Texts, Treebanks, Networks descriptions and analysis of languages

XINYING CHEN

University of Ostrava, Nov. 8, 2017

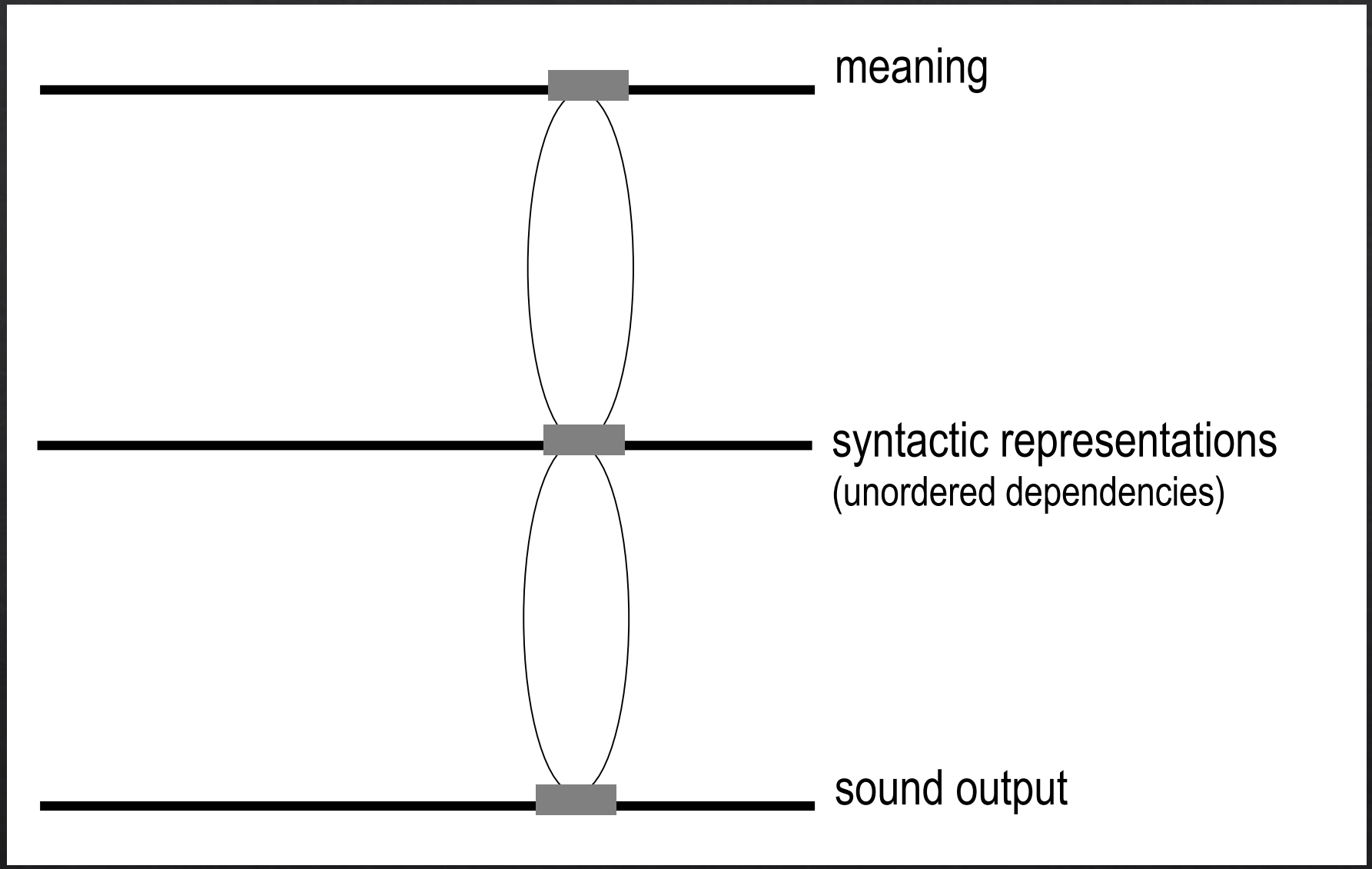
Outline

- Why different language descriptions and analysis
- Texts
- Treebanks
- Networks
- Questions & comments

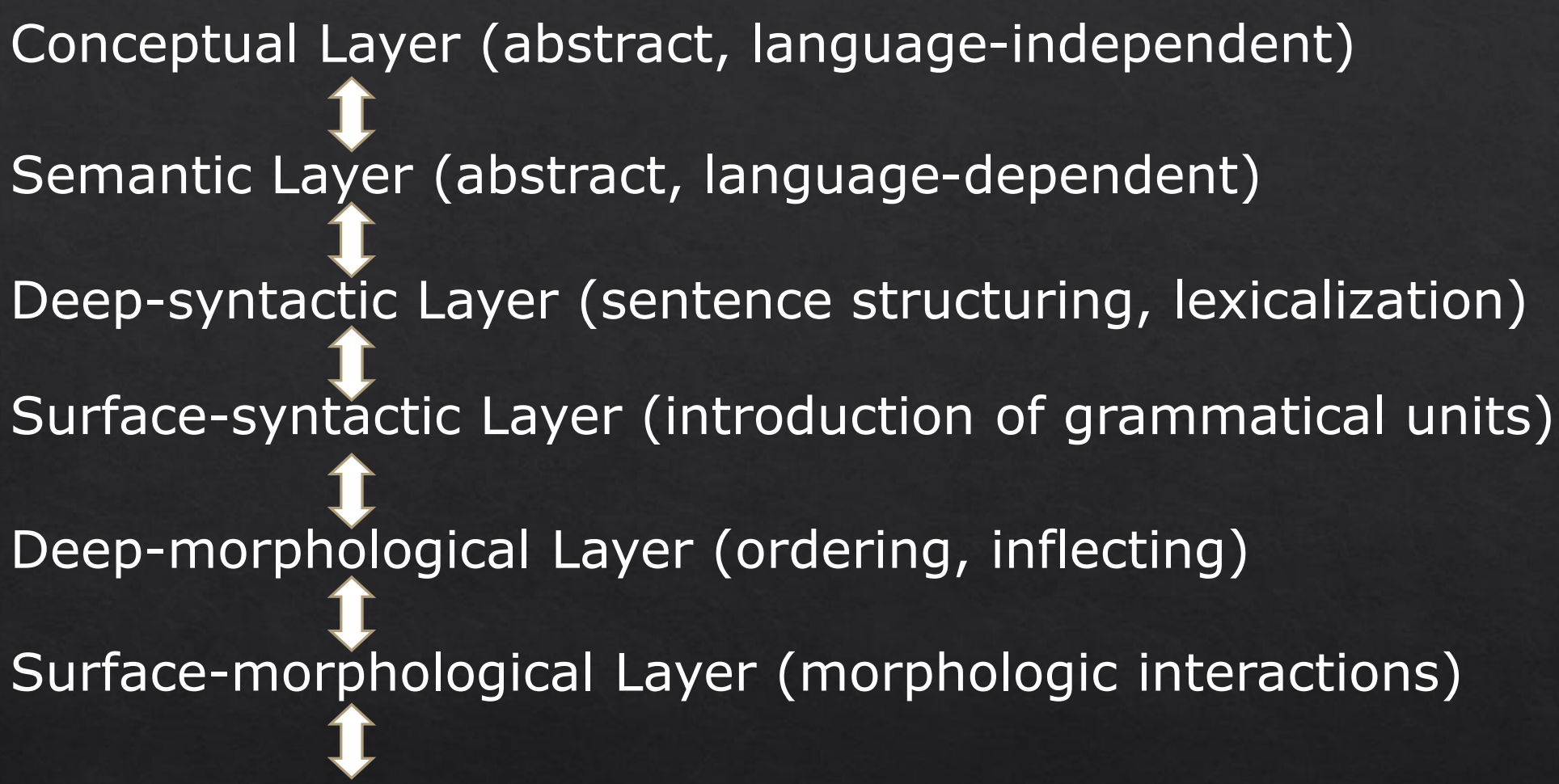
Outline

- Why different language descriptions and analysis

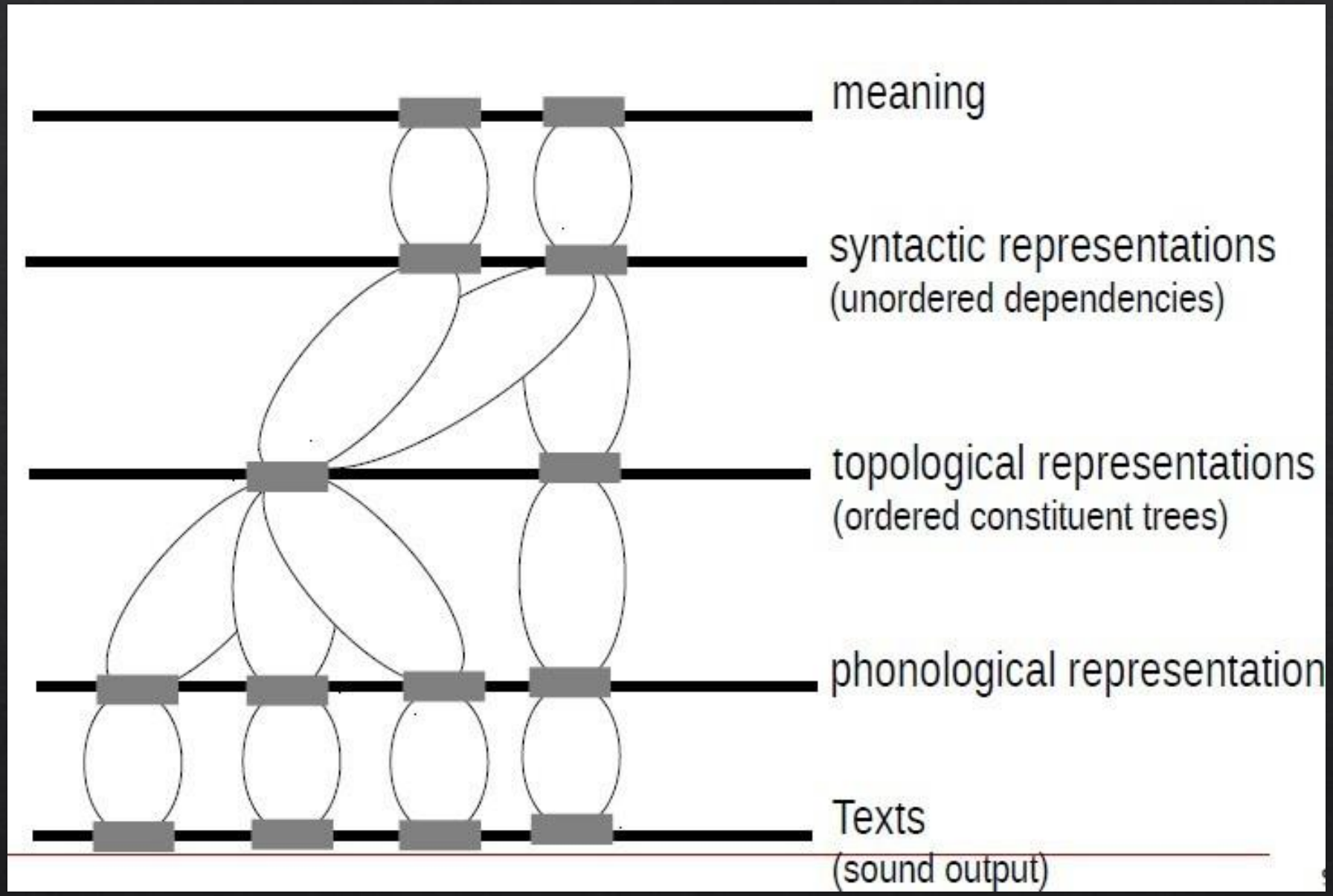
language is a system of signs, each of which is an arbitrary union of sound and meaning… Any given sign, is defined by its relationships with the others. (Saussure 1959)



Saussure’s signified and signifier (2011)
Mel’čuk’s meaning and text (1981)
Chomsky’s logical and phonetic structure (2002)


Conceptual Layer (abstract, language-independent)
Semantic Layer (abstract, language-dependent)
Deep-syntactic Layer (sentence structuring, lexicalization)
Surface-syntactic Layer (introduction of grammatical units)
Deep-morphological Layer (ordering, inflecting)
Surface-morphological Layer (morphologic interactions)
Text (full-fledged sentences)

Mel’čuk 1986



surface-deep hierarchy Mel’čuk 1986

Language understanding and generation require deconstructions or constructions of units on different language layers. Different language descriptions, therefore, yield to this nature of language communications.

Outline

- Why different language descriptions and analysis
- Texts

Texts

- You may have heard
  - word, token, type, lemma, phrase, ngram, sentence, paragraph…
  - word length, frequency, size…
  - genre, style, authorship…

Texts

- You may have heard
  - word, token, type, lemma, phrase, ngram, sentence, paragraph…
  - word length, frequency, size…
  - genre, style, authorship…
- Advantages
  - Relatively objective (word vs phrase…)
  - More available (OCR, Google books…)
  - Easier to test hypothesis (Zip’f law…)

Texts

- You may have heard
  - word, token, type, lemma, phrase, ngram, sentence, paragraph…
  - word length, frequency, size…
  - genre, style, authorship…
- Advantages
  - Relatively objective (word vs phrase…)
  - More available (OCR, Google books…)
  - Easier to test hypothesis (Zip’f law…)
- Disadvantages
  - One dimensional description (linear modals)
  - Missing details of language understanding or generation

Structural Complexity of Chinese Characters and Zipf’s Law

Structural Complexity of Chinese Characters and Zipf's Law

- Hypothesis
  - The relationship between the frequency of Chinese characters and structural complexity of characters should be in accordance with Zipf's law due to the principle of least effort

