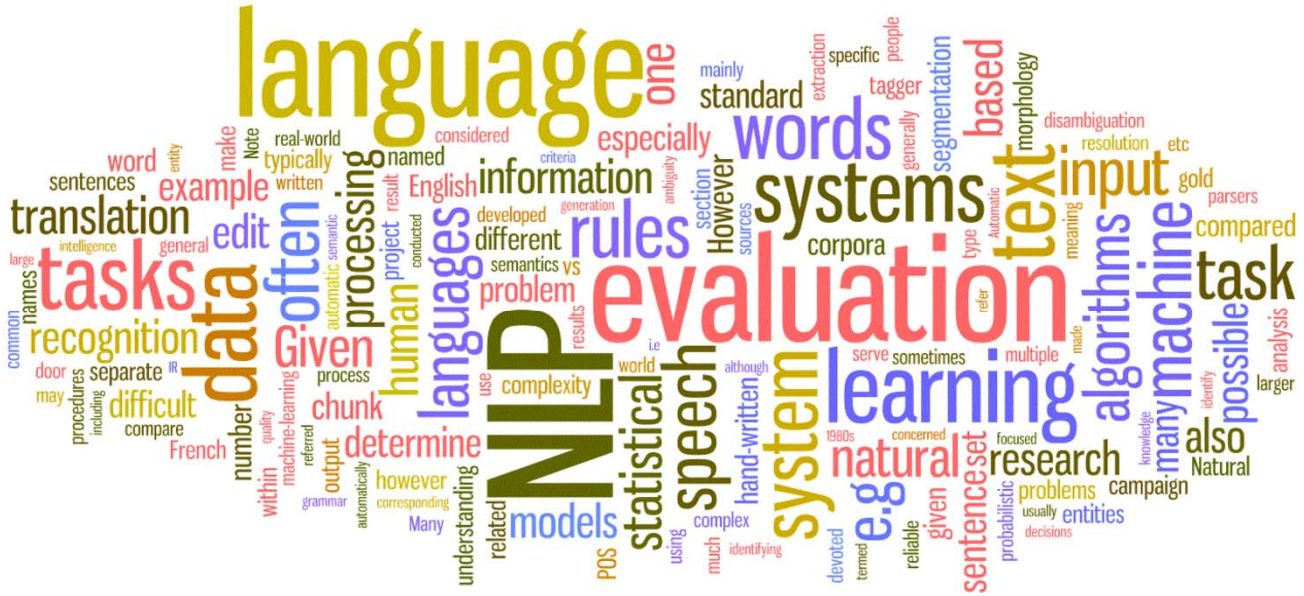


Lecture 2: Linguistics Foundations

Xu Ruifeng

Harbin Institute of Technology, Shenzhen

Last Time

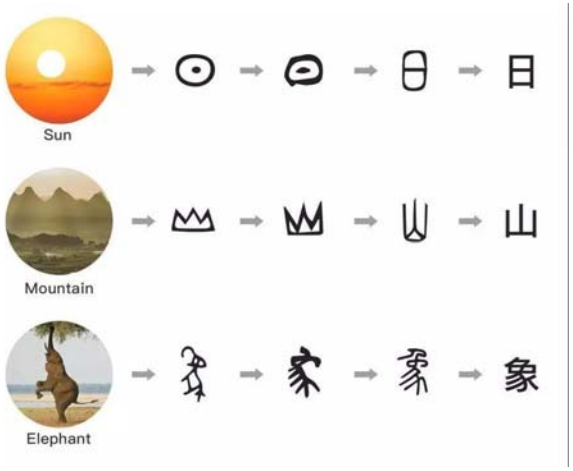
- Course Design
- What Is Natural Language Processing (NLP)
- What Can NLP Technique Do
- The Difficulties Faced
- The NLP History
- General NLP Approach

Today's class – Linguistics Foundations

- Chinese Characters, Morphemes and Words
- Part-of-Speech
- Challenges in Chinese Morphological Processing
- Phrase Grammar and Parsing
- Discourse Parser
- Corpus Linguistics

Chinese Characters 汉字

- Chinese character is non-alphabetic symbol
- The live Ideographic character



pictograph 象形文字

意音 logograph

表意文字/形意文字 (Ideograph)

- Chinese character information

Type	Example	Amount
Pictograph 象形	日月火人	Few
Ideographic 指示	上下	Very few
Compound indicative 会意	仁、信	Very few
Semantic-phonetic compounds 形声	请、情、清、晴 钾、钠、钙、镁	> 90%

- Chinese character is morphemes syllable words in ideographic writing system.

Glyph-vectors Representations

- Using Historical Scripts

Chinese	English	Time Period
金文	Bronzeware script	Shang and Zhou dynasty (2000 BC – 300 BC)
隶书	Clerical script	Han dynasty (200BC-200AD)
篆书	Seal script	Han dynasty and Wei-Jin period (100BC - 420 AD)
魏碑	Tablet script	Northern and Southern dynasties 420AD - 588AD
繁体中文	Traditional Chinese	600AD - 1950AD (mainland China). still currently used in HongKong and Taiwan
简体中文(宋体)	Simplified Chinese - Song	1950-now
简体中文(仿宋体)	Simplified Chinese - FangSong	1950-now
草书	Cursive script	Jin Dynasty to now

Table 1: Scripts and writing styles used in Glyce.

- The Tianzige-CNN Structure for Glyce

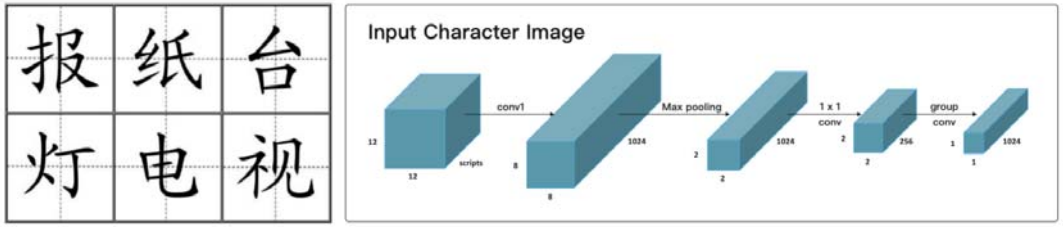


Figure 1: Illustration of the Tianzege-CNN used in Glyce.

- Image Classification as an Auxiliary Objective

Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li et al. Glyce: Glyph-vectors for Chinese Character Representations. NeurIPS 2019.

Large Number of Characters

- Much larger number of characters as compared with the number of letters
- The exact number of existent Chinese characters cannot be precisely ascertained
  - 康熙字典 (Kangxi dictionary 1716) 47, 035
  - 中华字海 (Zhonghua Zihai dictionary 1994) 87, 019
  - 1000 characters may cover 92% written materials, and 3000 characters cover more than 99%
- For computer processing purpose, Chinese characters are encoded in 16+ bits.

Traditional and Simplified Characters

- Traditional/simplified characters
  - Simp. Chinese: China, Singapore, Malaysia, United Nations.
  - Trad. Chinese: Taiwan, Hong Kong and Macau

- No one-one corresponding between
  - 乾 'dry' and 幹 'to do' -> Simplified 干
  - 遊 'travel' and 游 'swim' -> Simplified 游

    親不見，愛無心，產不生，廠空空，麵無麥，運無車，導無道，
    兒無首，飛單翼，有雲無雨，開関無門
    护(護)用手，爱(愛)有友，灶(竈)生火，显(顯)明明，龟(龜)
    有甲，笔(筆)有毛，宝(寶)有玉，众(眾)有人，网(網)像形，
    灭(滅)无需水，呼吁(呼籲)有口

Variant Character

- Characters that have the same meaning and sounds but different shapes 异体字
- Most of the characters in the Kangxi dictionary are variant character
  - Four variant characters of 回(回囘囲廻)
- Often share the same components as their standard counterparts
  - 裏/裡; 膀/髈; 杯/盃; 秘/祕; 毙/斃
- Becomes hot in Internet
  - 囧 (窘)

Dialect characters and Dialectal Use of Standard Characters

- The existence of dialectal characters 方言字

Cantonese	Meaning	Mandarin
而家	Now	现在
同埋	And	和
边个	Who	哪位
边度	Where	哪儿

Shanghai	Meaning	Mandarin
侬	You	你
伊	He/she	他/她
伐	Not	不
白相	Play	玩

Southern Min	Meaning	Mandarin
阮	I/We	我(们)
暗	Late	晚
郎	Person	人
呷	Eat	吃

Multiple Character Encoding Standards

- GB “National Standard” in Chinese
  - 7445 characters. It includes 6, 763 simplified characters. Class-1/Class-2 characters
  - BG2312-80, contained only one code point for each character.
  - MSB. Bit-8 of each byte, is set to 1, and therefore becomes a 8-bit character. Otherwise, the byte is interpreted as ASCII
  - Every Chinese character is represented by a two-byte code. The MSB of both the first and second bytes are set.

