The role of main verbs in subextraction of wh-phrases from NPs

Nick Huang & Zheng Shen
National University of Singapore

1. Introduction

Subextraction of a wh-phrase from a NP object has long been known to be sensitive to a variety of factors. Much work has focused on how the definiteness of the object affects subextraction (Chomsky 1973; Fiengo & Higginbotham 1987; Diesing 1992; Simonenko 2016; Huang 2022, among many others). In this study, we take a closer look at another factor: the choice of the main verb that selects the NP.[1] In the context of NPs headed by content and representational nouns, it has been often observed that verb choice can affect the acceptability of subextraction, regardless of whether the NP is definite (1). However, exactly how verbs come to have such an effect is still a matter of debate.

(1) a. What did John {see/*destroy} [a picture of __]?
b. What did John {write/*destroy} [that book about]?

Our contribution is to experimentally evaluate three hypotheses about the role of the main verb: collocational frequency (Müller et al. 2022), semantic relatedness, and verb class (verbs of creation or conception, see Davies & Dubinsky 2003, Lim 2022, also Erteschik-Shir 1981, Shen and Huang 2023). We should also be upfront that the three hypotheses do not cover the entire empirical or theoretical landscape. For one, as we elaborate in our review of these hypotheses in Section 2, they do not account for exactly the same subextraction phenomena. Our primary goal here is to evaluate them on their own terms, rather than to compare them against each other. In addition, there are other hypotheses on subextraction that we will not be discussing in this paper. One such hypothesis is information structure, appealing to notions like dominance, backgroundedness, or relevance (Erteschik-Shir 1973, 1981, Goldberg 2006, Chaves and Putnam 2020). While there are recent experimental studies evaluating this hypothesis (e.g. Cuneo and Goldberg 2023), they use tests that are much more suitable for measuring the information structure properties of embedded clauses, and are therefore more relevant for the study of extraction from complement clauses. And while there are tests for measuring "dominance" of NPs (e.g. Erteschik-Shir 1973, 1981), as far as we can tell, it is not apparent from the literature whether these provide good measures of backgroundedness. As a result, in the context of subextraction from NPs, it is not clear how one could fairly evaluate the backgroundedness theory, perhaps the most prominent and worked-out information structure-based theory in the recent literature.

We test our three hypotheses using 300 verb-noun pairs in English, using formal experiments to collect acceptability judgments for subextraction from indefinite and definite NPs. Our approach, presented in Section 3, contributes to and complements existing work in a number of ways. Existing accounts of verb choice and subextraction have typically relied on informal acceptability judgments and a small number of verbs to provide support for a hypothesis. In contrast, our formal acceptability judgment experiments provide quantitative measures of acceptability, which is important because the hypotheses examined here might make predictions about gradience in acceptability (see also Lim 2022). In addition, while our set of verb-noun pairs is certainly not exhaustive, it is much larger and arguably

Our thanks to Mohamed Firdaus b. Mohd Moner, Joel Tan, Spencer Lim, and Wong Zi Shu for research assistance. Author contributions: NH, ZS: conceptualisation, design, writing; NH: data collection and analysis. This project was supported by a Humanities and Social Science Seed Fund grant from NUS.

[1] For ease of reference, we will describe such nominal projections as NPs throughout the paper, rather than DPs, as one might under the DP Hypothesis (e.g. Abney 1987, Szabolcsi 1994).

more representative than the ones used in existing studies.

To preview the results in Section 4, the best-performing hypothesis in our analysis is the creation (conception) verb hypothesis. However, we find that all three hypotheses offer relatively weak accounts of subextraction, whether for indefinite or definite NPs. This finding echoes results reported in large-scale studies for wh-extraction and elsewhere (Huang et al. 2022, White & Rawlins 2018, White 2021), but more importantly, suggests that there is room for improvement for our theories about subextraction.

2. Hypotheses

In this section, we review in more detail the three hypotheses about subextraction. For scope reasons, we will restrict our discussion of subextraction from NPs headed by content nouns or representational nouns, rather than nouns that are nominalizations of verbs (e.g. *election*, *purchase*) or denote roles (*governor*, *mother*).

2.1. Collocational frequency

In this account, proposed by Müller et al. (2022), subextraction from indefinite NPs is sensitive to whether the verb and the head noun of the NP forms a "natural predicate." They suggest that collocational frequency is the main driver of whether a verb-noun pair is perceived as a natural predicate in a language. To support this hypothesis, they conducted a study of 60 verb-noun pairs (5 nouns and 12 verbs) in German, calculating several collocational frequency measures for each of these verb-noun pairs using a German corpus, and show that collocational frequency is correlated with informal judgements of subextraction.

Müller et al. incorporate this hypothesis in a Harmonic Grammar framework: due to differences in collocational strength, two sentences that are structurally identical can differ in their degree of well-formedness. Of course, it is also possible to assume a more standard analysis, in which structurally-identical sentences are either well-formed or ill-formed. In this analysis, subextraction would always be well-formed but vary in ease of processing: perhaps sentences featuring less frequent verb-noun pairs are harder to process.

