

# A Statistical Analysis of Chinese Compounds:

## Power-law Distribution and Morphological Productivity

CHEN Chao-Jan  
National Chi Nan University

International Workshop on Grammar & Evidence  
April 14, 2007

# Compound Words in Chinese

- What is a Chinese word?  
Is a character a word or just a morpheme?
- What is a compound in Chinese?  
Two-character word as a compound?

# Data Source of Compounds

All the 2-character words in the Chinese corpus **ASBC** (Academia Sinica Balanced Corpus 中央研究院平衡語料庫)  
/ version 3.0 (size: **5 million** words)

Total: **5486** characters à **66722** compounds

# Two-character Compounds

- five major constructions for compounds

Let **X-Y** be a compound,

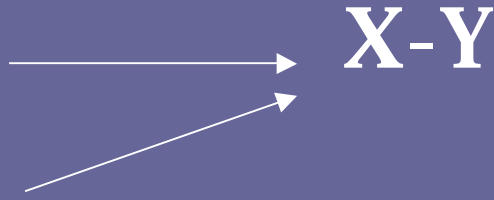
**(X,Y)** can be in one of the following grammatical relations:

- (1) 偏正 (modifier-head)
- (2) 並列 (coordination)
- (3) 主謂 (subject-predicate)
- (4) 述補 (verb-complement)
- (5) 述賓 (verb-object)

# Creation of Compounds : by rules or by examples?

- Rule-based creation:

- **semantic** constraints
- **syntactic** constraints  
(arguments,  $\theta$ -roles, ... ,etc.)



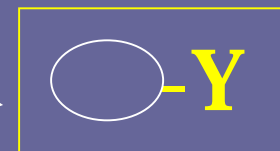
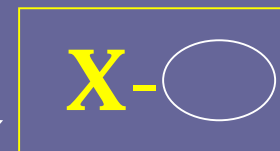
A diagram illustrating rule-based creation. Two arrows point from the text 'semantic constraints' and 'syntactic constraints' to the compound 'X-Y'. The arrow from 'semantic constraints' is a solid line, and the arrow from 'syntactic constraints' is a dashed line.

- Example-based creation:

- compounding **templates**

**X-A, X-B, X-C,...**

**P-Y, Q-Y, R-Y, ...**



**X-Y**

# Analogical Creation

e.g. 國手 ('nation hand') à 國腳 ('nation foot')

rule-based (by composition):

sense (國腳) = sense (國) + sense (腳) ?

example-based:

國手 : 國腳 = 手 : 腳 (the analogy formula)

sense(國手) : ? = sense(手) : sense(腳)

國-X as a template

sense(國手) : sense(國-X) = sense(手) : sense(X)

Probability(creation) ~ Productivity(template)

Productivity(template) ~ #type(examples)

# Character's Productivity in Compounding

- Why are some characters more productive in forming compounds?
- Ranging from 1 (1070 char) to 607 (1 char)  
à not a fair world at all !  
e.g. 死 occurs in 176 compounds,  
while 斃 only in 12 compounds
- Why can the difference be so huge?

# Character's Productivity: Data from ASBC

Top 10

大	607
出	553
子	513
上	482
下	418
人	415
到	392
成	383
水	340
開	316

some examples

打	201
擊	69
死	176
斃	12
破	106
裂	34
毀	38
壞	68
嘲	6
呻	1

# A Simple Statistical Analysis

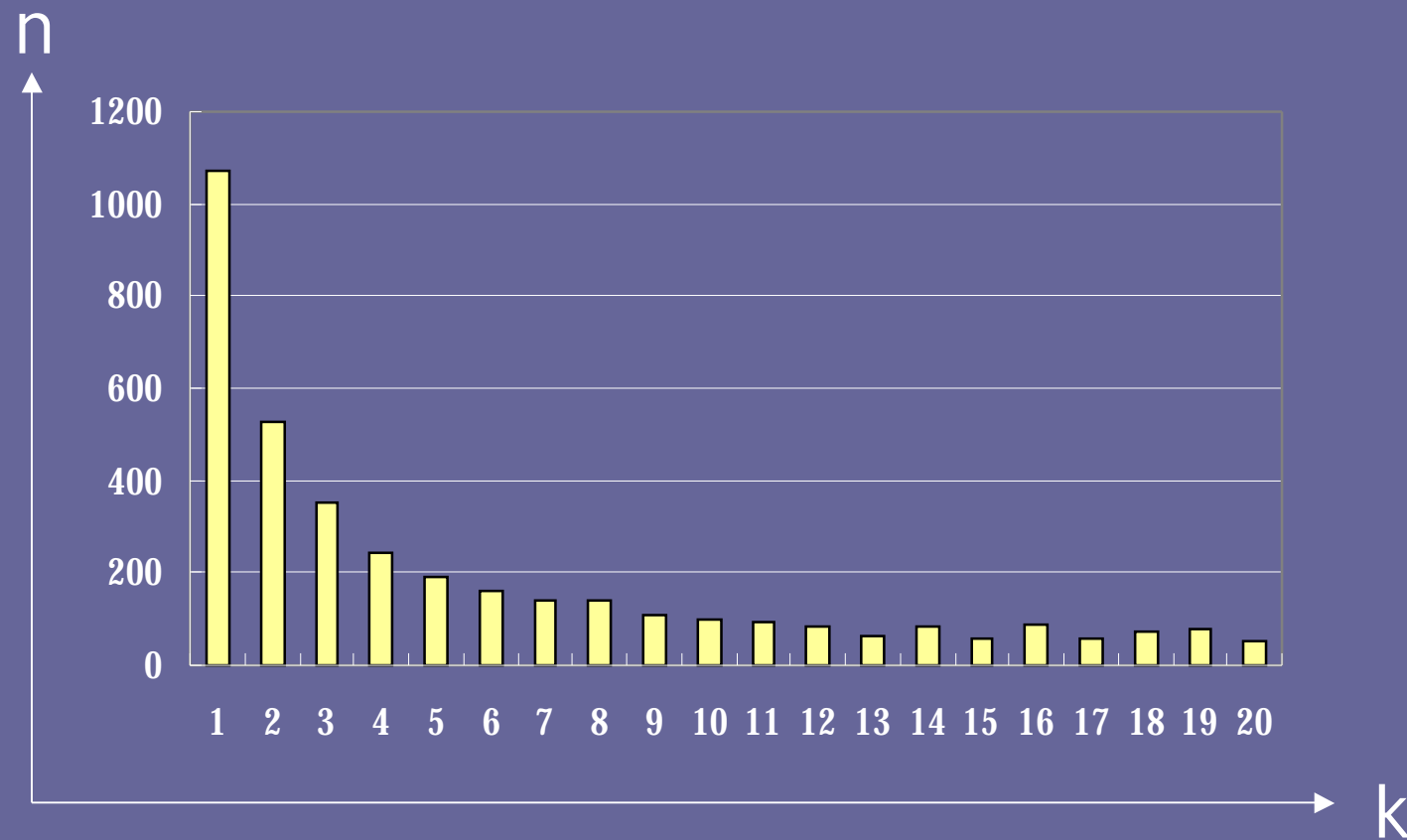
**n** = the number of characters that occur in **k** compounds  
total: #compounds = 66722, #characters = 5486 (in types)

k	n
1	1070
2	527
3	351
4	244
5	191
6	162
7	142
8	142
9	108
10	99

11	93
12	83
13	64
14	84
15	56
16	87
17	59
18	70
19	77
20	52
..	.....