Structural Complexity of Chinese Characters and Zipf's Law

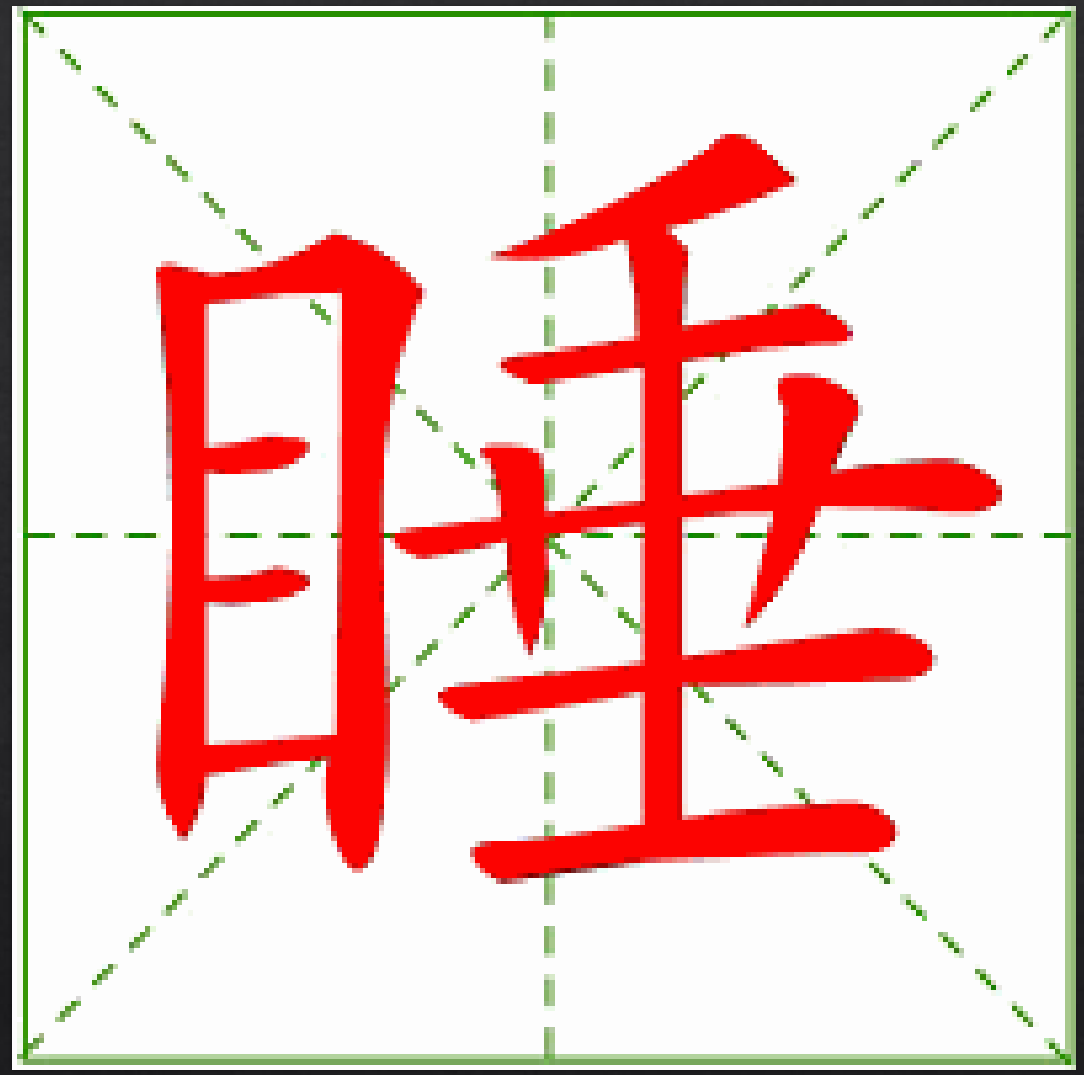
- Hypothesis
  - The relationship between the frequency of Chinese characters and structural complexity of characters should be in accordance with Zipf's law due to the principle of least effort

Structural Complexity of Chinese Characters and Zipf's Law

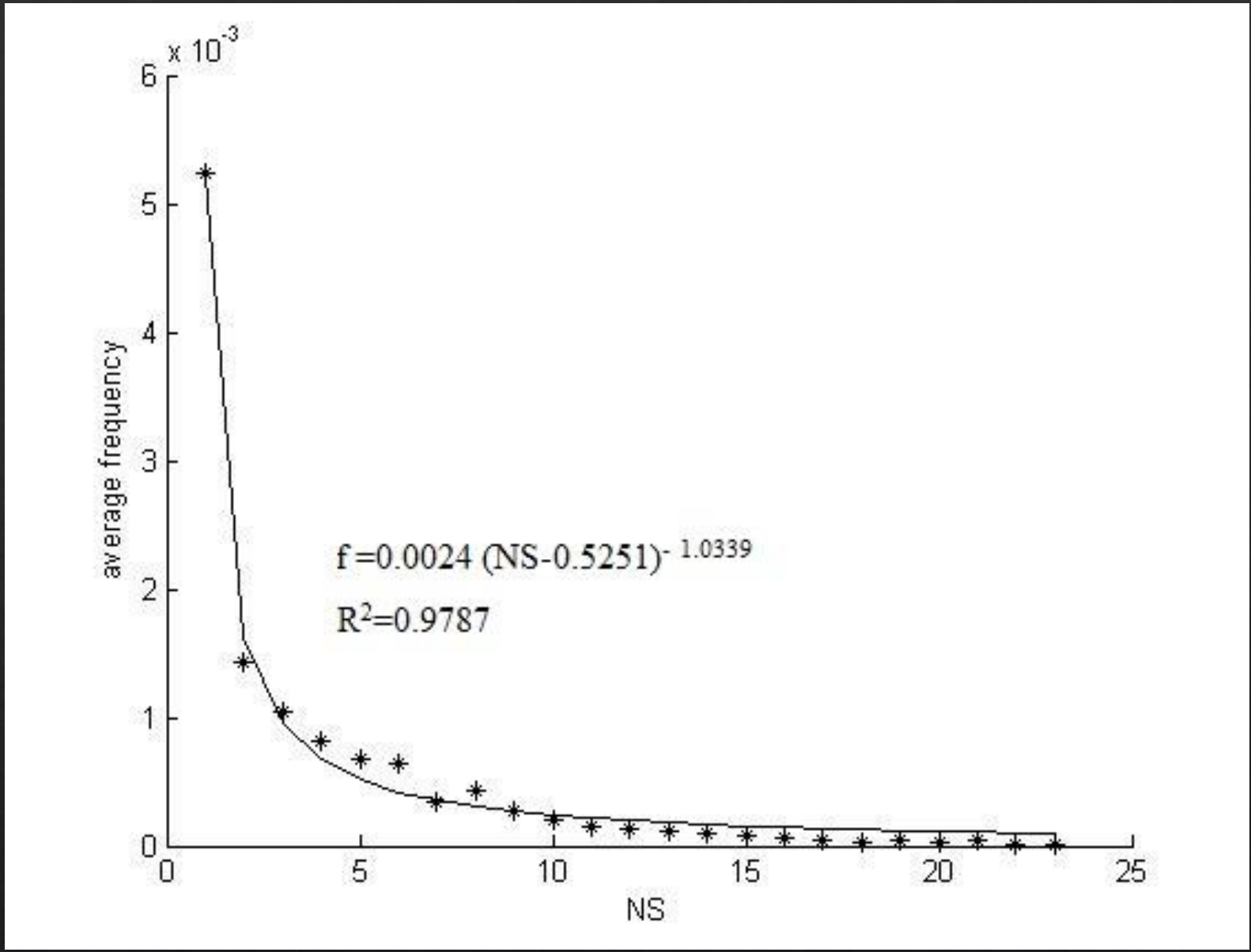
- Hypothesis
  - The relationship between the frequency of Chinese characters and structural complexity of characters should be in accordance with Zipf's law due to the principle of least effort
- Frequency
  - Chinese Characters' Frequency Dictionary
    - the most frequent 3,061 different Chinese characters and their frequency in People's Daily, a famous newspaper in China, Corpus. (99.43%)

Structural Complexity of Chinese Characters and Zipf's Law

- Hypothesis
  - The relationship between the frequency of Chinese characters and structural complexity of characters should be in accordance with Zipf's law due to the principle of least effort
- Frequency
  - Chinese Characters' Frequency Dictionary
    - the most frequent 3,061 different Chinese characters and their frequency in People's Daily, a famous newspaper in China, Corpus. (99.43%)
- Structural Complexity
  - Number of strokes
  - Number of components
    - Both counted based on dictionaries and issued formal documents.

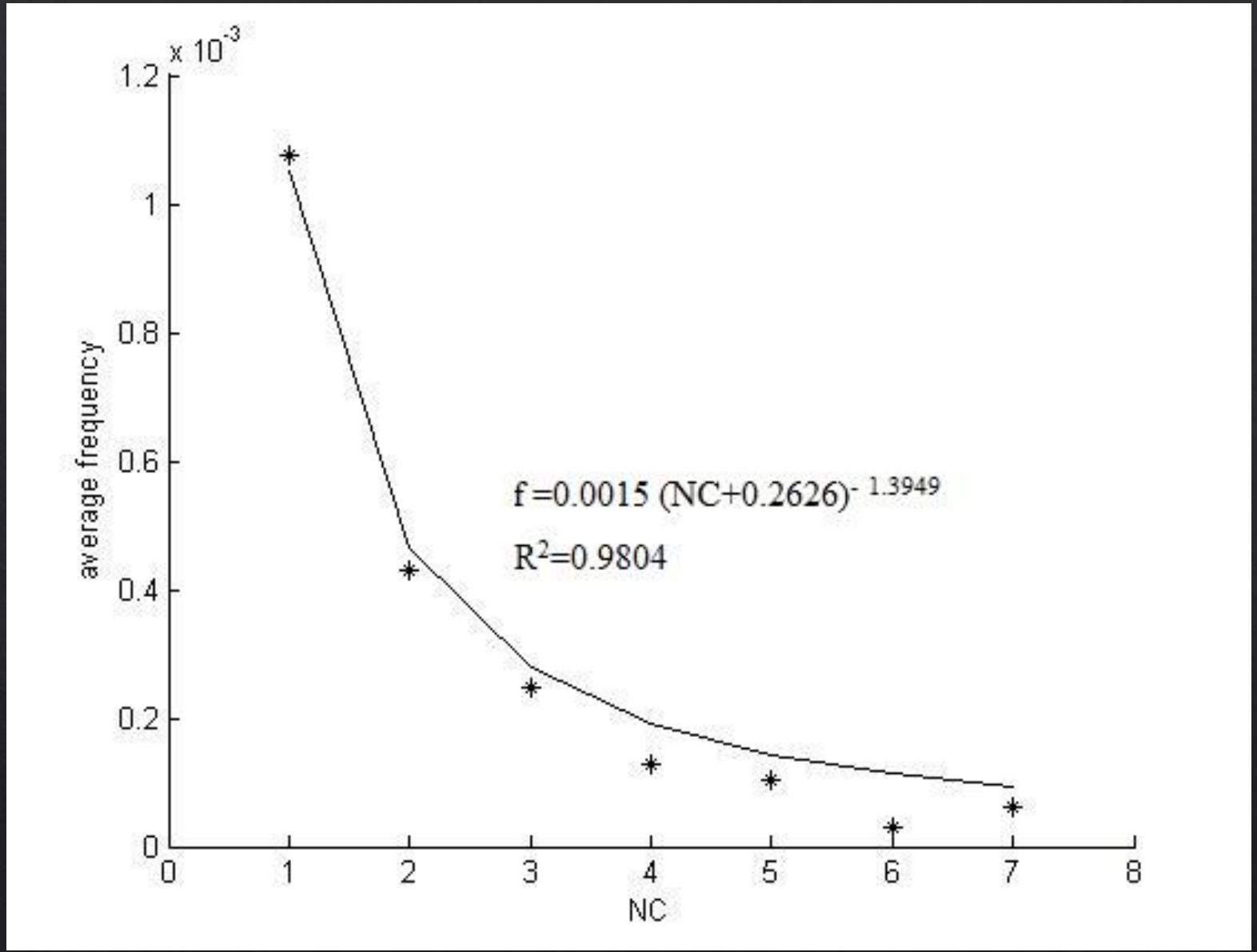

睡

Number of Stroks	Number of Chracters	Examples	Cumulative frequency	Average frequency
1	2	一,乙	0.0104818889	0.0052409444
2	19	人,了,力	0.0270823814	0.0014253885
3	51	大,上,工	0.0535283278	0.0010495751
4	113	中,不,为	0.0924586324	0.0008182180
5	145	发,业,民	0.0989700908	0.0006825524
6	237	在,有,年	0.1514154083	0.0006388836
7	307	作,这,来	0.1035224004	0.0003372065
8	379	的,国,和	0.1611448260	0.0004251842
9	368	是,要,政	0.1006057580	0.0002733852
10	343	家,部,展	0.0675444718	0.0001969227
11	290	理,得,基	0.0437326729	0.0001508023
12	276	就,等,提	0.0366750136	0.0001328805
13	177	新,意,解	0.0215748566	0.0001218918
14	124	道,管,赛	0.0111661176	0.0000900493
15	100	题,增,德	0.0084463093	0.0000844631
16	55	整, 融,器	0.0033965338	0.0000617552
17	33	藏,繁,疑	0.0012677566	0.0000384169
18	11	翻,覆,藤	0.0002244518	0.0000204047
19	13	警,疆,攀	0.0005720697	0.0000440054
20	11	籍,灌,耀	0.0002580577	0.0000234598
21	4	露,霸,髓	0.0001694056	0.0000423514
22	2	囊,镶	0.0000231615	0.0000115808
23	1	罐	0.0000151569	0.0000151569



The relationship between NS and the average frequency of Chinese characters that share the same NS

Number of Components	Number of Characters	Examples	Cumulative frequency	Average frequency
1	187	一,中,人	0.2008274347	0.0010739435
2	969	的,国,和	0.4164280467	0.0004297503
3	1186	在,是,发	0.2922298614	0.0002463995
4	534	高,能,说	0.0682191084	0.0001277511
5	150	题,领,解	0.0153812537	0.0001025417
6	32	歌,疑,衡	0.0010030980	0.0000313468
7	3	疆,凝,颤	0.0001869466	0.0000623155



The relationship between NC and the average frequency of Chinese characters that share the same NC

Diachronic observation of Chinese Words Based on Google Ngram

Diachronic observation of Chinese Words Based on Google Ngram

- Observation
  - the lexicon changes of a language during a time period
- Hypothesis verification
  - There is a multi-syllabification trend of Chinese words during Chinese evolution (Chen, H., Liang, J., & Liu, H. 2015 PLoS ONE)

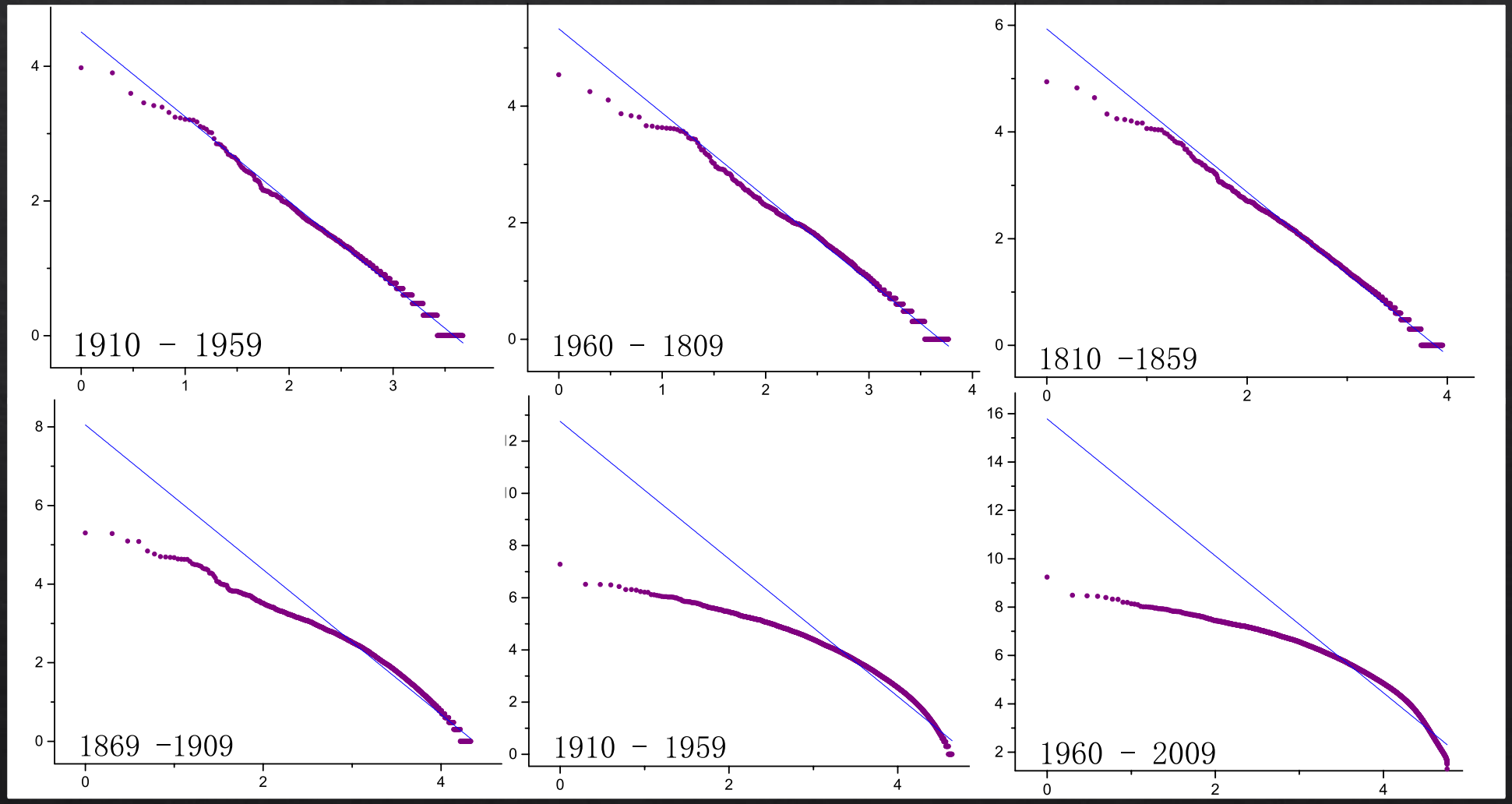
Diachronic observation of Chinese Words Based on Google Ngram