It is important to note that this Müller et al.'s account is based entirely on subextraction from indefinite NPs. While they do not address the issue of subextraction from definite NPs, it seems reasonable to assume that their account is not intended to cover such cases, given the general consensus that subextraction is much less compatible with definite NPs than indefinite NPs (Chomsky 1973, Fiengo & Higginbotham 1981, Simonenko 2016, among many others).

2.2. Verbs of creation (or conception)

Another hypothesis, first proposed by Davies and Dubinsky (2003), claims that subextraction from definite NPs is generally unacceptable unless the main verb has a verb of creation semantics, citing examples similar to (1b) above (see also Erteschik-Shir 1981 for very similar observations). In more recent work, Lim (2022) proposes a modification, suggesting that the relevant semantic property is not creation but the more specific notion of conception. Motivating this proposal are experimental results showing that not all creation verbs make subextraction from definite NPs acceptable; *What did Sally develop her picture of* is worse than *What did Sally snap her picture of*, even though both *develop* and *snap* involve creation. Lim suggests that in the case of *develop her picture*, the picture already existed prior to the event of developing it, while in the case of *snap her picture*, the snapping event is an event of conception that brought the picture into existence. Lim further introduced a conception test to determine whether a verb has conception semantics, and presented experimental results showing a correlation between conception semantics and subextraction acceptability for 16 verb-noun pairs in

English.

Beyond experimental support, Lim (2022) offers an explanation of why creation/conception semantics should affect subextraction, by adapting Truswell's (2007) Single Event Condition. Briefly, Truswell's condition predicts that island constraints are obviated when the event described in the island and the event described in the main verb are construed as a larger event grouping. Similarly, in the case of subextraction, the event described by a verb of conception and the existence of the object denoted by the NP are both construed as a larger event.

It is worth emphasizing again that both Davies and Dubinsky's and Lim's proposals were made in the context of subextraction from definite NPs. As far as we can tell, both proposals assume that definite NP objects are islands, except when selected by a creation/conception verb. This leaves open the question as to how and why the choice of main verbs affects subextraction from indefinite NPs.

2.3. Semantic relatedness

Finally, we consider a third possibility that subextraction from NPs in general is sensitive to how much the verb and the head noun are semantically related. For illustration, consider the examples in (1): *see* is more related than *destroy* to *picture*, in the sense that pictures are by their nature related to visual perception rather than destruction. Similarly, *write* is more related than *destroy* to *book*, since books by definition have to be created through writing.

To the best of our knowledge, this hypothesis has not been put forward in the literature, but it is worth articulating for the following reasons. First, this hypothesis can be seen as a variant of Müller et al.'s idea of "natural predicate," except that here, this notion is not dependent on collocational frequency. If it turns out that collocational frequency is a poor predictor of subextraction probabilities, adopting this hypothesis could serve to salvage the "natural predicate" account. Second, this hypothesis can be seen as a generalization of the creation/conception verb hypothesis, which attribute subextraction acceptability to one particular kind of semantic relatedness. Furthermore, because semantic relatedness can be gradient, this hypothesis could potentially offer an account of observations of gradience in acceptability of subextraction, e.g. as reported by Lim (2022).

3. Evaluating the hypotheses

Despite their differences, the three hypotheses make clear predictions that frequency or semantic properties should correlate with acceptability for certain kinds of subextraction. Our goal here is to evaluate these predictions: ideally, a good hypothesis should produce a strong correlation, at least for the subextraction domain that it is intended for: e.g. indefinite NPs for the collocation hypothesis and definite NPs for the creation (and conception) verb hypothesis.

3.1. Creating verb-noun pairs

To ensure comparability and representativeness, we wanted to evaluate the three hypotheses against one single set of verb-noun pairs that was relatively large. To that end, we compiled a set of 10 high-frequency nouns that have some kind of content semantics (*footage, image, map, photo, picture, portrait, review, statue, story, video*). For each noun, we parsed a random 20% subset of the Corpus of Contemporary American English (COCA, Davies 2020), identifying which verbs that had NP objects headed by these nouns. From this set of verbs, we selected 30 verbs for each noun to create a set of 300 verb-noun pairs. The process was pseudo-random, in that the selection was weighted by frequency, so that high-frequency verbs are more likely to be selected than low-frequency verbs, which are associated with typographical errors, mis-parses, and/or have metaphorical uses.

3.2. Measuring the acceptability of subextraction

3.2.1. Design

We ran an acceptability judgment experiment for all 300 verb-noun pairs. Our experiment used a 2x2 factorial design crossing subextraction and definiteness (specifically, demonstrative *that*) (2).

(2) a. Did Ben view a statue of Picasso? (No subextraction, Indefinite)
b. Who did Ben view a statue of? (Subextraction, Indefinite)
c. Did Ben view that statue of Picasso? (No subextraction, Definite)
d. Who did Ben view that statue of? (Subextraction, Definite)
(Note: we do not report our own judgments for these example sentences, which are intended as examples of the sentences to be judged by experiment participants.)

