語言大數據與人類集體行為改變：新冠盛行時的語言學研究

Language Big Data and Collective Human Behavioral Changes

第二屆中國語言學嶺南書院（2021年）

2021年12月8日 14:30

Why COVID? Each Neologism is Language at Work Reflecting Shared Concepts

Standard Answer: CORONAVIRUS DISEASE

- Unique stem , easy to pronounce etc.

- But why a 2+2_1 acronym?
  - Why not CD, CORD, CoD, Codi, CorDis, CVD, etc.?

- Corvid: A bird from the *Corvidae* family
  - That is, crows, ravens…… An omen, a harbinger of bad news

- Language *in vivo* involves multiple motivations, the convergence of many causes in form and meaning

嶺南書院語言大數據_CRHuang 08/12/21

聽其言,觀其行,人焉廋哉

A synthesis of two quotes from

- 聽其言也，觀其眸子，人焉廋哉？ Mencius
- 視其所以，觀其所由，察其所安，人焉叟哉？人焉叟哉？ Confucius

Listen to words said

Observe how they act

What can a person hide from you?

嶺南書院語言大數據_CRHuang 08/12/21

"URBAN BUNNY" BY JOHN BENSON, LICENSED UNDER CC BY 2.0
HTTPS://WWW.FLICKR.COM/PHOTOS/J_BENSON/48026729218

Language as Data: Quine's 'Gavagai!'
蒯因 的 詞義未定論

WHAT IS GAVAGAI ?

- WVO QUINE (1960) **WORD AND OBJECT**
-
  - RABBIT, HARE, SMALL ANIMAL, PRAIRIE FAIRY, PET, BUGS BUNNY…..
  - BIG EARS, BUCK TEETH, BIG EYES, ANGORA, SH
  - GAME,
  - …
  - 'LOOK' 'IT IS HOPPING' ,
  - "SO CUTE/FAST/FAT,…."
  - 'CATCH IT' 'CHASE IT AWAY!' 'WHERE IS MY PHONE!' 'XXX### (CURSE, SPELL, ….)"
- A SINGLE DATA POINT HAS MANY INTERPRETATIONS AND CAN ONLY BE 'UNDERSTOOD' WITH CONTEXT

What Is Language? Which Language?

- Language1: the collective sum of all linguistic activities and abilities of all human being in the world from past to future; (i.e. *Langage*)
- Language2: the shared system of linguistic communication by a specific people, throughout the duration of existence of the people and the system; (i.e. Langue, such as Chinese, French and English) or: *language1 as used by a certain people*
- **Language3: language2 as used by a specific people at a specific period of time (i.e. Langue, such as middle English, Archaic Chinese, Middle Chinese, Modern Chinese etc.)**
- Language4: (or dialect) language2/3 as used by a specific people at a specific region and developed to an extend that it has separate cultural identity and may (or may not) be mutually-unintelligible with other language4's
- Language5: language2/3 as used by a people at a specific region to the extent that it can be differentiate from others (i.e. American English, Brazilian Portuguese, Taiwan mandarin etc.)
- Language6: language3 as used by a specific group of people with related to their shared special purpose and/or profession (i.e. Academic English, tourist Japanese, business Chinese etc.)
- …
- Language_n: language3 as used for a special occasional, often with specific social functions or intension mind by the user (romantic language, polite language, persuasive language)
- Language_{n+1}: the word choices of a specific speaker… ('watch your language!')

上善若水
Language Behaves just like Water

Language as the sum of all 'languages'

- Permeates *in the tempo-spatial continuum and over the population over multiple individuals*
- Without a definitive shape
- To understand language, we must understand the dynamicity of language at work
  - Instead of any of the snapshots
  - Like claiming that the shape of water is like a missile, holding a Watson's bottle as evidence
- **To know language by how it works, not (just) by how it looks**

Adding water to water, the result is still water….

The Stories Continue, Language and Human Behavior Changes (during Covid-19)

- Jiang, M., Shen, X. Y., Ahrens, K., & Huang, C. R. (2021). Neologisms Are Epidemic: Modeling The Life Cycle Of Neologisms In China 2008-2016. *Plos One*, *16*(2), E0245984 Https://Doi.Org/10.1371/Journal.Pone.0245984
- Lei, S., Yang, R., & Huang, C. R. 2021. Emergent Neologism: A Study Of An Emerging Meaning With Competing Forms Based On The First Six Months Of COVID-19. *Lingua*, 103095. *Lingua Editor's Choice*. Https://Doi.Org/10.1016/J.Lingua.2021.103095
- Wang, X., & Huang, C. R. 2021. From Contact Prevention To Social Distancing: The Co-evolution Of Bilingual Neologisms And Public Health Campaigns In Two Cities In The Time Of COVID-19. *SAGE Open*, *11*(3), 21582440211031556. Https://Doi.Org/10.1177/21582440211031556
- Su, Q., Liu, P., Wei, W., Zhu, S., & Huang, C. R. 2021. Occupational Gender Segregation And Gendered Language In A Language Without Gender: Trends, Variations, Implications For Social Development In China. *Humanities And Social Sciences Communications*, *8*, 133. Https://Www.Nature.Com/Articles/S41599-021-00799-6

KUDOS: SHARING GREAT NEWS
Two PolyU Papers Are Lingua Editor's Choice 2021

- The Following Two Papers From PolyU (CBS And ENGL) Are Selected as *Lingua Editor's Choice*, Papers That Are Considered To Have Significant Potential Impacts In The Field: These papers are OA till December 2021

Https://Www.Journals.Elsevier.Com/Lingua/Editors-choice/Editors-choice-articles-lingua

- Framing Metaphor Use Over Time: 'Free Economy' Metaphors In Hong Kong Political Discourse (1997–2017)

Winnie Huiheng Zeng, Christian Burgers, Kathleen Ahrens

- Emergent Neologism: A Study Of An Emerging Meaning With Competing Forms Based On The First Six Months Of Covid-19

Siyu Lei, Ruiying Yang, Chu-Ren Huang

Papers And Books To Appear

- Huang, Chu-ren, Yen-hwei Lin, I-hsuan Chen, & Yu-yin Hsu. 2022 (In Press). *The Cambridge Handbook Of Chinese Linguistics*. Https://Www.Cambridge.Org/Core/Books/Cambridge-handbook-of-chinese-linguistics/033DC9E2ECFED54B9A8A42EFD5E33BBA
- Wang, Shan & Chu-ren Huang. To Appear. Social Changes Through The Lens Of Language: A Big Data Study Of Chinese Modal Verbs. Accepted By Plos One.
- Zhu Yongping & Chu-ren Huang. Accepted. A Student Grammar Of Chinese. Cambridge University Press.
- 蒋梦晗menghan Jiang,黄居仁 Chu-ren Huang. To Appear. 海峡两岸汉语动宾复合词的及物性差异--基于语料库驱动方法的对比研究transitivity Variations Of Mandarin Chinese VO Compounds: A Corpus-driven Comparative Study.*Zhongguoyuwen*中国语文

