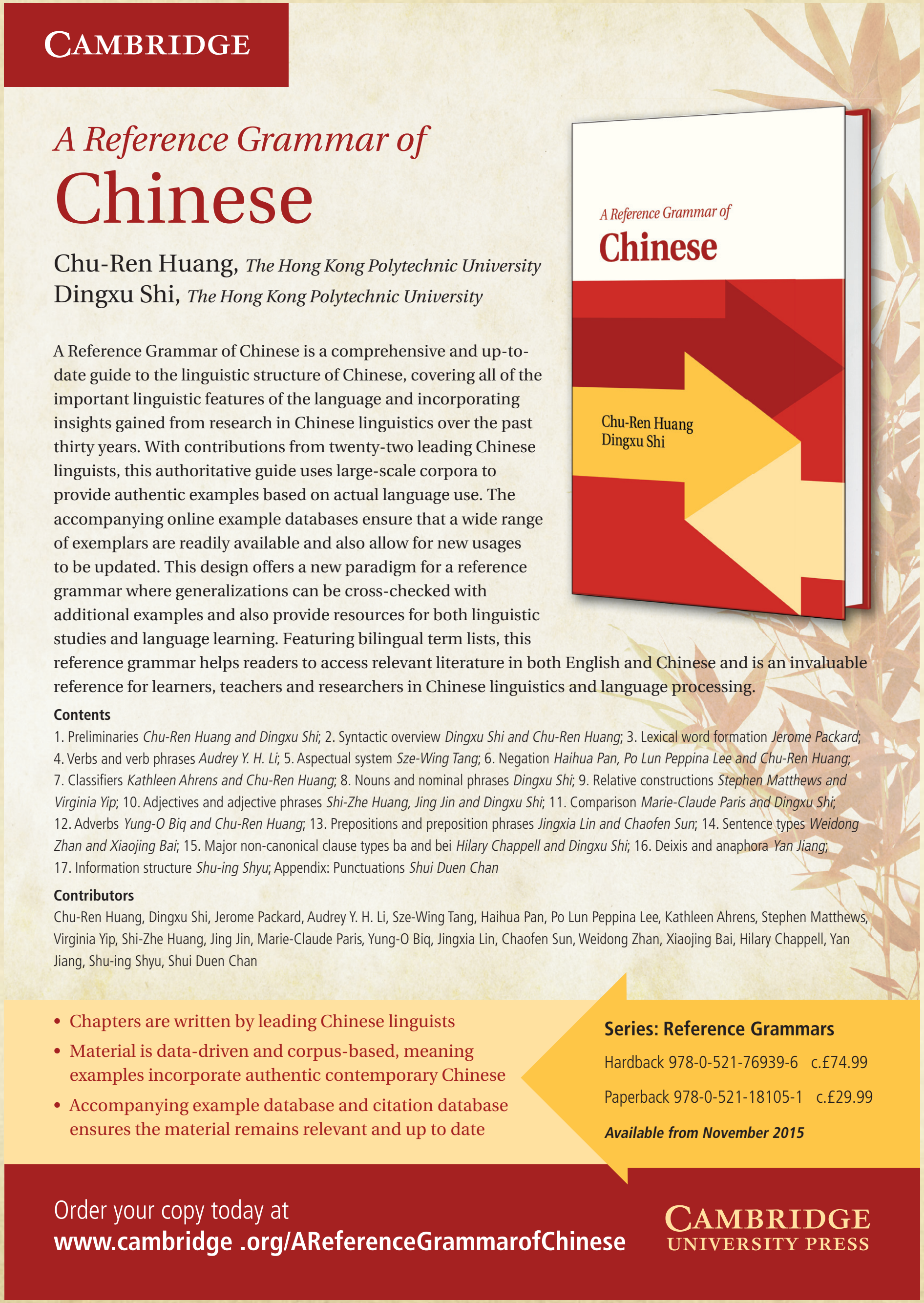
Language Resources from the LLT Group at the Hong Kong Polytechnic University: From phonological neighbourhood to semantic relata, from grammar to emotion

Chu-Ren Huang
The Hong Kong Polytechnic University
29 June 2017, LDC Institute, UPenn

