A construction-based approach to Cantonese classifiers

Francis Bond (凡土) and Joanna Ut-Seong Sio (蕭)
Department of Asian Studies
Palacký University, Olomouc
bond@ieee.org, joannautseong.sio@upol.cz

1 Introduction

Cantonese, a variety of Yue, belongs to the Sinitic branch of the Sino-Tibetan language family. Originating from southern China, it is named after Canton (Guangzhou), the capital city of the Guangdong province. Cantonese is spoken in Guangdong China, and the two Special Administrative Regions, Hong Kong and Macao, as well as in diaspora communities (e.g., Singapore, Malaysia, Australia, the United Kingdom and North America). There are over 82.4 million Cantonese native language speakers (Wikipedia contributors, 2024).

This paper focuses on Cantonese NPs, and contributes to their HPSG analysis in three ways. First, we account for the differences between NUME-CL-N and CL-N, despite both are interpreted as having the cardinality of 'one'; second, we propose a specifier-based analysis that does not require nouns to have two specifiers; third, this analysis allows us to correctly assign cognitive status to different types of Cantonese NPs. The analysis is implemented in an open-source Cantonese HPSG.[1]

2 Cantonese NPs

(Unmodified) Cantonese NPs have the following 4 schematic forms (demonstrative (D), numeral (X: numeral phrase or one of a small set of quantifiers), classifier (C), noun (N), and they have different definiteness interpretations (Cheng and Sybesma, 1999), as indicated in the table below:

Table 1: Definiteness (after Cheng and Sybesma, 1999)

Type	Example
D-(X)-C-N	definite
X-C-N	indefinite
C-N	(in)definite
N	indefinite

The following sentences illustrate the different types NPs in the object position:

(1) Cantonese (yue)

a. D-(X)-C-N

明恩	食咗	呢	(一)	個	蘋果。
Ming4jan1	sik6-zo2	nei1	jat1	go3	ping4gwo2
Ming-Jan	eat-PERF	this	one	CL	apple

'Ming-Jan ate this apple.'

b. X-C-N

明恩	食咗	一	個	蘋果。
Ming4jan1	sik6-zo2	jat1	go3	ping4gwo2
Ming-Jan	eat-PERF	one	CL	apple

'Ming-Jan ate one apple.'

[1] The implementation, using the DELPH-IN tools, is available at ⟨https://github.com/neosome/yue⟩.

c.	C-N	明恩	食咗	個	蘋果。
		Ming4jan1	sik6-zo2	go3	ping4gwo2
		Ming-Jan	eat-PERF	CL	apple

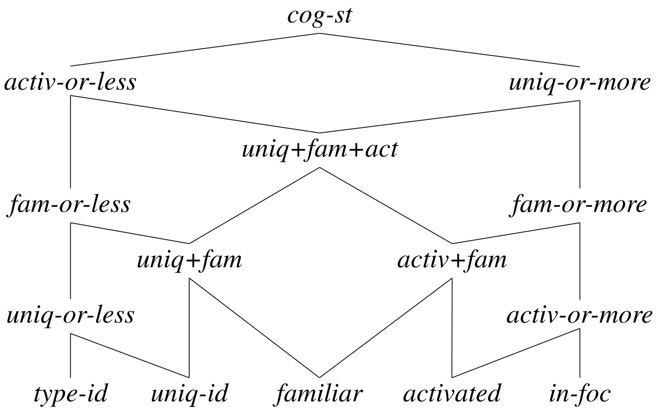
‘Ming-Jan ate an/the apple.’

d.	N	明恩	食咗	蘋果。
		Ming4jan1	sik6-zo2	ping4gwo2
		Ming-Jan	ate-PERF	apple

‘Ming-Jan ate an apple/apples.’

In Chinese, in general, only definite NPs can appear in the subject or topic position in a sentence (Li and Thompson, 1989). Definiteness is understood as the grammatical encoding of the pragmatic concept of identifiability (Chen, 2004). Identifiability is related to the assumptions made by the speaker on the cognitive status of a referent in the mind of the addressee in the context of an utterance (Gundel et al., 1993) . Borthen and Haugereid (2005) provide an HPSG-based type hierarchy of cognitive status, which was then refined by Bender and Goss-Grubbs (2008), as shown below:

(2)



Different languages have different inventories of referring expressions that can be used for different cognitive statuses. In Cantonese, we propose the interpretations in Table 2. We also show the desired semantics (using indexed MRS: Copestake et al., 2005).

