

PyCantonese: Cantonese linguistic research in the age of big data

Jackson L. Lee
University of Chicago
<http://jacksonllee.com>

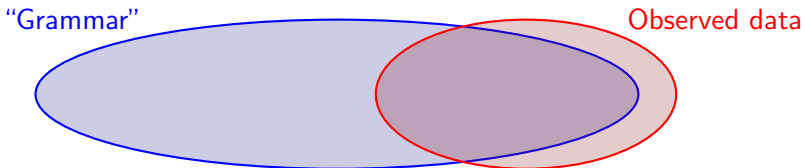
Childhood Bilingualism Research Center, CUHK
September 15, 2015



“Grammar” versus observed data



What is linguistics all about?



- ▶ **In grammar but not observed:**

Arguably the mainstream focus of linguistic research

- Why? Productivity, competence, etc
- How? Introspection, experiments, etc

- ▶ **Observed but not in grammar (?)**:

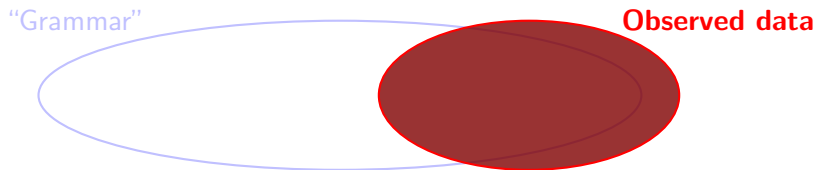
The noisy part of language

(slips of tongue, “I heard it but I’d never say that”, etc)

A bottom-up approach



But grammar is ultimately based on the observed data.



A strongly empirical view of linguistic research:

- ▶ Focus on what is **observed**.
- ▶ Where are the data? There's no shortage...
⇒ **big data research!**

Big data for linguistic research



...as the theme of the 2015 Linguistic Summer Institute at UChicago:



HOME
COURSES
INSTRUCTORS
EVENTS

NEWS & UPDATES
ABOUT
PARTICIPANT INFO
FAQS

**REGISTER
NOW**

Already registered? [LOGIN HERE.](#)

LINGUISTIC
THEORY IN A
WORLD OF BIG
DATA



<https://lsa2015.uchicago.edu/>

Big data + Cantonese?



Some (accessible) Cantonese corpora, by year of publication:

- ▶ The Hong Kong Cantonese Adult Language Corpus (Leung and Law 2001; Leung et al. 2004; Fung and Law 2013)
- ▶ Hong Kong Cantonese Child Language Corpus (Lee and Wong 1998)
- ▶ Cantonese Radio Corpus (Francis and Matthews 2005, 2006)
- ▶ The Hong Kong Bilingual Child Language Corpus (Yip and Matthews 2007)
- ▶ Early Cantonese Tagged Database (Yiu 2012)
- ▶ A Linguistic Corpus of Mid-20th Century Hong Kong Cantonese (Chin 2013)
- ▶ PolyU Corpus of Spoken Chinese (Yap et al. 2014)
- ▶ Hong Kong Cantonese Corpus (Luke and Wong 2015)

Big data + Cantonese?



To what extent are these resources usable and extensible for the general research community?

Issues:

- ▶ inconsistent/ad hoc data formats
- ▶ no general toolkits for handling data



PyCantonese is a toolkit for handling Cantonese corpus data.

- ▶ Evolving and expanding
- ▶ It is a **Python** library – why Python?
 - a general-purpose programming language
 - the lingua franca for computational linguistics and natural language processing
- ▶ Similar data structures as in NLTK (Bird et al. 2009)
- ▶ An open-source tool
- ▶ Current collaborators: Litong Chen, Tsz-Him Tsui
- ▶ Full documentation (with installation instructions):
<http://pycantonese.github.io/>

Accessing corpus data in PyCantonese



PyCantonese comes with builtin corpus data!
Currently, KK Luke's **HKCanCor** is included.

```
<info>
  1-TN-001
  2-DR-300497
  3-NS-2
  4-LS-AB
  5-A-F-34-HK
  6-B-F-37-HK
  INFO-END
</info>
<sent>
  <sent_head>
    A:
  </sent_head>
  <sent_tag>
    喂/e/wai3/
    遲/a/ci4/
    啲/u/di1/
    去/v/heoi3/
    唔/d/m4/
    去/v/heoi3/
    旅行/vn/leoi5hang4/
    啊/y/aa3/
    ? /w/VQ6/
```

The corpus provides word-segmented data with:

- ▶ characters
- ▶ part-of-speech tags
- ▶ Jyutping romanization

Accessing corpus data through PyCantonese



```
>>> import pycantonese as pc
>>> corpus = pc.hkccancor
>>> corpus.number_of_words()
160956
>>> corpus.number_of_characters()
210567
```



Jyutping → **onset, nucleus, coda, tone**

```
>>> import pycantonese as pc
>>> pc.jyutping('hou2')
[('h', 'o', 'u', '2')]
>>> pc.jyutping('hoeng1gong2')
[('h', 'oe', 'ng', '1'), ('g', 'o', 'ng', '2')]
```

Also provided: Conversion from Jyutping to **Yale** or to **LaTeX TIPA**

Basic search capabilities



Possible search queries depend heavily on what *is* encoded and annotated in the corpus data:

Jyutping elements? Part-of-speech tags? Characters?

