2nd International Conference on Spoken Chinese Corpora, Rice University, August 10, 2016

Exploiting Spoken Corpora for Chinese Language Research and Teaching

Hongyin Tao
University of California, Los Angeles
&
National Taiwan Normal University
tao@humnet.ucla.edu

Spoken Corpora

- What constitutes a spoken corpus?
- What does it entail for research and teaching?

Spoken Corpora: Key Elements

- Naturalistic
- With multiple varieties
- Multimodality
- For research and/or teaching

Degrees of Naturalness/Authenticity

- Theories
- Techniques

Authentic Language and Authentic Conversational Texts

Lana Rings
University of Texas at Arlington

ABSTRACT *Although there is a trend to advocate the use of authentic texts in the foreign language classroom, a consensus regarding the criteria determining textual authenticity has not been reached. Instead, researchers often provide varying, sometimes conflicting definitions as to what comprise authentic materials.*

This paper draws on research in discourse analysis in an attempt to determine text authenticity through text type authenticity and provide implications for classroom materials. A text may be considered a spoken or written verbal unit, and a text type may be described as a specific type of spoken or written unit. Thus, for example, the text type "textbook conversation," written by textbook authors for the purpose of teaching specific structures, can probably not be defined as the text type "authentic conversation," in which native speakers engage in speaking for purposes other than to teach their language. In addition, a ranking of types of conversational texts, from most to least authentic, provides a scale by which to judge the value of materials used in the classroom.

ICE-GB

Medium I	Medium II(?) or Interaction Type(?)	Super-genre or Function	Genres or Sub-genres
SPOKEN (300)	Dialogue (180)	Private (100)	face-to-face conversations (90) phone calls (10)
		Public (80)	classroom lessons (20) broadcast discussions (20) broadcast interviews (10) parliamentary debates (10) legal cross-examinations (10) business transactions (10)
	Monologue (100)	Unscripted (70)	spontaneous commentaries (20) unscripted speeches (30) demonstrations (10) legal presentations (10)
		Scripted (30)	broadcast talks (20) non-broadcast speeches (10)
	Mixed (20)		broadcast news (20)
WRITTEN (200)	Non-Printed (50)	Non-professional writing (20)	student essays (10) student examination scripts (10)
		Correspondence (30)	social letters (15) business letters (15)
	Printed (150)	Academic writing (40)	humanities (10) social sciences (10) natural sciences (10) technology (10)
		Non-academic writing (40)	humanities (10) social sciences (10) natural sciences (10) technology (10)
		Reportage (20)	press news reports (20)
		Instructional writing (20)	administrative/regulatory (10) skills/hobbies (10)
		Persuasive writing (10)	press editorials (10)
		Creative writing (20)	novels/stories (20)

Multimodal Corpora

Recording, annotation and analysis of several communication modalities such as speech, hand gesture, facial expression, body posture, etc.

–Multimodal-Corpora.org

(MM) Spoken Corpora for Research and/or Teaching

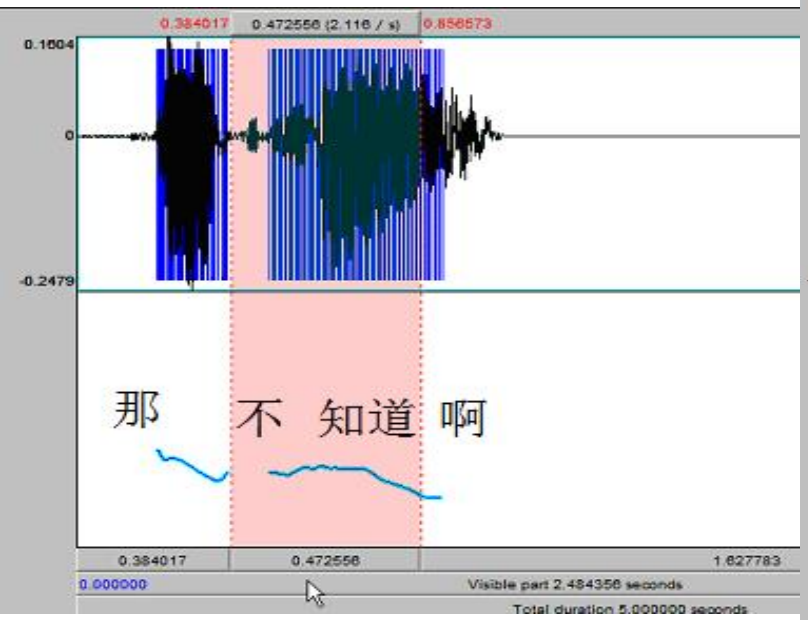
- A sound pedagogy involving spoken discourse needs a solid research foundation (understanding how spoken discourse works)
- Yet it takes effort to translate discourse research into useful and practical teaching practice