Table 2: Cognitive status

Type	Example	cog-st	Semantics
D-(X)-C-N	呢 (一) 個蘋果	*fam-or-more*	h_0:{呢 _q(x_1, h_2, h_3); card(e_4, x_1 ′1′), 個 _x(e_5, x_1); 蘋果 _n(x_1)}
X-C-N	一個蘋果	*type-id*	h_0:{exist_q(x_1, h_2, h_3); card(e_4, x_1 ′1′), 個 _x(e_5, x_1); 蘋果 _n(x_1)}
C-N	個蘋果	*fam-or-less*	h_0:{exist_q(x_1, h_2, h_3); card(e_4, x_1 ′1′), 個 _x(e_5, x_1); 蘋果 _n(x_1)}
N	蘋果	*type-id*	h_0:{exist_q(x_1, h_2, h_3); 蘋果 _n(x_1)}

In Sio and Song (2015), D-(X)-C-N covers all cognitive statuses except *type-id* in (2), i.e., *uniq-or-more*. In this paper, we restrict D-(X)-C-N to *fam-or-more*. D-(X)-C-N is not used in cases of *uniq-id*. *uniq-id* (uniquely identifiable) is defined as the addressee being able to identify the referent on the basis of the nominal alone. This covers cases which Schwarz (2009) calls *larger situation definites* (e.g., the moon), *immediate situation definites* (in a room with one door clearly open, e.g., close the door, please.) and part-whole bridging definites (e.g., I bought a shirt yesterday. The buttons are too big.). In these situations, C-N rather than D-(X)-C-N is used in Cantonese. In Sio and Song (2015), C-N is totally under-specified, compatible with all *cog-st*. In this paper, we restrict it to *fam-or-less*, excluding it from *activated* and *in-foc*. Activated is defined as represented in current working memory (Gundel et al., 1993); while *in-foc* is defined as ‘the referent is not only in short-term memory, but is also at the current center of attention’(Gundel et al., 1993). In these cases, D-(X)-C-N, or pronouns are used in Cantonese. Both D-(X)C-N and C-N can be used in *familiar* contexts. *familiar* is defined as a referent that already has a representation in memory (Gundel et al., 1993). In a context where both the speaker and the hearer know the neighbor downstairs has a dog, and the dog is currently barking, the speaker can utter the following (SFP = sentence-final-particle; QP = question-particle):

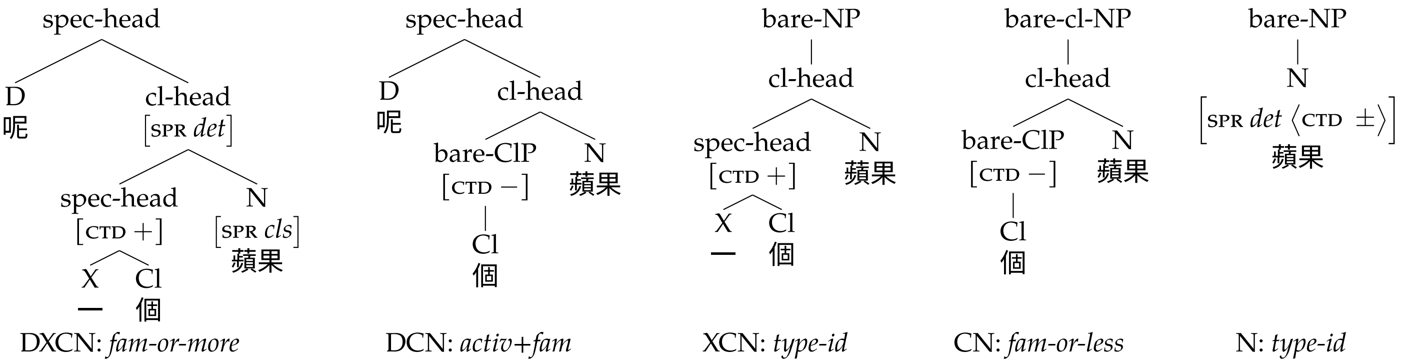


Figure 1: NPs eith and without demonstratives

(3)								
	聽到	嗎?	(嗰)	隻	狗	又	吠	啦 。
	teng-1dou2	maa3	go2	zek3	gau2	jau6	fai6	laa3
	listen	qp	that	cl	dog	again	bark	sfp

