THE PRAGMATICS OF QUANTIFIER SCOPE:
A CORPUS STUDY[*]

SCOTT ANDERBOIS[*]
ADRIAN BRASOVEANU[**]
ROBERT HENDERSON[**]
[*]*University of Connecticut,* [**]*UC Santa Cruz*

1 Introduction

One of the most well-studied phenomena in natural language semantics has long been the question of what readings are possible in various types of sentences with multiple quantifiers. Semanticists have generally been concerned with developing theories to capture the range of *possible* readings for such sentences. This is what Higgins & Sadock (2003) have dubbed the problem of **scope generation**. To take a simple example like (1), then, the job of semantics is (*i*) to argue that the sentence is truly ambiguous, i.e., it has two possible readings depending on which quantifier 'we think of first' (the need to postulate ambiguity is not immediately obvious, particularly for example (1)), and (*ii*) to capture the ambiguity by providing adequate representations for those readings.

(1) Every doctor talked to a patient.

a. *every* $\gg$ *a*: every doctor is such that s/he talked to a patient

b. *a* $\gg$ *every*: a patient is such that every doctor talked to her/him

While the semantic literature is rich with intuitions about the factors influencing the *actual patterns of usage* of sentences like (1), semantic theories are (quite rightly) not tasked with predicting which reading, (1a) or (1b), it has when it is used in a particular context. That is,semantics is generally not concerned with the problem of **quantifier scope disambiguation** (QSD) (scope prediction in Higgins & Sadock 2003's terms). While QSD is generally agreed to be beyond the purview of

[*]First of all, we would like to thank the Law School Admission Council (LSAC) for access to the practice test materials used in the analysis. We would also like to thank Chris Manning for first suggesting that logic puzzles would be a good corpus to study quantifier scope and Pranav Anand, Lucas Champollion, Robert Daland, Alexis Dimitriadis, Jakub Dotlacil, Donka Farkas, Chris Potts, Matt Wagers and the participants in the UCSC Corpus Linguistics Group and the audiences at CUSP 3, MIT Syntax Square, ESSLLI 2011 PUQOL, SuB 2011 and UCLA Linguistics Colloquium Series for very helpful feedback. This research was supported by an SRG (2009-2010) and an FRG (2010-2011) grant from the Committee on Research, UC Santa Cruz.

semantic theory, it has been argued to fall squarely within the domain of *pragmatics* (see especially Saba & Corriveau 2001).

The present paper seeks to contribute to the study of the pragmatics of QSD through statistical modeling of the factors influencing quantifier scope in sentences with two quantifiers in a naturally occurring (though controlled) body of text: LSAT Logic Puzzles. The paper is organized as follows: **§2** surveys previous literature, identifying likely predictors of quantifier scope; **§3** introduces our corpus; **§4** reports the main findings and the series of mixed-effects logistic regression models used to obtain them; and **§5** concludes.

2 Predictors and Previous Literature

While semanticists have generally not developed accounts of QSD, the semantics literature is filled with intuitions about various factors which influence quantifier scope in English and cross-linguistically, though we restrict our attention here to English. In addition to these, there are numerous psycholinguistic studies which have shown various factors to play a role in QSD. In §2.1, we review the main factors which have been proposed based on these literatures. §2.2 briefly discusses the only previous large-scale corpus study of quantifier scope of which we are aware: Higgins & Sadock (2003).

2.1 Possible predictors of quantifier scope

One of the earliest factors to receive attention in research on QSD is the syntactic position of the two quantifiers, and more specifically, **grammatical function**. One of the first works to deal explicitly with QSD is Ioup (1975), who conducted a survey of informants from various languages, with participants indicating the relative availability of the two possible scopes on a 5-point scale (1=unambiguous wide scope indefinite, 5=unambiguous narrow scope indefinite) for sentences like those in (3). While she does not report detailed numerical results from this survey, she concludes that QSD is sensitive to the grammatical function hierarchy in (2).

(2) **Ioup (1975)'s hierarchy:** S $\gg$ Prep $\gg$ IO $\gg$ O

(3) a. Joan told a child the story at every intersection. **Preferred scope:** *every* $\gg$ *a*
b. Joan told everyone the story at an intersection. **Preferred scope:** *a* $\gg$ *every*

One thing to note about Ioup (1975)'s methodology (a criticism which holds of many early works on QSD) is that the experimental task quite explicitly is for participants to determine whether and to what extent a given sentence is ambiguous. The validity of this sort of 'ambiguity judgment', however, is at best controversial. In the literature on semantic fieldwork methodology, for example, Matthewson (2004) argues convincingly that such tasks are not reliable sources of evidence since they conflate linguistic judgments with linguistic analysis. While they do not explicitly mention Ioup (1975) in this vein, Kurtzman & MacDonald (1993) argue that from a psycholinguistic perspective, experimental designs which draw attention to the potential ambiguity itself are unnatural in ways which might bias the data. With this said, Ioup's hierarchy provides valuable intuitions about the role of grammatical function in QSD, many of which have been borne out in subsequent psycholinguistic work (Micham et al. 1980, Gillen 1991, Kurtzman & MacDonald 1993, Tunstall 1998, and Anderson 2004 *inter alia*).