We use the demonstrative *that* instead of definite *the* because *the* is in principle ambiguous, having both an anaphoric and unique reading, which potentially introduces a confound. As Simonenko (2016) observes, while subextraction from anaphoric definites is generally unacceptable, subextraction from unique definites is judged to be better. We could have also used a possessor like *his* or *her* in place of *that*, but opted not to do so because of concerns over how the possessor would be interpreted and overall sentence plausibility. For instance, in a sentence like (2a) *Did Ben view his statue of Picasso*, it is unclear whether *his* refers to Ben or someone else and quite likely that Ben (or some other male individual) owns or created the statue. This in turn might raise questions about why (2a) would be uttered in the first place: if Ben owns or created the statue, presumably he must have also viewed it. In contrast, the demonstrative *that* avoids these confounds. More importantly, Davies and Dubinsky (2003) have also observed that subextraction from *that*-NP objects is sensitive to the choice of main verb, as (1b) illustrates.

These 4 conditions let us calculate two measures: a difference score (D score) between the two indefinite conditions (3a), as well as a difference-in-difference score (DD score) from all four conditions (3b). The D score quantifies how much less acceptable subextraction from indefinite NPs is compared with a yes/no question baseline: note that there is an implicit consensus in the literature that yes/no questions are not sensitive to the choice of main verbs. We follow Shen and Huang 2023 in using the DD score to quantify the acceptability of subextraction from definite NPs relative to an indefinite NP baseline. This captures the intuition that it is generally worse to subextract from a definite NP than from an indefinite NP (a definite island effect; see Neal and Dillon 2021, Shen and Lim 2022), setting aside the potential amelioration due to the choice of main verb, which is the phenomenon of interest here.

(3) a. Difference score (D score) for subextraction from indefinite objects = 2a−2b
b. DD score for subextraction from definite objects = (2c−2d)−(2a−2b)
(The higher the scores, the worse subextraction is from (in)definite NPs)

3.2.2. Materials and presentation

Because collecting acceptability judgments for all verb-noun pairs from a single participant would certainly incur fatigue and affect judgment quality, we decided to have each participant give ratings for only three verb-noun pairs in a survey. More specifically, we sorted the verbs into 100 different sets of three pairs each. For each set of three verb-noun pairs, we created 12 lexical frames per pair, with four variants for each of these frames, one per condition; (2) illustrates the four variants of one frame for the *view-statue* pair). Altogether this yielded 144 target sentences per set.

We then distributed these 144 target sentences into 12 different surveys using a Latin Square design, 12 sentences per survey. Because we wanted to be able to calculate D scores and DD scores for each verb-noun pair at a participant level, the 12 target sentences were distributed so that in each survey,

each of the three verb-noun pairs appeared four times, once per condition, and no lexical frames were repeated. Altogether, we created 1,200 different surveys.

We also added to the 12 target sentences in each survey a common set of 24 filler items. These fillers were taken and adapted from fillers used by Huang et al. (2022). 9 of these fillers appeared at the start in a fixed order; these were intended to prompt participants to use the full range of the acceptability scale. The remaining 15 fillers, also intended to span the full range of the acceptability scale, and the target sentences were then presented pseudo-randomly.

The surveys were hosted on PCIbex (Zehr and Schwarz 2018). Sentences were presented one at a time on the screen. Participants were to rate each sentence for acceptability using a slider scale provided on the screen. Prior to starting the surveys, participants first saw three example sentences with suggested acceptability ratings; these three sentences were completely unacceptable, of marginal acceptability, and completely acceptable (in that order). Participants were also asked whether they lived in the United States from birth until at least age 13, whether their parents spoke to them in English at home, and which state they grew up in.

3.2.3. Participants

We recruited 3,583 participants via the Prolific crowdsourcing platform, with the goal of recruiting about 36 participants per verb-noun pair (recall we had distributed our 300 verb-noun pairs into 100 sets of surveys), so that each verb-noun pair would be associated with D scores and DD scores for 36 native speakers. Participants were self-identified monolingual speakers of American English, born in the United States, and above the age of 21. Each participant received GBP 0.75 for completing the survey, based on the assumption that it would take about 5 minutes to complete the survey, and a GBP 9.00 hourly rate, as recommended by Prolific.

3.2.4. Data analysis

Acceptability judgments were z-scored at the participant level to control for differences in how each participant used the slider scale. For each of the 15 fillers that appeared with the target sentences, we checked each participant's judgment against the sample mean for that filler sentence, and counted the number of "extreme" judgments that were two standard deviations above or below the mean. We only included participants who gave at most 3 "extreme" judgments.

We imposed a few other filters on our participants. We selected only participants who answered that they lived in the United Sates from birth and their parents spoke to them in English at home. We also analyzed each participant's responses, selecting only participants whose median response time was at least 2 seconds, which provides some confidence that they had read each sentence before judging it for acceptability. Altogether, after applying these filters, we selected 2,264 participants' responses for analysis.

3.3. Compiling predictor measures

We next describe how we compiled predictor measures for each of the hypotheses under consideration.

