NUMBER WORDS AND THE OBJECT WIDE SCOPE PUZZLE

by

MARTA T. SUAREZ

A thesis submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Master of Science

Graduate Program in Psychology

written under the direction of

Julien Musolino, Ph.D.

and approved by

_____

_____

_____

New Brunswick, New Jersey

May 2011

ABSTRACT OF THE THESIS

Number Words and the Object Wide Scope Puzzle

by MARTA T. SUAREZ

Thesis Director:
Julien Musolino, Ph.D.

Sentences such as *Three girls are holding two balloons,* whose subjects and objects are quantified by bare numeral expressions, logically allow multiple readings. The semantics literature has reported that the so-called object wide scope distributive reading, (interpreted as having 6 girls and 2 balloons, each balloon held by 3 girls), is usually not accessible. Recent experimental studies showed the reading was accessible, albeit massively dispreferred (Musolino, 2009; Syrett & Musolino, in prep.) We report the findings of two experiments that tested competing theoretical accounts of why this reading should be disallowed. On one account, the syntactic configurations and operations required to generate the reading are not permitted by the grammar (Beghelli & Stowell, 1997). On the other, the reading is taken to be allowed by the grammar but rendered inaccessible by excessive processing costs (Reinhart, 2006). Crucially, this account involves the semantic nature of the subject; when it allows both a distributive and a collective interpretation, the computation of all possible readings exceeds working memory capacity. It straightforwardly predicts that if a collective reading can be forced by adding a modifier, e.g., *Three girls <u>together</u> are holding two balloons,* the object wide

scope reading should become acceptable. This was reported to be the case for a small number of informally consulted subjects.

Experiment 1 piloted 5 lexical items within subjects, all with singular indefinite subjects (N = 42). Results revealed that participants accepted the object wide scope reading, as predicted. However, clear item effects were found, contra the literature. Experiment 2 varied 3 types of subject noun phrase between groups: singular indefinites, bare numeral quantifiers and bare numeral quantifiers plus a collectivizing modifier. It also tested 4 lexical items from Experiment 1 within subjects (N = 132). Results revealed: 1- the object wide scope reading was acceptable to most participants, with or without modification, contra both theoretical claims and previous experimental findings; and 2- clear differences among lexical items, replicating Experiment 1. These findings suggest that participants did not treat all internal arguments equally. Rather, they were sensitive to the argument structure of the verbs, distinguishing true transitives from unaccusatives.

Acknowledgements

I would like to thank Julien Musolino for a happy and productive collaboration; Viviane Deprez for her helpful feedback and suggestions; Arnold Glass for his incisive critiques and continued encouragement; Kristen Syrett, Shigeto Kawahara and the members of the Rutgers Psycholinguistics Lab for their many questions and suggestions; and Asya Achimova and Justyna Grudzinska for countless hours of fruitful and enjoyable discussions.

Table of Contents

Lists of Tables

List of Figures

Introduction

Sentences whose subject and object noun phrases (NPs) are both quantified by bare numeral expressions allow multiple interpretations. However, not all interpretations are equally acceptable to native speakers. We report on two experiments that tested competing theoretical accounts of why a particular reading of English sentences of this type is dispreferred over others. The first account, owing to Beghelli and Stowell (1997), argued that the types of interpretation allowed by such sentences are constrained by the grammar of English in particular, and the grammar of quantification in general. The second account, owing to Reinhart (2006), argued instead that the observed pattern of interpretive preferences results from the different demands each interpretation places on cognitive processing capacity, especially working memory.

Consider sentence (1), which has bare numeral indefinite quantifiers in both its subject and object NPs. Such sentences logically allow at least 8 interpretations. This explosion of interpretations results, in part, from two interacting sources of ambiguity, quantifier scope and distributivity.

(1) Three girls are holding two balloons.

The notion of quantifier scope as used in formal semantics is familiar from algebra, where it is understood as the range over which an operator controls the arguments in an expression. For example, the value of the expression in (2) differs from that of the one in (3) by virtue of the relative scope of the operators + and ×. The parentheses determine the order of computation:

(2) $3 + (4 \times 5)$

(3) $(3 + 4) \times 5$

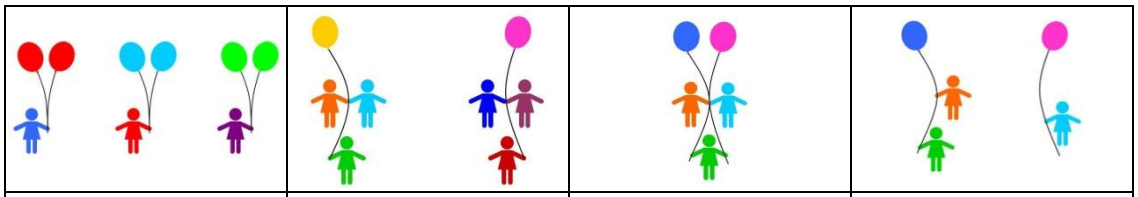
Similarly, the sentence in (1) has two distinct scoped interpretations, whose respective corresponding configurations are shown in Figure 1. These can be paraphrased as (4a) and (4b) and represented informally by the formulas in (4b) and (5b):

(4) a. Three girls are each holding two balloons.

b. 3x2y (x holding y), where x is a girl and y is a balloon

(5) a. Two balloons are such that for each balloon, three girls are holding it.

b. 2y3x (x holding y), where x is a girl and y is a balloon



(a) Subject Wide Scope	(b) Object Wide Scope	(c) Each-All	(d) Cumulative

Figure 1. Scoped and scopeless interpretations of *Three girls are holding two balloons*. The configurations in (a) and (b) depict the two scope-dependent readings. Two scopeless interpretations, which we refer to as the each-all (c) and cumulative (d) interpretations following Musolino (2009), are included for completeness. Note that the cumulative interpretation is actually a family of interpretations. The total number is a function of the number of ways the number of subjects and objects can combine.

(4) has been termed the subject wide scope reading of sentences such as (1). In this reading, corresponding to the configuration in Figure 1a, the number of girls is interpreted exactly as in the overt sentence – three. However, there are six balloons, the

two in the overt sentence multiplied by three, the number of girls. The subject NP *Three girls* is said to take scope over the object NP *two balloons*; hence subject wide scope. Most native speakers of English agree that the configuration in Figure 1a is a possible interpretation of (1). (5) represents the object wide scope reading of (1). In this reading, corresponding to the configuration in Figure 1b, the object NP *two balloons* takes scope over the subject NP *three girls.* Hence, the number of objects is interpreted as in the overt sentence, and the number of subjects is a multiple of the number of objects. The configurations in Figures 1c and 1d represent the scopeless (aka scope-independent) interpretations of (1). Note that in the scopeless cases, both the number of subject NPs and the number of object NPs is interpreted exactly as expressed in the overt sentence.

We know from both theoretical (Beghelli & Stowell, 1997; Reinhart, 2006) and experimental (Musolino, 2009; Syrett & Musolino, in prep.) studies that speakers have strong preferences for certain interpretations over others. For instance, most speakers accept the subject wide scope interpretation (Fig. 1a), but overwhelmingly reject the object wide scope interpretation of sentences with two bare numeral NPs (Fig. 1b).

The second source of ambiguity in sentences with two bare numeral NPs is distributivity. Distributivity is a property of events that (partially) defines how the event is construed. If the participants are understood to perform the action individually, the event is interpreted on its distributive reading. But if the participants are understood as performing the action as a group, the event is interpreted on its collective reading. For example, consider sentence (6) in a context where the two girls in the subject NP are named Alicia and Beth:

(6) Two girls visited Mary.

If Alicia visited Mary, then 20 minutes later Beth visited Mary, (6) is true on the distributive reading because there were two separate events of visiting Mary, one for each girl. But if Alicia and Beth met at the corner coffee shop and went to Mary's house together, (6) sentence is true on the collective reading because there was only one event of visiting Mary involving the two girls together.

The sentential object can be interpreted distributively, as well. Consider (7):

(7) Mary watered two plants.

(7) also has two interpretations. In the distributive interpretation – in this case, of the object – Mary watered each plant individually. There were two events of watering, one for each plant. In the collective reading, Mary adjusted her hose to a wide spray and watered both plants simultaneously. There was only one event of watering the two plants at the same time.

Quantifier scope interacts with the distributive/collective nature of subjects and objects to produce 8 interpretations for sentences like (1). However, as Figure 2 shows, the interaction is incomplete. Problems arise when the subject and object are both interpreted collectively (Figs. 2d and 2h). In these cases, the cardinality of the wide scope NP is interpreted as stated in the overt string, as expected. However, the cardinality of the narrow scope NP is interpreted as stated in the overt string, as well. If the interaction of scope and distributivity were fully productive, the expectation would be for the

cardinality of the narrow scope NP to be a multiple of the wide scope NP. For example, Figure 2d would depict a collection of 3 girls and a collection of 6 balloons, exactly equal to the number of balloons in Figures 2a – 2c. Likewise, Figure 1h would depict a collection of 6 girls, not 3. The collective-collective cases result in 2 readings that are indistinguishable from one another. Moreover, these are also independent of scope. In fact, Figures 2d and 2h represent the scopeless reading that Musolino (2009) refers to as each-all (aka strong symmetric (Grudzinska, 2010 citing May, 1985)).

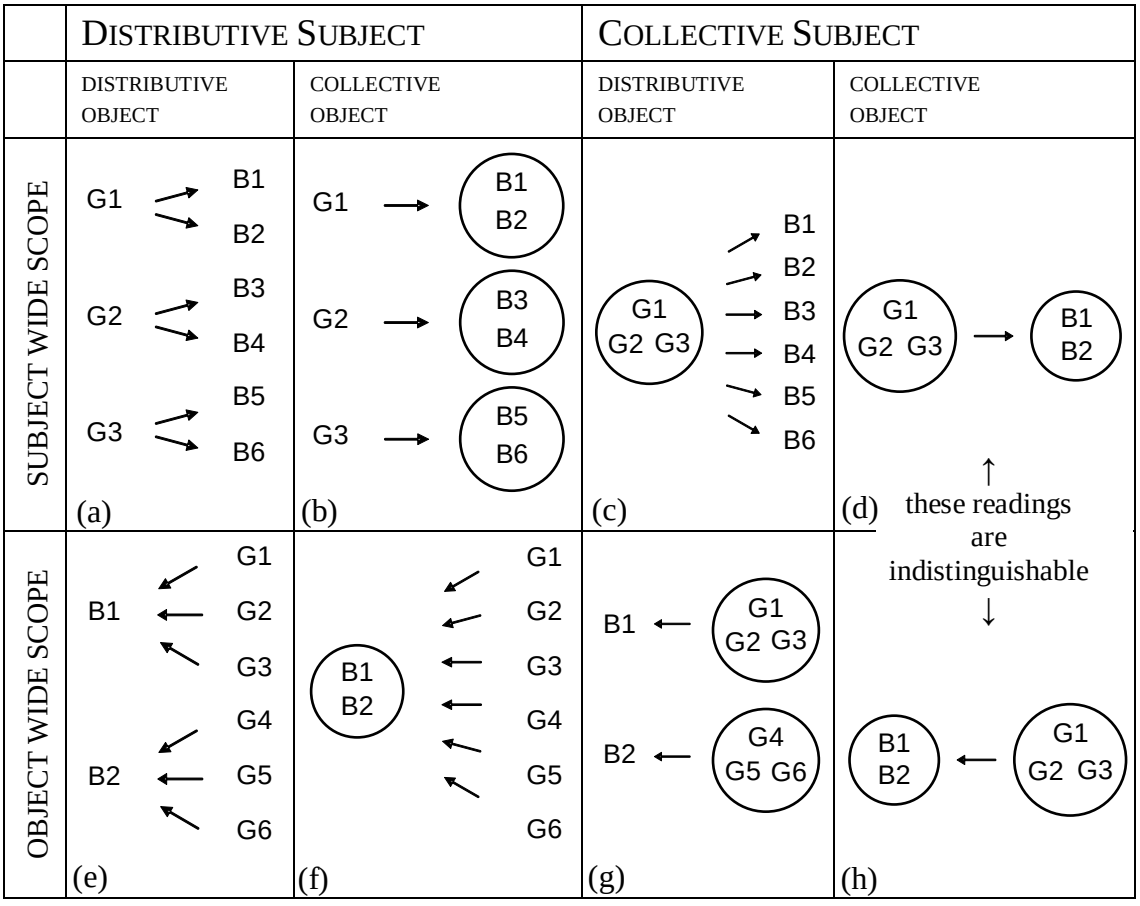


Figure 2. Map of scoped readings of *Three girls are holding two balloons.*

When the subject and object NPs are both construed as collective, the subject wide scope and object wide scope configurations are indistinguishable and result in the reading we refer to as the each-all (scopeless) reading in Figure 1. The cumulative reading is not represented here.