For English, it is difficult to separate the effect of grammatical function from another proposed predictor: **linear order**. This is because the two factors are heavily correlated given the relatively rigid word order of English.[1]

(4) Every professor saw a student. **Preferred scope:** *every* ≫ *a*

(5) A student saw every professor. **Preferred scope:** *a* ≫ *every*

The effect of linear order has been much debated in previous literature, with many studies positing an effect of linear order (e.g., Van Lehn 1978, Fodor 1982, Gillen 1991, Tunstall 1998), but others rejecting this idea (e.g., Ioup 1975, Micham et al. 1980, Kurtzman & MacDonald 1993). One source of this disagreement, we believe, is that the methodologies employed have led researchers to only consider particular kinds of closely related sentences rather than a broad sample across different types of constructions. To take an extreme example, the test sentences in Micham et al. (1980) always have one of the quantifiers occurring in goal PP headed by *to*, and generally have both quantifiers in postverbal position (e.g., double object constructions or PPs headed by *about*). By studying linear order in a corpus of naturally occurring text, the present study allows us to consider data from a wide range of different constructions in a way that has been impractical for more focused studies using surveys and other psycholinguistic methodologies.

The third factor which has been argued to influence QSD in prior literature is the **lexical realization** of the two quantifiers. One early work stressing the importance of lexical factors is once again Ioup (1975), who argues that lexical factors are the single most important factor in determining actual quantifier scope. For example, the quantifier *every* in (6) below takes wide-scope more readily than *all*. Ioup argues that the effects of lexical realization can be summarized by the quantifier hierarchy in (7).

(6) a. She knows a solution to every problem. **Preferred scope:** *every* ≫ *a*
b. She knows a solution to all problems. **Preferred scope:** *a* ≫ *all*

(7) **Ioup (1975)'s Quantifier Hierarchy:**
each ≫ *every* ≫ *all* ≫ *most* ≫ *many* ≫ *several* ≫ some_{pl} ≫ *a few*

While essentially all authors agree that lexical factors play some role in QSD, subsequent studies have emphasized its effect less than Ioup (1975) (the investigation of *each* vs. *every* in Tunstall 1998 is a notable exception). One of the benefits of mixed-effects regression analyses of the sort we present in §4 is that they allow us to directly address the question of the importance of lexical factors *relative* to other predictors without focusing only on a very small set of lexical items.

The fourth predictor found in previous literature to influence QSD is **world knowledge** and in particular, numerical typicality. For example, Saba & Corriveau (2001) point out that the narrow scope reading of *every* in (8) is dispreferred because it would require an individual to participate in the *living-in* relation with an atypically large number of cities.

(8) A doctor lives in every city. **Preferred scope:** *every* ≫ *a*

[1] Another factor that is highly correlated with both linear order and grammatical function in English is c-command. We do not treat c-command as a factor separate from linear order and / or grammatical function here and we leave to future work the question of whether c-command plays an independent role, leaving open the possibility that linear order and / or grammatical function may be serving as a proxy for c-command.

Based on such examples, Saba & Corriveau (2001) propose a formal model of the world knowledge used in QSD based on the number of restrictor entities that typically participate in the nuclear scope relation. Srinivasan & Yates (2009) build on this work by showing that numerical typicality can be extracted from a large corpus and applied successfully to QSD. Using a hand-picked corpus of 46 items, Srinivasan & Yates (2009) show that automatically extracted information about numerical typicality significantly improves prediction, especially for inverse scope.

While acknowledging that world knowledge plays a significant part in real world QSD, trying to incorporate it into a model of QSD – possibly together with other important factors like cross-sentential anaphoric relations between various quantifiers – is a much broader endeavor left for future research. In order to factor out world knowledge as much as possible, we have chosen to base our investigation on a body of text that is designed to minimize the assumed world knowledge; see §3 below for more discussion of the LSAT logic puzzles corpus.

To summarize, previous literature indicates that the following factors are plausible predictors for QSD, with the last two being more or less universally agreed to play some role:

(9) **Plausible Predictors of QSD:**
(*i*) linear order, (*ii*) grammatical function, (*iii*) lexical realization, and (*iv*) world knowledge.