Resources: ARGC

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


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

1. Preliminaries *Chu-Ren Huang and Dingxu Shi*; 2. Syntactic overview *Dingxu Shi and Chu-Ren Huang*; 3. Lexical word formation *Jerome Packard*; 4. Verbs and verb phrases *Audrey Y. H. Li*; 5. Aspectual system *Sze-Wing Tang*; 6. Negation *Haihua Pan, Po Lun Peppina Lee and Chu-Ren Huang*; 7. Classifiers *Kathleen Ahrens and Chu-Ren Huang*; 8. Nouns and nominal phrases *Dingxu Shi*; 9. Relative constructions *Stephen Matthews and Virginia Yip*; 10. Adjectives and adjective phrases *Shi-Zhe Huang, Jing Jin and Dingxu Shi*; 11. Comparison *Marie-Claude Paris and Dingxu Shi*; 12. Adverbs *Yung-O Biq and Chu-Ren Huang*; 13. Prepositions and preposition phrases *Jingxia Lin and Chaofen Sun*; 14. Sentence types *Weidong Zhan and Xiaojing Bai*; 15. Major non-canonical clause types ba and bei *Hilary Chappell and Dingxu Shi*; 16. Deixis and anaphora *Yan Jiang*; 17. Information structure *Shu-ing Shyu*; Appendix: Punctuations *Shui Duen Chan*

