A Construction-based Approach to Cantonese Classifiers

Francis Bond
Palacký University

Joanna Ut-Seong Sio
Palacký University

Proceedings of the 31st International Conference on
Head-Driven Phrase Structure Grammar

Palacký University Olomouc, Czech Republic

Stefan Müller, Rui Chaves (Editors)

2024

Frankfurt/Main: University Library

pages 60–75

Keywords: classifier, Cantonese, HPSG, specifier, cognitive status, givenness, computational grammar

Bond, Francis & Joanna Ut-Seong Sio. 2024. A Construction-based Approach to Cantonese Classifiers. In Stefan Müller & Rui Chaves (eds.), *Proceedings of the 31st International Conference on Head-Driven Phrase Structure Grammar, Palacký University Olomouc, Czech Republic*, 60–75. Frankfurt/Main: University Library. DOI: 10.21248/hpsg.2024.4.

Abstract

This paper focuses on the structure and interpretation of Cantonese NPs. We first map different cognitive statuses (Gundel et al. 1993) to different forms of Cantonese noun phrases, following the hierarchy proposed in Borthen & Haugereid (2005). We then provide an HPSG analysis for Cantonese noun phrases. We account for the differences between classifiers appearing with and without a numeral, where classifiers with no numeral are interpreted as having a cardinality of 'one'. We propose a Classifier Head Rule where the noun first takes a specifier containing the classifier, and the output further takes a determiner as its specifier. The analysis is implemented in an open-source Cantonese HPSG.

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 provides an HPSG analysis for Cantonese noun phrases with the following three implementations. First, we assign cognitive status to different types of Cantonese NPs, following the hierarchy proposed in Borthen & Haugereid (2005). Second, we account for the differences between classifiers appearing with and without a numeral. Classifiers with no numerals are interpreted as having a cardinality of 'one'. Third, we propose a Classifier Head Rule where the noun first takes a specifier containing the classifier, and the output further takes a determiner as its specifier. This is similar to a double specifier analysis except that the locus of the 'special treatment' (having to take both the classifier and then the determiner as specifiers) is built around the classifier. The analysis is implemented in an open-source Cantonese HPSG.[1]

2 Cantonese NPs

Cantonese NPs (unmodified) have the 4 schematic forms shown in Table 1 and they have different definiteness interpretations (Cheng & Sybesma 1999). In (1), we give example sentences illustrating the different types of NPs in the object position.

[†]We would like to thank the reviewers, Dan Flickinger, Emily Bender and Luis Morgado da Costa for their helpful comments and discussion.

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

Table 1: Definiteness (after Cheng & Sybesma 1999)

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

D: demonstrative, X: numeral,* C: classifier, N: noun

*X can in fact be a numeral phrase or one of a small set of quantifiers, in this paper, we only discuss X being a numeral.

(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.'

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 (Cantonese included), only definite NPs can appear in the subject or topic position in a sentence (Li & Thompson 1989), though not without exceptions (Li 1998).[2] Thus, it is important to include the definiteness information of the NPs when modeling Cantonese grammar. 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).

[2] Li (1998) argues that when the interpretation of a Chinese [X-C-N] phrase (indefinite) has only a quantity reading rather than an individual reading, it can appear in the subject position.

Gundel et al. (1993)'s analysis has been implemented as a type-hierarchy in (Borthen & Haugereid 2005) and adequately describes the distinctions needed for Cantonese. We adopt Gundel et al. (1993)'s model in this paper.

Gundel et al. (1993) proposes six cognitive statuses: 'type-identifiable', 'referential', 'uniquely identifiable', 'familiar', 'activated', and 'in focus'. Each cognitive status can be expressed with different forms of noun phrases in different languages. The different forms serve as processing signals to the addressee.[3] Each of the cognitive statuses will be discussed below with Cantonese example sentences.

'Type-identifiable' refers to cases where the addressee is able to access a representation of the type of objects described by the expression (Gundel et al. 1993). The English indefinite article is used in such cases. In Cantonese, the same cognitive status can be expressed by a bare noun, or a [(x)-c-n] phrase, with the numeral being optional.