2.2 Higgins & Sadock (2003)'s corpus study

To date, there is only one large-scale corpus study of QSD that we are aware of: Higgins & Sadock (2003). To investigate the factors influencing QSD, Higgins & Sadock (2003) build a corpus with sentences from the Wall Street Journal portion of the Penn Treebank hand-tagged for actual quantifier scope. As in the present study, all of the sentences considered are ones in which there are exactly two quantificational expressions. The resulting corpus consists of 893 sentences, an impressive size given the need for scope to be hand-coded. One caveat to this, however, is that 61.2% of the sentences in their corpus are tagged as having "no scope interaction", resulting in only 345 sentences where scope is determined (we return to this point shortly).

Given this corpus, Higgins & Sadock (2003) construct three different computational models (Naive Bayes, Maximum Entropy, Single Layer Perceptron) aimed at classifying the scope of a given example based upon a wide range of factors including lexical realization and the relative hierachical position of the two quantifiers as determined by c-command. Looking across the three models, they found both c-command and various factors relating to lexical realization to be important, alongside other factors to be discussed in a moment.

While Higgins & Sadock (2003) is noteworthy for being the first large-scale empirical study of QSD of its kind, there are two aspects of their corpus and analysis which we believe can be improved on, especially given our aim of studying QSD from a linguistic perspective rather than a natural language processing one. First, Higgins & Sadock (2003) explicitly omit NP/DPs where the quantifier is *a(n)*, a quantifier which is very much of interest to linguists.[2]

Second, many of the most active features in Higgins & Sadock (2003)'s models are things such as whether a conjoined node or various sorts of punctuation (commas, colons, quotation marks, etc.) intervene between the two quantifiers. All of these factors are found to be strong predictors of "no scope interaction", the majority of the sentences in their corpus. While these punctuation

[2] Their rationale for this decision has to do with the difficulties of determining scope for generics. While this is a real concern, we believe that since scope must be hand-tagged in any case, it is readily avoidable.

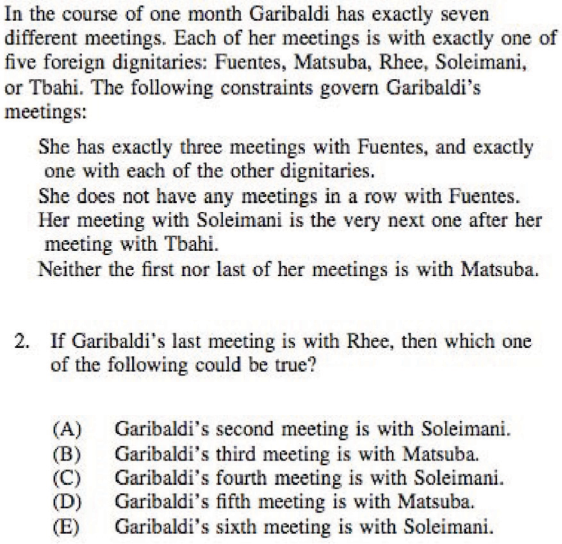
marks have various uses, this finding suggests that many of these examples involve appositive material, quotations, and other cases where the two quantifiers cannot interact, i.e., where the two quantifiers are in different *scopal domains*. Such sentences, therefore, do not contribute to our understanding of QSD any more than sequences of separate sentences would. Higgins & Sadock (2003)'s findings in this area may be quite useful for other applications such as determining when a possible quantifier scope ambiguity has arisen in a text, but this is separate from our current goal.[3]

3 A constrained, naturally occurring corpus

The data in our corpus come from prior instances of the Law School Admission Test (LSAT), a standardized test designed to assess the verbal, logical, reading, and reasoning skills of prospective law school students. In addition to a writing sample and a "variable" or "experimental" section, the LSAT consists of three type of sections: logical reasoning, reading comprehension, and analytical reasoning. Our data come from the third of these (analytical reasoning), questions colloquially known as "logic puzzles", a label we will use in what follows.

LSAT logic puzzles follow a standard format exemplified in (10) below. The Introduction establishes all of the individuals in question and what categories they belong to. Following this, the Laws section puts forth a set of statements which characterizes the facts of the world for the purposes of the questions that follow. Finally, there is a multiple choice Question and a set of four-five Answers (generally, there are several Question-Answers pairs per scenario). While the questions and answers are sometimes helpful in determining the intended scope, the sentences in our corpus come only from the first two sections.

(10)



In the course of one month Garibaldi has exactly seven different meetings. Each of her meetings is with exactly one of five foreign dignitaries: Fuentes, Matsuba, Rhee, Soleimani, or Tbahi. The following constraints govern Garibaldi's meetings:

} Introduction

She has exactly three meetings with Fuentes, and exactly one with each of the other dignitaries.
She does not have any meetings in a row with Fuentes.
Her meeting with Soleimani is the very next one after her meeting with Tbahi.
Neither the first nor last of her meetings is with Matsuba.

