An empirical, corpus-based, approach to Cantonese nominal expressions

Grégoire Winterstein and **David Vergnaud** and **Jérémie Lupien** and **Samuel Laperle**
Université du Québec à Montréal
Département de linguistique
winterstein.gregoire@uqam.ca,
{vergnaud.david,lupien.jeremie,laperle.samuel}@courrier.uqam.ca

Hannah Yu
The University of Hong Kong
hannahyu@connect.hku.hk

Christopher Davis
University of the Ryukyus
cmdavis@grs.u-ryukyu.ac.jp

Zoe Pei Sui Luk
The Education U. of Hong Kong
psluk@eduhk.hk

Abstract

This work focuses on the distribution and interpretation of Hong Kong Cantonese nominal expressions with the help of empirical, corpus-based, methods. We present the creation of a dataset containing all nominal expressions appearing in a corpus of naturally occurring Hong Kong conversations, along with annotations of inherent and contextual features of these expressions. We then compare the observed distribution of the different types of expressions with predictions made by existing theoretical analyses, and conclude that these predictions are largely supported by our data. We then focus on environments in which theory predicts that several types of expression are licensed, and thus in potential competition. For those cases, we find that inherent properties of the nominal head, in particular its length, play a significant role in determining which type of expression is used. We hypothesize that this constraint is related to the frequency of the head noun and to general linguistic principles that relate markedness and frequency in natural language production and interpretation.

1 Introduction

The inventory of Cantonese nominal expressions (NE) is rather large, and distinguishes itself in several respects from that of other Chinese languages.[1] Though NEs have been the object of various theoretical works, there is (to our knowledge) little to no work which approaches the question of the interpretation of Cantonese NEs from an empirical, corpus-based, point of view.

Here, we focus on Cantonese NEs which can be used to refer to indefinite referents, i.e. to elements that have not been previously introduced in the discourse and which are not common ground between the discourse participants. B's answer in example (1) shows three different types of NEs that can be used to refer to an indefinite referent in the context of A's utterance:

1. a **bare noun** ([BARE N]), e.g. *sin3* in (1)
2. a **bare classifier phrase** ([CL N]), e.g. *baa2 sin3* in (1)
3. a **numeral phrase** ([NUM CL N]), e.g. *jat1 baa2 sin3* in (1)

(1) A: It's so hot, I wish I had something to cool my face.
B: ngo5 jau5 daai3 ((**jat1**) **baa2**) **sin3** aa3
1SG have bring ((JAT1) CL) fan SFP
I brought a/some fan(s).

Each type of NE differs from the others in its overall distribution and compatibility with referents. For example, bare nouns cannot refer to specific indefinites, bare classifiers are inherently singular and can also refer to definite referents, unlike numeral phrases, and the three types of NE exhibit different behaviors in sentences with other scope-taking elements (e.g. negation, quantifiers, conditionals, see Davis et al. 2023). Nevertheless, all three are acceptable options in simple, non-embedded environments like the one in (1), where they appear to be largely synonymous.

This raises our main research question in this work, namely that of establishing which factors, if any, contribute to the choice of a particular type of NE in a given sentence by a given speaker. To address this question we rely on attested examples of Cantonese NEs, based on a corpus of everyday speech, manually annotated with various information about each NE.

[1] We use the term *Nominal Expression* to refer to phrases that have the distribution of a noun phrase/determiner phrase, aiming for a label that is theoretically neutral regarding the exact syntactic status of these phrases.

In section 2, we begin by giving an overview of the landscape of Cantonese NEs and their analysis, followed by a survey of other comparable phenomena which have been investigated via statistical, empirically driven, means. In section 3, we present the data and its annotation, and turn in section 4 to the analysis of the results and their interpretation. Section 5 concludes.

2 Related work

2.1 Theoretical analyses of Cantonese NEs

As a central element of Cantonese, Cantonese NEs have been extensively described (e.g. by Matthews and Yip 2011) and there is a significant amount of work that offers syntactic and semantic analyses of these elements.

2.1.1 Overview

Out of the three constructions exemplified in (1), the bare classifier one is striking in that it is very frequent in Cantonese, unlike in Mandarin, and that its interpretation appears to be the most flexible, ranging over both indefinite and definite interpretations. Most works dealing with this construction note the indefinite interpretation, but usually focus on the definite one, trying to establish whether the classifier acts as a (definite) determiner in those uses (Cheng and Sybesma, 1999; Wu and Bodomo, 2009; Cheng and Sybesma, 2008; Jenks, 2018). Davis et al. (2023) propose a unified analysis of the semantics of [CL N], based on its indefinite use, which is restricted to specific and singular referents. Contexts where these referents are contextually given give rise to definite *interpretations*, without the need for positing a distinct definite *reading*, i.e. without the need for positing a syntactic or semantic ambiguity for bare classifier phrases.