(2)	我	去	買	*(一)*	*(個)*	西瓜	。
	ngo5	heoi3	maai5	(jat1)	(go3)	sai1gwaa1	
	1sg	go	buy	(one)	(cl)	watermelon	

'I go buy a watermelon/watermelons.'

The only requirement on the addressee is that they understand the noun *sai1gwaa1* 'watermelon' to understand what it is to be bought. Note that when only a bare noun is used, it can be interpreted either as singular or plural.

'Referential' refers to cases where the speaker intends to refer to a particular object (or objects). The addressee needs to access a appropriate type-representation, plus either retrieving an existing representation of the speaker's intended referent or construct a new representation the time the sentence is uttered (Gundel et al. 1993). Borthen & Haugereid (2005) argue that cognitive status is speaker-oriented, and so

[3] One reviewer suggested other models of givenness, in particular, the four distinctions: (i) discourse-old and hearer-old; (ii) discourse-old and hearer-new; (iii) discourse-new and hearer-old, and (iv) discourse-new and hearer-new, as discussed in Birner (2021). We believe these four categories can be captured by Gundel et al. (1993)'s hierarchy, e.g., 'discourse-old and hearer-old' can be subsumed under 'familiar'; 'discourse-new and hearer-new' can be subsumed under 'type-identifiable'; 'discourse-new and hearer-old' are cases like e.g., 'the sun', 'the President', which can be subsumed under 'uniquely identifiable' (as these referents are unique in any particular context without prior introduction). 'Discourse-old and hearer-new' are cases where the referent is 'inferrable' (Prince 1981, Schwarz 2009), as in e.g., 'John put away all his grooming tools. *The combs* he put into the top drawer' (Birner 2021: 263). This can also be subsumed under 'uniquely identifiable', due to the prior introduction of an 'anchor' ('grooming tools' in the example), which makes the referent ('the combs') unique. In both 'discourse-new and hearer-old' and 'discourse-old and hearer-new' cases, [c-n] phrases are used in Cantonese, in other words, a distinction in forms is not made in Cantonese in these two cases. Given that the same forms of Cantonese noun phrases are often used for multiple categories of givenness, with the trend that the more 'given' (the more accessible in the addressee's mental representation) is associated with the use of the demonstrative (rather than clearly demarcated categorical 'forms to givenness' matching), we believe that Gundel et al. (1993)'s hierarchy provides a sufficient and fitting model for our purpose. Furthermore, as noted in the main text, Gundel et al. (1993) is adopted mainly because it has been implemented as a type-hierarchy.

'referential' is taken out of the hierarchy. They treat 'referential' as specificity which is speaker-oriented and introduce a separate feature (SPECI bool) that can be cross-classified with the different cognitive statuses (both definite and indefinite). We follow them to exclude 'referential' from our hierarchy of cognitive status.

'Uniquely identifiable' refers to cases where the addressee can identify referent on the basis of the nominal alone (Gundel et al. 1993). Identifiability does not have to be familiarity if enough descriptive content is provided (Gundel et al. 1993). In these situations, a [C-N] phrase is used in Cantonese, as shown in the example below (PERF = perfective marker; SFP = sentence-final-particle). The sentence can be used when there is only one open window in the non-linguistic context.

(3)	閂咗	個	窗	佢	丫	。
	saan1-zo2	go3	coeng1	keoi5	aa1	
	close-PERF	CL	window	3SG	SPF	

'Close the window.'

'Familiar' is when the addressee is able to uniquely identify the intended referent because they already has a representation of it in memory (in long-term memory if it has not been recently mentioned or perceived, or in short-term memory if it has) (Gundel et al. 1993). In these cases, both a [C-N] phrase and a [D-(X)-C-N] phrase can be used in Cantonese. For example, in a context where a particular student has been mentioned earlier on in the discourse, using *(go)2 go3 hok6saang1* 'the/that student' to refer to her/him would be appropriate.

(4)	*(*嗰*)*	個	學生	去咗	邊	呀？
	(go2)	go3	hok6saang1	heoi3-zo2	bin1	aa3
	(that)	CL	student	go-PERF	where	SPF

'Where does the student go?'