To Study Language
With All Its Permeation And Dynamicity

WE STUDY LANGUAGE

In Context

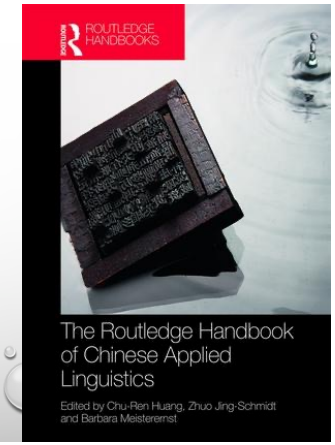
In Action

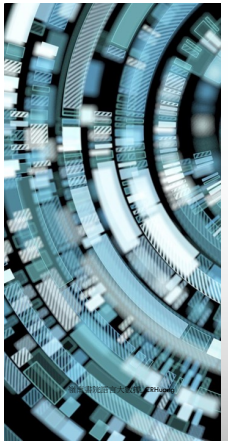
In Vivo

- ROUTLEDGE WEBSITE:
  - HTTPS://WWW.ROUTLEDGE.COM/THE-ROUTLEDGE-HANDBOOK-OF-CHINESE-APPLIED-LINGUISTICS/CHU-REN-JING-SCHMIDT-MEISTERERNST/P/BOOK/9781138450732
- AMAZON WEBSITE:
  - HTTPS://WWW.AMAZON.COM/ROUTLEDGE-HANDBOOK-CHINESE-APPLIED-LINGUISTICS/DP/1138650730/REF=SR_1_1?IE=UTF8&QID=1534211030&SR=8-1&KEYWORDS=THE+ROUTLEDGE+HANDBOOK+OF+CHINESE+APPLIED+LINGUISTICS.

The Routledge Handbook of Chinese Applied Linguistics

EDITED BY CHU-REN HUANG, ZHUO JING-SCHMIDT, & BARBARA MEISTERERNST





LANGUAGE *IN VIVO*

>Studies that are **in vivo** are those in which the effects of various biological entities are tested on whole, living organisms in their natural functioning environment (not *in vitro*, i.e. in a test tube or in a lab)

>>Language studies **in vivo** (huang et al. 2019, RHCAL) would then be studying language in its various natural environment of use, from mind and body, to communication and interaction, all the way to society and culture.

>>> In this view, language sciences should taking into consideration the aggregation of linguistic data (big data); the neuro-cognitive mechanism that produces, perceives and acts on such data; the function and operation of language in inter-personal interaction; and the role language plays within and between communities.

WHY Language Big Data

- Adding Big Data To Big Data Does Not Change The Nature Of The Data
- The Nature Of Permeation And Dynamicity Can Only Be Captured When Data Is Big Enough
  - To Cover Tempo-spatial Variation And Shared Conceptualization Of Multiple Individuals
  - And for robustness, in the sense of being able to take the sum of the data regardless of the so-called 'ungrammatical' usages
  - Recall that language by definition is a mean of mutually agree upon tools of communication and one of the 'most democratic' human devices. Which means that, when data is big enough,
    - A small fraction of 'mistakes' can be ignored
    - A significant representation of shared usage should be considered as the norm and cannot be ignored.

WHAT IS LANGUAGE BIG DATA

Language big data is simply a sizable sum of our language use
- The issue is not how big it is

1

The issue is which (fraction of the) SUM we should be looking at

2

And what are the appropriate tools to explore this particular 'sum'

IN TIMES OF CHANGES AND CHALLENGES

Looking At Language Big Data Will Be The Most Effective And Straight Forward Way To Look At

The Collective Behavior (And Behavioral Changes)	The Shared Observation And Perception Of Risks	And The Shared Patterns Of Different Coping/Decisions Made Facing Risks…

Language Big Data (Such As Online Searches And Posts) Are Available Before Pathological Or Epidemiological Data

Methodological Issues:

Linguistically Motivated and Big Data Driven

- All the books in the world > 130 million titles in 2010

- The Google Book corpus in 2019: >50 million titles
- Estimated to cover >30% of all books in the world

Google Ngram Viewer https://books.google.com/ngrams

Google Trends
Baidu Index 百度指數

Https://Index.Baidu.Com/

Https://Trends.Google.Com/Trends/

- The World on the Web
- Any time, all the time, Real time
- Any where and every where

FIRST TIME IN HUMAN HISTORY

A (NEARLY) FULL REAL TIME DOCUMENTATION OF WHAT HAPPENED IN REAL TIME

What We (And Other Linguists) Have Done WRT Covid-19

And Other Collective Human Behavior Or Environmental Changes

- Preliminaries
- Establish google book corpus as a tool for monitoring social changes

Li, Longxing, Chu-Ren Huang, & Vincent Xian Wang. 2020. Lexical Competition And Change: A Corpus-assisted Investigation Of Gambling And Gaming In The Past Centuries. SAGE Open:10.3. Https://Doi.Org/10.1177/2158244020951272

- Neologisms spread like epidemic as they require acceptance of human hosts

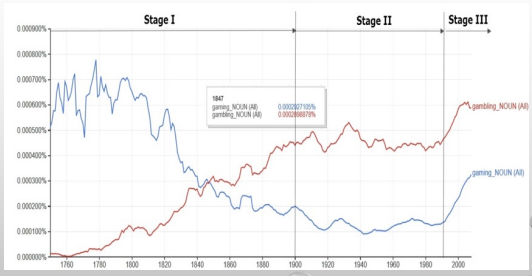
Jiang Menghan, Xiang Ying Shen, Kathleen Ahrens, Chu-ren Huang, 2021. Neologisms Are Epidemic: Modeling The Life Cycle Of Neologisms In China 2008-2016. Plos ONE 16(2): E0245984. Https://Doi.Org/10.1371/Journal.Pone.0245984

LI, LONGXING ET AL. (2020) SAGE OPEN
HTTPS://DOI.ORG/10.1177/2158244020951272

THIS PAPER INVESTIGATES THE INTERPLAY OF LEXICAL COMPETITION AND SOCIOHISTORICAL EVENTS THROUGH A CLOSE EXAMINATION OF THE USE OF GAMBLING AND GAMING BASED ON LARGE-SCALE SYNCHRONIC AND DIACHRONIC CORPORA. WE FIRST SET THE BACKGROUND FOR COMPARISON THROUGH A SYNCHRONIC STUDY OF THE COLLOCATIONAL PATTERNS AND GRAMMATICAL RELATIONS OF THE TWO WORDS USING SKETCH ENGINE. WE SHOW THAT GAMBLING TENDS TO BE ASSOCIATED WITH NEGATIVELY PERCEIVED ACTIVITIES AND STRONG DISAPPROVAL, WHILE GAMING TENDS TO COLLOCATE WITH RECREATIONAL ACTIVITIES, BUSINESS, AND TECHNOLOGY. USING GOOGLE BOOKS NGRAM VIEWER, WE FOCUS ON THE DRASTIC DIACHRONIC CHANGE IN USE OF THE TWO WORDS, FROM COMPETITION TO CO-DEVELOPMENT. BASED ON CORPORA TRENDS, WE CORRELATE THE RISE AND FALL OF THE TWO WORDS AND THE CHANGE IN THEIR COMPETITION RELATION TO PARTICULAR SOCIO-HISTORICAL EVENTS: GOLD RUSHES, SPORTS BETTING, THE POPULARITY OF VIDEO GAMES, AND THE GAMING INDUSTRY BOOM. THE CLASSICAL COMPETITION MODEL OF NEAR SYNONYMS REMAINED VALID UNTIL RECENT SOCIO-ECONOMIC EVENTS INTRODUCED ADDITIONAL AND UNIQUE MEANINGS FOR BOTH WORDS. THE PAPER THUS SHOWS THAT LINGUISTIC VARIATIONS AS COLLECTIVE HUMAN BEHAVIOR CHANGES CAN BE LEVERAGED TO EVIDENCE OTHER COLLECTIVE HUMAN BEHAVIOR CHANGES.