} Laws

2. If Garibaldi's last meeting is with Rhee, then which one of the following could be true?

} Question

(A) Garibaldi's second meeting is with Soleimani.
(B) Garibaldi's third meeting is with Matsuba.
(C) Garibaldi's fourth meeting is with Soleimani.
(D) Garibaldi's fifth meeting is with Matsuba.
(E) Garibaldi's sixth meeting is with Soleimani.

} Answers

3.1 Why Logic Puzzles?

There are several reasons why we have chosen LSAT logic puzzles as the source for our QSD corpus. First, from a practical perspective, LSAT logic puzzles contain a much higher number

[3] It should be noted that not all sentences in the "No scope interaction" category are of this sort. For example, there are sentences where two quantifiers may occur in the same scopal domain, e.g., *One woman bought one horse*, and yet there is no clear way to determine narrow or wide scope.

of sentences with two or more quantifiers relative to other registers of English. Second, in more neutral registers of English, it is not always the case that disambiguating between the various possible scopes is required of the hearer (see Tunstall 1998, for example, for more discussion of this issue with respect to universal-existential scope interactions). In contrast, LSAT test takers are expected to select a single correct answer, which generally means that they must settle on a particular reading for potentially ambiguous sentences with more than one quantifier.

Third, as an aptitude test, the LSAT is explicitly designed to avoid forcing test takers to rely on facts other than those which are present in the text itself. While there are undoubtedly some cases where test takers nonetheless bring world knowledge to bear, the writers generally err on the side of being overly explicit about the facts of the world. In essence, this property of logic puzzles serves to control for world knowledge, allowing for a clearer picture of the linguistic factors impacting QSD. This control is admittedly imperfect, but nonetheless serves to reduce the confounding impact of world knowledge in a way that would be quite difficult with more naturalistic corpora.

3.2 Tagging the data

Scopal Domains. One of the issues we discussed with Higgins & Sadock (2003) (again, given our fundamentally linguistic goals) was that many of their sentences had the two quantifiers in different clauses separated by conjunction, apposition, or quotation in ways that prevented any scopal interaction between them. In other words, Higgins & Sadock (2003) take the sentence to be the domain of quantifier scope regardless of its complexity. However, it is often clear that a sentence consists of multiple scopal domains, as in (11) below where the quantifiers *three* and *two* occur in different conjuncts of a coordinate structure. The lack of scopal interaction between them is therefore a consequence of the syntax/semantics of quantifier scope, not its pragmatics.

(11) **[** Joe ate three oranges **]** and **[** Pam ate two apples **]**.

In order to address such cases, which are quite numerous, we treat an example like (11) as consisting of the two scopal domains indicated in brackets, with each domain having one quantifier. That is, we do not include examples like (11) in our corpus because they do not contain a scopal domain with multiple quantifiers. Beyond coordinate structures, we consider quotations and appositives to also constitute separate scopal domains from the sentences in which they occur and take them into consideration only if there are multiple quantifiers in one or more of the resulting scopal domains.

Response variable: Scope. Tagging sentences for quantifier scope is by its nature time-consuming and requires a coder with linguistic training. Extracting and tagging the relevant sentences with multiple quantifiers out of the mass of sentences in our logic puzzle corpus took place in three steps. First, we separated the data into individual sentences and then further into scopal domains as outlined above. Then, we enlisted undergraduate semantics students to identify sentences with multiple quantifiers and to provide a first attempt at tagging them for the response variable (scope) and the three predictors we considered (linear order, grammatical function, and lexical realization). Each sentence was seen by several students, providing a reasonable preliminary tagging. Finally, the authors went through the corpus by hand adjudicating cases where the students disagreed and correcting cases where the students were mistaken. Given how involved this process was, no effort was made to quantify inter-annotator agreement since this would require additional skilled coders.[4]

The beginning of our tags is marked by **&**, and the end of a tag is marked by **#**, with subtags being separated by **_**. Scope was coded numerically, with **1** corresponding to widest scope and other numbers indicating narrower scope, as in (12) below. Quantifiers with no relative scope (mainly cumulative readings) were 'co-tagged' with the same number, as in (13). This is merely a convenience for the examples with only two quantifiers that we consider in this paper: we could have just as easily ignored such sentences. However, tagging scopal 'level' with the same number is necessary for sentences with more than two quantifiers, since two quantifiers may take wide or narrow scope relative to a third yet not have any scopal relation relative to one another, as in (14).

(12) Each**&1#** tape is to be assigned to a different**&2#** time slot.

(13) Exactly six**&1#** employees must be assigned to exactly three**&1#** committees.

(14) Exactly six**&2#** of seven**&1#** jugglers are each**&3#** assigned to exactly one**&4#** of three**&1#** positions.

In cases where no truth conditional difference was clear, we used the felicity of "such that" paraphrases as our ultimate criterion.