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

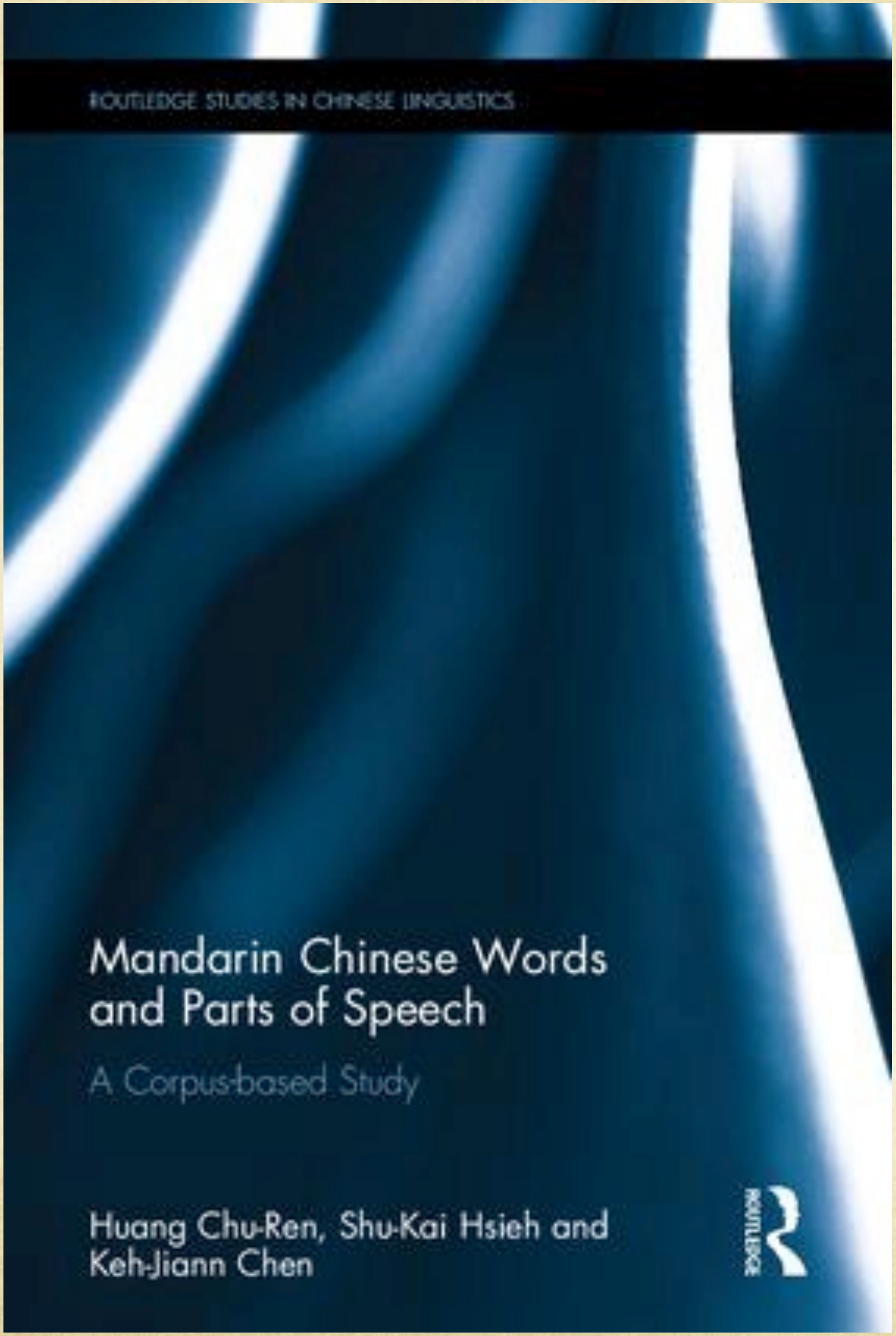
Hardback 978-0-521-76939-6 c.£74.99

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


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

Example Corpus for CUP's A Reference Grammar of Chinese

- http://crg.cbs.polyu.edu.hk

- Xu et al. (2012)

A Cognitive-based Annotation System for Emotion Computing

Ying Chen, *Sophia Lee and Chu-Ren Huang*

LAW III, ACL Workshop, *12 June, 2009*

Event and Emotion Annotation

Goals

- A unified linguistic model of emotion classification
- An annotated emotion corpus for both Chinese and English
- Evaluation of interactions between events and emotions
- Automatic detection and classification of emotions in text

The emotion annotation scheme (1)

- It is a layer just beyond a sentence, and encodes different-level emotion information for a sentence.
- *Elements*
    - *<emotion> element*
        - *<emotionType> element*
            - *<primaryEmotion> element*
    - *<neutral> element*

The emotion annotation scheme (2)

```
<emotion>
  <emotionType name =  "surprise"  keyword ="surprised">
        <primaryEmotion  order =  "1" name =  "surprise"  intensity = "moderate"></primaryEmotion>
  </emotionType>
  <emotionType name = "jealousy"  keyword = “jealousy”>
        <primaryEmotion  order =  "1"  name = "anger" intensity =  "moderate"></primaryEmotion>
        <primaryEmotion  order =  "2"  name =  "fear"   intensity =  "moderate"></primaryEmotion>
  </emotionType>
     <s n = "1"> Hari was surprised at the rush of pure jealousy that swept over her at the mention of
Emily Grenfell .</s>
</emotion>
<neutral>
   <s n = "2"> By law no attempts may be made to hasten death or prolong the life of the sufferer . </s>
</neutral>
<emotion>
  <emotionType>
         <primaryEmotion name =  "sadness"></primaryEmotion>
  </emotionType>
   <s n = "3">He looked hurt when she did n't join him , his emotions transparent as a child 's . </s>
</emotion>
```

The emotion annotation scheme (3)

- The presence of primary emotion can make our annotation scheme more robust.
- Our annotation scheme has the versatility to provide emotion data for different applications.

Corpus creation – analysis (1)

- Emotions expressed with an emotion keyword in the text.
  - An intuitive approach: emotion computing can be satisfactory when the collection of emotion vocabulary is comprehensive.
  - It cannot work well because of the ambiguity of some emotion keywords and the emotion context shift as sentiment (Polanyi and Zaenen, 2004).
- Emotions expressed without any emotion keyword in the text.

Corpus creation – analysis (2)

- The Natural Semantic Metalanguage (NSM) theory is chosen to support emotion computing.
  - One of the prominent cognitive models exploring human emotions.
  - It describes complex and abstract concepts into simpler and concrete ones.
  - It identifies the exact differences and connections between emotion concepts in terms of causes.
  - The proposed dichotomy of good and bad is very intuitive in sentiment analysis

Corpus creation – analysis (3)

- Assumptions of Natural Semantic Metalanguage (NSM)
  - Emotions can be decomposed into semantic primitives;
  - The cause event is essential to emotion classification;
  - Linguistic cues can be derived from different emotions.
- Emotion corpus becomes a collection of emotion stimuli with context.

