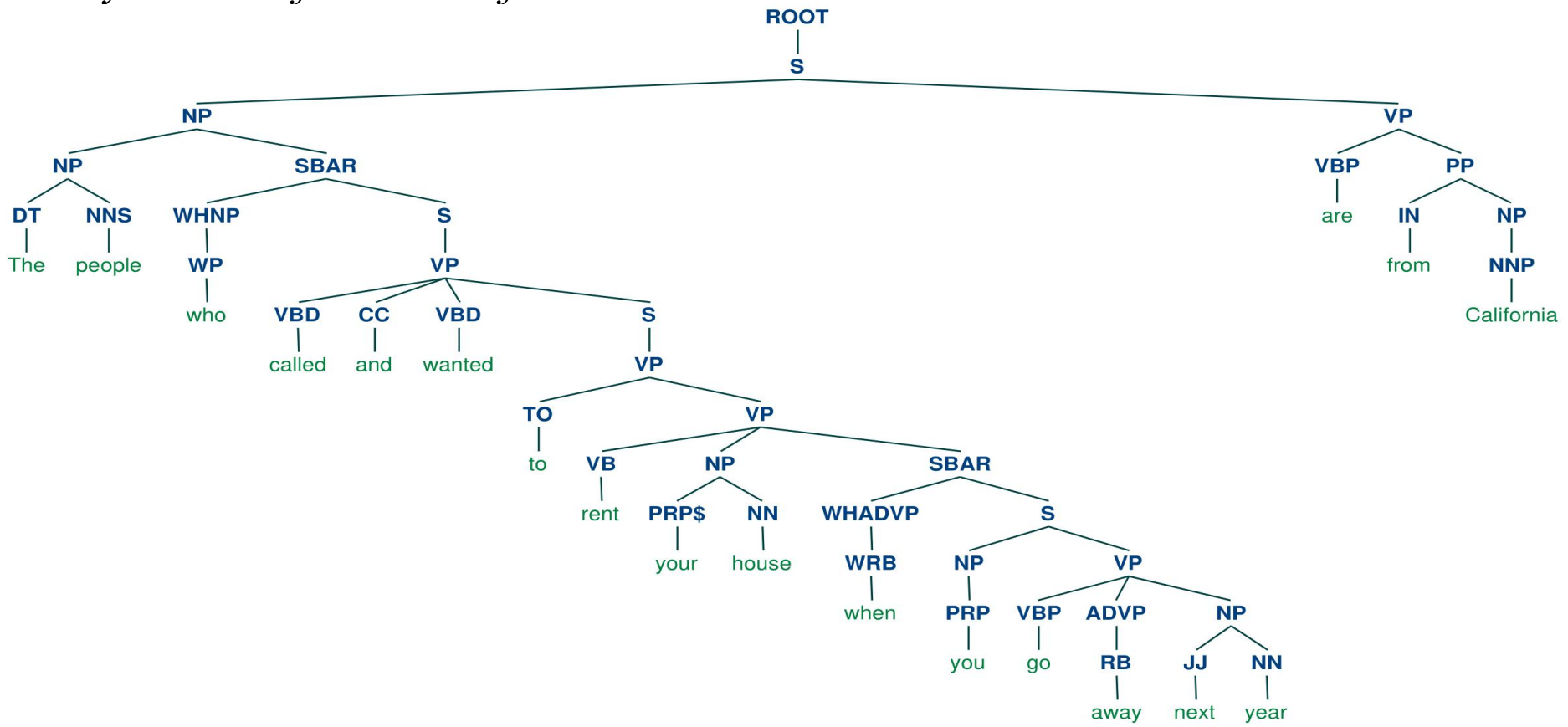
NP → NP and NP

VP → VP and VP

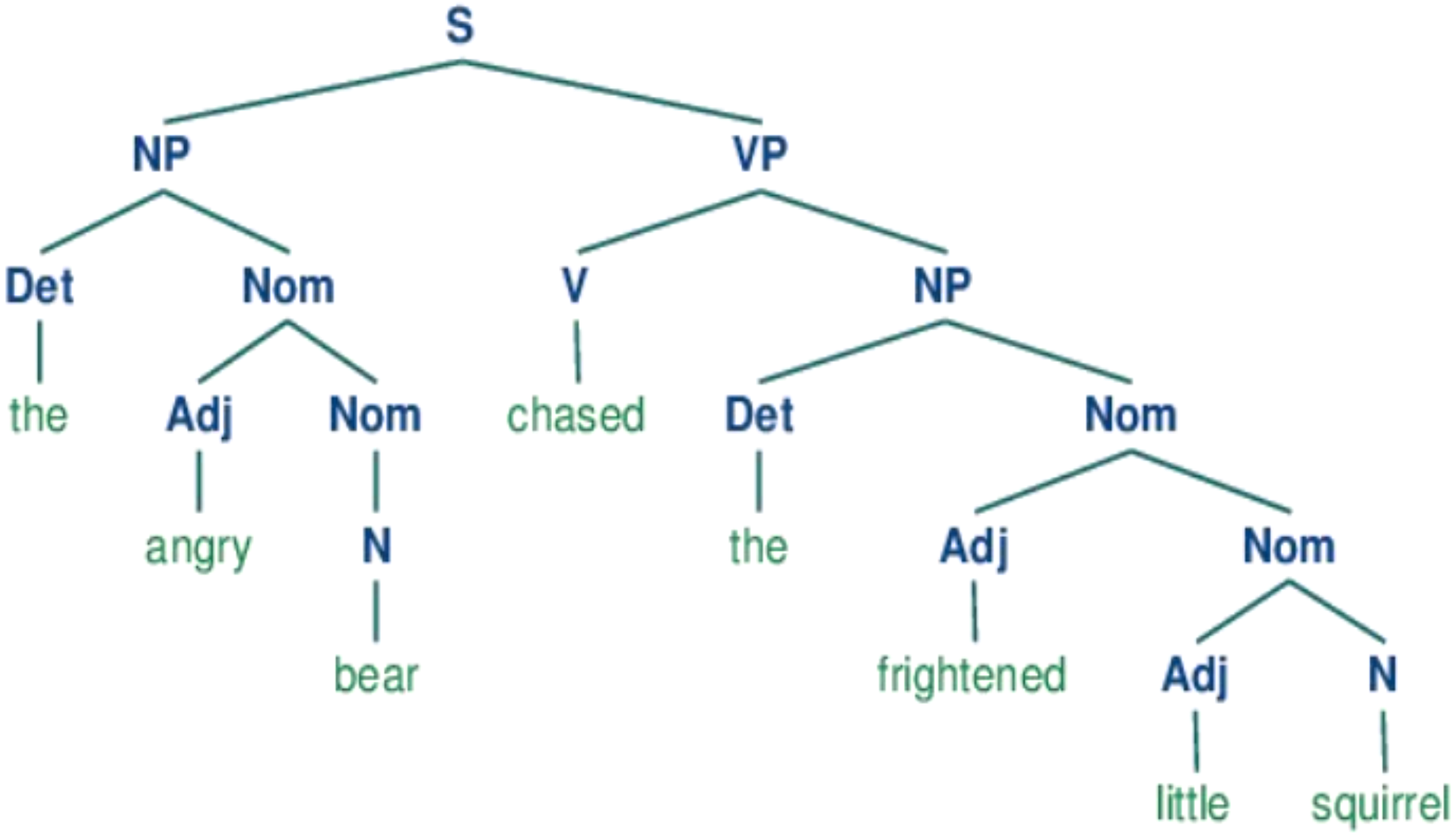
2.1. 自然语言的特点

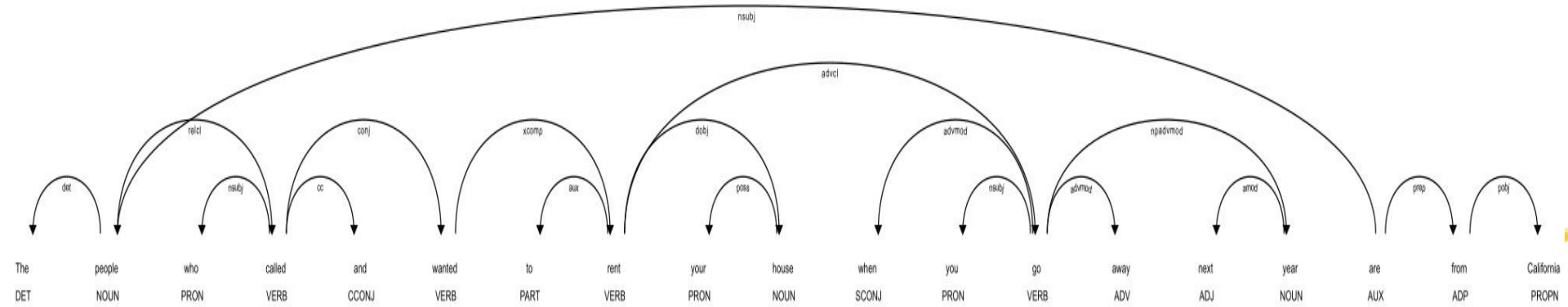
结构性

The people who called and wanted to rent your house when you go away next year are from California



2.1. 自然语言的特点


S
NP
VP
Det
Nom
V
NP
the
Adj
Nom
chased
Det
Nom
