A CORPUS OF CANTONESE CONNECTED SPEECH ON A SHOESTRING

ANNUAL RESEARCH FORUM 2016

REGINE LAI AND GRÉGOIRE WINTERSTEIN

THE EDUCATION UNIVERSITY OF HONG KONG

OUTLINE

- Introduction: a corpus of Cantonese connected speech
- Design of the corpus
  - The **Map Task**
  - Contents, setup
- Automatic transcription
  - How?
  - Issues
  - Future work

WHY YET ANOTHER CANTONESE CORPUS?

- There already are several Cantonese corpora available
- However:
  - Size remains limited (insufficient for data intensive applications)
  - Actual availability of the data is variable
  - Not all corpora encode the same information
- … and we want to test whether a rich corpus can be made on a budget

SOME EXISTING CANTONESE CORPORA

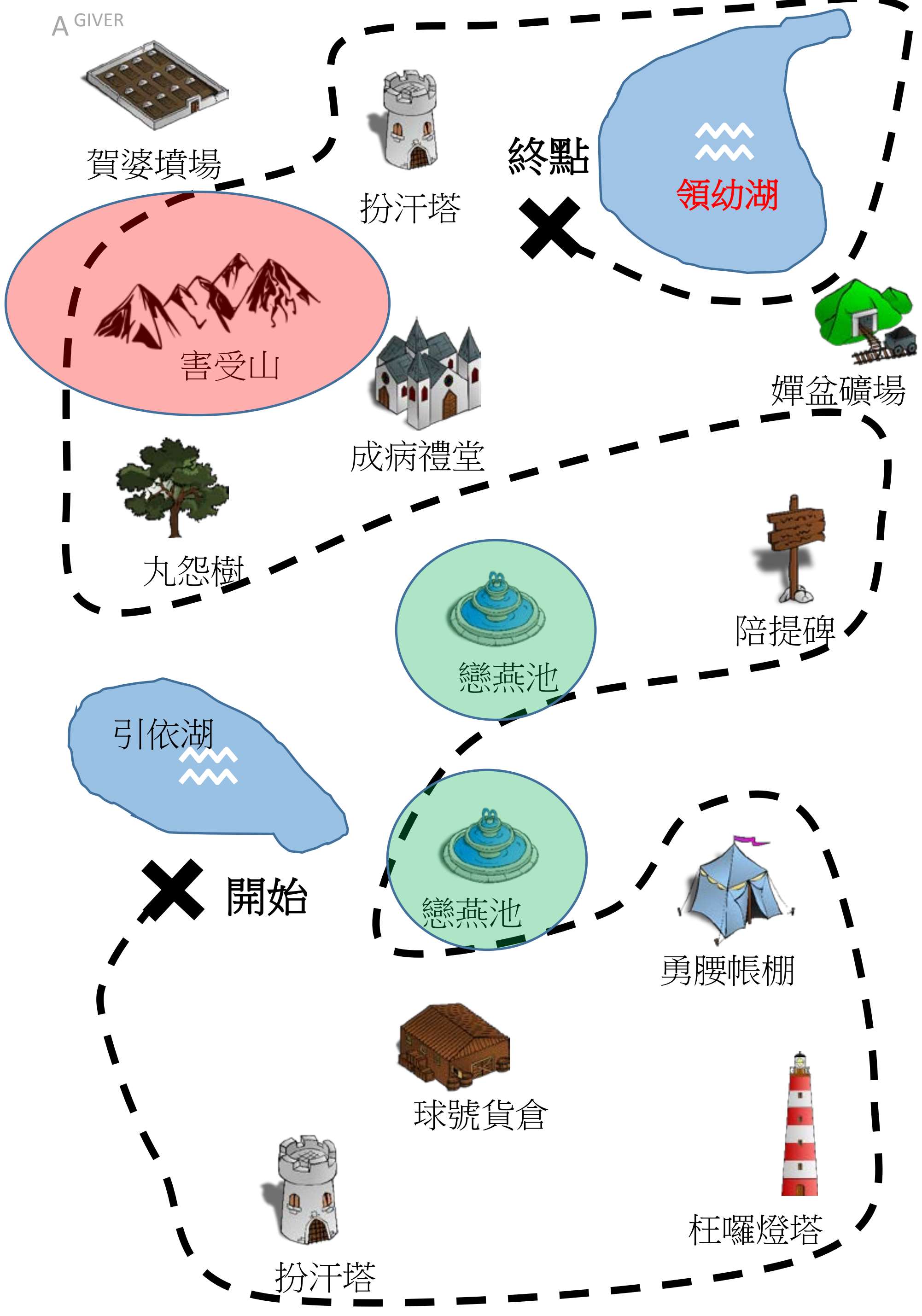
Corpus	*//*	*[]*	*字*	*Seg.*	*PoS*	*Size*	*Authentic*	*Open*	*Audio available*	*Notes*
HKCAC (Leung & Law, 2002)	✗	✓	✓	✗	✗	170,000 char.	✓	✓/≈	?	
HKCanCor (Luke & Wong, 2015)	✓	✗	✓	✓	✓	150,000 words	✓	✓	✗	
HK Mid-20th Cant. corpus (Chin, 2015)	✗	✗	✓	✓	✗	140,000 words	≈	≈	≈	Web queries only
HK Cant. Child Language Corp. (Lee et al. 1996)	✓	✗	✓	✓	✓	1,000,000 char.	✓/≈	✓	✓	Acq. Data; EN trans.
PolyU Corpus of Spoken Chinese	✗	✗	✓	✗	✗	?	≈	✓	≈	
Parallel Treebank of Cantonese and Mandarin (Lee et al., in prog.)	✓	✗	✓	✓	✓	75,000 char.	≈	?	✗	Dependency annot.
Our target	✓	✓	✓	✓	✓	200,000 words	✓	✓	✓	Currently ≈140,000char

