

# Extracting multiword units in China's English news articles

a presentation at the BAAL-OTA workshop  
Computing Services, Oxford University  
April 21, 2005

Li Wenzhong  
Faculty of International Studies, Henan Normal  
University, P.R. China  
&  
RDUES, The University of Central England

# 1 Introduction and rationale

- English in China must have its own unique contextualized features, both structural and functional, and such features are better captured and accounted for with quantitative investigations.

- “Language is phrasal and textual” (Scott, 1998: help).
- It is assumed that word clusters rather than single words in China English produce better illustrations on both linguistic and functional features.
- A word cluster is “a group of words which follow each other in a text” and they are words “which are found repeatedly in each other’s company. They represent a tighter relationship than collocates, more like groups or phrases” (Scott, 1998: help).
- Word clusters convey messages that are more complete and of easier access to identification and description, as they provide richer contextual information in a text.

## 2. Research questions

- What are the distributional patterns of n-word clusters in China's English news articles and in what aspects does the distribution characterizes China English texts in particular?
- To what extent are they unique in expressing things Chinese?

# 3. Methodology

- Source texts: *China Daily*, *Shanghai Star*, and *Beijing Weekend*;
- Year: 2002
- Construct:
  - over 2316 domestic news articles
  - 1, 281, 498 tokens and 33, 769 types
- The collection of texts is used as an observed corpus (referred to as China English News Article Corpus, shortened as CENAC).
- The reference corpus: a collection of British news articles (referred to as British English News Article Corpus, shortened as BENAC): 5, 683, 525 tokens and 88,969 types.

- Data processing:

--7 separate word cluster lists of 2 to 8 words are generated from the whole CENAC using *Wordsmith Tools v3.0* (Scott, 1998). In the lists each of the n-word cluster has been computed for its frequency and percentage ( $freq \geq 5$ ). The same procedures are applied to the BENAC texts so the same number of n-word cluster lists are generated

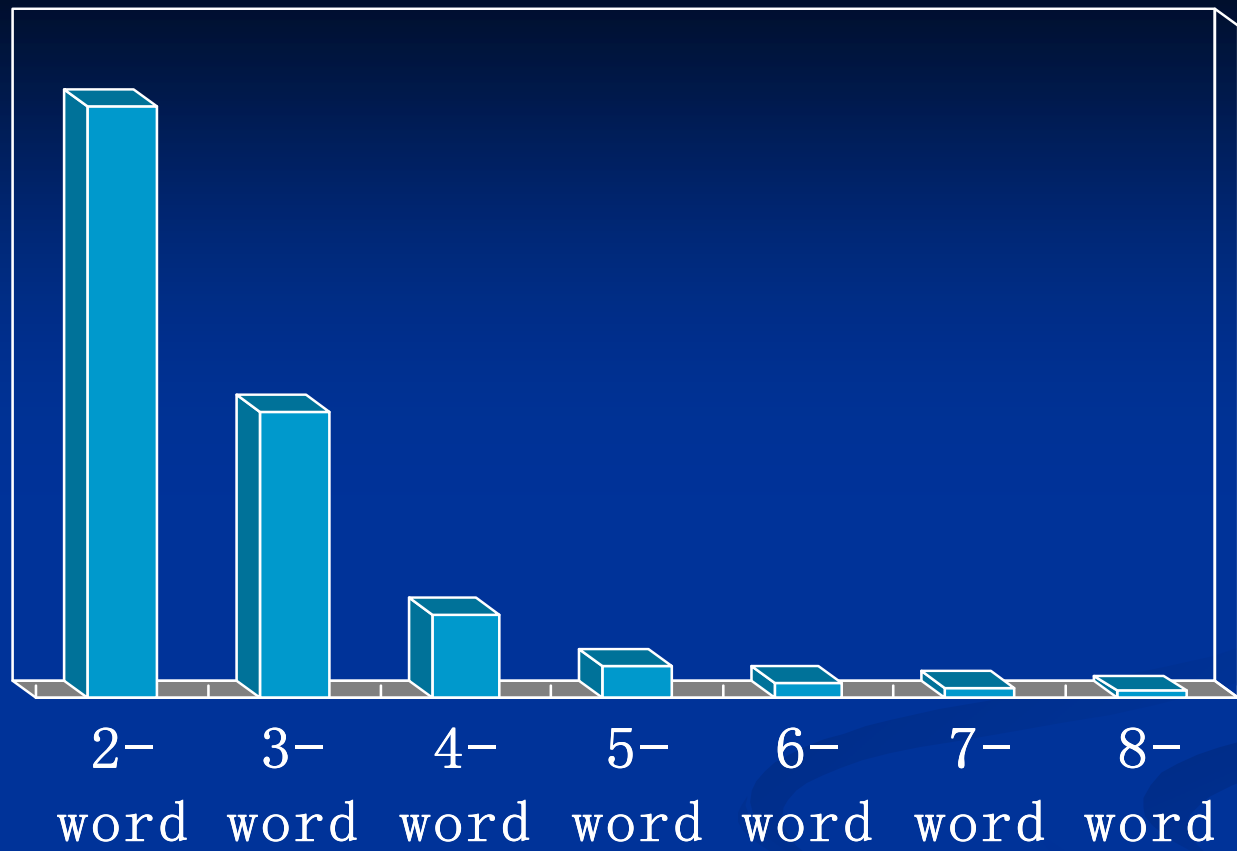
--to extract those clusters that are more frequently used in the CENAC, comparisons need to be made between the observed corpus and the reference corpus, so that those clusters have unusually high frequency of occurrence in the former can be picked out (see Li, 2003)

