1 Introduction

Previous studies have demonstrated that, cross-linguistically, the choice of NP form reflects the accessibility of the NP’s referent (Ariel 2001). Specifically, accessibility theory holds that given entities which are easier to retrieve are more likely to be referred to with NP forms on the high end of the accessibility scale, e.g., pronouns, while those entities which are more difficult to retrieve will be referred to with NP forms on the other end of the scale, such as definite descriptions. Factors proposed to affect the accessibility of an entity include global topicality, processing difficulty, and local syntactic prominence (Jaeger & Wasow 2008, Ariel 2001). Some researchers have suggested that, in addition to being reflective of the accessibility of an entity, NP form itself may also exert influence on the future accessibility of that entity, independent of other factors (Gernsbacher & Schroyer 1989).

The present corpus study examines which discourse factors implicated in accessibility, including previous NP form, may best predict NP form in Vernacular Written Cantonese (VWC), for both indefinite entities first introduced into the discourse and given entities. Cantonese is a particularly relevant language to study in the realm of accessibility because it contains an extensive set of possible NP forms that differ minimally in semantic content and are not strictly associated with a particular semantic status (Gundel et al. 1993:290). In examining the roles of discourse factors in predicting NP form, this study seeks not only to identify which factors may best predict NP form, but also to evaluate whether the relationships between significant predictive factors and NP forms are consistent with the claims of accessibility theory and with results of previous studies. In fact, analysis of the corpus data indicates that, while the relationship between accessibility factors and NP form for new entities is generally consistent with predictions, the picture is quite different for given entities. Entities predicted by global topicality and previous NP form to be higher on the accessibility scale are instead significantly more likely to be associated with NP forms lower on the scale. Thus, while these discourse factors are significant predictors of NP form, they are predictive in a direction that suggests other considerations are overriding the principle of economy underlying the accessibility scale.
2 Background and Previous Research

2.1 Accessibility

The notion of accessibility relates to the cognitive ease of retrieval of particular referents, and how this level of ease affects reference forms. In Mira Ariel’s account of accessibility (Ariel 1990, 2001), she posits that given referents which are more prominent in the discourse will be easier to retrieve, and therefore will be referred to with less linguistic material, assuming a principle of economy in which reference forms use the least amount of linguistic material necessary for retrieval. Ariel places reference forms on an accessibility marking scale in light of three factors: informativity, rigidity, and attenuation (Ariel 2001: 32); informativity relates to the amount of lexical information given in a particular form, rigidity is the ability of that form to select a unique referent, and attenuation refers to phonological size. Ariel’s conception of accessibility bears many similarities to the earlier notions of topicality and topic coherence prominent in the research of Givón (1978, 1983, 1995, a.o.) and the givenness hierarchy as discussed in Gundel et al. (1993).

Related to accessibility is the concept of cataphoricity, the boosting of the accessibility of a referent. While Ariel’s work focuses primarily on retrieving given information, other researchers have extended this framework to include the introduction of new items. Gernsbacher and Schroyer (1989) posit that NP forms used for introducing new items have different levels of cataphoricity, meaning that they boost the accessibility of their referent to varying degrees. This model is supported by the findings of their study of English indefinite ‘this,’ as in, “there was this guy walking down the street.” Participants who heard entities introduced with indefinite ‘this’ rather than with indefinite ‘a(n)’ in the first half of a narrative were more likely to mention this entity sooner and more often when asked to complete the narrative. Gernsbacher and Schroyer’s work suggests that, just as given items of higher accessibility are associated with less linguistic material because they are easier to retrieve, new items with high accessibility may be associated with more linguistic material (i.e., a more complex NP structure), which boosts their accessibility later in the discourse. In addition to NP form, Ariel argues that factors such as the syntactic position of an NP may also increase its cataphoricity (Ariel 2001:49).

2.2 Reference Forms in Chinese Discourse

2.2.1 Introduction
Varieties of spoken and written Chinese differ from languages such as English in their degree of pragmatically-driven variation in NP form. While canonical reference forms exist for entities of a particular semantic status in each Chinese variety, these forms have not yet been fully grammaticalized (Gundel et al. 1993:290, Givón 1978:306). As a result, a reader or listener must consider context to determine whether certain NP forms refer to a new or given entity, or are non-referential. Just as variation in the use of indefinite ‘a(n)’ and indefinite ‘this’ constructions in English may be addressed in an accessibility model (Gernsbacher & Schroyer 1989), variation in NP form in Chinese varieties, in both indefinite and definite contexts, can be examined in light of such factors. In order to clarify the relationships between the present research and previous studies, which have focused on Standard Written Chinese (SWC) and Mandarin, this discussion will introduce the NP forms used in those varieties, and then contrast them with those of Cantonese.

### 2.2.2 Indefinite NP forms

In Standard Written Chinese (SWC), indefinite NPs may be preceded by a numeral-classifier pair or appear simply as a bare noun. While not categorical, there are tendencies for certain NP forms to correlate with the referentiality status of an entity. Following Givón’s notion of referentiality and non-referentiality (Givón 1978:293-294), Sun (1988) posits the following canonical forms for SWC:

[see Table 1, p. 21]

Several authors have observed, however, that the marking of referential versus non-referential NPs in actual discourse often does not correspond to their roles in propositional semantics (Givón 1978, Sun 1988, Gundel et al. 1993, Li & Thomson 1981, et al.). In a similar case involving Krio, Givón found that referential-indefinites which are not salient in the discourse are often marked as if they were non-referential, and non-referentials which are salient are marked with the morphology of referentials (Givón 1978). Sun (1988) found a similar pattern for indefinite NPs in published SWC narratives. Sun supports Givón's explanation for this variation, that “thematic centrality,” meaning importance to the overall narrative, can override an NP’s semantic status in a discourse context (Sun 1988:301).

In spoken Mandarin, indefinites may appear as bare nouns or as numeral-classifier-noun phrases. In addition to these forms, the numeral-classifier construction may optionally appear without the numeral in less formal registers (Cheng & Sybesma 1999:511). Altogether, then, there are three basic indefinite NP forms in spoken Mandarin: bare noun, classifier-noun, and numeral-classifier-noun. Spoken and written Cantonese both follow this same three-option pattern (Table 2)³, although the classifier-
noun form is commonly perceived as being more common or robust than it is in Mandarin, perhaps because Cantonese is primarily a spoken language used in casual registers (ibid).