The interpretation of [BARE N] is also flexible to an extent, usually covering generics, kinds and non-specific definites, but also extending to "globally unique" definites (Hawkins, 1978), i.e. referents that are thought to be unique even outside of the confines of the conversation (such as astral objects like the Moon and the Sun). Unlike [CL N] constructions, [BARE N] are not restricted for number, and thus in the bare noun version of (1), the speaker could have brought more than one fan, unlike what would be the case in the bare classifier construction.

[NUM CL N] phrases are, unsurprisingly, restricted for number. If the numeral refers to a quantity higher than one, then the NE is plural. When the numeral is *jat1* ('one'), and the classifier is the sortal classifier associated with the noun, or a measure classifier, then the NE is singular. Another salient property of [NUM CL N] phrases is that their referent is necessarily indefinite: these phrases cannot be used to refer to anaphoric or unique definite entities.

2.1.2 The classifier *di1*

Beyond classic sortal and measure classifiers, Cantonese also has a so-called "plural" classifier (*di1*). When used in [CL N] and [NUM CL N] constructions, as in (2), the resulting NE will be interpreted as plural.

(2) ngo5 jau5 daai3 (jat1) di1 sin3 aa3
1SG have bring (JAT1) DI1 fan SFP
I brought some fans.

As shown in (2), we can see that *di1* is compatible with the numeral *jat1* ('one'), and yields a plural interpretation. Davis et al. (2023) analyze this data by attributing the singular interpretation of (non-*di1*) [CL N] constructions to the semantics of the sortal and measure classifiers, and by treating *jat1* as an indefinite marker, unmarked for number in cases like (2).

2.1.3 Position relative to the main verb

Another relevant property of NEs is that appearing in the preverbal domain restricts them to a definite interpretation. This is usually explained by the topic-prominent nature of Chinese languages (Li and Thompson, 1976), which organize constituents around the verb in terms of their informational status. Thus, though some indefinites do appear in the preverbal domain in Cantonese, they are restricted to generics, or to phrases explicitly introduced via an existential operator like *jau5* as in (3).

(3) *(jau5) **saam1 go3 jan4** lei4-zo2
(have) three CL person come-PFV
Three people came.

2.1.4 Summary

Table 1 summarizes the relevant properties of the three Cantonese NEs we have introduced, based on their theoretical descriptions. We indicate in bold the case of special interest in this work.

2.2 Data-driven approaches to grammaticality

In natural language, it is common to find cases in which different constructions are available to convey what appears to be similar meanings, or in

	Num.	*pre-verbal*	*post-verbal*
definite	SG	(CL) N	(CL) N
	PL	DI1 N	DI1 N
indefinite	SG	∃constr.	**((NUM) CL) N**
	PL	∃constr.	((JAT1) DI1) N

Table 1: Predicted NE types according to the (in)definite nature of the referent, number and position of the NE relative to the verb (definite bare nouns have to be globally unique, indefinites are understood as referential).

which only one construction is possible, but it is not clear why other options are ruled out. The case of the Cantonese NE illustrated in (1) and highlighted in table 1 is but one of many such examples.

Pioneering work by Bresnan (2007); Bresnan et al. (2007) approaches this question by treating linguistic knowledge, and in particular production, as probabilistic. The probability of using a given construction is seen as conditioned by a variety of factors, which might interact with each other. Bresnan and colleagues focused on the case of order of the complements of ditransitive verbs (i.e. "dative alternation") and found that various features of the complements of verbs (i.e. them being thematic, their discourse-givenness, but also their length etc.) played a significant role in predicting which construction would be used. This prediction was assessed by fitting and comparing mixed logistic regression models to data extracted from naturally occuring sentences that involve ditransitive verbs.

The same method was later used to address questions such as the order of adjectives in French (Thuilier et al., 2010) or the alternation between active and passive voice (Da Cunha and Abeillé, 2022).

As emphasized by Bresnan (2007), a picture like the one in table 1 might severely underestimate the space of grammatical possibilities for Cantonese NEs, and also does not help to understand why certain constructions appear more often than others when several seem to be in competition. Adopting a probabilistic stance can thus help shed light on complex issues related to the use of certain expressions such as Cantonese NEs.