3.3.1. Collocational frequency

We calculate three measures, DeltaP, Mutual Information, and t-scores, closely following Müller et al.'s (2022) analysis. We first split the Corpus of Contemporary American English (COCA) into sentences. Within each sentence, we identified all the nouns (based on the part-of-speech tags that come with COCA), and then checked whether there was a verb in the three words preceding the noun; if so, we consider that that noun to be the head noun of the verb's object. We recorded down all verbs and nouns in this configuration. We note that this three-word method, while adhering to Müller et al.'s

analysis, meant that we could not obtain estimates for eight verb-noun pairs. For six of these pairs, this was because the verbs are associated with a particle that could appear either before or after the NP (*pull up, beam down, black down, check out, clean up, cut out*); for the remaining two (*email-image, contribute-footage*), we believe that this is because the verb and noun appeared more than three words apart from each other.

With counts of verbs and nouns, we calculate the three collocational frequency measures for a verb-noun pair v, n using the formulas in (4). Read N(…) as "the total count of …" and "v occurring with n" as short for "verb-object relations between v and n."

(4) a. $\Delta P_{v|n}$ = (N(v occurring with n) ÷ N(all verbs occurring with n))
- (N(v occurring with all other nouns) ÷ N(all verbs occurring with all other nouns))[2]

b. Mutual information = N(v occurring with n) ÷ Expected count(v occurring with n)
where
Expected count(v occurring with n) = N(v occurring with all other nouns)
× N(all other verbs occurring with n)
÷ N(all verb-object relations)

c. t-score = (N(v occurring with n) – Expected count(v occurring with n))
÷ √N(v occurring with n)

Intuitively, as Müller et al. point out, $\Delta P_{v|n}$ measures how well the head noun of an object predicts the verb that selects it, compared to all other head nouns. Mutual Information (MI) measures how much more likely a verb and noun will be in a verb-object relation relative to chance, while the t-score serves to highlight the frequency of the co-occurrence of the verb and noun (see Gablasova et al. 2017 for a critique).

As Müller et al. also noted, these measures are based on counts obtained through this closeness heuristic, but the heuristic only gives an approximation of how often a given verb-noun pair appears in a verb-object configuration. Ideally, one would have first parsed each sentence to identify the verbs and noun occurring in a verb-object configuration, but doing so by hand is unfeasible, while using a statistical parser is computationally intensive (given the size of COCA). Furthermore, since statistical parsers are not perfectly accurate, using them would introduce a risk of misparsing that seems no worse than this closeness heuristic.

3.3.2. Creation/conception semantics

To determine whether the verbs have creation or conception semantics, we trained three undergraduate research assistants (RAs), without revealing to them the goal of the study. The RAs were presented with one example sentence per verb-noun pair based on those used in the acceptability judgment surveys. They were instructed that a verb is a creation verb if it creates an entity described by the noun; the entity can be a copy or a more abstract item, and that a verb is a conception verb if the action denoted by the verb makes the object denoted by the noun come into existence. RAs also saw two examples of Lim's (2022) conception test (5), to help them pick out conception verbs from the larger set of creation verbs.

(5) Jo {took / printed} a photo of the mountain. Did Jo's photo exist before she printed it?
If answer is "yes": *take* (*print*) is not a verb of conception.
If answer is "no": *take* (*print*) is a verb of conception.

[2] There is an alternative ΔP measure, $\Delta P_{n|v}$, that we do not use here. Müller et al. (2022: 1631) suggest that $\Delta P_{v|n}$ is the better predictor.

(Note: while *take-photo* and *print-photo* are in our list of 300 verb-noun pairs, these two examples did not specify whether *take* or *print* are creation verbs or conception verbs.)

The RAs annotated the set of verbs independently. They were further instructed to take frequent breaks, to minimize the risk of fatigue. We then compiled all three sets of annotations. We classified a verb as a creation (conception) verb only if at least a majority of RAs (two out of three) considered it as such. For statistical analysis purposes, we coded a creation (or conception) verb as 1 and all other verbs as 0.

3.3.3. Semantic relatedness

For this hypothesis, we relied on word embeddings, calculating the cosine similarity between the vector representations of a given verb and noun. We note that this approach is a commonly used one for calculating whether two words are semantically related, and has the advantage of producing gradient measures ranging from -1, which is interpreted as involving completely opposite meanings, to 1, which is interpreted as involving highly similar meanings.

We used two publicly-available word embedding data sets (Fares et al. 2017), created by applying GloVe (Pennington et al. 2014), an unsupervised learning algorithm, on a 2017 version of English Wikipedia and the 5th edition of the GigaWord corpus. Note that a small number of these verbs – e.g. those associated with a particle, like *pull up* – are absent in these word embedding data sets. For each of the verb-noun pairs present in the data sets, we calculated the cosine similarity between the verb and noun. This produced two sets of similarity measures, one based on Wikipedia and the other based on the GigaWord corpus, for 286 verb-noun pairs.

4. Results

To summarise, we calculated two different measures of z-scored subextraction acceptability: a D score for indefinite NPs and a DD score for definite NPs; and a total of seven predictor measures: three for collocational frequency, two for creation/conception semantics, and two for semantic relatedness. We calculate Pearson correlations for all combinations of subextraction acceptability and predictors, even though existing hypotheses in the literature – the collocational frequencies and the creation/conception verb hypotheses – only have clear predictions for either definite or indefinite NPs. To maximize comparability, we analysed only the 284 verb-noun pairs where we have values for all seven predictors.