Because (1) has 2 sources of ambiguity – scope and distributivity – it has (at least) 8 logically possible readings: 2 that result from scope (subject wide *vs.* object wide), 4 from distributivity (distributive *vs.* collective subject times distributive *vs.* collective object – with exceptions as noted above), the each-all reading, and as many cumulative readings as there are combinations of subjects and objects. Among all these interpretations, the object wide scope is least preferred.

One possible explanation for the object wide scope interpretation's dispreferred status is that there is a global ban on the object wide scope interpretation. But that possibility can be quickly dismissed. Consider the difference between (1), repeated here as (12), and (13):

(12) Three girls are holding two balloons.

(13) Three girls are holding each balloon.

Most speakers find the configuration in Figure 1d to be a natural interpretation of (13) but they find this interpretation to be unavailable for (12).

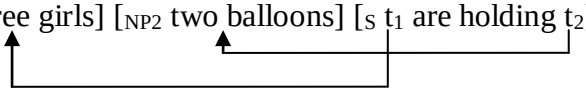
The scope-taking properties of *each* were recognized by linguists long ago (Beghelli & Stowell, 1997; May, 1977/1990), but were not tested experimentally until recently. Musolino (2009) found that the minimal change from *two* to *each* resulted in a nearly complete reversal of preferences. Note that (13) lacks a source of ambiguity inherent in (12). Its object NP can only be interpreted distributively, while the object NP of (12) can be interpreted either distributively or collectively.

The distributive reading of predicates like *hold* may not be obvious without context. But consider a situation in which two balloons must be displayed for some duration of time, say, to indicate where a party is being held. Several girls have volunteered to do the honors. They will take turns holding each balloon until all the guests arrive. In this situation, the relevant reading is clear. For an alternative in which the collective-distributive distinction is clearer, substitute a predicate that can distribute over both time and space, e.g., *Three girls are watering two plants.* Regardless of the choice of predicate, there is no context that can render *each* collective.

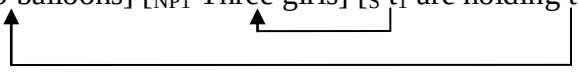
Theoretical Background

In the standard theory of quantifier scope from May (1977/1990), each reading of an structurally ambiguous sentence corresponds to a different syntactic configuration. At the level of logical form (LF), the syntactic level that forms the interface between the overt syntax and meaning, a quantified NP such as *a girl, no girl, each girl* and *three girls* obligatorily raises out of the position where it occurs in the speech stream and adjoins to the sentence, leaving behind a trace that marks its original relation to the verb. Because the quantified NPs may move in any order, this operation - called quantifier raising - yields the two scoped interpretations of (1) straightforwardly, where (14) and (15) are the syntactic analogues of the semantic representations in (4) and (5):

(14) $[_{NP1}$ Three girls$]$ $[_{NP2}$ two balloons$]$ $[_{S}$ t_1 are holding $t_2]$



(15) $[_{NP2}$ two balloons$]$ $[_{NP1}$ Three girls$]$ $[_{S}$ t_1 are holding $t_2]$



The relations that obtain among the moved noun phrases after movement at LF determine the relative scope of the quantifiers. The relevant relation, c-command, is not linear, as implied by (14) and (15), but hierarchical. C-command is a fundamental relation between the nodes of an inverted tree diagram parse of a sentence. Informally, the relation can be expressed as one between nodes of the same "generation" in the family tree, i.e., that are the same level down the hierarchy. A node dominates all nodes that are its children. It c-commands its siblings and all its siblings' descendents.

Formally, c-command is defined as follows:

Given two nodes, A and B, A c-commands B if and only if

1- The first branching node that dominates A dominates B; and

2- A and B do not dominate each other.

In the simplified tree diagram in Figure 3, the sentence node S dominates the subject noun phrase (NP_1) and its descendents, as well as the verb phrase (VP) and its descendents. S does not c-command anything. NP_1 c-commands VP and its descendents, V and NP_2. NP_1 also c-commands the descendents of NP_2, $Number_2$ and N_2. V c-commands NP_2 and its descendents but it does not c-command NP_1 because the first branching node that dominates V is VP, which does not dominate NP_1.

The relations that were expressed under a bracket notation in (14) and (15) are shown as tree diagrams in Figure 4. The tree diagrams make clear that the relation between constituents of a sentence is a hierarchical one and not a linear one.

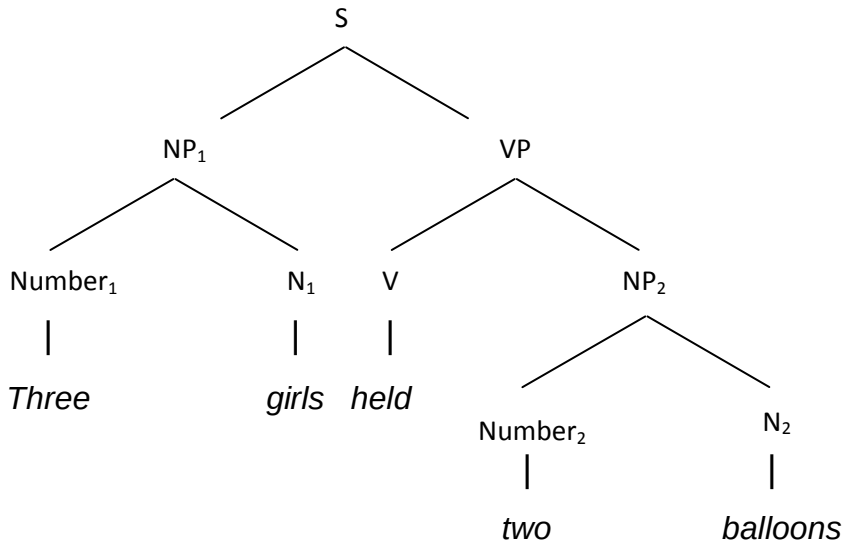


Figure 3. Simple inverted tree diagram of *Three girls held two balloons.*

The topmost node, S, dominates all other nodes. NP_1 dominates $Number_1$ and N_1 and no others. VP dominates V, and NP_2 and its descendents. NP_1 c-commands VP, V, and NP_2 and its descendents. V c-commands NP_2 and its descendents but V does not c-command NP_1.

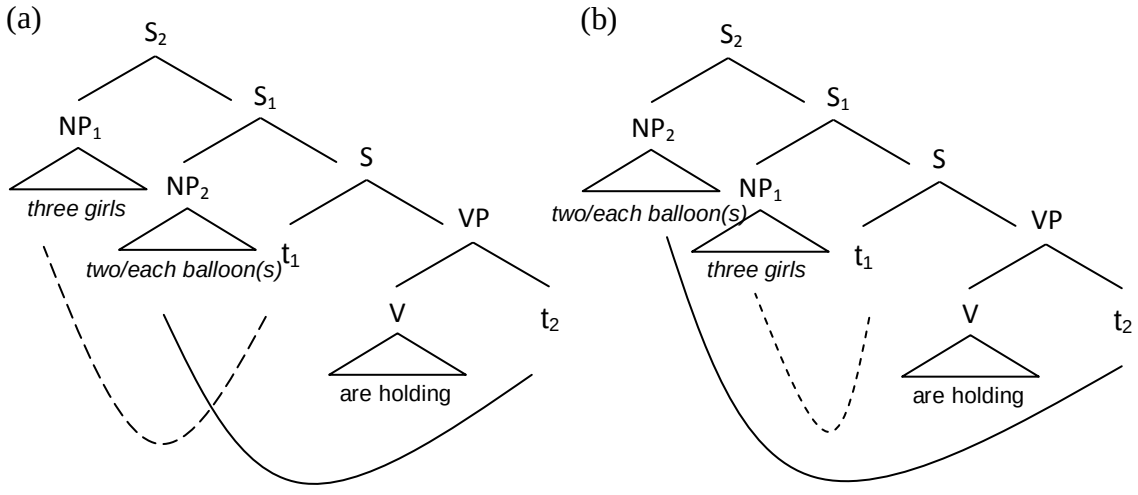


Figure 4. LFs of *Three girls are holding two/each balloon(s),* following May (1977/1990).

Moved quantified phrases adjoin to the sentence successively in any order. The scope of the quantifier is determined by its c-command relations.

The theory of quantifier raising in May (1977/1990) could account for a wide range of previously unexplained phenomena. However, it was empirically inadequate. It predicted that all quantified NPs should behave exactly in the same way, irrespective of the type of quantifier involved. Consequently, any such theory is unable to explain why the object wide scope interpretation is available for NPs quantified by *each*, but not for NPs quantified by *two*.

Beghelli and Stowell (1997)

To resolve this problem, and to tailor May (1977/1990)'s theory of quantification to take into account the differences among types of quantifiers, Beghelli and Stowell (1997) presented an alternative theory. In the 20 years since May's dissertation, developments in syntactic theory had elaborated a more complex map of clause structure that could account for cross-linguistic facts concerning movement and agreement (see Figure 5). X-bar theory had been the standard account of how words combine to form phrases. Phrases were understood as projections of lexical morphemes such that nouns headed noun phrases, verbs headed verb phrases, etc. Later, X-bar theory was applied to functional morphemes (plural markers, verbal inflection, etc.), and later still to the features that compose functional morphemes (gender, number, tense, aspect, etc.), extending the notion of phrasal projection. In parallel with the expansion of X-bar theory, the theory of checking was developed, in which all functional features must be checked for agreement in order for a sentence to be interpretable. Agreement was defined in terms of structural relations holding between a functional head and a dependent element that came to move to the specifier (Spec) position of the relevant head. Under Beghelli and Stowell (1997)'s working assumptions, the Subject Phrase Agreement node (AgrSP) is

roughly equivalent to the S(entence) node of the previous theory. The Object Phrase Agreement node (AgrOP) is assumed to be where the object is checked for agreement in languages where the object agrees with the verb. As we shall see, Beghelli and Stowell (1997) elaborate the structure even further. The central idea of their account is to take into account the different lexical natures of the quantifiers.

On Beghelli and Stowell (1997)'s account, movement at LF is obligatory for some types of quantified phrases and optional for others. [1] Movement is again motivated by the need to check for agreement between features, in this case the logico-semantic features of quantified phrases, such as universal or existential force, cardinality and distributivity, and their corresponding functional projections, which are part of the structure of a sentence. Each type of quantified phrase has its own position in the hierarchy of functional projections so each imposes different constraints on possible landing sites for moved quantified NPs. The functional projections of interest here are: Referential Phrase (RefP), the projection for sentential subjects; Distributed Share Phrase (ShareP), the projection where objects have their distributive feature checked; and Object Agreement Phrase (AgrOP), where objects are checked for case[2] features. In this theory, scope relations emerge as an epiphenomenon of having to check for feature agreement rather than from successive movement to scope-taking positions, as in May (1977/1990).

There are 3 positions in which bare numeral NPs such as *two balloons* (GNPs in Beghelli and Stowell (1997)'s typology) can be interpreted (see Fig. 5). If a bare numeral NP is the object of the sentence, it can move to the Specifier of Shared-Distributive

[1] Beghelli and Stowell classify quantified phrases into 5 types. In this paper, we will only be concerned with 2: Distributive-Universal phrases (DNPs), to which phrases such as *each balloon* belong and Group-Denoting NPs, to which phrases with bare numeral quantifiers such as *Three girls* belong.
[2] Case is a licensing condition on nouns.