[see Table 2, p. 22]

Given three possible NP forms and only two referentiality statuses, it is not possible to maintain a one-to-one correspondence between NP forms and canonical referentiality status. If a model is adopted in which bare nouns are canonical non-referential forms, and numeral-classifier-nouns are canonical referential forms, as in the Sun (1988) model of SWC, classifier-noun forms must then be treated as anomalous tokens with no apparent relationship to referentiality. Thus, rather than a model in which topicality or thematic centrality “overrides” a canonical reference form and causes a non-canonical form to be preferred in certain discourse contexts, this study will evaluate an alternative view in which factors of referentiality and topicality both exert influence on the choice of NP form, and more referential and more topical NPs are associated with more complex NP structures.

2.2.3 Definite NP Forms

While the bare noun form is described as the canonical form of non-referential indefinite NPs in SWC (Sun 1988), it is also commonly used for definite NPs. Another definite NP form is demonstrative-classifier-noun; SWC lacks a dedicated definite article such as English ‘the,’ and uses proximal and distal demonstratives for this function (Table 3) (Gundel et al. 1993:285).

[see Table 3, p. 23]

Spoken Mandarin follows this same pattern. Cantonese, however, introduces a third variant, classifier-noun, not present in any register of Mandarin (Cheng & Sybesma 1999:511, Matthews & Yip 2004:93). Thus, Cantonese has two NP forms, bare noun and classifier-noun, which may receive an indefinite or definite interpretation, depending upon context. In total, Cantonese has three definite NP forms: bare noun, classifier-noun, and demonstrative-classifier noun (Table 4):

[see Table 4, p. 24]
It should be noted that this analysis of Cantonese is not uncontroversial; Cheng and Sybesma (1999:512) claim that bare nouns in Cantonese cannot receive a definite interpretation, and that the classifier-noun form has “replaced” the bare noun form in its definite reference function. Based upon the corpus analysis presented here, and consultation with native speakers, this argument by Cheng and Sybesma would seem to be unsupported, at least in vernacular data. However, they may be correct in proposing that the classifier-noun form is replacing another form. Because Chinese has no determiners, the demonstrative-classifier-noun form serves two roles in SWC and in Mandarin, as a demonstrative and as a definite NP (Gundel et al. 1993:285). In Cantonese, it need not serve as a definite marker, given that the classifier-noun construction fulfills this role. Previous work debating the definite function of Cantonese classifiers suggests that the classifier-noun form in Cantonese fills many of the contexts that would employ demonstrative-classifier-noun forms in SWC and Mandarin, thus rendering demonstrative-classifier-noun more of a “true” demonstrative marker in Cantonese than it is in other Chinese varieties (Wu & Bodomo 2009, a.o.).

The observation that the demonstrative-classifier-noun construction in Cantonese may pragmatically correspond to the English demonstrative-noun construction suggests that the accessibility scale is an appropriate model with which to distinguish the pragmatic functions of these three definite NP forms. Following Ariel’s accessibility model, greater “attenuation,” or phonological size, would correlate with lower accessibility (Ariel 2001:32). Thus, in this framework, given entities with higher accessibility (i.e., higher saliency, topicality, etc.) are predicted to be realized as bare nouns, and those with lower accessibility are predicted to use classifier-noun or demonstrative-classifier-noun forms.

2.2.4 Zero Anaphora

All varieties of Chinese allow extensive use of zero anaphora. While considerable research has been done on zero anaphora in Chinese, the questions of where and why speakers choose to use an overt pronoun versus a zero pronoun have proved problematic to resolve. This issue relates to a more wide-ranging debate over the nature of anaphora; while some researchers in the Chomskyan tradition claim that only syntactic and semantic factors determine intrasentential anaphora (e.g., Chomsky 1982), most analyses argue that pragmatic factors primarily govern both discourse-level and intrasentential anaphora (Huang 1991, Levinson 1987, a.o.). In their influential 1979 study, Charles Li and Sandra Thompson found little consensus among native Chinese speakers as to where to put pronouns versus zeroes in a particular text. They concluded that pronoun use was largely a matter of an individual’s perception of how related or “conjoined” a clause was with the previous clause: the more related the clause, the more likely a zero would occur (Li & Thompson 1979:334). However, Liang Tao (1996:492) argues for a quite different view, that a zero may be used to indicate the return of a previous discourse referent not occurring in the immediately preceding clauses (a context known as a “return-pop”). Another contrary view comes from Yang et al. (1999), which finds that both pronouns
and zeroes contribute equally to discourse coherence, as measured by self-paced reading speed. These findings lead the authors to speculate that either a principle of least effort or a stylistic factor similar to the use of contractions in English underlies the choice of pronoun versus zero (Yang et al. 1999:741). In the realm of computational linguistics and natural language processing, researchers have taken various rule-based and statistical approaches to the identification and resolution of zero pronouns, with limited success. Zhao & Ng (2007), for example, use a machine learning approach for zero anaphora resolution that incorporates factors such as syntactic parallelism and distance, but are only able to accurately identify the referents of less than half of the zeroes in their test corpus. The main consensus points in the diverse literature on zero anaphora in Chinese are that zero pronouns are extremely common (to the extent that zeroes may be the default and overt pronouns the marked variant, as argued in Li & Thompson (1979:322)), and that there is a great deal of variation in usage that is influenced, to some extent, by issues of topic continuity.

While SWC, Mandarin, and Cantonese appear to have largely equivalent usage patterns for zero anaphora, one difference in the distribution of pronouns and zeroes is that speakers and writers of SWC and Mandarin are increasingly using overt pronouns to refer to third-person inanimates, while Cantonese preserves the more conservative preference for zero anaphora in these contexts (Chan 1985:67, Li 1997:291). In fact, while SWC has a character corresponding to a pronoun specifically denoting inanimates (它), written vernacular Cantonese does not.

### 2.2.4 Plural Classifier

With the exception of pronouns, Chinese varieties lack mandatory plural markers. Cantonese, however, makes extensive use of the classifier 啲 di, which indicates a plural quantity (Matthews & Yip 2004:98). Cantonese di behaves rather differently from the Mandarin classifier 些 xie (‘some’). Unlike xie, di can mark plurals or generic nouns, as in 1(a-b).