Three predictor variables. Following the discussion of the previous literature in §2 above, three explanatory variables were tagged: linear order, grammatical function, and lexical realization of the quantifier. Since linear order is implicit in the tagging, this was not explicitly tagged (i.e., the order is recoverable from the linear position of the tags themselves). For grammatical function, we distinguished four syntactic roles as follows: **S**ubject, **O**bject, **P**ivot, **A**djunct (including prepositional phrases). For prepositional phrases, we tagged individual prepositions separately.[5] As far as the final subtag, i.e., lexical realization, is concerned, we treated complex determiner material like **more.than.two** separately from **two** (see §4 for more examples of lexical tags).

Here are some examples of tagged sentences:

(15) a. Each**&1_S_each#** tape is to be assigned to a different**&2_to_a.different#** time slot, . . .
b. . . . and no**&1_S_no#** tape is longer than any**&2_than_any#** other tape.

(16) Each**&1_S_each#** professor has one or more**&2_O_one.or.more#** specialities.

(17) The judge of the show awards exactly four**&1_O_exactly.four#** ribbons to four**&1_to_four#** of the dogs.

[4] Higgins & Sadock (2003) did quantify inter-annotator agreement and found that their rate was fairly low relative to the normal standards for classification tasks (76.3%). They quite reasonably conclude that this rate of agreement is good enough given the complexity of the task. Note, however, that inter-annotator agreement may very well be (much) higher for our specialized, 'minimal ambiguity' LSAT corpus.

[5] Though the analysis presented in this paper focuses exclusively on sentences where at least one quantifier is a subject or a direct object.

4 Analysis

The Dataset. We focus on sentences with 2 quantifiers only in this paper. Furthermore, we remove the cumulative sentences and we focus on **S**ubject and **O**bject only, i.e., we drop the other grammatical functions. We are left with 497 observations.

However, these observations count some of scope-resolution events twice:[6] some sentences have both an **S** and an **O** quantifier and the scope of one completely determines the scope of the other. There are 139 such doubly counted sentences and we randomly sample one quantifier from each of them. The final number of observations is $N = 358$.

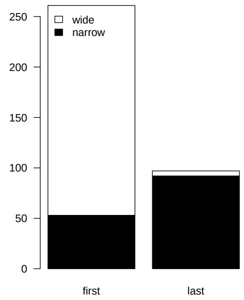
Response variable. Our response variable is SCOPE, a factor with 2 levels: **narrow** and **wide**. We designate **wide** scope as the 'success' level.

Fixed effects. We treat linear order (LIN.ORD) and grammatical function (GRAM.FUN) as fixed effects given the fact that we focus exclusively on binary contrasts for each of them.

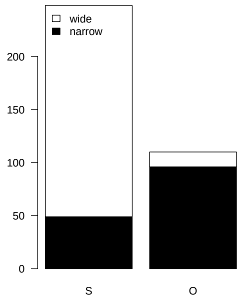
LIN.ORD is a factor with 2 levels, **first** (the quantifier comes first in the scopal domain) and **last** (the quantifier is not the first one in the scopal domain). We designate **first** as the reference level for two reasons. On one hand, more quantifiers come first in our corpus and we want our estimate of the reference level to be as accurate as possible. On the other hand, we expect quantifiers that are first in the scopal domain to exhibit a preference for wide scope / 'success', so it is natural to ask the question: what is the effect on scope (if any) of occurring in a non-initial position in the scopal domain?

GRAM.FUN is also a factor with 2 levels: **S**ubject and **O**bject. We designate **S** as the reference level, both because most quantifiers occur in subject position in our corpus and because we expect quantifiers in subject position to exhibit a preference for wide scope / 'success', so it is natural to ask the question: what is the effect on scope (if any) of occurring in a non-subject position?

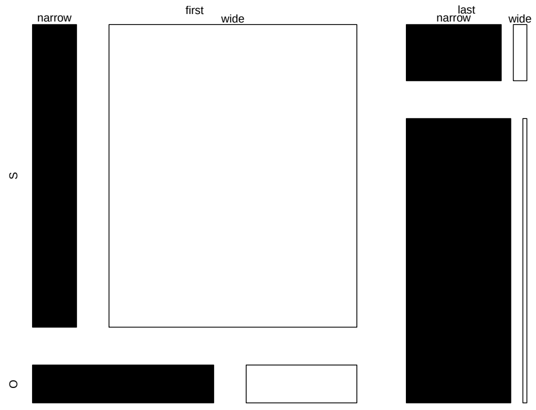
LIN.ORD by SCOPE



GRAM.FUN by SCOPE



LIN.ORD by GRAM.FUN by SCOPE



The bar plots of LIN.ORD by SCOPE and GRAM.FUN by SCOPE above indicate that quantifiers that come first in the scopal domain have a fairly strong preference for wide scope, just as quantifiers in

[6] We are indebted to Pranav Anand, Alexis Dimitriadis, Chris Potts and Matt Wagers for very helpful discussions of this point and for suggesting various solutions.

subject position do. In contrast, quantifiers that come last in the scopal domain have a very strong preference for narrow scope, just as quantifiers in object position do.

