Phonological Integrity or Orthographic Integrity: Challenging Traditional Views on How Sounds in a Language is Defined

Chu-Ren Huang 黃居仁

Dept. of Chinese and Bilingual Studies

The Hong Kong Polytechnic University,

http://llt.cbs.polyu.edu.hk/

How to Define a Language or a Linguistic System?

- エレベーターガール
- eɽebētā gāɽu ‘elevator girl’
- Phonological adaptation is predicted (and happened) because of the autonomy of the phonological system of Japanese that allows only 1) a restricted inventory of phonemes, and 2) CV structure

Phonological Adaptations Attested

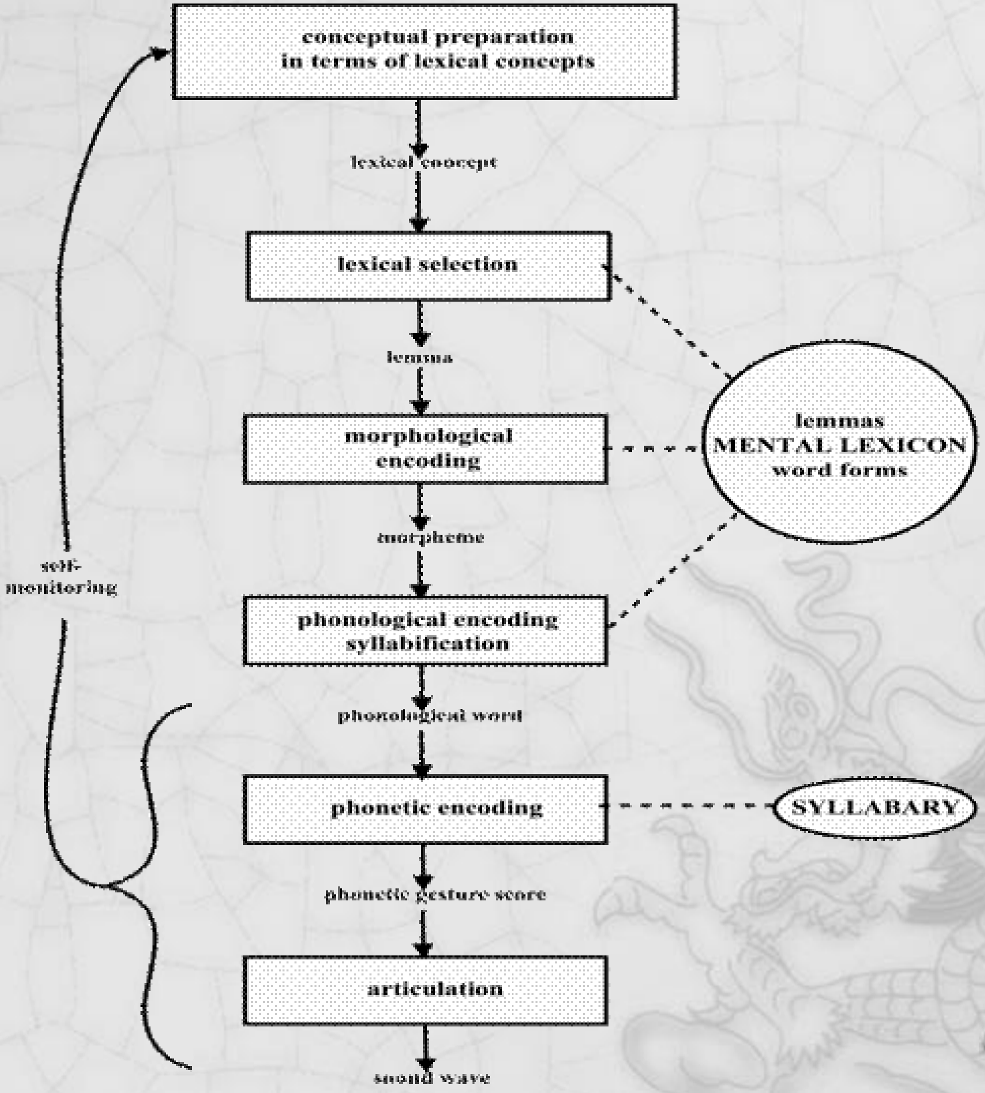
- Wien (Gr.) → Vienna (Eng.)
- ketchup (Eng.) →kechappu ケチャップ (Jp.)
- Paris (Fr.) → Paris (Eng.)
- guitar (Eng.)→結他[kit t'a] (Can.)
- cream (Eng.)→忌廉 [key lim] (Can.)

Models of Language Processing Theories of Phonology

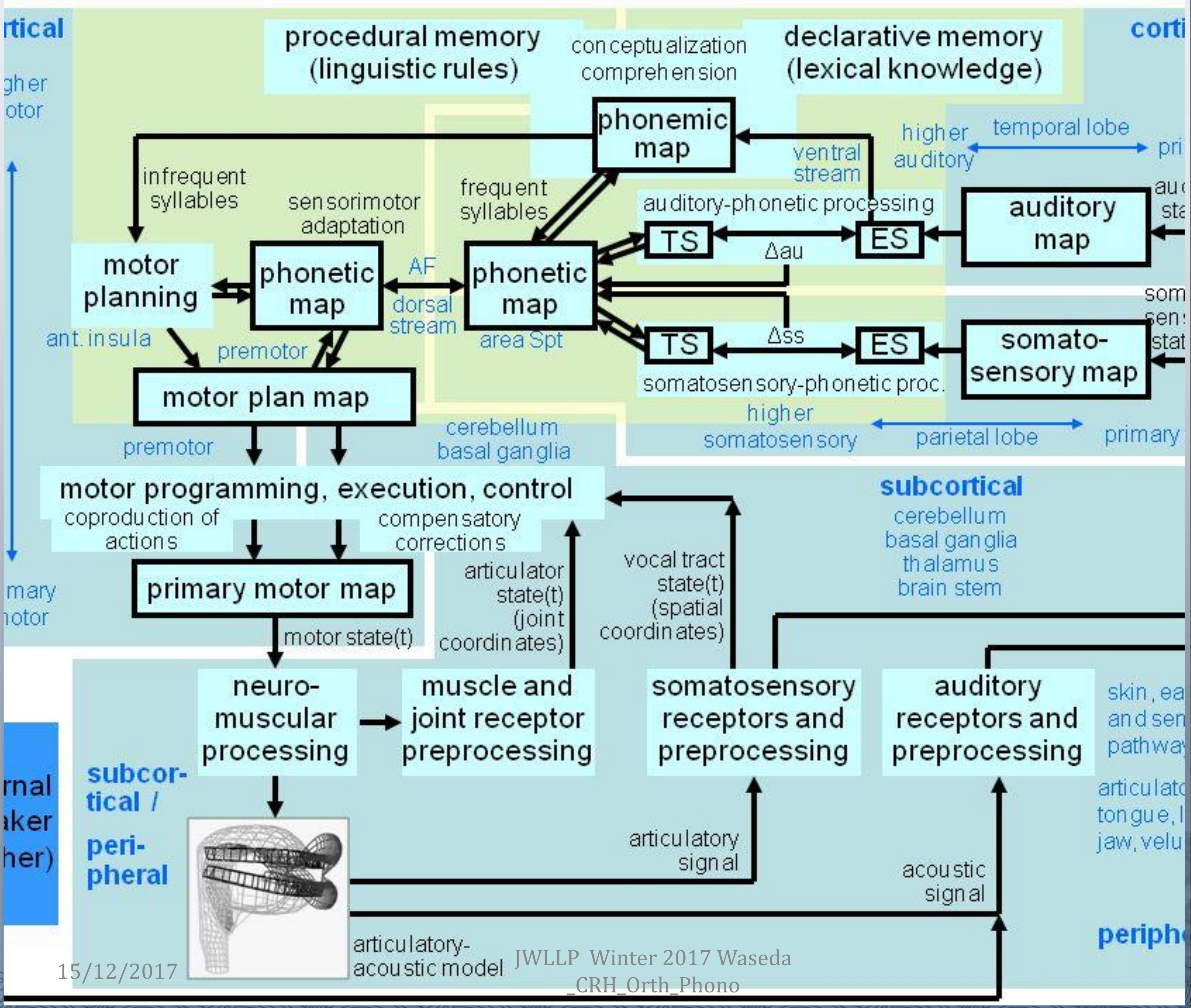
Models of language processing/linguistic theories of phonology are constructed on the premise that each language has an autonomous phonological system, which contains a set of well-organized phonemes.