'Activated' is defined as a referent being represented in current working memory; it can be retrieved from long term memory, or they may arise from the immediate linguistic or extra-linguistic contexts (Gundel et al. 1993). In these cases, a demonstrative has to be used, as in *go2 di1 seng1* 'that noise' below:

(5)	嗰	啲	聲	攪	到	我	成	晚	都	瞓唔着。
	go2	di1	seng1	gaau2	dou3	ngo5	seng4	maan5	dou1	fan3-m4-zoek6
	that	CL	noise	make	to.the.extent	1SG	whole	night	also	sleep-NEG-fall

'That noise made me unable to sleep the whole night'

In-foc is represented not only in short-term memory, but is also at the current center of attention' (Gundel et al. 1993). In these cases, again, a demonstrative needs to be used, as in *go2 tiu4 jyu2* 'that fish' below:

(6)	嗰	條	魚	實在	太	好味	啦	。
	go2	tiu4	jyu2	sat6zoi6	taai3	hou2mei6	laa3	
	that	CL	fish	indeed	too	yummy	SPF	

'That fish is indeed too yummy.'

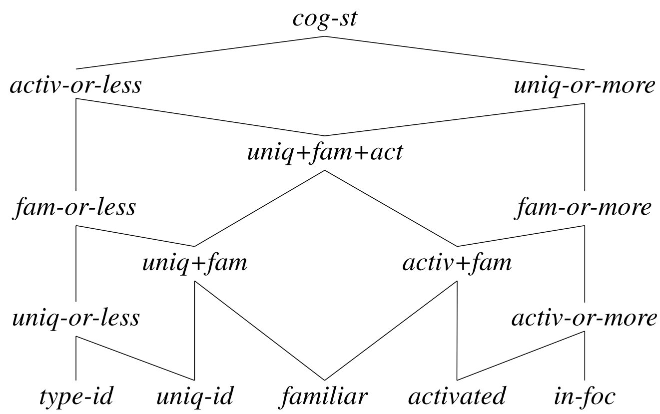


Figure 1: Cognitive Status Hierarchy

Borthen & Haugereid (2005) provide an HPSG-based type hierarchy of cognitive status, which was then refined by Bender & Goss-Grubbs (2008), as shown in Figure 1. 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.

Table 2: Cognitive status

Type	**Example**	**cog-st**	Definiteness
D-(X)-C-N	呢 (一) 個蘋果	*fam-or-more*	Definite
X-C-N	一個蘋果	*type-id*	Indefinite
C-N	個蘋果	*fam-or-less*	In/Definite
N	蘋果	*type-id*	Indefinite (or Generic)

In Sio & Song (2015), D-(X)-C-N covers all cognitive statuses except *type-id* in Figure 1, 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. We believe 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 & 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*.

The general tendency of the mapping between the cognitive statuses to the Cantonese NPs is such that the demonstrative, D-(X)-C-N, is required when the cognitive status reaches a certain level of prominence (i.e., 'familiar', 'activated', 'in-focus') while C-N spans over some less prominent cognitive status (i.e., 'type-identifiable', 'uniquely-identifiable' and 'familiar'). D-(X)-C-N and C-N overlap in covering 'familiar' cases. A note of caution is required here. The mapping between a certain cognitive status to a particular NP form is not always easy to determine, we follow the coding guidelines from the protocol for each cognitive status in Gundel et al. (1993) to the best of our understanding. It is possible that in some situations, the choice could just be a preference.[4]

3 Analysis

Following the majority of HPSG analyses on Chinese NPs (Wang & 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 & Liu (2007). However, instead of the nouns selecting two specifiers and modifying the HEAD-SPECIFIER rule, we add a new classifier construction (*cl-head*: §**??**) which requires another specifier after consuming the classifier. Empirical data from a wide range of languages does not require two specifiers for an adequate description of noun phrases and it is the classifier that is special in the Cantonese noun phrase, thus, we make the classifier-construction the locus of the unusual syntax. Currently, we have found no data that differentiates clearly between our one-specifier analysis and the two-specifier analysis. In future work, we will attempt to discover if there are different predictions from the two approaches.

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.

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

[4]It is not easy to determine whether [C-N] phrases can cover 'activated'; similarly, it is not easy to decide whether [D-(X)-C-N] phrases can cover 'uniq-id'. At any rate, the mapping can be easily adjusted for future work (e.g., for D-(X)-C-N, from *fam-or-more* to *uniq-or-more*.). We have chosen a more restrictive approach in this paper.