(1) a. 啲西瓜 食晒啦 (definite plural)

   di  watermelon eat all MOD

   “The watermelons have all been eaten.”

b. 啲西瓜 就快絕種啦 (generic)
Most notably for the present study, the structure of the *di* noun phrase is not consistent with other classifiers, in that *di*-noun is often the fullest possible form that can correspond to a particular interpretation. For example, in the case of 1(b), a generic interpretation is not possible with the addition of a demonstrative before the *di*. This poses a problem for coding along the accessibility scale, which otherwise operates under the assumption that classifier-noun is an intermediate form between bare noun and numeral-classifier-noun or demonstrative-classifier-noun. Due to the unusual behavior of *di*, it will be coded separately from other classifiers.

### 2.3 Previous Studies

While no previous research has examined indefinite and definite NP forms in Cantonese discourse, there has been work focusing on discourse factors in indefinite NP choice in SWC. Sun (1988) sought to account for variation between numeral-classifier and bare noun constructions in indefinite NP forms by appealing to discourse factors, using a corpus of five short stories (Sun 1988:303). Sun found that only 8% of topics marked as “minor” by independent raters were introduced with numeral-classifier constructions, as compared to 29% of intermediate and 80% of major topics (317). Both “thematic centrality” (importance to the overall narrative, as determined by independent raters) and “topic persistence” (how often a referent appears in the narrative) were significant predictors of NP form in the case of referential NPs (ibid). Sun also noted that plurality correlated with bare noun marking, and posited that this was because plural topics are low in individuation (307).

Li (2000) approached the same theoretical question as Sun (1988), taking issue with Sun’s claim that “thematic centrality,” the overall importance of a topic to the narrative as a whole, is the overriding factor in determining NP form. Instead of thematic centrality, Li hypothesized that the function of the numeral-classifier construction is to “foreground” an NP, i.e., indicate the saliency of an NP in its local discourse context, regardless of its overall thematic importance (1114). Li found a significant correlations between several foregrounding features, including modifying clauses and existential constructions, and the use of numeral-classifier marking (1120, 1122). Li also argued that non-referential NPs were marked with numeral-classifiers to foreground their attributes.

The two studies described above agree that discourse factors predict NP form in
Chinese. While there is some disagreement over precisely what pragmatic qualities are associated with the different NP forms, both studies conclude that the fuller NP form is associated with entities being more important in the discourse in some sense. Other scholars, however, disagree that the numeral-classifier construction marks increased salience in SWC. Chen (2004), for example, claims that numeral-classifier constructions de-inviduate the referent and mark them as less likely to serve as a topic in later discourse. Thus, the question of how NP form interacts with discourse factors in the case of indefinite NPs in SWC has not yet been entirely resolved.

The literature discussed here sheds light on the question of which discourse factors influence NP form in the case of indefinites, but it does not address how definite NPs may be predicted or how a previous NP form may affect the form of the next reference to that same entity. The conclusions of previous work are also limited in that they concern NP forms in SWC, and may or may not be applicable to other spoken and written varieties of Chinese. In addition, both studies looked exclusively at formal narration of traditional folktales, which is not equivalent to everyday speech. The present corpus study will examine a broader range of theoretical questions, using a more vernacular corpus in a different variety of Chinese. These findings may allow the conclusions from Sun and Li to be applied to a broader range of registers and varieties, as well as answer some theoretical questions not addressed in previous research.

3 Corpus Study

3.1 Corpus Description

3.1.1 Introduction

The present corpus study aims to address two major theoretical questions. First, it will examine possible factors that might predict the form of new items introduced into a discourse. The second goal is to test whether the form of a previous reference has any unique effect on the form of the subsequent reference to the same item. These research aims inform the selection of a corpus for analysis.

From a practical perspective, the best genre to analyze for repeated references to multiple entities is narrative, for the simple reason that, in narratives, speakers often introduce entities and refer back to them. At the same time, because this study seeks to examine the influence of processing-related factors in the choice of reference form, the need for narrative must be balanced with the desire to analyze naturally-occurring, spontaneous data. This requirement rules out the sort of published short story data used in Sun (1988) and Li (2000). Another obstacle in selecting a corpus is the fact that there are very few linguistic corpora in Cantonese. The corpora that are available are mostly drawn
from news text, which is not written in vernacular Cantonese. Therefore, a new corpus was collected and tagged for the purpose of this analysis. The data used in this corpus came from posts on an Internet message board called Mail2Love. Mail2Love is a message board hosted on the site of Yes! magazine, a popular entertainment magazine for young people in Hong Kong.

The primary language of Mail2Love is Vernacular Written Cantonese (VWC), a variety that closely corresponds to spoken Cantonese (Snow 2004). Table 5 illustrates some common lexical differences between SWC and VWC, as described in Yan (2005):

[see Table 5, p. 25]

One advantage of the VWC found on the Internet is its close conformity to spoken Cantonese. Unlike more formal forms of VWC, the VWC used on Internet message boards contains frequent use of modal final particles, discourse markers, and other features typical of casual speech. Posts were considered to be in VWC if they contained the VWC 佢 keuih rather than the SWC 他 ta for the third person, and the copula 係 haih rather than 是 shi. The other lexical items mentioned in Yan also proved to correlate with these two distinctions.

3.1.3 Mail2Love

As is suggested by its title, Mail2Love is a board devoted to questions of the heart. Authors post problems about their love life, and receive responses from other visitors to the site. While the majority of content posters are female, there is a moderate amount of diversity in gender and sexual orientation in the Mail2Love posts. Also, while some posts discuss marriages and jobs, the most typical messages are from female junior high school or high school students who are concerned about a potential, current, or former boyfriend. The posts contain narratives explaining the author’s dilemma, often followed by a request for advice about how to resolve the situation. Here is a typical example from the corpus in (2) (with the original Cantonese followed by an English translation):

(2)

Mail2Love Post 9
As soon as school began I started to like a boy (This boy is very handsome, he’s the “grass of the school” [hottest student]!), before the middle of the term, he sat next to me, so according to what the girls in our class say, he was most familiar with me. In the second half of the semester, the teacher planned the seating chart, and he became my deskmate, at that time I was so happy! During the time I was sitting with him, we talked and laughed, all day we’d be going back and forth, we also had many so-called physical contacts, we were even touching each other’s shoulders all day!! In the class, there were many people who thought that we had some relationship, but the whole class knew that he has a girl, in a nearby school, and he and his girlfriend have been together for a long time! I know that he’s a very focused person, and he is very cool, when he is with girls he doesn’t know, he doesn’t look at them very much. But this good situation didn’t last long, the teacher discovered that our grades were dropping, and also we were chatting all day and not paying attention in class, so he separated us. Before we were separated, we suddenly went from hot to cold, we weren’t as close as before, cold and indifferent, I don’t know how to fix it, my friends also don’t know how to fix it. His coldness is coming from his heart, I can feel it. Now that we are separated, things are even worse, I haven’t spoken one word with him, it’s down to the point where we don’t even say hello, it’s like we are strangers. How can I settle this?? What should I do now??