angry
N
the
Adj
Nom
bear
frightened
Adj
N
little
squirrel



http://stanza.run/

Stanza – A Python NLP Package for Many Human Languages

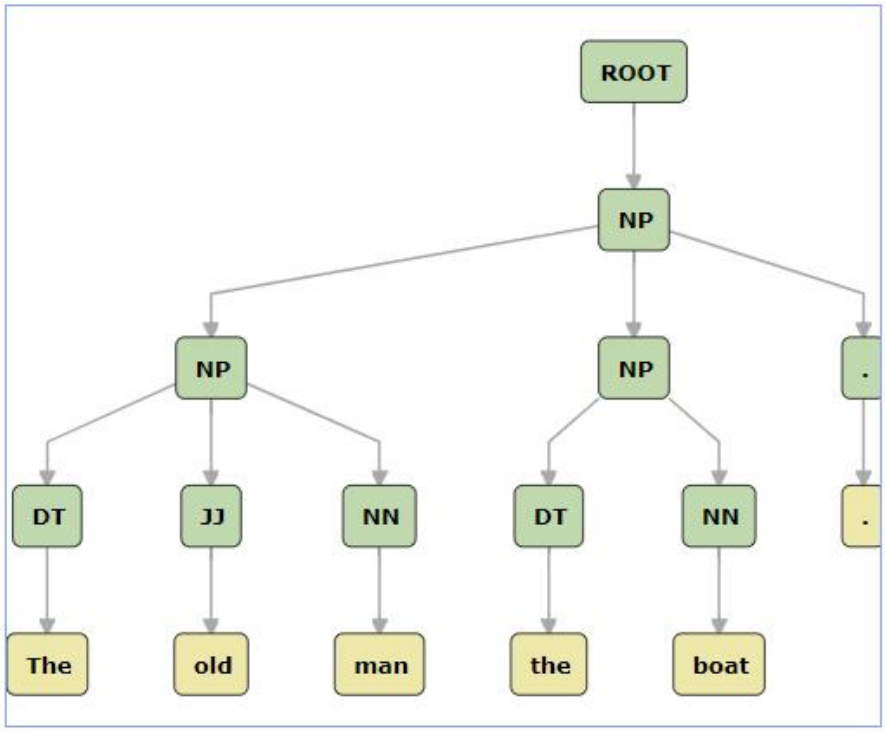
pypi v1.6.0 | conda v1.5.0 | python 3.6 | 3.7 | 3.8 | 3.9 | 3.10

Stanza is a collection of accurate and efficient tools for the linguistic analysis of many human languages. Starting from raw text to syntactic analysis and entity recognition, Stanza brings state-of-the-art NLP models to languages of your choosing.

https://spacy.io/

Garden path sentence, which is a sentence that is grammatically correct, but due to the way it's divided and structured, can seem ungrammatical or nonsensical.