We take this comprehensive approach because it is logically possible that the predictors might turn out to cover both indefinite and definite NPs: for example, perhaps collocational frequencies predict not only D scores (for subextraction from indefinite NPs), as proposed by Müller et al., but also DD scores, which measure subextraction acceptability for definite NPs. Although the hypotheses, as currently formulated, do not necessarily cover both types of NPs, we present these results here in the spirit of transparency and to encourage future research on this topic.

We should also point out that it is logically possible that some correlations, even if statistically significant, are relatively small and so have a very limited role to play in our theories of subextraction. To identify such correlations, we note that each correlation corresponds to a linear regression. We then calculate the R^2 of this regression model (in this case mathematically equivalent to the square of the Pearson correlation), which indicates the amount of variation in D (or DD) scores explained by the predictor. We also use the bayestestR package (Makowski et al. 2019) to calculate a Bayes factor for this model relative to a null hypothesis model that lacks the predictor, i.e. a model that only has an intercept. The Bayes factor here is a ratio indicating how much more likely the data is under this model compared with the null hypothesis model. The lower the ratio, the stronger the evidence is for the null hypothesis.

4.1. Subextraction from indefinite NPs

We first consider results for indefinite NPs. Pearson correlations and Bayes factor values are presented in Table 1. To help visualize the size of the correlations, scatterplots for selected predictors are presented in Figure 1.

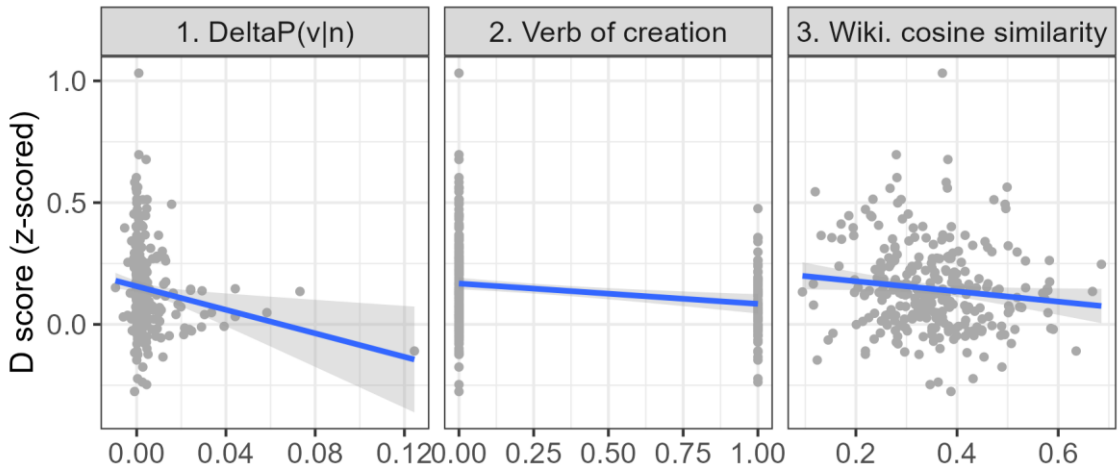
Table 1

Subextraction from indefinite NPs (D scores): correlation, R^2s, and Bayes factors for various predictors

	Pearson *r*	*p*	R^2	Bayes factor
Frequency: $\Delta P_{v\|n}$	-.16	.009	.02	2
Frequency: Mutual Information	-.09	.140	.01	0
Frequency: t-score	-.15	.011	.02	2
Creation verbs (1=creation verb, 0=others)	-.21	<.001	.04	32
Conception verbs (1=conception verb, 0=others)	-.21	<.001	.04	28
Semantic relatedness: Wikipedia cosine similarity	-.12	.042	.01	0
Semantic relatedness: GigaWord cosine similarity	-.10	.097	.01	0

Figure 1

Scatterplots of D scores with one selected predictor per hypothesis



The predictors that perhaps deserve the most attention here are those for the collocational frequency hypothesis, since that has been argued to account for variation in subextraction acceptability in German indefinite NPs. Here, we expect a negative correlation: high D scores (unacceptable subextraction) should be associated with low collocational frequencies. An examination of correlations, Bayes factors, and scatterplots, however, suggest that collocational frequencies are poor predictors of the variation in D scores in English. Although all three correlations are in the right direction, the correlation coefficients are all small (Pearson *r* -.09 to -.16); in fact, the correlation is not significant for Mutual Information. R^2 values are similarly small, around .01-.02, implying that these predictors explain only about 1-2% of all variation in D scores. Bayes factors are also small, at around 0-2. We note that this is below the ratio of 3 that is often suggested as indicating clear evidence against the null hypothesis.

Other predictors also show negative correlations, which are not implausible. For instance, for the creation/conception verb hypothesis, a negative correlation implies that creation or conception verbs tend to have low D scores, i.e. be associated with more acceptable subextraction. Likewise, for the

semantic relatedness hypothesis, a negative correlation implies that verb-noun pairs that are highly related semantically (high cosine similarity) have low D scores. Interestingly, we see that creation/conception verb predictors turn out to be the best-performing predictors, even though the creation/conception verb hypothesis makes no prediction about subextraction from indefinite NPs. These predictors produce the strongest correlations (and hence relatively high R^2s, although these are still low in absolute terms) and the highest Bayes factors (28-32). We return to this issue in the general discussion in Section 5.