- The significance level is  $p < 0.0000001$ , and the extracted word clusters are regarded as unique multiword units in China English news articles
- Finally, the extracted word clusters are used as indexes for concordance analysis to evaluate and categorize them and examine their use in the context.

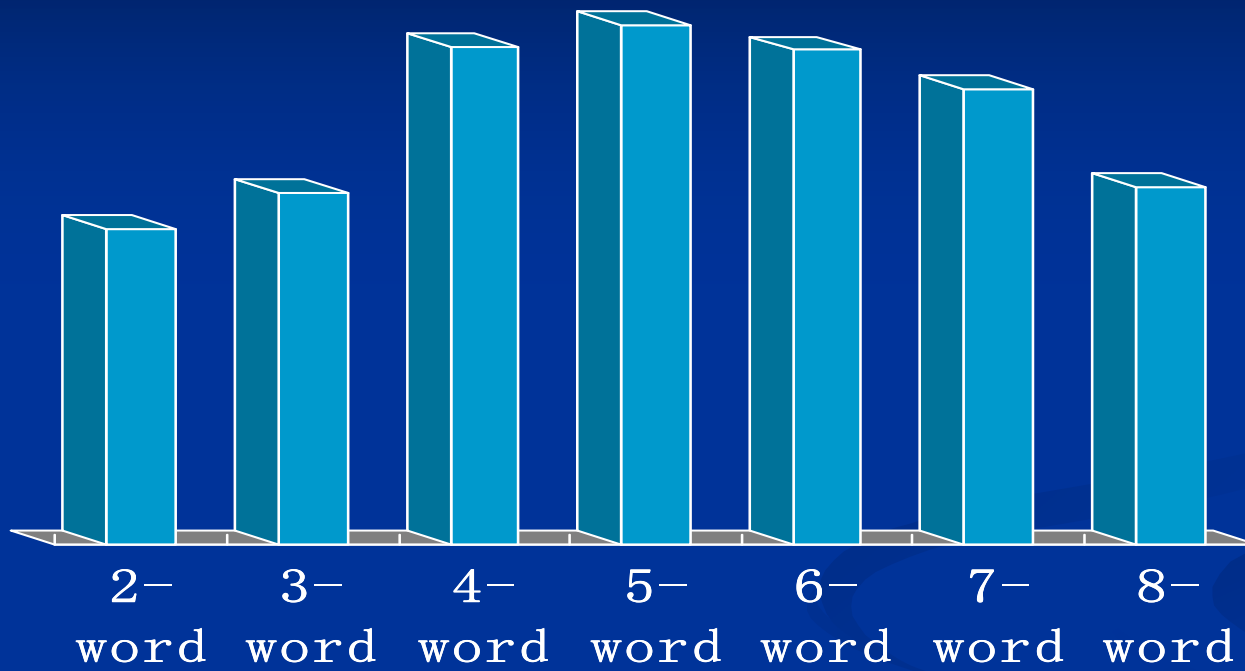
## 4. Findings and analysis

- 4.1 General features of n-word clusters
- To see to what extent each unique n-word clusters take up of the whole list, so that their the distribution can be revealed
- The recurrent word clusters in the CENAC:  
*f > 5, keyness > 25, p < 0.0000001*)