A screenshot of a web browser showing the 'Search functions' page of the PyCantonese documentation. The browser's address bar shows 'pycantonese.github.io/searches.html'. The page has a light blue header with 'PyCantonese 1.0 documentation' and navigation links 'previous', 'next', and 'index'. The main content area is titled 'Search functions' and contains two paragraphs explaining the search capabilities of the 'search()' function. A code block shows the import and initialization of the pycantonese module. On the right side, there is a 'Table Of Contents' section with links to 'Search functions' and 'Jyutping romanization: Parsing and conversion', and a 'Previous topic' section with a link to 'Jyutping romanization: Parsing and conversion'.

More here: <http://pycantonese.github.io/searches.html>

More search examples



Use filtering strategies for more complicated search queries.

Example: Find in HKCanCor **all verb+noun word pairs**.
(defined as 1st word tag = “V” and 2nd word tag == “N”)

Approach:

1. Find all words tagged as “V” together with the immediately following word.
2. Within the results from step 1, retain only cases where the second word has the tag of “N”.

Finding V+N word pairs



```
>>> import pycantonese as pc
>>> corpus = pc.hkccancor
>>> v = pc.search(corpus, pos="V", word_right=1)
>>> len(v) # number of words with "V"
25364
>>> vn = list()
>>> for wordpair in v:
...     if not wordpair or len(wordpair) < 2:
...         continue
...     if wordpair[1][1] and wordpair[1][1] == "N":
...         vn.append(wordpair) # save V+N
>>> len(vn) # number of V+N word pairs
1535
```

Some V+N pairs found



```
>>> for i in range(3):  
...     print(vn[i])  
[('聽_teng1', 'V'), ('朋友_pang4jau5', 'N')]  
[('跟_gan1', 'V'), ('旅行社_leoi5hang4se5', 'N')]  
[('搭_daap3', 'V'), ('飛機_fei1gei1', 'N')]
```

TODO: Allow regular expressions for search criteria.
e.g., the part-of-speech tag of interest could be anything in the tagset that begins with a “V” (= some sort of verb).

Recurrent problem: Part-of-speech tagging



Some issues of part-of-speech tagging:

1. How many tags do we use?
 - HKCanCor: **46+ tags**
 - Google universal tagset: **12 tags** (Petrov et al. 2011)
2. Relatedly, how fine-grained are the tags?
 - e.g., distinguish proper nouns and common nouns?
3. Human annotation work is time-consuming and costly.

鬼 gwai2 'ghost'



Examples from HKCanCor:

1. 好_hou2/D 鬼_gwai2/D1 細_sai3/A
“very GWAI small”
2. 有_jau5/V1 鬼_gwai2/D1 今日_gam1jat6/T
“resulting-in GWAI today”

What is the tag D1?

These two instances of gwai2 are very different.

(An expressive + negator in (2); see Beltrama and Lee (2015))

Current work:

Mapping HKCanCor to the universal PoS tagset by Petrov et al

A related issue: Word segmentation



Issues of word segmentation:

1. AB
→ compound or two separate words?
2. grammatical characters (e.g., aspect markers)
→ a separate word itself or part of a word?

Interrogatives A-not-A, A-not-AB



If we treat A-not-AB as three words...

What is **hap** in **hap-m-happy**?

Similarly, 鍾唔鍾意 “like or not”, etc.

(In HKCanCor, the first A is treated as an abbreviation, with a tag starting with “J”.)

Or perhaps things like A-not-AB should be treated as one word?

(Lee 2012)

Same problem: **aspect markers**

Ongoing work



- ▶ Corpus data prep
(The Leung-Law-Fung HKCAC, the Francis-Matthews CRCorpus)
- ▶ General tools thus derived

Comparing some Hong Kong Cantonese corpora



Both standard and non-standard data formats have been used.

HKCanCor

<info>

1-TN-001
2-DR-300497
3-NS-2
4-LS-AB
5-A-F-34-HK
6-B-F-37-HK
INFO-END

</info>

<sent>

<sent_head>

A:

</sent_head>

<sent_tag>

喂/e/wai3/

遲/a/ci4/

啲/u/di1/

去/v/heoi3/

唔/d/m4/

去/v/heoi3/

旅行/vn/leoi5hang4/

啊/y/aa3/

? /w/VQ6/

HKCAC

102	1	O	M	H1	我			聽	聽	下	一	位	聽	眾	
102	1	P	M	H1	O5	tei6	tHEN1	tHEN1	ha6	At1	wAi2	tHIN3	tsoN3		
102	2	O	M	H1	王	[生]	早	晨	王	生		
102	2	P	M	H1	wON4	[saN1]	tsou2	sAn4	wON4	saN1		
102	3	O	M	C	[x]						
102	3	P	M	C	[x]						
102	4	O	M	C	係	早	晨	早	晨	呀	[係	係		
102	4	P	M	C	hAi6	tsou2	sAn4	tsou2	sAn4	a3	[hAi6	hAi6		
102	5	O	M	H2	[x	你	好	係]			
102	5	P	M	H2	[x	lei5	hou2	hAi6]			

CRCorpus

@Font: Win95:Courier:-13:0

@Begin

@Participants: HS1 Host 1, JKC Jacky, SP1 speaker 1, SP2 speaker 2, SP3 speaker 3, CZK Can4zi2koeng4, CL1 caller 1, CL2 caller 2.