LANGUAGE AS BIG DATA I
GOOGLE BOOKS/NGRAM VIEWER

GOOGLE NGRAM VIEWER OF GAMBLING AND GAMING (1750-2008) [FROM LI ET AL. 2020]



LANGUAGE AS BIG DATA IIA SKETCHENGINE

Coordination relation of GAMBLING and GAMING

[Table 2 from Li et al. 2020]

"Gambling/gaming" and/or...	Counts		Typicality score (log2foe)	
	GAMBLING	GAMING	GAMBLING	GAMING
prostitution	845	16	8.6	2.6
sex	710	28	6.1	1.3
smoking	342	16	6.2	1.6
alcohol	1,202	54	6.8	2.3
drinking	1,271	71	8.4	4.1
pornography	275	52	8.0	4.3
addiction	281	60	6.9	3.5
porn	247	49	6.8	4.2
poker	646	214	7.4	5.7
casino	1501	546	8.1	6.5
lottery	364	233	7.5	6.6
gambling	1,086	660	8.6	7.6
betting	275	273	7.3	7.0
gaming	660	1,252	7.6	8.4
racing	159	326	5.4	6.3
entertainment	371	1,386	5.1	6.9
microgaming	10	194	2.7	6.6
playtech	9	181	2.6	6.5
networking	23	471	2.1	6.3
computing	0	219	--	5.8
playstation	0	140	--	6.9
online	0	169	--	6.1
multimedia	0	277	--	6.7
browsing	0	379	--	7.2

	Definitions		X is a ...		... is a X		Total	
Representation	GAMBLING	GAMING	GAMBLING	GAMING	GAMBLING	GAMING	GAMBLING	GAMING
game	116	21	269	67	20	32	405	120
activity	79	25	262	107			341	132
fun	93	49					93	49
entertainment	22	12	72	17			94	29
pastime	29	18	81	59			110	77
business	57	45	160	99			217	144
trend	9	9					9	9
industry	33	59	90	130			123	189
casino	19	41	57	45	21	45	97	131
hobby	9	63	50	147			59	210
addiction			128	17			128	17
pursuit			17	12			17	12
sector			12	18			12	18
market			24	48			24	48
passion			12	43			12	43
issue					17	13	17	13

LANGUAGE AS BIG DATA IIB SKETCHENGINE

- FREQUENCY OF COMMON PATTERNS OF GAMBLING AND GAMING IN THREE GRAMRELS
- [TABLE 7 FROM LI ET AL. 2020]

modifiers GAMBLING	of	freq.	score	modifiers GAMING	of	freq.	score
seahawks		7	4.0	alternate		9	4.5
football		7	4.0	football		6	3.8
texans		6	4.0	preseason		13	3.5
ravans		6	3.2	playoff		18	2.2
table		103	2.8	nfc		7	2.2
bet		11	2.7	rugby		6	0.9
fatigue		13	2.5	hocky		8	0.4
winner		12	1.5				
cyber		14	1.2				

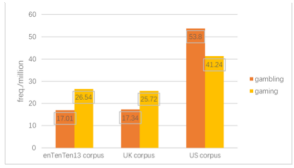
LANGUAGE AS BIG DATA IIC SKETCHENGINE

- ONLY PATTERNS OF 'MODIFIERS OF GAMBLING/ GAMING' IN THE US SUB-CORPUS
- [TABLE 7 FROM LI ET AL. 2020]

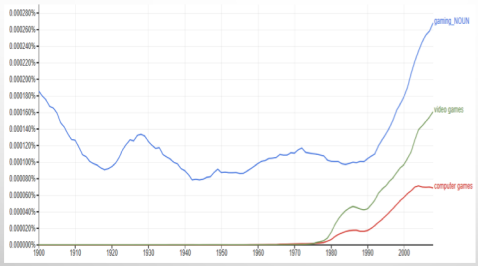
LANGUAGE AS BIG DATA IID SKETCHENGINE

	UK (1,182,251,470)	US (164,190,640)	enTenTen13 (19 billion)
GAMBLING	20,498	8,834	386,782
GAMING	30,402	6,772	603,436

Frequency of occurrence of GAMBLING and GAMING in enTenTen13
[Table 9 from Li et al. 2020]



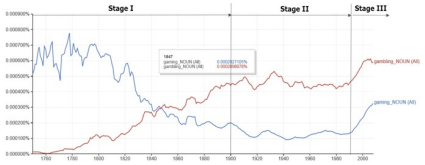
Normalized frequency of GAMBLING and GAMING in enTenTen13
[Figure 1 from Li et al. 2020]



THE RISE OF GAMING

- 'GAMBLING', 'VIDEO GAMES' AND 'COMPUTER GAMES' IN ENGLISH (1900-2008, GNV)
- [FIGURE 3 FROM LI ET AL. 2020]

LANGUAGE AS BIG DATA I GOOGLE BOOKS/NGRAM VIEWER



GOOGLE NGRAM VIEWER OF GAMBLING AND GAMING (1750-2008) [FROM LI ET AL. 2020]

SOCIO-HISTORICAL EVENTS:

- GOLD RUSHES (GLOBAL): 1848-1855
- SPORTS BETTING:
  - THE PROFESSIONAL AND AMATEUR SPORTS PROTECTION ACT OF 1992, (BAR SPORTS BETTING, BUT LEGALIZE IN 4 STATES IN USA: DELAWARE, MONTANA, NEVADA, AND OREGON). LEGALIZED IN 2018
- THE POPULARITY OF VIDEO GAMES, AND THE GAMING INDUSTRY BOOM
  - NINTENDO 1985 (SEVERAL WAVES OF GAMING PROGRAMMES BEFORE ALL CRASHED)

Jiang M, shen XY, ahrens K, huang CR (2021) neologisms are epidemic: modeling the life cycle of neologisms in china 2008-2016. PLOS ONE 16(2): e0245984. Https://doi.Org/10.1371/journal.Pone.0245984

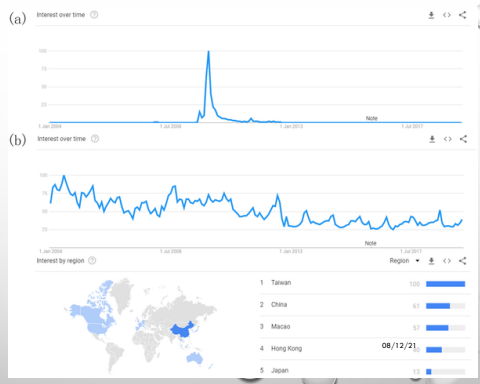


FIG. 1. INTERFACE OF GOOGLE TRENDS.

- (A) The Graph Shows The Search Result Of A Sample Neologism 蜗居 Wo1ju1 "Living Within A Snail's Shell" In Google Trends With A Typical Sharp Rise And Decay Pattern.
- (B) The Snapshot Of The Search Result For The Neologism 雷 Lei2 "Thunder, Describing A Person Getting Shocked By Something Absurd". This Word's Search Frequency Variation Cannot Be Included In The Paradigm Of The Sharp Rise And Decay Pattern. The Spatial Distribution Of The Search Frequency All Over The World Are Also Provided On The Web Page.