WHAT TO RECORD?

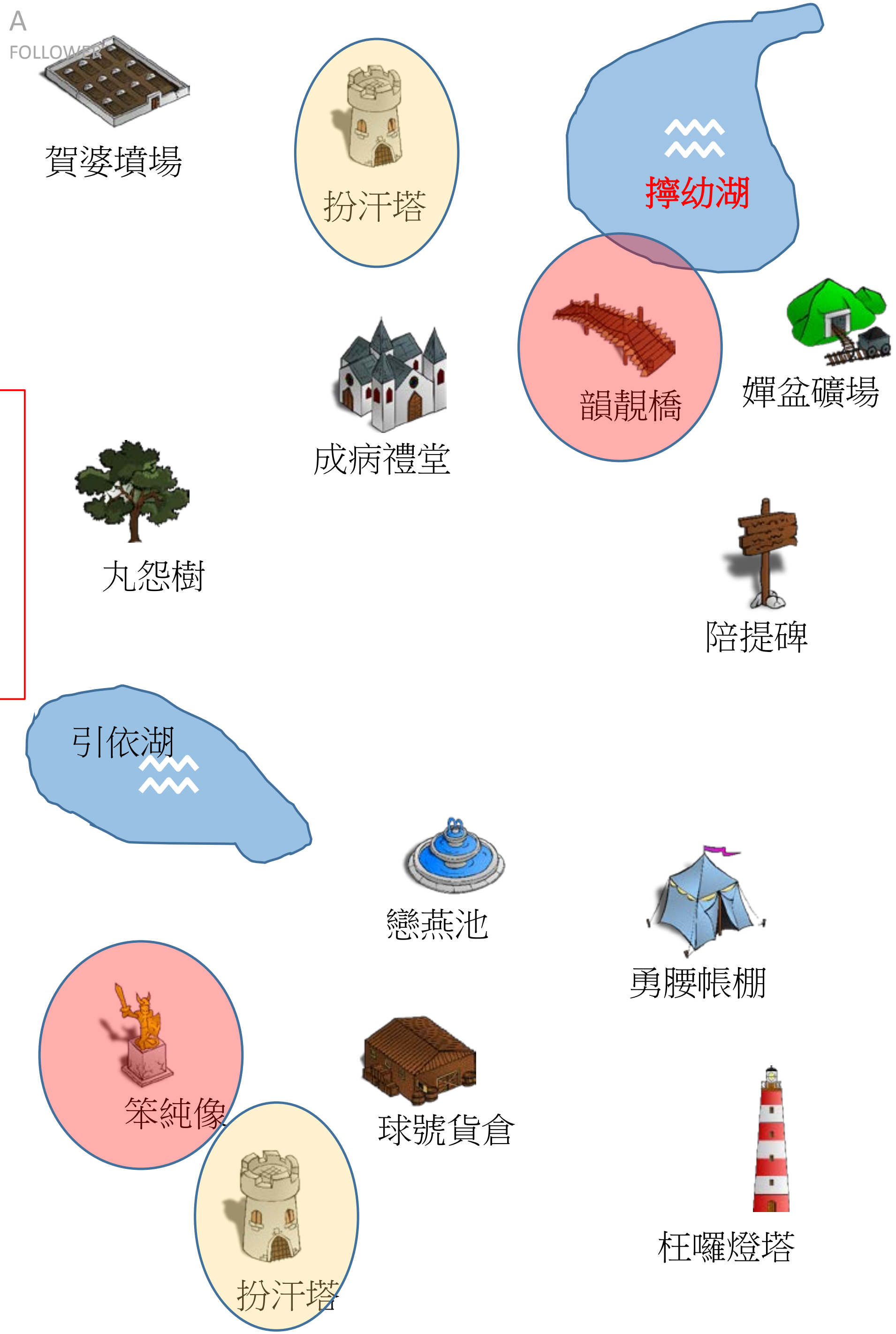
- What we want:
  - Authentic conversation, connected speech
  - Control elements of the conversation, e.g. elicit target words
  - Non-scripted, non-prepared discourse
  - Contemporary Cantonese
  - A “distracting” task
- Solution: do a **Map Task**

