Discontinuous constructions in spoken Chinese varieties: Extraction and contrastive covarying collexeme analysis

Ryan Ka Yau Lai
University of California, Santa Barbara
kayaulai@ucsb.edu

1 Introduction

Linguistic theory has long been interested in the co-occurrence of linguistic forms in discourse. Co-occurrence not only sheds light on individual forms, but also reveals *collocations* or *multi-word expressions* (MWEs), conventionalised strings of linguistic forms that co-occur frequently in discourse and are stored as chunks in memory. While most computational corpus-based studies of examine contiguous strings of co-occurring words, e.g. *kick the bucket* or *strong coffee*, some have examined co-occurrence between non-contiguous words, e.g. ***turn*** *the heater* ***off***, ***too*** *heavy* ***to*** *use* (e.g. Watanabe 2021, Dunn 2017). These non-contiguous strings often constitute partially lexically filled constructions in construction grammar: abstract constructions consisting of both concrete lexical items and a slot that can take a variety of expressions, e.g. [*turn* NP *off*].

This project extends such methodology to Chinese discontinuous constructions, focusing on constructions where a preverbal or earlier particle (usually adverbs or conjunctions, which I call *nonfinal particles* (NFPs)) tends to co-occur with a final particle (FP). Example (1) comes from the Hong Kong Cantonese Corpus (HKCanCor; Luke & Wong 2015):

(1) **唔通** 到 黎 明 **咩**
m⁴tung¹ dou³ lai⁴ ming⁴ **me¹**
NFP go Lai Ming **FP**
'What else, is it Leon Lai's turn?' [FC-105_v2]

Here, the NFP 唔通 *m⁴tung¹* expresses speculativeness, sarcasm or scepticism (Matthews & Yip 2011), all of which suggest the proposition is improbable; the FP 咩 *me¹* indicates a negatively-biased question (Yiu 2021). These two meanings are highly semantically compatible.

This study has two purposes. Firstly, I extend the methodology of detecting discontinuous multi-word expressions to the particularly difficult case of particle frames by combining and extending existing methods. In particular, I examine particle frames from the POS-tagged HKCanCor (mostly copresent conversations), plus the Mandarin CallHome corpus of telephone calls (Liu et al. 2004), POS-tagged and parsed using spaCy (Honnibal & Montani 2017). I first extract candidate combinations with a hybrid of existing methods, pruning results by statistically selecting only pairs with strong evidence for attraction/repulsion while controlling the false discovery rate.

Secondly, I carry out a contrastive *covarying collexeme analysis* (Stefanowitsch & Gries 2005), where I examine the co-occurrence properties of forms in two constructional slots – here, the NFP and the FP. I examine how NFPs and FPs co-occur within utterances (treated as constructions), through measures of attraction, repulsion and productivity. I then assess how well these statistics support the common claim in the literature that particle frames are much more common in Cantonese than Mandarin (Tang 2015).

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 324-328.
Irvine, California, June 27-29, 2024

2 Methodology

2.1 Extraction

Existing work. Three classes of methods are commonly used for non-contiguous collocations: window-based, POS-based and tree-based methods. None are fully sufficient for Chinese particle frames.

First, approaches based on fixed window sizes (e.g. Fissaha & Haller 2003, Watanabe 2021) are inadequate since there is no limit on the amount of text between the two particles of a particle frame. Consider (2):

(2) 嗷 **唔通** 即係 佢
gam^{2} **m^{4}tung^{1}** zek^{1}hai^{6} keoi^{5}
then **NFP** that.is 3sg
又 係 為咗 錢，
jau^{6} hai^{6} wai^{6}zo^{2} cin^{2}
also be for money
哩 樣 嘢 而 去 做 **咩？**
ni^{1} joeng^{6} je^{5} ji^{4} heoi^{3} zou^{6} **me^{1}**
this CL thing and go do **FP**
'Is it for this money thing that they did it then?' [FC-109a_v2]

Secondly, methods based on POS sequences[1] (e.g. Wible & Tsao 2010) cannot capture the fact that the POS of words within particle frames can be very free. While some POS-based methods (e.g. Baldwin & Villavicencio 2001, Dunn 2017) merge intervening words into chunks like noun phrases, i.e. ***turn*** *it* ***off*** and ***turn*** *the heater* ***off*** can both be detected as instances of ***turn*** [NP] ***off***, this is still insufficient since elements between FPs and NFPs cannot just be summarised by a grammatical constituent type. For example, (1) contains just a verb and an argument, (2) contains two coordinated clauses with multiple arguments, and there are cases (though not found in HKCanCor) where a single noun phrase may intervene between *m^{4}tung^{1}* and *me^{1}*.