Phrase (ShareP), the landing site for distributive object NPs. Alternatively, it can remain in the Specifier of the Object Agreement Phrase, its case feature-checking position. Finally, if a bare numeral NP is the sentential subject, (technically, "the logical subject of predication" (Beghelli & Stowell, 1997, p. 8)), it moves obligatorily to its only possible landing site: the Specifier position of the Referential Phrase. Because Spec of RefP is the

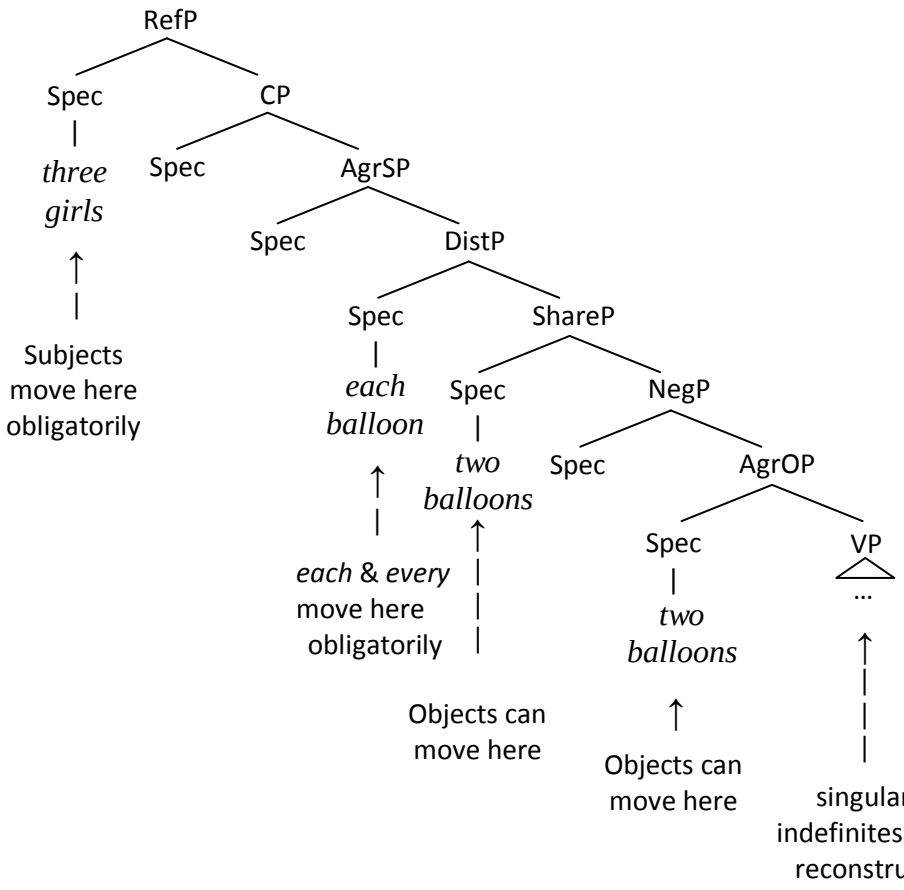


Figure 5. LF of *Three girls are holding two/each balloon(s)*, following Beghelli & Stowell (1997), details omitted.

RefP = Referential Phrase; CP = Complementizer Phrase; AgrSP = Subject Agreement Phrase; DistP = Distributive Phrase; ShareP = Distributed Share Phrase; NegP = Negative Phrase; AgrOP = Object Agreement Phrase; VP = Verb Phrase

Because subject quantified NPs must move to Spec of RefP, they will always c-command objects. This theory explicitly predicts that object wide scope interpretation is impossible in sentences with two bare numeral NPs.

highest node in the structure, Beghelli and Stowell (1997)'s answer to the puzzle of *each vs. two* is clear: "the *grammar* simply excludes" the object wide scope reading of sentences with two bare numeral NPs (italics original, p.9).

Beghelli and Stowell (1997) allow an additional mechanism for agreement-checking which only applies to "simple" indefinite (singular and bare plural) NPs: this mechanism is reconstruction. Reconstruction captures the idea that a phrase may be interpreted at any of the locations it has occupied within a given derivation. May (1977/1990) saw reconstruction as an instance of quantifier lowering. More recently, Hornstein (1995, cited in Beghelli & Stowell, 1997) construed reconstruction as an alternative to raising and lowering operations.

Current syntactic theories subsumed under Chomsky's Minimalist Program assume that Universal Grammar imposes a Principle of Full Interpretation at Logical Form (LF) and Phonological Form (PF), the interfaces between the grammar and the conceptual-intentional and articulatory-perceptual systems, respectively. In the course of a derivation, syntactic structures are checked for relevant features. All and only features that contribute to the interpretation of a sentence - for example, number, gender, tense, case, etc. - must be present at LF for a derivation to converge, otherwise the derivation crashes. Likewise for PF and features necessary to pronounce (or sign) a string. A derivation must converge at both LF and PF for an expression to be grammatical (Radford, 1997).

On Hornstein (1995)'s account, what were traces of movement under previous theories are *copies* of moved phrases. In order for a derivation to converge, only one copy of a moved constituent can be pronounced, i.e., remain at PF, so the other must copy be

deleted, i.e. not pronounced. The ones which remain after copy-deletion determine scope relations. For example, (16) is the subject wide scope reading and (17) the object wide scope reading of (1) after copy deletion:

(16) [$_{NP1}$ Three girls] [$_{NP2}$ ~~two balloons~~] [$_{S}$ ~~Three girls~~ are holding two balloons]

(17) [$_{NP1}$ ~~Three girls~~] [$_{NP2}$ two balloons] [$_{S}$ Three girls are holding ~~two balloons~~]

The scope relations obtained are identical to those in (14) and (15), although the mechanisms by which they were obtained differ.[3]

Beghelli and Stowell (1997) allow reconstruction of singular indefinites in order to rescue the derivation of sentences such as (18) on the object wide scope reading (which can be paraphrased as *There are two balloons; each balloon is being held by a different girl*) be would otherwise be ruled out by the theory.

(18) A girl is holding two balloons.

If *A girl* were not allowed the reconstruction escape hatch, it would have to move to Spec of RefP, where the sentence could only receive the object wide scope interpretation (Fig. 5). In fact, the derivation would be identical to (1)'s. By allowing singular indefinite (and bare plural) subjects to reconstruct to their originating VP, Beghelli & Stowell (1997) guarantee that they can be interpreted in a position below all the functional projections. Since object NPs must raise above VP to have their agreement features checked, singular indefinite NPs are predicted to have both a narrow scope and a wide scope interpretation.

[3] The two approaches make different predictions for other phenomena that will not be discussed here.

So, it is not only the distributive-collective nature of quantified object NPs that matters. On this account, the nature of the quantified subject NP matters, as well. The reason that the reading is available with singular indefinite subjects but not bare numeral ones is because only singular indefinites can reconstruct.

Reinhart (2006)

Reinhart (2006) presented an alternative theory that denied a grammatical ban on the object wide scope interpretation of bare plural NPs. Reinhart (2006) argued instead that the grammar makes the reading available, but it is massively dispreferred because the computation required by its derivation exceeds the limits of working memory. On this theory, the fundamental problem with the object wide scope reading is not scope, *per se*, but a quantified object taking wide scope over a subject that can be interpreted either collectively or distributively. The nature of the subject noun phrase is the crucial element in determining whether the object wide scope interpretation is available. If the subject can be construed as distributive, the derivation exceeds the limits of working memory and the sentence cannot receive an interpretation. But if the subject can only be construed as collective, the sentence is interpretable. This predicts that is a distributive subject can be forced into a collective reading, the interpretation should be acceptable.

Beghelli and Stowell (1997) offered a highly constrained, syntactic theory of movement in which every quantified phrase type had its own landing site (or in limited cases, sites). On that account, the grammar accounted for much of the sentential semantics. Reinhart (2006) assumes a minimalist grammar[4] that does not "code everything needed for the interpretation… (p. 309)" of a sentence. The grammar insures

[4] The grammar is referred to as the Computational System in the Minimalist Program, under whose assumptions Reinhart (2006) is working.

that syntactic requirements of a language are met, but leaves sentential semantics to an independent logic component. The grammar's job is to check lexical items for relevant features. Once the requirements of the grammar are met, syntactic information is output to the logic component which has its own requirements and carries out its own (semantic) computations.

In the grammar, overt movement such as occurs in forming *Where is Mary going?* from *Mary is going where*, can occur freely. Covert movement can also occur freely, as long as the movement is necessary for a derivation to converge, as in the feature-checking function described above. But quantifier raising is covert. And it is unnecessary for convergence, because the derivation of a scopally ambiguous sentence can converge on the subject wide scope reading without it. Nevertheless, the logic system clearly makes the object wide scope reading available, as (18) shows.[5] So although it is an "illicit operation (p. 300)" as regards the grammar, Reinhart (2006) allows quantifier raising to apply as a "repair strategy" at the syntax-semantics interface. However, this accommodation to the demands of the logic system comes at the cost of increased processing demands.

When there is no ambiguity, the logic system need not keep track of which interpretation corresponds to which derivation, as there is only one of each. But in the case of a one-to-many relationship between a derivation and possible interpretations, as occurs with quantifier raising, the system must generate of a reference set of all possible derivation-interpretation pairs. The reference set includes all possible interpretations that

[5] The claim is that the object wide scope reading is easier to access in sentences with singular indefinite subjects than in those with bare numeral subjects, not that all speakers will accept sentences such as (18). Compare (18) with *A flag was hanging in front of three buildings*, for which the object wide scope reading is claimed to be preferred (Reinhart, 2006, citing Hirschbühler, 1982).

can result from a given derivation - both scoped ones that arise from quantifier raising and scopeless ones, which do not arise from movement.

Only distinct interpretations are ultimately considered for inclusion in the reference set, due to the same principles of derivational economy that make quantifier raising an illicit operation. But as Figure 2 shows, not all interpretations are distinguishable. Therefore, every derivation-interpretation pair has to be held in working memory while it is checked against every other member of the set to make sure that it is not equivalent to any other member. Derivational economy also requires that each member of the reference set be checked to insure that its membership would not have been possible without quantifier raising. Naturally, NPs that allow both distributive and collective interpretations generate larger reference sets than NPs that allow only collective readings.

Reinhart (2006)'s central claim is that the reference set generated by a sentence that allows both distributive and collective object wide scope readings exceeds the limit of working memory. It is a processing failure which renders the sentence uninterpretable. In contrast, a sentence that allows only the collective wide scope reading does not exceed the limit of working memory (in adults). It can therefore be processed successfully and receive an interpretation.

The reference sets for (1), repeated below as (19), is computed as an example. Recall that Reinhart (2006) agrees with Beghelli and Stowell (1997) that bare numeral object NPs cannot be interpreted distributively, so that option is never considered. Reinhart (2006) used a different mechanism,[6] the details of which are not important here,

[6] The mechanism Reinhart (2006) used is choice functions. The details of choice functions do not matter here. Interested readers should consult Reinhart (1997), *Quantifier scope: How labor is divided between*

to derive scope relations of (existentially) quantified expressions that denote collections, such as sets of girls and balloons, without quantifier raising. We follow May (1977/1990) and Beghelli and Stowell (1997) in deriving scoped interpretations via quantifier raising regardless of whether they are construed as distributive or collective. For consistency of exposition, we chose to include the interpretation depicted in Figure 2h with Reinhart (2006)'s quantifier raising interpretations, but it should be noted that she considered it as representing an overt, non-quantifier raising, interpretation. Crucially, the size of the reference sets obtained by either mechanism is the same.

In the first step of the computation, all the interpretations that are possible without quantifier raising – those consistent with the overt string – are listed. These include what we have described as the subject wide scope and each-all[7] interpretations (in Figure 1). Each interpretation is informally described following Grudzinska (2010) in (19a) – (19c) and labeled according to its corresponding configuration in Figure 2. The *greater than* symbol indicates which NP takes wide scope.

(19) Three girls are holding two balloons.

a. There is a set of 3 girls and a set of 2 balloons;
   the 3 girls together are holding the 2 balloons together
   (collective subject > collective object, Figure 2d)

b. There is a set of 3 girls such that for each girl there is a set of 2 balloons,
   each girl is holding the 2 corresponding balloons together

---

QR and choice function, Linguistics and Philosophy, 20, 335-397 or Yoad Winter's *Choice functions and the scopal semantics of indefinites*, pp. 399-467 of that same volume.