'Do you hear that? The/That dog is barking again.' [yue]

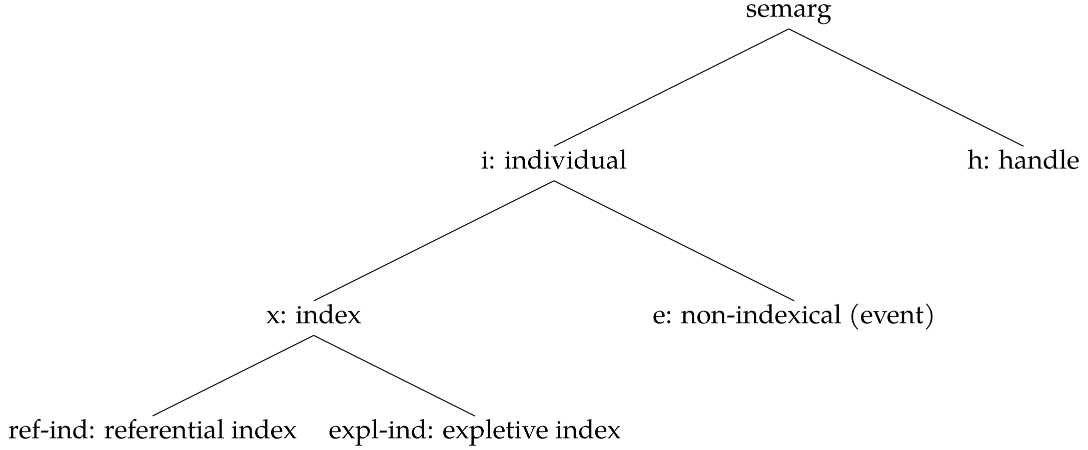
A note of caution is required here. The cognitive status of a referent is not always easy to determine, we follow the coding guidelines from the protocol for each cognitive status in Gundel (2010). It is possible that in some situations, the choice could just be a strong preference.

In Cantonese, when the numeral is omitted, both x-c-n and c-n have a cardinality of 'one'. However, in answering the question 'how many', only x-c-n can be used. This is, in part, similar to the contrast between 'one n' and 'a/an n' in English. The semantics represents this with the *card* relation, with a value of '1'. In addition, the well-formed semantics must have a quantifier for every referential index, if there is no explicit demonstrative, the grammar must supply this from a construction.

3 Analysis

Following the majority of HPSG analyses on Chinese NPs (Wang and Liu, 2007, and references therein), we adopt an NP analysis, where the numeral forms a constituent together with the classifier (Her, 2016). We treat both the demonstrative and classifier as specifiers, following the analysis of Mandarin by Ng (1997) and Wang and Liu (2007). However, instead of the nouns selecting two specifiers and modifying the head-specifier rule, we add a new classifier construction (*cl-head*: §3) which requires another specifier after consuming the classifier. This makes the classifier-construction the locus of the unusual syntax. Empirical data from a wide range of languages does not require two specifiers for an adequate description of noun phrases, so we attempt an analysis with one. In future work, we will attempt to discover if there are different predictions from the two approaches.

Our analysis requires one new lexical type (for sortal classifiers); one new feature used on classifier phrases to mark if they have been explicitly enumerated or not and three new constructions (classifier-head, bare-classifier and bare-classifier-np) as well as changes to numerals, head-specifier and the existing bare-np rule. Derivation trees are shown for the 5 NP types in Figure 1. We describe in more detail below. The descriptions given below are all only partial, we omit information we consider not relevant to the discussion at hand. Paths may also be shortened for clarity.