THE MAP TASK

- Based on a design by Brown et al. (1983), our corpus is inspired by Anderson et al. (1991) HCRC Map Task Corpus
- All MapTask dialogues have a similar goal which is known to the observer independently of what can be gleaned from participants' utterances: **reproducing a route of known form and controlled complexity on a map with comparable numbers of landmarks**.
- The goal can be achieved only by means of what the participants say to one another
- The **outcome is measurable-** the correct solution to the cooperative problem is well defined, successful communication can be measured in terms of the extent to which the achieved route corresponds to its model.
- Because **mismatches between landmarks**, their **names**, or their **locations** on a pair of maps are easy to arrange, the experimenter is in control of information initially shared by participants and can alter the difficulty of the task.



1. **Trick landmark**: minimal pair
2. **Repeated landmarks**: close/far away from each other
3. **Missing landmarks**

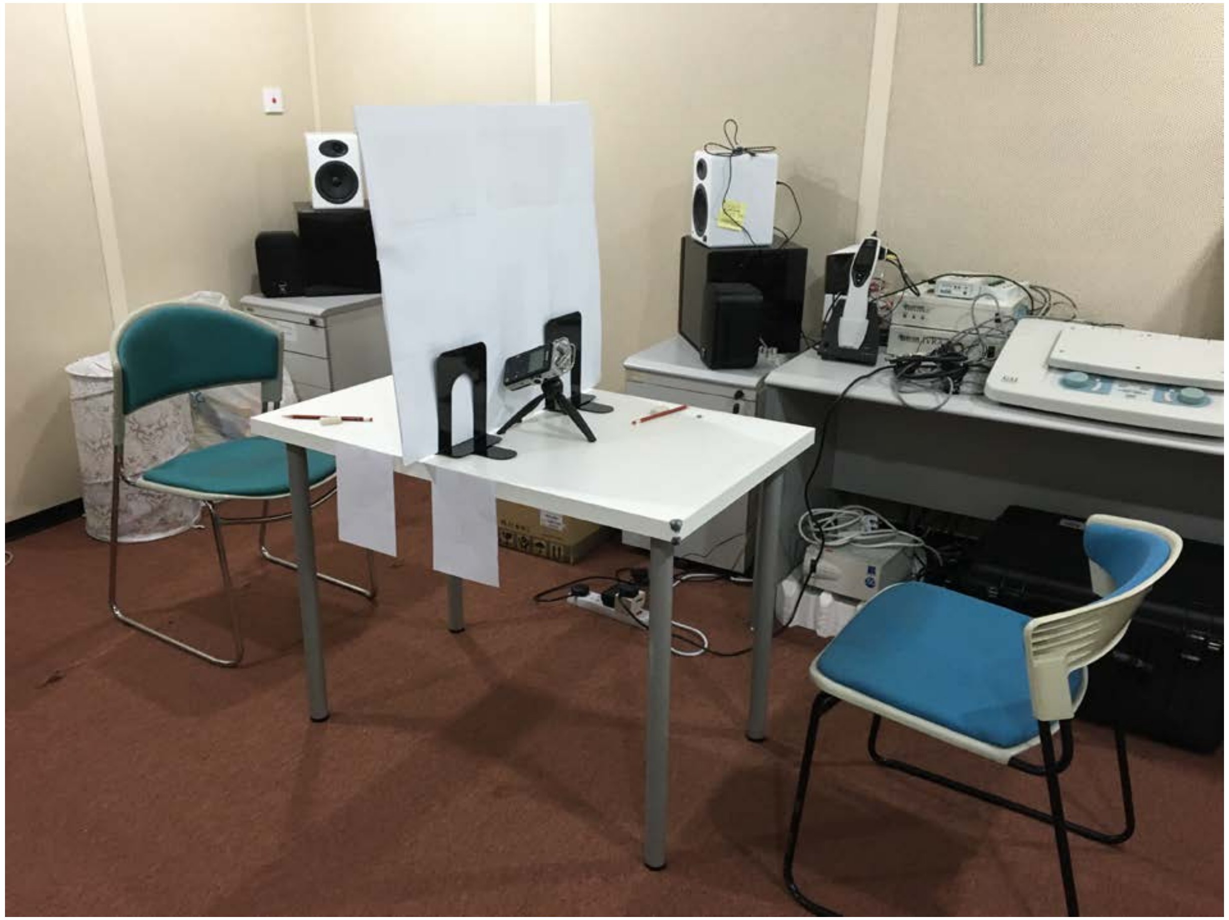


DESIGN OF THE MAPS

- **Maps**: 4 sets of maps, with a follower and a giver map for each set
- **Stimuli**: Each set consisted of 8 unique target stimuli and 8 unique fillers
- The landmarks on the maps were represented both graphically and orthographically in Chinese
- The images were downloaded from an open source image archive, and the label of each landmark was located directly below the image.
- The route of the map was controlled for its complexity across the four maps as each of them had 15 90° turns.
- The maps were printed in black and white.

SETUP

- Size: 40 participants (20 pairs). The duration of their recordings ranges from 18 min – 110 min. Total time recorded: 748.33min
- Each pair of participants completed all 4 maps. each participant took turns to be the Giver.
- They were given instructions that the goal of the task was to draw the route of Giver's map on the follower's map through verbal collaboration.
- The 2 participants were seated across from each other with approximately 1.5m apart in a soundproof booth.
- A cardboard was placed between the two participants to prevent any communication by eye contact and gestures.
- Each participant was recorded with a Sony PCM-D100 recorder
- Audio example