[7] Reinhart (2006) does not consider the scopeless cumulative interpretation (Fig. 1d). Doing so would increase the size of the reference set for (19) but not for (20).

(distributive subject > collective object, Figure 2b)

c. There is a set of 3 girls and a set of 2 balloons;
   each of the girls is holding each of the balloons
   (scopeless each-all interpretation, Figure 1c)

In the next step, each overt interpretation must be checked against every other one to insure uniqueness. In our example, (19a) is equivalent to (19c), so one interpretation (it is irrelevant which) is discarded. The reference set now has 2 derivation-interpretation pairs. These must be held in working memory while the object wide scope readings that result from the illicit quantifier raising strategy are listed and checked:

(19) Three girls are holding two balloons.

d. There is a set of 2 balloons and a set of 3 girls;
   the 3 girls together are holding the 2 balloons together
   (collective object > collective subject, Figure 2h)

e. There is a set of 2 balloons and a set of 3 girls such that,
   each of the girls is holding the 2 balloons together
   (collective object > distributive subject, Figure 2f)

f. There is a set of 2 balloons such that for each balloon there is a set of 3 girls;
   3 girls together are holding each of the balloons
   (distributive object > collective subject, Figure 2g)

g. There is a set of 2 balloons such that for each balloon there is a set of 3 girls;
   each of the girls is holding a corresponding balloon

(distributive object > distributive subject, Figure 2e)

Again, each member must be checked against every other. The interpretation in (19d) is equivalent to both (19a) and (19c), so it is discarded. However, the remaining 3 interpretations are unique. This can be verified by checking their corresponding configurations in Figure 2. Because we obtained all scoped interpretations through quantifier raising, Reinhart (2006)'s second condition, that all interpretations must be checked to insure that they could not have been possible without quantifier raising, only applies to the scopeless interpretation in 19c, which was already eliminated under non-uniqueness. This computation produces a final reference set with 5 members.

Earlier work by Reinhart and others (see Reinhart, 2006, p. 292) had shown the object wide scope reading to be more accessible in sentences with singular indefinite subjects than in those with bare numeral NPs. Compare (19) with (18), repeated here as (20):

(19) Three girls are holding two balloons.

(20) A girl is holding two balloons.

Both sentences have the same kind of quantified NPs subjects according to Beghelli and Stowell (1997)'s typology. But recall that because the subject of (20) is a singular indefinite, it can get a narrow scope interpretation by reconstruction of the singular indefinite NP to its original position within the verb phrase. Reinhart (2006) agreed with Beghelli and Stowell (1997)'s claim that the nature of the subject NP

mattered for whether the object NP could take wide scope. But she argued that their identification of the singular indefinite as the relevant aspect, and its subsequent rescue by reconstruction, was incorrect. Her insight was to notice that because a singular indefinite subject can only denote one thing, it cannot possibly be interpreted distributively, unlike a bare numeral subject which can. It is the potentially distributive nature of the bare numeral subject NP (in the presence of a bare numeral object NP) that accounts for the too-large-to-process reference set of 5 members generated by (19). By comparison, the reference set for (20) contains a manageable 2. Computing the reference set for (20) confirms that it only contains 2 members:

(20) A girl is holding two balloons.

a. There is a set of 1 girl such that there is a set of 2 balloons;
   the girl is holding the 2 balloons together
   (collective subject > collective object, Figure 2d)

b. There is a set of 2 balloons and a set of 1 girl;
   the girl is holding the 2 balloons together
   (collective object > collective subject, Figure 2h)

c. There is a set of 2 balloons such that for each balloon there is a set of 1 girl;
   the girl is holding each of the balloons
   (distributive object > collective subject, Figure 2g)

Checking for uniqueness, we find that (20a) and (20b) are equivalent, and discard one of them and leaving 2 members in the set.

Drawing on this observation, Reinhart (2006) predicted that if one could force a collective reading of the subject NP by adding a modifier such a *together* or *simultaneously*, the object wide scope reading would become more tractable and therefore more acceptable. Excluding the distributive interpretations of the subject produces a reference set with two fewer members, small enough to hold in working memory while the necessary computation is completed.

Reinhart (2006) tested her prediction on "a couple of non-linguist informants (p. 294)." She presented contrasting items such as (21) and (22) one at a time, and (orally) asked her informants how many flags there were (p. 304).

(21) Three flags were hanging in front of two buildings.

(22) Three identical flags were hanging in front of two buildings.

If they answered, "three," that is, if they gave the subject wide scope interpretation, she asked them whether 6 flags were possible. She reported being "able to elicit a yes answer to the second question in all her informal testing (p. 295)" of several items, some which used an adjective as the collectivizing modifier, and others an adverb. Reinhart (2006) tested native Hebrew speakers (in Hebrew, of course), but she asserted that semantic judgments of quantifier scope in Hebrew versus English are not known to differ.

Motivation and Predictions

The experiments were motivated by Beghelli and Stowell (1997)'s and Reinhart (2006)'s conflicting claims about the nature of the subject noun phrase. The two theories suggested comparing three subject NP types on the object wide scope reading:

1. singular indefinite
2. bare numeral
3. bare numeral plus a collectivizing modifier.

Beghelli and Stowell (1997)'s theory overtly predicts that type 1 will be accepted and type 2 will be rejected wholesale. They do not discuss type 3 explicitly, however, what makes type 1 acceptable is that only singular indefinite (*A girl*) and bare plural noun phrases (*Girls*) can reconstruct. A modified bare numeral noun phrase is neither, so it is predicted not to be able to reconstruct, and therefore be rejected (Reinhart, 2006).

Reinhart (2006)'s theory makes a clear prediction for acceptability judgments of the 3 subject types, and differs from Beghelli and Stowell (1997)'s predictions only as regards type 3. Type 1 is impossible to be construed distributively since the subject has a cardinality of 1, so it should be easily accessible; type 2 should be massively dispreferred because of its distributive nature, which creates a reference set that exceeds the capacity to process it; and because the additional cognitive load imposed by distributivity is cancelled by the modifier, type 3 NPs should pattern with type 1. Of course, these predictions rest on her assumption that intuitions related to scope are not affected by the choice of Hebrew *vs.* English.

Reinhart (2006) carried an implicit prediction that all the items she tested were identical with respect to the facts of scope and distributivity. Our intuitions did not agree.

We hypothesized that the choice of predicate mattered for whether people would accept a particular interpretation or not and predict to see item effects.

General Method

Overview

In the experiments reported below we used a modified version of the Truth Value Judgment Task (TVJT), an experimental procedure designed to test young children's sentence comprehension (Crain and Thornton, 1998). In the computerized version of the TVJT used in Musolino (2009), participants are shown a short PowerPoint animation whose soundtrack narrates the scene as it unfolds. In the critical cases, the final scene is composed of objects in a configuration that represents the object wide scope reading. At the final scene, the narrator summarizes his/her interpretation of the scene using the target sentence, "I know. Three girls are holding two kites," then asks, "Am I right?" Crucially, the target sentence will be either true or false on the intended reading for each story. The participant answers *yes* or *no*. The dependent measure is the percentage of *yes* answers.

We modified the computer-based TVJT for exclusive use with adults. Instead of a vignette that is narrated as it unfolds, adult participants were presented only with the final configuration - the one that participants see at the prompt in the version designed for use with children. We asked participants for their judgments directly rather than asking whether the narrator was right or wrong. Finally, in order to try to elicit finer-grained judgments than are allowed by *yes-no* truth value judgments, we asked participants to give their opinion of how well the target sentence matched a given configuration using a 7-point rating scale.

Procedure

Participants were recruited from the undergraduate subject pool and participated for partial course credit. After signing a consent form, they were asked to complete an optional demographic questionnaire (Appendix A) that included details of their language history.

Materials

All test items were taken from among Reinhart (2006)'s examples, and modified such that the cardinality of subjects and objects was held constant. Tense and aspect were held constant as much as possible, giving preference to naturalness when there was a conflict. This resulted in one item in the simple past and all others in the present progressive (Appendix B: Test Items). Test items were always paired with a configuration which rendered the target sentence true on the object wide scope reading.

Experiment 1

In this experiment, we piloted 5 lexical items and 2 presentation modalities for use in subsequent experiments. All test sentences had singular indefinite subjects; cardinality and tense/aspect were held constant as described above. The modalities compared were spoken *versus* written presentation of target sentences. The rating scale was presented visually in both conditions.

Method

Participants

Forty-two undergraduates at a large northeastern public university participated in the study for partial course credit. An additional participant's results were excluded because of technical problems with the stimulus presentation software. Forty-one

participants were self-reported native speakers of English. Although being a native English speaker was a condition for participation, one participant self-reported that he was not. However, on interview he revealed that he considered himself to have native proficiency, and was included in the study.[8] Sixteen participants were native speakers of at least one language in addition to English; ten reported speaking a language other than English at home (less than half the time). All but one had studied a language other than English for at least 2 years.

Procedure

Experiment 1 was conducted in-lab. Participants were tested individually in a quiet room without the experimenter present. Before beginning the experiment, the experimenter read a standardized set of instructions aloud (Appendix C), then randomly assigned the participant to either the spoken or written condition. Participants were seated in front of a Dell Latitude D830 laptop computer running Empirisoft DirectRT v. 1.0.11 stimulus presentation software for Windows XP, with Logitech v20 notebook speakers placed at either side of and slightly behind the computer. Additional self-paced instructions (Appendix D) guided participants through the mechanics of the experiment. Participants responded by pressing a number key from 1 and 7, which the presentation software automatically recorded before advancing to the next trial.

Design

This study represented a 5x2x2 mixed design. The within-subject factor was test item (5 lexical items, Appendix B, Experiment 1). Between-subject factors were presentation modality (spoken vs. written) and counterbalanced order of blocks of scope-

[8] Bilingual persons' self-described proficiency correlates well with their scores on formal assessments of language proficiency. (Judith Kroll, p.c.)

dependent configurations (subject wide scope block first vs. object wide scope block first). Following Musolino (2009), subject wide scope and object wide scope configurations were blocked separately, as these represent linguists' intuitions of the most and least accessible interpretations, respectively. Because the test items were always in the object wide scope configuration, participants would see all the test items in either the first or second block.

Participants saw 33 total items: 3 practice trials, 5 (object wide scope) test items, 5 false controls, 10 subject wide scope controls (5 true, 5 false), and 10 fillers that served as scopeless controls. Sample stimuli are shown in Figure 6.

Measures

The dependent variable was participants' acceptability of target sentence on the interpretation depicted by the visual configuration, as rating on a 7-point rating scale where 1 = Definitely No to 7 = Definitely Yes, with a midpoint of 4 = Either Way.

Materials

The 5 test sentences were paired with a configuration in which the target sentence, which always had a single indefinite subject and an object numerically quantified by 3, was true on the object wide scope reading, i.e., it depicted 3 subjects and 3 objects. For each test sentence there was a corresponding false object wide scope control that paired the target sentence with a configuration in which the sentence was false by virtue of an incorrect number of either subjects or objects.

