PyCantonese: Developing computational tools for Cantonese linguistics

Jackson L. Lee, Litong Chen, Tsz-Him Tsui
University of Chicago and The Ohio State University

The 3rd Workshop on Innovations in Cantonese Linguistics
The Ohio State University
March 12, 2016

What is missing in Cantonese linguistics?

Name subfields with lots of work on Cantonese!

phonetics, phonology, morphology, syntax, semantics, pragmatics, sociolingusitics, historical linguistics, discourse and conversation analysis...

How about...

Computational linguistics?

We are concerned with the strongly **empirical** and **data-driven** kind of computational linguistics.

Why computational linguistics? Why data?

Reproducible research

- Verifiable claims in linguistic research

Modeling learnability

- How does grammar come from data?

The socio-political status of Cantonese (?)

- Preserving data → Protecting and promoting the language

Apparent lack of computational linguistics for Cantonese

∵ Lack of data?

We *do* have data! (And we need more...)

Several Cantonese corpora

Adult Cantonese:

- The Hong Kong Cantonese Adult Language Corpus (Leung and Law 2001; Leung et al. 2004; Fung and Law 2013)
- Cantonese Radio Corpus (Francis and Matthews 2005, 2006)
- PolyU Corpus of Spoken Chinese (Yap et al. 2014)
- Hong Kong Cantonese Corpus (Luke and Wong 2015)

Child developmental data:

- Hong Kong Cantonese Child Language Corpus (Lee and Wong 1998)
- The Hong Kong Bilingual Child Language Corpus (Yip and Matthews 2007)

Non-contemporary Cantonese:

- Early Cantonese Tagged Database (Yiu 2012)
- A Linguistic Corpus of Mid-20th Century Hong Kong Cantonese (Chin 2013)

So, what *is* missing?

?????

corpora

custom formats!

divergent annotations!

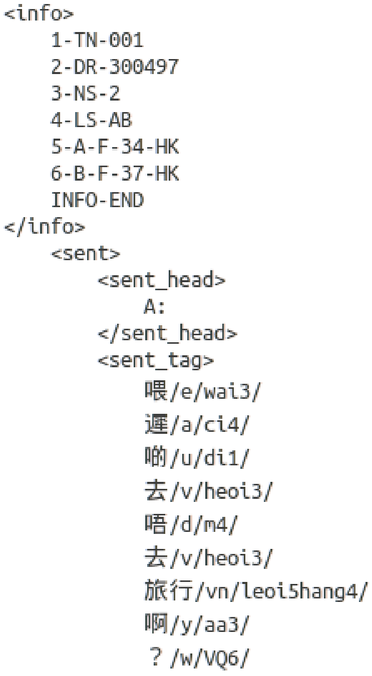
researchers

ARGH!

Comparing some Hong Kong Cantonese corpora

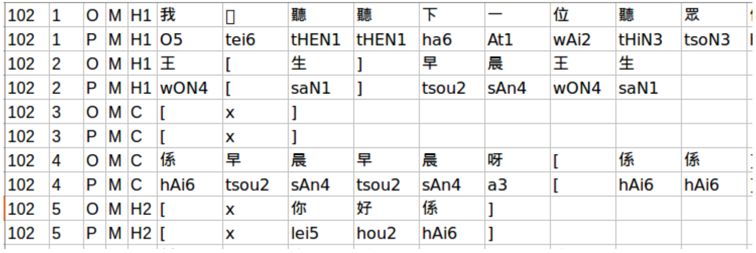
Both standard and non-standard data formats have been used.