- An average of 15.54% of all the n-word clusters under investigation are found unusually frequent of their occurrence in China English news articles
- The 4, 5, and 6-word clusters are the most frequent ones (18.1%, 18.88%, and 18% respectively), and the 2-word clusters the lowest (12.95%)



Raw distribution



Unusually frequent clusters

- The statistics also shows that the longer the clusters, the greater specificity of meaning, the more fixed of the cluster patterns.

(?) a (well-off) society (?)

(build(ing)) a well-off society

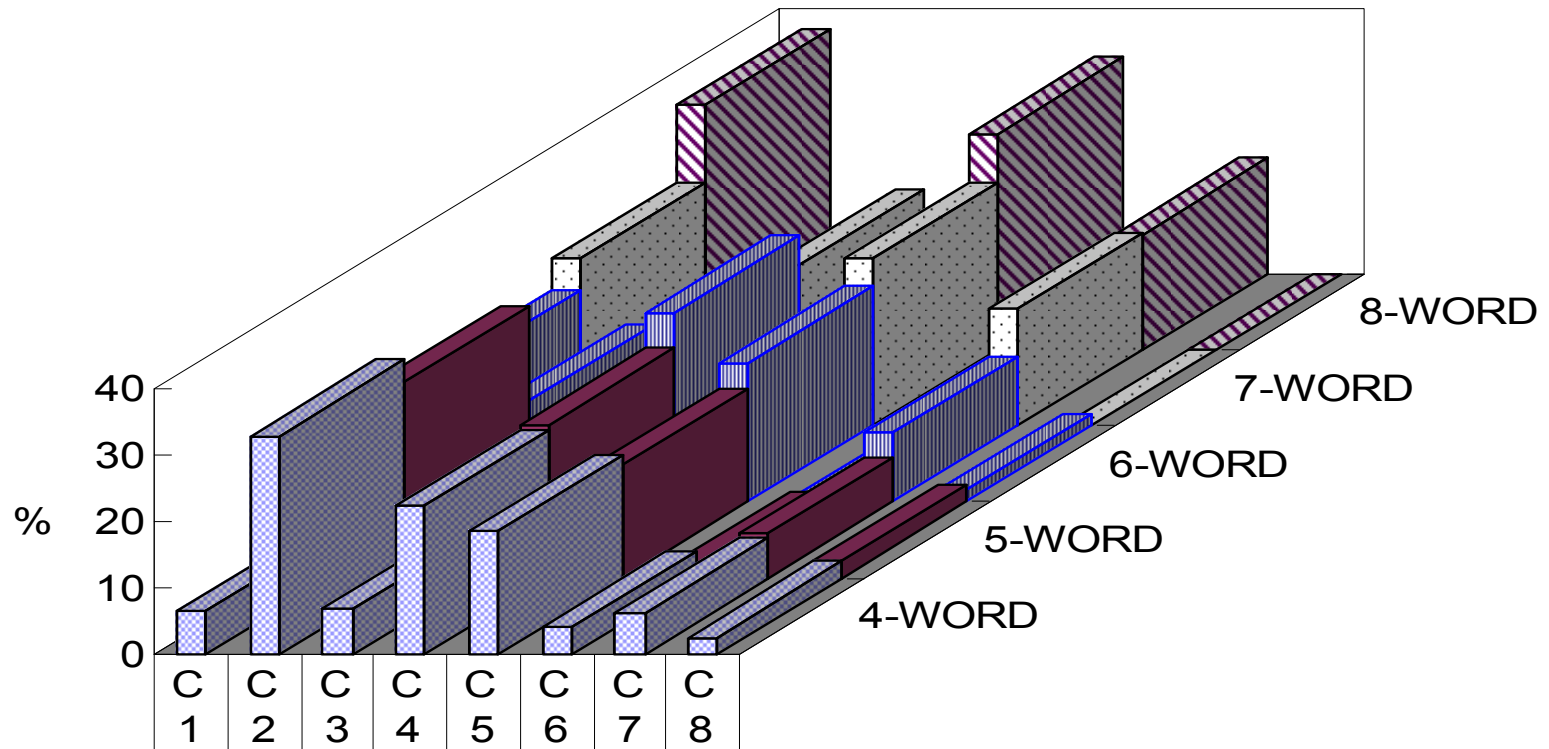
**Build(ing) a well-off society in a all-round way**