Finally, *dependency-based* methods (e.g. Martens & Vandeghinste 2010) use subtrees of dependency syntax trees. Aside from the lack

Figure 1: Dependency graph of (1) using Universal Dependencies. The discontinuous construction does not form a subtree.

of resources for Cantonese syntax, there is an additional problem: the two particles often do not form a dependency subtree. For example, in (1), both particles depend on *dou^{3}* 'go', which is not part of the construction (Figure 1).

Proposed method. My method combines these existing methods. Similar to Liu et al (2019), potential pairs were searched within *utterances*, since particle frames are utterance-level phenomena. I first segmented the text into utterances using punctuation in the corpus. For the Cantonese corpus, NFPs were extracted by taking adverbs (tagged `d` in HKCanCor) or conjunctions (`c`) separated from the end of the utterance by at least a verb (`v`), and FPs were extracted by taking words tagged as FPs (`y`) plus a few manually determined final adverbs:

(3) **唔通** 到 黎 明 **咩**
NFP go Lai Ming **FP**
d v nr nr **y**
'What else, is it Leon Lai's turn?'

For Mandarin, adverbs / conjunctions that precede their parents and have `advmod` relation with them were treated as NFPs, and words tagged as particles with a `discourse` dependency were treated as FPs (Figure 2).

Figure 2: Dependency parse of a Mandarin sentence with an NFP and two FPs.

[1] The term *part-of-speech* excludes sign languages; I have retained as it is most understandable to a computational audience.

In both varieties, each co-occurrence of a NFP and a FP within an utterance was treated as a candidate pair. When there were multiple NFPs and/or FPs, all possible pairings were extracted. I also recorded utterances with only one of the two, or neither. In this case, the NFP and/or FP absent from the utterance were recorded as NA, as illustrated in Table 1:

NFP	FP	Extracted pairs
A	q	{(A, q)}
A, B	q	{(A, q), (B, q)}
A	q, r	{(A, q), (A, r)}
A	/	{(A, NA)}
/	q, r	{(NA, q), (NA, r)}
/	/	{(NA, NA)}

Table 1 How pairs were extracted.

2.2 Pruning

To select only those pairs where there is strong evidence that the NFP and FP prefer to co-occur above chance level, the following contingency table was obtained for each extracted pair (including those with NAs), and Fisher's exact test was performed with `fisher.test` in R (R core team, 2023):

	A	Not A
q	#(A, q)	#(not A, q)
Not q	#(A, not q)	#(not A, not q)

Table 2: Sample contingency table for {A, q}.

For example, to calculate the *p*-value for *m⁴tung¹ … me¹* (the particle frame in (1)), #(A, q) is the frequency of *m⁴tung¹ … me¹*, #(not A, q) is the frequency of utterances with *me1* but not *m⁴tung¹*, #(A, not q) is the number of utterances with *m⁴tung¹* but not *me¹*, and #(not A, not q) is the number of utterances with neither particle. This was done for every logically possible pair of particles, even if it is unattested.

To determine the threshold at which to remove candidate pairs, I used the procedure in Benjamini & Yekutieli (2001). Given the smaller sample size, I control false discovery rate at 20%. Surviving pairs are those with significant evidence that the FP and NFP prefer or disprefer co-occurring with each other.

2.3 Covarying collexeme analysis

After extraction, several measures of the relationship between the two particles were calculated for the covarying collexeme analysis: pointwise mutual information (PMI) for bidirectional association, normalised Kullback-Leibler divergence (KLD)-based measures of unidirectional association, and the and entropy of the two particles normalised by log-sample size for productivity (Gries 2022).

3 Results

98 (Mandarin) and 108 (Cantonese) pairs survived the Benjamini-Yekutieli procedure. Of these, 55 (Cantonese) and 54 (Mandarin) pairs represent cases of attraction (PMI > 0) between two particles, which may be interpreted as particle frames. The other 'pairs' involve an absent particle, or are cases of repulsion (PMI < 0) (Figure 3).