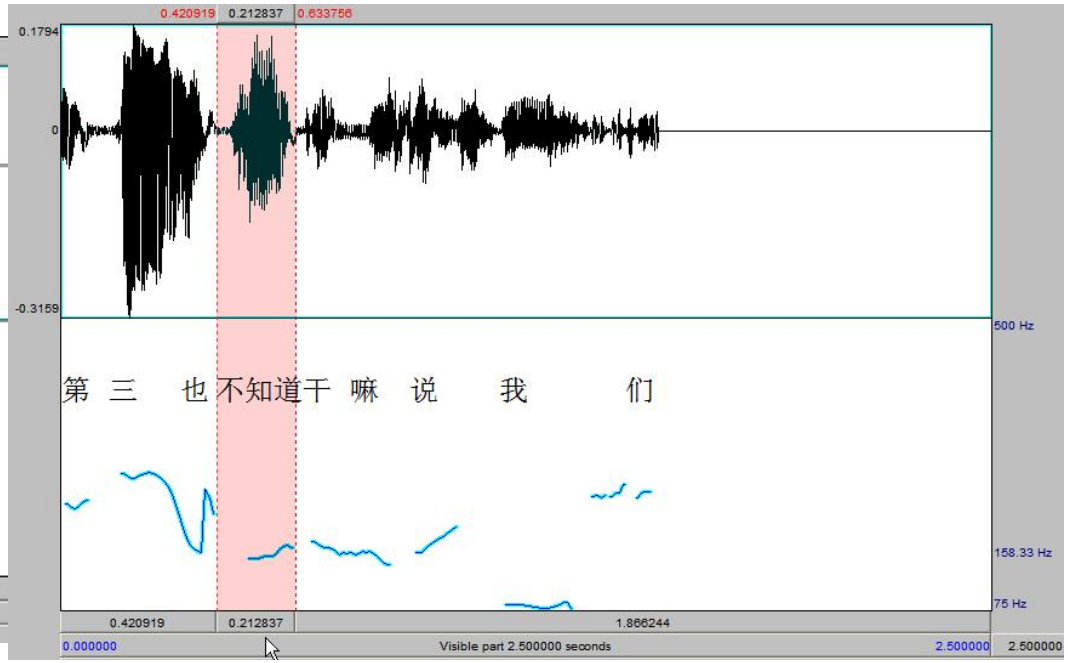
Prosody and discourse

- MM corpora afford prosodic information which is vital for understanding discourse and grammar (or grammar in interaction).