- Hence any speech sound not in the phonemic repertoire can either be ignore as non-speech sounds in processing or recognized as foreign.
- Under such assumption, phonological adaptation is predicted, attested, and used as motivation to account for both loan word phonology and for accent reduction.

Levelt et al 1999 Speech Production Model



ACT Model of Speech Processing: Kröger et al. 2009



Evidence to Support the Phonological Encoding Model

- Phonemic Restoration Effect (Warren 1970)
- McGurk Effect (McGurk and MacDonald 1976)
- Phonological Adaptation of Loan Words

Exceptions to Phonological Encoding

However, it is also true that speakers routinely produced sounds and units not included in the standard phonologic or lexical inventory of the language considered during language production. What happens when unexpected sounds are encountered?

But What About

- Vocal mimicries and/or onomatopoeia:
  - Chinese Taiwanese: biang, kiang, duaiñ, bonggiu, gu(M)gu(H)gu(L) etc.
- Non-lexical conversational sounds: such as variations of mh, nnh, enn, clicks (English)
  - They are in fact language specific
- Exceptional Lexical Items
  - Taiwanese reduced triple reduplication (3 to 2 syllable, 2 tones merged on the first syllable)
  - Mandarin Alphabetic Words

One Thing in Common

- These sounds are not typically rendered in orthography
  - With the exception of exceptional lexical items in Chinese, to be discussed later
- Could it be that it is orthography rather than phonology that is contributing to the integrity?
  - Or perhaps orthography reflects the fact that there are more than one sub-systems in our mental lexicon?