# Distribution Pattern (I)

- In histogram: the mean  $\langle k \rangle = 12.16$

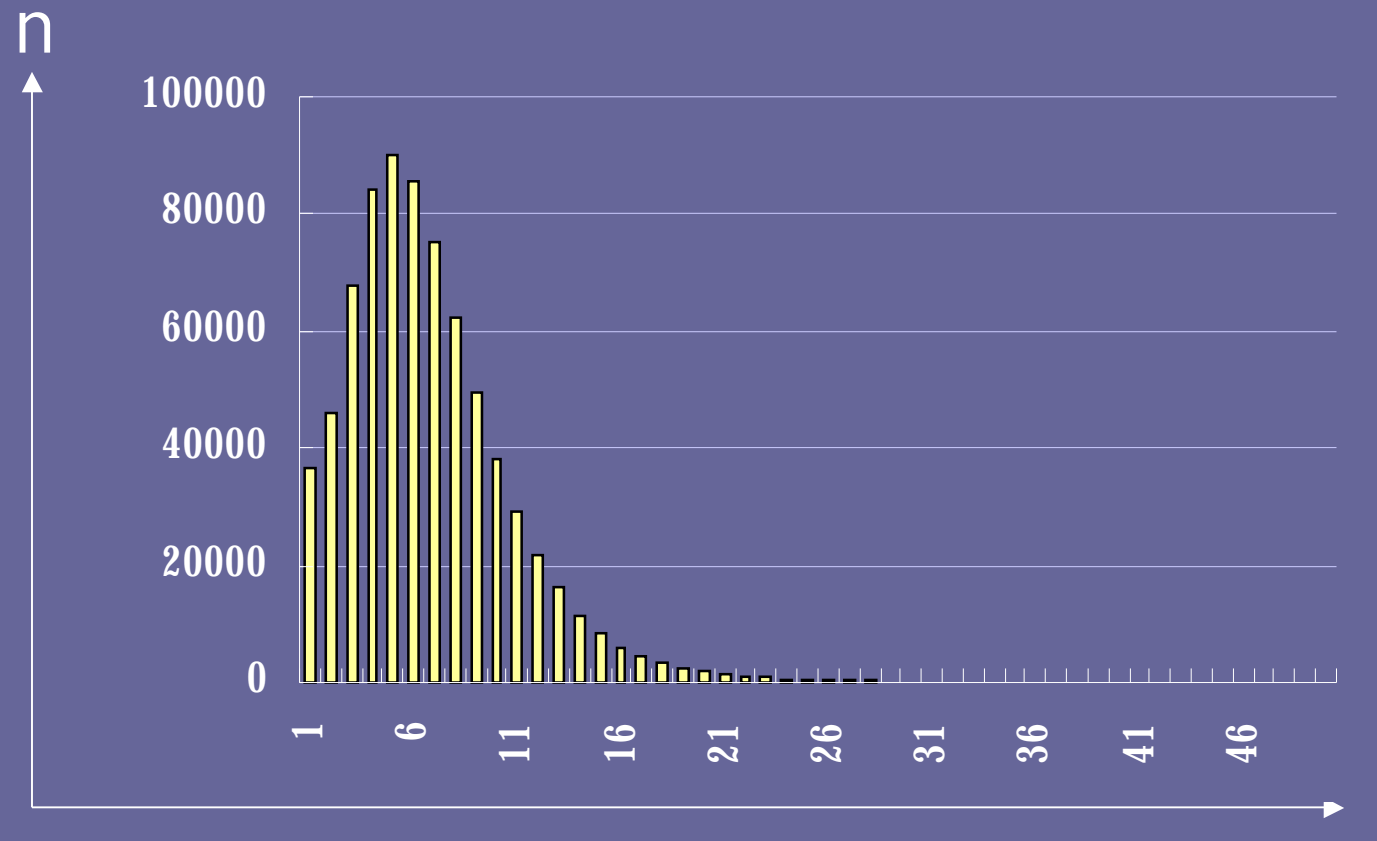


# Q: Why is it not a fair world?

- As  $\langle k \rangle = 12.16$  , why is the mode of the distribution not around 12.16? (as in a Poisson or Gaussian distribution)  
( #compounds = 66722, #characters = 5486 )
- Why most of them are so “poor” (unproductive)?
- Why few of them are extremely “rich” (productive)?
- Why the world of morphological productivity is so “aristocratic” instead of “equalitarian”?

# Sentence Length Distribution

- in histogram: mean length  $\langle l \rangle = 6.79$



# Sentence Length Distribution

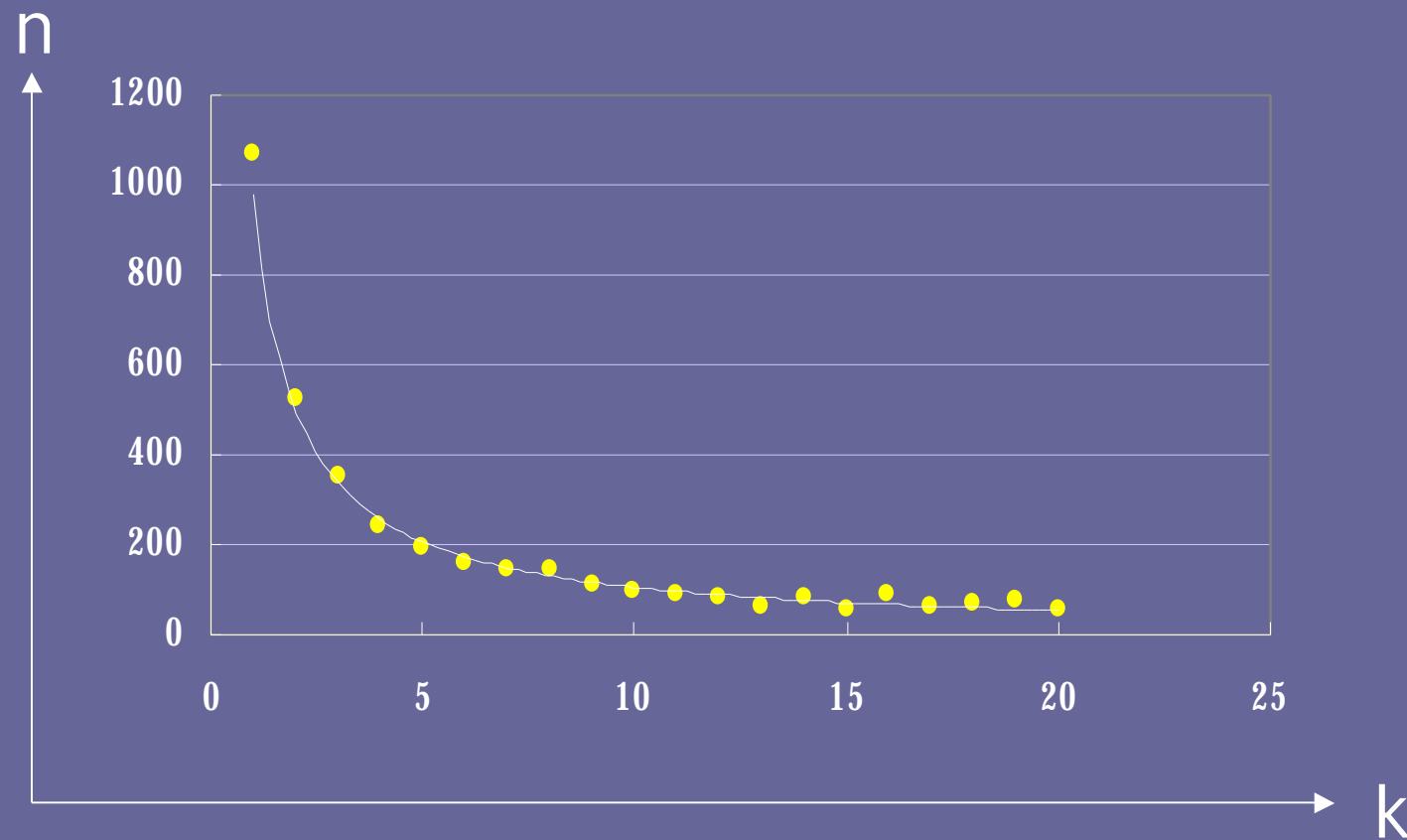
length / #sentence (748,693 sentences in total)

1	36804
2	45947
3	67745
4	83980
5	89894
6	85846
7	75163
8	62512
9	49629
10	38148

11	28969
12	21693
13	16093
14	11624
15	8526
16	6167
17	4692
18	3331
19	2496
20	1996
..	....