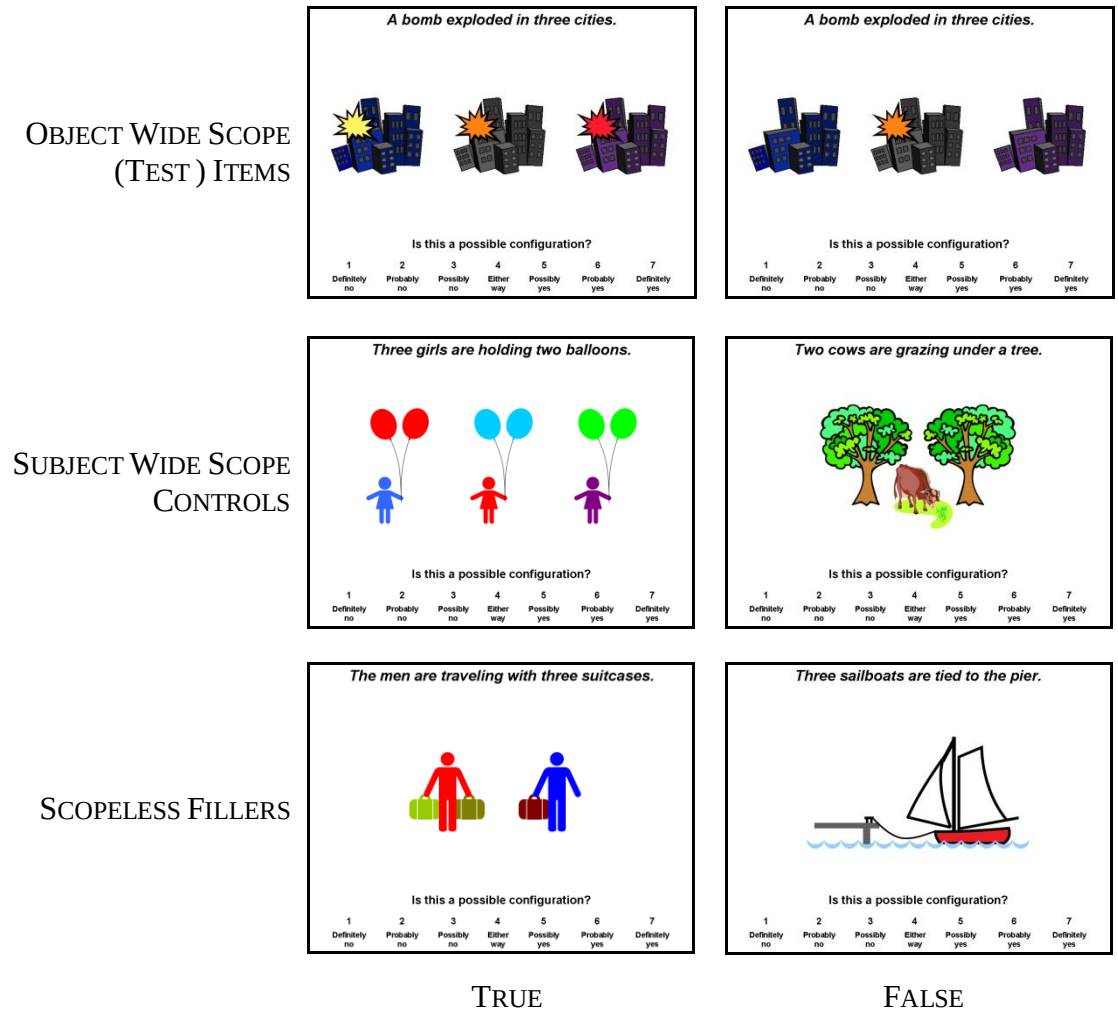


Figure 6. Sample stimuli used in Experiment 1.

To insure that subjects could access the full range of scoped readings, 10 control sentences with NQE subjects and objects were paired with their subject wide scope configuration, 5 true and 5 false. For example, in the true cases, the subject wide scope control sentences were presented with pictures that depicted 2 (or 3) subjects, each corresponding to 3 (or 2) objects, such that there was always a total of 6 objects. In the false cases, an additional 5 sentences were also paired with their corresponding subject wide scope configurations, but each picture was missing one or more objects.

Fillers consisted of 8 unambiguous sentences containing a single NQE (4 true and 4 false on its paired configuration), plus 2 tokens of the *Three girls are holding two balloons*, paired with its each/all and object wide scope configurations. The each/all and object wide scope interpretations had resulted in the highest and lowest acceptance rates, respectively, in previous experiments (Musolino, 2009; Syrett & Musolino, in preparation). They were included to calibrate participants' use of the rating scale to previous findings using an unmodified TVJT, which used percentage of *Yes* answers to a yes-no prompt as the dependent measure.

Test sentences were interlaced with one object wide scope false control sentence and one scopeless filler sentence. DirectRT was programmed to select each of the three sentence types at random without replacement from separate lists. In the subject wide scope block, true subject wide scope sentences were interlaced with one subject wide scope false control and one filler, selected at random as above.

DirectRT was additionally programmed to randomize block order, and order of false controls and fillers with respect to one another. Practice trials were presented in fixed order: one true, one false, one true.

To be able to take full advantage of DirectRT's randomization capabilities while insuring that the spoken and written conditions were minimally different, 20-sec animations (Windows Movie files) were created using Camtasia Studio 6 software. The resulting stimuli, however, appeared to be still images.

Visual stimuli were first created as PowerPoint slides using a combination of public domain clip art and figures drawn natively from basic shapes, including the rating scale along the bottom edge. Each slide was saved as a bitmap image. Auditory stimuli

were digitally recorded by a single female speaker over two sessions at the Rutgers Phonetics Laboratory using an AT4040 Cardioid Capacitor microphone with a pop filter in a sound-attenuated recording booth and amplified through an ART TubeMP microphone pre-amplifier, with Audacity 1.2.6 audio recording and editing software set to CD-quality minimal sampling rate. Extremes of pitch and loudness were reduced in Adobe Soundbooth CS3 so as to neutralize intonation contours as much as possible while preserving the subjective impression of natural speech.

In the spoken condition, soundtracks were placed to begin 200 ms after the picture appeared. In the written version, the movie file consisted of a continuous presentation of the bitmap image; the stimulus sentence was added in DirectRT, centered above the picture in large black italic font, and appeared simultaneously with the image.

Results and Discussion

Reinhart (2006) carried an implicit prediction that the choice of lexical items should not affect participants' interpretations given that all subjects NPs are in the singular indefinite. A mixed model ANOVA was used to compare participants' mean acceptability ratings (Table 1) using lexical item as the within-subjects variable, and condition and block order as between-subjects factors in SPSS v. 17. It revealed a significant main effect of item, $F(4, 152) = 16.86$, $p < .01$, contra Reinhart (2006). Pairwise comparisons with Bonferroni adjustment for multiple comparisons showed a statistically significant difference between: the mean of bomb-cities (highest) and the means of the 4 other items ($p \leq .01$); the mean of doctors-patients (lowest) and all other items except tablecloth-tables ($p \leq .05$) (Fig. 7). Because we wanted to create the most

favorable conditions possible for the object wide scope interpretation to emerge, we decided not to use the generally dispreferred doctors-patients item in Experiment 2.

TABLE 1. **EXPERIMENT 1**

OVERALL ACCEPTABILITY RATINGS, *N = 42*

Interpretation	***M***	***SD***
Object Wide Scope	4.6190	1.1982
Subject Wide Scope	5.1571	1.4063
Scopeless Filler	6.4429	0.6922
False Controls		
Object Wide Scope	1.9667	0.7473
Subject Wide Scope	1.7476	1.0227
Scopeless Filler	1.6810	0.8702

The difference between the means of the spoken and written conditions was not significant ($F(1,38) = 0.21$, $p = .652$). These results allowed us to proceed to Experiment 2 with increased confidence that using only a written presentation, rather than a spoken one as in Reinhart (2006), would not bias our results.

Block order was not significant and $F(1,38) = 0.24$, $p = .628$. No interaction effects were observed.

Although this experiment was designed primarily to pilot test items and conditions for future experiments, it also provided a preliminary test of Reinhart (2006)'s claim that the object wide scope reading is available with single indefinite subjects. Means are shown in Table 2. Mean responses to false controls (object wide scope, subject wide scope and scopeless) indicated that subjects rejected false cases.

A mixed model ANOVA was used to compare participants' overall acceptability ratings of the object wide scope (critical), subject wide scope (control) and scopeless

(true fillers) interpretations. Interpretation was entered as the within-subjects variable; condition (spoken vs. written) and block order as between-subject factors. There was a significant main effect of interpretation, $F(2, 38) = 31.32$, $p < .01$, with Bonferroni adjustment for multiple comparisons (Fig. 8). The effect was driven by the mean for the true scopeless fillers, which was close to ceiling.

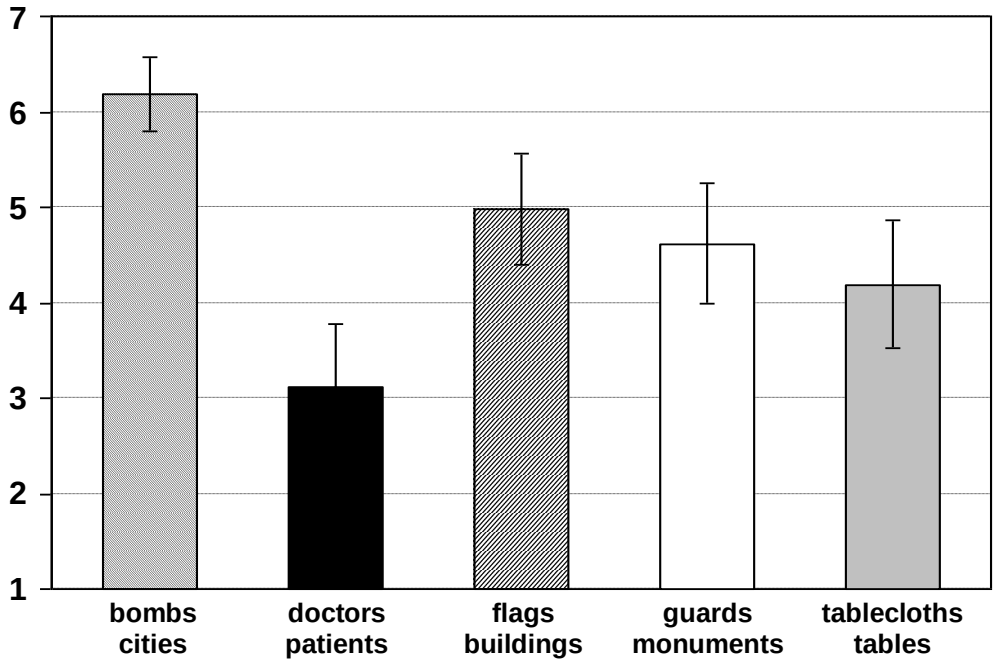


Figure 7. Experiment 1 item effects.

Clear item effects were found, contra Reinhart (2006). Bombs-cities and doctors-patients differed from all other items. Flags-building, guards-monument and tablecloth-tables differed from the first two items, but not from one another.

TABLE 2. **EXPERIMENT 1**

ACCEPTABILITY OF OBJECT WIDE SCOPE READING BY ITEM, *N = 42*

Critical Items	***M***	***SD***
A bomb exploded in three cities	6.1905	1.2923
A doctor examined three patients	3.1190	2.1437
A flag was hanging in front of three buildings	4.9762	1.9316
A guard was standing in front of three monuments	4.6190	2.1064
A tablecloth was covering three tables	4.1905	2.2332
False Controls		
A bomb exploded in three cities	1.4048	0.8571
A doctor examined three patients	1.9762	1.9316
A flag was hanging in front of three buildings	3.1190	1.9406
A guard was standing in front of three monuments	1.1190	0.3278
A tablecloth was covering three tables	2.2143	1.7466

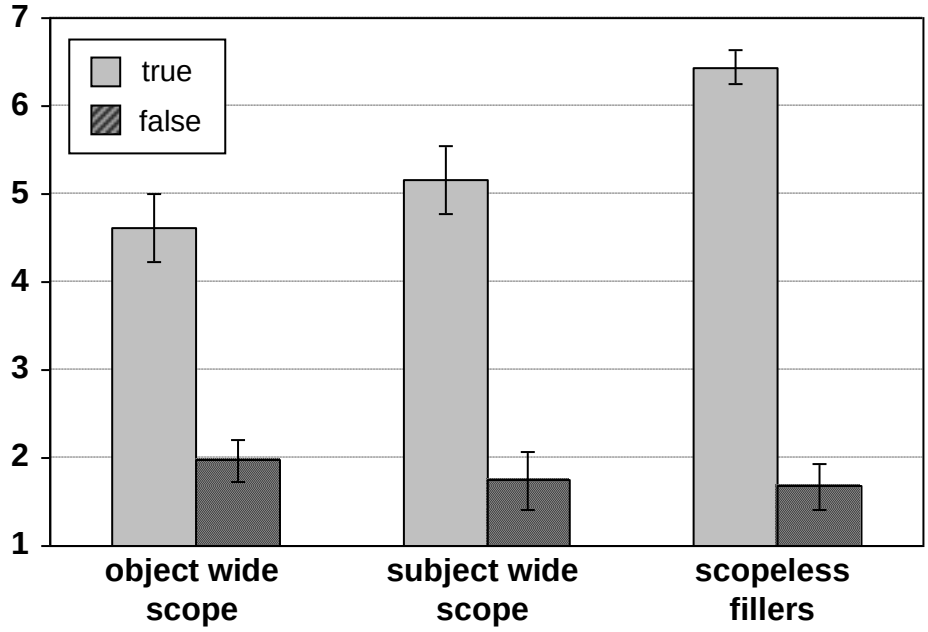


Figure 8. Experiment 1 overall effects of interpretation across conditions.