The old man the boat.



After Bill drank the water proved to be poisoned.

I told the girl the cat scratched Bill would help her.

2.1. 自然语言的特点

- **歧义性（Ambiguity）**

Lexical ambiguity

多音字(词)（polyphone）

朝辞白帝彩云间,
千里江陵一日还。

数风流人物,
还看今朝。

2.1. 自然语言的特点

多义词 (Polysemy)

同形异义字(词) （homograph）

“Minute”: (1) a unit for measuring time(noun); (2) to make a written record of what is said or decided. during a meeting(verb); (3) tiny(adj)

1a. One minute has sixty seconds.

1b. Part of the job of a secretary is to minute meetings.

1c. There is only minute difference between these pictures.

“编辑”

2.1. 自然语言的特点

结构歧义（Structural ambiguity）

亚洲语言学会　　n+n+n　　（句法结构歧义）

彩色铅笔盒子　　n+n+n　　（句法结构歧义）

关于鲁迅的书　　prep+n+的+n　（句法结构歧义）

他讲不清楚。　　v+不+adj　　（句法结构歧义）

漂亮的姑娘和小伙子 adj+的+n+的+n　（句法结构歧义）

小张的处理意见　　（语义结构歧义）

他在看病。　　（语义结构歧义）

他借我一本书。　　（语义结构歧义）

2.1. 自然语言的特点

他气死了。
他被气死了。
你气死我了!
诸葛亮气死了周瑜。

热爱人民的总理　v+n+的+n
咬死 猎人的鸡　咬死 | 猎人的鸡
咬死鸡的 狗　咬死鸡的 | 狗
咬死 猎人的狗　咬死 猎人的 | 狗　咬死 | 猎人的狗

2.1. 自然语言的特点

- 持续演化性
- 语言表达的非规范性（自由性）

他北京人。
快走吧，你!
他打开门，走了进来，悄悄地。

- 模糊性　“下半旗”
- 文化差异性

“lying on top of a bed in English and Chinese”:
“in bed” (English) and “在床上” (Chinese)

2.1. 自然语言的特点

- 经济性原则

最早来源于哈佛大学教授G.K.齐夫专著《人类行为与省力原则》（Human Behavior and Principle of Least Effort）中的“省力原则”