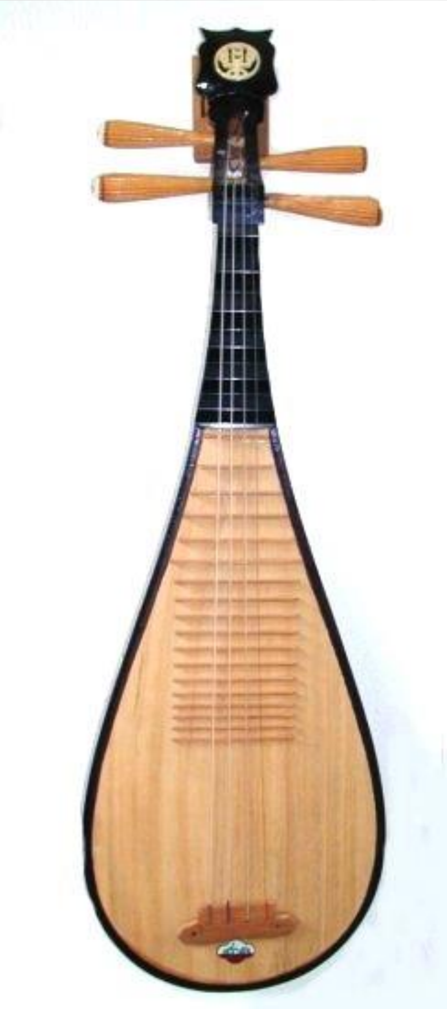
Role of Orthography: ORL

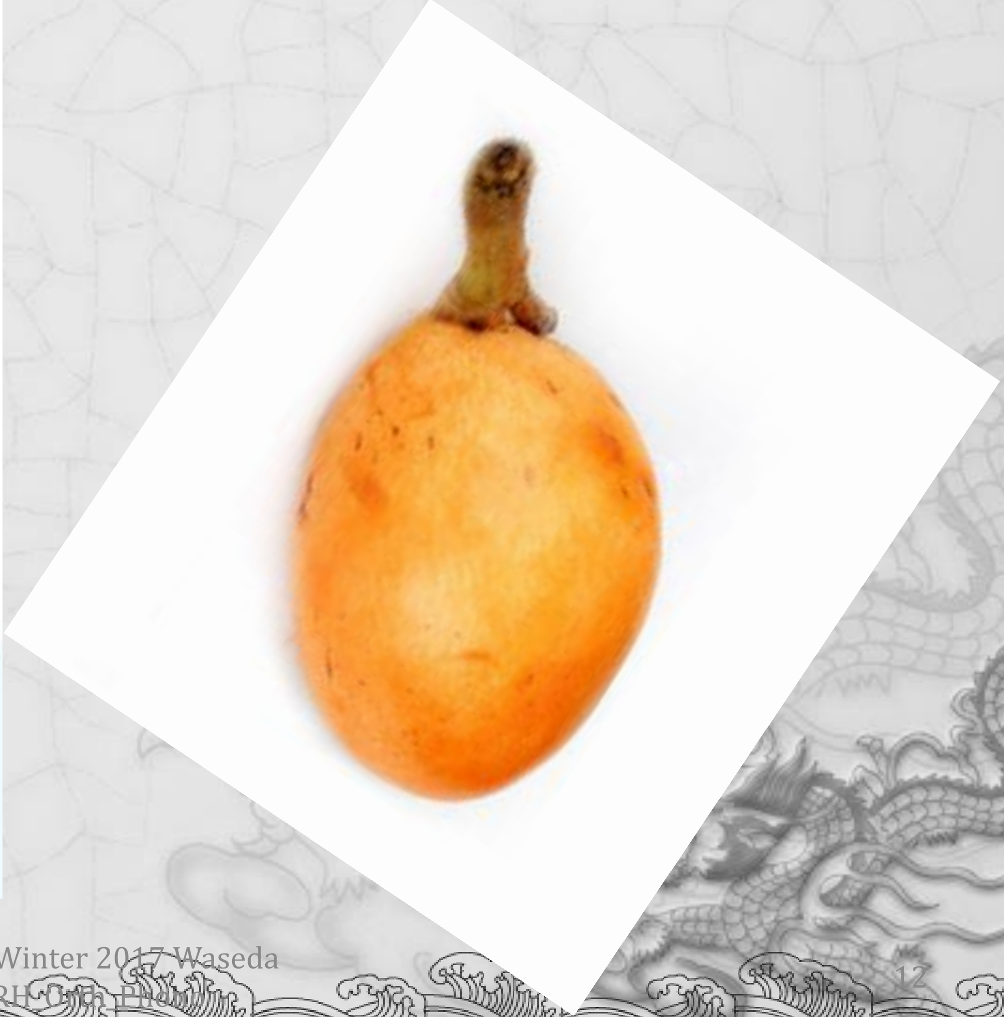
We will term the linguistic level represented by the orthography of a language the *Orthographically Relevant Level* – ORL…..The ORL is simply that linguistic level of representation at which those regular correspondences are most succinctly stated.

-Richard Sproat (2000.10)

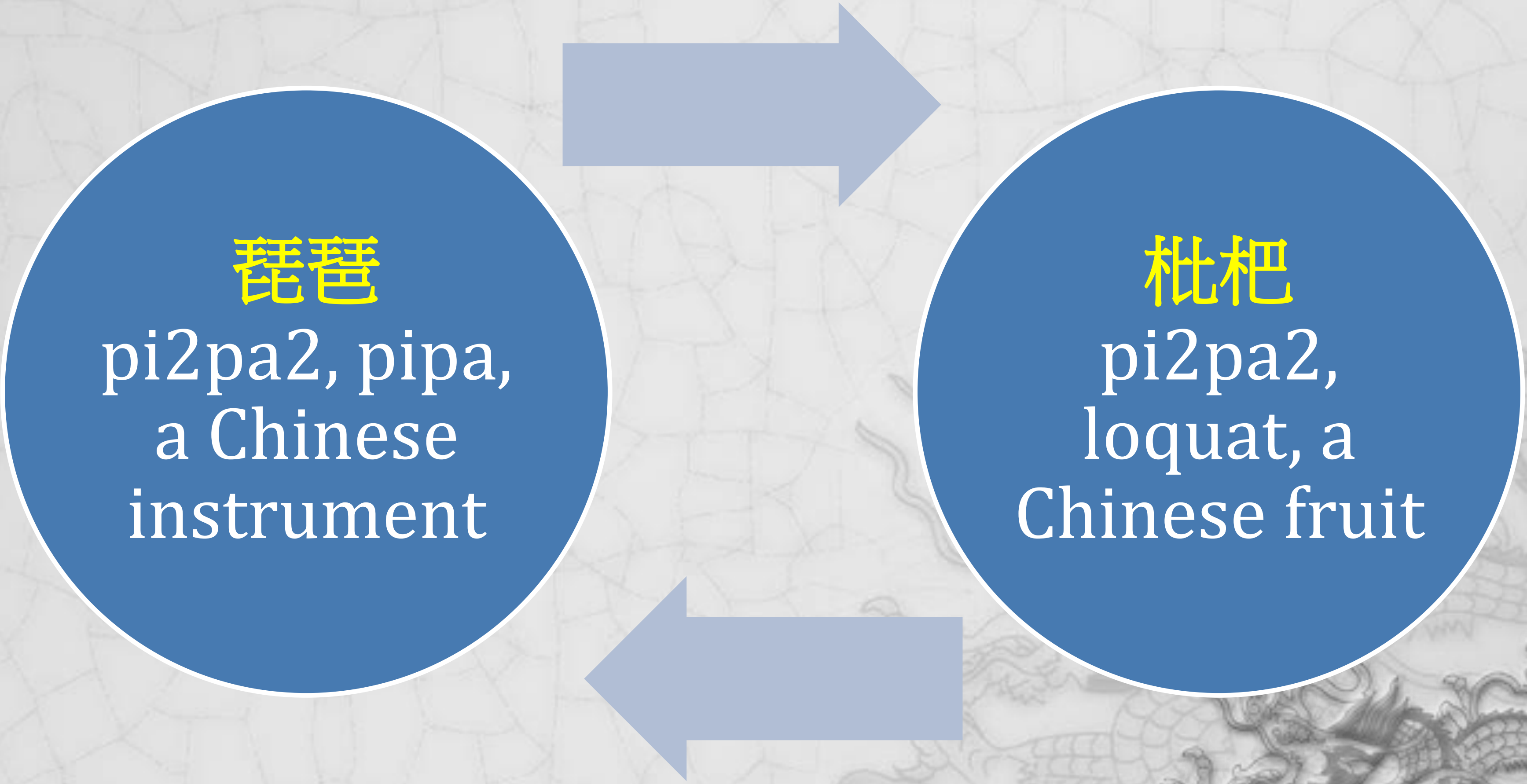
- ORL is the conventional representation of shared linguistic knowledge. -CRH