- Observation
  - the lexicon changes of a language during a time period
- Hypothesis verification
  - There is a multi-syllabification trend of Chinese words during Chinese evolution (Chen, H., Liang, J., & Liu, H. 2015 PLoS ONE)
- The Chinese Google 1-gram data (segmented-words frequency)
  - Symbols, Japanese, Roman letters, Arabic numerals
  - 1710 ~ 2009 (300 years)
    - × Before 1710 (sparse, segmentation accuracy)
    - 6 * 50 (years)
    - 1710~59, 1760~1809, 1810~59, 1869~1909, 1910~59, and 1960~2009

Diachronic observation of Chinese Words Based on Google Ngram

- Observation
  - the lexicon changes of a language during a time period
- Hypothesis verification
  - There is a multi-syllabification trend of Chinese words during Chinese evolution (Chen, H., Liang, J., & Liu, H. 2015 PLoS ONE)
- The Chinese Google 1-gram data (segmented-words frequency)
  - Symbols, Japanese, Roman letters, Arabic numerals
  - 1710 ~ 2009 (300 years)
    - × Before 1710 (sparse, segmentation accuracy)
    - 6 * 50 (years)
    - 1710~59, 1760~1809, 1810~59, 1869~1909, 1910~59, and 1960~2009
- Power law fitting

Power Law Fitting



Power Law Fitting: $y = a*x^{-b}$

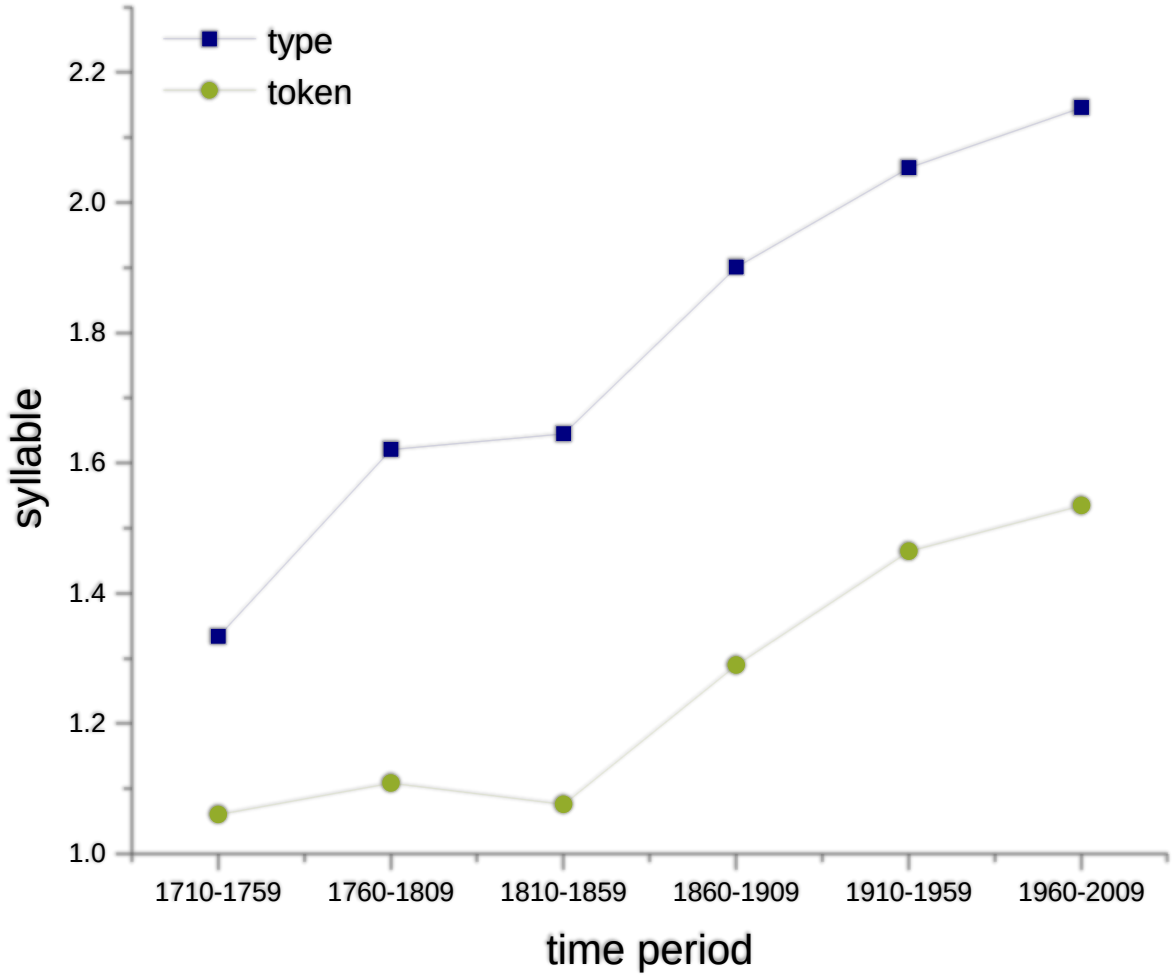
Time Period	a	b	R^2
1760-1809	34,791	-0.98122	0.98755
1810-1859	96,530	-0.94701	0.96039
1860-1909	249,946	-0.82656	0.92599
1910-1959	16,579,000	-1.09064	0.91207
1960-2009	1,512,270,000	-1.05617	0.90595

- all the six frequency lists fit the Power Law well (the minimum R2 is 0.90595) despite the huge data scale (the largest list includes more than 21 billion tokens)
- The Google Ngram data should be suitable for doing some linguistic studies

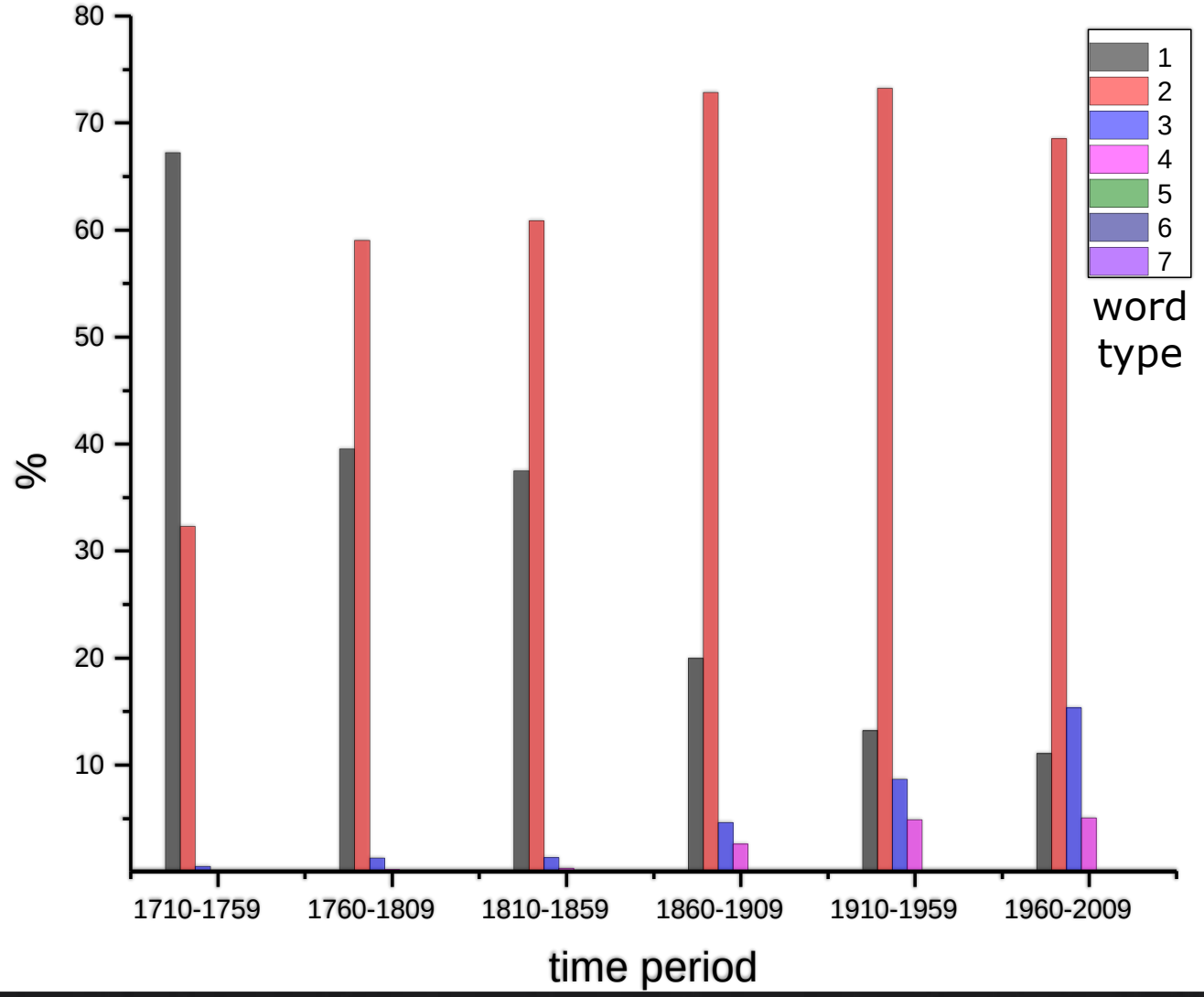
Diachronic observation of Chinese Words Based on Google Ngram

- Observation
  - the lexicon changes of a language during a time period
- Hypothesis verification
  - There is a multi-syllabification trend of Chinese words during Chinese evolution (Chen, H., Liang, J., & Liu, H. 2015 PLoS ONE)
- The Chinese Google 1-gram data (segmented-words frequency)
  - Symbols, Japanese, Roman letters, Arabic numerals
  - 1710 ~ 2009 (300 years)
    - × Before 1710 (sparse, segmentation accuracy)
    - 6 * 50 (years)
    - 1710~59, 1760~1809, 1810~59, 1869~1909, 1910~59, and 1960~2009
- Power law fitting
- Average word length

Average word length



Word length distribution



Outline

- Why different language descriptions and analysis
- Texts
- Treebanks

Treebanks

- You may have heard
  - Noam Chomsky
  - Grammar, syntax, semantics…
  - Predicator, agent, patient, subject, object, modifier, valence…
  - Machine translation, natural language processing, natural language understanding…

Treebanks

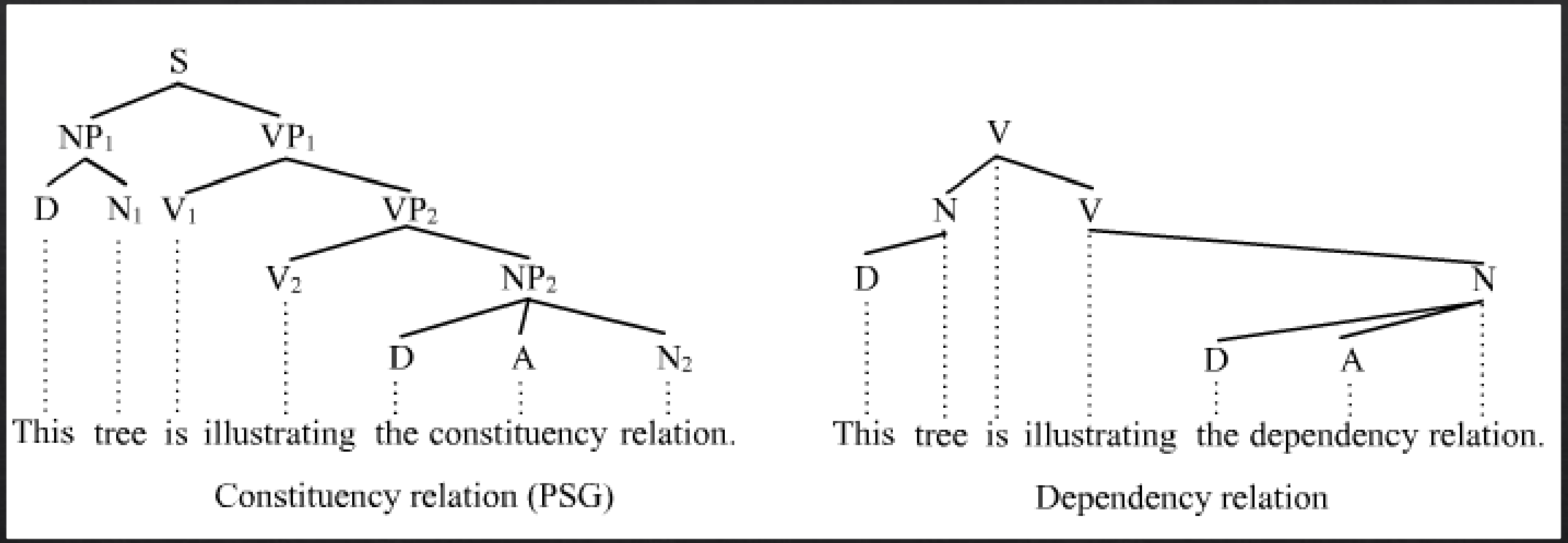
- You may have heard
  - Noam Chomsky
  - Grammar, syntax, semantics…
  - Predicator, agent, patient, subject, object, modifier, valence…
  - Machine translation, natural language processing, natural language understanding…
- Advantages
  - Two dimensional description (non-linear modals)
  - More detailed information of language understanding or generation
  - Practical uses for language teaching and language processing