This sort of narrative allows us to track how the same entity is referred to over time. For example, in this post, there is a love interest. He is introduced into the narrative as 一個男仔 ("a boy"), and is referred to in the post a total of 17 times. We can trace the
course of NP forms referring to him as follows in (3):

(3) num-cl-n; prox-cl-n; zero; pro; zero; pro; pro; pro; pro; pro; pro; zero; zero; pro; zero; pro

Because Mail2Love contains hundreds of posts, and because some systematic method was needed to rule out unacceptable posts, the following criteria were adopted:

1. The post must be in VWC, as defined by the standards outlined in the previous section.
2. The post must be at least four lines long.
3. The post must be relevant to the topic of the board (i.e., not an advertisement for English lessons)

The first criterion has been previously justified. The second criterion was included in order to maximize the chances that the post would contain a substantial narrative involving a number of entities referred to multiple times. Finally, the third criterion was meant to maximize the thematic consistency of the corpus (the implications of this will be discussed in a later section). Over the course of several days in 2005, the first 30 Mail2Love posts which met these three criteria were selected for inclusion in the corpus.

3.2 Topicality Scale

Sun (1988) found “thematic centrality” to be a good predictor of first-mention NP form. In Sun’s study, thematic centrality was determined by a group of raters blind to the hypothesis, who were instructed to pick out the important things in each narrative. In the case of the Mail2Love corpus, this analysis will make use of the strong thematic consistency of the content to formulate broad predictions about topicality.

As discussed in the introduction of the Mail2Love corpus, each post in the corpus contains a narrative about the author and a set of love interests. Due to the shared subject matter and cultural context, there are remarkable parallels between the narratives. As a result, the same sorts of entities are similarly central or non-central to the discourse throughout the corpus. Thus, for the purposes of this study, global topicality will be treated as a corpus-wide scale:
1. Speaker
2. Love interest (boyfriend, etc.)
3. Other animate entity
4. Inanimate entities

In this topicality scale, NPs referring to the speaker (or author of the post) are considered most central to the discourse. This high topicality level will also be applied to non-referential NPs in predicate constructions such as in (4):

(4) 我係 一 個 幾 fat 既 女仔
I am one CL pretty fat REL girl
“I am a pretty fat girl.” (Mail2Love Post 1)

Following Givón (1978), and Sun (1988), non-referentials with a high topicality level such as these are predicted to receive the fuller NP forms typical of referential NPs.

The most notable split predicted in the topicality scale is the distinction between love interest animates and other animates. A topicality scale blind to the content of the corpus might predict an animate versus non-animate contrast, but not this thematically relevant contrast. It will be of particular interest, therefore, to test for the significance of this contrast in the corpus analysis.

In addition to the topicality scale presented here, topicality will also be measured via entity frequency rank. If the proposed topicality scale is accurate, then these two operationalizations of topicality should correlate.

3.3 Model Factors

Every NP in the Mail2Love corpus, including zero anaphora occurring where a realized NP might have occurred, was hand-tagged for the following information by the author:

1) NP form (zero, pronoun, bare noun, classifier-noun, plural marker-noun, numeral-classifier-noun, proximal demonstrative-classifier-noun, distal
demonstrative-classifier-noun, proper name)
2) Referentiality (referential vs. non-referential (following Givon 1978))
3) Definiteness (definite, indefinite, non-definite (e.g., 視戲 tai-hei, ‘watch movies’) (following Givon 1978))
4) Givenness (new, given, inferred (following Prince 1981))
5) Topic (see discussion above in 3.2)
6) Plurality
7) Syntactic position (subject vs. non-subject)
8) Presence of a modifier / modifying clause.
9) Classifier Type (個 go, 只 jek, etc.)
10) Entity being referred to (identified by a unique number)
11) Distance (number of clauses away from previous reference to same entity)

Additional information was automatically calculated:

12) Raw number of occurrences of the same entity in the post.
13) Rank of the frequency of the entity in the post relative to all other entities.
14) Presence of intervening entity between present and previous reference.
15) NP form of intervening entity, if any.

3.3 Results

3.3.1 Tokens Collected
A total of 1,244 tokens of NPs were tagged. Of these, 1,086 were included in the analysis; the others were excluded because they were fixed expressions that could not alternate with other forms: specifically, non-definite complements in verb-complement constructions that cannot function independently (e.g., 分手 fan-sau (‘to break up’)), time phrases (e.g., ...既時候 ge sih-hauh) establishing setting such as “when I was in class,” and zeroes referring to abstract concepts that could not alternate with pronouns.

3.3.2 Preliminary predictions
3.3.2.1 Variation in NP form

In order to account for variation in NP form that may be predicted by accessibility factors, it must first be established that there is indeed variation in NP form that does not strictly correspond to the discourse status of the referent entity. In other words, it must be ascertained that not all referential indefinites are introduced using NPs of the type numeral-classifier-noun, etc. Cross-tabulations of the tokens in each cell for referential versus non-referential indefinites, and for indefinites versus definites, establishes that considerable variation is present in the corpus:

[see Table 6, p. 26]

[see Table 7, p. 27]

As predicted, both non-referential and referential indefinite NPs appear with all three possible NP forms, and there is a clear tendency for the referential NPs to be marked with more linguistic material. Examples of referential and non-referential uses of the bare noun form are shown in (5):

(5) a. non-referential: 仲 會 有 fd 做
   ZERO still will not have friend do
   “I still won’t have any friends to do it” (Mail2Love Post 4)

b. referential: 但 佢 已經 有 男朋友 呀！
   but she already has boyfriend MOD
   “But she already has a boyfriend!” (Mail2Love Post 6)

Also as predicted, definites appear in with all possible NP forms, including bare nouns, as in (6):

(6) 我 同 我 個 同學 一齊 搭 巴士 既 時候...