The object wide scope interpretation is available with singular indefinite NPs, as predicted by the theoretical literature, statistically no different from subject wide scope controls. The means of scoped interpretations differed from that of scopeless controls.

Pairwise comparisons revealed no statistically significant difference between the means of the object wide scope and subject wide scope interpretations ($p = .152$). This finding supports Reinhart (2006)'s claim insofar as: 1- the overall mean for the object wide scope reading was well above the means of all false controls (see discussion of controls below); and 2- the mean of the object wide scope reading did not differ from the mean of the subject wide scope interpretation. The latter has been shown to be generally well-accepted in studies that used percentage of *yes-no* trials accepted on the object wide scope interpretation as the dependent measure (Musolino (2009).

Paired t-tests were used to compare overall means of object wide scope critical cases (those that are true on that interpretation) and subject wide scope true controls to their respective false controls, as well as true *vs.* false scopeless fillers. The difference between the means of all 3 (true) interpretations and their false controls was statistically significant: object wide scope, $t(41) = 13.66$, $p < .01$; subject wide scope, $t(41) = 13.10$, $p < .01$; scopeless fillers, $t(41) = 22.88$, $p < .01$.

Results for the each/all ($M = 6.00$, $SD = 1.53$) and object wide scope ($M = 2.45$, $SD = 1.88$) configurations of *Three girls are holding two balloons* revealed that participants' responses were comparable to those reported previously. Musolino (2009) found that adults answered *Yes* to the each/all reading in 100% of trials and to the object wide scope reading 7.8% of trials. As a percentage of the rating scale[9], a mean of 6.00 is equal to approximately 83.33% Percentage *Yes* and a mean of 2.45 is approximately 24.17%. Although our results are not as extreme as Musolino (2009)'s, they do exclude

[9] The conversion from 7-pt. rating scale to percentage (answered *Yes*) was computed using the formula Percentage *Yes* = (rating score - 1) x 100/6. Alternatively, Percentage *Yes* can be rescaled using the formula 7-pt. score = (6/100 x Percentage *Yes*) + 1.

the middle range of the scale, indicating that our participants were fairly certain in their acceptability judgments of those interpretations. At the same time, the fact that our mean results are less extreme suggests that acceptability judgments may be more nuanced than what have been reported using coarser-grained measures.

These results of Experiment 1 indicated that participants are sensitive to differences in interpretation regardless of presentation modality. Participants behaved as expected on all false control items. The acceptance rate for true subject wide scope cases differed from true scopeless controls but was within the expected range. Musolino (2009) found that participants accepted the subject wide scope interpretation in 82.8% of trials. Converted to Percentage *Yes*, our participants accepted the subject wide scope reading in 69.28% of trials. We consider these findings comparable to Musolino (2009)'s, especially given that our converted participants' ratings of scopeless true controls was 90.67% *vs.* Musolino (2009)'s results of 98.9% for scopeless sentences containing *Two N* and 100% for scopeless sentences containing *Three N*. Most importantly, although we did not find a difference between the object wide scope true and subject wide scope true cases, as indirectly predicted by Reinhart (2006), we did find a difference among lexical items, contra the same.

Experiment 2

In this experiment, we investigated Reinhart (2006)'s claim that the availability of the object wide scope interpretation of multiply numerically quantified sentences depends on whether the subject is construed as distributive or collective. We compared 3 subject types - singular indefinite, NQE, and NQE plus collectivizing modifier - using the 4 lexical items from Experiment 1 that resulted in the highest acceptability ratings.

Cardinality and tense/aspect were held constant as in Experiment 1. Four additional cumulative configurations were included as filler items and served as pilot stimuli for subsequent experiments. This experiment was web-based.

Method

Participants

One hundred sixty-eight undergraduates participated in the study for partial course credit; 132 were included in the study. Thirteen were excluded because they self-identified as non-native speakers of English.[10] Because this experiment was conducted outside the lab, we applied more conservative inclusion criteria here that in Experiment 1. Participants' responses on unambiguous control sentence-picture pairs - scopeless true, scopeless false and subject wide scope false - were taken as an indirect measure of attention to task. We eliminated any subject whose response was less than 5 on a true control, or more than 3 on a false control. One participant's data set was missing a rating for 1 of the 4 subject wide scope false controls but because his other control items were within acceptable range, his data were included.

Procedure

Experiment 2 was web-based. Participants logged on to the university's Sona experiment management system, where they signed up and followed a link to the experiment. The welcome page informed participants of what to expect: the need to complete a consent form; an optional language history; and how to configure their computers before continuing. Participants indicated their consent electronically by providing a coded ID number and clicking on one of two buttons at the bottom of the

[10] We indicated "Native English Speaker" in the Sona system's "Eligibility Requirements" field, which is visible to participants. Although the Prescreen Questionnaire included questions about participants' native language, Prescreen responses were not used as exclusion criteria.

page. If after reading the consent form a prospect chose not to participate, he/she could opt out by clicking on a clearly marked alternative, next to the consent button. The alternative button was a link to a page that thanked prospects for having considered participating in the experiment.

Participants were assigned to one of six paths via a hidden page containing randomization code. Each path corresponded to a cross of the three conditions and two blocks orders (see Experiment 2 Design, below). The optional language history form and all instructions were identical to those in Experiment 1, except that the instructions given verbally in Experiment 1 were presented as an additional page of instructions in Experiment 2. The additional page also included instructions to maximize the screen and close all other programs so as to minimize distractions and prevent memory-related computer problems. Participants proceeded at their own pace through all instructions and stimuli.

Each stimulus was presented on a separate web-page. Participants indicated their responses by selecting a button from a radio group, then clicking a *Submit* button, which automatically advanced them to the next item. If *Submit* was clicked without having selected a rating, an error message stating that an answer is required was returned. The experimental pages were coded so that it was difficult, though not impossible, to return to the previous screen. After the last item was completed, participants' responses were posted to a data file as the participant was simultaneously directed to a debriefing page.

Design

This study represented a 4x3x2 mixed design. The within-subject factor was lexical item (Appendix B, Experiment 2). Between-subject factors were subject NP type

(singular indefinite *vs.* bare NQE (always *Two) vs.* collective NQE (always *Two* plus a collectivizing modifier) and counterbalanced order of blocks of scope-dependent configurations (subject wide scope block first *vs.* object wide scope block first).

Participants saw 31 total trials: 3 practice trials, 4 (object wide scope) test items, 4 object wide scope false controls, 8 subject wide scope controls (4 true, 4 false), and 8 scopeless fillers (4 true, 4 false) that also served as scopeless controls. Four additional items paired a single sentence with four possible cumulative (scopeless) readings. These were pilot items for future studies on the acceptability of various cumulative readings based on the semantics of cumulativity developed by Grudzinska (2010). The results are not reported here.

Measures

The dependent variable was as in Experiment 1. Additional individual variables were recorded in accordance with participants' answers on the language history questionnaire. Other than for exclusion criteria, individual variables were not used in this study. Time stamps were recorded on the first page of instructions and on the first and last pages of each stimulus block as a gross measure of attention to task; no participants were excluded based on time stamp data.

Materials

In each condition, critical items were constructed by pairing the 4 highest-rated test sentences from Experiment 1 with a configuration in which the target sentence is true on the object wide scope reading. Because we wanted to create the most favorable conditions possible for Reinhart (2006)'s theory, we excluded the lowest-rated item from Experiment 1, doctors-patients.

Condition 1 (singular indefinite subject) used the identical stimuli used in Experiment 1. The configurations depicted in Conditions 2 (NQE) and 3 (NQE + modifier) were identical in number of objects and location on the screen to those in Condition 1, but with twice the number of subjects. The target sentences were modified according to condition, as well (Appendix B, Experiment 2). For each test sentence there was a corresponding false object wide scope control that paired the target sentence with a configuration in which the sentence was false by virtue of an incorrect number of either subjects or objects.

Subject wide scope controls and scopeless fillers were as in Experiment 1 except there were 4 tokens of each instead of 5. In order to maintain the same proportion of critical and control items, there were 4 of each type of false controls, as well. Experiment 2 also included 4 pilot items that paired the sentence *Three girls are holding two balloons* with four possible cumulative (scopeless) configurations.

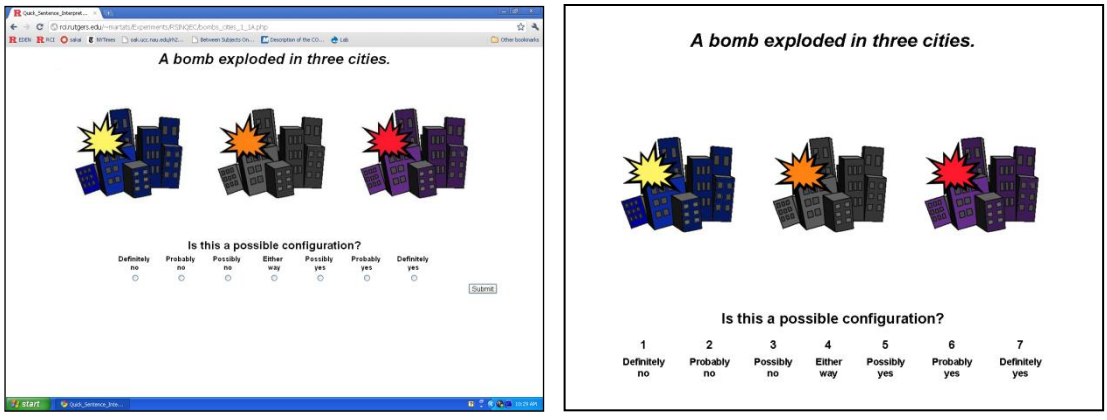


Figure 9. Singular indefinite condition in Internet Explorer (left) and on DirectRT.

All web-pages were programmed to simulate the presentation on DirectRT (Fig. 9). Unlike Experiment 1, in which DirectRT fully randomized the choice of stimulus from lists of stimulus types, Experiment 2 represented a pseudorandom ordering of stimulus, control and filler items. However, the order of presentation of critical, control and filler items remained as in Experiment 1, such that critical items were interlaced between one false control and one filler item.

Results and Discussion

This experiment investigated the predictions of two competing theories: 1- Reinhart (2006)'s claim that by modifying the subject noun phrase so as to force a collective reading, the object wide scope reading becomes accessible, predicting that the singular indefinite and NQE + modifier conditions should pattern together, and; 2- Beghelli and Stowell (1997)'s claim that the object wide scope reading is barred with numerically quantified objects, predicting that participants will unequivocally reject the object wide scope reading in the bare NQE condition. The findings reported below are contrary to Beghelli and Stowell (1997) but lend partial support to Reinhart (2006).

TABLE 3. **EXPERIMENT 2**

OVERALL ACCEPTABILITY RATINGS

Subject Noun Phrase Type	***M***	***SD***	***N***
Singular Indefinite	4.6957	1.7175	46
NQE	4.3221	1.6182	52
NQE + modifier	4.5662	1.2558	34
False Controls			
Singular Indefinite	1.2717	0.5959	46
NQE	1.8702	0.9923	52
NQE + modifier	1.3824	0.5911	34

Mean ratings for the three experimental conditions and their respective false controls are shown in Table 3. A univariate ANOVA was used to analyze the effect of subject NP type on the availability of the object wide scope interpretation using the mean rating of the interpretation of interest as the dependent measure and condition and counterbalanced block order as between-subjects factors. Block order did not show a significant effect ($F(1,126) = .169$, $p = .720$) and was excluded from all subsequent analyses. Of greater interest is that the analysis revealed no effect of subject NP type ($F(2,126) = 1.411$, $p = .415$, Figure 10).