- When the size of the word clusters increases, the number of clusters as phrases also increases steadily. In the 8-word clusters, 67% are phrases and express independent meanings. For the 2-word and 3 word clusters, very small portions of the word lists can stand alone and be identified as phrases of explicit meanings (22% and 26% respectively), whereas a great majority of the clusters are but fragments of which the meanings can only be figured out with the help of context.

## 4.2 Categorization of the n-word clusters

- Since a large majority of clusters below 4-word are fragments and hard to be categorized into specific domains, a tentative categorization starts from 4-word and above.
- In our case 8 broad categories are made and labeled with numbers (from C1 to C8):
  - (C1) **national, political and governmental organizations**
  - (C2) common word combinations with no clear categorical identity
  - (C3) local administrations and organizations
  - (C4) **economic, commercial, and business organizations, events, and expressions**
  - (C5) **political expressions**
  - (C6) international names and events
  - (C7) **historical and socio-cultural, life, sports events and expressions**
  - (C8) academic organizations and universities; academic expressions (see Fig. 3).

Comparison of n-word clusters across category



	C 1	C 2	C 3	C 4	C 5	C 6	C 7	C 8
4-WORD	7	33	7	22	19	4	6	3
5-WORD	13	30	6	23	17	2	7	3
6-WORD	20	15	2	29	21	1	11	2
7-WORD	25	7	1	24	25	0	18	0
8-WORD	37	0	0	13	33	0	17	0

Category

- Over 30% of the 4-word and 5-word clusters belongs to C2 (common word combinations that have no clear categorical identity); the percentage of C2 decreases significantly as the word clusters grows in size and it drops to 6.6% and zero in the 7-word and 8-word clusters, which further supports the statement that longer word clusters carry more explicit meanings and are therefore more likely to stand alone as unique phrases.
- An average of 27.2% of all the n-word clusters (4-word and above) are political expressions (C5), 22.32% indicate national, political and governmental organizations (C1), 22.2% are words of economic, commercial, and business organizations, events, and expressions (C4), and 11.76% express historical and socio-cultural, life, sports events and expressions (C7).
- The four categories (C5, C1, C4, and C7) make up about 83% of all the word clusters from 4-word and above

- The political and economic terms are more likely to contribute to the multiword units. The longer the units, the more oriented to politics they are. The multiword expressions convey intended focus and image.
- Longer phrases and expressions are made up of shorter clusters that are strongly mutually attracted
- Sometimes the multiword lexical units are so fixed that they turn into something like idioms, of which the meanings are not transparent and it is paradoxically true that the meanings of long idioms are both specific and vague at the same time.



round way and create a new situation in building socialism with Chinese characteristics," Jiang d  
l-round way and create a new situation in building socialism with Chinese characteristics," was pu  
d work hard to create a new situation in building socialism with Chinese characteristics." The  
ation as well as the core of leadership in building socialism with Chinese characteristics and that i  
rt and one mind to advance the cause of building socialism with Chinese characteristics and crea  
d Leninism and formulated the theory of building socialism with Chinese characteristics after revi  
s have placed China on the right path of building socialism with Chinese characteristics, and have  
ese people to advance along the road of building socialism with Chinese characteristics. Jian  
It provides basic solutions for the building of socialism with Chinese characteristics and the r  
id that to open up new prospects for the cause of socialism with Chinese characteristics, the Part  
g and supporting the concept of market-oriented "socialism with Chinese characteristics."  
tinued to forge ahead triumphantly on the road to socialism with Chinese characteristics in spite of  
rejuvenation of the Chinese nation on its road to socialism with Chinese characteristics. Achi