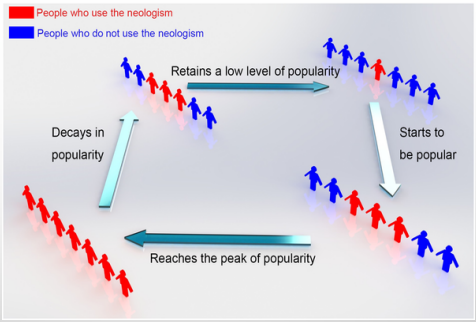
https://journals.Plos.Org/plosone/article?Id=10.1371/journal.Pone.0245984

Why neologisms spread like epidemic: The SIR Model :

- Susceptible
- Infected
- Recovered

The similarity in human mediation; and the human 'decision' to take or not

FIG 2. FOUR STAGES OF NEOLOGISMS' LIFETIME.

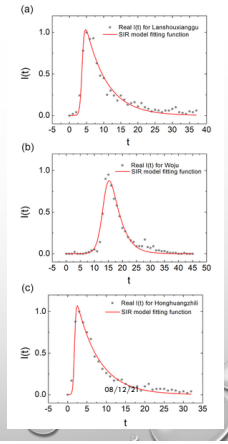


Jiang M, Shen XY, Ahrens K, Huang CR (2021) Neologisms are epidemic: Modeling the life cycle of neologisms in China 2008-2016. PLOS ONE 16(2): e0245984. https://doi.org/10.1371/journal.pone.0245984

The SIR model fitting results are as illustrated. In the picture, the *P*(*t*) data obtained from Google Trends are grey stars. The SIR model fitting functions are denoted by the red lines. (a),(b),(c) gives the data of 蓝瘦香菇 lan2shou4xiang1gu1 "too sad to cry", 蜗居 wo1ju1 "living within a snail's shell/small room", 洪荒之力 hong2huang1zhi1li4 "the force from primitive period" respectively, as well as the SIR fitting functions.

Fig 4. The performance of the model fitting

Jiang M, Shen XY, Ahrens K, Huang CR (2021) Neologisms are epidemic: Modeling the life cycle of neologisms in China 2008-2016. PLOS ONE 16(2): e0245984. https://doi.org/10.1371/journal.pone.0245984



JIANG ET AL (2020)

- The Life Time Of Neologisms Are More Like Epidemics, And Not Like Memes
  - Note That Many Call Neologisms Word Memes
  - But Memetic Models Predict That It Will Continue To Propagate Until It Dies
- The Similarity In Life Time Model Suggest That Neologisms Can Be Used To Model Epidemics

LINGUISTIC STUDIES ON COVID-19 BY OTHERS

- CHATER, NICK. "FACING UP TO THE UNCERTAINTIES OF COVID-19." *NATURE HUMAN BEHAVIOUR* 4, NO. 5 (2020): 439-439.
- VAN BAVEL, JAY J., KATHERINE BAICKER, PAULO S. BOGGIO, VALERIO CAPRARO, ALEKSANDRA CICHOCKA, MINA CIKARA, MOLLY J. CROCKETT ET AL. "USING SOCIAL AND BEHAVIOURAL SCIENCE TO SUPPORT COVID-19 PANDEMIC RESPONSE." *NATURE HUMAN BEHAVIOUR* 4, NO. 5 (2020): 460-471.
- ASLAM, FAHEEM, TAHIR MUMTAZ AWAN, JABIR HUSSAIN SYED, AISHA KASHIF, AND MAHWISH PARVEEN. "SENTIMENTS AND EMOTIONS EVOKED BY NEWS HEADLINES OF CORONAVIRUS DISEASE (COVID-19) OUTBREAK." *HUMANITIES AND SOCIAL SCIENCES COMMUNICATIONS* 7, NO. 1 (2020): 1-9.
- GALLOTTI, RICCARDO, FRANCESCO VALLE, NICOLA CASTALDO, PIERLUIGI SACCO, AND MANLIO DE DOMENICO. "ASSESSING THE RISKS OF 'INFODEMICS' IN RESPONSE TO COVID-19 EPIDEMICS." *NATURE HUMAN BEHAVIOUR* 4, NO. 12 (2020): 1285-1293.

OUR (AND PARTNERS') WORK RELATED TO COVID-19

- Lang, J., Erickson, W. W., & Jing-schmidt, Z. (2021). # Maskon!# Maskoff! Digital Polarization Of Mask-wearing In The United States During COVID-19. *Plos One*, *16*(4), E0250817.
- Meng, J. I., Huang, C. R., & Hall, B. (2021). Flu Vaccination Due To COVID 19 As A Risk-aversion Health Measure Among Low-risk Populations.
- Ngai, Cindy Sing Bik, Rita Gill Singh, W. Z. Lu, And Alex Chun Koon. "Grappling With The COVID-19 Health Crisis: Content Analysis Of Communication Strategies And Their Effects On Public Engagement On Social Media." (2020).
- Chen, Xi, Vincent Xian Wang, Chu-Ren Huang. 2021. *Themes And Sentiments Of Online Comments Under COVID-19: A Case Study Of Macau*. Presented At *The 22nd Chinese Lexical Semantics Workshop (CLSW2021)*. 15-16 May 2021

MODELING COVID DEVELOPMENTS WITH LINGUISTIC ANALYSIS

- Lei, Siyu Lei, Ruiying Yang, Chu-Ren Huang. 2021. Emergent Neologism: A Study Of An Emerging Meaning With Competing Forms Based On The First Six Months Of COVID-19. *Lingua*. 258: 103095. Https://Www.Journals.Elsevier.Com/Lingua/Editors-choice/Editors-choice-articles-lingua.
- Wang, Xiaowen, & Chu-Ren Huang. 2021. From Contact Prevention To Social Distancing: The Co-evolution Of Bilingual Neologisms And Public Health Campaigns In Two Cities In The Time Of COVID-19. *Sage Open*. 11.3. HTTPS://DOI.ORG/10.1177/21582440211031556.
-

LEI ET AL. 2021 EMERGENT NEOLOGISMS

- What Is Emergent Neologism: No Previous Forms
  - Not Replacement: S-curve/Snowball Effect
- But What Model?
  - What Are The Types Of Emergent Neologisms
  - What Are The Word Formation Strategies
  - And Their Life Cycle
- Relation Between Neologism And Instigating Events

Categories	Terms
Under-specification	疫情*yiqing* 'situation of pandemic'
	肺炎*feiyan* 'pneumonia'
	病毒*bingdu* 'virus'
Pre-official names	不明原因肺炎*buming yuanyin feiyan* 'pneumonia of unknown sources'
	病毒性肺炎*bingduxing feiyan* 'viral pneumonia'
	新型病毒*xinxing bingdu* 'novel type virus'
	新型肺炎*xinxing feiyan* 'novel type pneumonia'
	冠状病毒*guanzhuang bingdu* 'corona virus'
Stigmatizing names	武汉肺炎*wuhan feiyan* 'Wuhan pneumonia'
	武汉病毒性肺炎*wuhan bingduxing feiyan* 'Wuhan viral pneumonia'
	中国病毒*zhongguo bingdu* 'Chinese virus'
	武汉新型肺炎*wuhan xinxing feiyan* 'Wuhan novel type pneumonia'
	武汉病毒*wuhan bingdu* 'Wuhan virus'