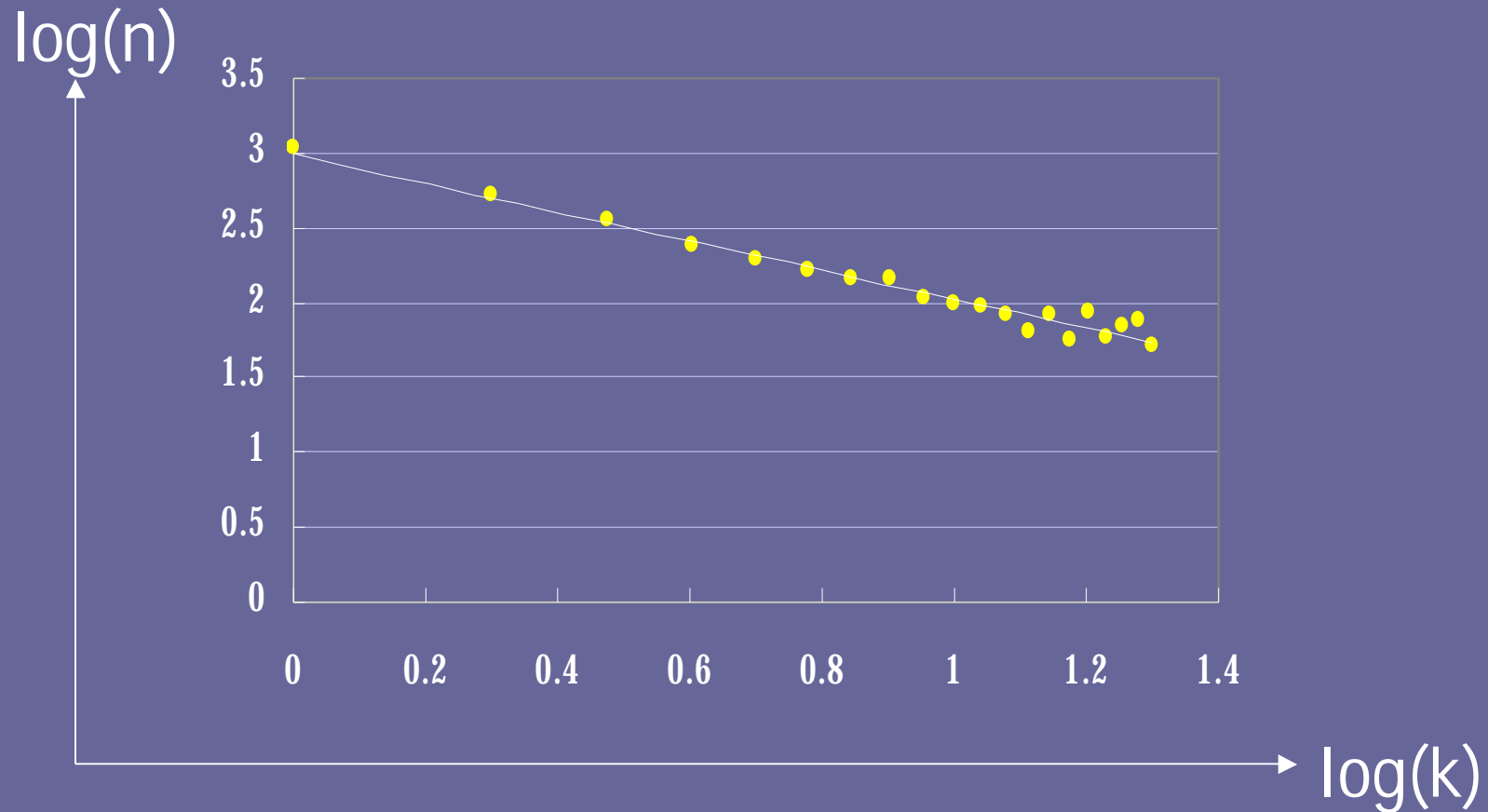
# Distribution Pattern (II)

- plot in ordinary coordinates:



# Distribution Pattern (III)

- plot in log-log coordinates:



“Wow, a power-law distribution !”

# The Best Fitting Curve

$$\log(N(k)) = 3.028 - 1.003 \log(k)$$

by least-square regression

( with  $r = -0.9846$ ,  $r^2 = 0.9694$  )

$$\Rightarrow N(k) = 1066 k^{-1.003} \quad ( N(k) \sim 1/k^{1.003} )$$

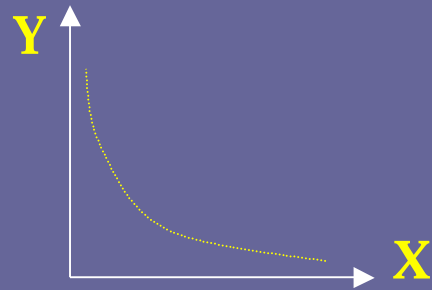
$N(k)$  : the number of nodes with  $k$  links

$$N(\mathbf{k}) = C \mathbf{k}^{-\alpha}$$

	<i>#Comp</i>	<i>#Char</i>	<i>C</i>	$\alpha$	$r^2$
<b>Total</b>	66772	5686	1066	-1.003	0.9694
<b>V/</b>	29316	3820	879	-1.012	0.9867
VC	10431	2088	299	-0.886	0.9655
VA	6107	2566	110	-0.660	0.9684
VH	8028	2566	153	-0.716	0.9712
<b>N/</b>	36513	4449	538	-0.893	0.9772
Na	25078	3855	437	-0.870	0.9807
Nb	5808	1969	90	-0.653	0.9629

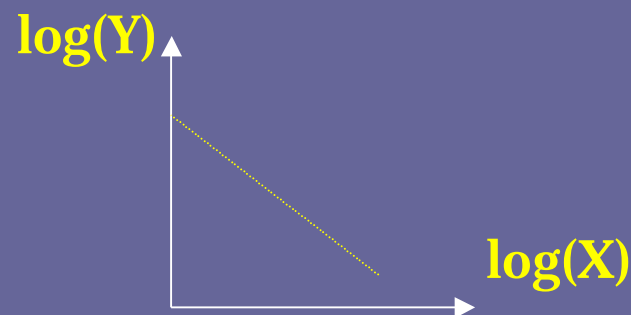
# What is a power-law distribution?

when  $Y = C X^{-\alpha}$ , we have a power-law distribution



as  $\log(Y) = \log(C) - \alpha \log(X)$ ,

we can see a straight line with a slope of  $-\alpha$  in a log-log plot



# Power-law Distributions in our Worlds

- Power-law distributions have been found in real or virtual networks such as

WWW sites

Movie actors

Coauthors

Citation

Food chain web

Airports

Synonyms

these networks are called  
“scale-free”

their #node **N** is a function  
of #link **k** as follows:

$$N(k) = C k^{-\alpha}$$

see also Barabasi (2002), Buchanan (2002), Barabasi et al. (2004),  
Steyvers and Tenenbaum (2005), etc.