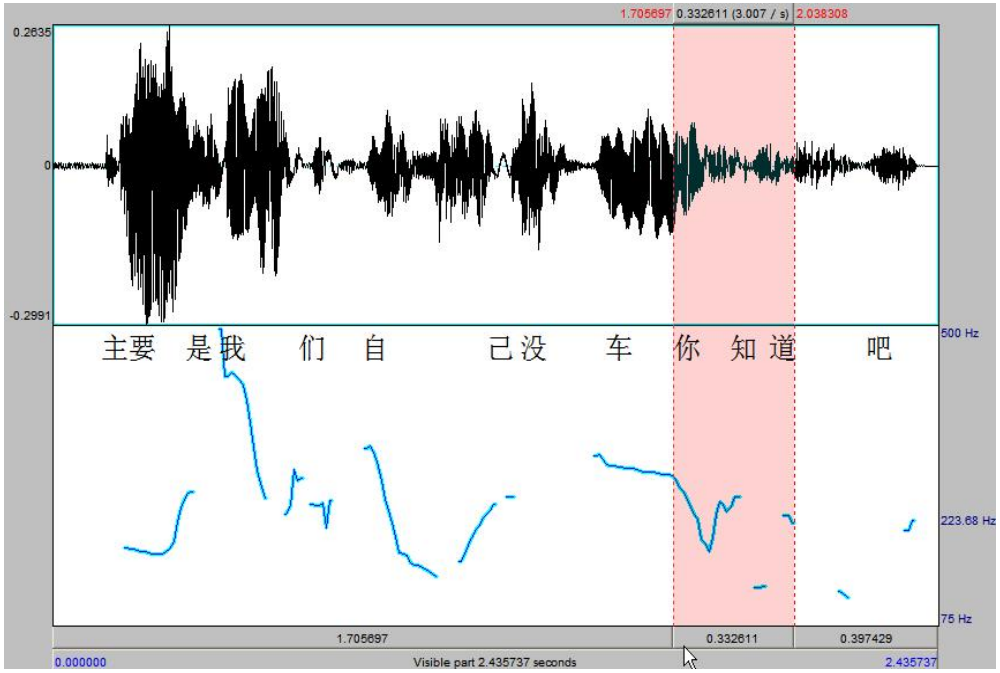
Ex. 1: Discourse/epistemic markers

- *Zhidao* 知道 ‘to know’ constructions
- Prosodic change reflects pragmatic strengthening & syntactic change


0.384017
0.472556 (2.116 / s)
0.856573
0.1604
0
-0.2479
那 不 知道 啊
0.384017
0.472556
1.627783
0.000000
Visible part 2.484356 seconds
Total duration 5.000000 seconds


0.420919
0.212837
0.633756
0.1794
0
-0.3159
第 三 也 不知道 干 嘛 说 我 们
500 Hz
158.33 Hz
75 Hz
0.420919
0.212837
1.866244
0.000000
Visible part 2.500000 seconds
2.500000
2.500000

45%


1.705697
0.332611 (3.007 / s)
2.038308
0.2635
0
-0.2991
主要 是我 们 自 己没 车 你 知道 吧
500 Hz
223.68 Hz
75 Hz
1.705697
0.332611
0.397429
0.000000
Visible part 2.435737 seconds
2.435737

70%

Ex 2: Gesture in interaction

- Gestures are used to serve various rhetorical and interpersonal functions
- These functions may not be reflected in speech alone
- Co-speech gestures: speech & gesture mutually elaborate each other

Offering help to a friend

R：..你打电话？哎，让我老头陪你去吧。

Y：…为什么让你老头陪我去?

R：..因为我-我现在急着录这个嘛。因为我那个什么，我今天晚上有家教。我老头正好 在研究室。他正好没事干。你赶快让他陪你去吧。

[他无聊]死了。

Y：..[你让-] 你让他上我们宿舍，

R: Are you still on the phone? Why don't you let my man go with you there?

Y: Why him?

R: 'Cause I'm busy at the moment with this recording thing. I have a private tutoring session this evening also. He's in his office, doing nothing. Let him go with you. He's probably bored to death right now.

Y: OK tell him to go to my dorm then.

Thus

- Multimodal spoken corpora provide new avenues for research
  - New dimensions for understanding the way human communication works
  - How language structure works together with other modalities and communicative resources

Issues for Researchers/Teachers

- How to translate MM spoken corpus research into teaching?
- How to expose learners to multimodal patterns?
- Language learning beyond lexical grammar (doing and acting)

Prosody in Discourse & Pronunciation Teaching

Understandable Reasons

- From single words/simple sounds to strings of words/complex patterns
- Chinese individual tonal patterns are important & difficult to acquire
- Intonation & prosody are difficult to teach

Characteristics of Textbook Pronunciation

- Slow speed
- Clear enunciation
- Standard accent

Any problems?

ACTFL Standards: Communication

- **Standard 1.1** Students engage in conversations, provide and obtain information, express feelings and emotions, and exchange opinions.