Multiple Character Encoding Standards

- GBK “National Standard Extension” in Chinese
  - An extension of GB2312
  - Includes 14, 240 traditional characters
  - The scheme is used by Simplified Microsoft Windows 95 and 98

- GB18030
  - Released by the China Standard Press, 2000
  - BG18030 supersedes all previous versions of GB
  - Officially mandatory for all software products sold in the PRC
  - Supports both simplified and traditional Chinese characters

Big 5

- Traditional Chinese characters 13, 000

- Every Chinese Character is represented by a two byte code.
  - The first byte ranges from 0xA0 to 0xF9
  - The second byte ranges from 0x40 to 0x7E, 0xA0 to 0xFE.
  - ASCII characters are still represented with a single byte.
  - The MSB of the first byte of a Big5 character is always 1;
  - Big5 is and 8-bit encoding with a 15-bit code space.

Unicode

- Industry standard, Universal Character Set
  - More than 100, 000 characters
  - Originates from East Asia
- Implemented by different character encodings
  - UTF-8: uses 1 byte for all ASCII characters, up to 4 bytes for other characters
  - UCS-2: uses 2 bytes for all characters, but does not include every character in the Unicode
  - UTF-16: using 4 bytes to encode characters missing from UCS-2
- Simplified and traditional characters as part of the project of Han unification

Example

	自	然	语	言	处	理
GB2312	D7 D4	C8 BB	D3 EF	D1 D4	B4 A6	C0 ED
Big5	A6 D8	B5 4D	BB 79 (語)	A8 A5	B3 A2 (處)	B2 7A
UTF-8	E8 87 AA	E7 84 B6	E8 AF AD	E8 A8 80	E5 A4 84	E7 90 86
UTF-16	EA 81	36 71	ED 8B	00 8A	04 59	06 74

Character-based Language Processing

- Character-based Chinese Processing 字本位

- Transliteration Detection

  Mississippi – 密西西比
  France – 法兰西 vs. 发懒西 发烂西
  America – 阿美利加 美利坚 vs. 没力坚
  Internet – 英特网

  - Character/phonetic based mapping and combination testing

Character-based Language Processing

- Word Sentiment Guessing

  佳人 佳期 佳信 上佳 – 佳缘
  劣势 劣行 劣性 劣等 – 劣根性 劣马
  喜事 喜爱 喜庆 喜色 – 喜兴 喜羊羊

- Lexicon semantic similarity computing

- Character-based word segmentation and new word detection

Character Embedding

- To morphology rich language (English, French)
  - Word lookup table is large
  - Large parameter space for word embedding
  - Word embedding cannot describe the relations between morphological related words
  - Word embedding cannot handle OOV/new word/typos

Xiang Zhang, Junbo Zhao, Yann LeCun. Character-level convolutional networks. NIPS 2015
Yoon Kim, Yacine Jernite, David Sontag, Alexander M.Rush. Character-Aware Neural Language Models. ArXiv.

Morphemes 词素

- The basic morphological units, i.e., the smallest meaningful elements
- Cannot be further analyzed into smaller parts and still retain meaning
- Generalization that one syllable is one morpheme represented by one character
- One character morphemes
  - 吧：酒吧 清吧 咖啡吧
- Multi-character morphemes
  - 葡萄 菩萨 马达：葡萄酒 葡萄干

English Word

- Consider the word “unhappiness”
  - Composed of **three morphemes**, each carrying a certain amount of meaning:
    - “un” here means opposite of [or not in many other cases]
    - “ness” means being in a state or condition
    - “happy” the familiar word (slightly modified by being combined on the right).
- One classification of morphemes:
  - “happy” is a ***free*** morpheme because it can appear on its own and still mean the same as in the word above.
  - “un” and “ness” are ***bound*** morphemes as they have to be attached to a free morpheme – they can’t mean what they do above when standing on their own.
- But:
  - There is a completely unrelated word “ness”.
  - There is a rather informal word “un” derived from the “un” morpheme, meaning something like “unimportant, characterless, …”
  - “happy” can act as part of a bigger word in other ways, as in “trigger-happy”.

Chinese Word

- Distinct from both morphemes at a lower level and from phrases at a higher level

- Distributional restriction
  - Can occur freely by itself vs. bound morphemes
  - Grammatical morphemes 的、地、得
  - Content morphemes 饮 冷饮 饮食

- Integrity of word meanings
  - Have meanings that are not predictable from the meanings of their component morphemes
  - 大人 vs. 小人 打人 vs. 打手

Disyllabic Chinese Words

- Chinese words was regarded as monosyllabic language
  - Chinese morphemes are one syllable long
  - Multi-syllabic morphemes constitute 11% of the total number of morphemes
  - Only 44% of monosyllabic morphemes can occur by themselves as words.
  - Therefore, monosyllabicity is not true for Chinese words.
- Most Chinese words are disyllabic

Chinese Word Formation – Disyllabic compounds

- Modified noun compounds, over 53%
  - Have descriptive modifiers in front of nouns
  - The meaning of the whole may different from the simple addition of the meanings of the parts
  - 小人 热心 水手
- Modified verb compounds
  - Have modifiers in front of verbs
  - Specifying the manner in which the verbal action is carried out
  - 寄生 飞驰
- Coordinative compounds:27%
  - Synonymous compounds: the two component morphemes have similar or identical meanings
  - Do not seem to be English equivalent
  - 报告 声音

- Antonymous compounds: the component have either opposite or contrary meanings
  - Obvious differences in meanings
  - The parts of speech of the whole can be different
  - 买卖 左右 大小 开关
- Verb-object compounds: 13%
  - can be either verbal or nominal
  - 放心 鼓掌 司机
- Verb-complement compounds (English not)
  - Verbal morpheme is followed by a complement
  - Indicating the direction or the result of the verb
  - 进来 进去

- Subject-predicate compounds
  - A subject and a predicate
  - English: earthquake
  - 地震 心疼 民主
- Noun-measure complement compounds
  - Morphemes do not exhibit any of the usual grammatical relationships
  - Denote a generic kind that the noun is a member of
  - 人口 羊群 书本

Tri-syllabic compounds

- Possibility of hierarchical structure
- In a string of ABC, there can be closer grouping between AB or BC
  - 口香糖 vs. 开玩笑
- The greater number of syllables gives rise to the possibility of ambiguous analyze
  - 大学生
    - 大/学生 大学生为主，小学生去拔草
    - 大学/生 大学生毕业后走上工作岗位

- Modifier-noun
  - Constitute three-fourth of the total number
  - 情人节 大学生
- Verb-object tri-syllabic compounds:
  - Common type of constitute three-fourth of the total
  - 开玩笑 吹牛皮
- Subject-verb-object
  - These compounds with three grammatical parts are naturally only possible with at least three syllables
  - 胆结石 鬼画符
- Descriptive + noun.
  - With their reduplicative or onomatopoetic disyllables preceding the head noun
  - Unique to tri-syllabic form
  - 乒乓球 棒棒糖

Quad-syllabic compounds

- With pairs of synonymous disyllable compounds conjoined
  - 骄傲自满 艰难困苦
- Pairs of synonyms interwoven
  - 花言巧语 油腔滑调
- Pairs of antonyms interwoven
  - 大同小异 口是心非
- Pairs of synonyms and antonyms interwoven
  - 大惊小怪 生离死别
- Pairs of directional morphemes and synonyms
  - 南腔北调 东奔西跑

- Pairs of directional morphemes and antonyms
  - 南来北往
- Pairs of numbers and synonyms interwoven
  - 一干二净 四平八稳
- Pairs of numbers and antonyms interwoven
  - 七上八下
- Pairs of interwoven with synonyms
  - 全心全意 称王称霸
- Reduplication interwoven with antonyms
  - 自生自灭 不破不立
- Reduplication interwoven with numbers
  - 不三不四

Chinese Word Formation - Affixation

- Words can be formed by adding affixes to the root
- prefix
  - 第一/第二 老婆/老公
- Suffix
  - Noun ending
    - 儿子/桌子 物理学/化学
  - Verb ending
    - 吃了/吃着/吃过

Chinese Word Formation - Reduplication

- Either monosyllabic or disyllabic root morphemes can be copied
- AABB vs. ABAB vs. A里AB
- Not Applicable to English
- Verbal
  - 商量 -> 商量商量 高兴 -> 高兴高兴
- adjectival
  - 漂亮 -> 漂漂亮亮 高兴 -> 高高兴兴
  - 流气 -> 流里流气 糊涂 -> 糊里糊涂
  - Nominal
    - 人 -> 人人 个 -> 个个

Chinese Word Formation - Ionization

- A morphs-syntactic phenomenon

- The process by which component morphemes of words are separated from each other
  - 理发 ‘cut-hair’
  - 理短发 ‘have short haircut’
  - 理一个发 ‘have a haircut’
  - 发理了吗 ‘hair has been cut?’