@sex of HS1: male

@sex of CKC: male

@comment: RTHK1:

@TOP: interview

@Location: HK

@Date: 10-NOV-2000

@ID: can.hk00.JackyChan.1011(Date)=HHH

@Dependent: eng

@Time Duration: 2:56-3:56

@Tape Location: tape 2, side A

*HS1: zeihai6 kei4sat6 lei5 lei4 dou3 gam1jat6.

*mor: conj|zeihai6=that_is advs|kei4sat6=actually nnpr|lei5=you

dir|lei4=come vt|dou3=arrive advs|gam1jat6=today

*pos: conj|zeihai6=that_is advs|kei4sat6=actually nnpr|lei5=you dir|lei4=come

vt|dou3=arrive advs|gam1jat6=today

*eng: 'You have reached,

Potential new tools in PyCantonese



...and a call for arms!

- ▶ A part-of-speech tagger
- ▶ Conversion between Jyutping and characters, both directions (Issues: Homophony and homography)
- ▶ Word segmentation (with all the usual problems!)

Data, data, data

Ultimately, what is observed *is* the data.



Data format:

- ▶ General direction for PyCantonese:
Adopting the CHILDES **CHAT** format (MacWhinney 2000)
- ▶ Reasons:
 - Rich annotations
 - It is well documented and supported.
 - XML format available by conversion
⇒ readable by NLTK – and PyCantonese!
- ▶ What about non-conversational data?

Data prep:

- ▶ Other (publicly available) datasets out there?
- ▶ Audio(-visual) data?
- ▶ What annotations are desirable?

[Update 2015-09-22]

Additional notes and code snippets are available here:

<http://jacksonllee.com/papers/Lee-pycantonese-2015.html>

References I



- Beltrama, Andrea and Jackson L. Lee. 2015. Great pizzas, ghost negations: The emergence and persistence of mixed expressives. In *Proceedings of Sinn und Bedeutung* 19.
- Bird, Steven, Edward Loper and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.
- Chin, Andy C. 2013. New resources for Cantonese language studies: A linguistic corpus of mid-20th century Hong Kong Cantonese. *Newsletter of Chinese Language* 92(1): 7–16.
- Francis, Elaine J. and Stephen Matthews. 2005. A multi-dimensional approach to the category 'verb' in Cantonese. *Journal of Linguistics* 41: 269–305.
- Francis, Elaine J. and Stephen Matthews. 2006. Categoriality and object extraction in Cantonese serial verb constructions. *Natural Language and Linguistic Theory* 24: 751–801.
- Fung, Suk-Yee and Sam-Po Law. 2013. A phonetically annotated corpus of spoken Cantonese: The Hong Kong Cantonese Adult Language Corpus. *Newsletter of Chinese Language* 92(1): 1–5.
- Lee, Jackson L. 2012. Fixed-tone reduplication in Cantonese. In *McGill Working Papers in Linguistics* 22(1). *Proceedings from the Montreal-Ottawa-Toronto (MOT) Phonology Workshop 2011: Phonology in the 21st Century: In Honour of Glyne Piggott*.

References II



- Lee, Thomas Hung-Tak and Colleen Wong. 1998. CANCORP: The Hong Kong Cantonese Child Language Corpus. *Cahiers de Linguistique Asie Orientale* 27(2): 211–228.
- Leung, Man-Tak and Sam-Po Law. 2001. HKCAC: The Hong Kong Cantonese adult language corpus. *International Journal of Corpus Linguistics* 6: 305–326.
- Leung, Man-Tak, Sam-Po Law and Suk-Yee Fung. 2004. Type and token frequencies of phonological units in Hong Kong Cantonese. *Behavior Research Methods, Instruments, and Computers* 36(3): 500–505.
- Luke, Kang-Kwong and May Lai-Yin Wong. 2015. The Hong Kong Cantonese Corpus: Design and uses. *Journal of Chinese Linguistics* .
- MacWhinney, Brian. 2000. *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Petrov, Slav, Dipanjan Das and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086* .
- Yap, Foong Ha, Ying Yang and Tak-Sum Wong. 2014. On the development of sentence final particles (and utterance tags) in Chinese. In Kate Beeching and Ulrich Detges (eds.), *Discourse functions at the left and right periphery*, 179–220. Leiden: Koninklijke Brill NV.
- Yip, Virginia and Stephen Matthews. 2007. *The Bilingual Child: Early Development and Language Contact*. Cambridge University Press.
- Yiu, Carine Yuk-Man. 2012. Reconstructing early Chinese dialectal grammar: A study of directional verbs in Cantonese. Talk at the Workshop on Innovations in Cantonese Linguistics, March 16–17, Columbus: The Ohio State University.