4.2. Subextraction from definite NPs

We next consider results for definite NPs. Correlations with DD scores, indicating the relative acceptability of subextraction from these NPs, and Bayes factors are presented in Table 2, and scatterplots are presented in Figure 2.

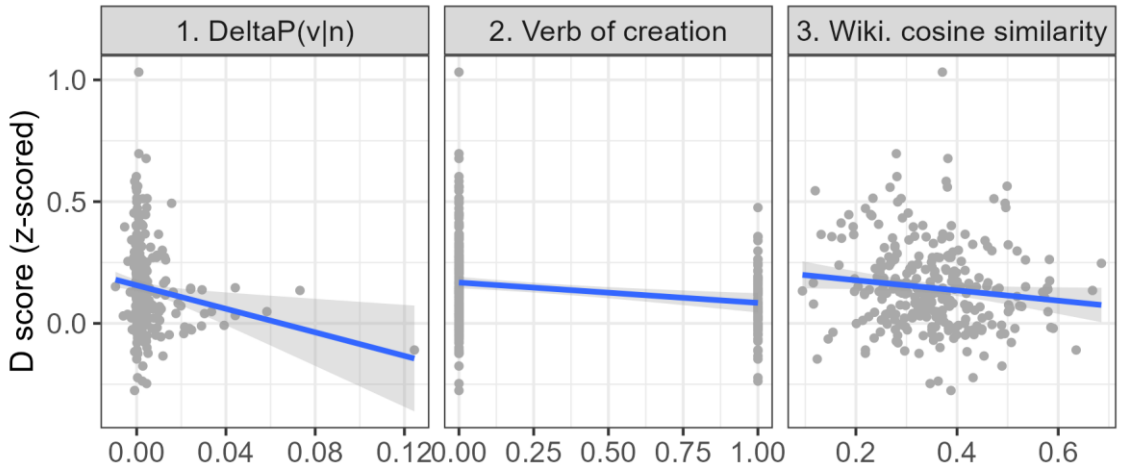
Table 2

Subextraction from definite NPs (DD scores): correlation, R^2s, and Bayes factors for various predictors

	Pearson *r*	*p*	R^2	Bayes factor
Frequency: ΔP(v|n)	-.11	.071	.01	0
Frequency: Mutual Information	-.26	<.001	.07	1,224
Frequency: t-score	-.07	.21	.01	0
Creation verbs (1=creation verb, 0=others)	-.36	<.001	.13	>100,000
Conception verbs (1=conception verb, 0=others)	-.33	<.001	.11	>100,000
Semantic relatedness: Wikipedia cosine similarity	-.13	.030	.02	1
Semantic relatedness: GigaWord cosine similarity	-.07	.235	.01	0

Figure 2

Scatterplots of DD scores with one selected predictor per hypothesis



Here, the predictors for interest are the creation verb and conception verb predictors, as they have been linked in the literature to acceptable subextraction from definite NPs. We expect a negative correlation: high DD scores (unacceptable subextraction) should be correlated with noncreation or nonconception verbs, which are coded as 0 in our data. Results are consistent with this prediction: we observe statistically significant negative correlations for both predictors, with stronger correlations (Pearson *r* -.33 and -.36) and much higher R^2s and Bayes factors than in the indefinite NP cases. But as

the scatterplots show, even though there is a much clearer correlation, there is nonetheless a lot of variation in DD scores within each verb class.

Turning to the other predictors, we also see negative correlations in general, which are not implausible. These correlations mean that unacceptable subextraction from definite NPs (high DD scores) is associated with low collocational frequency and low semantic relatedness. However, the size of correlations here are much smaller, around -.1, and Bayes factors are generally far lower, suggesting that these predictors account for very little, if any, of the variation in DD scores. The one exception to this pattern is the Mutual Information predictor, whose correlation, R^2s and Bayes factors are comparable to those for both creation/conception verb predictors.

5. Discussion

Our results indicate that collocational frequency is not a good predictor of the acceptability of subextraction from indefinite NPs, as measured through D scores (*pace* Müller et al), even though at least one measure (Mutual Information) is a relatively good predictor of DD scores, which are intended to reflect the acceptability of subextraction from definite NPs. Our results also show that general semantic relatedness is not a good predictor of either case of subextraction.

The results are more favorable for the creation/conception verb hypothesis: verb class predictors produce statistically significant correlations with DD scores (subextraction from definite NPs) in the predicted direction, corroborating informal observations by Davies & Dubinsky (2003) and experimental results reported by Lim (2022).