Most Chinese spoken Textbooks: Favor Information over Social Interaction

- Provide and obtain information;
- Express feelings and emotions;
- Exchange opinions；

- *Engage in interaction;*
- *Building coherent discourse (linking).*

Natural Speech

B：不过一般是不是这边衣服，
ˆ都是，
ˆ皮衣的话二百块钱左右啊，都是。
A：嗯=，我不知道，
但是可能跟国内价格要便宜点儿吧。
B：便宜多=了，/ 我--
A：因为我上回听国内同学说，
她就想买一件儿嘛，
B：嗯=。
A：但是因为她可能也没有机会去纽约那边，
B：对。
A：所以那就也就算了。
B：对。
A：因为我们这边儿没有。
我们这边儿那个小的那种超市都没有，
B：[哦，对。]

A：[没有卖这种东西的。]
嗯。
B：反正国内现在买一件就说的，
… 就说ˆ你能看着好点儿，
[不是说外边儿摊儿上]的那种皮衣，
A：[嗯嗯，嗯。]
嗯哼。
B：四五ˆ千=呢=。
A：对呀。
B：特贵=。
A：你还是应该买一件儿。
@@@
B：啊？
A：ˆ应该买一件儿。
B：那这不没看着嘛，
给我难受死了。
A：两千。

“Artificial Dialogue”

B：这边衣服，

皮衣的话二百块钱左右

A：我上回听国内同学说，
她就想买一件儿

B：嗯=。

A：她可能也没有机会去纽约那边，

B：对。

A：就算了。

A：我们这边儿没有。

A：没有卖这种东西的。

B：国内现在买一件

那种皮衣，

A：四五 千

B：对呀。

A：特贵=。

B：你还是应该买一件儿。

A：啊？

B：买一件儿

A：没看着

B：两千。

Teaching Conversational Prosody

- Based on large collections of spoken data
- Selected features and segments
- Selected levels
- Selected varieties/accents as needed
- Implicit/holistic and explicit teaching

Explicit Instructions

Listen to the audio clip one last time and decide if each of the following utterance or part of the utterance is lengthened (L), stressed (S), has a rising intonation (R), or has an expressive intonation (E).