Emotion corpus creation – Approach (1)

- A pattern-based approach
- Procedure
  - Extract emotion sentences: sentences containing the given emotion keywords are extracted by keyword matching.
  - Delete ambiguous structures: some ambiguous sentences, which contain structures such as negation and modal, are filtered out.
  - Delete ambiguous emotion keywords: if an emotion keyword is very ambiguous, all sentences containing this ambiguous emotion keyword are filtered out.
  - Give emotion tags: each remaining sentence is marked with its emotion tag according to the emotion type which the focus emotion word belongs to.
  - Ignore the focus emotion keywords: for emotion computing, the emotion word is ignored from each sentence.

Emotion corpus creation – Approach (2)

- Polanyi and Zaenen (2004) addressed the issue of polarity-based sentiment context shift, and the similar phenomenon also exists in emotion expressions.
- Two kinds of contextual structures are handled with: the negation structure and the modal structure.

S1 (Neg_Happiness): I am not happy about that.
S2 (Netural): Though the palazzo is our family home, my father had never been very happy there.
S3 (Pos_Happiness): I 've never been so happy.
S4 (Netural): I can die happy if you will look after them when I have gone.
S5 (Netural): Then you could move over there and we'd all be happy.

Neutral corpus creation

- A naïve yet effective algorithm to create a neutral corpus.
  - A sentence is considered as neutral only when the sentence itself and its context (i.e. the previous sentence and the following sentence) do not contain any of the given emotion keywords.

Corpus creation – Corpus selection

- Corpus selection is very important for emotion analysis, such as emotion distribution.
- Three corpora: the Sinica Corpus (Chinese), the Chinese Gigaword Corpus, and the British National Corpus (BNC, English)

Corpus analysis (1)

- The high accuracy of neutral corpus proves that our neutral sentence extraction is effective.

- The accuracy of English emotion corpus is much lower than Chinese emotion corpus.

- It is important for real emotion computing to deal with the emotion expressions which contain emotion keywords and yet are ambiguous.

	Emotion corpus	Neutral corpus
Gigaword	82.17	98.61
Sinica	77.56	98.39
BNC	69.36	99.50

Table 2: The accuracy of the emotion-driven corpora

Corpus analysis (2)

- No-emotion-keyword sentences
  - They do not contain any given emotion keyword.
  - Only about 1% of those sentences express emotions.

E.g.

emotion-keyword sentence - Hari was surprised at the rush of pure jealousy that swept over her at the mention of Emily Grenfell .

no-emotion-keyword sentence - He looked hurt when she did n't join him , his emotions transparent as a child 's .

Emotion computing system-Feature

- Feature:{1,2}-gram words in the focus sentences.
- Emotions are mostly human attitudes or expectations arising from situations, where situations are often expressed in more than a single word.

Emotion system-Corpus (1)

- Emotion type selection for emotion corpus
  - The five primary emotions
  - Nine complex emotions (Chinese), and four complex emotions (English).
  - Other emotion types are renamed as “Other Emotions.”
- Emotion-driven corpus:
  - We combine partial neutral sentences with emotion sentences.
  - 80% as the training data, 10% as the developed data, and 10% as the test data
- Test data
  - *Test data set 1 (TDS 1)*: contains about 10% of the sentences from the complete emotion-driven corpus.
  - *Test data set 2 (TDS 2)*: contains the sentences used for Table 2, which is checked by two annotators.

Emotion system- Experiments (1)

- Our corpus creation approach is effective for emotion computing.
- From the error analysis, it is surprising that for Chinese, “emotion” vs. “neutral” is a common error, whereas, for English, “emotion” vs. “neutral” and “focus emotions” vs. “Other emotions” occupy at least 50%.

	1-gram words	{1,2}-gram words
Chinese TDS 1	53.92	58.75
English TDS 1	44.02	48.20
Chinese TDS 2	37.18	39.95
English TDS 2	33.24	36.31

Table 3: The performances of our system for the test data

Emotion system-Corpus (2)

- Emotion type selection for emotion corpus
  - The five primary emotions
  - For a complex emotion, choose the first primary emotion involved.
- Emotion-driven corpus:
  - We combine partial neutral sentences with emotion sentences.
  - 80% as the training data, 10% as the developed data, and 10% as the test data

Emotion system- Experiments (2)

- There are a big improvement for Chinese (7%), but a little improvement for English (3%).