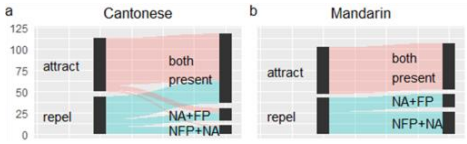

Figure 3: Distribution of types of pairs extracted in Cantonese and Mandarin.

The following subsections present the covarying collexeme analysis in detail.

3.1 Particle frames

As seen above, the number of particle frames found for Cantonese and Mandarin are similar despite the Mandarin corpus being around

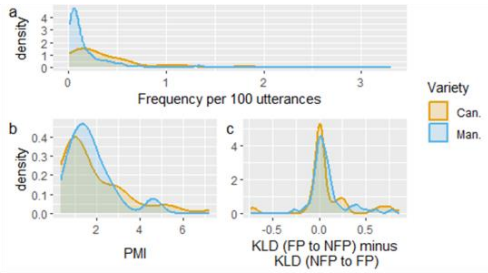

Figure 4: Kernel density estimates (KDEs) of (a) normalised token frequencies of extracted frames (b) PMIs between particles of extracted frames, and (c) differences between the FP-to-NFP and NFP-to-FP KLD-based measures of unidirectional attraction.

twice as large (21189 utterances in HKCanCor vs 40460 in CallHome). With regards to token frequency, Cantonese particle frames tend to be used more often (Figure 4).

However, attraction-related properties of such pairs in Cantonese and Mandarin seem very similar: Most PMIs are below 4 (though >4 values are somewhat more common in Cantonese). From the KLDs, most attraction is either symmetric (i.e. the two particles attract each other about equally), or the FP is much more attracted to the NFP than vice versa (consistent with general pressures on interactive language use, where utterance-final particles are most vulnerable to overlap).

Thus, Mandarin and Cantonese appear to have comparable numbers of particle frames at similar level and directionality of attraction, seemingly contradicting the impression that Cantonese is far richer in particle frames.

3.2 Particle repulsion

Repulsion data paints a very different picture than attraction: Cantonese has 27 repelling pairs of particles, while Mandarin has none.

3.3 Properties of particles

Cantonese has 2 NFPs and 6 FPs that prefer *not* to co-occur with the other particle type (i.e. are attracted to NA); Mandarin, none. Mandarin has 26 NFPs and 16 FPs particles that disprefer *not* appearing with the other particle type; Cantonese only 8 NFPs and 9 FPs.

If one examines normalised entropy (excluding *hapax legomena*, for which normalised entropy cannot be calculated), the normalised entropy of the NFP slot for FPs is generally higher in Mandarin than in Cantonese: 22/50 (44%) of FPs have normalised entropy <.5 in Cantonese, compared to 6/25 (24%) FPs in Mandarin. By contrast, 65/276 (24%) of NFPs have FP entropy greater than <.5, compared to 214/372 (58%) in Mandarin. In other words, predicting NFPs from FPs is easier in Mandarin than in Cantonese; predicting FPs from NFPs is easier in Mandarin than in Cantonese.

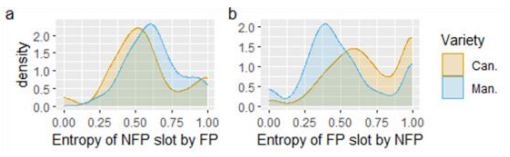

Figure 5: Entropies of (a) the NFP slot for each FP and (b) the FP slot for each NFP.

4 Discussion

The results show how using multiple corpus statistics can bring further nuance to claims in comparative syntax. While Cantonese is typically thought to be much richer in particle frames, we find a set of highly associated particles in both varieties with similar attractional strength and direction.

What distinguishes Cantonese from Mandarin is the 'exclusivity' of such constructions. Cantonese particles are more likely than in Mandarin to reject particles they 'dislike', as shown in statistics for repulsion and preference for appearing without the other particle type. This seems to be primarily driven by the presence of more specific final particles, as shown in lower normalised entropy for NFPs given the FP slot.