法国语言学家马丁内（Martinet）进一步提出了“经济原则”：人的生理及精神上的自然惰性与人的交际表达需要之间存在着矛盾

缩略语

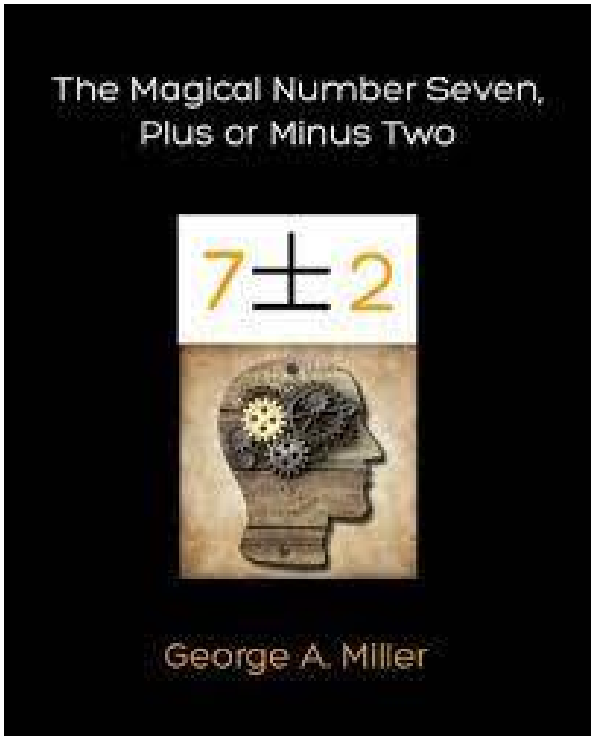
诸葛亮气死了周瑜。

他喝醉了酒。

他气坏了身子。

2.1. 自然语言的特点

Miller's law:
The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information



+ People's maximum performance on a one-dimensional absolute judgment can be characterized as an information channel capacity with approximately 2 to 3 bits of information, which corresponds to the ability to distinguish between four and eight alternatives.
+ Memory span is not limited in terms of bits but rather in terms of chunks. (A chunk is the largest meaningful unit in the presented material that the person recognizes)
+ A coincidence between the limits of one-dimensional absolute judgment and the limits of short-term memory.

2.1. 自然语言的特点

- 统计性

我爱吃红 _____

Shannon proposed an interesting scheme to generate text according to a Markov model of order 1.

To construct [an order 1 model] for example, one opens a book at random and selects a letter at random on the page. This letter is recorded. The book is then opened to another page and one reads until this letter is encountered. The succeeding letter is then recorded. Turning to another page this second letter is searched for and the succeeding letter recorded, etc. It would be interesting if further approximations could be constructed, but the labor involved becomes enormous at the next stage.

2.2. Complexity of Natural Languages

Grammar:

$$G = (V_N, V_T, S, P)$$

where

V_N: Finite non-empty set of non-terminal symbols

V_T: Finite set of terminal symbols

P: Finite non-empty set of production rules

S: Start symbol, $S \in V_N$

2.2. Complexity of Natural Languages

$$G: \quad S \rightarrow aA$$
$$A \rightarrow cA$$
$$A \rightarrow b$$

$$G = (V_N, V_T, S, P)$$

$$V_N = \{S, A\} \qquad V_T = \{a, b, c\}$$

$$P = \{S \rightarrow aA,\ A \rightarrow cA,\ A \rightarrow b\}$$

2.2. Complexity of Natural Languages

$$G: \quad S \rightarrow aSb$$
$$S \rightarrow \varepsilon$$

$$G = (V_N, V_T, S, P)$$

$$V_N = \{S\}$$

$$V_T = \{a, b, \varepsilon\}$$

$$P = \{S \rightarrow aSb,\ S \rightarrow \varepsilon\}$$

2.2. Complexity of Natural Languages

Sentential Form:

A sentence that contains non-terminals and terminals

Example:

$$S \Rightarrow aA \Rightarrow acA \Rightarrow accA \Rightarrow accb$$

Sentential Forms sentence

Derivation:

$$S \overset{*}{\Rightarrow} accb$$

2.2. Complexity of Natural Languages

Example2:

$$S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaaSbbb \Rightarrow aaabbb$$

Sentential Forms

sentence

Derivation: $S \overset{*}{\Rightarrow} aaabbb$

2.2. Complexity of Natural Languages

Language of a Grammar

For a Grammar G

With start symbol S

$$L(G) = \{w : S \overset{*}{\Rightarrow} w\}$$

String of terminals

2.2. Complexity of Natural Languages

Chomsky hierarchy

The Chomsky hierarchy is an ordering of types of grammar according to generality. The classification in fact only depends on the type of grammar rule (rewrite rule) used. The grammar types include:

unrestricted grammars(type 0): rules of the form $\alpha \rightarrow \beta$ with no restrictions on the sequence of symbols α and β .

context sensitive grammars(type 1): rules of the form $\alpha X \beta \rightarrow \alpha \Psi \beta$ where X is a non-terminal symbol, α and β are (possibly empty) sequences of symbols, and Ψ is nonempty sequence of symbols.