Treebanks

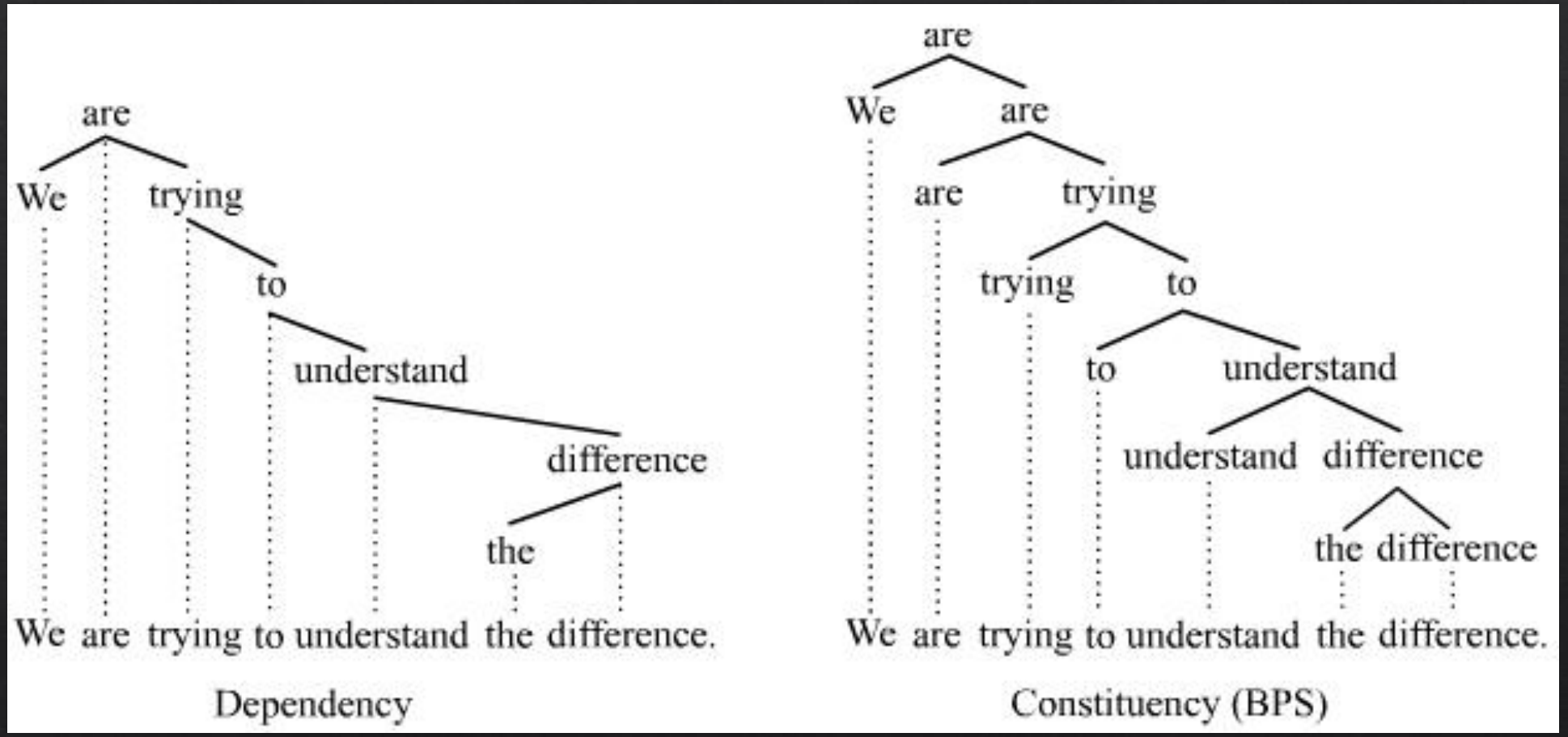
- You may have heard
  - Noam Chomsky
  - Grammar, syntax, semantics…
  - Predicator, agent, patient, subject, object, modifier, valence…
  - Machine translation, natural language processing, natural language understanding…
- Advantages
  - Two dimensional description (non-linear modals)
  - More detailed information of language understanding or generation
  - Practical uses for language teaching and language processing
- Disadvantages
  - Disagreements on description frameworks (less objective)
  - Less available
  - More difficult to test hypothesis (Null models/hypothesis)

Treebanks

- You may have heard
  - Noam Chomsky
  - Grammar, syntax, semantics…
  - Predicator, agent, patient, subject, object, modifier, valence…
  - Machine translation, natural language processing, natural language understanding…
- Advantages
  - Two dimensional description (non-linear modals)
  - More detailed information of language understanding or generation
  - Practical uses for language teaching and language processing
- Disadvantages
  - Disagreements on description frameworks (less objective)
  - Less available
  - More difficult to test hypothesis (Null models/hypothesis)
- What is treebank

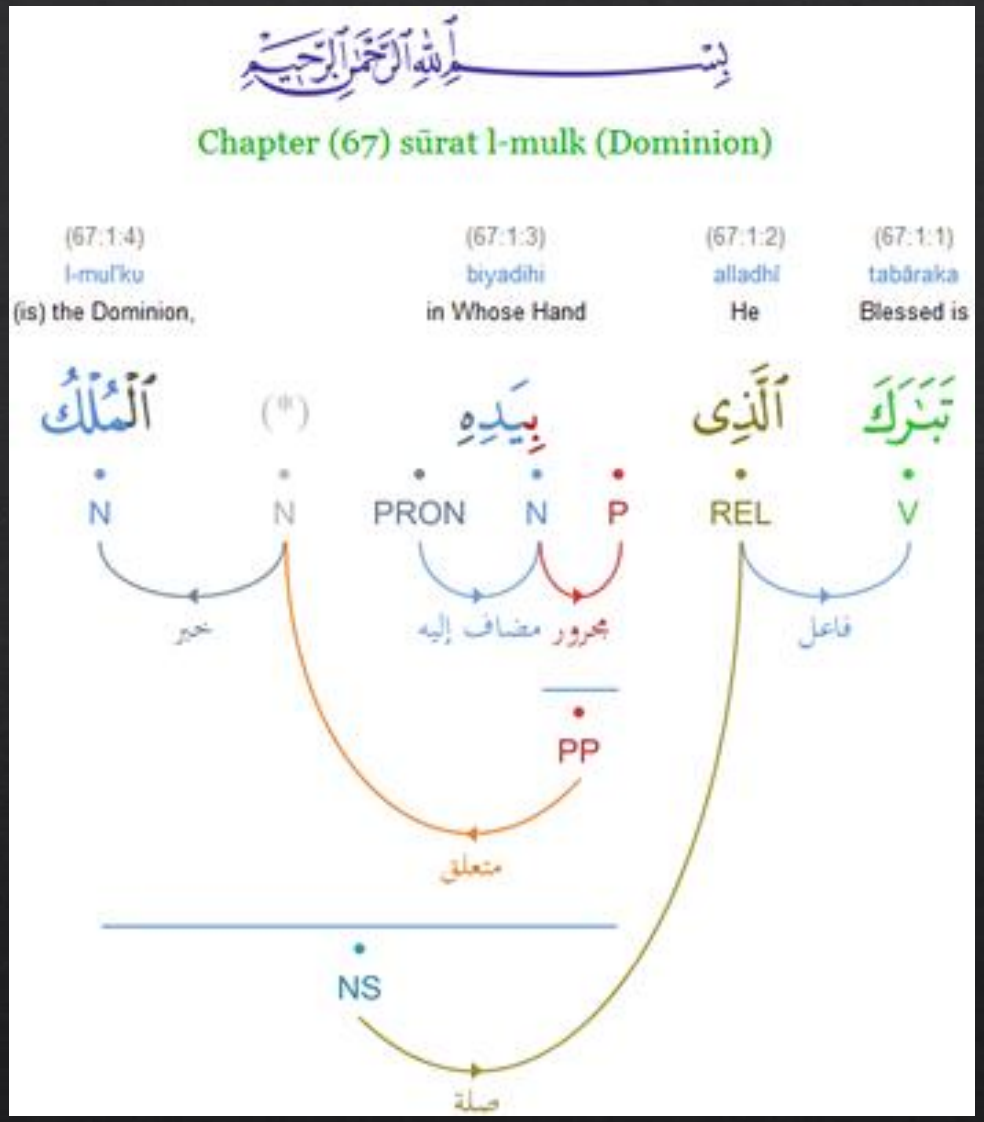


http://en.wikipedia.org/wiki/Phrase_structure_grammar



http://en.wikipedia.org/wiki/Dependency_grammar

Hybrid constituency/dependency tree from the Quranic Arabic Corpus



http://en.wikipedia.org/wiki/Treebank

```
1    Pricing        pricing        NN    _    2    NMOD
2    details        detail         NNS   _    3    SBJ
3    were           be             VBD   _    0    ROOT
4    n't            not            RB    _    3    ADV
5    immediately    immediately    RB    _    6    AMOD
6    available      available      JJ    _    3    PRD
7    .              .              .     _    3    P

1    She            she            PRP   _    2    SBJ
2    bought         buy            VBD   _    0    ROOT
3    a              a              DT    _    4    NMOD
4    car            car            NN    _    2    OBJ
5    .              .              .     _    2    P
```

CoNLL-X

Other format (xml)

```
<?xml-stylesheet type="text/xsl" href="nlp_style.xsl" ?>↓
<xml4nlp>↓
    <doc>↓
        <para id="0">↓
            <sent id="0" cont="迈向充满希望的新世纪——一九九八年新年讲话（附图
片1张）">↓
                <word id="0" cont="迈向" pos="v" parent="-1" relate="HED" />↓
                <word id="1" cont="充满" pos="v" parent="3" relate="DE" />↓
                <word id="2" cont="希望" pos="n" parent="1" relate="VOB" />↓
                <word id="3" cont="的" pos="u" parent="5" relate="ATT" />↓
                <word id="4" cont="新" pos="a" parent="5" relate="ATT" />↓
                <word id="5" cont="世纪" pos="n" parent="0" relate="VOB" />↓
                <word id="6" cont="——" pos="wp" parent="-2" relate="PUN" />↓
                <word id="7" cont="一九九八年" pos="nt" parent="8" relate="ATT"
```

https://code.google.com/p/clearparser/wiki/DataFormat

Classifying Languages by Dependency Structure

Typologies of Delexicalized Universal Dependency Treebanks

Classifying Languages by Dependency Structure
Typologies of Delexicalized Universal Dependency Treebanks

- Observation
  - Syntactical differences between different languages
- Question
  - Can UD be used for language typology study and reveal the similarity and diversity between language families? If so, then how?

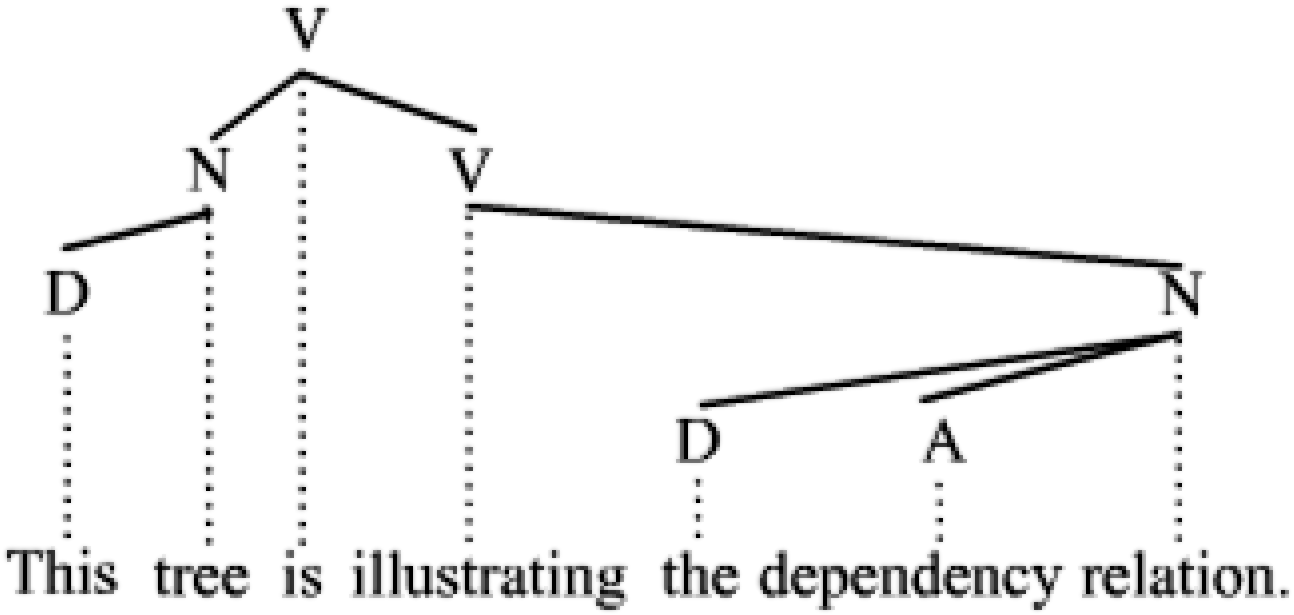
Classifying Languages by Dependency Structure
Typologies of Delexicalized Universal Dependency Treebanks

- Observation
  - Syntactical differences between different languages
- Question
  - Can UD be used for language typology study and reveal the similarity and diversity between language families? If so, then how?
- UD treebanks
  - UD 2.0 70 treebanks of 50 languages
  - 63 of which have more than 10,000 tokens.

Classifying Languages by Dependency Structure
Typologies of Delexicalized Universal Dependency Treebanks

- Observation
  - Syntactical differences between different languages
- Question
  - Can UD be used for language typology study and reveal the similarity and diversity between language families? If so, then how?
- UD treebanks
  - UD 2.0 70 treebanks of 50 languages
  - 63 of which have more than 10,000 tokens.
- Dependency distance & word order (head initial, head final, mixed)
- Distributions of frequencies
- Clustering algorithm

Dependency Length



DL(this-is)=3-1=2

DL(is-illustrating)=3-4=-1

Dependency Length

$$DDD(R) = \frac{\sum_{r \in R} distance(r)}{frequency(R)}$$

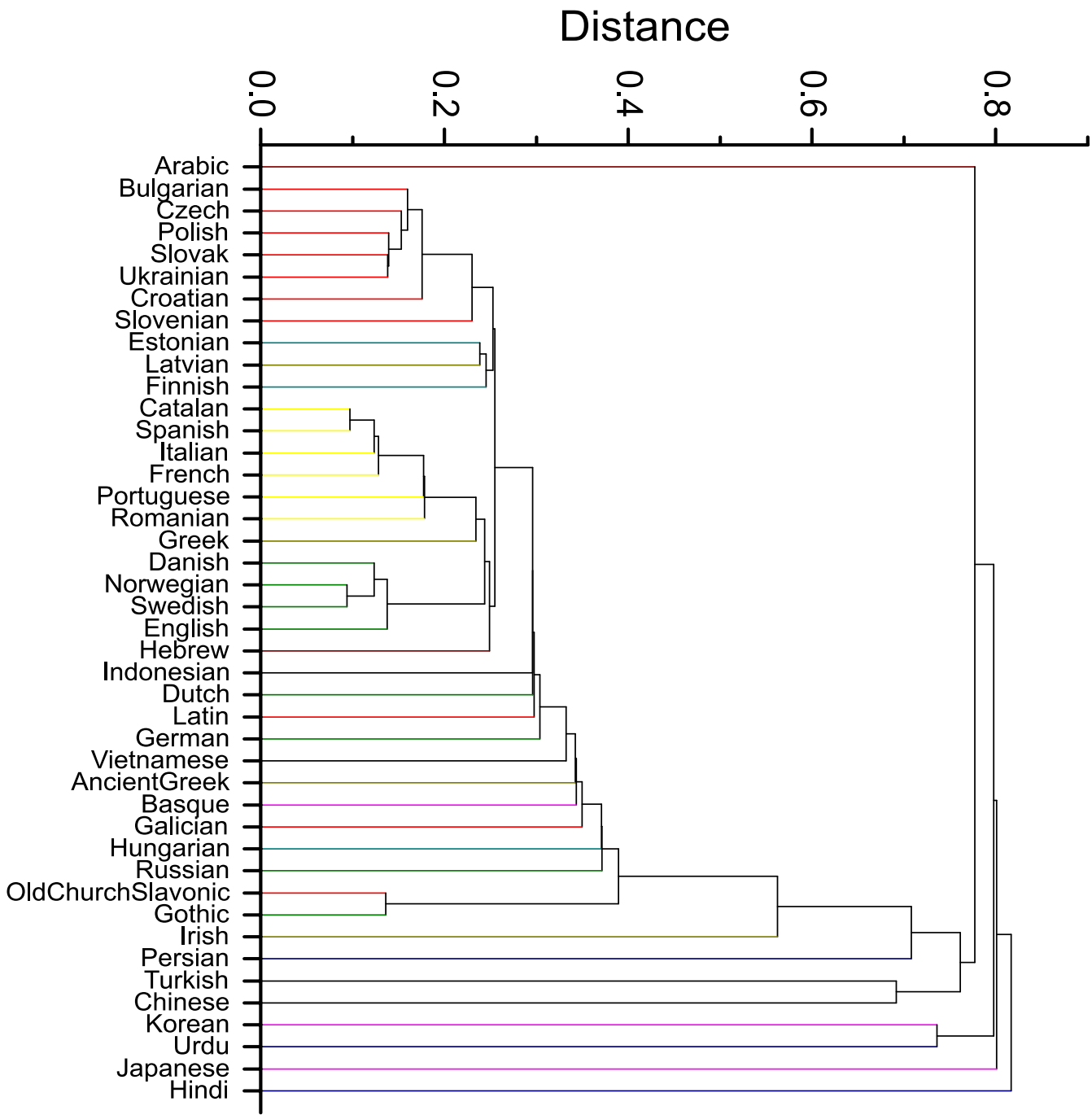
R: type of dependencies r: dependencies

Classifying Languages by Dependency Structure
Typologies of Delexicalized Universal Dependency Treebanks