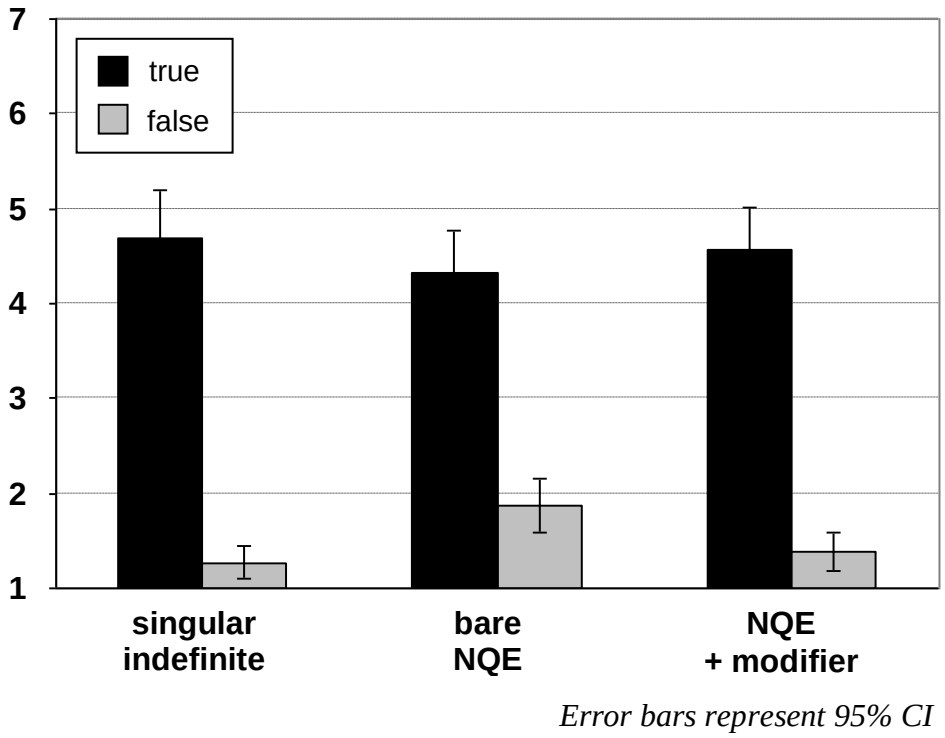


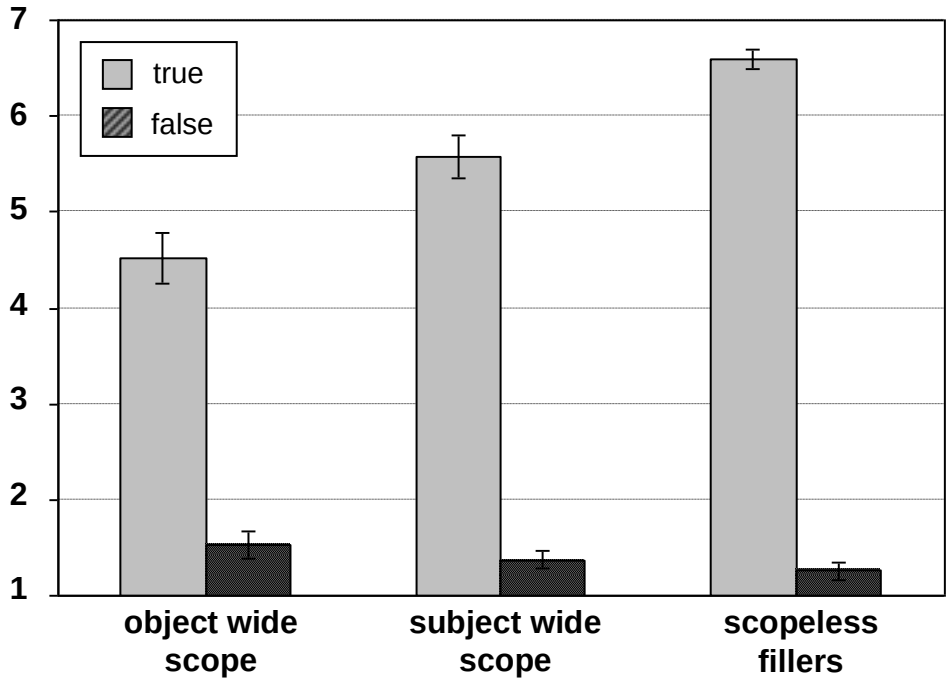
Figure 10. Experiment 2 overall results of experimental condition.

There was no significant effect of subject NP type. Means of singular indefinite and NQE + modifier conditions did not differ, supporting Reinhart (2006). The mean of the bare NQE condition did not differ from that of the other two conditions, which was not expected.

With 132 participants, this study had at least 80% power to detect a large effect, but not a medium sized one (according to Cohen's interpretation of effect size).
However, the effect of subject NP type is claimed to be robust: Reinhart (2006) found the effect using only "a couple… of informants (p.294)." We continue to gather data.

Paired t-tests revealed the differences between critical and control items to be statistically significant: $t(1,45) = 13.783$, $p < .01$ for the singular indefinite; $t(1,51) = 9.366$, $p < .01$ for bare NQE; $t(1,33) = 14.661$, $p < .01$ for NQE + modifier. This finding partially supports Reinhart (2006) in that singular indefinite subjects patterned with (collective) NQE + modifier subjects, as predicted (but see below for results of item analysis). It should be noted that Reinhart (2006)'s tests were conducted in Hebrew. She asserted that "the area of semantic judgments of quantifier scope is not, to my knowledge, subject to variations between Hebrew and English, (p.295), but this assumption alone, if incorrect, could account for the difference between her findings and ours.

A mixed ANOVA using mean interpretation type (object wide scope critical and false control, subject wide scope true and false controls, scopeless true and false controls) as the within-subjects variable and condition as the between subjects factors revealed that participants behaved as expected relative to both scoped and scopeless controls, with the mean of the subject wide scope true control above that of the object wide scope, but less than that of the scopeless true control ($F(5,645) = 843.46$, $p < .01$, pairwise comparisons, $p < .01$, Figure 11). False object wide scope controls did not differ from false subject wide scope controls ($p = 1.00$), but there was a statistically significant difference between object wide scope false and unambiguous, scopeless false scope controls ($p < .026$).



Error bars represent 95% CI

Figure 11. Experiment 2 overall results for interpretation across conditions. Means represent ratings collapsed across experimental conditions, with did not differ significantly. Comparison of means revealed that subjects behaved as expected on all controls. Scoped conditions differed with respect to each other as well as scopeless fillers.

Object wide false control items consisted of target sentences paired with a configuration that made the sentence false on the object wide scope reading. It is possible that the false configuration of bombs-cities, which depicted a single bomb in one city in none in the other two, could be read as true on the subject wide scope reading. If so, that item would receive a high rating. The difference between subject wide scope false and scopeless false controls did not differ significantly (p = .437).

The results of Experiment 2 are contra Beghelli and Stowell (1997). Table 3 shows the mean rating in the bare NQE condition to be above 4. While 4 out of 7 is

hardly a strong endorsement of the object wide scope interpretation of sentences with NQE subjects, mean ratings for bare NQE critical items were higher than those of corresponding false controls, which participants clearly rejected, and closer to subject wide scope controls, an interpretation that has been shown to be generally well-accepted (Musolino, 2009). If the reading were barred by the grammar, ratings for sentences with NQE subjects would not differ from their unambiguously bad false controls. Worse yet for that account is that 11 out of 52 (21.15%) participants in the bare NQE condition had mean ratings for the object wide scope reading of 5 or higher, so it is cannot be the case that the grammar bars the object wide scope interpretation in sentences with numerically quantified NPs. This finding is also contrary to Reinhart (2006)'s claim that the reason the object wide scope reading is dispreferred is because of the extra processing load imposed by distributivity. Reinhart (2006) clearly predicts that participants in this condition should reject the reading. Acceptance ratings should be as low as those for false controls, since computing the interpretation would exceed working memory capacity and the derivation would crash. Even allowing for scaling factors secondary to our new dependent measure, participants in the bare NQE group were predicted to produce mean acceptability ratings higher than those of false controls, but also significantly lower than those of the singular indefinite and NQE + modifier groups. So Reinhart (2006) appears to be right about the facts concerning distributive *vs.* collective object wide scope, but not about why those facts obtain.

Overall, 33 out of 132 participants accepted the object wide scope interpretation for all items, rating each critical item 5 or higher. An additional 7 participants consistently rejected the object wide scope interpretation, rating each critical item 3 or

lower. The 33 participants who always accepted the interpretation were spread across all three experimental conditions: 17, 11 and 5 in conditions 1, 2 and 3 respectively. There were 3 consistent rejectors in the singular indefinite condition and 4 in the bare NQE condition, but none in NQE + modifier. We leave open the question of whether these participants represent populations with different grammars.

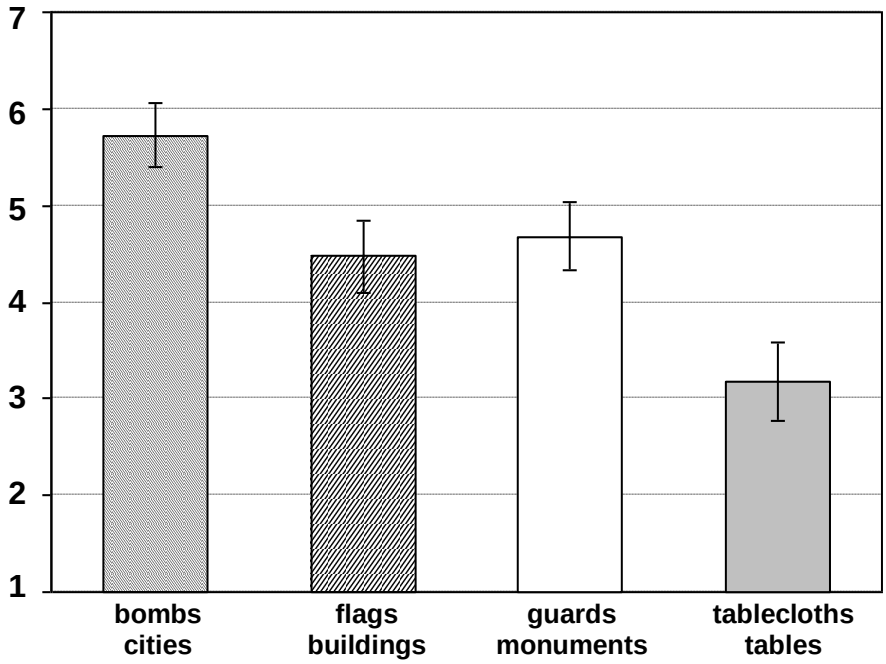


Figure 12. Experiment 2 item effects across conditions.

Item effects found in Experiment 1 were replicated in Experiment 2. When the doctors-patients item was excluded, tablecloths-tables was least preferred. The difference between tablecloth-tables and the other three items was statistically significant.

Experiment 2 largely replicates the item effects found in Experiment 1 (doctors-patients was not included in Experiment 2, Table 4). Recall that Experiment 1 revealed clear differences among acceptability ratings of lexical items, with bombs-cities rated higher than all other items, doctors-patients rated lower than all other items, and the other

three items not differing from one another. A mixed model ANOVA using lexical item as the repeated measure and condition as the between-subjects variable revealed a significant main effect of item ($F(3,387) = 50.19$, $p < .01$, Figure 12). There was no effect of condition ($F(2,129) = 0.71$, $p = .49$). Pairwise comparisons with Bonferroni correction for multiple comparisons revealed a statistically significant difference between bombs-cities and all other items ($p < .01$) and between tablecloth-tables and all other items ($p < .01$) but the difference between flags-buildings and guards-monuments was not significant ($p = 1.00$).