## Multiword units of the same core expression attract different collocates

vention. When all the goals of modernization and a well-off society are met, the dream for national r  
e been increasing enrollments since 1999. To build a well-off society, China must increase the proport  
ernization. His reading of the report, entitled "Build a well-off society in an all-round way and create a  
l Congress of the CPC. In the report entitled "Build a well-off society in an all-round way and create a  
Party's cause, keep pace with the times, help build a well-off society and speed up socialist moderniz  
use into the future, keep pace with the times, build a well-off society in an all-round way, speed up so  
s, to start a new phase of development for building a well-off society in an all-round way and to speed  
as started a new phase of development for building a well-off society in an all-round way and speeding  
people, the scholar pointed out. The goal of building a well-off society has been closely related to the r  
safeguarding China's security and unity and building a well-off society in an all-round way, says the pap  
modernization by and large on the basis of building a well-off society in an all-round way, Jiang said.  
y of the rule of law as goals on the way to building a well-off society. China's legal system, seriously  
pidly. Meanwhile, China entered a stage of building a well-off society. Jiang Zemin, the core of the Par  
higher standard China will concentrate on building a well-off society of a higher standard in an all-rou  
cialist culture and spiritual civilization while building a well-off society in an all-round way in China, Jian  
and shares weal and woe with the people. Building a well-off society of a higher standard China will c  
tion drive and its magnificent blueprint to establish a well-off society in an all-round way will inevitably  
ng people's health, Yang said. The ultimate goal for a well-off society is to usher in a better and happie  
xiaokang 2002/11/06 The ancient Chinese ideal of a well-off society, xiaokang, revitalized by the cou  
as 1979, Deng Xiaoping proposed the concept of a well-off society and associated it with China's m  
d mainly on economics, Yang said. The concept of a well-off society proposed by the Party no longer  
epresents" and to forge ahead with construction of a well-off society in an all-round way. Jiang define  
ate food and clothing to its people, Deng's ideal of a well-off society was based mainly on economics,

reshold does not mean that China has become a xiaokang society, said Li Zhongjie of the Party School Association. Xue said China's decision to build a xiaokang, or well-off, society "of a higher standard", said Li Zhongjie of the Party School. To build a xiaokang society in an all-round way in China is a task that has been undertaken in recent years. But details of how to build a xiaokang society have not yet been spread at large. At the 14th National Congress (CPC) on November 8 was the call for building a xiaokang, or well-off society "of a higher standard". The 15th National Congress decided to expound on ways and means of building a xiaokang society in sectors such as politics, economy, culture, science and technology, and social construction, so as to enable all Chinese people to attain the goal of building a xiaokang (well-off) society in an all-round way in an all-round way, which will provide the material base for a xiaokang society. As always, China should greatly contribute to the world, the more it will contribute to world peace. A xiaokang society in China will also help accelerate the development of the world market. If it grows in accordance towards a xiaokang society, its trade volume will reach up to the level of the world market. Domestic product and build an all-inclusive affluent - xiaokang - society by 2020. In a resolution adopted at the 15th National Congress, the goal of building a xiaokang society is to promote social progress because xiaokang constitutes the second phase of the building of a well-off society is to promote social progress because xiaokang denotes not only material comfort, but also spiritual well-being. The 15th National Congress likely to discuss xiaokang 2002/11/06 The ancient Chinese ideal of a well-off society is to promote social progress because xiaokang denotes not only material comfort, but also spiritual well-being. The 15th National Congress likely to discuss xiaokang

14th and 15th Party congresses. In Deng's eyes, xiaokang is a key element in China's modernization strategy to improve people's standard of living. Deng's economic xiaokang strategy, designed to raise most Chinese people's standard of living, by Deng Xiaoping in 1979, the concept of xiaokang has found its way into a string of key Party documents. Beijing-based expert in the study of the concept of xiaokang. However, ordinary Chinese tend to see xiaokang as a goal for the 21st century. He further elaborated on his concept of xiaokang in 1984, saying "Xiaokang means that, enabling the people to enjoy an even higher level of xiaokang. In 2000, the Fifth Plenary Session of the 15th Central Committee decided, "We can say that we have reached the level of xiaokang." The third step is to basically complete the modernization of the country, are taken into account as new norms of xiaokang. China may not have much difficulty in