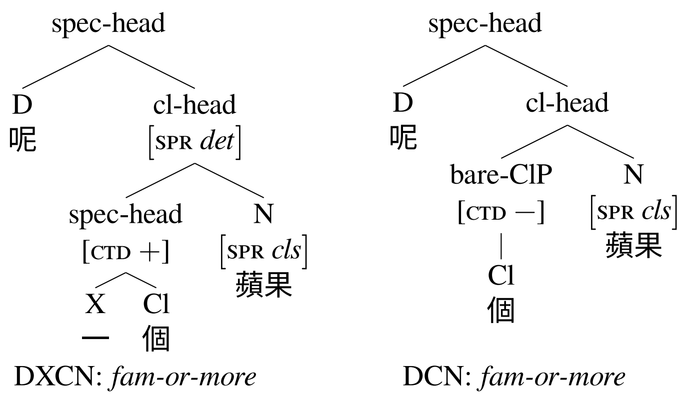


Figure 2: NPs with demonstratives

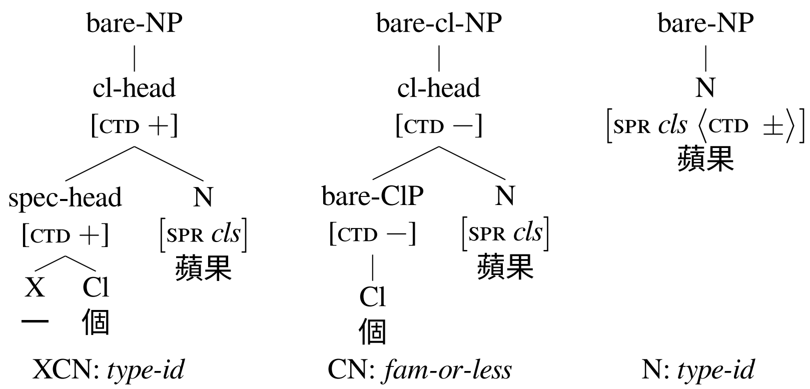


Figure 3: NPs without demonstratives

three new constructions (classifier-head, bare-classifier and bare-classifier-np) as well as changes to the existing lexical types for numerals, and the head-specifier and bare-np rules. Derivation trees are shown for the two NP types with demonstratives in Figure 2 and the three types without in Figure 3. 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.

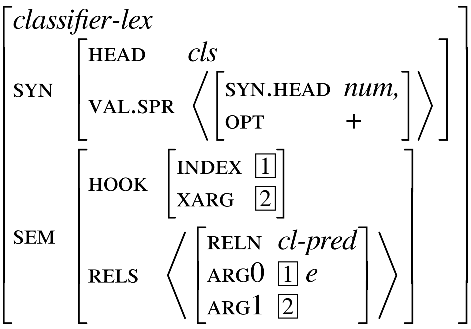
3.1 Lexical types

3.1.1 Classifier lexical type

The sortal classifier lexical type is shown in (7). The category is *cls* for classifier. The *cl-pred* 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.

(7)



Because sortal classifiers do not introduce a referent, their type is *e* (non-indexical). This contrasts with referential noun-phrases, which will be *ref-ind* (referential index). The types we use are shown in Figure 4.