Part-of-Speech (POS) Defined in English

Tag set	Basic size	Total tags
Brown	87	179
Penn	45	
CLAWS1	132	
CLAWS2	166	
CLAWS c5	62	
London-Lund	197	

Sizes of various tag sets.

Category	Examples	Claws c5	Brown	Penn
Adjective	happy, bad	AJ0	JJ	JJ
Adjective, ordinal number	sixth, 72nd, last	ORD	OD	JJ
Adjective, comparative	happier, worse	AJC	JJR	JJR
Adjective, superlative	happiest, worst	AJS	JJT	JJS
Adjective, superlative, semantically	chief, top	AJ0	JJS	JJ
Adjective, cardinal number	3, fifteen	CRD	CD	CD
Adjective, cardinal number, one	one	PNI	CD	CD
Adverb	often, particularly	AV0	RB	RB
Adverb, negative	not, n t	XX0	*	RB
Adverb, comparative	faster	AV0	RBR	RBR
Adverb, superlative	fastest	AV0	RBT	RBS
Adverb, particle	up, off, out	AVP	RP	RP
Adverb, question	when, how, why	AVQ	WRB	WRB
Adverb, degree & question	how, however	AVQ	WQL	WRB
Adverb, degree	very, so, too	AV0	QL	RB
Adverb, degree, postposed	enough, indeed	AV0	QLP	RB
Adverb, nominal	here, there, now	AV0	RN	RB
Conjunction, coordination	and, or	CJC	c c	c c
Conjunction, subordinating	although, when	CJS	c s	IN
Conjunction, complementizer *that*	that	CJT	c s	IN

Part-of-Speech (POS) Defined in Chinese

ag	adjective morpheme	绿色/n似/d锦/ag	a	adjective	重要/a 步伐/n
ad	adverb-adjective	积极/ad 谋求/v	an	adnoun	克服/v 困难/an
bg	distinguish morpheme	一个/m 次/bg 地区/n	b	distinguish word	女/b 司机/n
c	conjunction	合作/vn 与/c 伙伴/n	dg	adverb morpheme	了解/v 甚/dg 深/a
d	adverb	进一步/d 发展/v	e	exclamation	啊/e
f	position word	贵州/ns 南部/f	h	heading element	非/h 主角/n
i	idiom	一言一行/i	j	abbreviation	德/j 外长/n
k	tail element	朋友/n 们/k	l	habitual word	落到实处/l
mg	numeral morpheme	让/v 乙/mg 背上/v	m	numeral	三/m 个/q
ng	noun morpheme	出/v 两/m 天/q 差/ng	n	noun	科技/n 文献/n
nr	person's name	朴/nr 贞爱/nr	ns	toponym	安徽/ns
nt	organization proper noun	联合国/nt	nx	foreign character	24/m K/nx
nz	other proper noun	满族/nz	o	onomatopoeia	哈哈/o 笑/v
p	preposition	对/p 子孙/n 负责/v	q	quantifier	首/m 批/q
rg	pronoun morpheme	成长/v 于/p 斯/rg	r	pronoun	本/r 地区/n
s	location word	西部/s 交通/n	tg	time morpheme	3 日/t 晚/tg
t	time	下午/t 2时/t	u	auxiliary	填平/v 了/u
vg	verb morpheme	洗/v 了/u 澡/vg	v	verb	编辑/v文献/n
vd	adverb-verb	持续/vd 好转/v	vn	gerund	收费/vn 电话/n
w	punctuation	"/w	yg	modal morpheme	致/v 之/u 耳/Yg
y	modal word	又/d 何在/v 呢/y	z	state word	短短/z 几/m 年

Challenges in Chinese Processing-Few formal morphological markings

- No verbal inflections:
  - No tense: a verb will have the same form
    - 我过去 是 学生。 I **was** a student.
    - 我现在 是 学生。 I **am** a student.
    - 我将来 是 学生。 I **will be** student.
  - No personal and number agreements
    - 我 去。 I **go**.
    - 她 去。 She **goes**.

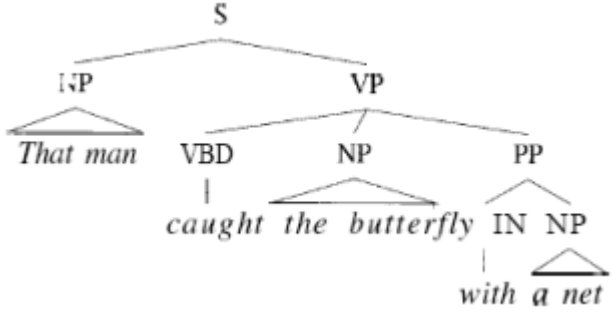
- No nominal endings
  - No number marking
    - 我的书 My book(s)
  - No gender marking
  - No case marking
    - **I** love **her**. Vs. **She** loves **me**.
    - (I=subject case; me = object case; she=subject case; her=object case)
    - 我爱她 她爱我。
    - (我=both subject and object case; 她=both subject and object case)

Challenges in Chinese Processing-Ambiguities in Words

- Lexical Ambiguities:
  - 他很好吃
  - 炸鸡很好吃
- Structural Ambiguities
  - Overlapping (crossing) ambiguity 交集型歧义
    - 网球场 美国会
  - Combinatorial ambiguity 组合型歧义
    - 才能 学生会
  - Mixed type 混合型歧义
    - 太平淡 (too dull), 太平 (peaceful), 平淡

Phrase

- Words are organized into phrases, then comes to sentence



- Syntax studies the regularities and constraints of word order and phrase structure

- Major phrase categories in English
  - Noun phrase. Prepositional phrases
  - Verb phrases. Adjective phrases

English Phrase

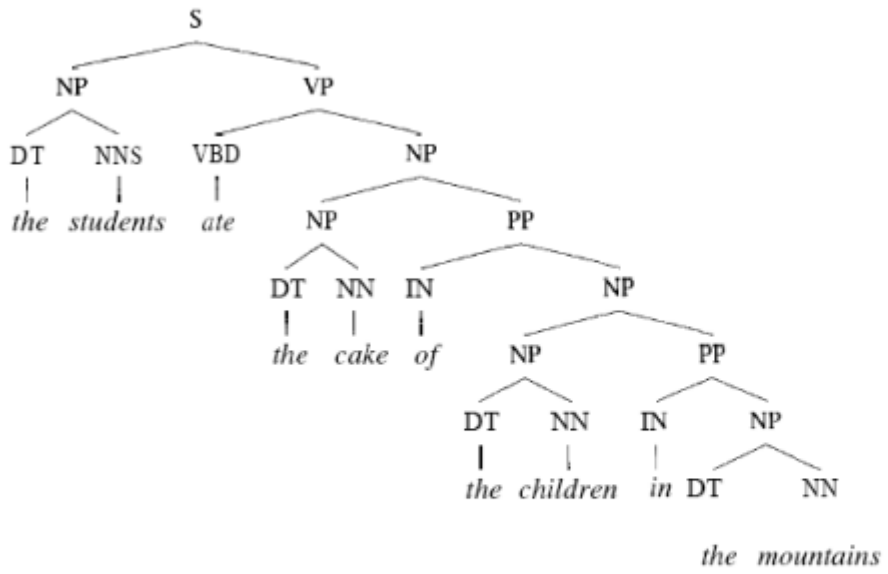
Category	Description	Examples
Noun Phrase (NP)	A noun and all its modifiers	**The bewildered tourist** was lost.
Verb Phrase (VP)	A verb and all its modifiers	He **was waiting for the rain to stop**.
Gerund Phrase (GP)	A noun phrase that starts with a gerund.	**Taking my dog for a walk** is fun.
Infinitive Phrase (IP)	A noun phrase that begins with an infinitive verb.	**To make lemonade**, you have **to start with lemons**.
Appositive Phrase (AP)	It restates and defines a noun. It consists of one or more words.	**My favorite pastime**, needlepoint, surprises some people.
Participial Phrase (PP)	Begins with a past or present participle.	**Washed with my clothes**, my cell phone no longer worked.
Absolute Phrase	It modifies the whole sentence, not just a noun.	**His tail between his legs**, the dog walked out the door.