# What can power-law distributions tell us about network structures?

It is regarded as the signature of

- (1) incremental growth  
of nodes and links
- (2) preferential attachment  
à a “rich-get-richer” effect

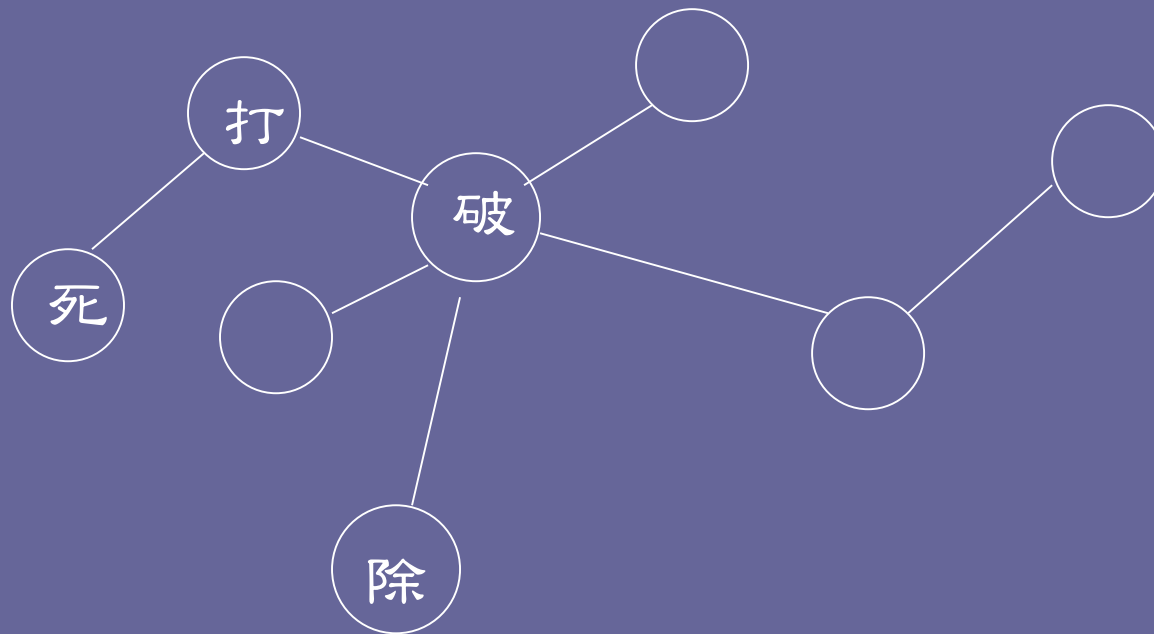
Barabasi & Albert (1999)

# Why do the characters observe a power-law distribution?

- (1) compounding as a link
  - à a compound **X-Y** à a link between **X** and **Y**
- (2) compounding connectivity forms a network
  - à characters as nodes (vertices)
  - à compounding relations as links (edges)

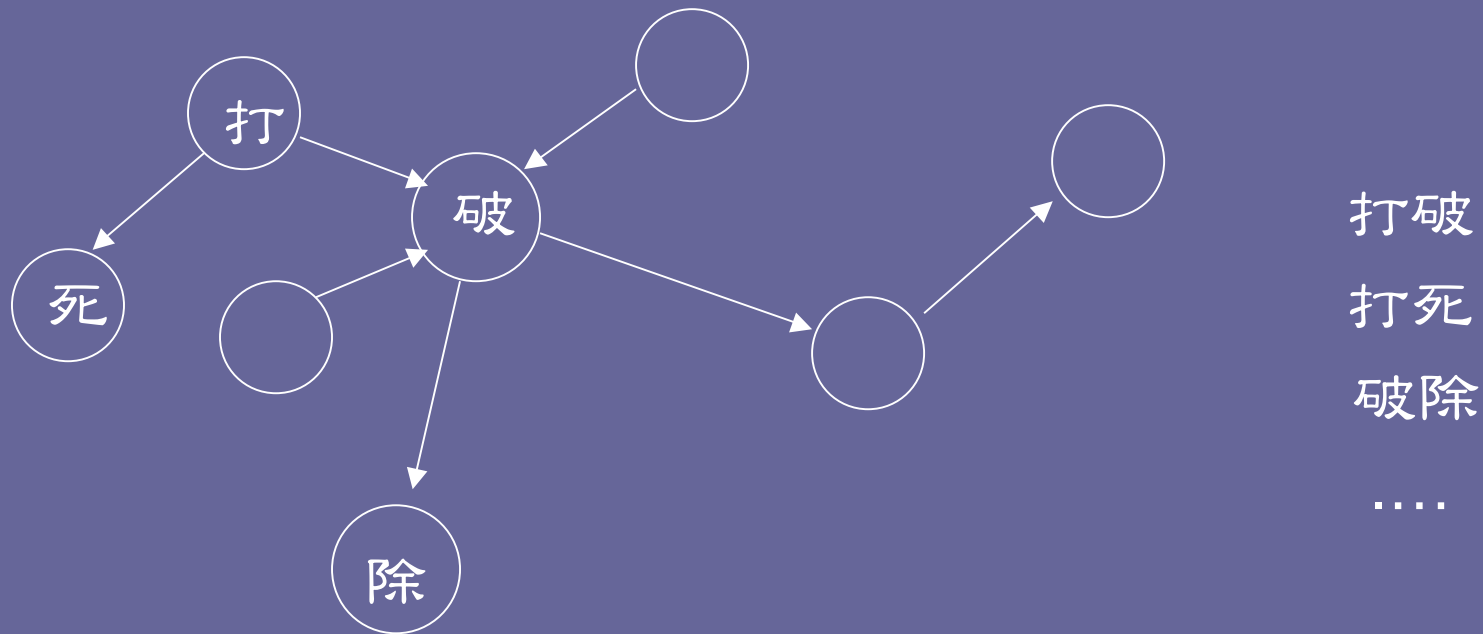
see also Chen (2005 : 92-96)

# Character Network (I): compounding as connectivity



打破  
打死  
破除  
....

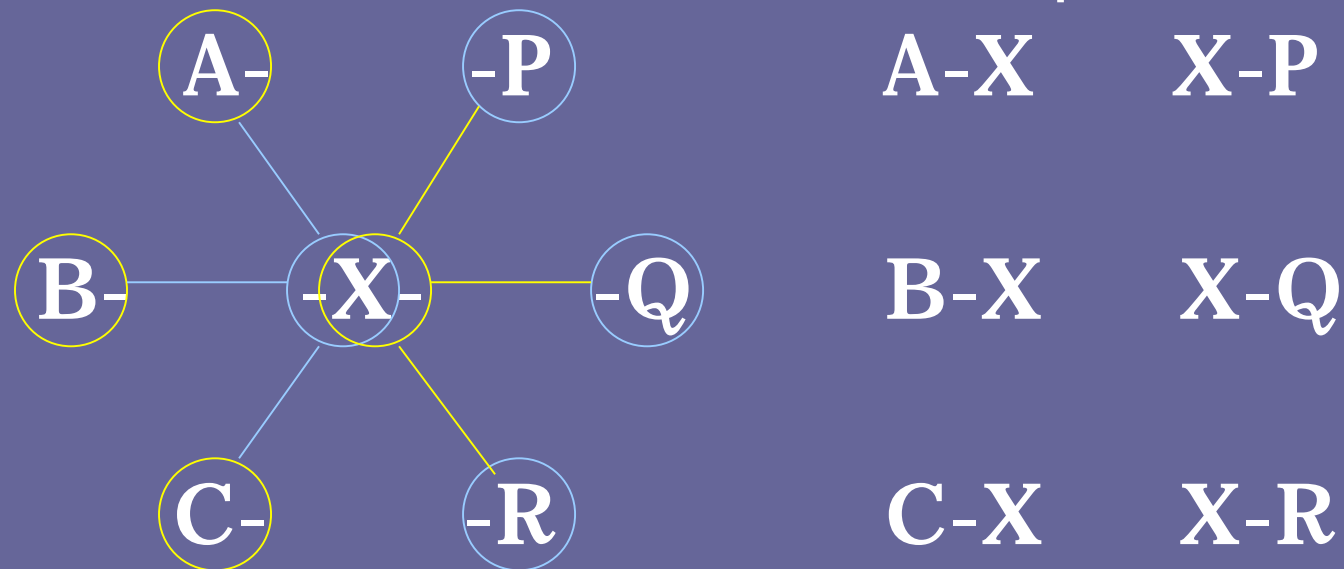
# Character Network (II): compounding as connectivity