Categories	Terms
Official names	新型冠状病毒肺炎*xinxing guanzhuang feiyan* 'novel type crown-shape virus pneumonia'
	新冠肺炎*xin guan feiyan* 'novel corona pneumonia'
	新冠疫情*xin guan yiqing* 'novel corona pandemic'
	新冠病毒*xin guan bingdu* 'novel corona virus'
	新冠*xin guan* 'novel corona'
	新型冠状病毒*xinxing guanzhuang bingdu* 'novel-type corona-shape virus'
	2019新型冠状病毒*xinxing guanzhuang bingdu* '2019 novel-type crown-shape virus'
English abbreviations	COVID-19
	2019-nCov
	Coronavirus
	SARS-CoV-2

FIGURE 1 PERCENTAGES OF COVID-19 NEOLOGISMS CATEGORIES ON BAIDU INDEX

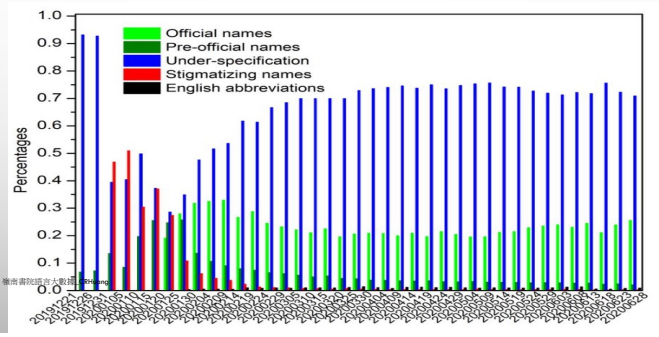


FIGURE 2A DEVELOPMENT OF NATIONAL NEW CONFIRMED AND SUSPECTED CASES

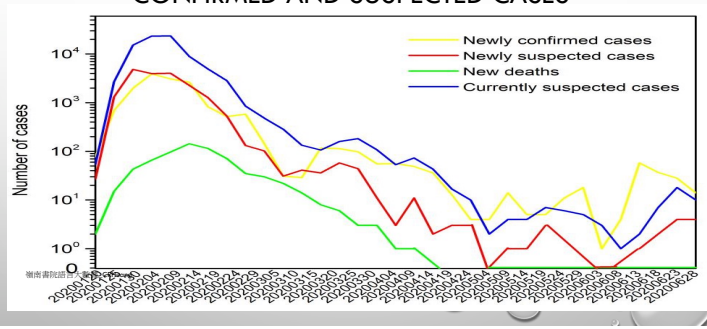
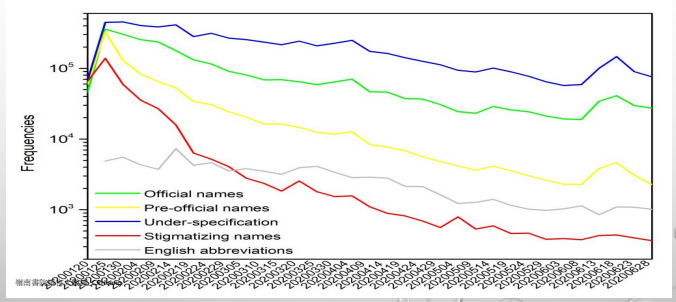


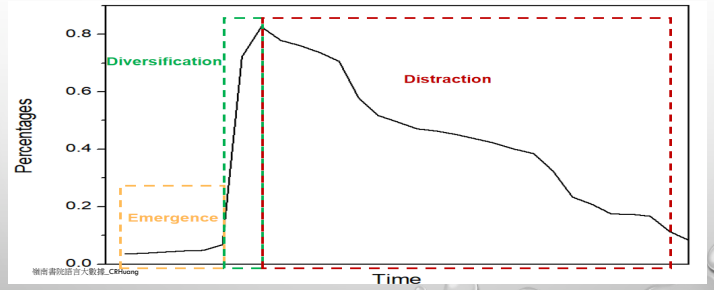
FIGURE 2B DEVELOPMENT IN FREQUENCIES OF COVID-19 WORD CATEGORIES



PREDICTING COVID WITH NEOLOGISM USES

IV	DV	Expression	R^2
Official names	Newly suspected cases	Binomial ($y = ax + bx^2 + \varepsilon$)	0.970
Stigmatizing names	Newly confirmed cases	Binomial ($y = ax + bx^2 + \varepsilon$)	0.964
Official names	Newly suspected cases	Linear ($y = ax + b + \varepsilon$)	0.924
Stigmatizing names	Currently suspected cases	Binomial ($y = ax + bx^2 + \varepsilon$)	0.908
Under-specifications	Newly suspected cases	Binomial ($y = ax + bx^2 + \varepsilon$)	0.903
Under-specifications	Currently suspected cases	Logistic ($y = 1/(1 + e^x)$)	0.903

MODELLING EMERGENT NEOLOGISMS



QUICK SUMMARY OF PUBLIC HEALTH CAMPAIGNS IN TWO CITIES AND IN TWO LANGUAGES

- Wang, Xiaowen, & Chu-ren Huang. 2021. From Contact Prevention To Social Distancing: The Co-evolution Of Bilingual Neologisms And Public Health Campaigns In Two Cities In The Time Of COVID-19. *Sage Open*. 11.3.

HTTPS://DOI.ORG/10.1177/21582440211031556

FROM BIG DATA TO LINKED DATA

- LINKED DATA REFERRING TO THE LINKING OF TWO OR MORE DATASETS FOR KNOWLEDGE INTEGRATION AND DISCOVERY
  - NOTE THAT DATA BEING LINKED DO NOT HAVE TO BE OF THE SAME TYPE. WHEN LINKING HETEROGENOUS DATA, THE MEDIATION OF ONTOLOGY IS CRUCIAL
- NOTE THAT SCIENTIFIC DISCOVERY RARELY (IF EVER) INVOLVES SOMETHING UNHEARD OF,…
- SCIENTIFIC DISCOVERIES OFTEN INVOLVE PUTTING TWO PIECES OF PUZZLE TOGETHER IN A NEW WAY…THE EUREKA! MOMENT

BEYOND EPIDEMICS

WHAT DOES LANGUAGE BIG DATA TELL US ABOUT OUR ENVIRONMENT AND OUT SOCIETY

- Su, Qi, Pengyuan Liu, Wei Wei, Shucheng Zhu, & Chu-ren Huang. 2021. Occupational Gender Segregation And Gendered Language In A Language Without Gender: Trends, Variations, Implications For Social Development In China. *Humanities And Social Sciences Communications*. Nature. 8:133. Https://Doi.Org/10.1057/S41599-021-00799-6
- Historical Trends And Regional Variations Based On Markedness

LANGUAGE AND THE NATURAL ENVIRONMENT

- Huang, Chu-Ren, Sicong Dong, Yike Yang, & Ren He. 2021. From Language To Meteorology: Kinesis In Weather Events And Weather Verbs Across Sinitic Languages. *Humanities And Social Sciences Communications*. Nature. 8:4. Https://Doi.Org/10.1057/S41599-020-00682-w
- Dong, Sicong, Yike Yang, Ren He, & Chu-ren Huang. 2021. Directionality Of Atmospheric Water In Chinese: A Lexical Semantic Study Based On Linguistic Ontology. *SAGE Open*. 12.1. Https://Doi.Org/10.1177/21582440209882 93
- Dong, Sicong, Chu-Ren Huang & Ren He. 2020. Towards a New Typology of Meteorological Events: A Study Based on Synchronic and Diachronic Data. *Lingua*: 247. #102894. https://doi.org/10.1016/j.lingua.2020.102894

Weather and Language

Q1. Why some types of weather waters rise and others fall?