3.1 Lexical types

3.1.1 Classifier lexical type

The sortal classifier lexical type is shown in Figure 2a. The category is *cls* for classifier. The *?* shows where the predicate would be for an actual entry of a word. They optionally take a number as their specifier. The head-specifier rule will link the XARG to the INDEX of the specified constituent.

The sortal classifier lexical type doesn't say anything about cognitive status, nominals containing the classifier are compatible with all cognitive status. The ultimate cognitive status of a nominal containing a classifier is determined by (i) whether it is preceded by a numeral; (ii) whether the nominal contains a demonstrative.

$$\begin{bmatrix} \textit{classifier-lex} & \\ \textsc{syn} & \begin{bmatrix} \textsc{head} & \textit{cls} \\ \textsc{val.spr} & \left\langle \begin{bmatrix} \textsc{syn.head} & \textit{num,} \\ \textsc{opt} & + \end{bmatrix} \right\rangle \end{bmatrix} \\ \textsc{sem} & \begin{bmatrix} \textsc{hook} & \begin{bmatrix} \textsc{index} & \boxed{1} \\ \textsc{xarg} & \boxed{2} \end{bmatrix} \\ \textsc{rels} & \left\langle \begin{bmatrix} \textsc{reln} & \textit{?} \\ \textsc{arg0} & \boxed{1}\ \textit{non-ref} \\ \textsc{arg1} & \boxed{2} \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix}$$

(2a) Classifier lexical type

$$\begin{bmatrix} \textit{numeral-lex} & \\ \textsc{syn} & \begin{bmatrix} \textsc{head} & \textit{numeral} \\ \textsc{val.spec} & \left\langle \begin{bmatrix} \textsc{syn.head.ctd} & + \\ \textsc{sem.index} & \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix} \\ \textsc{sem} & \begin{bmatrix} \textsc{hook} & \begin{bmatrix} \textsc{index} & \boxed{2} \end{bmatrix} \\ \textsc{rels} & \left\langle \begin{bmatrix} \textsc{reln} & \textit{card-relation} \\ \textsc{carg} & \textit{?} \\ \textsc{arg0} & \boxed{2}\ \textit{non-ref} \\ \textsc{arg1} & \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix}$$

(2b) Numeral lexical type

3.1.2 Numeral lexical type

Their semantics is somewhat special, using CARG (Constant Argument) to introduce the value of the number. The index of the thing it will specify over (the classifier) is the same as ARG1 on the relation it introduces. That is, it counts the classifier. Further, it sets it's head to *ctd* +.

In the implementation we reuse the PRON feature it is only used on NPs and we only use it on ClPs

3.1.3 Noun lexical type

$$\begin{bmatrix} \textit{noun-lex} & \\ \textsc{syn} & \begin{bmatrix} \textsc{head} & \textit{noun} \\ \textsc{val.spr} & \left\langle \begin{bmatrix} \textsc{syn.head} & \textit{cls,} \\ \textsc{opt} & + \end{bmatrix} \right\rangle \end{bmatrix} \\ \textsc{sem} & \begin{bmatrix} \textsc{hook} & \begin{bmatrix} \textsc{index} & \boxed{1} \end{bmatrix} \\ \textsc{rels} & \left\langle \begin{bmatrix} \textsc{arg0} & \boxed{1}\ \textit{ref-ind} \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix}$$

The Cantonese *noun-lex* sets its specifier to be a classifier, not a determiner.

3.1.4 Demonstrative

A demonstrative constrains the index of the noun it specifies to be *fam-or-more*, it does not care about the CTD value of its specifier.

$$\begin{bmatrix} \textit{noun-lex} & \\ \textsc{syn} & \begin{bmatrix} \textsc{head} & \textit{det} \\ \textsc{val.spec} & \left\langle \begin{bmatrix} \textsc{local.sem.hook.index} & \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix} \\ \textsc{sem} & \begin{bmatrix} \textsc{hook} & \begin{bmatrix} \textsc{index} & \boxed{1}\ \begin{bmatrix} \textsc{cog-st} & \textit{activ+fam} \end{bmatrix} \end{bmatrix} \\ \textsc{rels} & \left\langle \begin{bmatrix} \textsc{arg0} & \boxed{1}\ \textit{ref-ind} \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix}$$