**X-Y: X à Y** directed links

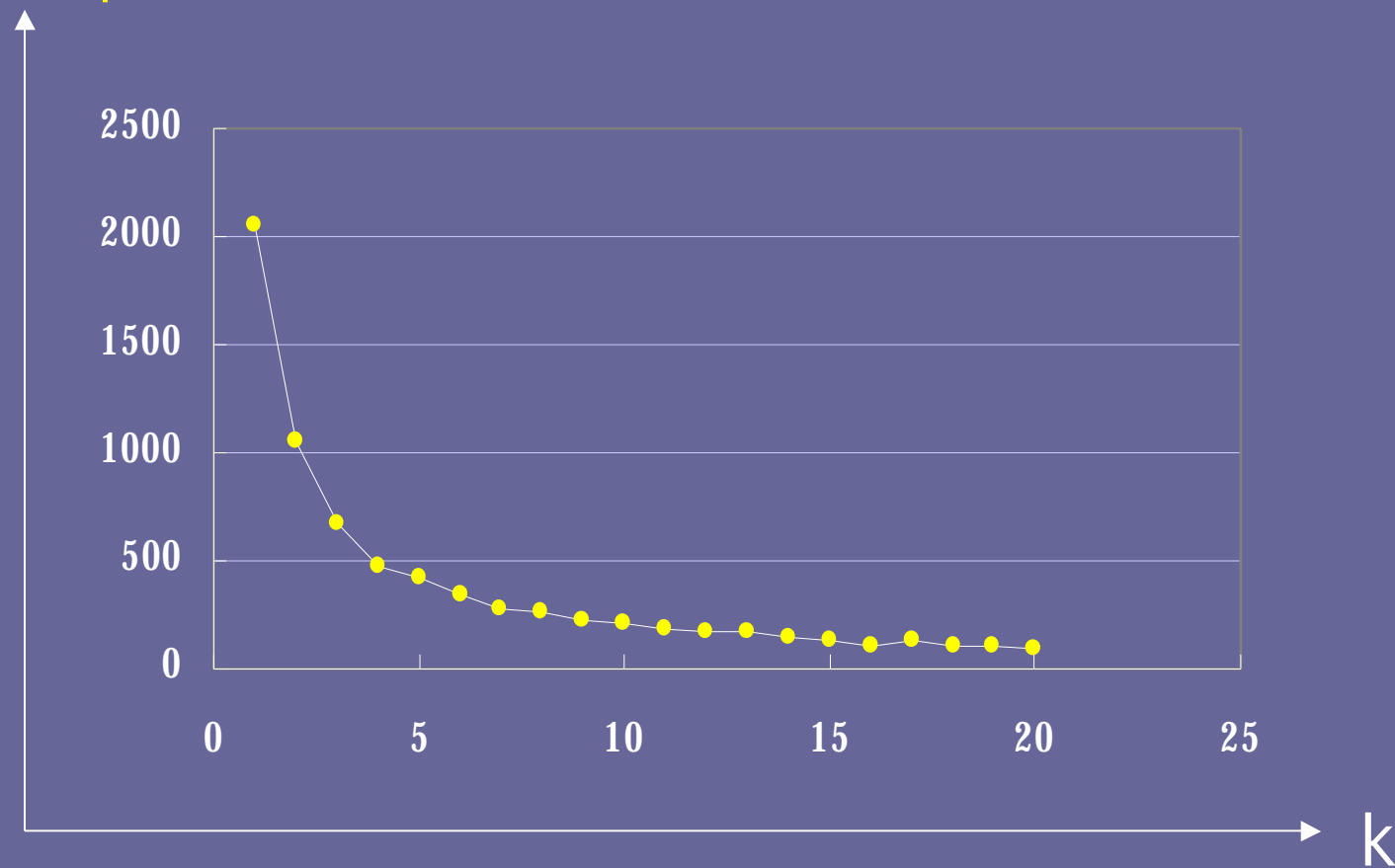
# More Statistics

- distinguishing heads (**X-**) and tails (**-X**)
  - à the distribution still follows power law



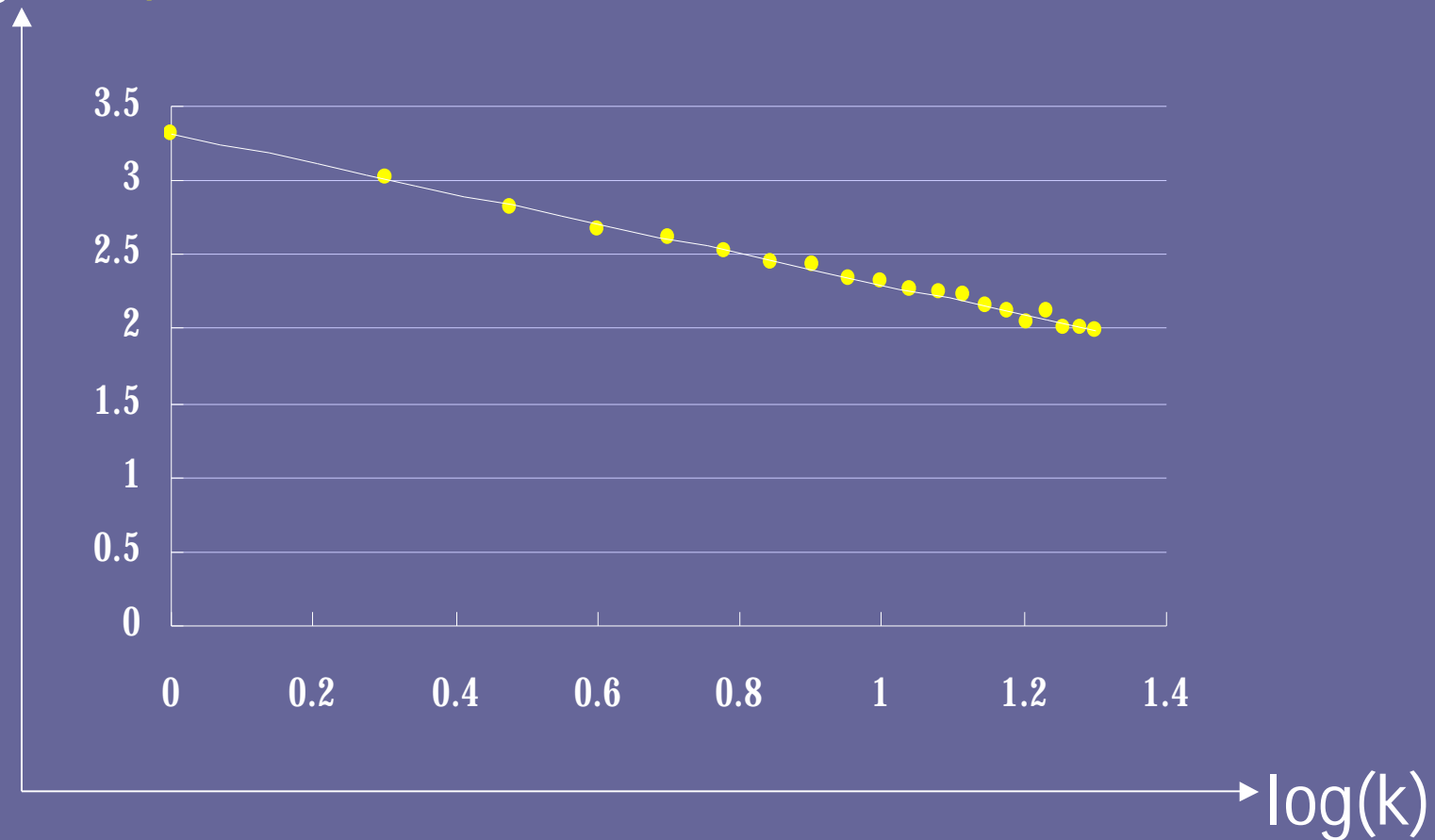
# Distribution Pattern : distinguishing heads and tails (I)