	1-gram words	{1,2}-gram words
Chinese TDS 1	61.64	65.19
English TDS 1	47.54	51.00
Chinese TDS 2	45.34	47.01
English TDS 2	36.50	37.94

Table 4: The performances of our system for the test data

EVALution 1.0

An Evolving Semantic Dataset for Training and Evaluation of *Distributional Semantic Models*

Enrico Santus, Frances Yung,
Alessandro Lenci & Chu-Ren Huang

EVALution 1.0

- **Freely downloadable** dataset designed for the **training** and the **evaluation** of *DSMs*
  - 7.5K pairs
  - 1.8K relata (63 of which: MWE)
  - 9 semantic relations
  - 10 types of additional information for PAIRS
  - 7 types of additional information for RELATA

Methodology

- **Tuples were**:
  - **extracted** from ConceptNet 5.0 + WordNet 4.0 (8.8M pairs)
  - **filtered** through **automatic methods** to exclude (13K pairs):
    - useless pairs (i.e. !relevant relations, mirrors, !alpha char, etc.)
    - pairs in other resources (i.e. BLESS and Lenci/Benotto).
    - pairs which relata do not occur at least in 3 relations
  - **paraphrased**: "W1 is a kind of W2", "W1 is the opposite of W2"…
  - **judged** through Crowdflower (7.5K pairs)
    - **5 subjects** → 1 (strongly disagree) to 5 (strongly agree)
      - Threshold: 3 positive judgments (>3)
  - **annotated**
    - 5 subjects → PAIRS → semantic tags
    - 2 subjects → RELATA → semantic tags
    - Corpus-based info (frequency, POS, forms, etc.)

Relations, Pairs and Relata

Relation	Pairs	Relata	Template Sentence
IsA	1880	1296	X is a kind of Y
Ant	1600	1144	X can be used as the opposite of Y
Syn	1086	1019	X can be used with the same meaning of Y
Mero - PartOf - MemberOf - MadeOf	1003 *654* *32* *317*	978 *599* *52* *327*	X is... ...part of Y ...member of Y ...made of Y
Entailment	82	132	If X is true, then also Y is true
HasA (possession)	544	460	X can have or can contain Y
HasProperty (attribute)	1297	770	Y is to specify X

Graph Theoretic Approach to Mandarin Syllable Segmentation

- Karl Neergaard, Chu-Ren Huang

The 15th International Symposium on Chinese Languages and Linguistics
(IsCLL-15)

DoWLS: Database of Word Level Statistics

Karl Neergaard, Hongzhi Xu and Chu-Ren Huang
Hong Kong Polytechnic University

- Phonological Neighbourhood Density
- What is the effect of tones in PND?

✓ Goal

✓ A free online database of lexical statistics for Mandarin

✓ Users will be able to query according to

  ✓ Orthography (character or pinyin)

  ✓ Phonology (X-Sampa or IPA)

  ✓ Syllable segmentation

✓ Lexical statistics provided: phoneme length, syllable length, syllable structure (ex: xiang4 = CVVX), dominant POS, frequency, phonological neighborhood density, phonological neighborhood frequency, list of neighbors

Various Chinese language datasets

Shichang Wang, Chu-Ren Huang, Yao Yao and Angel Chan
Hong Kong Polytechnic University

- Establishing the psychological reality of crowdsourced data by comparative study with parallel behavior experiments

✓ SemTransCNC 1.0

  ✓ 1,200 nominal Chinese compounds

  ✓ both overall semantic transparency and constituent semantic transparency data collected by crowdsourcing

✓ WordSegCHC 1.0

  ✓ manual word segmentation data of 152 Chinese sentences (20 to 46 characters without punctuations)

  ✓ 120+ segmentation responses for each sentence

✓ Confirmed that crowdsourcing is a promising tool for linguistic data collection

Phonological networks

Database of Mandarin Neighborhood Statistics (Neergaard & Huang, 2016):

- 14 database files based each on a segmentation schema
  - 7 with tone
  - 7 without tone
- Lexical Statistics relevant to this study
  - Subtitle movie frequency
  - Homophone density (HD)
  - Phonological neighborhood density (PND)
  - Neighborhood frequency (NF)