Q2. Why different languages (dialects) encodes the directionalities of weather waters differently?

	BCC	CCL	Sinica
降露 *jiànglù*	7	2	0
下露 *xiàlù*	7	4	0
起露 *qǐlù*	1	0	0
上露 *shànglù*	0	0	0
降霜 *jiàngshuāng*	66	13	0
下霜 *xiàshuāng*	119	26	1
起霜 *qǐshuāng*	14	0	0
上霜 *shàngshuāng*	16	2	0
降霧 *jiàngwù*	11	8	0
下霧 *xiàwù*	174	10	0
起霧 *qǐwù*	352	22	4
上霧 *shàngwù*	5	0	0

BCC/CCL: China mainland, Sinica: Taiwan

霧 *wù* 'fog'	Taiwan	突起 'to rise suddenly', 出現 'to appear', 形成 'to form', 飄 'to drift', 起 'to rise'
	Mainland China	出現 'to appear', 形成 'to form', 降 'to fall', 突起 'to rise suddenly', 起 'to rise', 漸起 'to rise gradually', 驟起 'to rise abruptly'
	Singapore	形成 'to form', 發生 'to occur'
霜 *shuāng* 'frost'	Taiwan	結成 'to condense to', 結 'to condense', 凍成 'to freeze to', 出現 'to appear', 凝結成 'to condense to'
	Mainland China	凍成 'to freeze to', 出現 'to appear', 結 'to condense'

	Down	Up	Both	None
Snow	100	0	0	0
Rain	98.3	0	0	1.7
Hail	97.0	0	0	3.0
Fog	51.7	24.7	13.5	10.1
Dew	50.0	15.2	4.5	30.3
Frost	45.5	0	2.0	52.5

Some Generalizations

In Mandarin Chinese, The Three Weather Types Exhibit Three Different Patterns In Terms Of Directionality.

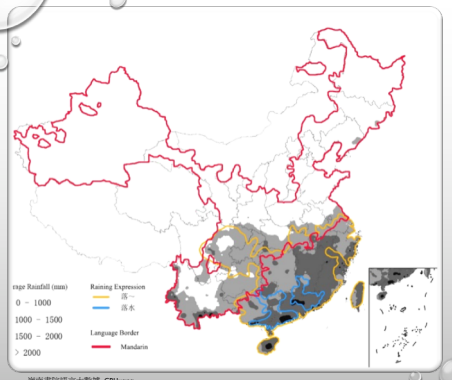
- Precipitation: Rain, Snow And Hail Are Consistently Expressed With Downward Movement. ↓
- Condensation: Dew And Frost Can Be Expressed Both As Moving Downwards And Upwards, But The Downward Meaning Has Bigger Proportions. ↓ > ↑
- Suspension: Fog Can Also Be Said To Move Downwards And Upwards, But More Cases Were Found To Indicate The Upward Meaning. ↓ < ↑

The Sinica Corpus, partly because of corpus size, has Clear-cut Tendencies: Fog Can Only Be 'Rising' And Frost Can Only Be 'Falling'..

	Down	Up	Both	None
Snow	100	0	0	0
Rain	98.3	0	0	1.7
Hail	97.0	0	0	3.0
Fog	51.7	24.7	13.5	10.1
Dew	50.0	15.2	4.5	30.3
Frost	45.5	0	2.0	52.5

4 DISCUSSION

In Sinitic languages, fog tend to be expressed as 'falling'. However, fog has the biggest proportions of 'up' and 'both' among all the six phenomena, which can serve as evidence that fog is most likely to be described as moving upwards by Chinese people, although it is more often said to move downwards. This makes fog standing out in these phenomena with seeming uncertainty about directions.



LINKING LINGUISTIC TYPOLOGY DATA WITH WEATHER MAPS

Map Of Distribution Of Raining Expressions, Yearly Average Rainfall And Mandarin Border.

(Huang et Al. 2021)

ADDITIONAL EVIDENCE FROM THE THREE PAPERS

- Directionality of verbs used with weather waters can be predicted by the mass of the typical mass and size of the kind of weather phenomenon…..
- The distribution of two special nouns for hail (冷leng 'cold', or 蛋 dan 'egg'; instead of the standard 雹 bao) roughly correspond to the regions with the most severe hail damage in China.
- The area of using highly transitive verbs (e.g. 打 da 'to hit') instead of the typical low transitivity weather verbs (e.g. 下 xia, or 降 jiang 'to fall') for frost 霜 also correspond roughly to the areas of known to suffer more severe frost damage in China
- Three documented usages of the term下凌 xialing 'rain freezing rain' happen to be at the location of the three cities with the highest probability for having freezing rain in China: Guizhou, Changsha, and Wuhan,.

*Sleet: frozen rain drops; Freezing rain: very cold rain that freezes upon landing

From Linking Facts to Proving of Hypothesis

1. Showing correlational patterns
   - Note that patterns do not equal causal relations (or explanation)
2. Show that such patterns cannot be accounted for and typically contradict the prediction of the best available theory
   - In this study, it is crucial to establish that these patterns of lexical choices cannot be predicted by isoglosses (i.e. do not match the typical patterns of language variations according to local dialect/language)

From Linking Facts to Proving of Hypothesis II

3. Corroborating with multiple sets of correlations (and show that they can be explained by the same causal relation)
   - One set of data may be coincidental; but multiple sets of data in different environments showing same pattern corroborate the hypothesis
   - Provide supporting data from rigorously controlled experiment to show that the proposed cognitive motivation is valid. In this paper, we used hypothetical novel weather events for speaker to choose different weather verb to verify that the choices of weather verbs depends on kinesis…..
     - Again, this experiment is corroborated by well established studies in typology that verbs encode different degrees of kinesis via transitivity and other linguistics devices …

Conclusion: How To Design Effective Research In Humanities And Social Sciences

And In Applied Language Sciences

- Beware of Galton's Problem
  - Can we show whether a correlation is functional or environmental
  - I.e. in cross-cultural studies, is it by the cause proposed or by borrowing or by natural evolution etc.
  - In experimental studies, how to exclude spurious correlations

Experimental Studies

- Robust Experimental Design
  - Sample Size And Power Analysis
  - Measures Of Statistical Significance
  - Good Theoretical Explanation
  - Error Analysis

But what can we do facing a complex and challenging issue that cannot be reduced to a single (or a few) experimental controllable variables?