3 Data

To address the question of the choice and use of Cantonese NEs, we extracted all NEs from a Cantonese corpus and annotated them with features relevant to the dimensions introduced in the previous section. The complete extraction, along with the part that was annotated, is available on the following anonymous OSF repository in the form of a `CSV` file: `https://osf.io/wncj9/?view_only=673e8af11bba4ab6b8559ffe29e5d8ac`. The same repository also contains the `R` scripts used for the statistical analysis (see section 4).

3.1 Corpus and extraction

To extract the data, we relied on the Hong Kong Cantonese Corpus (HKCanCor, Luke and Wong 2015), which we accessed through the `PyCantonese` library (Lee et al., 2022). Our choice was motivated both by the nature of the data (unscripted, daily conversations), and by the fact that the corpus is annotated with part of speech information, which greatly facilitates the extraction of NEs in it.

The algorithm for automatically extracting NEs is based on the description of the structure of Cantonese NEs given by Matthews and Yip (2011). It basically assumes that NEs are always head final, and that an NE's constituents fall into specific categories. We thus looked for any sequence of elements which matched the following pattern:

(4) (Demonstrative) ((Numeral) (Modifiers) Classifier) (Modifiers) N

Note that this pattern ensures that every noun in the corpus is extracted, along with the maximal structure of the NE headed by the noun in question. Though that structure served for the automatic extraction, it was manually checked and revised by an annotator, allowing for some deviation with the pattern in (4) and for correcting erroneous structures.

Several decisions were made to align the PoS labels used in HKCanCor with the ones needed for the extraction. We thus excluded nouns that were not in Cantonese (i.e. code switched) along with proper nouns. We also manually detected demonstratives, which were not tagged as a specific category. We extracted NEs with adjectival modifiers, but left out those with relative phrases as they would require a more complex syntactic analysis to be properly identified.

NEs that were contained within other NEs (e.g. in compounds or genitive constructions) were extracted both as part of their embedding phrase and as standalone NEs, e.g. both *saang1jat6* and *jat1 fan6 hou2 hou2 ge3 saang1jat6 lai5mat6* are

entries in the dataset for the expression in (5).[2]

(5) jat1 fan6 hou2 hou2 ge3 saang1jat6 lai5mat6
one CL very good GEN birthday present
a very nice birthday present

In total 10 979 NEs were extracted from the corpus. To facilitate their annotation and allow future statistical analysis, additional data were extracted for each NE occurrence. Those were: the file from which the NE was extracted (which serves as the identifier for conversations), the speaker's ID, their age and gender, the whole sentence in which the NE appears, and the two preceding conversational turns to give an approximation of the context in which the NE appeared. In addition to this information, some features were automatically pre-annotated before being manually checked. Those are presented in the next section along with all other annotated features.

3.2 Annotated features

In line with the works discussed in section 2, we annotated the extracted NEs with two types of information: (i) features that pertain to lexical properties of the head noun of the NE, and (ii) features that are not idiosyncratic to the noun, such as its informational status and its position relative to the main verb. The list of annotated features along with their possible values is given in table 2. Features indicated with a * are those that were automatically pre-annotated and later manually checked. Entries in the dataset were annotated by one trained native Cantonese speaker annotator, and checked by two members of the team, including another Cantonese native speaker.

Note that some of the NEs initially extracted were removed from our data, and that we extracted a subset of our data containing odd combinations of annotations and further revised and updated them as required. Not all extracted NEs were annotated for logistical reasons. In the end, a total of 4 469 NEs were annotated with the features in table 2, including the contextual, structural and semantic

Name	Values
`Type*`	the type of NE ([BARE N], [CL N], etc., see sec. 3.3 for details)
`Head noun*`	the head noun of the NE
`Classifier`	the classifier used for the head noun (if any)
`Numeral`	the numeral used in the NE (if any)
`pos.V`	position relative to the verb (BEFORE / AFTER / ∅ for verbless sentences and parenthetical expressions)
`Info. Status`	the informational status of the referent (OLD / NEW / GENERIC / BRIDGE)[3]
`Global.Unique`	whether the referent of the NE is globally unique (YES/NO)
`Bridge.Ant`	the antecedent of a bridged NE
`Genitive`	whether the phrase is used in a genitive construction (either via a classifier or the particle *ge3*)
`Sent.type`	the syntactic type of the host sentence of the NE (DECLARATIVE / INTERROGATIVE / IMPERATIVE)
`Embed.`	whether the NE is embedded in a complex construction like negation, antecedent of a conditional or other potential scope island operators (YES/NO)
`Length*`	the length of the NE in number of characters (one character being equivalent to a syllable)
`LengthHead*`	the length of the head of the NE in number of characters
`Abstractness`	the ABSTRACT/CONCRETE nature of the head noun
`Animacy`	the animacy of concrete referents (ANIMATE / INANIMATE)
`Mass`	whether the head noun is a mass noun or not (MASS / COUNT)[4]