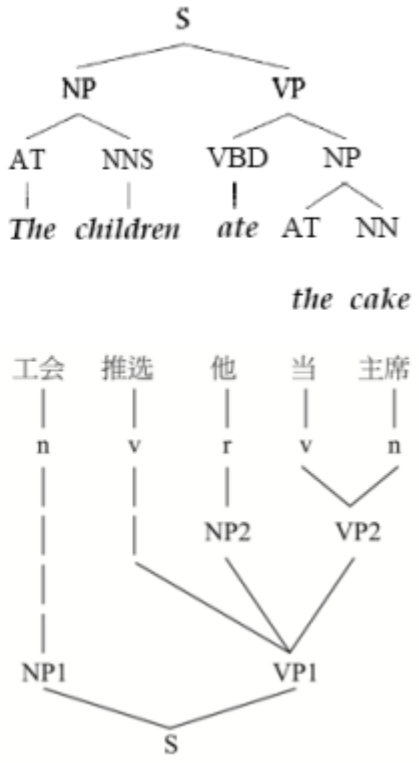
Chinese Phrase

Category	Description	Example
BNP	Base noun phrase	[市场/n 经济/n]NP *market economy*
BAP	Base adjective phrase	[公正/a合理/a]BAP *fair and reasonable*
BVP	Base verb phrase	[顺利/a启动/v]BVP *successfully start*
BDP	Base adverb phrase	[已/d 不再/d]BDP *no longer*
BQP	Base quantifier phrase	[数千/m名/q]BQP 士兵/n *several thousand soldiers*
BTP	Base time phrase	[早上/t 8 时/t]BTP *8:00 in the morning*
BFP	Base position phrase	[内蒙古/ns东北部/f]BFP *North-east of Inner Mongolia*
BNT	Name of an organization	[烟台/ns 大学/n]BNT *Yantai University*
BNS	Name of a place	[江苏省/ns铜山县/ns]BNS *Jiangsu Province, Tongshan Country*
BNZ	Other proper noun phrase	[诺贝尔/nr奖/n]BNZ *The Nobel Prize*
BSV	S-V structure	[领土/n 完整/a]BSV *territorial integrity*

Phrase Structure Grammar

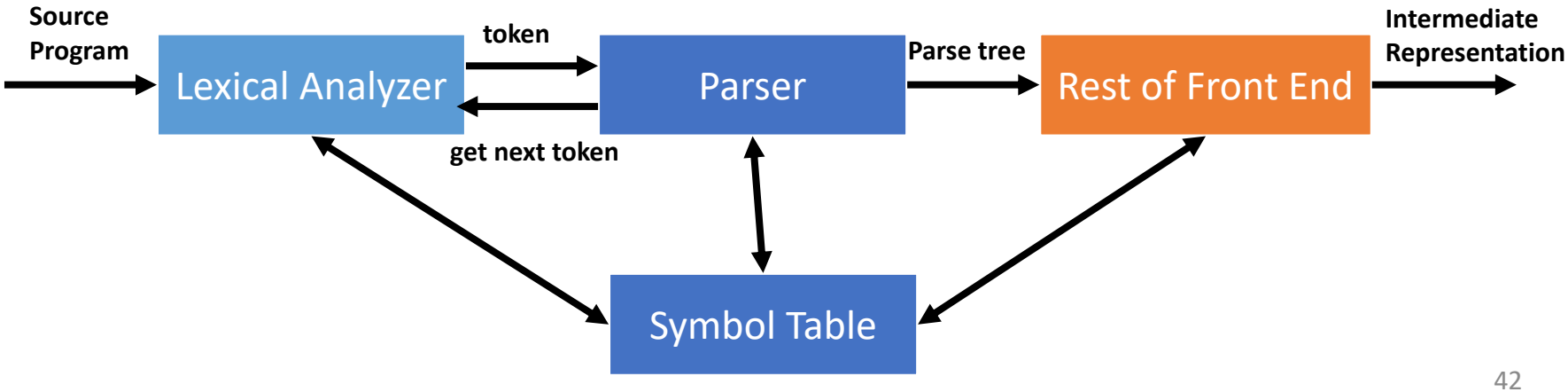
- Tells us how to determine the meaning of the sentence from the meaning of the words





Parsing (Syntax Analysis)

- Syntax
  - Is one of the major components of **grammar**.
  - Is the proper order of words in a phrase or sentence.
  - Is a tool used in writing proper grammatical sentences
  - Native speakers of a language learn correct syntax without realizing it.
  - The complexity of a writer's or speaker's sentences creates a formal or informal level of diction that is presented to its audience.



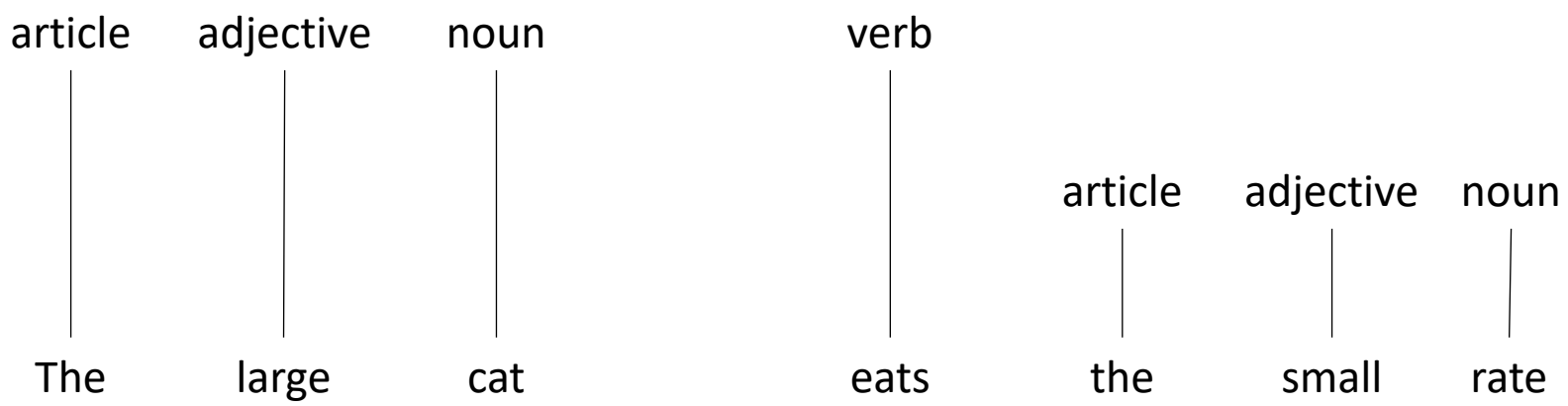
Parse tree (Syntactic Tree)

"The large cat eats the small rat"