14 / 40
I and my CL classmate together ride bus REL time

“When my classmate and I were riding the bus together...” (Mail2Love Post 7)

Another prediction supported in these results is the claim that proximal and distal demonstratives are relatively rare in Cantonese, and that the construction classifier-noun serves the canonical definite role filled by demonstrative constructions in Mandarin (Gundel et al 1993:285). Hence, in contrast to Mandarin, the data support the view that demonstratives in Cantonese serve much the same function as they do in English, at least in the case of this construction. (7) gives contrasting examples of the use of demonstratives versus the classifier-noun construction:

(7) a. dem-cl-n: 我覺得呢 個世界 個 個人 都 只 會 單單 睇
   I feel this CL world CL CL person all only will simply look
   外表
   outside appearance

   “I feel that, in this world, everyone only looks at what’s on the outside.” (Mail2Love Post 21)

b. cl-n: 如果個 仔 仔 真係 可以 除傳除到
   if CL boy call her she really can ready anytime

   “If the guy asked her out then she would really be able to go anytime.” (Mail2Love Post 16)

As for plural classifier di, the construction di-noun is found in both indefinite and definite contexts, as in (8):

(8) a. indefinite: 成 日 撩 d 女仔 ge
   ZERO all day flirt PL girl MOD

   “He flirts with girls all day.” (Mail2Love Post 20)
b. definite:  但 d fd 都話 佢係鍾意我.

but PL friend all say he does like me

“But [my] friends all say he does like me.” (Mail2Love Post 4)

Although the construction numeral-di-noun is possible in more formal Cantonese, this pattern never appeared in the corpus. This analysis will therefore treat di-noun as the fullest indefinite di form. In the analysis of definites, however, di forms will be excluded because their position in the accessibility scale in relation to other forms is too ambiguous.

Now that it been established that variation does exist in NP form, a model can be developed to predict NP form based on the various factors coded for in the Mail2Love corpus.

3.3.2.2 Topicality scale vs. Popularity Rank

As predicted, the topicality scale positing a global ranking of topics for the corpus shows a significant positive correlation with entity popularity rank (Graph 1). The data indicate that speaker was indeed the most popular entity, followed by boyfriend. The considerable gap between boyfriend and other animates is symptomatic of the thematic uniformity of the corpus; other than themselves and their boyfriends, these authors have little else on their minds.

[see Graph 1, p. 29]

3.3.3 New Entities

The first subject of analysis will be indefinite items first introduced into the discourse. In light of previous studies by Sun (1988) and Li (2000), the same general pattern of influence of discourse factors in NP form should hold for both non-referentials and referentials, with more salient or topical entities being marked with more linguistic material. The following fixed effects were tested for significant unique predictiveness of NP form: referentiality, topicality, plurality, subjecthood, and modifier presence. In addition, the random effect of post number was included to account for variation between authors.
A stepwise multinomial logistic regression analysis was performed in SPSS. The best predictive model found in this statistical analysis ($\chi^2 = 59.921, p < .0001$) included the fixed factors of referentiality ($\chi^2 = 26.905, p < .0001$), topicality ($\chi^2 = 20.014, p = .003$), and plurality ($\chi^2 = 10.376, p = .006$). In a model including all of the factors, neither of the other fixed factors—modifier presence and subjecthood—reached significance, and collectively reduced the predictive power of the model.

The fact that referentiality was a significant factor in the model indicates that, while NP forms are variable and each form is potentially ambiguous between referential and non-referential readings, non-referential NPs and referential NPs are marked significantly differently. Referring back to Table 6, it seems that the referential NPs form a predictable scale from bare noun up to numeral-classifier-noun, such that the frequency of referential NPs increases regularly as linguistic material increases. Non-referential NPs, however, behave somewhat differently; the frequency of non-referential NPs drops to one for the classifier-noun form, and then back to 13 for numeral-classifier-noun. We would have expected them to decrease linearly. Why are there so few tokens of the intermediate form? These data suggest that the relationship between the three forms examined here may not be strictly linear. One possible explanation is that the particular context these non-referential NPs were often found in within the Mail2Love corpus, the predicative construction, prefers fully marked numeral-classifier-noun NPs rather than the classifier-noun construction, as in (9):

(9) 我係 一 個 幾 fat 既 女仔
I am one CL pretty fat REL girl

“I am a pretty fat girl.” (Mail2Love Post 1)

As noted in Li (2000:1126), there is generally a lower frequency of marked non-referentials in shorter narratives such as the stories in Mail2Love, perhaps because the primary function of these non-referentials is to describe a scene rather than directly perpetuate the central narrative. For this reason, the disproportionately high occurrence of non-referentials in these predicative constructions may result in the overall distribution of non-referentials not reflecting their distribution in actual discourse. Alternately, there may be something genuinely marked about the classifier-noun construction for non-referentials.

The second two predictive factors, topicality and plurality, are the same factors that proved significant in predicting the form of referential indefinites in Sun (1988). However, contrary to Sun’s finding that plurality reduced the likelihood of a noun being marked with a numeral-classifier construction, in the Mail2Love corpus, plurals are more
likely to be marked with more linguistic material than singular nouns. This appears to result from a linguistic difference between Cantonese and Mandarin/SWC; specifically, this confirms previous work (e.g., Au-Yeung 2007) suggesting that the *di* plural classifier in Cantonese is used more frequently than the *xie* plural classifier in Mandarin and functions more as a generic or default plural marker than as a measure word meaning “some.”

Topicality, which was shown to correlate with popularity rank of entity in the previous section, is the discourse factor with the highest predictive power in the model. Significantly, this factor refers to the overall topicality of the entity within the entire genre of Mail2Love postings (i.e., whether the entity is the speaker, the boyfriend, other animate, or other), not that entity’s relevance in that particular narrative or its prominence or “foregrounding” in a given local context. In fact, both of the “local context” factors meant to evaluate Li (2000)’s claims, modifier presence and subjectxood, proved to be insignificant predictors of NP form in the case of indefinites. Given the small number of indefinites available in this corpus, this result does not demonstrate conclusively that these local factors are of no significance in determining NP form in VWC, but the data do suggest that overall topicality is a more influential factor.