- Observation
  - Syntactical differences between different languages
- Question
  - Can UD be used for language typology study and reveal the similarity and diversity between language families? If so, then how?
- UD treebanks
  - UD 2.0 70 treebanks of 50 languages
  - 63 of which have more than 10,000 tokens.
- Dependency distance & word order (head initial, head final, mixed)
- Distributions of frequencies

Classifying Languages by Dependency Structure
Typologies of Delexicalized Universal Dependency Treebanks

- Observation
  - Syntactical differences between different languages
- Question
  - Can UD be used for language typology study and reveal the similarity and diversity between language families? If so, then how?
- UD treebanks
  - UD 2.0 70 treebanks of 50 languages
  - 63 of which have more than 10,000 tokens.
- Dependency distance & word order (head initial, head final, mixed)
- Distributions of frequencies
- Clustering algorithm


Distance
0.0
0.2
0.4
0.6
0.8
Arabic
Bulgarian
Czech
Polish
Slovak
Ukrainian
Croatian
Slovenian
Estonian
Latvian
Finnish
Catalan
Spanish
Italian
French
Portuguese
Romanian
Greek
Danish
Norwegian
Swedish
English
Hebrew
Indonesian
Dutch
Latin
German
Vietnamese
AncientGreek
Basque
Galician
Hungarian
Russian
OldChurchSlavonic
Gothic
Irish
Persian
Turkish
Chinese
Korean
Urdu
Japanese
Hindi

0.8
0.6
0.4
0.2
0.0
ar
bg
cs
cs_cac
pl
sk
uk
hr
ru
cs_cltt
sl
ca
es_ancora
es
fr
fr_partut
it_partut
it
pt
fr_sequoia
gl_treegal
ro
el
da
no_bokmaal
no_nynorsk
sv
sv_lines
en_lines
en
en_partut
he
et
fi
lv
pt_br
de
nl
id
la_itt
vi
eu
hu
grc
la
gl
cu
got
la_proiel
grc_proiel
fi_ftb
ru_syntagrus
sl_sst
ga
ar_nyuad
nl_lassysmall
tr
zh
fa
ko
ur
la
hi

Outline

- Why different language descriptions and analysis
- Texts
- Treebanks
- Networks

Networks

- You may have heard
  - six degrees of separation, social network, small world…
  - graph theory, seven bridges, complex theory…

Networks

- You may have heard
  - six degrees of separation, social network, small world…
  - graph theory, seven bridges, complex theory…
- Advantages
  - Three dimensional description
  - Macro system perspective (break the sentence boundaries)

Networks

- You may have heard
  - six degrees of separation, social network, small world…
  - graph theory, seven bridges, complex theory…
- Advantages
  - Three dimensional description
  - Macro system perspective (break the sentence boundaries)
- Disadvantages
  - Too abstract to be understood or interpreted?
  - More difficult to test hypothesis (Null models/hypothesis)
  - Disappeared details?

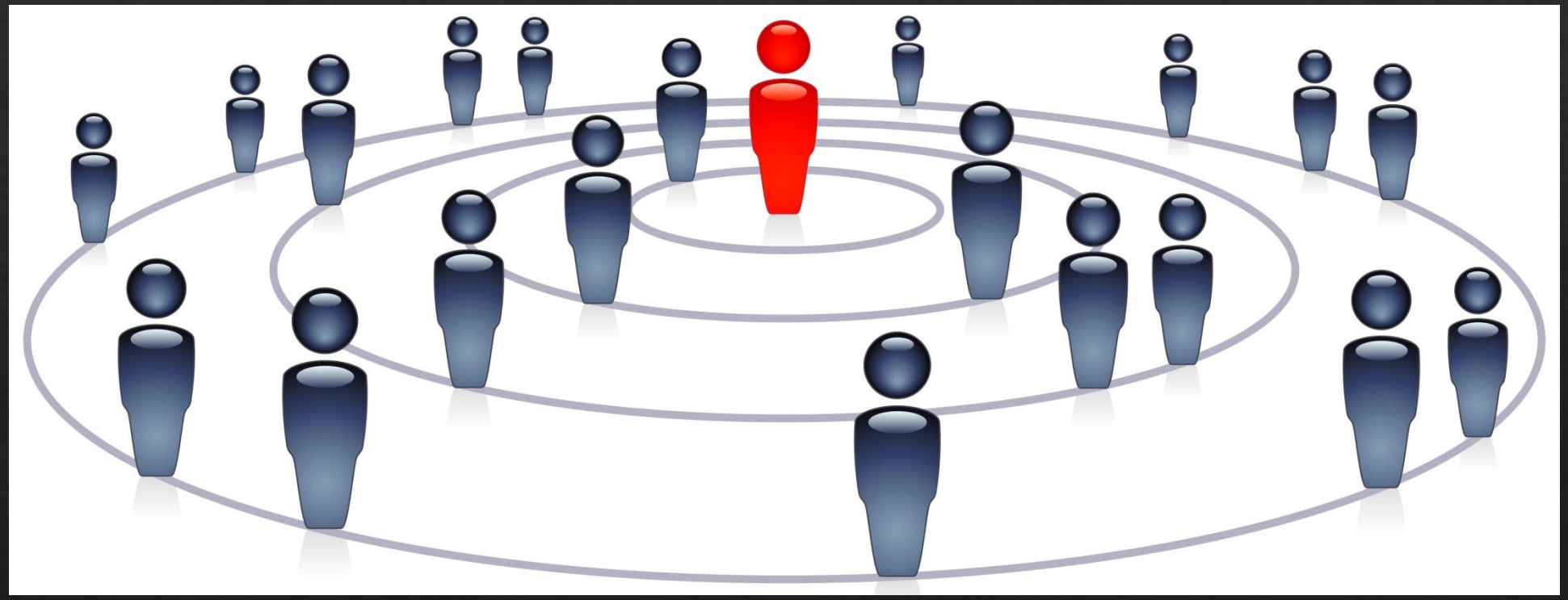
Networks

- You may have heard
  - six degrees of separation, social network, small world…
  - graph theory, seven bridges, complex theory…
- Advantages
  - Three dimensional description
  - Macro system perspective (break the sentence boundaries)
- Disadvantages
  - Too abstract to be understood or interpreted?
  - More difficult to test hypothesis (Null models/hypothesis)
  - Disappeared details?
- What is network

Origin and development

Origin and development

- Six Degrees of Separation



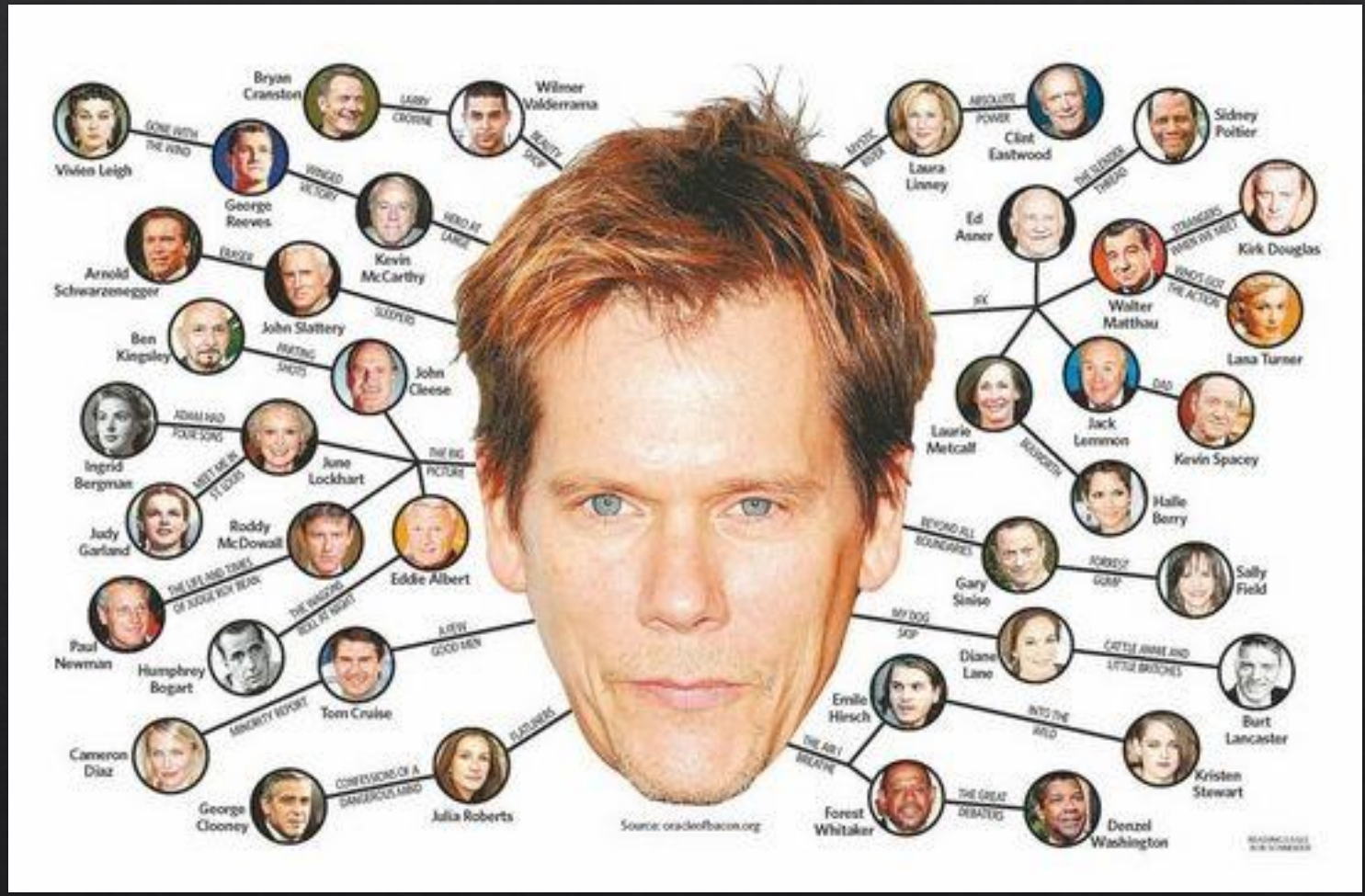
http://indianvox.com/articles/news/143667.php

Origin and development

- Six Degrees of Separation
  - Harvard psychologist Stanley Milgram 1967
    - U.S. / 300 letters / >60 / 6

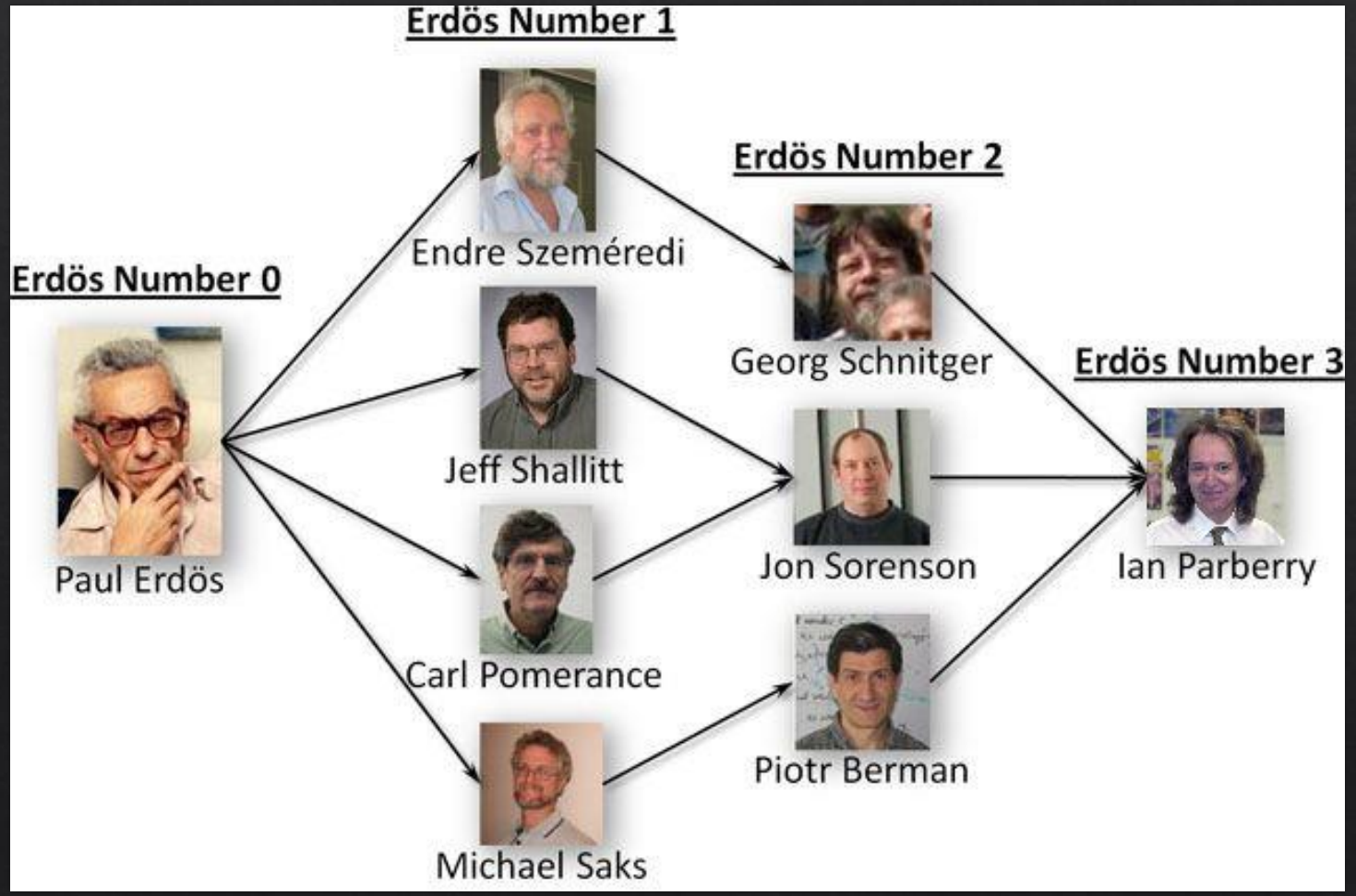
https://en.wikipedia.org/wiki/Six_degrees_of_separation