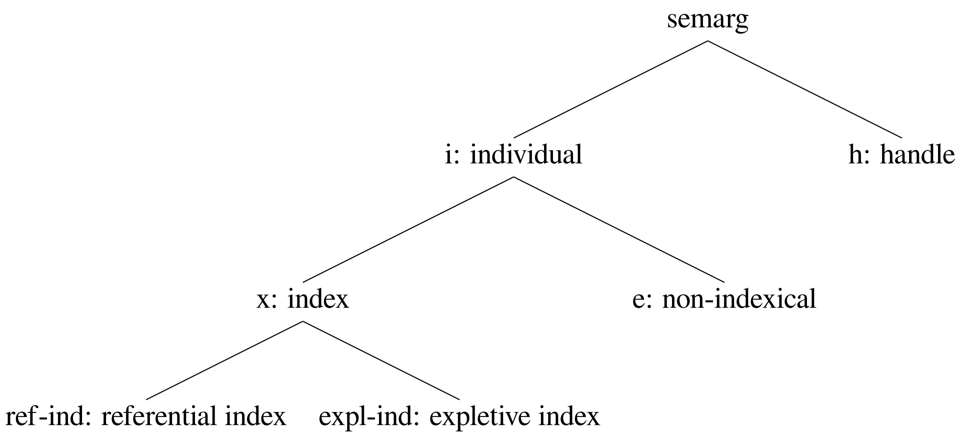


Figure 4: Types of semantic objects

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 +: it has been explicitly counted.

In the implemented grammar, rather than defining a new feature, we reuse the PRON feature. This makes the size of the feature structure smaller. Because PRON was originally only used on NPs and we only use it here on ClPs, its interpretation is never ambiguous.

(8)
$$\begin{bmatrix} \textit{numeral-lex} & \\ \text{SYN} & \begin{bmatrix} \text{HEAD} & \textit{numeral} \\ \text{VAL.SPEC} & \left\langle \begin{bmatrix} \text{SYN.HEAD.CTD} & + \\ \text{SEM.INDEX} & \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix} \\ \text{SEM} & \begin{bmatrix} \text{HOOK} & \begin{bmatrix} \text{INDEX} & \boxed{2} \end{bmatrix} \\ \text{RELS} & \left\langle \begin{bmatrix} \text{RELN} & \textit{card-relation} \\ \text{CARG} & \textit{?} \\ \text{ARG0} & \boxed{2}\ \textit{non-ref} \\ \text{ARG1} & \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix}$$

3.1.3 Noun lexical type

The Cantonese *noun-lex* (9) sets its specifier to be a classifier, not a determiner. This means it must either pick up a specifier, or have the specifier discharged by the bare NP rule.

(9)
$$\begin{bmatrix} \textit{noun-lex} & \\ \text{SYN} & \begin{bmatrix} \text{HEAD} & \textit{noun} \\ \text{VAL.SPR} & \left\langle \begin{bmatrix} \text{SYN.HEAD} & \textit{cls,} \\ \text{OPT} & + \end{bmatrix} \right\rangle \end{bmatrix} \\ \text{SEM} & \begin{bmatrix} \text{HOOK} & \begin{bmatrix} \text{INDEX} & \boxed{1} \end{bmatrix} \\ \text{RELS} & \left\langle \begin{bmatrix} \text{ARG0} & \boxed{1}\ \textit{ref-ind} \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix}$$

3.1.4 Demonstrative

A demonstrative (10) constrains the index of the noun it specifies to be *fam-or-more*, it does not care about the CTD value of its specifier.

(10)
$$\begin{bmatrix} \textit{dem-lex} & \\ \text{SYN} & \begin{bmatrix} \text{HEAD} & \textit{det} \\ \text{VAL.SPEC} & \left\langle \begin{bmatrix} \text{LOCAL.SEM.HOOK.INDEX} & \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix} \\ \text{SEM} & \begin{bmatrix} \text{HOOK} & \begin{bmatrix} \text{INDEX} & \boxed{1} \begin{bmatrix} \text{COG-ST} & \textit{activ+fam} \end{bmatrix} \end{bmatrix} \\ \text{RELS} & \left\langle \begin{bmatrix} \text{ARG0} & \boxed{1}\ \textit{ref-ind} \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix}$$

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 its 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 (and thus can inherit from a common super-type).

(11)
$$\begin{bmatrix} \textit{cl-head-phrase} & \\ \text{SYN} & \begin{bmatrix} \text{VAL.SPR} & \left\langle \begin{bmatrix}\text{HEAD} & \textit{det} \begin{bmatrix}\text{CTD} & \boxed{0}\end{bmatrix}\end{bmatrix}\right\rangle \\ \text{SEM.INDEX} & \boxed{1}\end{bmatrix} \\ \text{NHD}\ \boxed{2} & \begin{bmatrix} \text{VAL.SPR} & \left\langle \begin{bmatrix}\text{HEAD} & \textit{cls} \begin{bmatrix}\text{CTD} & \boxed{0}\end{bmatrix}\end{bmatrix}\right\rangle \\ \text{SEM.XARG} & \boxed{1}\end{bmatrix} \\ \text{HD} & \begin{bmatrix} \text{VAL.SPR} & \left\langle \boxed{2} \right\rangle \\ \text{SEM.INDEX} & \boxed{1}\end{bmatrix} \end{bmatrix}$$

3.2.2 Head Specifier Rule (*spec-head*)

The head specifier rule has one change: we do not allow a classifier as specifier — in this case, the classifier head rule should be used instead.