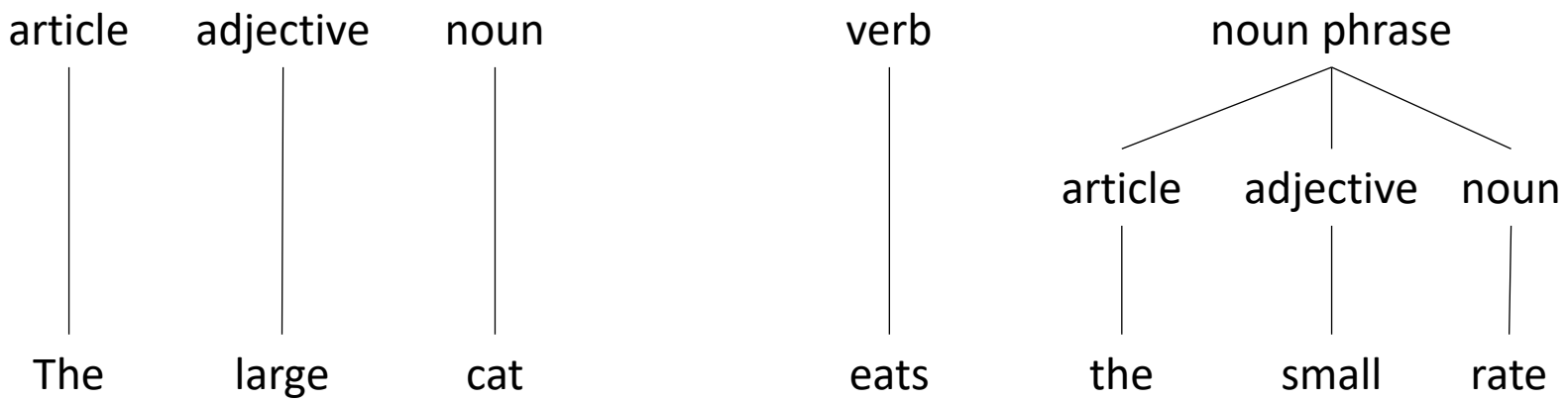
Parse tree (Syntactic Tree)

The large cat eats the small rate

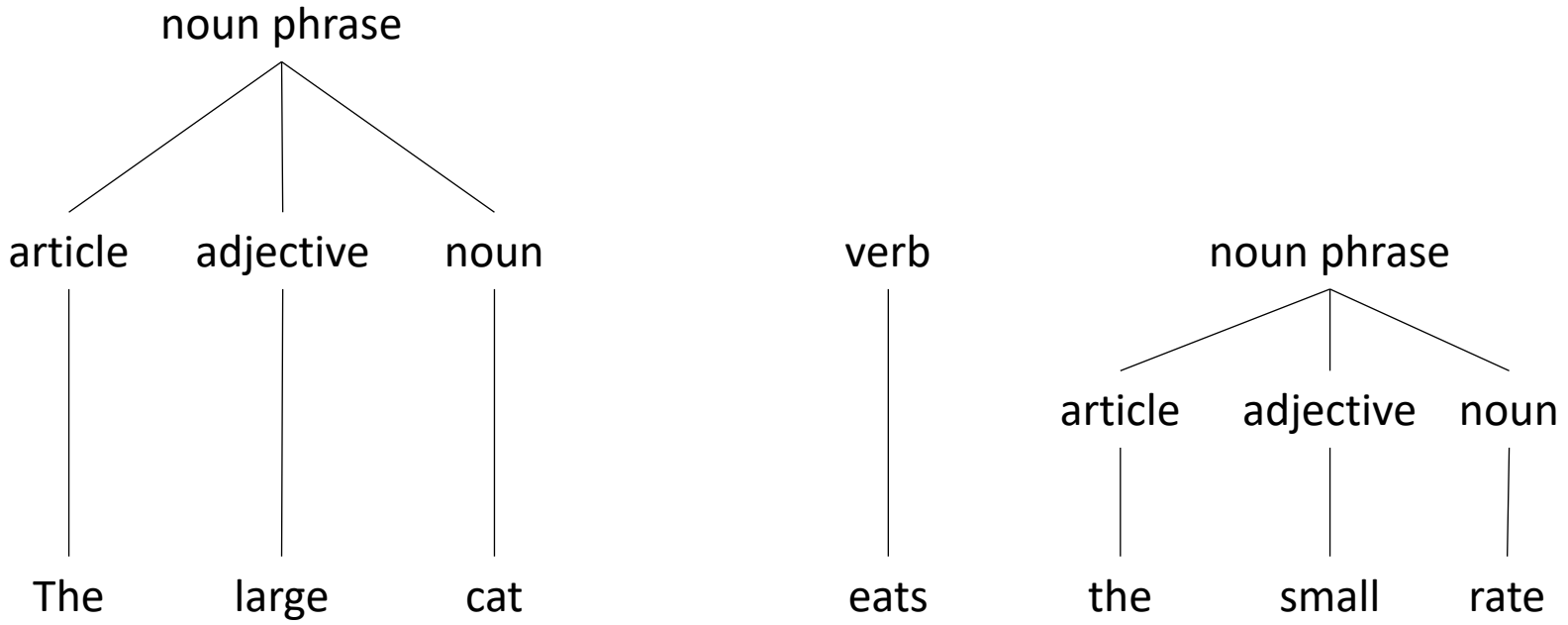
Parse tree (Syntactic Tree)



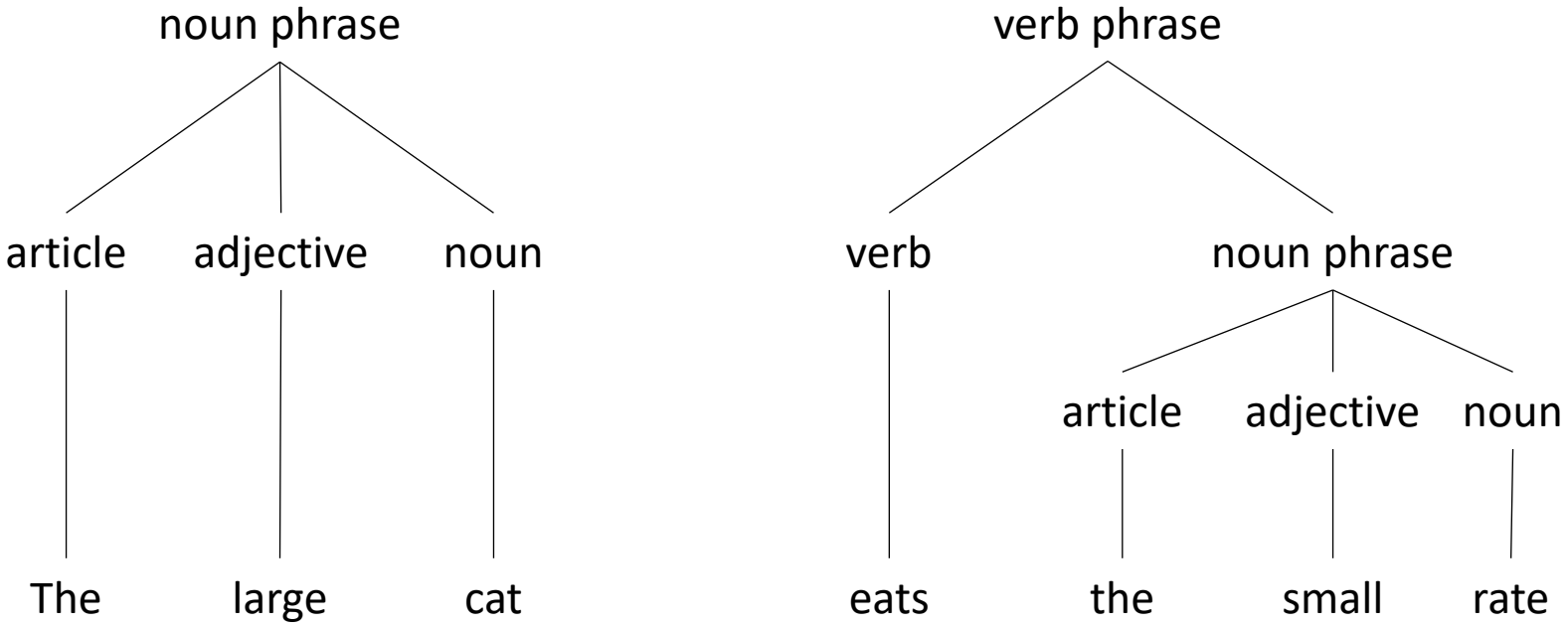
Parse tree (Syntactic Tree)