Table 2: Annotated features

[2] Example (5) is also an example in which the results of the automated script had to be manually revised: the candidate entries were initially *jat1 fan6 hou2 hou2 ge3 saang1jat6* (which is incorrect) and *jat1 fan6 hou2 hou2 ge3 saang1jat6 lai5mat6* (which is correct).

[3] We consider that generic NEs are not used referentially. "Bridging" refers to definite NEs which are related to a previously mentioned referent (the "antecedent") by a part/whole or producer/product relation.

[4] Though part of the dataset, we opted not to analyze the `Mass` feature, since the boundaries between these types of nouns are less clear in Cantonese, and that annotators disagreed on several cases. As it stands, we only typed as MASS elements which lacked stable discrete boundaries, and were clearly cumulative (Deal, 2017).

information noted above, as well as unique IDs and additional notes for the organization of the data.

3.3 Overview of the NE types

Table 3 gives an overview of the major types of constructions found in the data.

Type	Frequency
[N]	2591
[CL N]	639
[DEM CL N]	518
[GE N]	183
[NUM CL N]	168
[DEM N]	86
[QUANT CL N]	80
[QUANT N]	72
Other	132
Total	4469

Table 3: Distribution of the annotated NEs by Type

In table 3, DEM refers to demonstratives, GE to the genitive particle *ge3*, and QUANT to quantifiers of different sorts, including classifier reduplication (Lee, 2020). The *Other* category groups rare occurrences of combination of those elements, such as demonstrative and quantified expressions etc.

Overall, all the annotated data conform to the expected pattern for Cantonese NE, with minor exceptions. The most flagrant exception to usual descriptions is that we found some cases of [DEM N], without the use of a classifier. Examples include the following: *ni1 hong4* ('this industry'), *go2 baan2* ('that version'), *ni1 fong1min6* ('this aspect'). Those were confirmed by several native speakers as being natural without a classifier, and degraded with it. Those cases however seem to be very idiosyncratic, for example the noun *hong4jip6*, which also means industry, requires a classifier with a demonstrative, unlike its one syllable version. This could point to a rhythmic constraint, in line with some results we present below (section 4.2), but cannot account for all the outliers, e.g. *fong1min6* which is disyllabic and disallows a classifier with a demonstrative.

4 Data analysis

In this section, we focus on two types of constructions in the data: the [CL N] cases and the [NUM CL N] ones. As mentioned in section 2.1, these constructions largely overlap in terms of meaning and distribution: their number features are identical (singular for sortal and measure classifiers, plural with the classifier *di1*), and [CL N] phrases are compatible with indefinite referents in the same way as [NUM CL N] ones. Their main difference is that [CL N] phrases are compatible with definite referents, unlike [NUM CL N] phrases.

The analyses in this section are thus restricted to a particular subset of our annotated data. Concretely, the observations in this subset have the following properties:

- their `Type` feature is either [CL N] or [NUM CL N]
- they are not embedded under negation, in the antecedent of conditionals or other potential scope islands, or in the *jau5* constructions (feature `Embed.` set to NON.EMBEDDED and feature JAU5.CONTEXT to empty)
- [CL N] phrases are not used with generic-like readings (e.g. *di1 jan5* 'people'), or in bridging constructions (feature `Info.Status` set to OLD or NEW)

These filters yielded a total of 551 observations for the statistical analyses.

In section 4.1, we check that our data reflects the theoretically motivated differences between these constructions in terms of placement relative to the verb and the informational status of the referent. In section 4.2, we then focus on post-verbal indefinite environments in which the two constructions are predicted to be synonymous and in free variation.

4.1 Theoretical predictions

The effects of the informational status (`Info.Status`) and position of the NE relative to the verb (`pos.V`) on the type of NE (`Type`) were assessed in two complementary ways.

To begin, we ran a series χ^2 tests in order to test the independence of the three variables under study and examined the Pearson residuals of a fit of the contingency tables (`vcd` package for `R` Zeileis et al. 2007).