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. The first (12) 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*.

(12)
$$\begin{bmatrix} \textit{bare-np-phrase} & \\ \text{SYN} & \begin{bmatrix} \text{VAL.SPR} & \langle\rangle \\ \text{SEM.INDEX} & \boxed{1} \begin{bmatrix}\text{COG-ST} & \textit{id-type}\end{bmatrix}\end{bmatrix} \\ \text{HD} & \begin{bmatrix} \text{HEAD} & \textit{noun} \\ \text{VAL.SPR} & \left\langle \begin{bmatrix}\text{HEAD} & \textit{cls-or-det} \begin{bmatrix}\text{CTD} & +\end{bmatrix}\end{bmatrix}\right\rangle \\ \text{SEM.INDEX} & \boxed{1}\end{bmatrix} \\ \text{C-CONT} & \begin{bmatrix} \text{RELS} & \left\langle \begin{bmatrix}\text{RELN} & \textit{exist\_q} \\ \text{ARG0} & \boxed{1}\end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix}$$

The second (13) 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 $+$.

(13)
$$\begin{bmatrix} \textit{bare-cl-np-phrase} & \\ \text{SYN} & \begin{bmatrix} \text{VAL.SPR} & \langle\rangle \\ \text{SEM.INDEX} & \boxed{1}\begin{bmatrix}\text{COG-ST} & \textit{fam-or-less}\end{bmatrix}\end{bmatrix} \\ \text{HD} & \begin{bmatrix} \text{HEAD} & \textit{noun} \\ \text{VAL.SPR} & \left\langle\begin{bmatrix}\text{HEAD} & \textit{det}\begin{bmatrix}\text{CTD} & -\end{bmatrix}\end{bmatrix}\right\rangle \\ \text{SEM.INDEX} & \boxed{1}\end{bmatrix} \\ \text{C-CONT} & \begin{bmatrix}\text{RELS} & \left\langle\begin{bmatrix}\text{RELN} & \textit{exist\_q} \\ \text{ARG0} & \boxed{1}\end{bmatrix}\right\rangle\end{bmatrix}\end{bmatrix}$$

In the implemented grammar, both of these rules inherit from a single supertype *bare-np-super* which contains the shared structure.

3.2.4 Bare Classifier Rule

This non-branching rule (14) takes a classifier, and creates a classifier phrase. As the interpretation is always that there is one thing being classified, 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 & Song (2015: 189), 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.

(14)
$$\begin{bmatrix} \textit{bare-cl-phrase} & \\ \text{SYN} & \begin{bmatrix} \text{VAL} & \begin{bmatrix}\text{SPR} & \langle\rangle, \\ \text{SPEC} & \boxed{3}\left\langle\begin{bmatrix}\text{INDEX} & \boxed{0}\end{bmatrix}\right\rangle\end{bmatrix} \\ \text{SEM} & \begin{bmatrix}\text{INDEX} & \boxed{0}, \\ \text{XARG} & \boxed{1}\end{bmatrix}\end{bmatrix} \\ \text{HD} & \begin{bmatrix} \text{HEAD} & \textit{cls}\begin{bmatrix}\text{CTD} & -\end{bmatrix} \\ \text{VAL} & \begin{bmatrix}\text{SPR} & \langle\text{UNEXPRESSED}\rangle \\ \text{SPEC} & \boxed{3}\end{bmatrix} \\ \text{SEM} & \begin{bmatrix}\text{INDEX} & \boxed{2} \\ \text{XARG} & \boxed{1}\end{bmatrix}\end{bmatrix} \\ \text{C-CONT} & \begin{bmatrix}\text{RELS} & \left\langle\begin{bmatrix}\text{RELN} & \textit{card-relation} \\ \text{ARG1} & \boxed{2} \\ \text{CARG} & \textit{1}\end{bmatrix}\right\rangle\end{bmatrix}\end{bmatrix}$$