### 3.3.4 Given Entities

The second portion of this analysis seeks to model the selection of NP form for given items, and specifically to determine whether the previous NP form used to refer to that same entity has any significant predictive power independent of other discourse factors. Ideally, this model would incorporate every possible given NP form into a scale forming one dependent variable. Looking again at the distribution of tokens along the scale for definites in Table 7, however, it is quickly apparent that a single scale is not advisable. At the “high accessibility” end of the scale, there are 182 tokens of zero and 590 pronouns, while at the low end, there are only eight proximal demonstrative constructions and six distals. Rather than lumping the forms at the higher end together, thus losing some of the distinctions that make VWC so interesting, two separate analyses will be conducted for the two distinctions with reasonably high and comparable token counts: pronoun versus zero and noun versus classifier-noun. The following factors were considered in both analyses: topicality scale, modifier presence, subjecthood, plurality, NP form of previous reference, syntactic agreement between previous and current reference, distance from previous reference to the same entity, and presence of intervening entity references between the previous and current reference.

In the case of bare noun versus classifier-noun, a stepwise logistic regression found the best model ($\chi^2 = 23.77, p = .003$) to have two significant factors: the NP form of the previous instance ($\chi^2 = 12.09, p = .034$), and topicality ($\chi^2 = 12.64, p = .002$).

According to the theory of accessibility boosting, a previous NP form with more linguistic material should boost the accessibility of that entity such that it can be referred
to with less linguistic material later on. This prediction is upheld in the model: a previous NP form with more linguistic material is associated with significantly less linguistic material in the current NP form.

It is not surprising that topicality is a significant predicting factor here, considering that it was the biggest predictor of NP form in the case of indefinites. What is unexpected, however, is that the relationship between topicality and NP form in this case is the opposite of what is predicted based upon the accessibility scale, as shown in Graph 2. Both entities of lower popularity rank and those considered less thematically important on the topicality scale are significantly more likely to be marked with a bare noun rather than a classifier-noun form. According to Ariel’s accessibility scale, the opposite should be true, meaning more popular entities should be more easily accessible and therefore referred to with less linguistic material. Before the significance of this result can be evaluated, it is necessary to determine whether this is a fluke occurring only on one part of the scale, or whether it can be duplicated in the zero versus pronoun analysis.

[see Graph 2, p. 30]

The second contrast, pronoun versus zero, yields a logistic regression model ($\chi^2 = 134.54, p < .0001$) with five significant factors: subjecthood ($\chi^2 = 39.18, p < .0001$), syntactic agreement with previous reference ($\chi^2 = 24.77, p < .0001$), topicality scale ($\chi^2 = 14.92, p < .0001$), distance from previous reference ($\chi^2 = 13.63, p < .0001$), and previous NP form ($\chi^2 = 5.2, p = .023$). Given that there are many more tokens to work with in this case, it is not unexpected that this model has found several more significantly predictive factors than in the previous contrast.

As was the case in the noun versus classifier-noun analysis, the effect of topicality is the opposite of what was expected: more topical entities are more likely to be referred to with pronouns, while less topical ones are more likely to get zeroes. In this particular contrast, it may be relevant that the entities near the top of the topicality scale are all animate, while the bottom of the scale is inanimate. In Cantonese, it is very unusual to refer to inanimate objects with pronouns, thus making it more likely that inanimates will be referred to with full nouns or zeroes while animates may be referred to with pronouns. While this may account for the phenomenon in the situation of the pronoun versus zero contrast, it does not account for the similar pattern seen in the bare noun versus classifier-noun contrast. Also, the fact that this trend holds for both popularity rank and for the topicality scale suggests that it is not reflective of some flaw in the operationalization of topicality, but is a genuine phenomenon that presents certain challenges to the notion of
an accessibility scale.

In the previous contrast, it was found that a previous NP with more linguistic material would boost the accessibility of its referent so that it would be referred to with less linguistic material in its next instantiation. For pronoun versus zero, however, there is a significant trend in the opposite direction, as shown in Graph 3:

[see Graph 3, p. 31]

These results indicate that zeroes prefer to follow other zeroes, or at least reference forms with less linguistic material. This runs contrary to what one might expect, that a zero would be more likely to follow a lengthy or explicit reference so that it could easily be traced back to its source. However, the contribution of this factor is quite small in comparison with other discourse factors, including distance from previous reference and syntactic agreement, which make clear contributions to a speaker’s ability to trace back the referent of a zero.

As predicted by accessibility theory, greater distance from the previous reference increased the probability of the occurrence of a pronoun versus a zero (the previous reference was an average of 1.207 clauses away in the case of zeroes, and 1.620 clauses away for pronouns). When the current reference instance and the previous reference instance agreed on their syntactic position (subject versus non-subject), the current reference was more likely to be a zero than if there was syntactic disagreement. This phenomenon is consistent with theories presented in previous literature (Li & Thompson 1979, Ariel 2001, et al.), since consistency in syntactic position should make the referent easier to trace and therefore more likely to be realized as a zero rather than as an overt pronoun. There was also a significant main effect of subject position; when the current reference is in subject position, it is more likely to be a zero than when it is not in subject position. As Graph 4 illustrates, there is a significant interaction (F=9.286, p=.002) between these two factors such that references which are in subject position are affected more by the syntactic agreement factor than references not in subject position. The reason for this interaction may be that NPs in the non-subject position have a variety of syntactic functions, and the measure used in the current study, subject vs. non-subject, did not capture these differences. Thus, NPs that are coded as being in syntactic agreement in the sense that they are both non-subjects are in fact not entirely syntactically parallel, making it more likely that the current NP would be realized as a pronoun.

[see Graph 4, p. 32]
Given that previous attempts to predict zeroes in Chinese have been somewhat unsuccessful due to supposed inter-speaker inconsistency in zero usage (Li & Thomson 1979), it seems relevant to evaluate the success of the present best-fit model in predicting pronouns and zeroes. As discussed in the preliminary results section, there were 182 zero tokens and 590 pronoun tokens in the corpus. Thus, if a model predicting pronoun versus zero were to guess ‘pronoun’ in every single instance, it would be right approximately 76% of the time. The current model is correct 80% of the time, which is only a small improvement over this baseline rate. The model yields the specific predictions shown in Table 8:

[see Table 8, p. 28]

When the model predicts a pronoun to occur, it is correct 82.7% of the time. When it predicts a zero, however, it only has a 59.5% chance of being correct. This suggests that the model has successfully identified several positions in the discourse, as defined by contextual factors accounted for in this model, where pronouns are extremely common or obligatory, but has been unable to identify many such positions for zeroes.