In addition, the mosaic plot of LIN.ORD by GRAM.FUN by SCOPE raises the possibility that LIN.ORD might interact with GRAM.FUN: if a quantifier comes last, it has a strong preference for narrow scope whether it is a subject or not; but if a quantifier comes first, it has a strong preference for wide scope only if it is a subject – if it is an object, it does not exhibit a clear scopal preference.

Importantly, however, arguing for or against such an interaction cannot proceed on the basis of visual inspection alone: the mosaic plot does not provide clear evidence one way or another. In addition, we need a way to quantify if the interaction is simply an artifact of the sample we happen to have or if we can confidently claim that it is an actual contribution of the fixed effects.

Similarly, given that LIN.ORD and GRAM.FUN are correlated in English, we cannot disentangle their main effects by visual inspection of the bar and mosaic plots or by introspective intuitions of the kind standardly used in linguistic research.

Finally, whatever generalizations we make about LIN.ORD and GRAM.FUN have to take into consideration the fact that quantifiers might have lexically encoded preferences for wide or narrow scope. That is, to determine whether linear order and grammatical function actually have an effect on scopal preference and if they do, to determine the extent of that effect, we need to also take into account the lexical realization of various quantifiers .

For all these reasons, we cannot obtain reliable empirical generalizations about QSD without the help of adequate statistical models, to which we will turn presently.

The sentences below exemplify wide and narrow scope for each of the four GRAM.FUN × LIN.ORD combinations.

(18) GRAM.FUN=S, LIN.ORD=first
   a. SCOPE=wide: **Each chair** is occupied by exactly one representative.
   b. SCOPE=narrow: **Exactly one child** sits in each chair.

(19) GRAM.FUN=S, LIN.ORD=last
   a. SCOPE=wide: Every week **six crews** – A,B,C,D,E,F – were ranked from first (most productive) to sixth (least productive).
   b. SCOPE=narrow: On each day of other days of hiring, **exactly one** worker was hired.

(20) GRAM.FUN=O, LIN.ORD=first
   a. SCOPE=wide: He did not wash **any two of the objects** at the same time.
   b. SCOPE=narrow: The nine flowers used in the corsages must include **at least one flower** from each of the four types.

(21) GRAM.FUN=O, LIN.ORD=last
   a. SCOPE=wide: Exactly three girls perform **each dance**.
   b. SCOPE=narrow: The official will also assign each runner to represent **a different charity**.

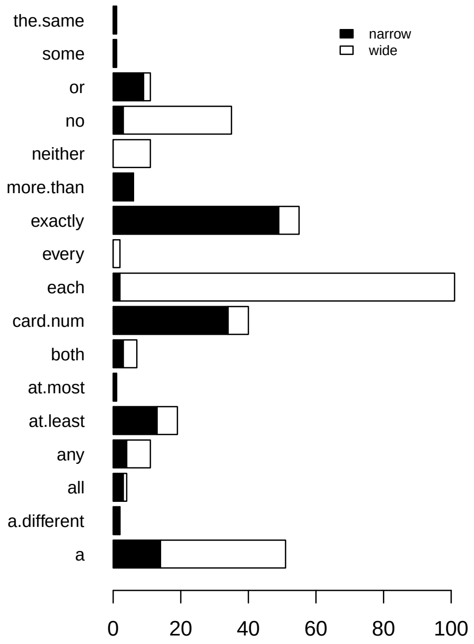
Random effects. We treat the lexical realization LEX.REAL of the quantifier we want to predict scope for as a random effect. LEX.REAL is a factor with 17 levels: **a**, **a.different**, **all** . . . ; see the figure below for the complete list.

We treat LEX.REAL as a random – as opposed to a fixed – effect for three reasons. First, some quantifiers, e.g., *every* or *at most*, occur very few times in our corpus, so their estimated contributions to scope preference would be highly biased / unstable if they were estimated as fixed effects. As Gelman & Hill (2007) among others argue, treating a variable as a random effect moderates this because the estimation of a particular quantifier / level of LEX.REAL 'borrows strength' from the observations made about other quantifiers / levels of LEX.REAL.