3.3 Semantics

The semantics for the 5 types are given in Table 3 (using indexed MRS: Copestake et al. 2005). We use the jyutping transliteration for the predicate names, in the actual grammar they are written with Chinese characters. In all cases save the bare noun, there are four predicates: a quantifier (either from the determiner or the Bare Noun

Type	Indexed MRS	cog-st of x_1
D-X-C-N	nei1_q(x_1, h_2, h_3); card(e_4, x_1, 'x'), go3_x(e_5, x_1); ping4gwo2_n(x_1)	*fam-or-more*
D-C-N	nei1_q(x_1, h_2, h_3); card(e_4, x_1, '1'), go3_x(e_5, x_1); ping4gwo2_n(x_1)	*fam-or-more*
X-C-N	exist_q(x_1, h_2, h_3); card(e_4, x_1, 'x'), go3_x(e_5, x_1); ping4gwo2_n(x_1)	*type-id*
C-N	exist_q(x_1, h_2, h_3); card(e_4, x_1 '1'), go3_x(e_5, x_1), ping4gwo2_n(x_1)	*fam-or-less*
N	exist_q(x_1, h_2, h_3), ping4gwo2_n(x_1)	*type-id*

Table 3: MRS for the various combinations

Phrase rule), the head noun, the classifier (which takes the head noun as its external argument (ARG1) and the cardinality relation (which has two arguments, the classifier as its external argument and the amount as a value). The semantics expresses the situation where the noun is being measured out in units of the classifier, to the amount of the number. If the cardinality is not explicitly given, then the default value of one comes from the Bare Classifier Rule. For the bare noun, there is no measurement, so the noun appears with just the default determiner.

This is compatible with the analysis of Takao (2005) for Japanese, where he combines the cardinality and classifer into a single *measure* relation, as in (15). We choose to encode it as two different relations to retain compatibility with the analyses of other languages in the DELPH-IN framework (Uszkoreit 2002). However, it would be trivial to transform one to the other.

(15)
$$\begin{bmatrix} \text{RELN} & \textit{measure} \\ \text{ARG1} & x_1 \\ \text{NUM} & \textit{1} \\ \text{DIMENSION} & \text{個}\_x \ (\text{IN OUR CASE GO3\_X}) \end{bmatrix}$$

It differs from the analysis of Bender & Siegel (2004) who co-index the classifier and noun, and have no representation for the classifier. This models the intuition that they are not separate referents, but loses the opportunity to represent the classifier semantics. As the choice of classifier has some effect on the interpretation of the meaning, this is undesirable. We model this intuition by making the index (ARG0) of the classifier (and cardinality) relation non-referential.

Kim & Yang (2007) co-index the classifier and noun, and have a representation for the classifier. This means that two referential predicates share an ARG0. We avoid this to retain compatibility with the characteristic variable property of Dependency MRS (Copestake 2009). Again this could be converted easily as our analysis captures the same intuition (that we need two predicates and that they have only one quantifier).

Our grammar is bi-directional, it can parse from a string to the semantics or from the semantics to a string (or strings). In the fully specified semantics, the classifier is given. In many applications (such as machine translation), it would need to be generated, which could be done with a generative language model or an ontology such as wordnet (Mok et al. 2012, Morgado da Costa et al. 2016).

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 & 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 in more depth to those of other classifier languages, especially those with computational analyses like Indonesian (Moeljadi et al. 2015), Japanese (Siegel et al. 2016), Korean (Kim et al. 2011) and Mandarin (Müller & Lipenkova 2013, Fan et al. 2015).

References

Bender, Emily M. & 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*, 17–29.

Bender, Emily M. & Melanie Siegel. 2004. Implementing the syntax of Japanese numeral classifiers. In *Proceedings of the IJC-NLP-2004*.

Birner, Betty J. 2021. *Pragmatics: a slim guide*. Oxford University Press, USA.