1) ____L ____S ____R____E 便宜多了

2) ____L ____S ____R____E 四五千呢

3) ____L ____S ____R____E 你还是应该买一件儿

4) ____L ____S ____R____E 给我难受死了

睫毛增生劑

讓妳會掉…讓妳會長

Vocabulary / Lexico-Grammar

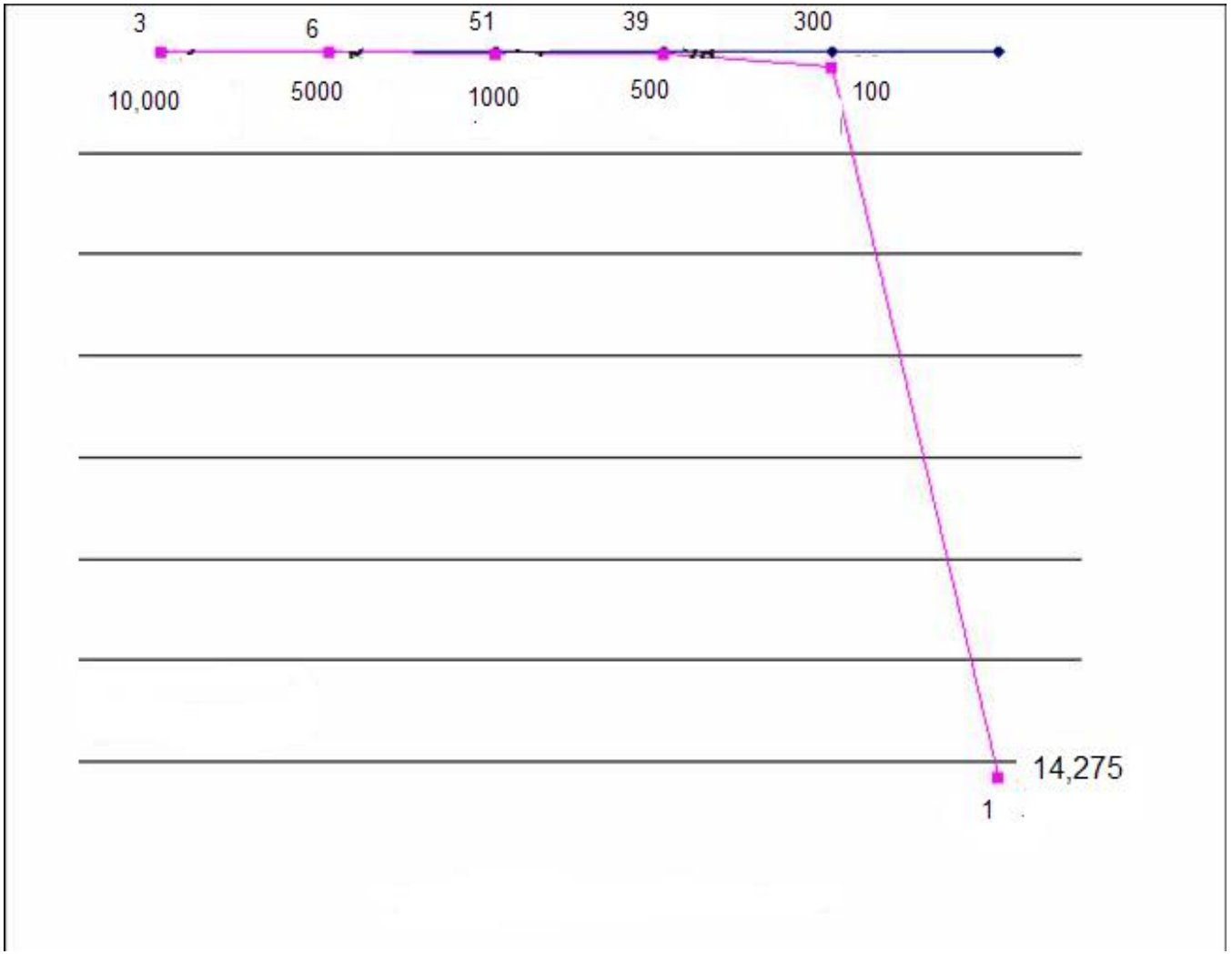
Tao 2013-15: High frequency items

1) 的.....13245
2) 是.....12047
3) 我.....10052
4) 就.....7782
5) 不.....7743
6) 你.....7658
7) 了.....7484
8) 那.....6846
9) 啊.....5792
10) 个.....4696
11) 他.....4385
12) 对.....4285
13) 就是.....3920
14) 有.....3816
15) 都.....3760
16) 说.....3677
17) 一.....3497
18) 也.....3186
19) 那个.....3154
20) 然后.....3076
21) 在.....3067
22) 什么.....3064
23) 这.....3027
24) 这个.....2772
25) 很.....2373
26) 哦.....2245
27) 看.....2197
28) 人.....2100
29) 还.....2093
30) 嗯.....1953
31) 好.....1939
32) 要.....1871
33) 我们.....1847
34) 去.....1824
35) 一个.....1814
36) 觉得.....1694
37) 到.....1666
38) 她.....1606
39) 没.....1590
40) 吧.....1539
41) 多.....1490
42) 它.....1474
43) 没有.....1438
44) 得.....1412
45) 呢.....1384
46) 跟.....1336
47) 他们.....1335
48) 儿.....1326
49) 上.....1235
50) 吗.....1200
51) 现在.....1176
52) 知道.....1135
53) 嘛.....1112
54) 但是.....1082

Intriguing Patterns

of Items

Frequency



The findings suggest

- A very small number of words are doing most of the work in everyday conversation
- (for comparable numbers from modern spoken English, see McCarthy & Carter 2003)
- What exactly are they then and what does this imply for vocabulary teaching?