HKCanCor



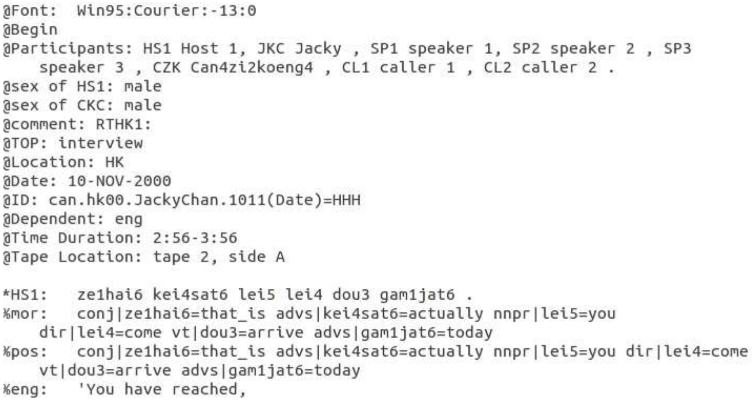
```
<info>
    1-TN-001
    2-DR-300497
    3-NS-2
    4-LS-AB
    5-A-F-34-HK
    6-B-F-37-HK
    INFO-END
</info>
    <sent>
        <sent_head>
            A:
        </sent_head>
        <sent_tag>
            喂/e/wai3/
            遲/a/ci4/
            啲/u/di1/
            去/v/heoi3/
            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:   ze1hai6 kei4sat6 lei5 lei4 dou3 gam1jat6 .
%mor:   conj|ze1hai6=that_is advs|kei4sat6=actually nnpr|lei5=you
    dir|lei4=come vt|dou3=arrive advs|gam1jat6=today
%pos:   conj|ze1hai6=that_is advs|kei4sat6=actually nnpr|lei5=you dir|lei4=come
    vt|dou3=arrive advs|gam1jat6=today
%eng:   'You have reached,
```

Using multiple corpora in research?

It's hard!

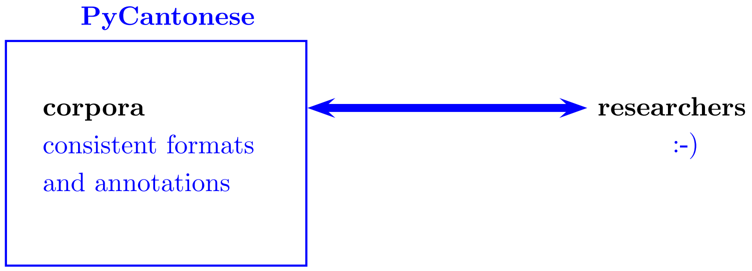
∵ Individual corpora are usually compiled for specific purposes

⇒ Different foci in annotations and formatting

Some recent work that could have benefited from more data:

- Chen (2015): phonological variation of keoi5 's/he' in HKCAC
- Tsui (2014): functional load of Cantonese tones in HKCanCor

PyCantonese – General goals



PyCantonese

corpora
consistent formats
and annotations

researchers
:-)

Data format

PyCantonese adopts the CHILDES **CHAT** format (MacWhinney 2000).

- Rich annotations for conversational data
- Well documented and supported
- PyCantonese piggybacks on PyLangAcq (Lee et al. 2016) for handling the CHAT format.

(How about non-conversational data?)

PyCantonese – Background

PyCantonese is a growing toolkit for computational work in Cantonese linguistics.

- 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)
- A free and open-source tool
- Full documentation (with installation instructions): `http://pycantonese.org/`

Basic functionality

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

For some given corpus data, we can ask about its basic information...

Let’s begin...

```
>>> import pycantonese as pc
>>> corpus = pc.hkcancor()
>>> corpus.number_of_files()
58
>>> corpus.number_of_utterances()
15938
```

Accessing corpus data

words()

```
>>> all_words = corpus.words()
>>> len(all_words)
149781
>>> all_words[:10]
```

```
[’喂’, ’遲’, ’o的’, ’去’, ’唔’, ’去’, ’旅行’, ’啊’, ’?’, ’你’]
```

characters()

```
>>> all_characters = corpus.characters()
>>> len(all_characters)
186888
>>> all_words[:10]
```

```
[’喂’, ’遲’, ’o的’, ’去’, ’唔’, ’去’, ’旅’, ’行’, ’啊’, ’?’]
```

Word-level annotations

tagged_words()

a tagged word =
(word, part-of-speech tag, Jyutping, grammatical relations)

```
>>> all_tagged_words = corpus.tagged_words()
>>> all_tagged_words[:4]
```

```
[(’喂’, ’E’, ’wai3’, ”), (’遲’, ’A’, ’ci4’, ”), (’o的’, ’U’, ’di1’, ”), (’去’, ’V’, ’heoi3’, ”)]
```

(More on grammatical relations in a minute!)