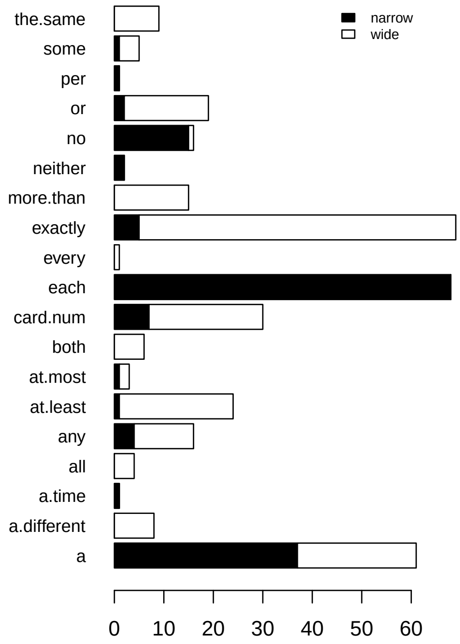
Second, it makes theoretical and empirical sense to model LEX.REAL as a random effect, i.e., to model scope preferences of various quantifiers as coming from the same underlying probability distribution: quantifiers are part of the same class of meanings and our statistical model should reflect that at some level. In addition, various quantifiers are clearly related to one another, e.g., the universals *every*, *each*, *all* and *both* are likely not independent of one another, the indefinites *a*, *some*, *a different* and cardinal numerals are not either, cardinals with numeral modifiers like *at least*, *at most* and *exactly* are not, etc.

Finally, and most importantly, treating LEX.REAL as a random effect, i.e., modeling the lexical realization of quantifiers jointly as part of a single abstract class, enables us to evaluate and compare the different scope-preference contributions made by lexical realization as a whole vs. linear order and grammatical function.

LEX.REAL by SCOPE



LEX.REAL.OTHER by SCOPE



In addition to the lexical realization of the quantifier we want to predict scope for, we will also consider the lexical realization of the other quantifier LEX.REAL.OTHER and the contributions this makes to scopal preference. LEX.REAL.OTHER is a factor with 19 levels: **a**, **a.different**, **a.time**, **all** . . . ; see the figure below for the complete list. We model LEX.REAL.OTHER as a random effect for the same three reasons provided above for LEX.REAL. Lexical effects on scope-taking could be analyzed in other ways, e.g., as pairs of lexical realizations for the two quantifiers,[7] but we leave a more in-depth investigation of this issue for a future occasion.

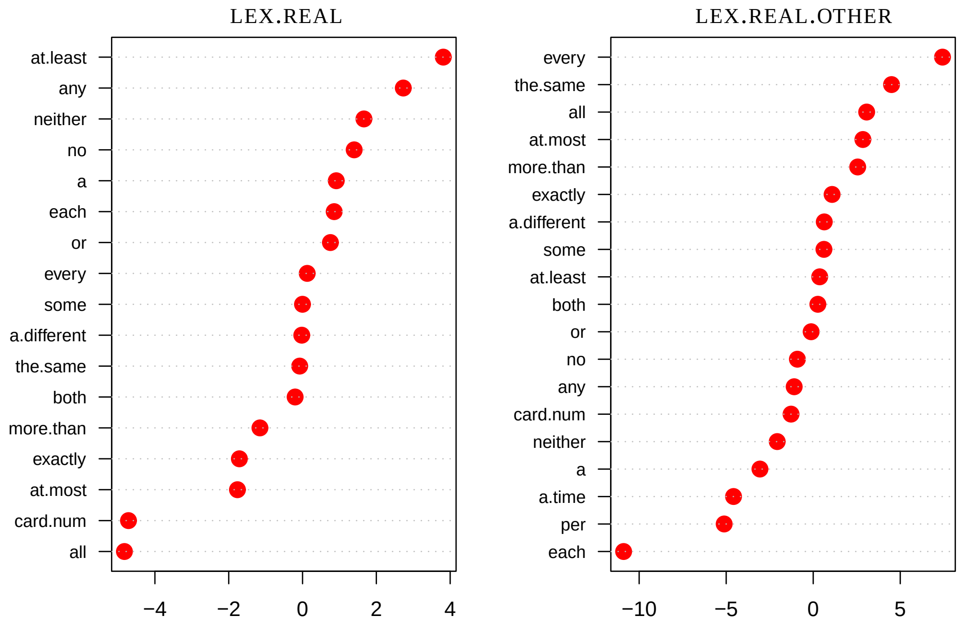
[7] We are indebted to Alexis Dimitriadis and Chris Potts for insightful suggestions and comments about this.

Statistical modeling and the resulting generalizations. We start with the full model for the fixed effects (the two main effects and their interaction) and intercept-only random effects for LEX.REAL and LEX.REAL.OTHER. The interaction of LIN.ORD and GRAM.FUN is not significant ($p = 0.70$), so we drop it. But adding GRAM.FUN to the model with LIN.ORD as the only fixed effect significantly reduces deviance ($p = 0.005$) and similarly, adding LIN.ORD to the model with GRAM.FUN as the only fixed effect significantly reduces deviance ($p = 1.04 \times 10^{-7}$). Adding random effects for the LIN.ORD and / or GRAM.FUN slopes is not significant – at least when the maximum likelihood estimates (MLEs) of the resulting models can be estimated, which is not always possible. But dropping the intercept random effects for LEX.REAL or LEX.REAL.OTHER significantly increases deviance ($p = 3.21 \times 10^{-11}$ and $p = 2.08 \times 10^{-13}$, respectively).