These facts can be explained qualitatively. Mandarin frames often have semantically general FPs compatible with many different NFPs, so the semantic relation is relatively loose. In 挺 *tǐng* … 的 *de* ('quite…FP'), the particle frame with the lowest *p*-value, the FP expresses certainty and the adverb moderately high degree. These are semantically compatible, but each particle is compatible with many other contexts. Cantonese particle frames tend to have semantically specific FPs. For example, the particle frame with lowest *p*-value was 咪 *mai*^{6} …囉 *lo*^{1}. *lo*^{1} suggests that the current assertion should be obvious given something else the speaker knows (Wakefield 2011), while the sole purpose of *mai*^{6} seems to be adding emphasis to *lo*^{1}-statements and defining the scope of focus as the content between *mai*^{6} and *lo*^{1} (Wakefield 2020). This specificity may make Cantonese particle frames more psychologically salient, supporting the intuition that Cantoense has more particle frames.

References

Baldwin, Timothy & Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *COLING-02: The 6th Conference on Natural Language Learning 2002* (CoNLL-2002).

Benjamini, Yoav & Daniel Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*. 1165–1188.

Cook, Angela. 2023. Reconsidering the *shi…(de)* construction in spoken Mandarin. *Chinese Language and Discourse. An International and Interdisciplinary Journal* 14(2). 209–231. https://doi.org/10.1075/cld.18003.coo.

Dunn, Jonathan. 2017. Computational learning of construction grammars. *Language and Cognition* 9(2). 254–292. https://doi.org/10.1017/langcog.2016.7.

Gries, Stefan Th. 2022. Multi-word units (and tokenization more generally): a multi-dimensional and largely information-theoretic approach. *Lexis. Journal in English Lexicology*. Université Jean Moulin Lyon 3 (19).

Fissaha, Sisay & Johann Haller. 2003. Application of corpus-based techniques to Amharic texts. In Proc. MT Summit IX Workshop on Machine Translation for Semitic Languages

Honnibal, Matthew & Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Liu, Yi, Pascale Fung, Shudong Huang, Christopher Cieri, Lufeng Zhai & Benfeng Chen. 2004. Development of a Chinese telephony conversational corpus for speech processing. In *2004 International Symposium on Chinese Spoken Language Processing: Proceedings: December 15-18, 2004, the Chinese University of Hong Kong, Hong Kong*, 197. Institute of Electrical & Electronics Engineers (IEEE).

Liu, Xiaoxia, Degen Huang, Zhangzhi Yin & Fuji Ren. 2019. Recognition of collocation frames from sentences. *IEICE Transactions on Information and Systems*. 102(3). 620–627.

Luke, Kang Kwong & May LY Wong. 2015. The Hong Kong Cantonese corpus: design and uses. In Zhou Bin, Simon Smith & Michael Hoey (eds.), *Linguistic corpus and corpus linguistics in the Chinese context* (Journal of Chinese Linguistics Monograph Series 43), 312–333.

Matthews, Stephen & Virginia Yip. 2011. *Cantonese: a comprehensive grammar* (Routledge Comprehensive Grammars). 2nd ed. London ; New York: Routledge.

Martens, Scott & Vincent Vandeghinste. 2010. An efficient, generic approach to extracting multi-word expressions from dependency trees. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, 85–88.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Stefanowitsch, Anatol & Stefan Th. Gries. 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory* 1(1). 1–43. https://doi.org/10.1515/cllt.2005.1.1.1.

鄧思穎 [Tang, Sze-Wing]. 2015. *粵語語法講義 [Lecture Notes on Cantonese Grammar]*. Hong Kong: 商務印書館(香港)有限公司.

Wakefield, John C. 2011. The English equivalents of Cantonese sentence-final particles: A contrastive analysis. Hong Kong Polytechnic University.

Wakefield, John C. 2020. *Intonational Morphology* (Prosody, Phonology and Phonetics). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-15-2265-9.

Watanabe, Kohei. 2021. Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages. *Communication Methods and Measures* 15(2). 81–102. https://doi.org/10.1080/19312458.2020.1832976.

Wible, David & Nai-Lung Tsao. 2010. StringNet as a computational resource for discovering and investigating linguistic constructions. In *Proceedings of the NAACL HLT workshop on extracting and using constructions in computational linguistics*, 25–31.

Yip, Ka-Fai. 2023. A compositional account of “only” doubling. Syntax-Semantics Reading Group (LFRG), Massachusetts Institute of Technology.

Yiu, Carine Yuk-man. 2021. The origin and development of the question particle me1 in Cantonese. *Lingua* 254. 103049. https://doi.org/10.1016/j.lingua.2021.103049.