We found a significant effect ($\chi^2 = 26.703, p < 0.0001$) between `Type` and `Info.Status`, with [NUM CL N] phrases strongly associated with new referents rather than old ones (cf. Fig. 1).

Similarly, the informational status and position relative to the verb were equally significantly related ($\chi^2 = 39.886, p < 0.0001$). Figure 2 shows

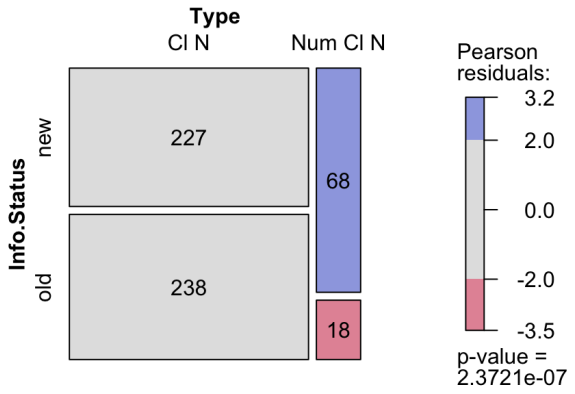

Figure 1: Distribution of the observations by `Type`/`Info.Status` and evaluation of the Pearson residuals

the residuals of a fit of the contingency table between `Info.Status` and a binary synthetic variable `before.V` which collapses cases in which the NE appears either after the verb or in a verbless clause.

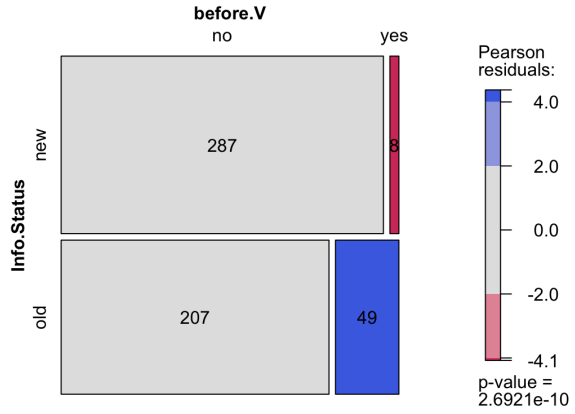

Figure 2: Distribution of the observations by `before.V`/`Info.Status` and evaluation of the Pearson residuals

The tests for `Type` and `before.V` did not find any significant relation between the two variables ($\chi^2 = 2.2556, p = 0.12$)

To further investigate these variables, we fitted several generalized linear mixed models for logistic regression with `Info.Status` and `pos.V` as predictors of the `Type` of the NE, and using the conversation ID as a random factor. We then used model comparison to assess the significance of the two independent factors for predicting `type`. In line with the χ^2 analysis, we found a significant effect of `Info.Status` ($\chi^2 = 25.96, p < 0.001$), and no effect of `pos.V` ($\chi^2 = 1.47, p = 0.48$). In addition, we found a significant interaction between `Info.Status` and `pos.V` ($\chi^2 = 11.934, p < 0.001$).

Overall the results confirm the theoretical picture described in section 2.1: OLD referents (i.e. anaphoric definites) are mostly expressed via [CL N] phrases, and NEW ones are mostly found in the post-verbal domain. An examination of the few NEW referents found pre-verbally showed that most of these are best understood as accommodated definite referents rather than newly introduced indefinite referents (e.g. a speaker referring to the tour guide of a trip in the general context of discussing travel, but with no previous mention of the guide or a tour, making it hard to analyze it as a case of bridging reference).

The fact that `Type` and `before.V` appear independent might be surprising since one would expect not to find [NUM CL N] in the preverbal domain. This absence of a significant result can be attributed to a number of factors. First, there are relatively few preverbal cases in general, compared to other cases, and in the postverbal domain we do not expect to find significant differences between the types (and indeed we find none). There might thus simply be a problem of statistical power, and we might be able to find a significant effect with additional observations (e.g. after the dataset has been completely annotated). Second, the more elaborate analyses using GLMM showed that `before.V` has a significant interaction with `Info.Status` for the prediction of `Type`, further suggesting that the sole knowledge of the position relative to the verb is not enough to predict the type of an NE, but that it brings significant information when combined with the informational status of the NE.

