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

黃居仁

香港理工大學

第二屆中國語言學嶺南書院（2021 年）

2021 年 12 月 8 日 14:30

聽其言,觀其行,人焉廋哉

*The lecture will be delivered in Mandarin, with the main references and PPT content (mostly) in English. 報告以漢語進行，佐以英文投影片與參考資料。

Most References are Open Access. Please click on the link to access the full paper (unless marked otherwise) 除非特別說明，引用資料為開源期刊，請直接點入閱讀。

Huang, C-R., S. Dong, Y. Yang, & H. Ren. 2021. From Language to meteorology: Kinesis in weather events and weather verbs across Sinitic languages. *Humanities and Social Sciences Communications*. 8:4. https://doi.org/10.1057/s41599-020-00682-w

Jiang, M., X. Shen, K. Ahrens, & C-R. 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

Lei, S., R. Yang, & C-R. 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://doi.org/10.1016/j.lingua.2021.103095 Editor's Choice: Open Access till end of December 2021. 主編特選論文，2021/12 前可免費開放閱讀

Su, Q., P. Liu, W. Wei, S. Zhu, & C-R. 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*. 8:133. https://doi.org/10.1057/s41599-021-00799-6

Wang, X., & C-R. 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

To Appear 待刊

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

Other References 其他參考文獻

Huang, C-R., B. Meisterernst, & Z. Jing-Schmidt. (Eds.). 2019. *Routledge Handbook of Chinese Applied Linguistics*. London: Routledge.

Huang, C. R. (2009). Tagged Chinese gigaword version 2.0, ldc2009t14. *Linguistic Data Consortium*. https://doi.org/10.35111/9bhh-2s82

綱要

十億字級的語料庫(gigaword corpus, 如 Huang 2009)是否能算大數據?二來大數據之大，是否一定需用統計式的數據分析法(data analytics),而非手動分析？目前最常用的語言大數據包括谷歌圖書(Google Books)：以幾個重要語言中過去數百年所有人類出版品為範圍，涵蓋了絕大多數無版權限制的出版品。另外 Google Trend,或 Baidu Index 等則是這兩個搜索引擎有紀錄以來，所有搜尋詞語的逐日統計。這些語言大數據資料的特色，除了規模之外，還有涵蓋時間深度與使用區域信息的特色。谷歌圖書是以每年為單位。而 Google Trend, Baidu Index 或大數據級的新聞資料庫，則可以細到每天的統計與變化。基本上，我們可以把這些語言大數據看成是人類群體行為留下最直接，最難作假的證據。語言是人們最主要的共同工具，而且只有語言文字能夠突破時空的限制，聯繫眾人。任何與人相關的重大事件，都會大量的語言證據；而網路語言大數據更有幾乎在時間發生時同時留下紀錄的即時特性。以新冠病毒或類似的流行病為例，醫學上的證明需要等到流行到了一定程度才可能進行採樣，取得病株並隔離後分析實驗等。總需要起碼一個月以上的時間。但是，網路搜尋與使用語言的趨勢是即時的，早在剛剛流行時就有明顯的證據了。比較這兩樣證據，起碼有一到兩個月的時間差。語言大數據另一個特色，就是能夠凸顯整個

社會共同的態度與觀點。以 WHO 與全球有志之士全力防堵的‘二次傷害語言’(stigmatizing language)作為例子。AIDS 早先翻譯為‘愛滋病’，反應的就是恐同的心理。‘西班牙流感’，‘香港腳’，‘自閉症’，‘老人痴呆症’這些目前國際主流媒體與醫學文獻避免使用的二次傷害名詞，表現的是怪罪得病者與‘死道友，不死貧道’ 的心態。雷司宇等(Lei et al. 2021)以新冠肺炎的各種不同新造名稱在疫情最初六個月網路使用的頻率改變與競爭關係，來建立湧現新詞的發展規律，以及不同新造詞的競爭起伏與疫情發展的相關性。另一個例子是而中文裡常用‘女醫師’，‘女工‘卻少用‘男醫師’或‘男工’，表現的就是社會中特定職業的性別區隔(occupational gender segregation) 現象。語言大數據，為各種敏銳語言觀察，提供了社會與人類集體行為改變的具體科學證據，以及用來發掘與預測重要認知與行為改變的原始資料。蘇祺等(Su et al. 2021) 利用過去六十年的報紙數據，歸納出中國職場性別區隔的歷史趨勢與區域差別。而黃居仁等(Huang et al. 2021)則從氣象用語的類型學差異，如動詞的及物性及動能高低及以動詞或名詞來表達特定氣候現象等，與氣象圖套疊，建立漢語方言氣象詞分佈，與氣象型態分佈的相關性。

‘聽其言,觀其行,人焉廋哉’。語言大數據的研究，凸顯了語言是反映人類認知觀察，情感，動機，立場等的即時證據。必須能充分掌握個別語言，才能正確判讀這些深層意義。