- Such as many of the urgent issue we face and many of the important HSS research questions

For HSS and for Emerging Challenges

- Find interesting and challenging correlations and propose hypothesis
- Show better prediction than the gold standard: The currently accepted standard theory or the best available account
  - In Jiang et al. (2021) we showed that the SIR epidemic model makes better prediction than that commonly accepted memetic model for neologisms
  - In Huang et al. (2021) we showed the kinesis and cognitive experience based account correctly predict lexical choices that are not predictable by distribution of Sinitic languages [i.e. in cases of both different dialects sharing the same unusual form; or the same dialects selecting different forms]
- Corroborate the hypothesis with multiple sets of independent correlations
  - I.e. Showing that the same hypothesis can correctly predict different sets of facts
  - If possible, design experiments to support parts of the hypothesis
  - Corroborating by contextualizing: develop your hypothesis or corroborating studies based on well-accepted, sate-of-the-art theory or scientific facts.…

THE END
Linguistics and Language Sciences *In Vivo*

With Language Big Data

Facing The Challenging Complexity Of Our Contemporary Society

To design a research question that leads to proving of a clearly defined clausal relation or a meaning correlational pattern

To show improvement over 'gold standard' and/or to corroborate with data, theory, and experiments

Be ware of Galton's Problem

BONUS SLIDES

- NOT GIVEN IN THE LECTURE
- FOR REFERENCE ONLY

WHERE TO GET BIG DATA:
LDC: THE LINGUISTIC DATA CONSORTIUM

- HTTPS://WWW.LDC.UPENN.EDU/

LDC Linguistic Data Consortium

- LDC IS A CONSORTIUM OF MEMBER ORGANIZATIONS THAT POOL RESOURCES TO SUPPORT LANGUAGE-RELATED RESEARCH, EDUCATION AND TECHNOLOGY DEVELOPMENT. MEMBERS INCLUDE UNIVERSITIES, RESEARCH LABS, COMPANIES AND GOVERNMENT ORGANIZATIONS FROM AROUND THE GLOBE. LDC MEMBERSHIPS ARE BASED ON THE CALENDAR YEAR AND MEMBERS RECEIVE PERPETUAL RIGHTS TO DATA ACQUIRED DURING THEIR MEMBERSHIP YEARS.

CONSORTIUM: SHARABLE AND INTEROPERABLE DATA

Membership can be paid (covered by project, for instance)

Or received as contribution of databases

Contribute one language resources, enjoying unlimited access (by all members of your institution) in perpetual to all databases published in the same calendar year

*PolyU is a member for at least 5 years….

SHARABLE LANGUAGE RESOURCES UNDERPINNING BIG DATA I

- HUANG, CHU-REN. 2009. TAGGED CHINESE GIGAWORD VERSION 2.0. PHILADELPHIA: LEXICAL DATA CONSORTIUM. ISBN 1-58563-516-2 HTTPS://CATALOG.LDC.UPENN.EDU/LDC2009T14
- HUANG, CHU-REN. 2007. TAGGED CHINESE GIGAWORD LDC2007T03. PHILADELPHIA: LINGUISTIC DATA CONSORTIUM. ISBN 1-58563-409-3 HTTPS://CATALOG.LDC.UPENN.EDU/LDC2007T03

SHARABLE LANGUAGE RESOURCES UNDERPINNING BIG DATA II

- NEERGAARD, KARL DAVID, HONGZHI XU, AND CHU-REN HUANG. 2020. DATABASE OF WORD LEVEL STATISTICS - MANDARIN LDC2020L01. PHILADELPHIA: LINGUISTIC DATA CONSORTIUM. ISBN: 1-58563-914-1 HTTPS://CATALOG.LDC.UPENN.EDU/LDC2020L01
- WANG, SHICHANG, ET AL. 2020. SEMTRANSCNC LDC2020T12. PHILADELPHIA: LINGUISTIC DATA CONSORTIUM. ISBN: 1-58563-931-1

HTTPS://CATALOG.LDC.UPENN.EDU/LDC2020T12

SHARABLE LANGUAGE RESOURCES UNDERPINNING BIG DATA III

- SANTUS, ENRICO, HONGCHAO LIU, AND CHU-REN HUANG. 2020. EVALUTION LDC2020T06. PHILADELPHIA: LINGUISTIC DATA CONSORTIUM. ISBN: 1-58563-921-4

HTTPS://CATALOG.LDC.UPENN.EDU/LDC2020T06

*THIS DATASET CONTAINS RELATA FOR BOTH CHINESE AND ENGLISH (RELATA: PAIRS OF WORDS WITH THEIR LEXICAL SEMANTIC RELATIONS MARKED AND VERIFIED)

Trends

OTHER EMERGING TOPICS

AS ATTESTED BY PUBLICATIONS

Sensory Lexicon and Synaesthesia: A Linguistic Window to Sense and Cognition

- ZHONG, YIN, CHU-REN HUANG, AND SICONG DONG. BODILY SENSATION AND EMBODIMENT: A CORPUS-BASED STUDY OF GUSTATORY VOCABULARY IN MANDARIN CHINESE. TO APPEAR IN JOURNAL OF CHINESE LINGUISTICS.
- ZHONG, YIN, AND CHU-REN HUANG. SWEETNESS OR MOUTHFEEL: A CORPUS-BASED STUDY OF THE CONCEPTUALIZATION OF TASTE. TO APPEAR IN LINGUISTIC RESEARCH.
- CHEN, I-HSUAN, QINGQING ZHAO, YUNFEI LONG, QIN LU, AND CHU-REN HUANG. "MANDARIN CHINESE MODALITY EXCLUSIVITY NORMS." *PLOS ONE* 14, NO. 2 (2019): E0211336.
- ZHAO, QINGQING , **CHU-REN HUANG**, KATHLEEN AHRENS. 2019. DIRECTIONALITY OF LINGUISTIC SYNESTHESIA IN MANDARIN: A CORPUS-BASED STUDY. *LINGUA*. 232. #102744. HTTPS://DOI.ORG/10.1016/J.LINGUA.2019.102744 ***LINGUA EDITOR'S CHOICE***
- JO, CHARMHUN. 2019. A CORPUS-BASED ANALYSIS OF SYNESTHETIC METAPHORS IN KOREAN. *LINGSUISTIC RESEARCH, 36*(3), 459-483.
- 赵青青ZHAO, QINGQING, 董居仁CHU-REN HUANG, 熊佳娟 JIAJUAN XIONG. 2019.《通感、隐喻与认知-通感现象在汉语中的系统性表现与语言学价值》LINGUISTIC SYNAESTHESIA, METAPHOR AND COGNITION: THE SYSTEMATICITY AND SIGNIFICANCE OF LINGUISTIC SYNAESTHESIA IN CHINESE.《中国语文》ZHONGGUOYUWEN. 2019.2.
- ZHAO, QINGQING, CHU-REN HUANG, YUNFEI LONG. 2018. SYNAESTHESIA IN CHINESE: A CORPUS-BASED STUDY OF GUSTATORY ADJECTIVES IN MANDARIN. *LINGUISTICS*. HTTPS://DOI.ORG/10.1515/LING-2018-0019

Complex Adaptive Systems: Language Systems

- LIESENFELD, ANDREAS, MEICHUN LIU, AND CHU-REN HUANG. PROFILING THE CHINESE CAUSATIVE CONSTRUCTION WITH RANG (讓), SHI (使) AND LING (令) USING FRAME SEMANTIC FEATURES. TO APPEAR IN CORPUS LINGUISTICS AND LINGUISTIC THEORY.
- XU, HONGZHI, MENGHAN JIANG, JINGXIA LIN, AND **CHU-REN HUANG**. 2020. LIGHT VERB VARIATIONS AND VARIETIES OF MANDARIN CHINESE: COMPARABLE CORPUS DRIVEN APPROACHES TO GRAMMATICAL VARIATIONS. *CORPUS LINGUISTICS AND LINGUISTIC THEORY*. AHEAD OF PRINT. HTTPS://DOI.ORG/10.1515/CLLT-2019-0049
-
-
-
41598-019-52433-W.
ANG, 2019. CONSTRUCTING THE MANDARIN GICAL NETWORK. COMPLEXITY (SPECIAL ISSUE), COGNITIVE NETWORK SCIENCE: A NEW FRONTIER.
- ENGYU FANG, AND **CHU-REN HUANG**. 2019. THE ATIVENESS OF INTERNAL SYNTACTIC REPRESENTATIONS IN TOMATIC GENRE CLASSIFICATION. *JOURNAL OF AL OF QUANTITATIVE LINGUISTICS*. 1-34.