3.4 Discussion of Results

This analysis of the Mail2Love corpus has addressed two major aims: predicting the form of new items, and predicting the form of given items. For new items, some of the findings supported earlier results, such as Sun (1988)’s conclusion that overall thematic centrality significantly influences the choice of NP form. Other findings contradict earlier results and point to differences between Cantonese and Mandarin, such as plurality predicting longer NP forms in these data, in contrast to shorter forms in Sun (1988). It was also found that certain local contextual factors, including the presence of modifiers on the NP, had no significant predictive power in the Mail2Love corpus, thus failing to support the findings of Li (2000). Finally, some anomalous trends suggested that the classifier-noun form may not form a linear scale with the bare noun and numeral-classifier NP forms, at least in the case of non-referentials.

The results for indefinite NPs should be regarded as preliminary findings in the study of indefinite NP behavior in VWC. Because each narrative in the Mail2Love corpus contained very few items overall and did not have as many new items being introduced as we might find in some more traditional narration genres, it presents a limited setting in which to examine the influence of discourse factors on new items. One advantage of the Mail2Love corpus, however, is that it provides a perfect opportunity to examine differences between animates with different levels of topicality. Indeed, significant differences were found between the NP forms associated with love interests and those associated with friends and other less topical animates.
In the second half of this analysis, the given NPs were split into two separate contrasts. Both the classifier-noun versus bare noun contrast and the pronoun versus zero contrast yielded some surprising results; foremost among these was that topicality influenced NP form in the opposite direction than that predicted by accessibility theory. Rather than more topical items being referred to with less linguistic material, they were referred to with more. In the case of pronoun versus zero, this phenomenon might be due to the lack of pronouns used to refer to inanimates, as previously discussed. However, this explanation cannot account for the entire finding, given that the vast majority of entities in the corpus were animate, and that we see the same trend in the contrast of classifier-noun versus bare noun.

One possibility that must be addressed in considering why these findings were inconsistent with accessibility theory is that both the topicality scale and the entity popularity ranking used to evaluate topicality in the Mail2Love corpus are somehow not linked to accessibility. Given that topicality did correlate highly with the NP form of new items in both this study and in Sun (1988), and that it is generally considered to be associated with accessibility (Gernsbacher & Schroyer 1989, Ariel 2001, et al.), there is no evidence to support an assertion that more topical and more frequent entities are not more highly accessible. This strongly suggests that the analysis of given items did indeed show that more accessible items were referred to with NP forms associated with lower accessibility in the canonical accessibility scale, contrary to the predictions of accessibility theory.

A similar problem for accessibility theory is posed by the behavior of the previous NP factor. While small, there was a significant trend indicating that a previous NP form with more linguistic material made it more likely that the following NP reference would have more linguistic material, not less. However, if the previous finding is incorporated into the present analysis, this result is actually consistent with a modified version of the accessibility-boosting hypothesis, in the sense that higher accessibility in these data consistently leads to the use of forms with more linguistic material. In the case of the present factor, a previous NP form with more linguistic material boosted the accessibility of the entity, and made it more likely that it would be referred to next with more linguistic material, just as those entities of higher topicality and popularity appeared with more linguistic material. Thus, when these two factors are taken together, they both reinforce the finding that the predictions of the accessibility scale are not upheld, and confirm the prediction that previous NP form will have a significant accessibility-boosting effect on the current NP form.

4 Conclusions

In this study of a corpus of Vernacular Written Cantonese, combinations of various discourse factors have been shown to predict NP form of new entities first introduced into
the discourse and given entities that have already been introduced. For new items, referential, topical, and plural items are more likely to be referred to with NP forms containing more linguistic material. For given entities referred to with bare noun or classifier-noun forms, topical items and items whose previous reference have a longer NP form are more likely to be referred to with longer NP forms. In the case of pronouns and zeroes, topical entities, items whose previous reference has a longer NP form, items in non-subject positions, items whose syntactic position do not agree with the previous reference’s position, and items whose previous reference is further away are more likely to be referred to with longer NP forms.

While the predictions of accessibility theory have not generally been upheld in the case of given items, it is not the case that every factor has gone entirely in the opposite direction from the one predicted. The fact that zeroes become less likely the further a reference gets from its previous reference, for example, indicates that at least some of the behavior of the data in the Mail2Love corpus is consistent with previous claims about the relationship between ease of retrieval and NP form. In other words, these results do not warrant the conclusion that Cantonese is somehow uniquely uninfluenced by concerns of retrievability and therefore is not subject to accessibility effects. One possible explanation of these data is that there is some idiosyncratic style at work here specific to the genre of advice-seeking personal narrative in the Mail2Love corpus. Given that the global topicality scale posited for the entire Mail2Love corpus turned out to be highly significant, it is possible that, just as animates are more likely than inanimates to be referred to with pronouns in Chinese, in this particular genre more highly topical entities such as boyfriends may be given more NP material to highlight their saliency, even though they are more easily retrievable. Thus, given entities would behave like new entities, in that more topical entities would be given more linguistic material. We can see this style at work in the very beginning of Post 9:

(10) 我一開學既時候就中意左一個男仔（哩個男仔好靚仔架）

I once began school REL time then like CMPL num-CL-boy this-CL-boy

very handsome MOD

“As soon as school began I started to like a boy (This boy is very handsome...”
The author uses the complete proximal demonstrative form “this-CL-boy” in her second mention of her love interest, even though she has just introduced him in the preceding clause with the fullest possible indefinite form, numeral-classifier-noun. If economy were her only concern, she could have used a zero in place of “this-CL-boy,” and her readers would have had no problem retrieving the correct referent. Instead, she chose to use a form on the opposite end of the accessibility scale. This stylistic factor might be viewed as a sort of highly-ranked constraint that tells the speaker or writer not to attenuate topical entities. If this constraint is particular to a genre, it remains to be seen how far-reaching it is; it may be specific to personal narration, or to any sort of narration in Cantonese. It may well be a feature of written language that does not appear in speech. Additional studies of additional narrative styles in Cantonese would shed light on this question.

While this corpus study of VWC has answered several questions about the role of discourse factors in predicting NP form, it has raised several more relating to how these factors predict forms referring to given entities. The increasing presence of VWC on the Internet means it is less difficult than in the past to conduct research on this variety of Chinese through publicly-available corpora. Further research in a similar narrative genre may resolve questions about stylistic factors that could contribute to a model which predicts NP form and is consistent with the claims of accessibility theory. Given the data available in the present study, it is evident that several discourse factors, including global topicality and previous NP form, play a significant role in predicting NP form for both new and given entities in VWC.