Pipa and Loquat





Pipa and Loquat: Two Related Disyllabic Stems



琵琶
pi2pa2, pipa, a Chinese instrument

枇杷
pi2pa2, loquat, a Chinese fruit

字词：
Characters vs. Words, a Reminder

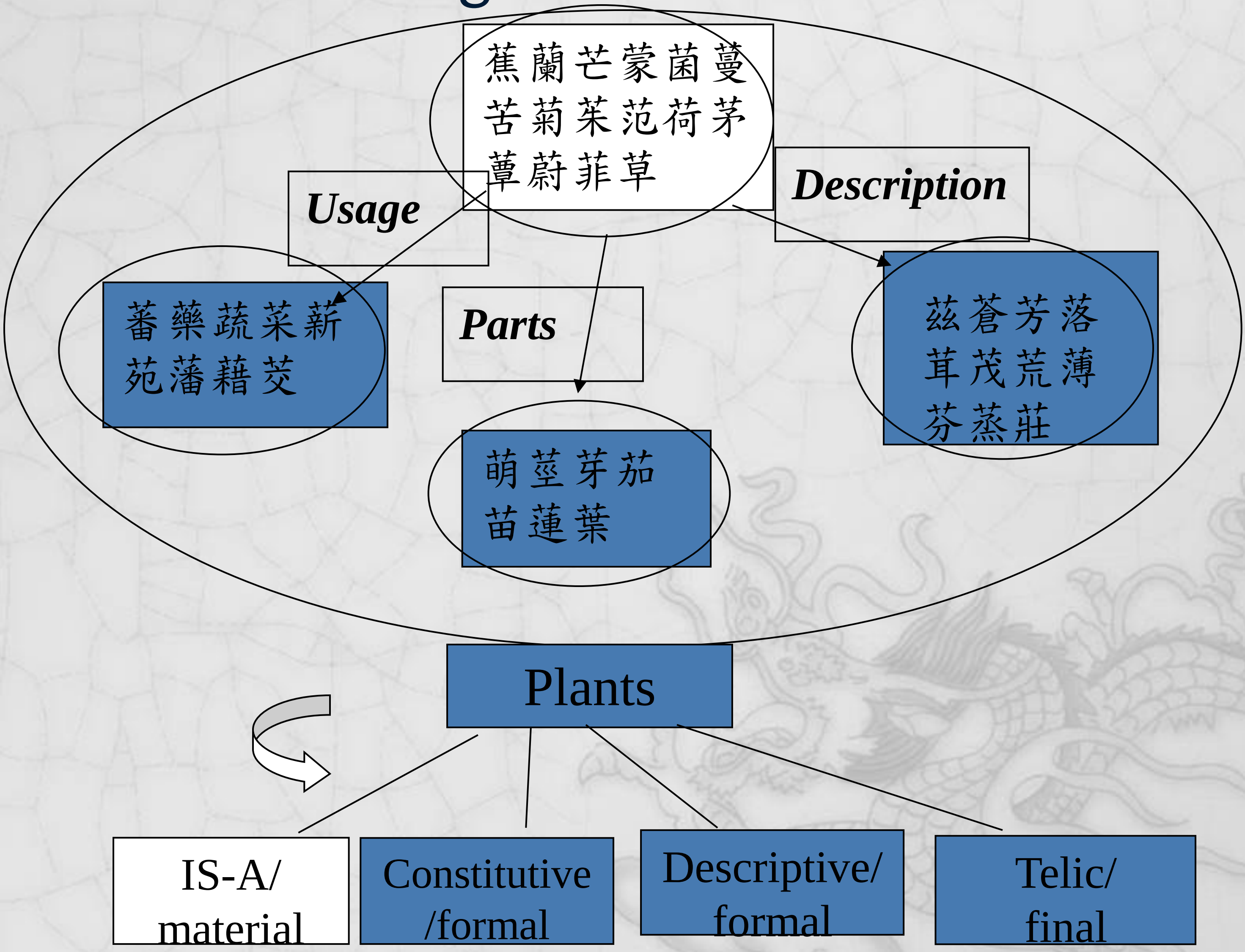
- There are many multi-syllabic roots in Chinese
- Radicals are component parts of characters
- Each pair of characters share the same radical, without exceptions
  - Why 枇杷 is not 琵琶 (while we use *blackberry* to refer to both the berry and the gadget;
  - Why there is neither 琵杷 nor 枇琶 (while knight/night time/tyme are common mistakes)?
- Following data was extracted from ROCLING and United Daily Corpus (over 20 million characters in total)

60 Bi-syllabic Stems (Updated Based on Sproat 2000)

Orthography Analysis		Pronunciation	Gloss
鴛鴦	<BIRD+YUĀNYĀNG>	yuānyāng	‘mandarin duck’
狡猾	<DOG+JIĀOHUA>	jiǎohuā	‘cunning’
蕃薯	<GRASS+FĀNSHÙ>	fānshǔ	‘yam’
葫蘆	<GRASS+HÚLÚ>	húlú	‘gourd’
蘿蔔	<GRASS+LUÓFÙ>	luóbō	‘daikon’
葡萄	<GRASS+PÚTÁO>	pútáo	‘grape’
恍惚	<HEART+GUĀNGHŪ>	huǎnghū	‘illusionarily’
慷慨	<HEART+KĀNGJÌ>	kāngkǎi	‘generous’
蝴蝶	<INSECT+HÚDIE>	húdié	‘butterfly’
螞蟻	<INSECT+MǍYÌ>	mǎyǐ	‘ant’
螃蟹	<INSECT+PÁNGXIÈ>	pángxiè	‘crab’
蟑螂	<INSECT+ZHĀNGLÁNG>	zhāngláng	‘cockroach’
琥珀	<JADE+HÚBÓ>	hǔpò	‘amber’
琳瑯	<JADE+LÍNLÁNG>	línláng	‘kind of jade’
玻璃	<JADE+PÍLÍ>	bōlí	‘glass’
尷尬	<LAME+JIĀNJIÈ>	gāngà	‘awkward’
咆哮	<MOUTH+PAOXIÀO>	páoxiào	‘roar’