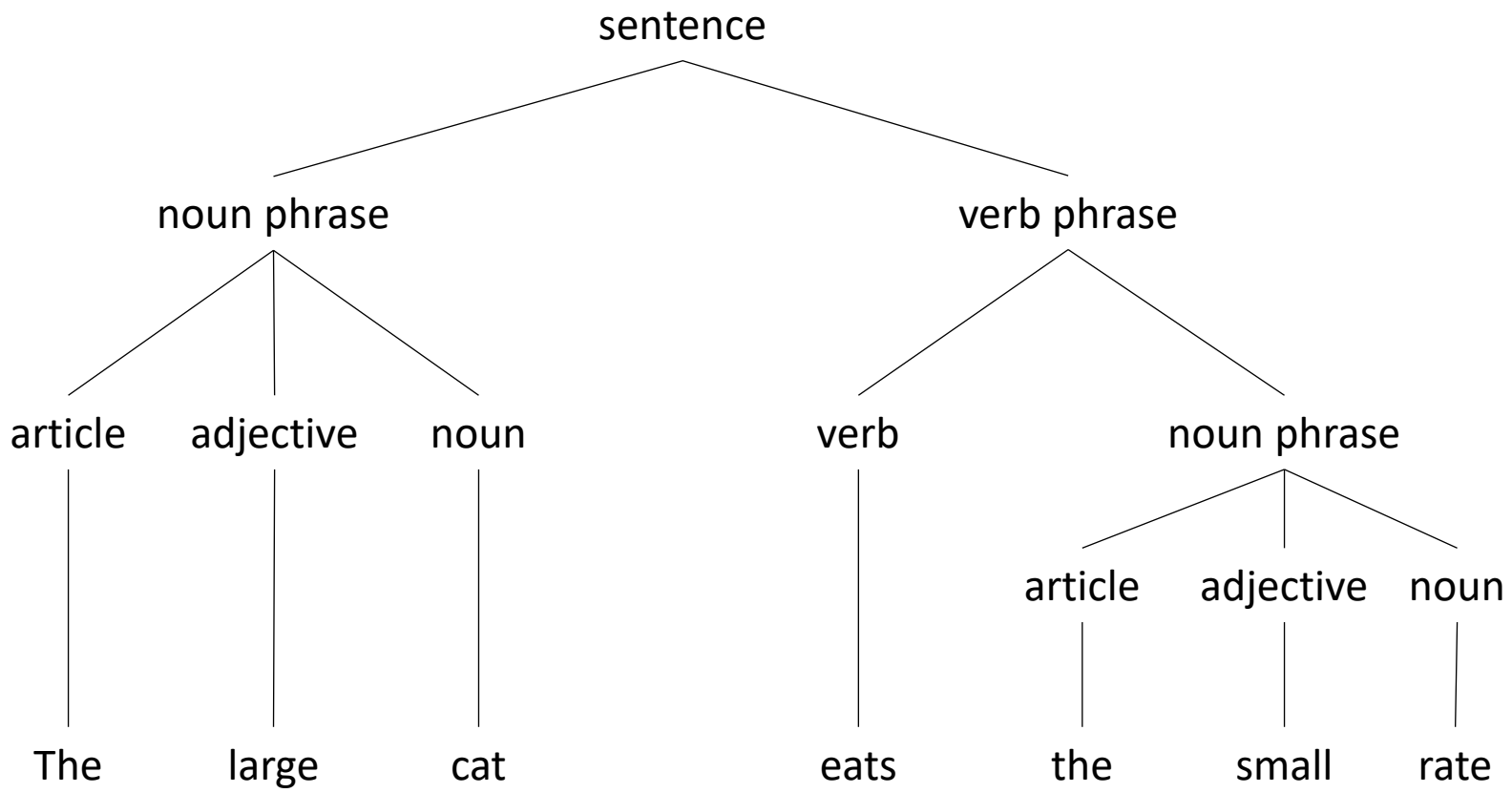
Parse tree (Syntactic Tree)



Parse tree (Syntactic Tree)

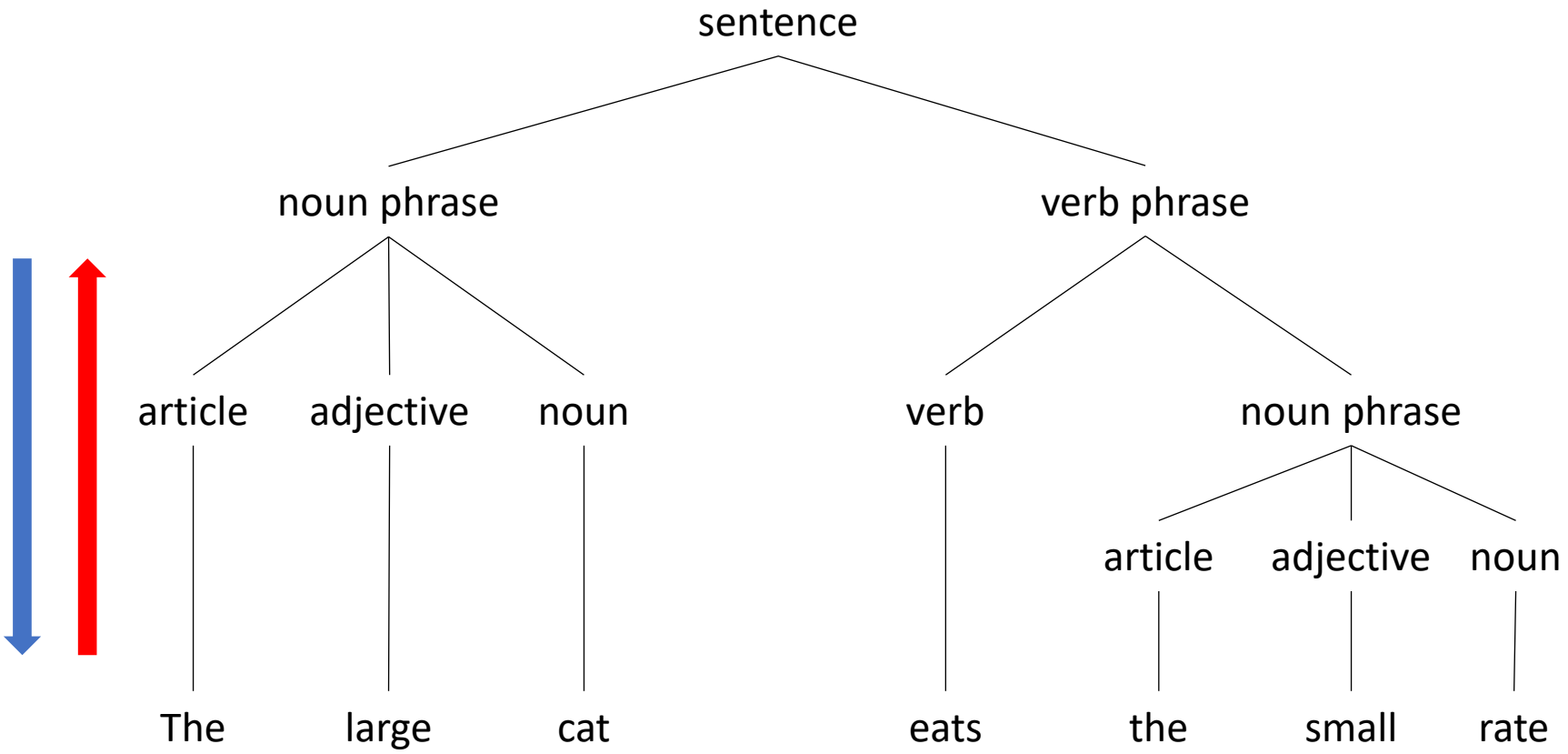


Parse tree (Syntactic Tree)



Parse tree (Syntactic Tree)