Verb class predictors also show a significant negative correlation with D scores, around -.2. This is actually not a problem for existing statements of this hypothesis, which does not make clear predictions about whether verb semantics matter for subextraction from indefinite NPs. But to the extent that creation/conception verb semantics do matter in this case, we believe it will be challenging to extend whatever account developed for DD scores for subextraction from definite NPs (such as Lim's adaptation of the Single Event Condition) to also cover D scores. This is because intuitively, DD scores are defined as the impact on acceptability of subextraction from definite NPs above and beyond the impact of acceptability of subextraction from indefinite NPs. An account for DD scores by definition should explain the difference between the two kinds of subextraction, but logically that will not guarantee an explanation for facts around subextraction from indefinite NPs.

Another point that is worth highlighting is how our results seem much weaker than what has been suggested in previous work. We will have to leave for future investigation the exact reasons for this discrepancy between our results and previous work. However, it seems not implausible that one reason might be the relatively small number of verb-noun pairs studied: for instance, Müller et al. (2021) looked at 60 verb-noun pairs, while Lim (2022) considered 16 verb-noun pairs. Whether in German or English, the actual number of transitive verbs that take NP objects with content or representational head nouns is certainly much larger. Even with best efforts, it would have been very difficult to ensure a sample of 16-60 verb-noun pairs that is representative of the entire population. Our 300 verb-noun pairs helps to address this methodological issue by providing a larger set of data that is hopefully more representative.

6. Conclusion

In this study, we evaluated three hypotheses about subextraction from NP objects with large scale experiments in English. Among the three hypotheses – collocational frequency, creation (conception) verbs, and semantic similarity – the creation (conception) verbs performed the best, in producing a correlation with DD scores (reflecting subextraction acceptability for definite objects) in the predicted direction, along with a high Bayes factor suggesting clear evidence against the null hypothesis.

However, the R^2 for even this hypothesis is low, indicating that it does not offer a full account of variation in DD scores. We further pointed out that these results appear worse than what was reported previously on these hypotheses, and suggested that this might be related to the smaller samples used in previous work, which might not have been as representative as intended.

Consistent with our speculation about data representativeness, the lack of clear results in favor of existing hypotheses seem to be typical of large-scale experimental studies. In a similarly large-scale study of how clause-embedding verbs affect long-distance wh-dependencies, Huang et al. (2022) also found weak support for existing accounts, including those positing a link between extraction and frequency, semantic similarity, and information structure. Outside of wh-dependencies, similarly weak results have been obtained in large-scale studies testing claims about the selection of interrogative and declarative clauses in attitude verbs (White 2021 and White and Rawlins 2018).

Finally, it is also important to keep in mind that we have not evaluated information structure theories in this paper. It is possible that, while all three hypotheses here provide at best a weak account for variation in subextraction acceptability, information structure can deliver much better empirical coverage for both indefinite and definite NPs. We will set this empirical question aside, noting again that it is not immediately clear how to use existing tests to measure the currently theoretically-important notion of backgroundedness in the context of subextraction from NPs.

For the time being, focusing only on the results reported above, we argue that these results indicate room for improvement for theories about how main verbs affect subextraction. For instance, perhaps definitions of verb of creation/conception could be refined further, in order to better account for the variation of DD scores within the class of creation (conception) verbs and noncreation (nonconception) verbs; perhaps the operationalization of collocation has to be reconsidered: maybe instead of counting individual verb-noun pairs, it might be helpful to consider combinations of verb classes and noun classes. Or given that single-factor hypotheses are relatively weak, perhaps it would be fruitful to explore multifactorial hypotheses, e.g. modeling variation in subextraction with both collocational frequency and verb class.

References

Abney, Steven P. 1987. The English noun phrase in its sentential aspect. Doctoral dissertation, MIT.

Chaves, Rui P., and Michael T. Putnam. 2020. *Unbounded dependency constructions: Theoretical and experimental perspectives*. Oxford: Oxford University Press.

Chomsky, Noam. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, ed. by Stephen Anderson and Paul Kiparsky, 232–286. New York: Holt, Rinehart & Winston.

Cuneo, Nicole, and Adele E. Goldberg. 2023. The discourse functions of grammatical constructions explain an enduring syntactic puzzle. *Cognition* 240: 105563.

Davies, Mark. 2020. The Corpus of Contemporary American English (March 2020). Online: https://www.english-corpora.org/coca/ (accessed 2023).

Davies, William D., and Stanley Dubinsky. 2003. On extraction from NPs. *Natural Language and Linguistic Theory* 21:1–37.

Diesing, Molly. 1992. Bare plural subjects and the derivation of logical representations. *Linguistic Inquiry* 23: 353–380

Erteschik-Shir, Nomi. 1973. On the nature of island constraints. Doctoral dissertation, MIT.

Erteschik-Shir, Nomi. 1981. On extraction from noun phrases (picture noun phrases). In *Theory of Markedness in Generative Grammar: Proceedings of the 1979 GLOW Conference*, ed. by Andrea Belletti, Luciana Brandi, and Luigi Rizzi (eds.), 147-69, Scuola Normale Superiore di Pisa, Pisa.