Finally, we note that the picture is nevertheless not as perfect as the theory would predict. We do observe some cases of new, non definite, referents introduced by [NUM CL N] phrases, including in the pre-verbal domain. We will not present an analysis of these few cases, and will remain neutral regarding their status as production errors, or counter-examples to the theory.

4.2 Indefinites in free variation

We now turn to a subset of our data: NEs that appear in environments in which theory predicts that both [NUM CL N] and [CL N] phrases are licensed. Concretely, this means that we only consider NEs which belong to the dataset described earlier in this section and have the additional properties:

- their `Info.Status` is OLD (i.e. indefinite) and they are not found in the preverbal domain (feature `pos.V` different from BEFORE)
- if they are [NUM CL N] phrases, their numeral is *jat1* (which makes them freely alternate with [CL N] phrases)
- if they are [CL N] phrases, they are not involved in a genitive construction (feature `Genitive` set to NO).

Those additional filters yielded 279 observations for the analysis.

We further restricted the data by excluding NEs that use the *di1* classifier. As shown in figure 3, though *jat1 di1 N* phrases do exist in the data, they are rare (only 4 observations), making the use of *di1* a strong predictor of the [CL N] `Type` for NEs that are not singular.[5]

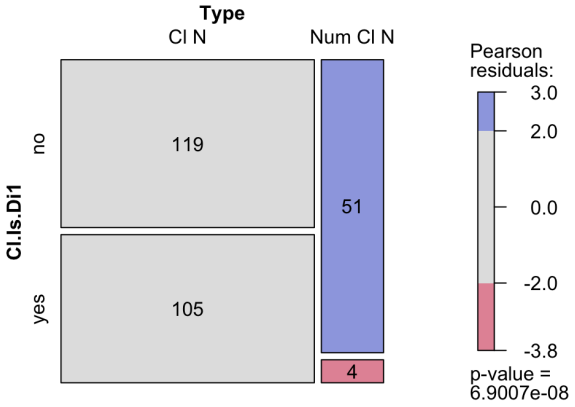

Figure 3: Distribution of the observations by `Type` and nature of the classifier (*di1* or not)

Thus, after restricting our dataset to singular NEs (feature `Is.Cl.Di1` set to NO) we had 170 observations to analyze.

We used the same method as in section 4.1 for our analyses. The overall goal was the same: we assessed the significance of different features for the prediction of the `Type` of the NEs in the dataset. We relied both on the analysis of Pearson residuals, and on comparing model generalized linear mixed models fitted for logistic regression.

Specifically, we looked at features inherent to the nominal head: its length, measured in number of characters (`LengthHead`), and whether the head noun denotes (i) a concrete or abstract individual (`Abstractness`), and (ii) an animate individual (`Animacy`). We also assessed the significance of the `Classifier` as a random factor in our models, as well as that of the head noun itself, also by using it as a random factor in the models.

The most significant effect we found was the length of the head noun ($\chi^2 = 14.96, p < 0.001$, via model comparison). As shown on figure 4, longer head nouns tend to be used more frequently as [NUM CL N] phrases, and significantly less as [CL N] phrases, which favor monosyllabic heads.

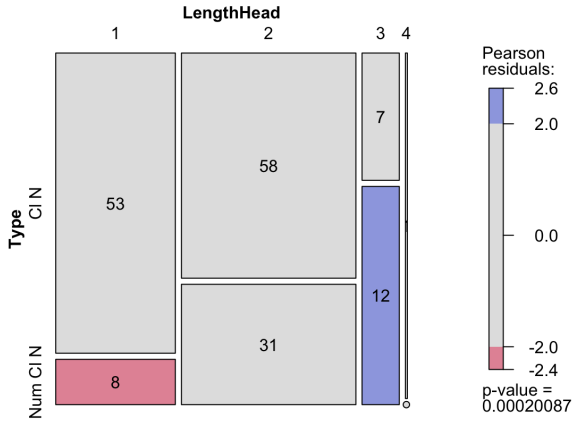

Figure 4: Distribution of the observations by `Type` and length of the head noun of the NE

Overall, there were 145 different noun types in the dataset (for 170 tokens), which means that very few of them were observed as multiple tokens. Statistical analysis shows that the identity of the head noun itself does not play a significant role in the prediction of `Type` ($\chi^2 = 1.042, p = 0.307$). This is consistent with the hypothesis that it is indeed the length of the noun that plays a role in our data, rather than noun-specific idiosyncrasies. In line with that idea, we found that the identity of the classifier used in the NE, when treated as a random factor in our regressions, has a marginally significant effect ($\chi^2 = 1.042, p = 0.078$). This is additional evidence that some properties of the nominal head, reflected in the semantics of the classifier, might be relevant in the preference for a structure over another.