PARTNERSHIP BETWEEN PARTICIPANTS

- The duration of the task varies, possibly due to the friendship status of the participants who were paired up
- 5 out of 20 pairs were friends
- Friendship status tends to shorten the task
  - Mean duration = 37 min 27 sec
  - 4 out of 5 friend pairs' duration is below mean

DIFFICULTIES: CHARACTERS → PRODUCTION

- Words that are specific to Cantonese are difficult to elicit
- Participants are reluctant/unable to pronounce the Cantonese pronunciations of such words
- 𤓓 lo3, 囉 lo1, 𡃁 leng1: 60%-70% error
  - These words are likely to be pronounced as their visually similar counterparts, i.e. 攞lo2, 羅lo4 and 靚leng3
- 擰 ling2: 25-35% error
  - Similar to the above error pattern above, the radical seemed to be disregarded by the participants, and the most common mispronunciation is 寧ling4
- 燕 jin3, 冤 jyun1, 怨 jyun3: 5-20% error
  - The non-target pronunciation for these words are more surprising: 燕jin1 (very uncommon pronunciation of the word);冤 jyun1 was sometimes pronounced as 怨 jyun3 and vice versa.
- Possible reasons for mispronunciations: (1) Formality of the recording session discourages Cantonese pronunciations; (2) Font size might be too small

AUTOMATIC TRANSCRIPTION

- Manual transcription is long, hard and costly
- There are plenty of available tools of voice recognition, some of them free of charge
- These tools may not be perfect, but might speed up the transcription process
- We tested Google API, which offers an off the shelf solution

GOOGLE CLOUD SPEECH API

- https://cloud.google.com/speech/
- Intended usage: a voice recognition solution for mobile apps
  - Voice transcription in Chinese characters
  - Adapted to short utterances (e.g. voice commands) or voice to text typing usage.
- With some minor tweaking, it can be used on voice recordings:
  - Python scripts already exist (SpeechRecognition https://pypi.python.org/pypi/SpeechRecognition/)
  - For data intensive usage, Google charges $0.006 per minute after the first 60 minutes (per month)
  - Google Cloud offers 300$ for the first 60 days of usage
  - This allows us to automatically transcribe more than 800 hours of speech for free

EXAMPLE

Google API	Manual Transcription
好過起點係喺人醫護嘅下面	好個起點係喺引依湖嘅下面
跟住呢就向下行	跟住呢就向下行
行去辦看他	行去扮汗塔
跟住一路向右行去	跟住呢一路向右行去
∅來燈塔	枉囉燈塔
跟住呢再兜個個燈卡啦向上行	跟住呢再兜過個燈塔啦向上行
跟住你行到去	跟住呢行到去