Orthography Analysis		Pronunciation	Gloss
囹圄	<SURROUND+LÌNGWÚ >	língyú	‘imprisoned’
囫圇	<SURROUND+WÙLÚN>	húlún	‘swallow whole’
轇轕	<CART+LIÀOGĚ>	jiūgé	‘entwined’
窈窕	<CAVE+YÒUTIAO>	yǎotiǎo	‘graceful’
魍魎	<DEMON+WǍNGLIǍNG >	wǎngliǎng	‘roaming ghost’
妯娌	<FEMALE+ZHOULǏ >	zhóulǐ	‘sister in laws’
餛飩	<FOOD+KŪNTÚN>	húntún	‘wonton’
蹉跎	<FOOT+CUŌTUŌ>	cuōtuó	‘procrastinate’
踉蹌	<FOOT+LÁNGQIANG>	lángqiāng	‘hobble’
蹂躪	<FOOT+RÓULÌN>	róulìn	‘trample’
躊躇	<FOOT+CHÓUZHÙ>	chóuchú	‘hesitate’
躑躅	<FOOT+ZHÌSHǓ>	zhízhú	‘hesitate’
氤氳	<GAS+YĪNYUN>	yīnyūn	‘misty atmosphere’
邂逅	<GOING+XIÈHÒU>	xièhòu	‘encounter’
迤邐	<GOING+YÍLÌ>	yǐlǐ	‘trailing’
荸薺	<GRASS+BÓQÍ>	bíqí	‘water chestnut’
萵苣	<GRASS+GUǍJÙ>	wōjù	‘lettuce’
菡萏	<GRASS+HÁNXIÀN>	hàndàn	‘lotus’
蒹葭	<GRASS+JIĀNJIǍ>	jiānjiā	‘type of reed’
苜蓿	<GRASS+MÙSÙ>	mùsù	‘clover’
揶揄	<HAND+YĒYÚ>	yéyú	‘tease’
顢頇	<HEAD+MǍNHAN>	mánhān	‘muddleheaded’
慫恿	<HEART+CÓNGYǑNG>	sǒngyǒng	‘egg on’
忸怩	<HEART+NIUNÍ>	niǔní	‘coy’
慇懃	<HEART+YĪNQÍN>	yīnqín	‘attentively’
蝙蝠	<INSECT+BIǍNFÙ>	biānfú	‘bat’
蜉蝣	<INSECT+FÚYÓU>	fúyóu	‘mayfly’
蚯蚓	<INSECT+QIŪYǏN>	qiūyǐn	‘earthworm’
璀璨	<JADE+CUĪCÀN>	cuǐcàn	‘brilliant’
玳瑁	<JADE+DÀIMÀO>	dàimào	‘tortoise shell’
鞦韆	<LEATHER+QIŪQIĀN>	qiūqiān	‘swing’
耄耋	<OLD+MÁOZHÌ>	màodié	‘old people’
旖旎	<OVERHANGING+YINÍ>	yǐnǐ	‘fluttering’
倥傯	<PERSON+KŌNGZǑNG>	kǒngzǒng	‘busy’
疙瘩	<SICKNESS+GEDÁ>	gēdā	‘cyst, boil’
徬徨	<STEP+PÁNGHUÁNG>	pánghuáng	‘roam aimlessly’
徜徉	<STEP+SHÀNGYÁNG>	chángyáng	‘roam leisurely’
齟齬	<TEETH+JŪWÚ>	jǔyǔ	‘bickering’
枇杷	<TREE+PIBĀ>	pípá	‘loquat’
檸檬	<TREE+NÍNGMÉNG>	níngméng	‘lemon’
酩酊	<WINE+MÍNGDĪNG>	míngdǐng	‘drunk’
醍醐	<WINE+TÍHÚ>	tíhú	‘clear wine, butterfat’
匍匐	<WRAP+PUFÙ>	púfú	‘crawl’

Knowledge System of the Radical 艸/艹 Encoding Four Causes



蕉蘭芒蒙菌蔓
苦菊茱范荷茅
蕈蔚菲草

Usage

Description

蕃藥蔬菜薪
苑藩藉茭

Parts

茲蒼芳落
茸茂荒薄
芬蒸莊

萌莖芽茄
苗蓮葉

Plants

IS-A/ material

Constitutive /formal

Descriptive/ formal

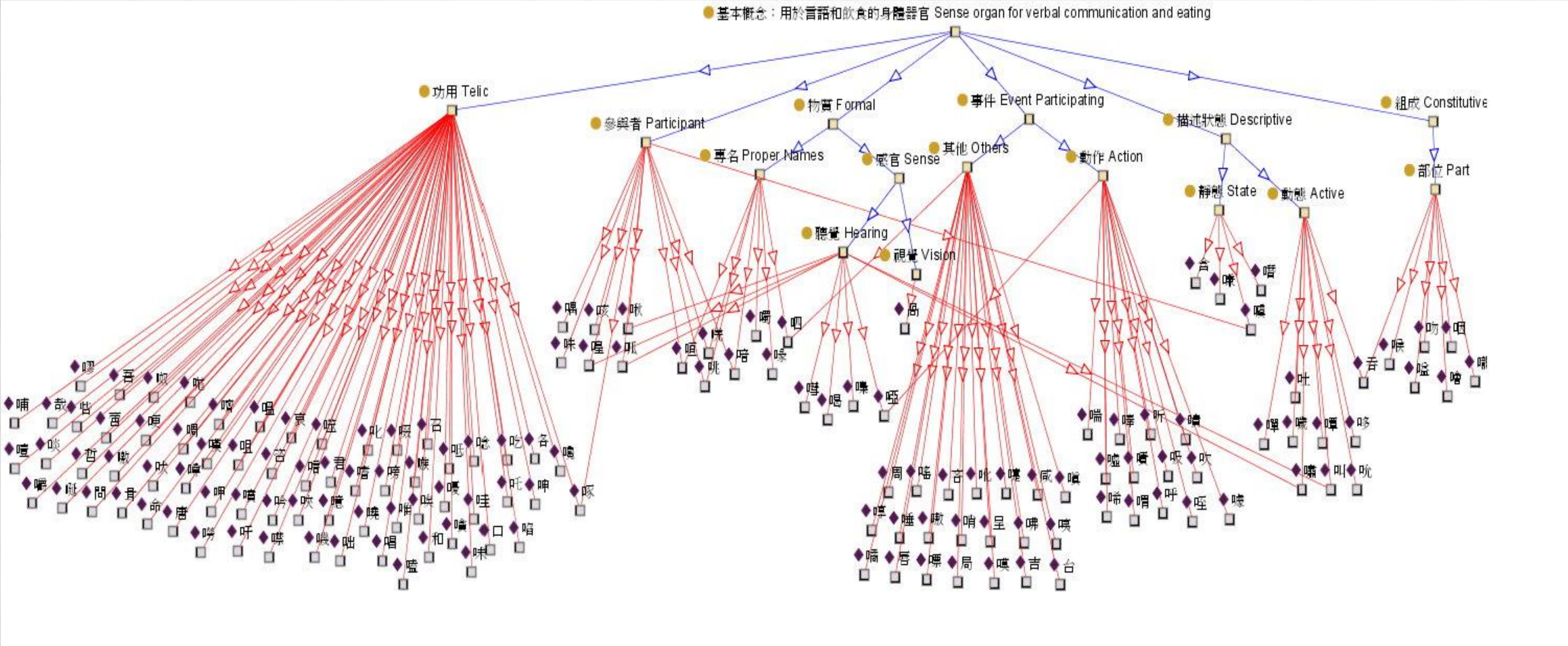
Telic/ final

口 Mouth: To Eat or to Speak

口：所以言者，所以食者。

- The definition of 口 in 說文解字 *ShuoWen Jiezhi* is
  - That which one speaks with
  - That which one eats with