An Initial Taxonomy of the Core Lexicon

Pronouns（我，你，他）
Low content verbs (是，有，沒有）
Speech act verbs （說)
Cognitive verbs （覺得，知道，(看)）
Adverbs （就，就是，都，也,很，還）
Numeral/Classifiers (一，一個)
Modal expressions （要）
Negation （不，沒有）
Demonstratives （這，這個，那，那個）
Temporal deictic (然後，現在)
Reactive tokens (哦，嗯，啊, 對)
Particles (吧，呢，嘛，啊)
Interrogatives （什麼）
Conjunctions (所以，而且，但是)

Lexical Bundles in Spoken Corpora

Three-word lexical bundles in conversation

1	35	你 說 你
2	442	是 不 是
3	125	不 是 說
4	122	對 對 對
5	116	也 不 是
6	75	不 是 啊
7	72	對 不 對
8	70	不 是 不
9	66	是 是 是
10	42	你 看 你
11	42	就是 說 你
12	40	那 你 就
13	39	你 知道 嗎

(By contrast: Academic Chinese)

1 53 是 一 種
3 34 的 一 種
4 32 並 不 是
5 28 的 基礎 上
9 19 而 不 是
17 15 如 圖 所 示
18 15 更 多 的
19 15 有 不同 的
20 15 的 情況 下
25 14 之間 的 關係
26 14 所 說 的
27 14 相 結合 的
33 13 的 過程 中
35 13 這 兩 個
36 12 一些 新 的
37 12 兩 個 方面
38 12 主要 表現 在

Discourse Grammar

- Most grammars list items such as: 因为/所以；不但/而且；由于；尽管；无论/也；虽然/但是； 既然/就；
- These are valid items that need to be taught and learned but not always typical of everyday language use;
- Few talk about context and function;
- A discourse approach: find patterns in discourse and orient teaching based on these patterns

CALPER Chinese, PSU

- http://calper.la.psu.edu/ → Chinese
- http://calper.la.psu.edu/chinese.php
- Pennsylvania State University

- Integration of
  Audio/(Video)/Text/Concordance/Discourse Analysis

Variety: Categories & Features

- ***Discourse features*** *话语功能*: story telling & presenting events, expressing agreements and disagreements, introducing and negotiating topics, reacting to and collaborating with the main speaker, etc.
- ***Topical types*** *话题类别*: travel, TV shows, shopping, sports, playing cards, campus tour, etc.
- ***Speech events*** *言语活动类别*: chat, talking in action, narration, demonstration, etc.

- **Unit 1 Travel Adventures** 旅行奇遇

  Story telling • Elements of a Story (Who, When, Where, etc.) • Addressee Behaviors in Story Sessions
- **Unit 2 Fashion Fever** 流行装

  Comments on Fashion Styles • Expressing Agreement / Disagreement • Approval / Disapproval • Support Arguments with Examples
- **Unit 3 Global Citizen**地球公民

  Discussing Attitude toward International Experience • Contrasting Information • Elaborating Information • Supporting Ideas with Examples
- **Unit 4 Who Wants to be a Millionaire** 百万富翁

  Talk about a TV Game Show • Compare Similar and Unlike Things • Degrees and Qualities • Qualifying an Opinion • Vague Expressions
- **Unit 5 Insurance Abuse** 滥用医疗保险

  Commenting on Insurance-Related Social Phenomena • Making Negative Statements • Vague References and Negative Events • Demonstratives and Degree Expressions

- **Unit 6 Las Vegas 拉斯韦加斯**

  Reflecting on a Tourist Destination • Admiration of Places • Contrasting Features of Places • Expressions for Praising • Listener Behavior in Praising
- **Unit 7 Tough Job Market 就业问题**

  Discussing Job Opportunities • Presenting Multiple Reasons / Perspectives • Expressing Past, Present. and Future Temporal Frames of Events
- **Unit 8 People and Personalities 表情与性格**

  Commenting on People's Appearances and Qualities • Opinions with Different Degrees of Certainty • Personal Interest • Emotional State • Listener Reactions to Comments
- **Unit 9 Driving Cautiously 小心开车**

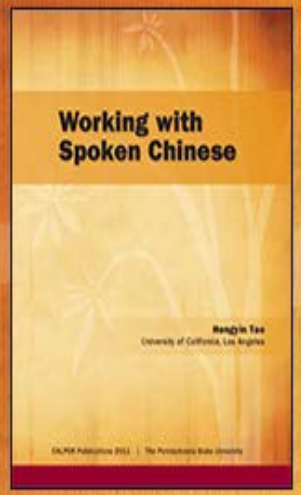
  Reflecting on Driving Experiences • Different Accounts of Shared Experience • Humor • Alluding to Cultural Concepts
- **Unit 10 Horror Movies 恐怖片**