ISSUES WITH THE AUTOMATIC TRANSCRIPTION

- **Nonce** words
- **Homophonous** words
- **Discourse particles**
  - 係嘞 → 系列
- **Gap-fillers/interjections** are ignored
- Problems related to the **diarization** of speakers

NONCE WORDS

- One of the goals of this project is to collect natural production **specific phonological targets** in connected speech, hence, a list of **nonce words** were included as the landmarks to facilitate elicitation.
- As predicted, these words are problematic for the automatic transcription.
- Google API was trained on natural authentic data, and will infer the most probable word if it has to transcribe a word that it never encountered before.

- Some examples:
  - 引依湖 → 人醫護
  - 扮汗塔 → 辦看他
  - 枉囉燈塔 → ∅來燈塔
  - 勇腰帳棚 → 重要將牌
  - 戀燕池 → 暖現時 / 軟件事
  - 害受山 → 害羞山
  - 賀婆墳場 → 婆婆墳場
  - 擰夏店 → 令下店
  - 擁腰沙漠 → 重要沙漠
  - 共間馬戲團 → 敢行馬戲團
  - 誰幣灘 → 稅費餐
  - 領幼湖 → 名又糊

HOMOPHONOUS WORDS AND UNEXPECTED ERRORS

- (Near) homophonous words:
  - 好個起點係喺引依湖嘅下面 → 好過起點係喺人醫護嘅下面
  - 係嘞跟住呢就再 → 係啦跟住呢就在
  - 兜過 → 都講 / 讀過 / 透過
- Unexpected errors
  - 右面行 → 又問嚇
  - 無共閒馬戲團唔緊要 → 冇咁係咪喺屯門你若

SENTENCE-FINAL PARTICLES AND GAP-FILLERS/INTERJECTIONS

- 跟住呢行到去 → 跟住你行到去
- 你兜過佢啦 → 喱透過佢∅
- 咁,呃,就唔使兜過去架嘞,就喺佢下面行過啦 → 咁∅ 就唔使多過去 ∅ ∅ 但係佢下面行過啦
- 戀燕池附近有啲咩㗎 → 邊段時附近有啲咩嘅
- 係嘞 → 系列

SPEAKER DIARIZATION

- The system has problems with floor change: when the speaker changes, it sometimes does not transcribe anymore
  - This might be related to a problem of volume
- Solution: use a **speaker diarization** system before, i.e. a system that indicates "Who spoke when"
- The setting is ideal for such applications: the number of speakers is known and small, and each speaker has its dedicated microphone (Anguera et al., 2012).
- Besides improving the transcription, using it will also facilitate the further encoding of the conversations.

SUMMARY, OUTLOOK

- Existing tools offer imperfect results, but a sound basis to speed up the transcription task.
- More and more readily usable tools are available to ease up the transcription process.
- Future work:
  - Speech diarization to improve the results of the automatic transcription
  - Train our own speech recognition systems rather than Google API
    - With specific training for our target words (e.g. CMU Sphinx)
    - To test automatic narrow transcription (in IPA)
  - Add additional layers of annotation: word segmentation, PoS
    - → Also rely on tools for an automatic first pass
  - Release the corpus under the CC-BY-SA 4.0 International license for the community

REFERENCES

- A. Anderson, Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S. and Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, 34, pp. 351-366.
- X. Anguera et al. (2012) Speaker Diarization: A Review of Recent Research. IEEE Transactions on Audio, Speech, and Language Processing archive, 20(2), pp. 356-370 .
- Brown, C., Anderson, A., Yule, G., and Shillcock, R. (1983). *Teaching Talk*. Cambridge, U.K.: Cambridge University Press.
- A. Chin (2015) *A Linguistics Corpus of Mid-20th Century Hong Kong Cantonese*, Department of Linguistics and Modern Language Studies, The Education University of Hong Kong.
- M.T. Leung and S.P. Law (2001) HKCAC: The Hong Kong Cantonese Adult Language Corpus, *International Journal of Corpus Linguistics*, 6:2, 305-325
- J. Lee, K. Gerdes, H. Leung, and T-S Wong (in prog.) Toward a Parallel Treebank of Cantonese and Mandarin.
- K. K. Luke and May L.Y. Wong (2015) The Hong Kong Cantonese Corpus: Design and Uses. *Journal of Chinese Linguistics*.