- Bottom-up or Top-down

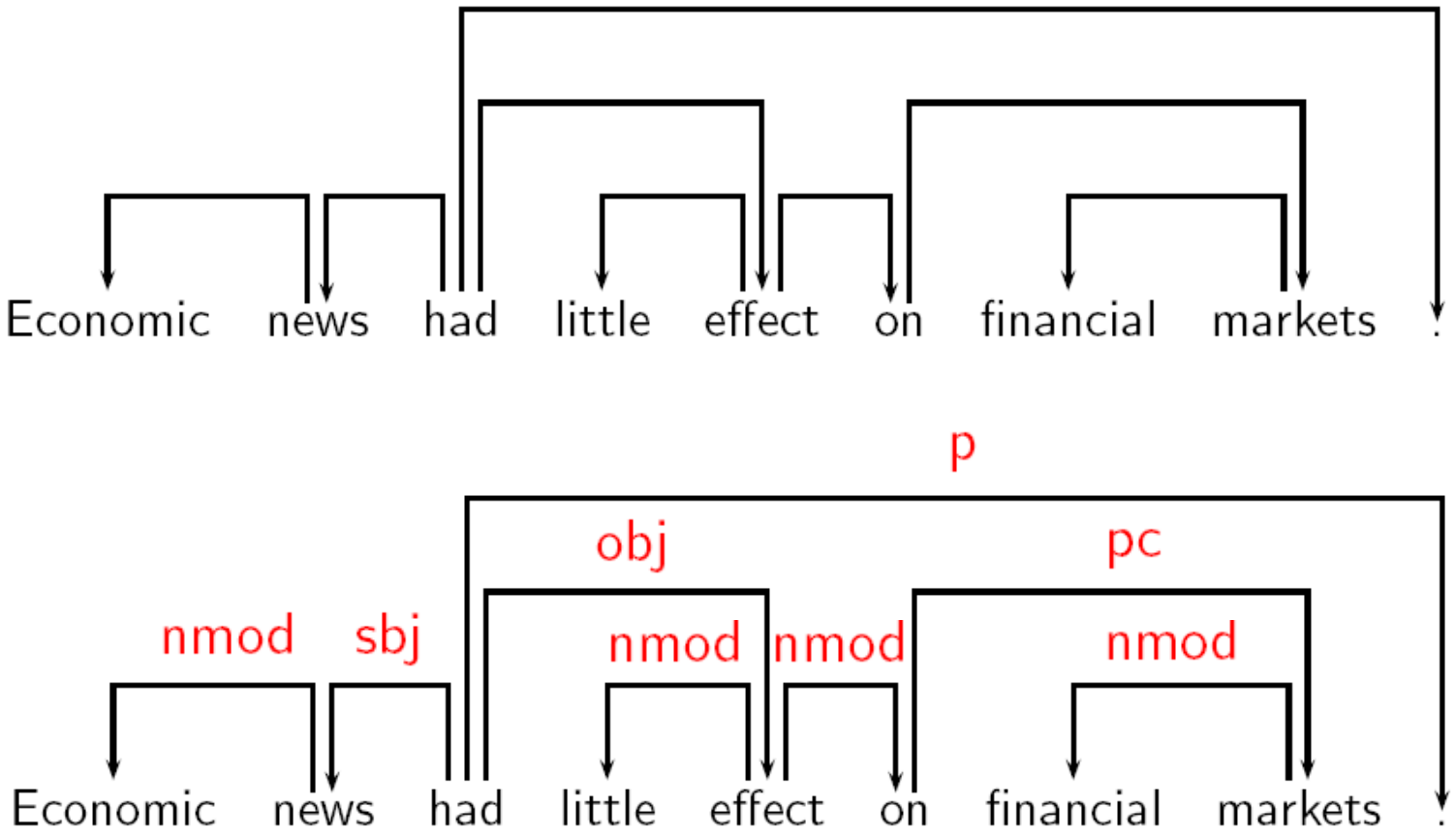


Dependency

- Syntactic structure consists of lexical items, linked by binary asymmetric relations called dependencies. [Tesniere 1959]

> The sentence is an *organized whole*, the constituent elements of which are *words*. Every word that belongs to a sentence ceases by itself to be isolated as in the dictionary. Between the word and its neighbors, the mind perceives *connections*, the totality of which forms the structure of the sentence. The structural connections establish *dependency* relations between the words. Each connection in principle unites a *superior* term and an *inferior* term. The superior term receives the name *governor*. The inferior term receives the name *subordinate*.

Dependency Structure



Phrase Structure vs. Dependency Structure

	Phrase structure	Dependency structure
word relation	phrasal constitution	head-dependent
categories	syntactic functional	syntactic structural
new node (Y/N)	multiple nodes per word	one node per word
operation	waiting for complete phrase	word-at-a-time

Problems in Parsing - Ambiguity

- "One morning, I shot an elephant in my pajamas. How he got into my pajamas, I don't know."
- Syntactical/Structural ambiguity
  - Several parse trees are possible e.g. above sentence
- Semantic/Lexical ambiguity
  - Several word meanings. e.g. bank (银行) or bank (河流)
- Different word categories
  - "He **books** the flight" vs. "The **books** are here."
  - "Fruit flies from the balcony" vs. "Fruit flies are on the balcony"
- Attachment
  - In particular PP (prepositional phrase) binding; often referred to as "binding problem"
  - "One morning, I shot an elephant **in my pajamas**."

Discourse Processing 篇章/语篇

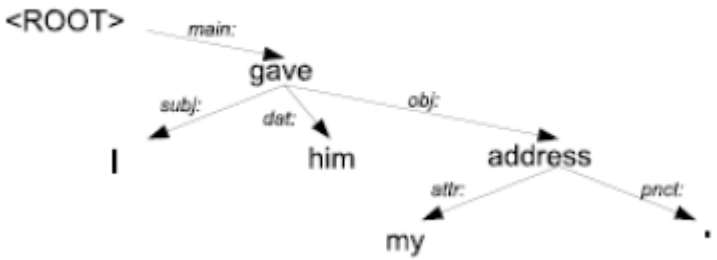
- **Discourse** is a group of collocated and coherent sentences
- **Discourse theory** deals with language phenomena that operate beyond the single sentence
- **Discourse analysis/processing** is a suite of **Natural Language Processing (NLP) tasks** to uncover linguistic structures from multi-sentential texts at several levels, which can support many "downstream" NLP applications.

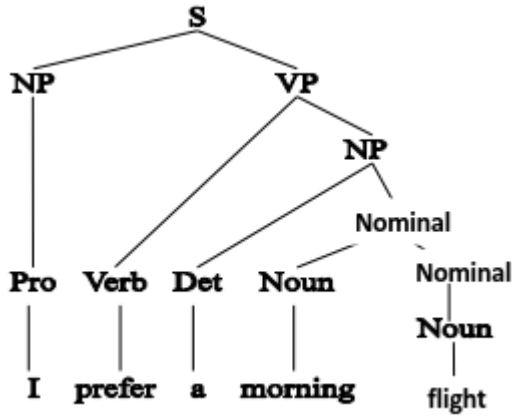
- Coherence structure
- Conversation structure
- Co-reference structure
- Topic structure

- Text summarization
- Essay scoring
- Sentiment analysis
- Machine translation
- Information extraction
- Question answering
- Thread recovery

Modeling Coherence Structure

- **Grammaticality:** distinguishes well-structured sentences from random sequences of words. Models are specified as groupings and relations between words





- **Coherence** plays the same role at the multi-sentence level. Models are also specified as groupings and relations between …?

Discourse/Coherence Relations

- Specify the relations between **sentences** or **clauses**.
- Due to the relations, two adjacent sentences can look coherent.

What is the discourse relation between the following two sentences?

John hid Bill's car keys. He was drunk

"Explanation" relation

vs.