Distributions of lexical statistics for xiang3 (想) /ɕiaŋ3/

#Units	Schema	Phonological word	Subtitle frequency	Homophone density	PND	Neighborhood frequency
5	C_V_V_X_T	ɕ_i_a_ŋ_3	0.213711	4	9	179,533
4	C_V_V_X	ɕ_i_a_ŋ	0.292855	20	23	328,129
4	C_V_C_T	ɕ_ia_ŋ_3	0.213711	4	8	144,533
3	C_V_C	ɕ_ia_ŋ	0.292855	20	33	260,903
4	C_V_VX_T	ɕ_i_aŋ_3	0.213711	4	12	197,939
3	C_V_VX	ɕ_i_aŋ	0.292855	20	36	387,156
4	CV_V_X_T	ɕi_a_ŋ_3	0.213711	4	23	217,281
3	CV_V_X	ɕi_a_ŋ	0.292855	20	40	722,008
3	CV_VX_T	ɕi_aŋ_3	0.213711	4	26	236,904
2	CV_VX	ɕi_aŋ	0.292855	20	47	804,290
3	C_VVX_T	ɕ_iaŋ_3	0.213711	4	16	212,064
2	C_VVX	ɕ_iaŋ	0.292855	20	51	460,056
2	CVVX_T	ɕiaŋ_3	0.213711	4	262	6,431,641
1	CVVX	ɕiaŋ	0.292855	20	576	21,149,615

Experimental paradigm

Auditory shadowing task

- Procedure
  - Exposure to auditory stimuli
  - Participant repeat stimuli
- Outcome variable
  - Reaction time
- Two processes
  - Sound in: Perception of the auditory stimuli
  - Sound out: Production of the selected lexical item

Method

Participants

- 36 native Mandarin speakers (F: 20)

Stimuli

- 195 Mandarin monosyllabic words (415ms in duration)
  - 4-segment length: 46
  - 3-segment length: 99
  - 2-segment length: 44
  - 1-segment length: 3

Procedure

- 10 practice words
- “下个词” (next word) for 1000ms
- Pause of 3000ms

Model Selection Across Multiple Segmentation Schemas Reveals Mandarin Chinese Phonological Neighborhood Effects

Karl Neergaard and Chu-Ren Huang

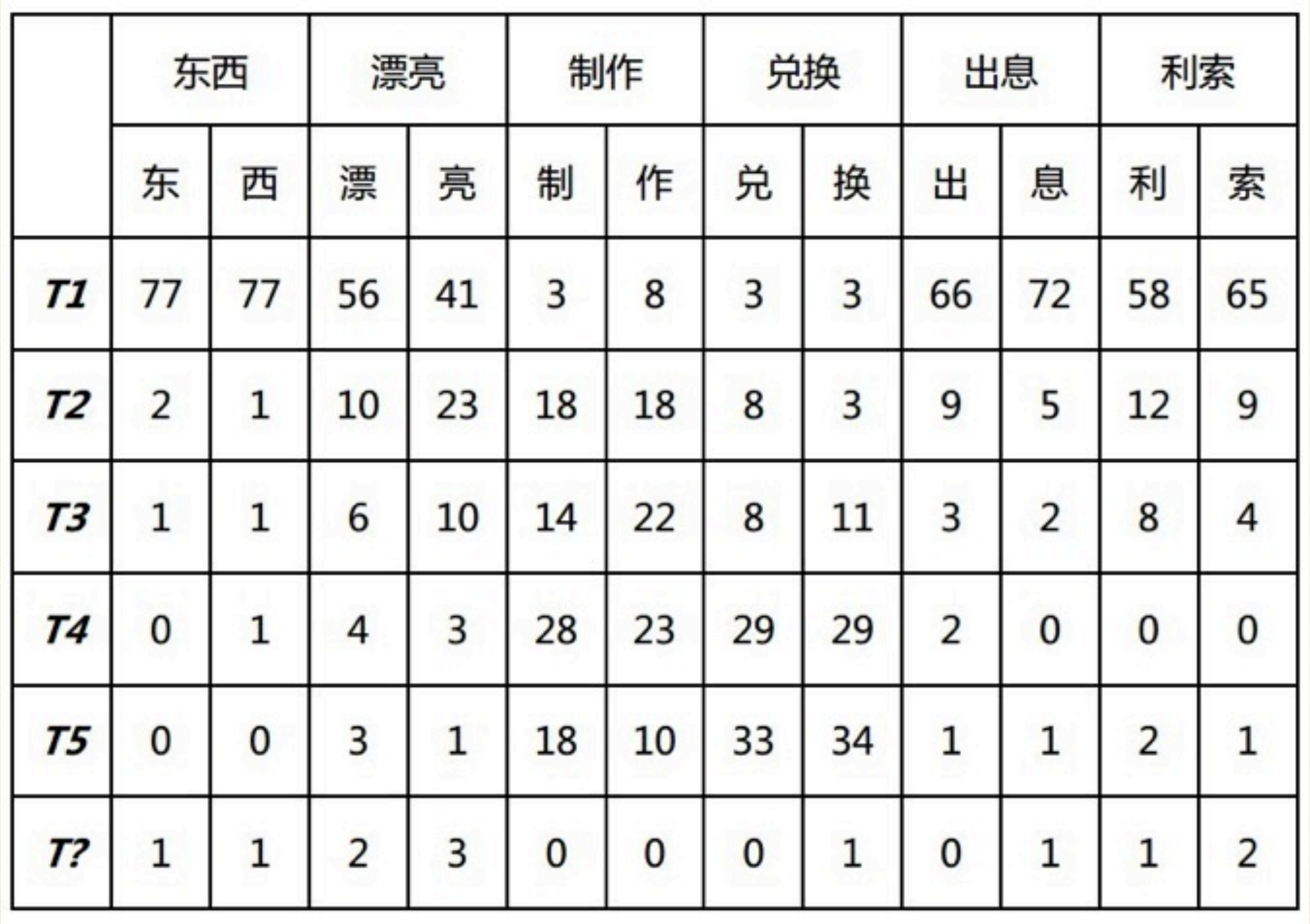
2016 International Meeting of the Psychonmic Society