COMPLEX ADAPTIVE SYSTEMS: MA LAW

- HOU, RENKUI, AND **CHU-REN HUANG**. 2020. ROBUST STYLOMETRIC ANALYSIS AND AUTHOR ATTRIBUTION BASED ON TONES AND RIMES. *JOURNAL OF NATURAL LANGUAGE ENGINEERING*. 26.1.49-71. HTTPS://DOI.ORG/10.1017/S135132491900010X
- HOU, RENKUI, AND **CHU-REN HUANG**. 2020. CLASSIFICATION OF REGIONAL AND GENRE VARIETIES OF CHINESE: A CORRESPONDENCE ANALYSIS APPROACH BASED ON COMPARABLE BALANCED CORPORA. *JOURNAL OF NATURAL LANGUAGE ENGINEERING*. HTTPS://DOI.ORG/10.1017/S1351324920000121.
- HOU, RENKUI, **CHU-REN HUANG**, KATHLEEN AHRENS, YAT-MEI SOPHIA LEE. 2020. LINGUISTIC CHARACTERISTICS OF CHINESE REGISTER BASED ON THE MENZERATH—ALTMANN LAW AND TEXT CLUSTERING, DIGITAL SCHOLARSHIP IN THE HUMANITIES. 35.1:54–66. HTTPS://DOI.ORG/10.1093/LLC/FQZ005
- HOU, RENKUI, CHU-REN HUANG, AND HONGCHAO LIU. (2018). A STUDY ON CHINESE REGISTER CHARACTERISTICS BASED ON REGRESSION ANALYSIS AND TEXT CLUSTERING. CORPUS LINGUISTICS AND LINGUISTIC THEORY. PUBLISHED ONLINE: 30 MARCH 2017. DOI: HTTPS://DOI.ORG/10.1515/CLLT-2016-0062.
- HOU, RENKUI, CHU-REN HUANG, HUE SAN DO, AND HONGCHAO LIU. (2017). A STUDY ON CORRELATION BETWEEN CHINESE SENTENCE AND CONSTITUTING CLAUSES BASED ON THE MENZERATH-ALTMANN LAW. JOURNAL OF QUANTITATIVE LINGUISTICS: 24(4).350-366. PUBLISHED ONLINE: 26 APR 2017. DOI: HTTP://DX.DOI.ORG/10.1080/09296174.2017.1314411

LANGUAGE TECHNOLOGY AND ENGINEERING

- LIU, HONGCHAO, EMMANUELE CHERSONI, NATALIA KLUYEVA, ENRICO SANTUS, **CHU-REN HUANG**. 2019. SEMANTIC RELATA FOR THE EVALUATION OF DISTRIBUTIONAL MODELS IN MANDARIN CHINESE. *IEEE ACCESS 7:145705-145713.*
-
HE, **CHU-REN HUANG.** 2019. A STRUCTURED DISTRIBUTIONAL MODEL OF SENTENCE MEANING AND COMPLEXITY. *JOURNAL OF URAL LANGUAGE ENGINEERING, 25* (4).483-502.
- *INEERING, 25* (4).483-502.
REN HUANG, AND MINGLEI LI. 2019. IMPROVING ATTENTION MODEL BASED ON COGNITION GROUNDED DATA FOR SENTIMENT
- IMENT ANALYSIS. JOURNAL OF IEEE TRANSACTIONS ON AFFECTIVE COMPUTING
AND CHU-REN HUANG. 2019. METAPHOR DETECTION: LEVERAGING CULTURALLY GROUNDED EVENTIVE INFORMATION. IEEE ACCESS.
- RMATION. IEEE ACCESS. HTTPS://IEEEXPLORE.IEEE.ORG/DOCUMENT/8410071
, BI, C., & LI, M. (2018). LEARNING HETEROGENEOUS NETWORK EMBEDDING FROM TEXT AND LINKS. *IEEE ACCESS*, *6*, 55850-55860.
- NETWORK EMBEDDING FROM TEXT AND LINKS. *IEEE ACCESS*, *6*, 55850-55860.
N LU, **CHU-REN HUANG**, ELVIRA PEREZ VALLEJOS, YUNFEI LONG. 2020. DUAL MEMORY NETWORK MODEL FOR SENTIMENT ANALYSIS

CORE "REFERENCES" FOR ANY CHINESE LANGUAGE RELATED RESEARCH

- HUANG, CHU-REN, SHU-KAI HSIEH, AND PENG JIN. 2019 (IN PREPARATION). CHINESE LANGUAGE RESOURCES: *DATA COLLECTION, LINGUISTIC ANALYSIS, ANNOTATION, AND LANGUAGE PROCESSING*. SPRINGER.
- LU, QIN, NIANWEN XUE, AND CHU-REN HUANG. (IN PREPARATION) *COMPUTER PROCESSING OF CHINESE. STUDIES IN NATURAL LANGUAGE PROCESSING* BOOK SERIES. CAMBRIDGE UNIVERSITY PRESS.
- HUANG, CHU-REN, YEN-HWEI LIN, AND I-HSUAN CHEN. IN PREPARATION. (EDS.) CAMBRIDGE HANDBOOK OF CHINESE LINGUISTICS. CAMBRIDGE: CAMBRIDGE UNIVERSITY PRESS.
- HUANG, CHU-REN, BARBARA MEISTERERNST, AND ZHUO JING-SCHMIDT. 2019. *ROUTLEDGE HANDBOOK ON CHINESE APPLIED LINGUISTICS*. LONDON: ROUTLEDGE.
- HUANG, CHU-REN, SHU-KAI HSIEH, AND KEH-JIANN CHEN. (2017). *MANDARIN CHINESE WORDS AND PARTS OF SPEECH: A CORPUS-BASED STUDY*. LONDON: ROUTLEDGE
- HUANG, CHU-REN, AND DINGXU SHI. (EDS.). (2016). *A REFERENCE GRAMMAR OF CHINESE*. CAMBRIDGE: CAMBRIDGE UNIVERSITY PRESS.
- WANG, WILLIAM S. Y. AND CHAOFEN SUN. (2015). OXFORD HANDBOOK IN CHINESE LINGUISTICS. OXFORD: OXFORD UNIVERSITY PRESS.

SHARABLE LANGUAGE RESOURCES UNDERPINNING BIG DATA II

ONLINE RESOURCES

- **SINICA CORPUS**: ACADEMIA SINICA BALANCED CORPUS FOR MANDARIN CHINESE中央研究院現代漢語平衡語料庫. NOVEMBER 1996 (FIRST WEB VERSION). HTTP://ASBC.IIS.SINICA.EDU.TW/
- **SINICA BOW:** ACADEMIA SINICA BILINGUAL ONTOLOGICAL WORDNET中央研究院中英雙語知識本體詞網. OCTOBER 2003. HTTP://BOW.LING.SINICA.EDU.TW
- SINICA TREEBANK中文句結構樹資料庫. APRIL 2004. HTTP://TREEBANK.SINICA.EDU.TW/
- CHINESE WORDNET (PROTOTYPE) 2005. HTTP://CWN.LING.SINICA.EDU.TW
- HANTOLOGY 漢字知識本體. 2006. HTTP://HANTOLOGY.LING.SINICA.EDU.TW