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

Chen, Ping. 2004. Identifiability and definiteness in Chinese. *Linguistics* 42(6). 1129–1184. DOI: 10.1515/ling.2004.42.6.1129.

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

Copestake, Ann. 2009. Slacker semantics: why superficiality, dependency and avoidance of commitment can be the right way to go. In *Proceedings of the 12th conference of the european chapter of the ACL (EACL 2009)*, 1–9. Athens.

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

Fan, Zhenzhen, Sanghoun Song & Francis Bond. 2015. Building Zhong [|], a Chinese HPSG meta-grammar. In *Proceedings of the 22nd international conference on Head-Driven Phrase Structure Grammar (HPSG 2015)*, 97–110. DOI: 10.21248/hpsg.2015.6.

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

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

Kim, Jong-Bok & Jaehyung Yang. 2007. Syntax and semantics of Korean numeral classifier constructions. In Stefan Müller (ed.), *Proceedings of the 14th International Conference on Head-Driven Phrase Structure Grammar, Stanford Department of Linguistics and CSLI's LinGO Lab*, 163–172. Stanford, CA: CSLI Publications. DOI: 10.21248/hpsg.2007.10.

Kim, Jong-Bok, Jaehyung Yang, Sanghoun Song & Francis Bond. 2011. Processing of Korean and the development of the Korean resource grammar. *Linguistic Research* 28(3). 635–672.

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

Li, Yen-hui Audrey. 1998. Argument determiner phrases and number phrases. *Linguistic Inquiry* 29(4). 693–702.

Moeljadi, David, Francis Bond & Sanghoun Song. 2015. Building an HPSG-based Indonesian resource grammar (INDRA). In *ACL 2015 workshop on grammar engineering across frameworks (GEAF 2015)*.

Mok, Hazel Shuwen, Eshley Huini Gao & Francis Bond. 2012. Generating numeral classifiers in Chinese and Japanese. In *Proceedings of the 6th global wordnet conference (GWC 2012)*. 211-218. Matsue.

Morgado da Costa, Luís, Francis Bond & Helena Gao. 2016. Mapping and generating classifiers using an open Chinese ontology. In *Proceedings of the 8th Global Wordnet Conference (GWC 2016)*, 248–254.

Müller, Stefan & Janna Lipenkova. 2013. Chingram: a TRALE implementation of an HPSG fragment of Mandarin Chinese. *Sponsors: National Science Council, Executive Yuan, ROC Institute of Linguistics, Academia Sinica NCCU Office of Research and Development*. 240.

Ng, Say Kiat. 1997. *A double specifier account of Chinese NPs using head-driven phrase structure grammar*. University of Edinburgh: Department of Linguistics. (MSc Thesis).

Prince, Ellen F. 1981. *Toward a taxonomy of given-new information*.

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

Siegel, Melanie, Emily Bender & Francis Bond. 2016. *Jacy — an implemented grammar of Japanese*. CSLI Publications.

Sio, Joanna Ut-Seong & Sanghoun Song. 2015. Divergence in expressing definiteness between Mandarin and Cantonese. In Stefan Müller (ed.), *Proceedings of* *the 22nd International Conference on Head-Driven Phrase Structure Grammar, Nanyang Technological University (NTU), Singapore*, 177–194. Stanford, CA: CSLI Publications. DOI: 10.21248/hpsg.2015.11.

Takao, GUNJI. 2005. Measurement and quantification revisited. 8(8). 21–36.

Uszkoreit, Hans. 2002. New chances for deep linguistic processing. In *19th international conference on computational linguistics: COLING-2002*, XIV–XXVII. Taipei.

Wang, Lulu & Haitao Liu. 2007. A description of Chinese NPs using Head-Driven Phrase Structure Grammar. In Stefan Müller (ed.), *Proceedings of the 14th International Conference on Head-Driven Phrase Structure Grammar, Stanford Department of Linguistics and CSLI's LinGO Lab*, 287–305. Stanford, CA: CSLI Publications. DOI: 10.21248/hpsg.2007.17.

Wikipedia contributors. 2024. *Cantonese — Wikipedia, the free encyclopedia*. [Online; accessed 24-March-2024]. https://en.wikipedia.org/w/index.php?title=Cantonese&oldid=1214658644.