n p-l distribution:  $N(k) = 1790 k^{-0.979}$  ( $r^2 = 0.9693$ )



# Distribution Pattern : distinguishing heads and tails (II)

$\log(n)$  p-I distribution:  $N(k) = 1790 k^{-0.979}$  ( $r^2 = 0.9693$ )

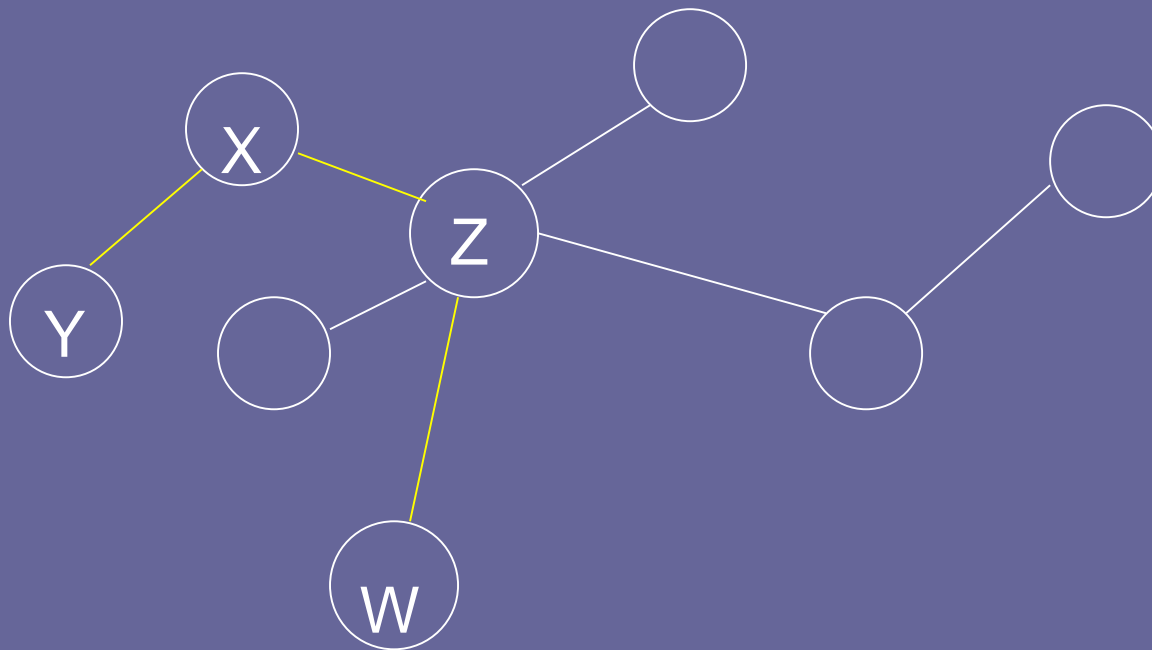


# Preferential Attachment in our Networks

- different from the Web network:
  - à growth of links, but no growth of nodes
- preferential attachment
  - à the more used characters get more productive in compounding (“rich get richer” )

# A Preliminary Stochastic Model: Compounding as Link

random or preferential attachment (linking)



X-Y

X-Z

Z-W

....

# A Preliminary Stochastic Model

Prob (**X-Y**)  $\sim$  Prob(**X**) Prob(**Y**)

for preferential attachment

à Prob(**X**)  $\sim$  #example(**X**)

Prob(**Y**)  $\sim$  #example(**Y**)

for random attachment

Prob(**X**) = Prob(**Y**) = **p** (**1/n**, **n=#char**)

#example(**X**) is the number of compound types that contain X

Prob (**X-Y**) is the probability of a pair **X,Y** getting linked

Prob (**X**) is the probability of **X** getting chosen to be linked

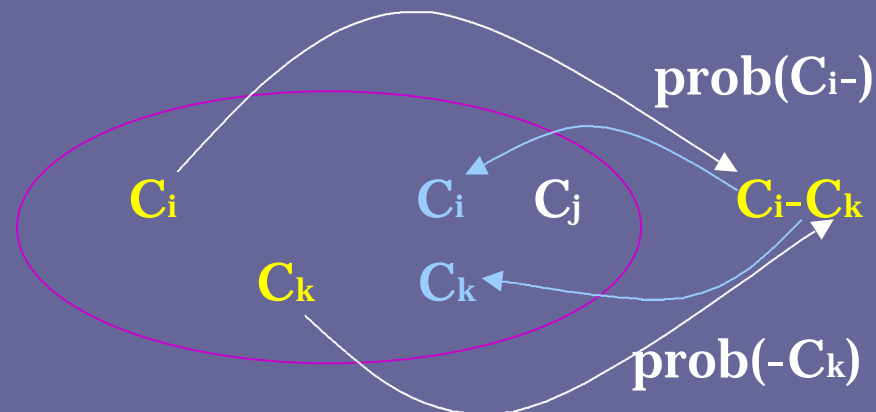
# A Preliminary Stochastic Model

random attachment:

$$\text{Prob}(C_i) = 1 / n \quad (n = \# \text{original char})$$

preferential attachment

$$\text{Prob}(C_i) = \text{freq}(C_i) / N \quad (N = \# \text{char in time } t)$$



# Simulation Experiments

- Experiment design:

Exp1: **Random Attachment** Model:

5000 characters

à to form 10000 compounds (by **R**-attachment)