TABLE 4. **EXPERIMENT 2**

ACCEPTABILITY OF OBJECT WIDE SCOPE READING BY ITEM, *N* = 132

Critical Items	***M***	***SD***
bombs-cities	5.7273	1.9618
flags-buildings	4.4773	2.1701
guards-monuments	4.6818	2.0465
tablecloths-tables	3.1742	2.3751
False Controls		
bombs-cities	1.9318	1.6261
flags-buildings	1.3788	1.2013
guards-monuments	1.4091	1.1849
tablecloths-tables	1.4242	1.1859

Our failure to replicate previous findings that the object wide scope interpretation is (at best) massively dispreferred was surprising given the literature reporting both intuitions (Beghelli & Stowell, 1997; Reinhart, 2006) and experimental results (Musolino, 2009). Overall, participants in the bare NQE condition rated the object wide

scope interpretation as high as participants in the other two conditions did. We believe that the reason for this result lies in our choice of lexical items.

These experiments were the first to use a rating scale instead of a polar response. In an effort to calibrate the results using the rating scale to previous results, Experiments 1 and 2 included a single item from Musolino (2009), *Three boys are holding two balloons,* changed minimally to *Three girls are holding two balloons,* paired with both Musolino (2009)'s highest-rated each-all configuration and lowest-rated object wide scope configuration. This item allowed us to perform post hoc comparisons of our critical items to an item that has been shown to be dispreferred and for which we had data. A repeated-measures ANOVA using the 4 planned critical items plus *Three girls are holding two balloons* revealed a significant effect of item, $F(4,524) = 66.48$, $p < .01$. Not only did girls-balloons have the lowest mean rating both overall and in each condition (Table 5), pairwise comparisons with Bonferroni correction for multiple comparisons revealed that girls-balloons differed significantly from all items ($p < .01$) except tablecloths-tables ($p = .105$), suggesting that the choice of lexical item is driving our unexpected results in the bare NQE condition.

TABLE 5. **EXPERIMENT 2**

ACCEPTABILITY RATINGS FOR

THREE GIRLS ARE HOLDING TWO BALLOONS, N = 132

	M	***SD***
Overall	2.5758	2.0308
singular indefinite	2.3478	1.9232
bare NQE	2.8269	2.2468
NQE + modifier	2.5000	1.8299

General Discussion

This paper tested experimentally two competing accounts of why the object wide scope interpretation of sentences with multiply-quantified NPs is difficult, if not impossible, to obtain. Beghelli and Stowell (1997) presented a strictly syntactic account based on the requirement that all quantified NPs must have their features checked against their functional projections for agreement at LF. Because bare numeral NPs that are also sentential subjects must move to the specifier position of the highest functional projection, the subject NP will always and only take wide scope, predicting that the object wide scope interpretation is impossible. The results of Experiment 2 provide strong evidence against this account. The object wide scope interpretation of bare numeral NPs received a mean rating above 4, significantly higher than the means of false controls and approaching that of the well-accepted subject wide scope controls. Worse yet for this account, there was a subset of participants in the bare NQE condition who unequivocally accepted the object wide scope reading for all 4 critical items.

On Reinhart (2006)'s account, the general failure of the object wide scope reading is not due to the grammar, but to the additional cost incurred by covert movement, which ultimately results in processing failure. Bare numeral object NPs may move covertly to take wide scope at LF. All interpretations that result from covert movement are then paired with their derivations to form a reference set. This reference set must be checked against two requirements: feature agreement and non-identity with interpretations resulting from overt movement. The resulting reference set for sentences with subject NPs that can receive both a collective and a distributive reading is too large to hold in working memory while its features and non-identity are checked. Therefore, their

derivation crashes. However, the addition of a collectivizing modifier reduces the size of reference set (since it eliminates at least one possible interpretation), predicting that the reading will be easier to access. The findings reported here support Reinhart (2006)'s claim that a collectivizing modifier improves the object wide scope reading of sentences with numeral indefinite object NPs. With a sample large enough to detect a large effect, we did not find a statistically significant difference between participants' mean acceptability ratings in the singular indefinite and bare numeral + modifier conditions.

However, our findings do not support Reinhart (2006)'s claim that the dispreferred status of the object wide scope reading results from processing failure. Participants in our bare numeral condition showed no such processing failure, or at least no more than that incurred by participants in the other two groups, as the mean acceptability ratings of the three groups did not differ.

Our result for sentences with two bare numeral quantified NPs constitutes a failure to replicate a well-established finding reported in both the theoretical and experimental literatures. Most participants in the study reported here accepted the object wide scope reading, if not emphatically, at least on a par with sentences with singular indefinite subject NPs. The results of our item analysis suggest an explanation.

We propose that our failure to replicate the inaccessibility of the object wide scope interpretation is due to lexical effects. We tested a subset of Reinhart (2006)'s examples (with slight modifications to hold non-critical features constant). The predicates we used were *explode (in), fly, stand* and *cover*. In contrast, Musolino (2009) used three tokens of *hold* and one of transitive *walk.* In an experiment designed to test children's interpretations of all possible readings, scope-independent as well as scoped distributive

and collective, Syrett and Musolino (in prep) used *painted*, as it allows for readings that are distributed in both space and time. Their adult controls behaved comparably to the adult controls in Musolino (2009), rejecting the object wide scope reading regardless of its distributive/collective nature.

We did, however, replicate Musolino (2009)'s results with the one lexical item we used in common, *hold,* which we included as a calibration item. And as importantly, our Experiment 2 replicated our item results from Experiment 1.

The pattern of acceptability ratings reported here and in previous studies suggests that participants may be sensitive to whether a verb is a true transitive or an intransitive of the unaccusative type. Although the two classes of verb have subtle differences in the relation between the verb and its argument (Levin and Rappaport Hovav, 1995), those differences were treated as irrelevant by the theories tested in our experiments. Both Beghelli and Stowell (1997) and Reinhart (2006) treated transitive verbs on a par with unaccusatives. What mattered was that the object noun phrase be required by the verb, and therefore originate within the verb phrase. Yet findings reported here and elsewhere for the transitive verbs *examine, hold* and *paint* were lower – much lower – than our findings for *explode, hang* and *stand.* Only *cover,* a transitive verb, patterned with the unaccusatives, and the extent to which it was dispreferred depended on whether the least preferred predicate of all, *examine*, was among the other items. Also, *cover,* as used here, has an unaccusative flavor. In the sentence *A tablecloth covered three tables, tablecloth* is the sentential subject, but its role is non-agentive - it is an instrument. Comparing *A tablecloth covered the table* with *John covered the table with a tablecloth* makes the agent-instrument distinction clearer. The subjects of the other transitive verbs clearly are

agentive. Whatever the facts turn out to be, it is clear that our participants did not treat all verb phrase-internal noun phrases equally, as the theories we tested predicted they would.

There is support in the literature for the lexical hypothesis. Lexical effects are well-established in the literature on structural ambiguity resolution, for example (Frazier & Fodor, 1978). We think that lexical effects may be driving ambiguity resolution in the distributive object wide scope case, as well.

Appendix A

DEMOGRAPHIC INFORMATION SHEET
Participant Demographics – VERSION I, Subject Pool

Recall from the *Consent Form to Participate in a Research Study* you just signed that this research is confidential. The information requested here may help us to better interpret your experimental data, but providing it is optional.

Gender: F ☐ M ☐

Do you consider yourself Hispanic/Latino/Latina? Yes ☐ No ☐

Is English one of your native languages? Yes ☐ No ☐
Please list all other native languages.
_____

For all languages that you speak at home *besides English*, please list the age of acquisition, and percentage of time spoken.

_____

_____

Have you studied any other languages? Yes ☐ No ☐
If yes, please indicate which one(s), at what age you started, and how long studied.

_____

_____

_____

What is your ethnicity? (Choose all that apply.)
African American or Black ☐ American Indian/Alaska Native ☐ Asian ☐ White ☐
Native Hawaiian or Pacific Islander ☐ I do not wish to indicate my ethnicity. ☐

Age: _____ I do not wish to indicate my age. ☐

Appendix B

Test Items

Experiment 1

A bomb exploded in three cities.

A doctor is examining three patients.

A flag is hanging in front of three buildings.

A guard is standing in front of three monuments.

A tablecloth is covering three tables.

Experiment 2

Condition 1

A bomb exploded in three cities.

A flag is hanging in front of three buildings.

A guard is standing in front of three monuments.

A tablecloth is covering three tables.

Condition 2

Two bombs exploded in three cities.

Two flags are hanging in front of three buildings.

Two guards are standing in front of three monuments.

Two tablecloths are covering three tables.

Condition 3

Two bombs exploded simultaneously in three cities.

Two identical flags are hanging in front of three buildings.

Two guards are standing together in front of three monuments.

Two coordinated tablecloths are covering three tables.

Appendix C

Verbal Instructions for Experiment 1

As you know, this experiment is titled Quick Sentence Interpretations. We will be presenting you with pairs of sentences and pictures and asking you whether the picture and sentence match. We are interested in the possible configurations of people or objects that the sentences allow, not in the details in the pictures themselves.

For example, some sentences will be about flags flying, girls holding balloons, and guards and monuments. When making your decision, disregard whether the flags look static or in motion, or if the guards are wearing uniforms or not. Some of the pictures are highly stylized because they are only meant to provide you with a map of the configuration. And people will usually be shown as stick figures.

Remember, it's your opinion of the configurations that interests us.

The computer's instructions will guide you through the mechanics of the experiment. Feel free to call the experimenter if you have any questions.

Appendix D

Automated Instructions for Experiment 1

Screen 1

By participating in this experiment,
you will be helping us to select the items
we use in future experiments.

On each screen, you will see a sentence, a picture and a rating scale.
Your job is to use the rating scale to indicate whether
the configuration in the picture matches the situation described by the sentence.

Screen two

The question you will be answering, "Is this a possible configuration?"

and

a 7-point Likert scale will appear at the bottom of every screen. The scale ranges from

1 = *Definitely no*

to

7 = *Definitely yes.*

A LOW rating means that you think the picture DOES NOT MATCH the sentence well.

The scale will remain visible until you make your selection.

Screen 3

Some of the items might seem a bit odd, but they are not meant to be tricky.

Please take each one at face value, consider it carefully, and give us your honest opinion.

You'll respond by pressing the number key that corresponds to your answer.

Each subsequent item will appear automatically.

Screen 4

The first three items are for practice.

If you have questions or need assistance, ask the experimenter.

Remember, you must use the number keys to enter your answer.

References

Beghelli, F. & Stowell, T. (1997). Distributivity and negation: The syntax of *each* and *every*. In A. Szabolcsi, (Ed.), *Ways of Scope Taking* (pp. 71-107). Dordrecht: Kluwer.

Crain, S., & Thornton, R. (1998). *Investigations in universal grammar: A guide to research on the acquisition of syntax and semantics*. Cambridge: The MIT Press.

Frazier, L. & Fodor, J.D. (1978). The sausage machine: A new two-stage parsing model*. *Cognition,* 6, 291-325.

Grudzinska, J. (2010). Empirically adequate semantics for sentences with two NQE. Talk presented at Rutgers University, March 18, 2010.

Kempson, R.M., & Cormack, A. (1981). Ambiguity and quantification. *Linguistics and Philosophy,* 4(2), 259-309.

Levin, B., & Rappaport Hovav, M. (1995). *Unaccusativity: At the syntax-lexical semantics interface*. Cambridge, MA: MIT Press.

Link, G. (1998). *Algebraic semantics in language and philosophy*. Stanford, CA: CSLI Publications.

May, R.C. (1990). The grammar of quantification. In Hankamer, J. (Ed.) *Outstanding dissertations in linguistics: A Garland series*. New York: Garland. (Reprinted MIT Dissertation, 1977)

Musolino, J. (2009). The logical syntax of number words: Theory, acquisition and processing, *Cognition,* 111, 24-45.

Radford, A. (1997). *Syntactic theory and the structure of English: A minimalist approach.* New York: Cambridge UP.

Reinhart, T. (2006). Scope shift with numeral indefinites: Syntax or processing?. In S. Vogeleer & L. Tasmowski (Eds.), *Non-definiteness and plurality* (pp. 291-310). Philadelphia: John Benjamins.

Syrett, K., & Musolino, J. (in preparation). The linguistic expression of numerical relations.