Bacon number



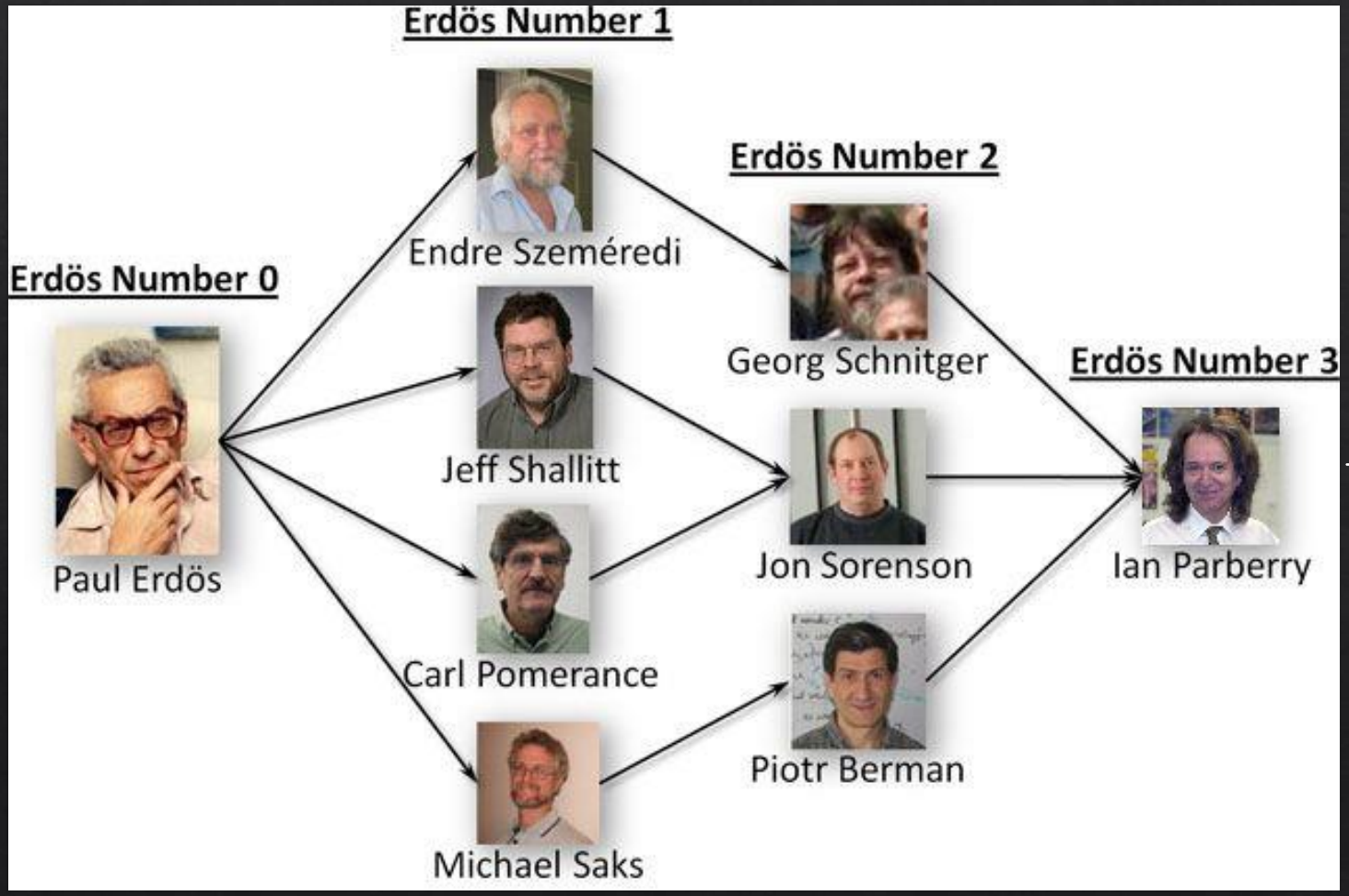
http://badegreedraza.blogspot.jp/2015/04/6-degrees-of-kevin-bacon.html

Erdős number



https://larc.unt.edu/ian/claimtofame.html

Erdős number



Erdős number 4

https://larc.unt.edu/ian/claimtofame.html

Origin and development

- Six Degrees of Separation
  - Harvard psychologist Stanley Milgram 1967
    - U.S. / 300 letters / >60 / 6
  - Columbia professor Duncan Watts 2001
    - 19 email targets (157 countries) / 48,000 senders / 6

https://en.wikipedia.org/wiki/Six_degrees_of_separation

Origin and development

- Six Degrees of Separation
  - Harvard psychologist Stanley Milgram 1967
    - U.S. / 300 letters / >60 / 6
  - Columbia professor Duncan Watts 2001
    - 19 email targets (157 countries) / 48,000 senders / 6
  - Microsoft researchers Jure Leskovec and Eric Horvitz 2007
    - 30 billion Skype messages / 240 million people / 6

https://en.wikipedia.org/wiki/Six_degrees_of_separation

Origin and development

- Six Degrees of Separation
- Big Data Area & Quantitative Method
  - UCINET (Borgatti et al. 2002)
  - PAJEK (Nooy et al. 2005)
  - NETDRAW (Borgatti 2002)
  - CYTOSCAPE (Shannon 2003)
  - …

Origin and development

- Six Degrees of Separation
- Big Data Area & Quantitative Method
- Complex System

Origin and development

- Six Degrees of Separation
- Big Data Area & Quantitative Method
- Complex System
  - Simplification & Decomposition

Origin and development

- Six Degrees of Separation
- Big Data Area & Quantitative Method
- Complex System
  - ~~Simplification & Decomposition~~

Origin and development

- Six Degrees of Separation
- Big Data Area & Quantitative Method
- Complex System
  - Duncan J Watts Small World
  - Albert-László Barabási Scale Free

Origin and development

- Six Degrees of Separation
- Big Data Area & Quantitative Method
- Complex System
  - Duncan J Watts Small World
  - Albert-László Barabási Scale Free
  - Universal Features of Existing Systems

Origin and development

- Six Degrees of Separation
- Big Data Area & Quantitative Method
- Complex System
  - Universal Features of Existing Systems

Origin and development

- Six Degrees of Separation
- Big Data Area & Quantitative Method
- Complex System
  - Universal Features of Existing Systems
    - Brain network, Nation network, Social network, Traffic network, Epidemic network…

Origin and development

- Six Degrees of Separation
- Big Data Area & Quantitative Method
- Complex System
  - Universal Features of Existing Systems
    - Brain network, Nation network, Social network, Traffic network, Epidemic network…
    - Language network

Origin and development (Language)

Origin and development (Language)

- Language is an ‘organic’ complex system
  - Ramon Ferrer i Cancho & Ricard Solé

Origin and development (Language)

- Language is an ‘organic’ complex system
  - Ramon Ferrer i Cancho & Ricard Solé
- Different language networks
  - Syntactic dependency networks, language development or language evolution, language clustering and linguistic categorization, manual and machine translation, word sense disambiguation, communication and interaction, semantic networks, phonetics, morphology, parts of speech, knowledge networks, cognitive networks …

Origin and development (Language)

- Why do we need the network approach in linguistic research?

Origin and development (Language)

- Why do we need the network approach in linguistic research?
  - General advantages of the network approach
    - Graphic view, intuitive and easy to understand
    - Overall perspective, not only focusing on the sentence level
    - General method, easier to communicate and to combine with other disciplines, such as neurology
    - Tools developed for graphs and other domains can be applied

Origin and development (Language)

- Why do we need the network approach in linguistic research?
  - General advantages of the network approach
    - Graphic view, intuitive and easy to understand
    - Overall perspective, not only focusing on the sentence level
    - General method, easier to communicate and to combine with other disciplines, such as neurology
    - Tools developed for graphs and other domains can be applied
  - Limitations of current methods of treebank-based linguistics study
    - The boundary between sentences

Origin and development (Language)

- Why do we need the network approach in linguistic research?
  - General advantages of the network approach
    - Graphic view, intuitive and easy to understand
    - Overall perspective, not only focusing on the sentence level
    - General method, easier to communicate and to combine with other disciplines, such as neurology
    - Tools developed for graphs and other domains can be applied
  - Limitations of current methods of treebank-based linguistics study
    - The boundary between sentences

1	Pricing	pricing	NN	_	2	NMOD
2	details	detail	NNS	_	3	SBJ
3	were	be	VBD	_	0	ROOT
4	n't	not	RB	_	3	ADV
5	immediately	immediately	RB	_	6	AMOD
6	available	available	JJ	_	3	PRD
7	.	.	.	_	3	P

1	She	she	PRP	_	2	SBJ
2	bought	buy	VBD	_	0	ROOT
3	a	a	DT	_	4	NMOD
4	car	car	NN	_	2	OBJ
5	.	.	.	_	2	P

1	Pricing	pricing	NN	_	2	NMOD
2	details	detail	NNS	_	3	SBJ
3	were	be	VBD	_	0	ROOT
4	n't	not	RB	_	3	ADV
5	immediately	immediately	RB	_	6	AMOD
6	available	available	JJ	_	3	PRD
7	.	.	.	_	3	P
1	She	she	PRP	_	2	SBJ
2	bought	buy	VBD	_	0	ROOT
3	a	a	DT	_	4	NMOD
4	car	car	NN	_	2	OBJ
5	.	.	.	_	2	P

language is a system of signs, each of which is an arbitrary union of sound and meaning… Any given sign, is defined by its relationships with the others. (Saussure 1959)

1	Pricing	pricing	NN	_	2	NMOD
2	details	detail	NNS	_	3	SBJ
3	were	be	VBD	_	0	ROOT
4	n't	not	RB	_	3	ADV
5	immediately	immediately	RB	_	6	AMOD
6	available	available	JJ	_	3	PRD
7	.	.	.	_	3	P

1	She	she	PRP	_	2	SBJ
2	bought	buy	VBD	_	0	ROOT
3	a	a	DT	_	4	NMOD
4	car	car	NN	_	2	OBJ
5	.	.	.	_	2	P

language is a system of signs, each of which is an arbitrary union of sound and meaning… Any given sign, is defined by its relationships with the others. (Saussure 1959)

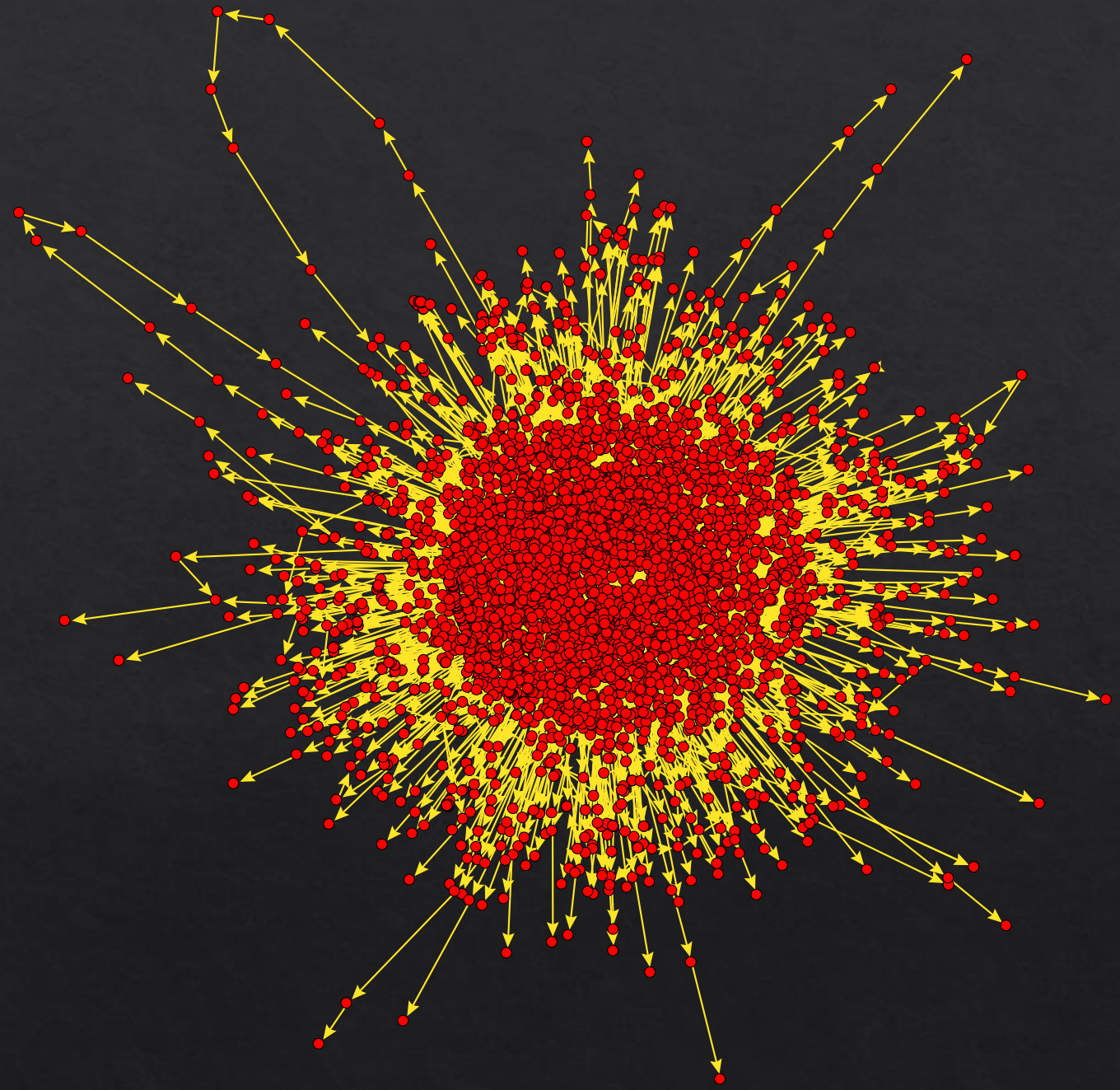
language is a system of signs, each of which is an arbitrary union of sound and meaning… Any given sign, is defined by its relationships with the others. (Saussure 1959)