  Describing a Movie • Types of Movies • Personal Taste about Movies • Liking and Disliking • Listener Corroborating with the Other Speaker

Working with Spoken Chinese

by Hongyin Tao, UCLA

Companion Website to the Textbook



ISBN: 978-0-9793950-8-6
Purchase the textbook

This site accompanies the textbook "Working with Spoken Chinese." For each of the ten units of the
(without annotations, with Pinyin, with selected line-by-line commentaries), a vocabulary list, and cor

Users of this site need to be aware that the site contains copyrighted materials. Individuals who prov
only be used for the development of learning materials for "Working with Spoken Chinese."

Publisher: CALPER Publications, Center for Advanced Language Proficiency Education and Resea
University Park, PA, USA. Email: calper@psu.edu

Unit Overview

Unit 1 Travel Adventures 旅行奇遇
Story telling, Elements of a Story (Who, When, Where, etc.), Addressee Behaviors in Story S

Unit 2 Fashion Fever 流行裝
Comments on Fashion Styles, Expressing Agreement/Disagreement, Approval/Disapproval,

Working with Spoken Chinese

by Hongyin Tao, UCLA

Companion Website to the Textbook

Unit 1 Travel Adventures 旅行奇遇

Story telling Elements of a Story (Who, When, Where, etc.), Addressee Behaviors in Story Sessions

- Audio Clip
- Transcript
- Transcript (with Pinyin)
- Transcript (with selected line-by-line commentary)
- Vocabulary List
- Concordance

Return to the Unit Overview

Wordlist
(in sections)

Whole list

的 (17)
一 (13)
就 (12)
个 (12)
那 (10)

uh (9)
了 (9)
是 (9)
我们 (8)
在 (8)
他 (8)
也 (7)
什么 (6)
然后 (5)
对 (5)
去 (4)
呢 (4)
地方 (4)
好 (4)
跑 (4)
住满 (4)

huh (3)
走 (3)
不 (3)
没有 (3)
哦 (3)

The Concordance

Next section

的.....17

因为 意想不到 **的** 事情 太 多 了 。	2
所有 **的** 宾馆 也 是 爆满 ,	24
几 十 分钟 **的** 一 个 小城 ,	28
在 那儿 **的** 宾馆 也 住 得 很 满 ,	32
uh 已经 好像 是 一 两 点钟 **的** 事情 ,	36
uh 然后 就 是 , 很 无辜 **的** 看 着 他 ,	46
后来 他 呢 , 倒 还 挺 好 **的** ,	48
就 把 我们 带 到 他 **的** 餐厅 ,	50
他 **的** 餐厅 , 那种 窄 的 椅子 ,	52
他 的 餐厅 , 那种 窄 **的** 椅子 ,	52
F2: 那 个 也 是 薄薄 **的** 窄窄 的 。	54
F2: 那 个 也 是 薄薄 的 窄窄 **的** 。	54

The Text

F2: 对 , 这个 出国 的话 , 比较 锻炼 人 。
因为 意想不到 的 事情 太 多 了 。
(several lines deleted)
比如说 有 一 次 我们 去 威尼斯 ,
那 个 是 周末 ,
然后
uh
没 想到 会
全部 都 住满 , 旅馆
F1: 哦 .
F2: 每 一 间 旅馆 都 住满 。
在 那边 走 了 一 大 圈 ,
每 一 个 都 住满 了 ,
好 , 那 我们 就 跑 去
先 开始 是 在 什么 地方 ?
哦 , 我们 就 跑 去 另外 一 个 城市 ,
结果 发现 那 个 地方 呢 有 一 个 什么 uh

Conclusions

- (MM) spoken corpora afford a wide range of information for us to understand human communication, of which language structure is but one part of the ecology.
- Spoken corpora raise questions about units and use of prosody, vocabulary, grammar, and discourse pragmatic patterns.
- Spoken corpora can be exploited for both research & language teaching from multiple perspectives.