We now turn to possible explanations for the data shown in figure 4 and the role played by the length of the head noun in the selection of a construction. A first hypothesis is that Cantonese has a rhythmic constraint favoring phrases with an even number of syllables. This would be consistent with the significant preference for [CL N] types with monosyllabic heads, and the comparatively higher proportion of

[5] As of now, we have no clear explanation for this finding, and will investigate it in future work.

disyllabic heads in [NUM CL N] phrases. However, figure 4 suggests that the relationship at hand is that the longer the head is, the more chances there are of using a [NUM CL N] phrase: crucially, with nouns that contain 3 syllables there are significantly more cases of [NUM CL N] phrases than [CL N] ones, which would not be expected if we assumed a preference for an even number of syllables.

Rather than a rhythmic constraint, we hypothesize that the results observed stem from universal linguistic constraints relating frequency and the length of linguistic expressions. Since the early work of Zipf (1935), there have been numerous observations that there seems to be a negative correlation between the length and frequency of words across a wide variety of language (Bentz and Ferrer-i Cancho, 2016; Kanwal et al., 2017). So, in the case of Cantonese as well, we expect that, on average, shorter words will be more frequent, and longer words will be less frequent. Concretely, in terms of character/syllable length, this means that on average the frequency of one-character words is higher than the frequency of two-characters words, which is higher than the frequency of three-characters words. Another well-attested principle of language use, articulated by Levinson (2000) in the guise of his M-principle (traceable at least to Jakobson's markedness theory), is that more surprising material tends to be more formally marked than less surprising material. For example, a phrase like *cause to die* will be understood as denoting indirect, non conventional, ways to bring about the death of someone, unlike the more prototypical verb *kill*.

Putting these two principles together leads us to expect that longer words will tend to be more formally marked. In other words: longer words are on average less frequent, and hence on average more surprising or less expected when they occur. The opposite holds for shorter words. We expect then that, on average, shorter words will be less formally marked than longer ones. The use of a numeral is a kind of formal marking since it increases the complexity of the phrase. Hence, all else being equal, we expect that the numeral will be used more frequently with longer words. Finally, we motivate the pressure to mark indefinite NEs with the observation that, in Cantonese, indefinites, rather than definites, appear to be semantically marked (unlike in a language like English). This is especially salient in the cases at hand here: while [CL N] cases are underspecified with regards to the (in)definite status of their referent, [NUM CL N] phrases can only refer to indefinite referents. One could thus justify using a [NUM CL N] phrase by the desire to formally block a definite interpretation of the referent. Under the hypotheses at work here, we would assume that definite interpretations would thus be more salient for less frequent, longer nouns.

5 Conclusion

We presented how we created and annotated a dataset for the study of the use and production of nominal expressions in Cantonese. The results of our statistical analyses are two-fold. On one hand, we were able to confirm that theoretical descriptions of the interpretation and distribution of two major types of Cantonese NEs ([CL N] and [NUM CL N] phrases) were largely correct in terms of information structure and syntactic features (with some potentially interesting outliers).

On the other hand, for contexts in which [CL N] and [NUM CL N] types are both licensed, the statistical analysis suggests that properties of the head have an effect in determining which form is used, in particular the length of the head noun. We proposed to relate this constraint to general linguistic principles about markedness and frequency.

Many issues still remain open. In some cases, we surmised that adding statistical power might result in finding additional significant effects in our data. This basically calls for the annotation of more data, a project that is currently ongoing with the annotation of all the extracted NEs in the dataset.

Another avenue of research is to look at bare noun phrases, since those NEs are also compatible with both definite and indefinite referents (though in restricted ways), and can, in contexts like (1) alternate with [CL N] and [NUM CL N] phrases.

We also plan to analyze genitive cases, which offer yet another example of an environment in which two alternatives seem to coexist, without any clear factor determining which version will most likely be used. Specifically, we will look at the alternance between genitive uses of [CL N] phrases and those that rely on the particle *ge3*.

Finally, we plan to further investigate our hypothesis about the effect of the frequency of the head noun (rather than its length) by including information about frequency in our models. For this purpose, we will rely on recent frequency databases

for Cantonese (Lai and Winterstein, 2020; Li et al., 2023) and use them both to verify the relationship between frequency and word length in Cantonese, and the direct effect of frequency on the probability of producing different types of NEs. We also plan to evaluate whether our analysis applies to other constructions that appear sensitive to the number of syllables of linguistic expressions (see a.o. the work of Sio and Sze-Wing 2020 on *aa3* nominals).