## Multiword units convey intended focus and image

The decision was made between 1955 and 1956, the People's Republic of China was young and winners of the First Time Film-makers project in the People's Republic of China, launched by Diurnal. Friess, you see, has lived and worked in the People's Republic of China - other than as a city was in 1949 again made Beijing, capital of the People's Republic of China. The fresh lease chairman of the Central Military Commission of the People's Republic of China, president of the chairman of the Central Military Commission of the People's Republic of China and president chairman of the Central Military Commission of the People's Republic of China, president of the movement of women in China. The Constitution of the People's Republic of China, the labour Law, he said. Article 29 of the Constitution of the People's Republic of China says that "the Ministry of Foreign Affairs for the first two decades of the People's Republic of China, 1949-1968. It is discussed as a gynaecologist after the establishment of the People's Republic of China, it is discussed as a gynaecologist after the establishment of the People's Republic of China in 1949, when models. New China After the establishment of the People's Republic of China in 1949, model through three reform phases since the foundation of the People's Republic of China: 1949-78: The women as part of celebrations for the founding of the People's Republic of China, the late crafts magazine," Han said. After the founding of the People's Republic of China in 1949, Han went on to say that as a result of a revolution, the founding of the People's Republic of China only changed the buildings were renovated after the founding of the People's Republic of China in 1949. The National People's Congress began on October 1, 1949 with the founding of the People's Republic of China. Shijitan was the country's social structure since the founding of the People's Republic of China in 1949, Yang Jie has said that events have occurred there since the founding of the People's Republic of China in October 1949. The columns used at the ceremony of the founding of the People's Republic of China, as well as promoting women's status since the founding of the People's Republic of China is remarkable.

top 10 distinguished Chinese women by the All-China Women's Federation. Yang is now studying in Beijing. The urban residents interviewed said they believe China can "basically" succeed in hosting the 2008 Olympics and the public. The following stories, by China Daily feature writers YU NAN, WANG SHIRONG, and China Unicom bets its future on CDMA. China Unicom is unlikely to surpass China Mobile's tennis team rise to the occasion (TANYI) and degradation of the ecological environment, China will be faced with a contradiction between economic growth and environmental protection. Telephone: 6512-8321 Under the hammer China Guardian Auction Co will hold a pre-auction of art. The number is 1588-9010. Airlines in China and the ROK are offering more flights to South Korea. China plan to organize a biennale of new media art in China. Experts believe that new media art is lifting the level of Chinese art. Yu Feng, who is one of the best conductors in China." Many participating students said they were surprised. The event was well-attended. Some universities and high schools in China called off afternoon classes to let students attend. "said Chen, who is from Harbin in Northeast China's Heilongjiang Province. A graduate of the Chinese Academy of Sciences. A senior manager of China Telecom, who preferred to be unnamed, said that the figures do not tell a whole story of China's agriculture trade in the context of the global market. The potential to shine at the next two Olympics. China dispatched its youngest-ever Asiad delegation to the 2002 World Volleyball Federation's (FIVB) world rankings. China is considered a title contender at the World Cup. China Information Industry. The figures also show that China had 196 million fixed-line telephone users and 100 million mobile phone users. China Mobile. Unicom Vice-President Shang Bain told China Daily that the number of new users grew rapidly. China Netcom, China Mobile, China Unicom, China Satcom and China Railcom and over 4,000 new users. The surging and far-reaching exploitation of West China, according to Zhang. Lack of trained personnel. The government official proclaim that although in recent years China's urban population has had trouble maintaining

- **Linguistically, the findings seem to support the assumptions that China English as a variety demonstrates more of its explicit features in vocabulary and phrases (Li, 1993; Widdowson, 1994; Wen, 2003).**
- **To put the point further, the most explicit features of China English are found in the multiword units and context, as in which the complete message is conveyed.**

Thank you  
for your  
attention!