3.2 Rules

3.2.1 Classifier Head rule (*cl-head*)

This rule is the main new construction. It takes two daughters. The left-hand, non-head daughter (NHD) takes a classifier phrase as its daughter. The right-hand, head daughter (HD), takes a noun or nominal that requires a classifier as it's specifier. Crucially, the parent also requires a specifier, this time a determiner: in this way a noun phrase can effectively have two specifiers, so long as the first is a classifier, and the second a determiner, even though the noun has only one specifier. The value of CTD is passed from the non-head daughter (the specifier) to the new specifier slot, making it visible to the bare NP rules. In most other ways it is identical to the *spec-head* rule.

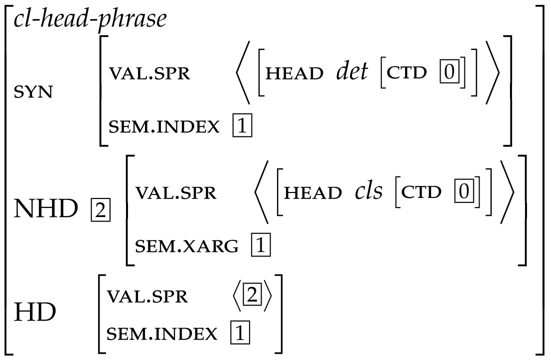


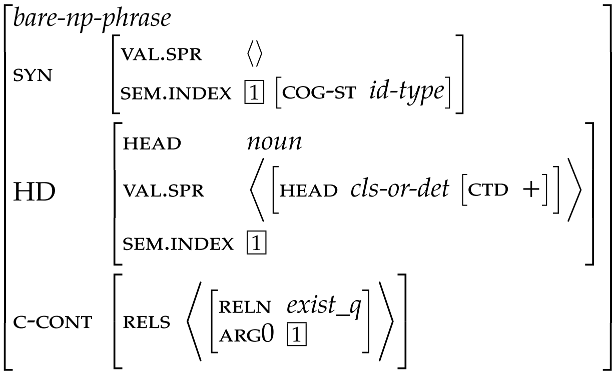
Figure 3: Classifier Head rule

3.2.2 Head specifier rule (*spec-head*)

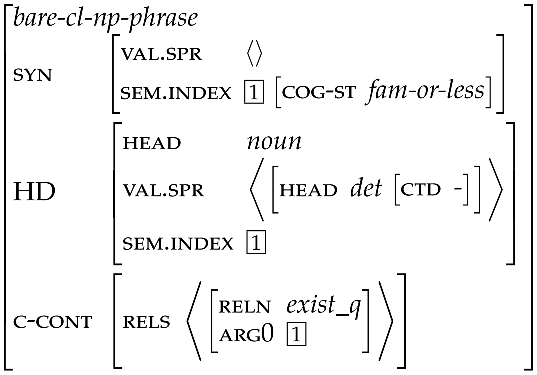
we fix to not allow classifier as specifier

3.2.3 Bare NP rules (*bare-NP*, *bare-cl-NP*)

We introduce two bare NP rules, for the two different cognitive statuses we want.


bare-np-phrase
SYN
VAL.SPR ⟨⟩
SEM.INDEX 1 [COG-ST id-type]
HD
HEAD noun
VAL.SPR ⟨[HEAD cls-or-det [CTD +]]⟩
SEM.INDEX 1
C-CONT
RELS ⟨[RELN exist_q ARG0 1]⟩

(4a) Bare NP rule (N or X-C-N)


bare-cl-np-phrase
SYN
VAL.SPR ⟨⟩
SEM.INDEX 1 [COG-ST fam-or-less]
HD
HEAD noun
VAL.SPR ⟨[HEAD det [CTD -]]⟩
SEM.INDEX 1
C-CONT
RELS ⟨[RELN exist_q ARG0 1]⟩

(4b) Bare NP rule for bare ClP (C-N)

The first (4a) is a headed unary rule, which makes an NP with the specifier satisfied, if the head daughter's specifier is *cls-or-det* and *ctd* $+$. This will be true for nouns with a numeral and classifier as input, or just for a noun, as its CTD is unspecified. The *cog-st* of the resulting NP is set to *type-id* . The second (4b) restricts the value of the head daughter's spec to a determiner (DET) with CTD $-$, and the NP's *cog-st* is set to *fam-or-less*. This excludes bare nouns, whose specifier is *cls* and nouns specified with a classifier and no numeral, which will be CTD $+$. In the grammar, they both inherit from a single supertype *bare-np-super* which contains the shared structure.