Other methods: http://pycantonese.org/reader.html
— utterance-level structures, word frequency info, etc.

Parsing Jyutping

parse_jyutping()

Jyutping → (onset, nucleus, coda, tone)

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

Search queries

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

Jyutping elements? **Part-of-speech tags**? **Characters**?
A combination of any of these?

Additional features:

- Search by a word/sentence range
- Search by a regular expression

Details — http://pycantonese.org/searches.html

Example: jau5 ‘have’, C. Lam (2016a) 1 hour ago
Example: *aa* is the only onsetless syllable with all 6 tones in HKCanCor, cf. Z. Lam (2016b) 2 hours ago

Ongoing work

- Corpus reformatting (currently the HKCAC dataset)
- Devising tools for filling in the gaps in formatting and annotations across corpora

Anticipated functionality

- Jyutping ↔ characters (issues: homophony and homography)
- word segmentation (a perennial problem for CJK languages)
- part-of-speech tagging (depending on tagset etc)

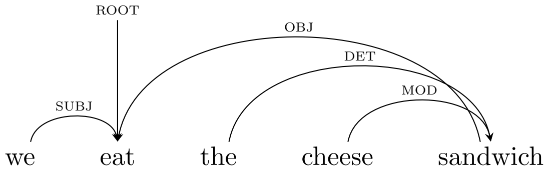
We’d need these for preparing a usable corpus dataset based on, say, the novel 男人唔可以窮 from the HK Golden Forum!

More on the to-do list

- Forced alignment (cf. Peters and Tse (2016) 30 min ago)
- Dependency and grammatical relations

English (example from the CHILDES CLAN menu)

```
*TXT:   we        eat        the       cheese    sandwich
%mor:   pro|we    v|eat      det|the   n|cheese  n|sandwich
%gra:   1|2|SUBJ  2|0|ROOT   3|5|DET   4|5|MOD   5|2|OBJ
```



Moving Cantonese linguistics forward

- We all need one another.
- PyCantonese opens the door for *shared* and *open-access* resources.
- Call for arms!
  PyCantonese is a *collaborative* project.
- Questions, comments, bug reports, feature requests etc are more than welcome.

References

References

Bird, Steven, Edward Loper and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Chen, Litong. 2015. Variations of the third-person singular pronoun in Hong Kong Cantonese. In *University of Pennsylvania Working Papers in Linguistics*, vol. 21, 1.8, 1–5.

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.

Lam, Charles. 2016a. Multiple functions of HAVE in Cantonese: a corpus study. Presented at the 3rd Workshop on Innovations in Cantonese Linguistics (WICL-3), The Ohio State University.

Lam, Zoe. 2016b. Temporal location of perceptual cues for Cantonese tone identification. Presented at the 3rd Workshop on Innovations in Cantonese Linguistics (WICL-3), The Ohio State University.

Lee, Jackson L., Ross Burkholder, Gallagher B. Flinn and Emily R. Coppess. 2016. Working with CHAT transcripts in Python. Tech. Rep. TR-2016-02, Department of Computer Science, University of Chicago.

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 Computer* 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.

Peters, Andrew and Holman Tse. 2016. Evaluating the efficacy of Prosody-lab Aligner for a study of vowel variation in Cantonese. Presented at the 3rd Workshop on Innovations in Cantonese Linguistics (WICL-3), The Ohio State University.

Tsui, Tsz-Him. 2014. Tonal variation in Hong Kong Cantonese: acoustic distance & functional load. In Andrea Beltrama, Tasos Chatzikonstantinou, Jackson L. Lee, Mike Pham, and Diane Rak (eds.), *Proceedings of the Forty-eighth Annual Meeting of the Chicago Linguistic Society*, 579–588. Chicago: Chicago Linguistic Society.

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.