https://www.postersessiononline.eu/173580348_eu/congresos/PS2016/aula/-PS3_59_PS2016.pdf

Semantic Transparency and Word Segmentation Judgment Databases

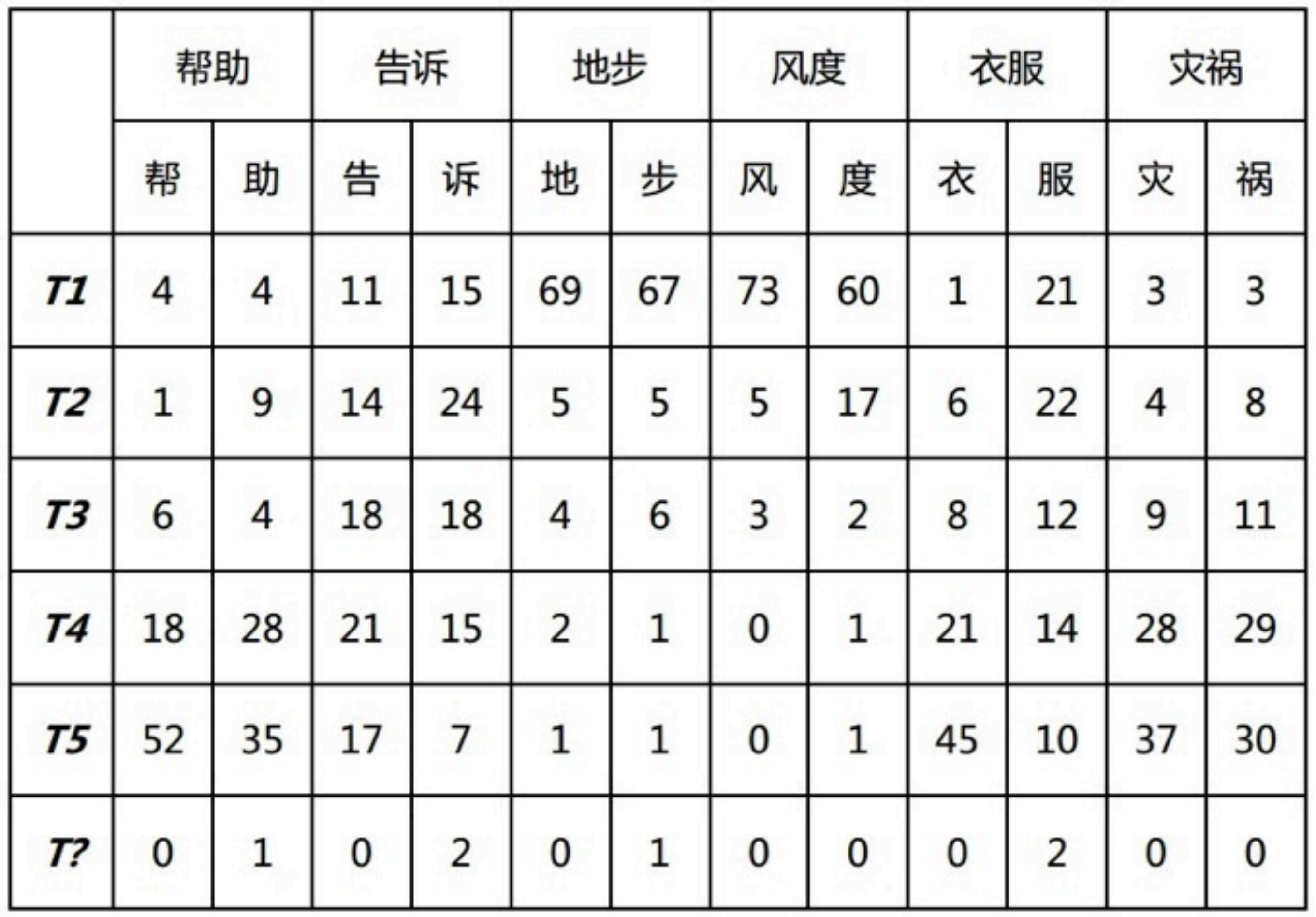
Shichang Wang, Chu-Ren Huang, Angel Chan, and Yao Yao