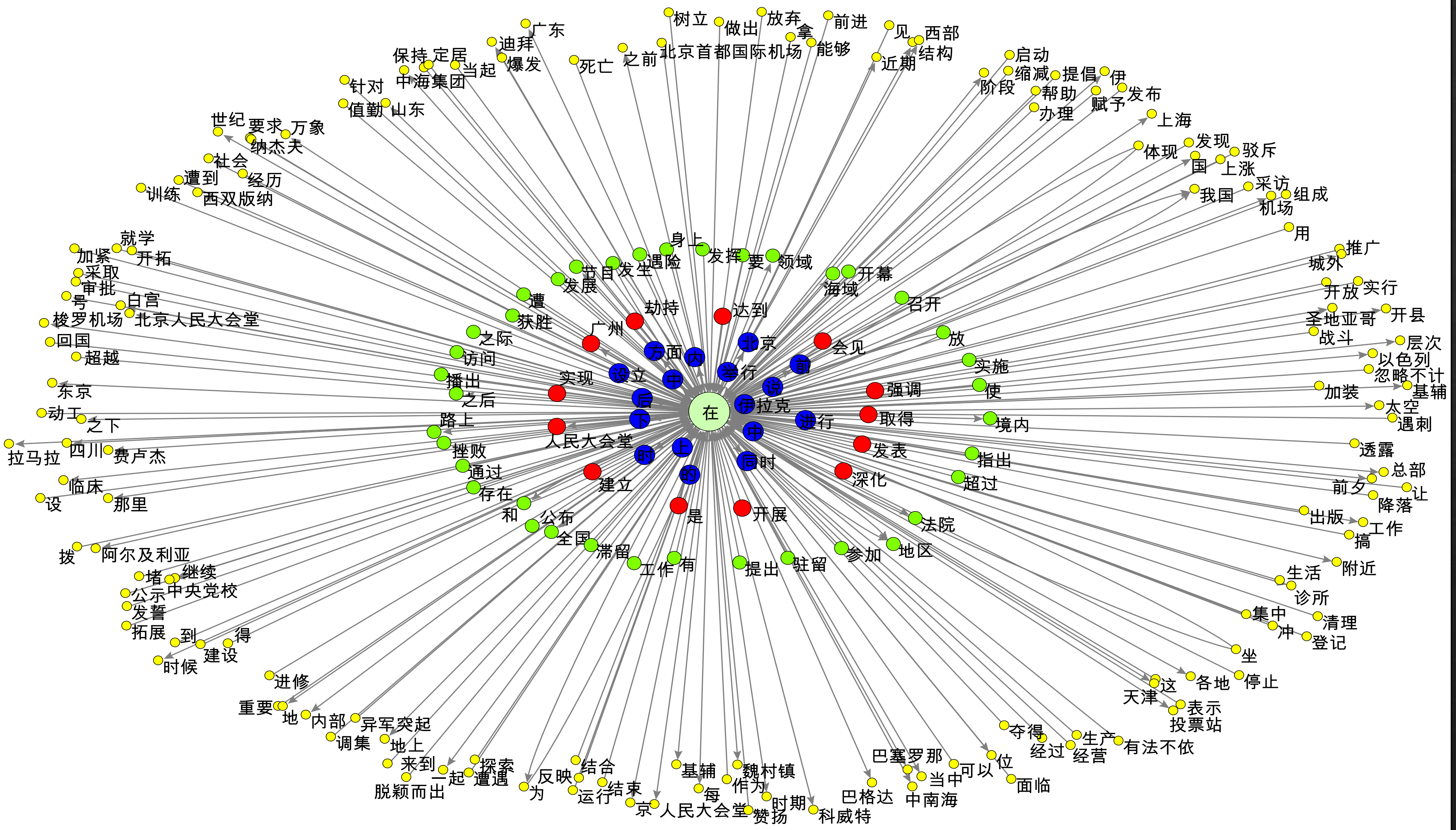
language is a system of signs, each of which is an arbitrary union of sound and meaning… Any given sign, is defined by its relationships with the others. (Saussure 1959)

whole ≠ sum of parts

Origin and development (Language)

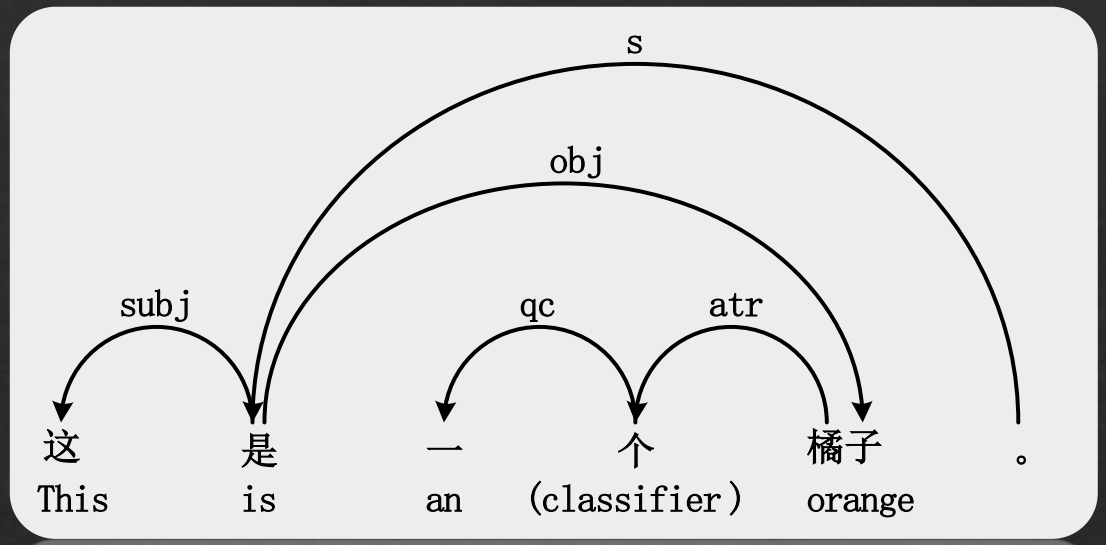
- Why do we need the network approach in linguistic research?
  - General advantages of the network approach
    - Graphic view, intuitive and easy to understand
    - Overall perspective, not only focusing on the sentence level
    - General method, easier to communicate and to combine with other disciplines, such as neurology
    - Tools developed for graphs and other domains can be applied
  - Limitations of current methods of treebank-based linguistics study
    - ~~The boundary between sentences~~

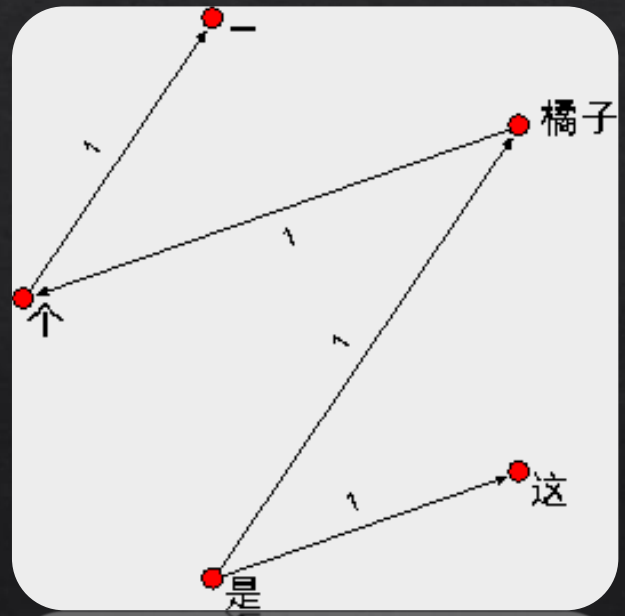




在
方面
内
中
举行
北京
前
设
伊拉克
进行
中
同时
的
上
时
下
后
设立
劫持
达到
会见
强调
取得
发表
深化
开展
是
建立
人民大会堂
实现
广州
身上
发挥
要
领域
开幕
海域
召开
放
实施
使
境内
指出
超过
法院
地区
参加
驻留
提出
有
工作
滞留
全国
公布
和
存在
通过
挫败
路上
之后
播出
访问
之际
获胜
遭
发展
节目
发生
遇险
树立
做出
放弃
拿
前进
能够
北京首都国际机场
之前
死亡
广东
迪拜
爆发
保持
定居
当起
中海集团
针对
值勤
山东
世纪
要求
万象
纳杰夫
社会
遭到
经历
训练
西双版纳
就学
加紧
开拓
采取
审批
号
白宫
梭罗机场
北京人民大会堂
回国
超越
东京
动工
之下
拉马拉
四川
费卢杰
临床
设
那里
拔
阿尔及利亚
堵
继续
公示
中央党校
发誓
拓展
到
得
建设
时候
进修
重要
地
内部
异军突起
调集
地上
来到
一起
探索
遭遇
脱颖而出
为
反映
结合
运行
结束
京
人民大会堂
基辅
每
魏村镇
作为
时期
赞扬
科威特
巴格达
巴塞罗那
当中
中南海
可以
位
面临
夺得
经过
生产
经营
有法不依
天津
这
各地
表示
投票站
停止
坐
集中
冲
清理
登记
生活
诊所
附近
出版
工作
搞
前夕
降落
让
总部
透露
遇刺
太空
基辅
加装
忽略不计
以色列
层次
战斗
圣地亚哥
开县
开放
实行
城外
推广
用
机场
组成
我国
采访
国
上涨
驳斥
体现
发现
上海
赋予
发布
办理
帮助
提倡
伊
阶段
缩减
启动
近期
见
西部
结构

How to Build the Networks

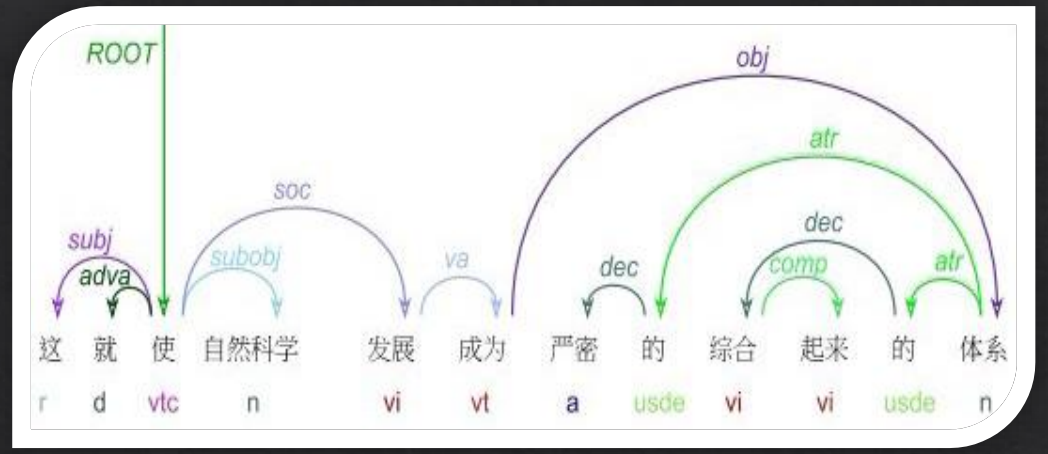

s
obj
subj
qc
atr
这
是
一
个
橘子
。
This
is
an
(classifier)
orange

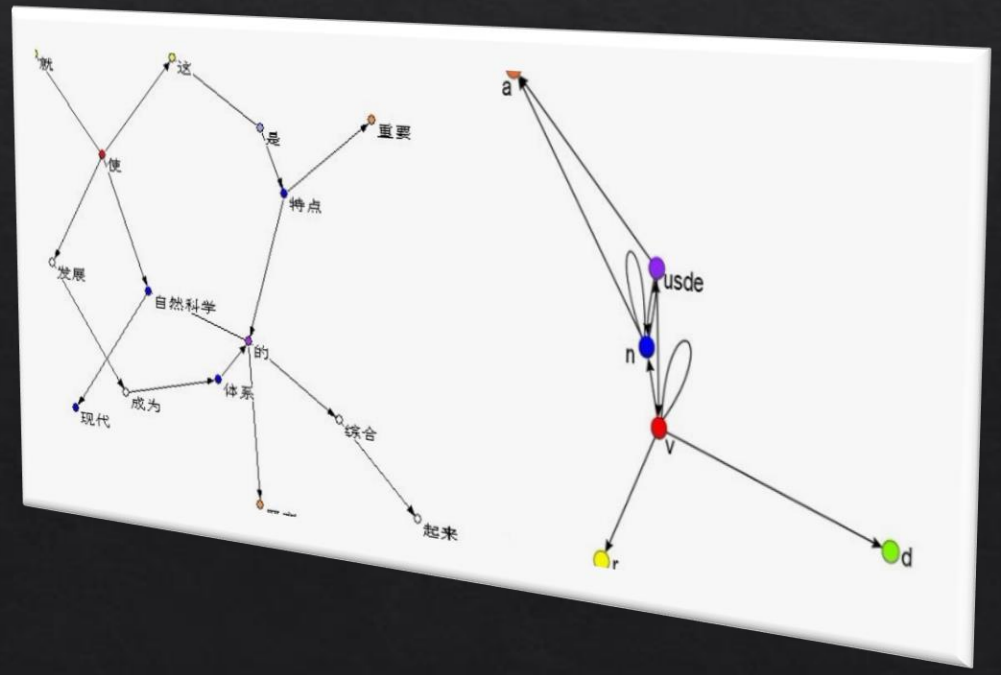

一
橘子
个
这
是
1
1
1
1

Annotation of a sample sentence in the Treebank.
这是一个橘子 *zhe-shi-yi-ge-ju-zi* 'this is an orange'

Sentence Order	Dependent			Governor			Dependency Type
	Order	Character	POS	Order	Character	POS	
S1	1	这 zhe	pronoun	2	是 shi	verb	Subject
S1	2	是 shi	verb	6	。	punctuation	main governor
S1	3	一 yi	numeral	4	个 ge	classifier	complement of classifier
S1	4	个 ge	classifier	5	橘子 juzi	noun	Attributer
S1	5	橘子 juzi	noun	2	是 shi	verb	Object
S1	6	。	punctuation				

sno	wno	wd	wtag	gno	gwd	gtag	syt
S1	1	这	r	3	使	vtc	subj
S1	2	就	d	3	使	vtc	adva
S1	3	使	vtc	0	s	s	s
S1	4	自然科学	n	3	使	vtc	subobj
S1	5	发展	vi	3	使	vtc	soc
S1	6	成为	vt	5	发展	vi	va
S1	7	严密	a	8	的	usde	dec
S1	8	的	usde	12	体系	n	atr
S1	9	综合	vi	11	的	usde	dec
S1	10	起来	vi	9	综合	vi	comp
S1	11	的	usde	12	体系	n	atr
S1	12	体系	n	6	成为	vt	obj
S2	1	这	r	2	是	vl	subj
S2	2	是	vl	0	s	s	s
S2	3	现代	n	4	自然科学	n	atr
S2	4	自然科学	n	5	的	usde	dec
S2	5	的	usde	7	特点	n	atr
S2	6	重要	a	7	特点	n	atr
S2	7	特点	n	2	是	vl	obj

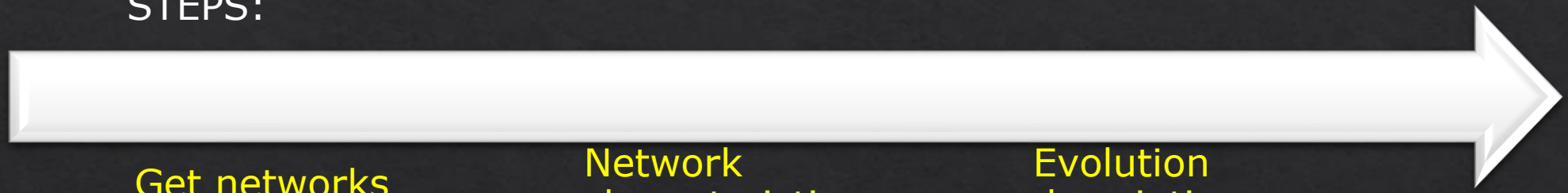




Chinese Function Characters/Words Evolution

Chinese Function Characters/Words Evolution

STEPS:



Get networks

- 4 periods
  - Ancient Chinese
  - Middle Ancient Times Chinese
  - Modern Times Chinese
  - Modern Chinese
- 2 single-character words
  - 在 zai ‘to exist, be living, to stay or remain, (to be located) in, at, ’
  - 人 ren ‘people’