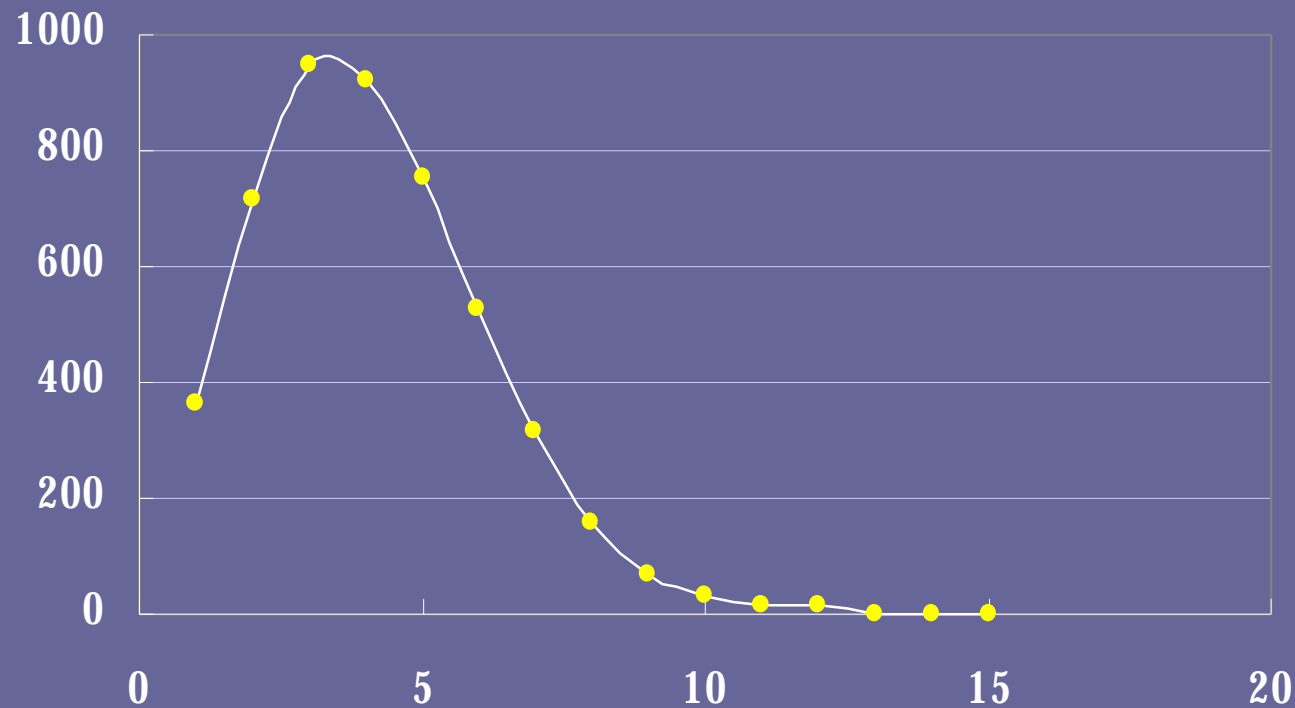
Exp2: **Preferential Attachment** Model:

5000 characters

à to form 10000 compounds (by **P**-attachment)

# Results of the Experiments (I)

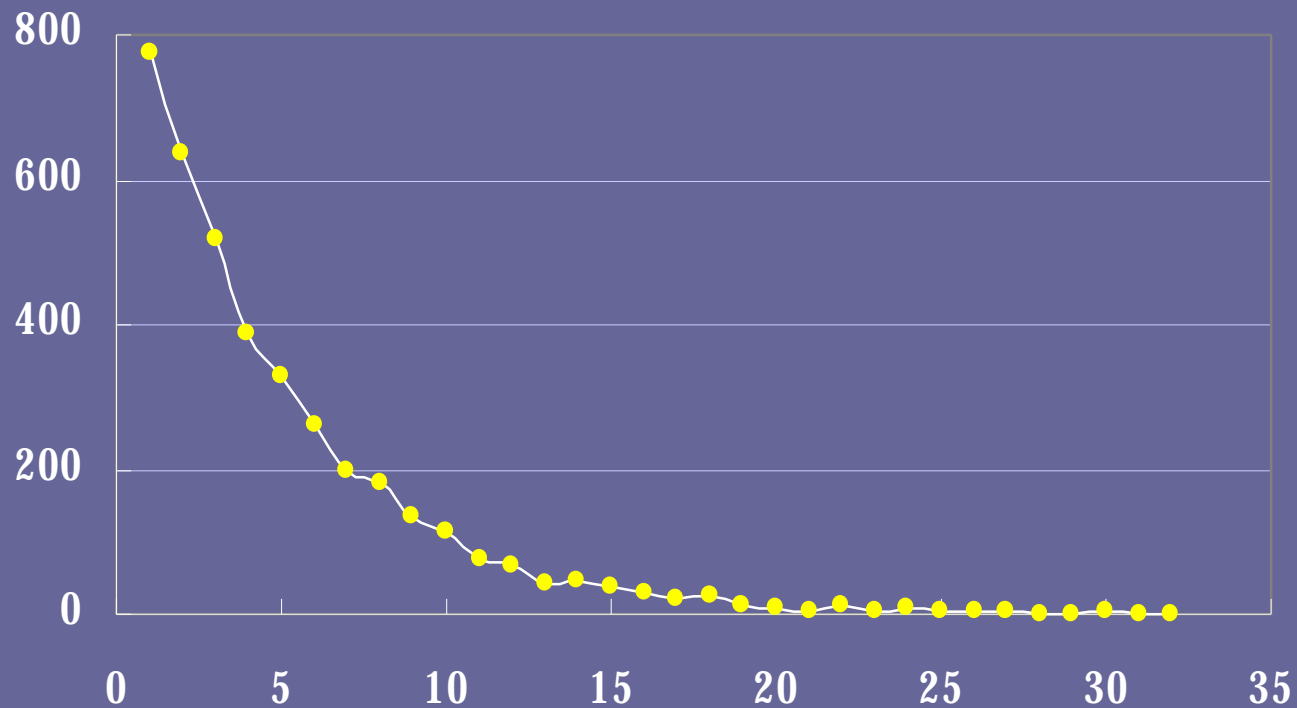
Degree Distribution for the **Random Attachment** Model:



a **Poisson** distribution (without distinguishing heads/tails)

# Results of the Experiments (II)

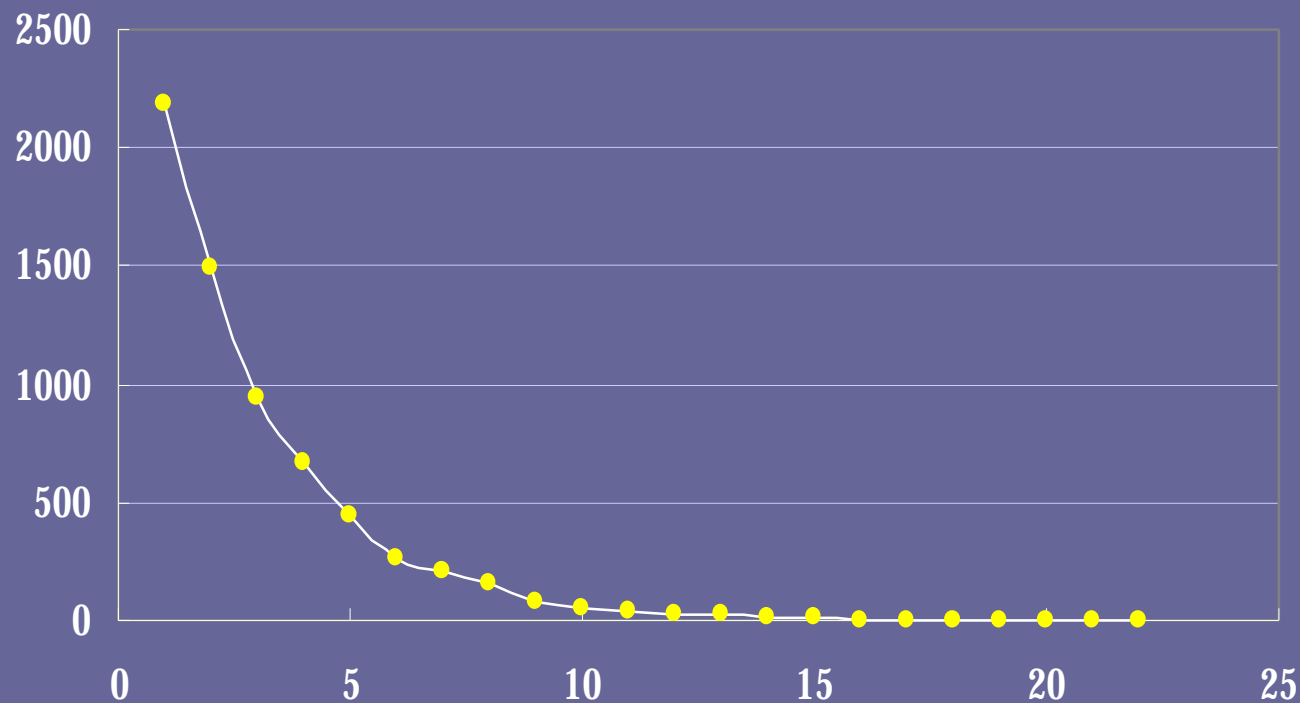
Degree Distribution for the Preferential Attachment Model:



a power-law distribution (without distinguishing heads/tails)