Fares, Murhaf, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. *Proceedings of the 21st Nordic Conference on Computational Linguistics*, 271-276.
Fiengo, Robert, and James Higginbotham. 1981. Opacity in NP. *Linguistic Analysis* 7: 395–421.
Gablasova, Dana, Vaclav Brezina, and Tony McEnery. 2017. Collocations in corpus-based language learning research: identifying, comparing, and interpreting the evidence. *Language Learning* 67:155-179.
Goldberg, Adele. 2006. *Constructions at work*. Oxford: Oxford University Press.
Huang, Nick, Diogo Almeida, and Jon Sprouse. 2022. How good are leading theories of bridge verbs? An experimental evaluation. Talk given at WCCFL 40, Stanford University.
Huang, Nick. 2022. How subjects and possessors can obviate phasehood. *Linguistic Inquiry* 53:427-458.
Lim, Meghan. 2022. The bridging effects of verbs of creation: An experimental look: National University of Singapore MA thesis.
Makowski, Dominique, Mattan S. Ben-Shachar, and Daniel Lüdecke. 2019. bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software* 40, 1541.
Müller, Gereon, Johannes Englisch, and Andreas Opitz. 2022. Extraction from NP, frequency, and minimalist gradient harmonic grammar. *Linguistics* 60: 1619-1662.
Neal, Anissa, and Brian Dillon. 2021. Definitely islands? An investigation into the offline and online status of definite islands. Poster presented at the annual meeting of Linguistic Society of America 2021.
Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532–1543.
Shen, Zheng & Meghan Lim. Accepted. The definite NP island in wh-questions and relative clauses. *Syntax*.
Shen, Zheng, and Nick Huang. 2023. The role of phases and specificity in definite islands. National University of Singapore manuscript. Available at https://ling.auf.net/lingbuzz/007746.
Simonenko, Alexandra. 2016. Semantics of NP islands: the case of questions. *Journal of Semantics* 33: 661–702.
Szabolcsi, Anna. 1994. The noun phrase. In *The syntactic structure of Hungarian*, ed. by Ferenc Kiefer and Katalin É. Kiss, 179–275. San Diego, CA: Academic Press.
Truswell, Robert. 2007. Extraction from adjuncts and the structure of events. *Lingua* 117. 1355–1377.
White, Aaron Steven, and Kyle Rawlins. 2018. The role of veridicality and factivity in clause selection. In Sherry Hucklebridge & Max Nelson (eds.), *Proceedings of the 48th Annual Meeting of the North East Linguistic Society*, 221–234. Amherst, MA: GLSA Publications.
White, Aaron Steven. 2021. On believing and hoping *whether*. *Semantics and Pragmatics* 14: 1–18.
Zehr, Jeremy, and Florian Schwarz. 2018. *PennController for Internet Based Experiments (IBEX)*. https://doi.org/10.17605/OSF.IO/MD832.23

glow

Proceedings of

The 14th Generative Linguistics in The Old World in Asia (GLOW in Asia XIV)

Edited by Xiangyu Li, Zetao Xu, Yuqiao Du, Zhuo Chen, Chenghao Hu, Zhongyang Yu & Victor Junnan Pan

March 6-8, 2024

The Chinese University of Hong Kong

Hosted by

Department of Linguistics and Modern Languages,
The Chinese University of Hong Kong

Sponsored by

The Chinese University of Hong Kong
Faculty of Arts
Department of Linguistics and Modern Languages
New Asia College

Sponsors

香港中文大學
The Chinese University of Hong Kong

文學院

Department of Linguistics and Modern Languages
語言學及現代語言系

香港中文大學新亞書院
New Asia College, The Chinese University of Hong Kong

GLOW in Asia XIV Organizing Committee

Victor Junnan Pan
Zhuo Chen
Yuqiao Du
Chenghao Hu
Xiangyu Li
Zetao Xu
Zhongyang Yu

Acknowledgement

This conference could not have been possible without the participation and assistance of so many people, especially our anonymous reviewers, whose names may not be enumerated here. Their contributions are sincerely appreciated and gratefully acknowledged. Also, the organizers of GLOW in Asia XIV would like to express our special gratitude to the session chairs (not in any particular order): Kwang-sup Kim (Hankun University of Foreign Studies), Zhuo Chen (Chinese University of Hong Kong), Haihua Pan (Chinese University of Hong Kong), Gladys Wai-lan Tang (Chinese University of Hong Kong), Yoichi Miyamoto (Osaka University), Toru Ishii (Meiji University), Hamida Demirdache (Nantes Université/CNRS/IUF), Thomas Hun-tak Lee (Chinese University of Hong Kong), Paul Law (Chinese University of Hong Kong), Nobu Goto (Toyo University), Keiko Murasugi (Nanzan University).

Finally, we would like to give our heartful thanks to the following colleagues: Professor Mamoru Saito from Nanzan University of Notre Dame, Professor Anoop Mahajan from University of California Los Angeles, and the other GLOW in Asia executive committee members: Professor Wei-Tien Dylan Tsai, Professor Tanmoy Bhattacharya, Professor Myung-Kwan Park, Professor Koji Sugisaki and Professor Yuji Takano for their great support.

Proceedings of the 14th Generative Linguistics in the Old World in Asia (*GLOW in Asia XIV*) 2024

First published in 2024
Published by Department of Linguistics and Modern Languages,
The Chinese University of Hong Kong.