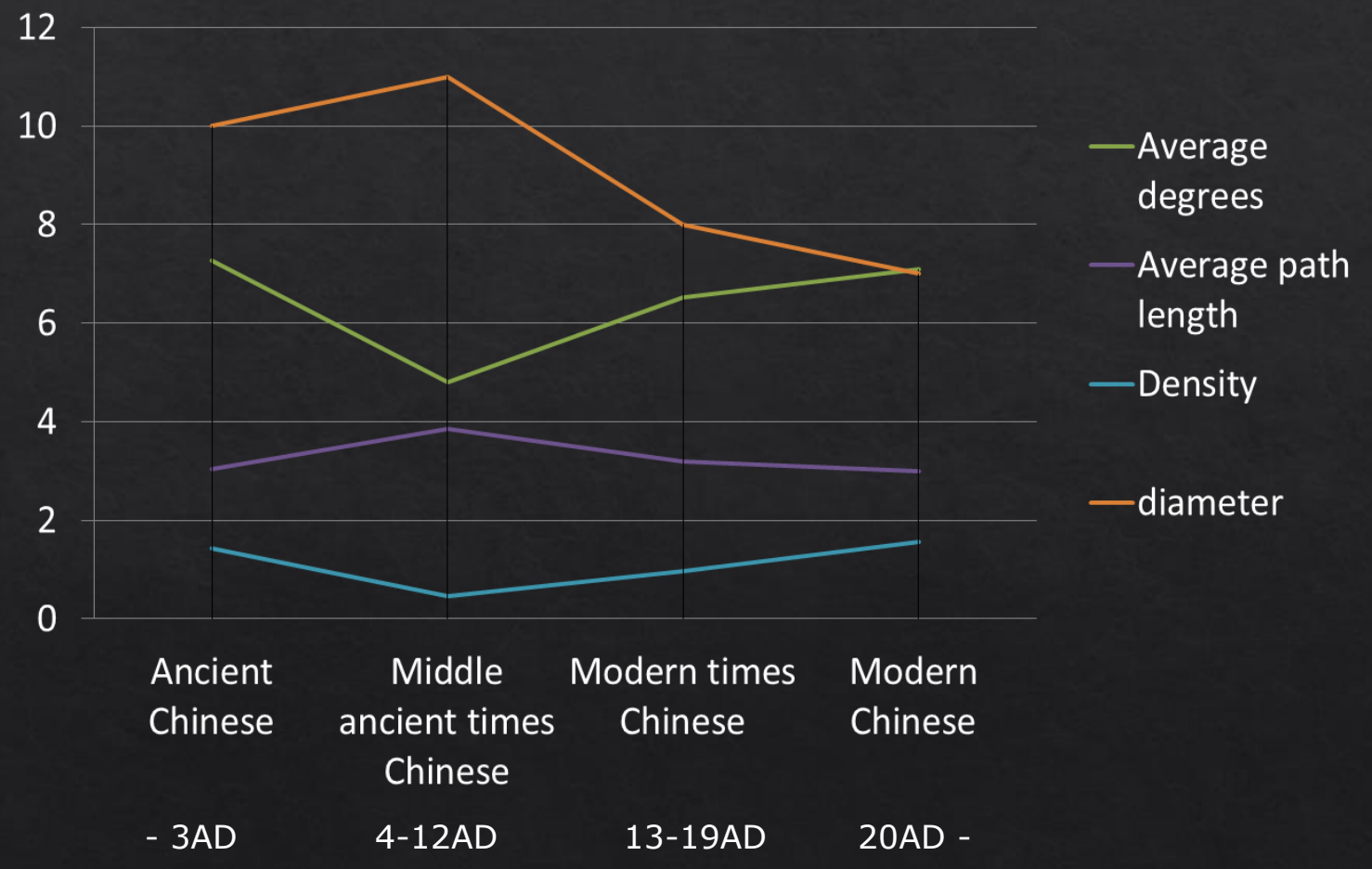
Network characteristics

- average degrees
- average path length
- density
- diameter
- degrees of ‘zai’
- degrees of ‘ren’

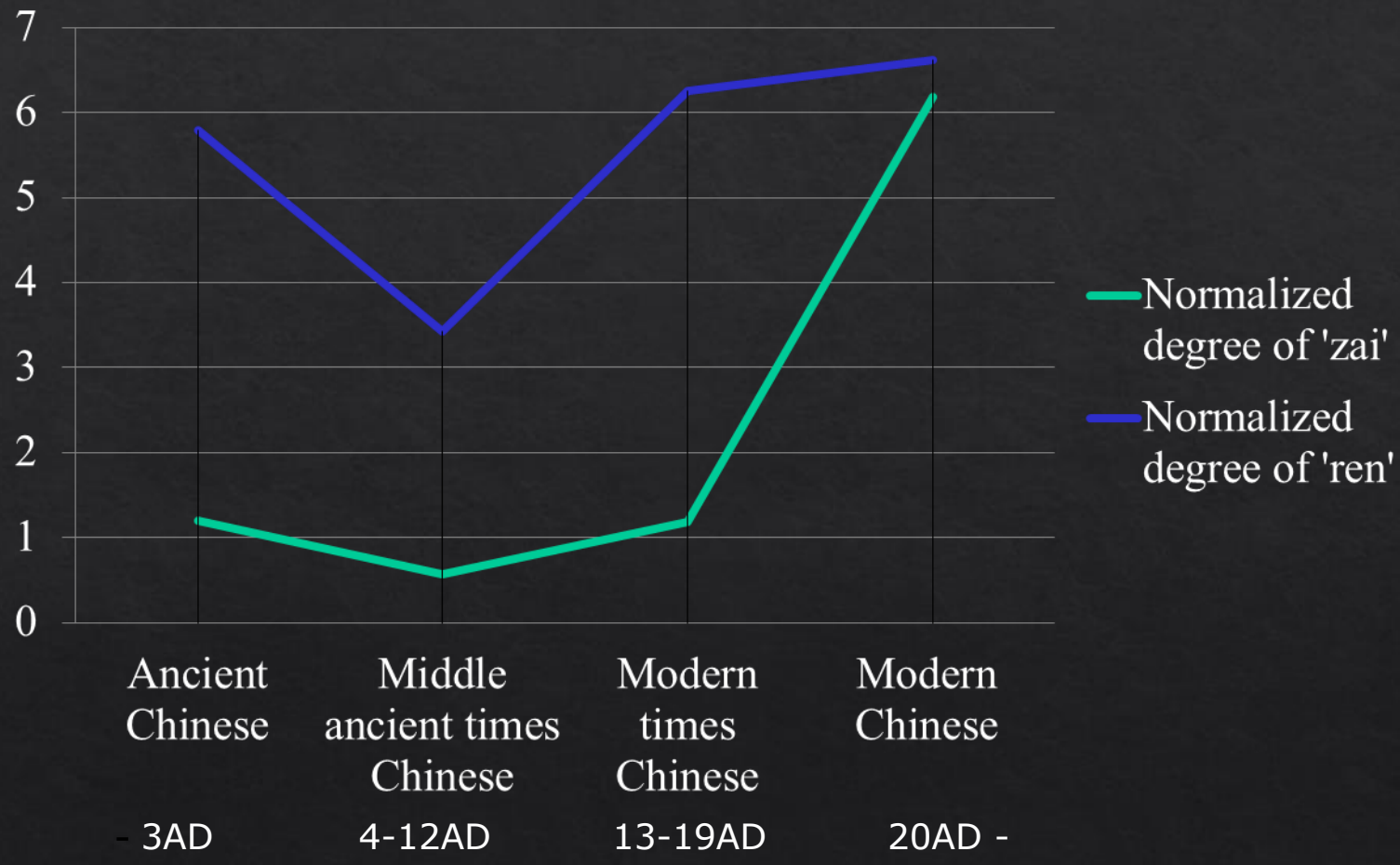
Evolution description

- quantitatively describe the evolution process of Chinese in 4 periods
- quantitatively describe the evolution process of 2 single-character words of Chinese in 4 periods

Co-occurrence Character Networks



Comparison of Single-character words



Comparison of Single-character words

Comparison of Single-character words

- Chinese is an isolating language: syntactic structure relies primarily on function words and word order rather than on rich morphological information to encode functional relations between elements (Levy & Manning 2003).

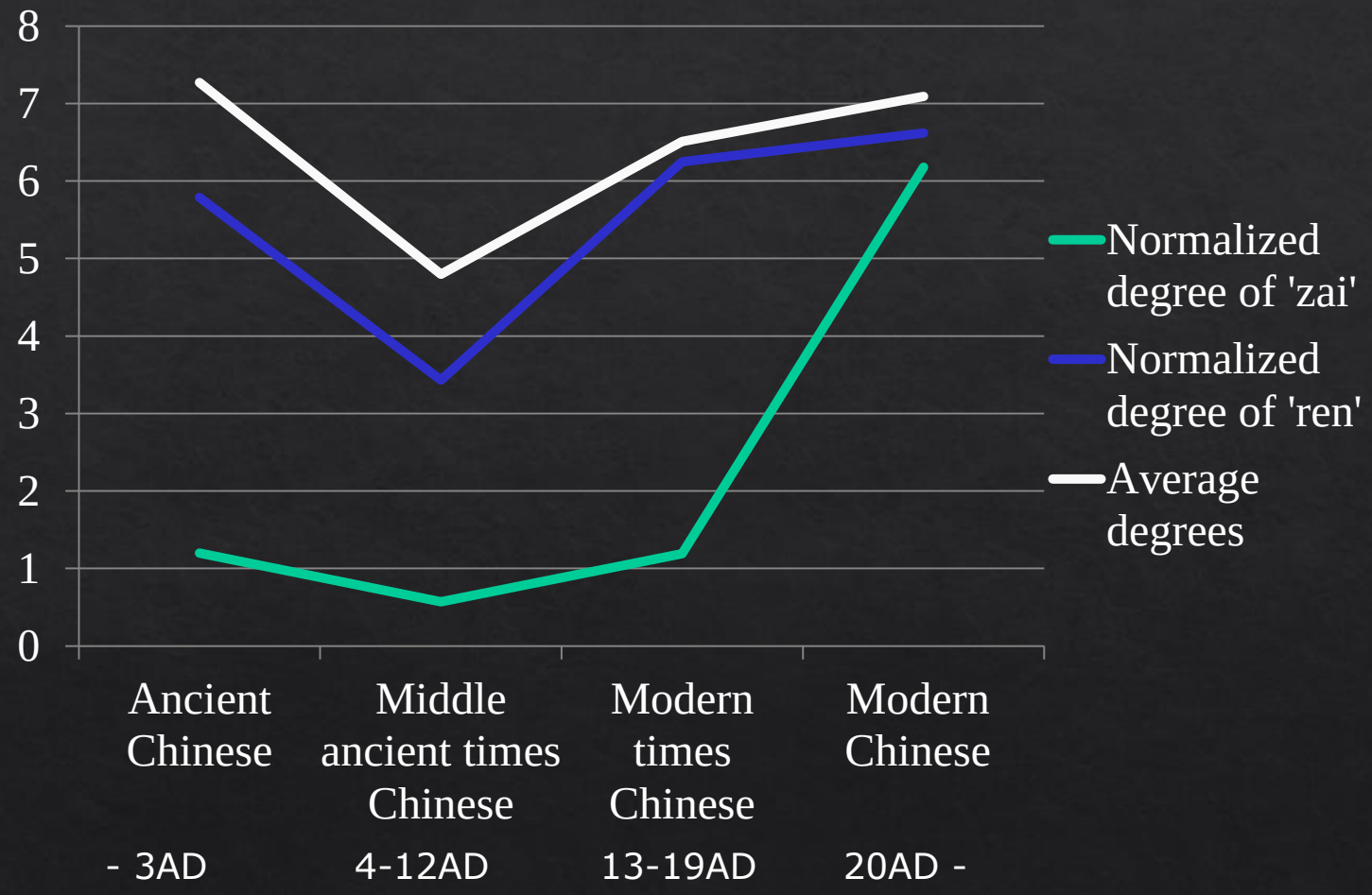
Comparison of Single-character words

- Chinese is an isolating language: syntactic structure relies primarily on function words and word order rather than on rich morphological information to encode functional relations between elements (Levy & Manning 2003).
- 2 single-character words
  - Frequent characters/words both in the corpus and in general
  - zai underwent a grammaticalization process, whereas ren remains a content word

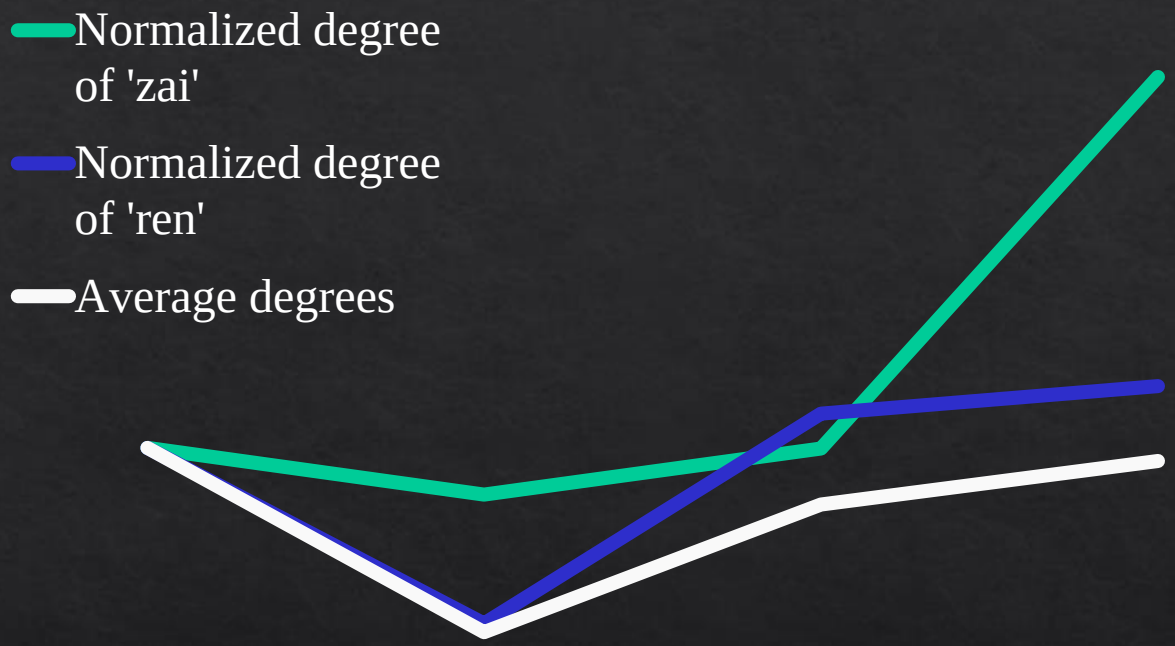
Comparison of Single-character words

- Chinese is an isolating language: syntactic structure relies primarily on function words and word order rather than on rich morphological information to encode functional relations between elements (Levy & Manning 2003).
- 2 single-character words
  - Frequent characters/words both in the corpus and in general
  - zai underwent a grammaticalization process, whereas ren remains a content word
- Hubs could indicate the grammaticalization process and its starting points. Hubs could be functional or potential functional words to undergo future grammaticalization. (Solé et al. 2002)

Grammaticalization Process in Networks



Grammaticalization Process in Networks



Conclusion

- The network features offer a new source of information to distinguish evolutions of different characters
- Observing evolution process in networks allows us to consider the evolutions of some specific units and the evolution of the whole system simultaneously.

Questions & Comments

Thank you ！