Ontology of 口 mouth



Semantics is the ORL for Chinese

- Characters are organized by semantic concepts and form a linguistic ontology
- The ontology is organized by principles very similar to qualia
- The integrity of the radical orthography system cannot be violated as evidenced by the disyllabic roots

Mandarin Alphabetic Words (MAW)

- They are NOT letter words (as often referred to in the literature)
  - Each Chinese word is a letter word when written in Pinyin romanization (zì is a two letter word, etc.)
  - MAWs in fact cannot be rendered in Pinyin romanization (or in Chinese characters)
  - What defines them is their alphabetic orthography and exceptional phonology

MAW, 字母词

- 爱克斯射线 : calque loan word of X-ray,
- Aìkèsīshèxiàn: lettered word (in Pinyin)

- X射线 (documented 1903)
- [eɪk^{h}s] or [eɪk^{h}ʊ̯sʊ̯] – shèxiàn

  -Calque form (transliteration-translation) should be favored loan word strategy in Chinese yet the competing calque form lost out.

MAW,字母詞

- XianHan 1996: 39 entries
- XianHan 2012: 239 entries
- >60,000 types from Chinese Gigaword Corpus (Huang and Liu 2017)

B股	354	T恤	292	H股	195	B组	165				
G七	151	H5N1型	135	X光	120	B型	100				
A组	93	M2年	90	B組	84	A股	79	K仔	65	G八	62
H5型	59	C组	48	A型	44	C型	43	C組	43		
D组	39	F1赛	36	A级	35	A錢	34				

Some Well-known MAWs

- 阿Q [a kju]
- **吃NG chi1[en ji]**
- A錢 **[eI] qian2**
- AA 制 **[eI eI] zhi4**
- CCTV (China Central TV station 中央电视台)
- KTV **[keI ti vi]**
- **PK**

A seeming anomaly of Chinese, where phonemes NOT in the language's phonological system are freely introduced without creating any stress/change of the phonological system

Idiosyncrasies

Phonologically,

- lack typical lexical tones,
- introducing non-Mandarin syllables (e.g. [k^{h}eɪ] for “K”; [k^{h}jy] for “Q”)
- and phonemes (e.g., /v/ in “V”: [vi])

Tonal or Atonal?

- MAWs bear tone like pitch; but not any of the typical lexical tones of Chinese

-Data by Yuan Jiahong (UPenn) from HKUST speech corpus

- Pitch contour of MAWs show variations and (potentially) dialect influenced adaptation (Ding et al. 2017)

Idiosyncrasies II

- Many are not loan words
  - AA 制; PK, Q, 阿Q,, QQ
- Typically monosyllabic, but alphabet-dependent: x光， wto
- Takes aspects:PK過
- Word formation rules apply
  - productive stem represented either by a character (e.g. A類, B類 C類 with類 lèi ‘type’)
  - an alphabetical letter (e.g. ‘K’ stands for ‘ketamine’ in 拉K, K仔, K粉, K膏, K毒, K癮, K他命).

Speculation on (Non)Adaptation

Differences in ORL in Japanese and Chinese

- Japanese Hiragana/Katakana orthography system is phonology based
  - And Hiragana/Katakana are alternative representations of the same system
- Chinese character orthography is semantics based
  - The introduction of alphabetic writing introduces a diagonal system

Orthographic Integrity

-On surface, at least, what we have been taught to be evidence of phonological integrity of a language system in fact does not hold

-What can be shown instead is orthographical integrity. That is, what is encoded by the standard orthography must follow the regular phonological system, but what is not encoded in standard orthography does not

e.g. *naïve* in English

Challenges, Research Issues

- How are non-phonemic sounds encoded/decoded?
    - Non-lexical conversational sounds
    - Exceptional lexical sounds
- Is there a single homogeneous phonological system; a dominant regular system with small subsystems; or several heterogeneous systems?

How to meet the challenges?

- Corpus-based and experimental phonetic studies to find out the exact nature of distribution/variations of MAWs and conversational sounds in terms of their phonetic properties
- Conduct experiments to construct phonological neighborhood density (PND) model to explore the neighborhood distribution of these exceptional sounds vis-à-vis the ‘regular’ system

Preliminary Research

- 黄居仁 Huang, Chu-Ren,刘洪超 Hongchao Liu. 2017. 基于语料库的汉语字母词自动抽取与分析 Corpus-based Automatic Extraction and Analysis of Mandarin Alphabetic Words. 《云南师范大学学报》(哲学社会科学版) https://www.researchgate.net/publication/318645716_jiyuyuliaokudehanyuzimucizidongchouquyufenxi_Corpus-based_Automatic_Extraction_and_Analysis_of_Mandarin_Alphabetic_Words
- Ding, H.W., Zhang, Y.Y., Liu H.C. and Huang C.R. A Preliminary Phonetic Investigation of Alphabetic Words in Mandarin Chinese. In Proceedings of Interspeech 2017, August 20-24, 2017. Stockholm, Sweden. https://www.researchgate.net/publication/318645592_A_Preliminary_Phonetic_Investigation_of_Alphabetic_Words_in_Mandarin_Chinese

LiNCR: Linguistic and Neuro - Cognitive Resources

http://lincr2018.cbs.polyu.edu.hk/LiNCR_workshop/

lincr2018@gmail.com

- A new generation of language resources which link and aggregate cognitive behavioral, neuroimaging measurement data to a shared set of richly annotated linguistic data.

A LREC 2018 Workshop

- 8 May 2018, co-located with LREC
- The Phoenix Seagaia Resort, Miyazaki, Japan
- Submission deadline: January 15, 2018

- Submission Website:

https://www.softconf.com/lrec2018/LiNCR/

In Vivo Language Resources

Language Resources are

  - Documentation of language use
  - With (linguistic) annotation

- How about *in vivo* data of language use?
  - Brain Activities
  - Behavioral Measurement
  - Hearers’ Reaction/Judgement
- What would *in vivo* Language Resources Look Like
  - How to document, link, use?

The Potsdam Sentence Corpus

- Eye-tracking dataset in English, German, Chinese (two varieties)
- Shared with annotation linkable to experimental result
- Potential to add additional annotation

-Boston, M.F., Hale, J., Kliegl, R., Patil, U. and Vasishth, S., 2008. Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*(1).

-Chinese by Ming Yan, Hua Shu, Jieli Tsai

Wehbe, L., B. Murphy, P. Talukdar, A. Fyshe, A. Ramdas, and T. Mitchell. 2014. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one* 9. 11: e112575.