# Results of the Experiments (III)

Degree Distribution for the Preferential Attachment Model:



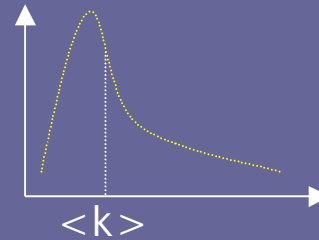
a power-law distribution (distinguishing heads/tails)

# Rule-governed & Poisson Distribution (I)

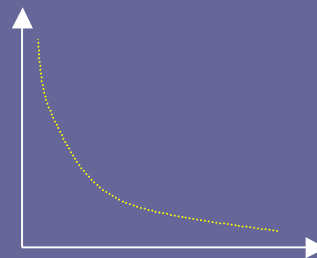
- If compounding is purely rule-governed ...  
suppose that every rule-licensed pair of compound components **X, Y** has the same probability  $p$  of getting compounded,  
then with  $n$  compounds formed, the character connectivity is supposed to follow a Poisson distribution with mean  $\langle k \rangle$   
( $\langle k \rangle = n p$ )

# Poisson Distribution or Power-law Distribution?

- If purely rule-governed ...
  - à what kind of productivity distribution will we see? ... a Poisson distribution



- But actually... .
  - we see a power-law distribution... .



# Two Factors: Templates & Rules

in a **hybrid** model combining **preferential** attachment and **random** attachment,

5000 characters  $\rightarrow$  form 10000 compounds

Exp-n:  $n = 0 \rightarrow 10$

each time create 100 compounds

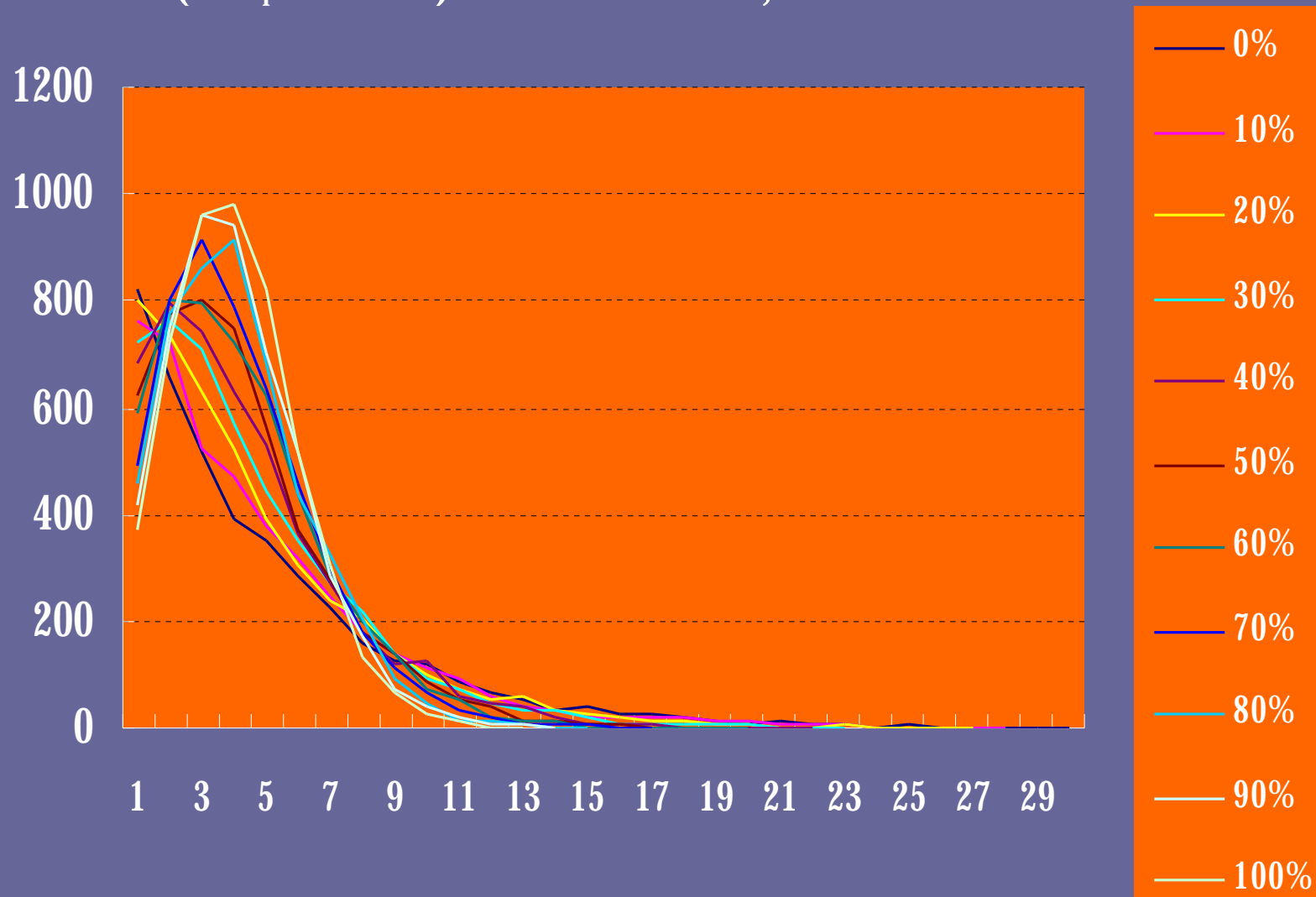
$n \cdot 10\%$  among them are created by

random attachment; the rest of them

are created by preferential attachment

# Two Factors: Templates & Rules

When  $f_{\text{Rule}} / (f_{\text{Template}} + f_{\text{Rule}})$ : 0% à 100% , Power-law à Poisson



# Which factor keeps the shape of a power-law distribution?

- What's the moral of the previous plot?
- Template-based creation must play a more important role in compounding mechanism than rule-based creation.

The power-law distribution cannot be accounted by a pure rule-governed creation mechanism.

# Analogical Creation & “Richer-get-richer” Effect

a plausible linguistic explanation for the power-law distribution:

analogical creation based on existent templates

à source of templates: compound types

# productivity(X-)  $\sim$  #type (X-C<sub>i</sub>)

à preferential attachment Model

# Concluding Remarks

A corpus-based empirical analysis provide an evidence supporting the **preference** of a **template-based** mechanism in compound creation over a pure rule-governed one... ..

à morphological productivity can be explained as the result of **positive feedback** in the history of the evolving lexicon of compounds

The features of the character compounding network in Chinese is to be explored in the future research... ..

It's probably a **Small World!**

The End