3.2.4 Bare classifier rule

This non-branching rule takes a classifier, and creates a classifier phrase. As the interpretation is always that there is one thing being classifies, the rule adds a *card-relation* with CARG of *1*. It also sets CTD to $-$ so that the classifier phrase will pass through the Bare NP rule for bare classifiers (3.2.3). The rule is similar to the NO-SPR-CL-RULE proposed by (Sio and

Song, 2015, p189), but differs in two important ways. The first is that it explicitly models the cardinality. The second is that it marks the head so that the cognitive status can be restricted.

4 Conclusion and future work

In this paper, we presented our preliminary attempt in generating different nominal types in Cantonese (with construction-specific rules) as well as mapping them to different cognitive statuses in HPSG. In the future, we want to expand our investigation in the following directions. Our analysis does not investigate the effects of modification on the semantics or cognitive status, nor the anaphoric use of the classifier (in the absence of the head noun). We also have only looked at sortal classifiers, not mensural or kind. With the inclusion of cognitive statuses, we would like to model the restriction on banning indefinite NPs (i.e., *type-id*) appearing in subject and topic position in Chinese (Li and Thompson, 1989). We would like to extend the analysis to cover these, and test against naturally occurring texts. Finally, although we have focused on Cantonese here, we would like to compare our analysis to those of other classifier languages, especially those with computational analyses like Indonesian, Japanese and Mandarin.

References

Emily M. Bender and David Goss-Grubbs. 2008. Semantic Representations of Syntactically Marked Discourse Status in Crosslinguistic Perspective. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, pages 17–29. Association for Computational Linguistics.

Kaja Borthen and Petter Haugereid. 2005. Representing Referential Properties of Nominals. *Research on Language and Computation*, 3(2-3):221–246.

Ping Chen. 2004. Identifiability and definiteness in Chinese. *Linguistics*, 42(6):1129–1184.

Lisa Lai-Shen Cheng and Rint Sybesma. 1999. Bare and not-so-bare nouns and the structure of NP. *Linguistic inquiry*, 30(4):509–542.

Ann Copestake, Dan Flickinger, Ivan A. Sag, and Carl Pollard. 2005. Minimal Recursion Semantics. An introduction. *Research on Language and Computation*, 3(4):281–332.

Jeanette Gundel. 2010. Reference and accessibility from a givenness hierarchy perspective. *International Review of Pragmatics*, 2(2):148–168.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive Status and the Form of Referring Expressions in Discourse. *Language*, 69(2):274–307.

One-Soon Her. 2016. Structure of numerals and classifiers in Chinese. *Historical and typological perspectives and cross-linguistic implications*, pages 28–29.

Charles N Li and Sandra A Thompson. 1989. *Mandarin Chinese: A functional reference grammar*. Univ of California Press.

Say Kiat Ng. 1997. A double specifier account of Chinese NPs using head-driven phrase structure grammar. Msc thesis, Department of Linguistics, University of Edinburgh.

Florian Schwarz. 2009. *Two types of definites in natural language*. University of Massachusetts Amherst.

Joanna Ut-Seong Sio and Sanghoun Song. 2015. Divergence in expressing definiteness between Mandarin and Cantonese. In *Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar, Nanyang Technological University (NTU), Singapore*, pages 177–194, Stanford, CA. CSLI Publications.

Lulu Wang and Haitao Liu. 2007. A description of Chinese NPs using Head-Driven Phrase Structure Grammar. In *Proceedings of the 14th International Conference on Head-Driven Phrase Structure Grammar, Stanford Department of Linguistics and CSLI's LinGO Lab*, pages 287–305, Stanford, CA. CSLI Publications.

Wikipedia contributors. 2024. Cantonese — Wikipedia, the free encyclopedia. [Online; accessed 24-March-2024].