References


Zhang, Nina. 2004. Representing specificity by the internal order of indefinites. Taiwan: National Chung Cheng University, ms.
<table>
<thead>
<tr>
<th>type</th>
<th>canonical form</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-referential indefinite</td>
<td>bare stem</td>
<td>我 没 看到 人 wo mei kandao ren I NEG see man “I didn't see a man.”</td>
</tr>
<tr>
<td>referential indefinite</td>
<td>numeral-classifier</td>
<td>我 看到 一 個 人 wo kandao yi ge ren I see one CL man “I saw a man.”</td>
</tr>
</tbody>
</table>

*Table 1: Marking of NP propositional semantic properties for indefinites in Standard Written Chinese (Sun 1988:301-303)*
<table>
<thead>
<tr>
<th>form</th>
<th>example</th>
</tr>
</thead>
</table>
| bare stem       | 我 見到 人  
ngoh gihndou yahn  
I see man  
“I saw a man”   |
| classifier-NP   | 我 見到 個 人  
ngoh gihndou go yahn  
I see CL man  
“I saw a man”   |
| numeral-classifier-NP | 我 見到 一個 人  
ngoh gihndou yat go yahn  
I see one CL man  
“I saw a man”   |

*Table 2: Forms of indefinite NPs in Cantonese*
<table>
<thead>
<tr>
<th>form</th>
<th>example</th>
</tr>
</thead>
</table>
| bare stem               | 我看到人<br>
|                          | *wo kandao ren*<br>
|                          | I see man<br>
|                          | “I saw the man”              |
| demonstrative-classifier-NP | 我看到那个人<br>
|                          | *wo kandao na ge ren*<br>
|                          | I see that CL man<br>
|                          | “I saw the [that] man”      |

*Table 3: Forms of definite NPs in Standard Written Chinese (Sun 1988:301-303)*
<table>
<thead>
<tr>
<th>form</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>bare stem</td>
<td>我見到人 ngoh gihndou yahn</td>
</tr>
<tr>
<td></td>
<td>I see man</td>
</tr>
<tr>
<td></td>
<td>“I saw the man”</td>
</tr>
<tr>
<td>classifier-NP</td>
<td>我 見到 個人 ngoh gihndou go yahn</td>
</tr>
<tr>
<td></td>
<td>I see CL man</td>
</tr>
<tr>
<td></td>
<td>“I saw the man”</td>
</tr>
<tr>
<td>demonstrative-classifier-NP</td>
<td>我 見到 嗰 個人 ngoh gihndou go go yahn</td>
</tr>
<tr>
<td></td>
<td>I see that CL man</td>
</tr>
<tr>
<td></td>
<td>“I saw the [that] man”</td>
</tr>
</tbody>
</table>

*Table 4: Forms of definite NPs in Cantonese*
<table>
<thead>
<tr>
<th>English</th>
<th>SWC</th>
<th>VWC</th>
</tr>
</thead>
<tbody>
<tr>
<td>copula (‘to be’)</td>
<td>是 shi</td>
<td>係 haih</td>
</tr>
<tr>
<td>relativizer</td>
<td>的 de</td>
<td>嘢 ge</td>
</tr>
<tr>
<td>‘very’</td>
<td>很 hen</td>
<td>好 hou</td>
</tr>
<tr>
<td>3rd person pronoun</td>
<td>他 ta</td>
<td>佢 keuih</td>
</tr>
<tr>
<td>‘to see’</td>
<td>看 kan</td>
<td>睇 tai</td>
</tr>
</tbody>
</table>

*Table 5: SWC vs. VWC on the Internet (Yan 2005)*
Table 6: Crosstabulation of indefinite NP form by referentiality

<table>
<thead>
<tr>
<th>NP Type</th>
<th>Referentiality</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>noun</td>
<td>36</td>
<td>11</td>
</tr>
<tr>
<td>classifier-noun</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>di-noun</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>numeral-classifier-noun</td>
<td>8</td>
<td>24</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>52</td>
</tr>
<tr>
<td>NP Type</td>
<td>Definite</td>
<td>Total</td>
</tr>
<tr>
<td>------------------</td>
<td>----------</td>
<td>-------</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>zero</td>
<td>0</td>
<td>182</td>
</tr>
<tr>
<td>pronoun</td>
<td>0</td>
<td>590</td>
</tr>
<tr>
<td>noun</td>
<td>47</td>
<td>73</td>
</tr>
<tr>
<td>classifier-noun</td>
<td>14</td>
<td>43</td>
</tr>
<tr>
<td>di-noun</td>
<td>9</td>
<td>20</td>
</tr>
<tr>
<td>num-classifier-noun</td>
<td>32</td>
<td>0</td>
</tr>
<tr>
<td>proximal-classifier-noun</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>distal-classifier-noun</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>name</td>
<td>0</td>
<td>36</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>102</td>
<td>961</td>
</tr>
</tbody>
</table>

*Table 7: Crosstabulation of NP form by definiteness*
<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>zero</td>
<td>pro</td>
</tr>
<tr>
<td>zero</td>
<td>47</td>
</tr>
<tr>
<td>pro</td>
<td>32</td>
</tr>
</tbody>
</table>

*Table 8: Predictions of logistic regression model for pronoun vs. zero.*
Graph 1: Relationship of topicality scale and popularity rank.
Graph 2: Mean topicality scale value by NP form for definites.
Graph 3: Mean value of previous NP form (0 = zero, 1 = pro, 2 = noun) by current NP form.
Graph 4: Interaction of subject agreement and subjecthood for zero and pronoun NPs
Thanks to Drs. Sik Lee Dennig, David Beaver, Chaofen Sun, Huihsin Tseng, Dan Jurafsky, and Chung-chieh Shan. All errors are my own.

This analysis will not discuss syntactic or word-order constraints on various interpretations of NP forms. For accounts of syntactic constraints on the interpretation of NPs in Chinese, see Zhang (2004), Cheng & Sybesma (1999), and Givón (1978).

Cantonese data is presented using Yale romanization.


Because this study is meant to account for variation in reference form, it was not of interest to include zeroes in contexts where the zero could not alternate with a pronoun or some other NP.