Semantic Transparency Judgments (1/2)



	东西		漂亮		制作		兑换		出息		利索	
	东	西	漂	亮	制	作	兑	换	出	息	利	索
T1	77	77	56	41	3	8	3	3	66	72	58	65
T2	2	1	10	23	18	18	8	3	9	5	12	9
T3	1	1	6	10	14	22	8	11	3	2	8	4
T4	0	1	4	3	28	23	29	29	2	0	0	0
T5	0	0	3	1	18	10	33	34	1	1	2	1
T?	1	1	2	3	0	0	0	1	0	1	1	2

Semantic Transparency Judgments (2/2)



	帮助		告诉		地步		风度		衣服		灾祸	
	帮	助	告	诉	地	步	风	度	衣	服	灾	祸
T1	4	4	11	15	69	67	73	60	1	21	3	3
T2	1	9	14	24	5	5	5	17	6	22	4	8
T3	6	4	18	18	4	6	3	2	8	12	9	11
T4	18	28	21	15	2	1	0	1	21	14	28	29
T5	52	35	17	7	1	1	0	1	45	10	37	30
T?	0	1	0	2	0	1	0	0	0	2	0	0

Discussion

- Quality control strategy
  - Pre-submission measures
    - Account restrictions (1 judgment per account)
    - IP restrictions (1 judgment per IP)
    - Channel restrictions (only enable reliable channels)
    - Regional restrictions (only enable reliable regions)
    - Checkpoints (insert checkpoints into the questionnaire, check if the answer is right or well-formed, prevent obviously bad submissions)
    - Pricing strategy (avoid high prices, high prices attract cheating, low price strategy)
    - Screening questions (especially open ended questions)
    - Interface language (interface language is a natural barrier)
  - Post-submission measures
    - Majority voting
    - Detect and resist spammers automatically (the monitor program, spammers continually and rapidly submit invalid data points, can fill the dataset with garbage, dangerous!)
    - Manual rejection (BUT Crowdflower doesn't support it)
    - Manual filtration (filter out invalid data points before analysis)

- How to calculate the semantic transparency value of a compound word?

Counter-examples

Human Segmentation Result	Count	%
他/将/来/上海/工作	88	69.29%
他/将来/上海/工作	35	27.56%
他/将/来上海/工作	2	1.57%
他将/来/上海/工作	1	0.79%
他将来/上海工作	1	0.79%
Total:	127	100%