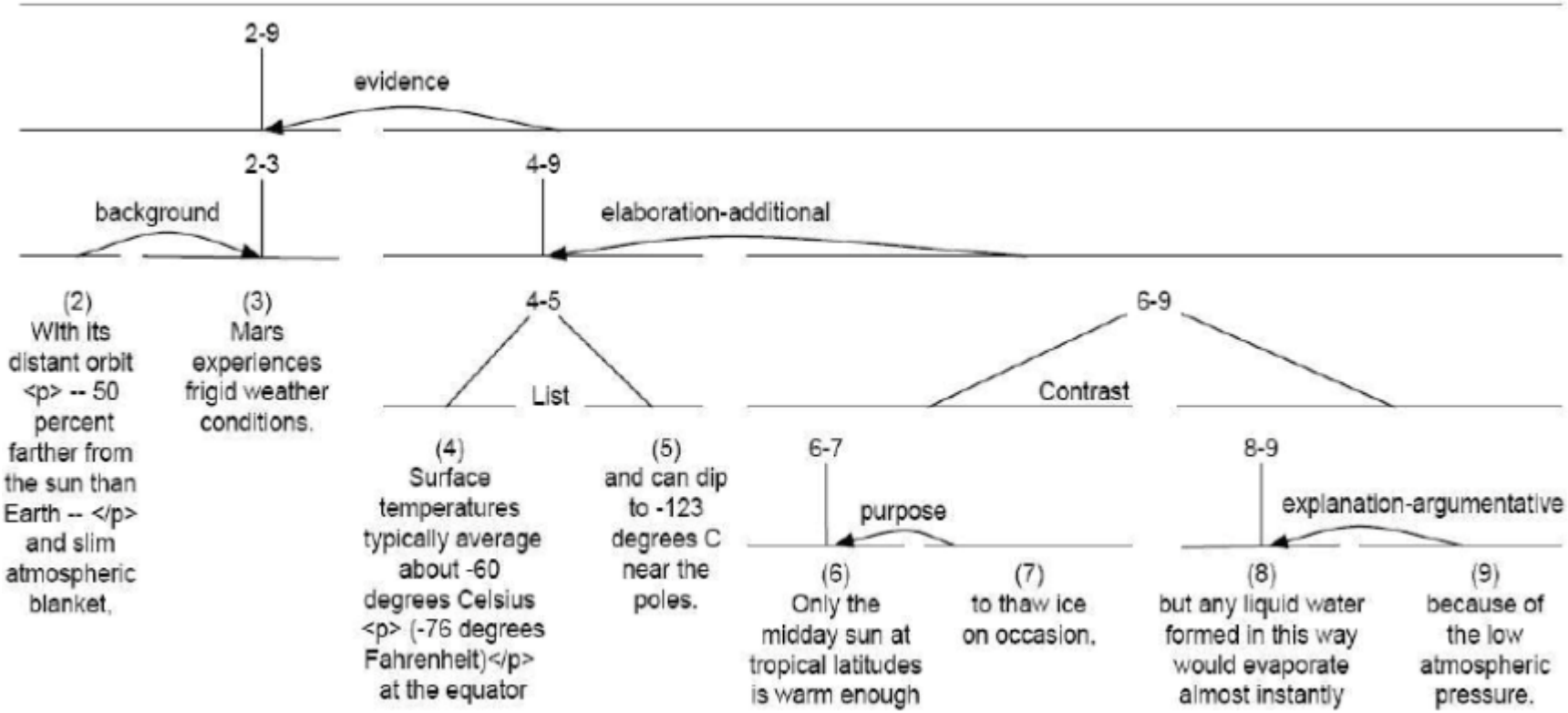
John hid Bill's car keys. He likes spinach

More Discourse Relations

- Elaboration 细化
  - Dorothy was from Kansas. She lived on the Kansas prairies.
- Result 因果
  - The tin woodman was caught in the rain. His joints rusted.
- Parallel 并列
  - The scarecrow wanted some brains. The tin woodsman wanted a heart.

- How many relations? Which ones?
- What kind of structures? Flat? Trees? Graphs?

RST Discourse Parse (Tree Structure)



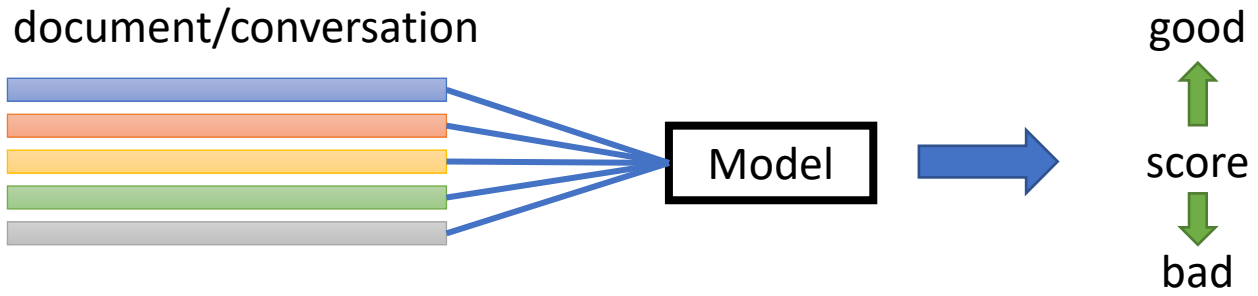
Mann W C, Thompson S A. Rhetorical structure theory: A theory of text organization[M]. Los Angeles: University of Southern California, Information Sciences Institute, 1987.

PDTB Discourse Parse (Flat Structure)

- Either two clauses / sentences connected by an explicit connective
  - The federal government suspended sales of U.S. savings bonds because Congress hasn't lifted the ceiling on government debt.
    cause-reason
  - The subject will be written into the plots of prime-time shows, and viewers will be given a 900 number to call. Conjunction
- Or two adjacent sentences connected by an implicit connective
  - Some have raised their cash positions to record levels. High cash positions help buffer a fund when the market falls.
    Implicit=because(cause-reason)

Webber B L, Joshi A K. Anchoring a lexicalized tree-adjoining grammar for discourse[J]. arXiv preprint cmp-lg/9806017, 1998.

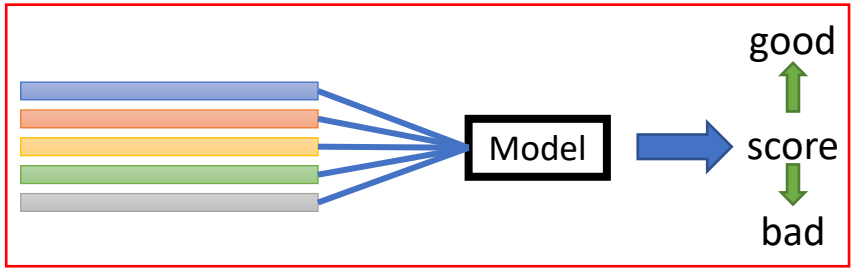
Coherence Models



- Helps predict which sentences are pragmatically appropriate
- Tells us which sentences are closely related

- Applications
  - Essay scoring
  - Summarization (sentence selection and ordering)
  - Generation (including MT)
  - …

Coherence Models – Outline



Evaluation tasks

Applications

- Entity-based models
- Models based on discourse relations
- Graph-based models
- Syntax-based models
- Neural models

- Discrimination
- Insertion

- Essay Scoring
- Summary Coherence Rating
- MT

Corpus

- A **Corpus** (plural corpora) is a large and structured set of texts (nowadays usually electronically stored and processed).
- Used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory.
- A corpus may contain texts in a single language (**monolingual corpus**) or text data in multiple languages (**multilingual corpus**).

Corpus

- Multilingual corpora that have been specially formatted for side-by-side comparison are called Aligned Parallel Corpora.
- Some corpora have further *structured* levels of analysis applied. In particular, a number of smaller corpora may be fully parsed. Such corpora are usually called Treebank
- More corpora
  - Domain specific
  - Sentiment

Famous Corpora

- The Brown Corpus of Standard American English
  - The first of the modern, computer readable, general corpora. The corpus consist of one million words of American English texts printed in 1961
  - Consists of 500 samples, distributed across 15 genres in rough proportion to the amount published in 1961 in each of those genres.

- The London-Lund Corpus of Spoken English
  - Derives from two projects:
    - The Survey of English Usage at University College London
    - The Survey of Spoken English started at Lund University in 1975
  - Consists of 100 spoken texts of 5000 words each (500, 000 in total)

Corpus Linguistics

- Corpus linguistics is the study of language as expressed in samples (corpora) of “real world” text.
- This method represents a digestive approach to deriving a set of abstract rules by which a natural language is governed or else relates to another languages.
- Originally done by hand, corpora are now largely derived by an automated process.

Corpus Linguistics

- Landmark in modern corpus linguistics
  - Henry Kucera and W. Nelson Francis of Computational Analysis of Present-Day American English in 1967
  - A work based on the analysis of the Brown Corpus
- American Heritage Dictionary
  - The first dictionary to be compiled using corpus linguistics
- Collins' COBUILD monolingual learner's dictionary
  - Compiled using the Bank of English.

Corpus Linguistics

- Wallis and Nelson (2001) 3A perspective: Annotation, Abstraction and Analysis
  - Annotation: the application of a scheme to texts
  - Abstraction consists of the translation (mapping) of terms in the scheme to terms in a theoretically motivated model or dataset.
  - Analysis consists of statistically probing, manipulating and generalizing from the dataset.
  - Analysis might include statistical evaluations, optimization of rule-bases or knowledge discovery methods.

The next lecture

- Lecture 3
  Machine Learning Foundations