- Experiment by Machine Learning
- Richly annotated data (Harry Potter Novel): segmentation, syntax, semantics, …
- Global measurement of brain activity in normal reading
- Identification of different brain location for different linguistic sub-processes

Applications in NLP

- Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. 2016. Neural sentiment classification with user and product attention. EMNLP.
- Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. ACL: Short Papers
- Long, Yunfei, Lu Qin, Rong Xiang, Minglei Li and Chu-Ren Huang. 2017. A Cognition Based Attention Model for Sentiment Analysis. *EMNLP 2017*. September 7–11, 2017. Copenhagen, Denmark.

A Few Other Linkable LiNCR's : Linking Behavioral Experiment Data with Corpora

Text_Synaesthesia_SenseExusivity

- Chen, I.-H., Q. Zhao, S. Wang, Y. Long, and C.-R. Huang. 2017. Exclusivity and Competition of Sensory Modalities: Evidence from Mandarin Synaesthesia. Presented at the 2017 International Cognitive Linguistic Conference (ICLC) 10 July. Tartu, Estonia.

Text_Word Segmentation

- Wang, S. C..R. Huang, Y. Yao, and A. Chan. 2017. Word Intuition Agreement among Chinese Speakers: A Mechanical Turk-Based Study. Lingua Sinica

Text_Semantic Transparency

- Wang, S., C.-R. Huang, Y. Yao and A. Chan. 2015. Mechanical Turk-based Experiment vs Laboratory-based Experiment: A Case Study on the Comparison of Semantic Transparency Rating Data. PACLIC-29.

Sample Topics/Challenges

- Corpus selection (Mono/Multi-lingual)
- Ontology/framework for linking annotations in different modalities
- Linking experimental results to linguistically annotated data
- Design for multiple neuro-cognitive experimental platforms to share same linguistic data set
- Aggregation and normalization of data between population with special cognitive

Challenges

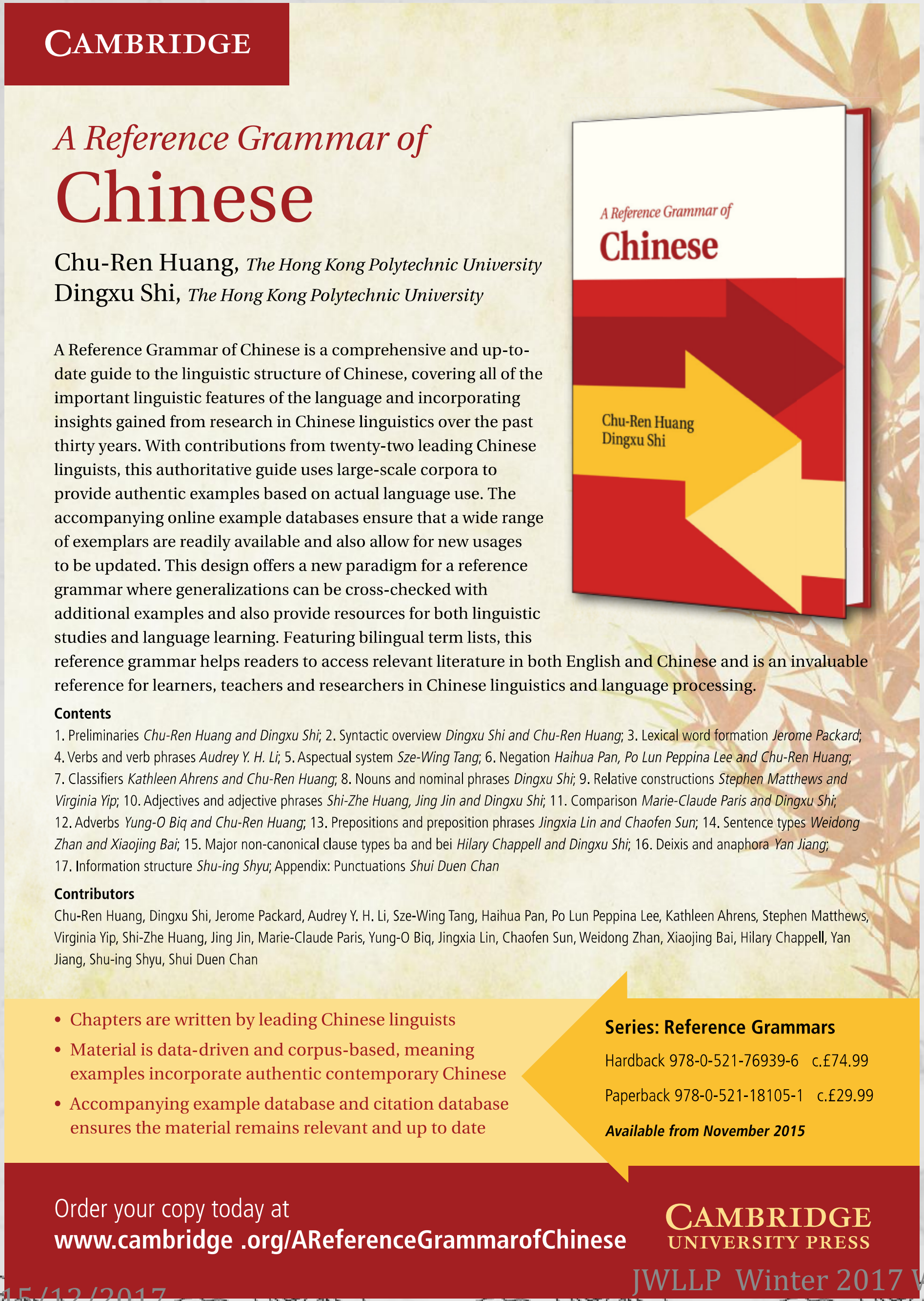
- What to annotate: from phonetics, sub-lexical to discourse level, and probably to the environment and interactive:
- What data to collect and how to collet *in vivo* data (language and its living environment in longitudinal data)
  - http://www.brainwavebank.com/ personalized cognitive activity data collection and aggregation with portable ERP
- How to link/interpret brain activity/behavioral data from different experimental design/paradigm
- Hoe to 'look up' the linked data to check/find possible alternative hypothesis based on result of a purpose-designed experiment and its design (or to check alternative account before design)…

Thank You!