References

Christian Bentz and Ramon Ferrer-i Cancho. 2016. Zipf's law of abbreviation as a language universal. In *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*.

Joan Bresnan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston and Wolfgang Sternefeld, editors, *Roots: Linguistics in Search of Its Evidential Base.*, Studies in Generative Grammar, pages 77–96. Mouton de Gruyter, Berlin.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the Dative Alternation. In G. Boume, I. Kraemer, and J. Zwarts, editors, *Cognitive Foundations of Interpretation*, pages 69–94. Royal Netherlands Academy of Science, Amsterdam.

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.

Lisa Lai-shen Cheng and Rint Sybesma. 2008. Classifiers in four varieties of Chinese. In Guglielmo Cinque and Richard Kayne, editors, *The Oxford Handbook of Comparative Syntax*, pages 259–292. Oxford University Press.

Yanis Da Cunha and Anne Abeillé. 2022. Objectifying women? A syntactic bias in French and English corpora. In *Proceedings of LATERAISSE Workshop – LREC2022*, pages 8–16.

Christopher Davis, Zoe Pei-sui Luk, and Grégoire Winterstein. 2023. Quantificational and choice-functional noun phrases in cantonese. In *Proceedings of Sinn und Bedeutung 28 (to appear)*.

Amy Rose Deal. 2017. Countability distinctions and semantic variation. *Natural Language Semantics*, 25(2):125–171.

John A. Hawkins. 1978. *Definiteness and indefiniteness: A study in reference and grammaticality prediction*. Croom Helm, London.

Peter Jenks. 2018. Articulated definiteness without articles. *Linguistic Inquiry*, 49(3):501–536.

Jasmeen Kanwal, Kenny Smith, Jennifer Culbertson, and Simon Kirby. 2017. Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, 165:45–52.

Regine Lai and Grégoire Winterstein. 2020. Cifu: a frequency lexicon of Hong Kong Cantonese. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3062–3070, Marseille, France. European Language Resources Association.

Jackson L. Lee, Litong Chen, Charles Lam, Chaak Ming Lau, and Tsz-Him Tsui. 2022. Pycantonese: Cantonese linguistics and NLP in Python. In *Proceedings of The 13th Language Resources and Evaluation Conference*. European Language Resources Association.

Peppina Po-Lun Lee. 2020. On the semantics of classifier reduplication in Cantonese. *Journal of Linguistics*, 56(4):701–743.

Stephen C. Levinson. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. MIT Press, Cambridge.

Charles N. Li and Sandra A. Thompson. 1976. Subject and topic: A new typology of language. In Charles N. Li, editor, *Subject and topic*, pages 457–489. Academic Press, New York.

Jane S.Y. Li, Heikal Badrulhisham, and John Alderete. 2023. Lexical and sub-lexical frequency effects in cantonese. *Taiwan Journal of Linguistics*, 21(2):45–99.

Kang Kwong Luke and May L.Y. Wong. 2015. The Hong Kong Cantonese Corpus: Design and Uses. *Journal of Chinese Linguistics*, 25:312–333.

Stephen Matthews and Virginia Yip. 2011. *Cantonese A Comprehensive Grammar*, 2nd edition. Routledge, London.

Joanna Ut-Seong Sio and Tang Sze-Wing. 2020. Two types of aa3-nominals in Cantonese. *Language and Linguistics*, 21(1):80–103.

Juliette Thuilier, Gwendoline Fox, and Benoît Crabbé. 2010. Approche quantitative en syntaxe, l'exemple de l'alternance de la position de l'adjectif épithète en français. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles*, pages 71–80, Montréal. ATALA.

Yicheng Wu and Adams Bodomo. 2009. Classifiers $\neq$ Determiners. *Linguistic Inquiry*, 40(3):487–503.

Achim Zeileis, David Meyer, and Kurt Hornik. 2007. Residual-based Shadings for Visualizing (Conditional) Independence. *Journal of Computational and Graphical Statistics*, 16(3):507–525.

George K. Zipf. 1935. *The psycho-biology of language*, volume ix. Houghton Mifflin, Oxford, England.

A Appendix

All supplementary material can be accessed on the following anonymous OSF repository: https://osf.io/wncj9/?view_only=673e8af11bba4ab6b8559ffe29e5d8ac.