Questions and Comments

In Vivo

Resources: ARGC


A Reference Grammar of
Chinese
Chu-Ren Huang
Dingxu Shi

CAMBRIDGE

A Reference Grammar of
Chinese

Chu-Ren Huang, *The Hong Kong Polytechnic University*
Dingxu Shi, *The Hong Kong Polytechnic University*

A Reference Grammar of Chinese is a comprehensive and up-to-date guide to the linguistic structure of Chinese, covering all of the important linguistic features of the language and incorporating insights gained from research in Chinese linguistics over the past thirty years. With contributions from twenty-two leading Chinese linguists, this authoritative guide uses large-scale corpora to provide authentic examples based on actual language use. The accompanying online example databases ensure that a wide range of exemplars are readily available and also allow for new usages to be updated. This design offers a new paradigm for a reference grammar where generalizations can be cross-checked with additional examples and also provide resources for both linguistic studies and language learning. Featuring bilingual term lists, this reference grammar helps readers to access relevant literature in both English and Chinese and is an invaluable reference for learners, teachers and researchers in Chinese linguistics and language processing.

Contents

Contributors

Chu-Ren Huang, Dingxu Shi, Jerome Packard, Audrey Y. H. Li, Sze-Wing Tang, Haihua Pan, Po Lun Peppina Lee, Kathleen Ahrens, Stephen Matthews, Virginia Yip, Shi-Zhe Huang, Jing Jin, Marie-Claude Paris, Yung-O Biq, Jingxia Lin, Chaofen Sun, Weidong Zhan, Xiaojing Bai, Hilary Chappell, Yan Jiang, Shu-ing Shyu, Shui Duen Chan

- Chapters are written by leading Chinese linguists
- Material is data-driven and corpus-based, meaning examples incorporate authentic contemporary Chinese
- Accompanying example database and citation database ensures the material remains relevant and up to date

Series: Reference Grammars

Hardback 978-0-521-76939-6 c.£74.99

Paperback 978-0-521-18105-1 c.£29.99

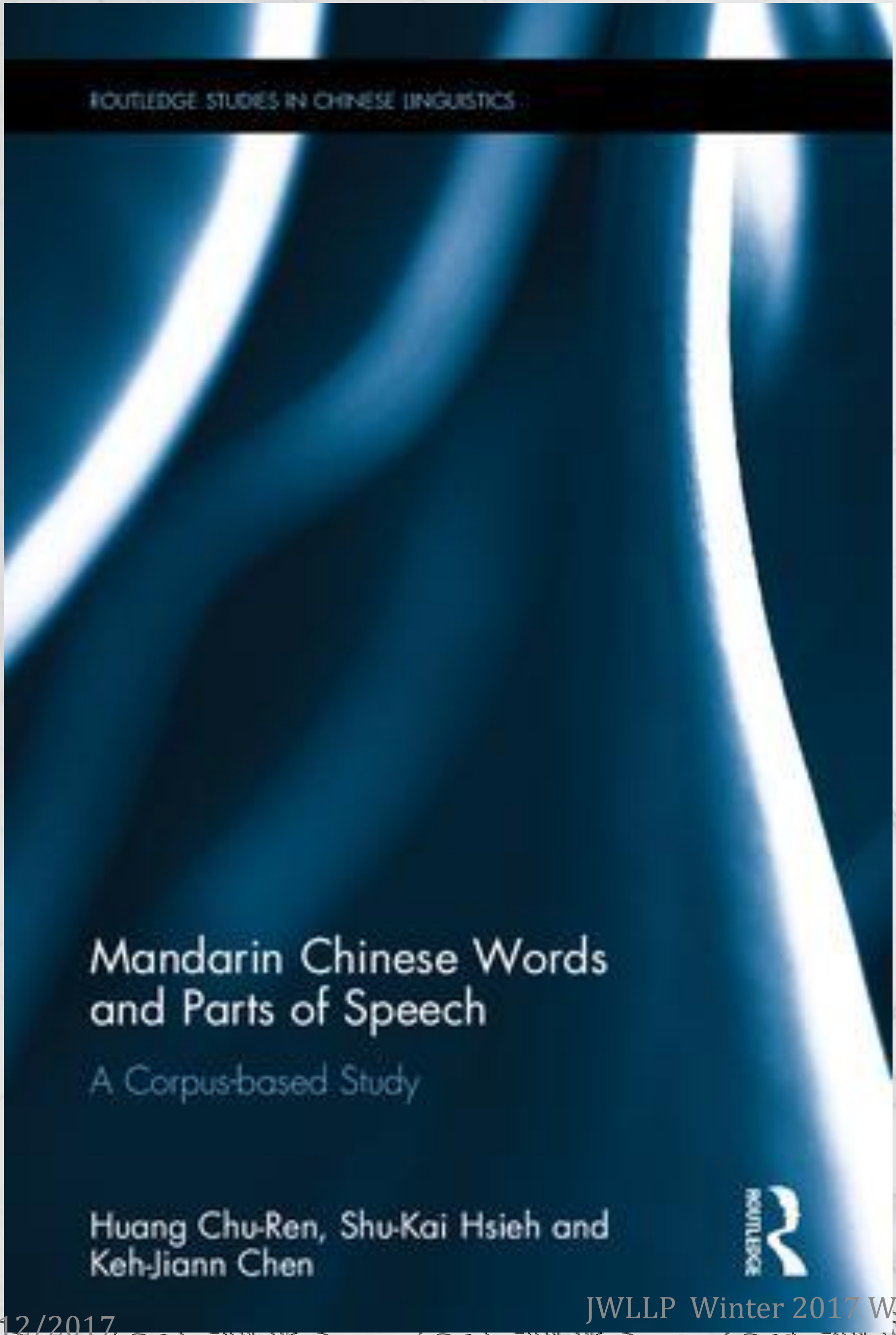
Available from November 2015

Order your copy today at
www.cambridge .org/AReferenceGrammarofChinese

CAMBRIDGE
UNIVERSITY PRESS

Huang, Chu-Ren and Dingxu Shi. 2016. A Reference Grammar of Chinese. Cambridge University Press.

Primary source (unless specified otherwise)



ROUTLEDGE STUDIES IN CHINESE LINGUISTICS

Mandarin Chinese Words and Parts of Speech

A Corpus-based Study

Huang Chu-Ren, Shu-Kai Hsieh and Keh-Jiann Chen

Huang, Chu-Ren, Shu-Kai Hsieh, and Keh-Jiann Chen. 2017. *Mandarin Chinese Words and Parts of Speech: A corpus-based study*. London: Routledge

For Reference

- **LLT Group at PolyU**

  http://llt.cbs.polyu.edu.hk/

- **Google Scholar** https://scholar.google.com.hk/citations?user=zP4DNqgAAAAJ&hl=en

- **ResearchGate**

https://www.researchgate.net/profile/Chu-Ren_Huang