Thus, our final mixed-effects logistic regression model has (*i*) fixed effects for LIN.ORD and GRAM.FUN (no interaction), and (*ii*) intercept random effects for LEX.REAL and LEX.REAL.OTHER. The MLEs for this model are provided in the table below. We see that even when we control for lexical effects, both linear order and grammatical function are significant, in line with the results in the previous literature. If a quantifier comes last, this greatly increases its preference for / probability of narrow scope. Similarly, if a quantifier occurs in object position, this increases its probability of narrow scope but to a more moderate extent.

(22)

RANDOM EFFECTS		std.dev.		
	LEX.REAL	3.45		
	LEX.REAL.OTHER	5.55		
FIXED EFFECTS		estimate	std.error	p-value
	INTERCEPT	4.60	1.86	0.014
	LIN.ORD-**last**	-6.16	1.42	1×10^{-5}
	GRAM.FUN-**O**	-2.49	0.93	0.007


LEX.REAL

LEX.REAL.OTHER

The figure above plots the MLEs of the random effects for both LEX.REAL and LEX.REAL.OTHER. Note that these random effects are roughly on the same scale as the fixed effects, i.e., roughly between -8 and 8, and of similar magnitude.[8]

Model fit. One way to evaluate how well the model fits the data is to examine the C and Somers' Dxy statistics for several models. C is an index of concordance between predicted probability and observed response; Somers' Dxy is a rank correlation between predicted probabilities and observed responses related to C. The final mixed-effects logistic regression model has a C statistic of 0.996 and a Dxy statistic of 0.992, very close to 1 (the maximal value, indicating perfect model fit). In contrast, the C and Dxy statistics for the model with fixed effects only are 0.859 and 0.717, respectively. This indicates that lexical realization significantly improves scope predictions, which is confirmed by the C and Dxy values for the model with lexical random effects only: they are 0.982 and 0.965, very close to the values for our final mixed-effects model.

Another measure of model fit is Nagelkerke's R-squared, again with a maximum value of 1. This a common pseudo-R-squared measure for (mixed-effects) logit models assessing how much of the 'variance' in the response is accounted for by the predictors, i.e., Nagelkerke's R-squared assesses the quality of a model with regard to the model with only the intercept. Nagelkerke's R-squared for our final model relative to the 'intercept random-effects' only model is 0.404 , indicating that the LIN.ORD and GRAM.FUN fixed effects do account for some of the variation in scope-taking behavior observed in our corpus. Nagelkerke's R-squared for our final model relative to the ordinary intercept (null) model is a much higher 0.847, indicating that the LEX.REAL and LEX.REAL.OTHER random effects account for at least as much variation in scope taking behavior as the fixed effects. This is confirmed by the fact that Nagelkerke's R-squared for the model with intercept random-effects only relative to the ordinary intercept model is 0.743.

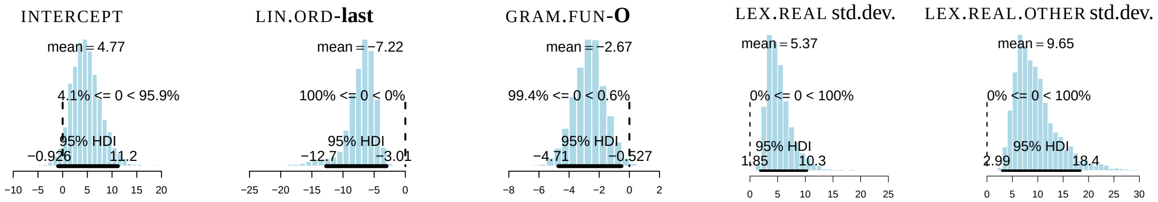
Bayesian estimation. We can estimate the parameters of our final logistic regression model more precisely based on MCMC samples from their posterior distributions.[9] We assume fairly low information priors for the intercept and fixed-effect slopes: independent normals $N(0, 10^2)$ (recall that we are dealing with logistic regression models here, so the coefficients are relatively small). Also, we assume independent normals $N(0, \sigma^2)$ and $N(0, \tau^2)$ for the two random effects; the priors for the standard deviations σ and τ are independent uniforms $\mathit{Unif}(0, 30)$. The means and standard deviations of the posterior distributions for the fixed and the random effects are fairly close to the MLEs, as the table and five figures below show (the R code for the figures is from Kruschke 2011).

(23)

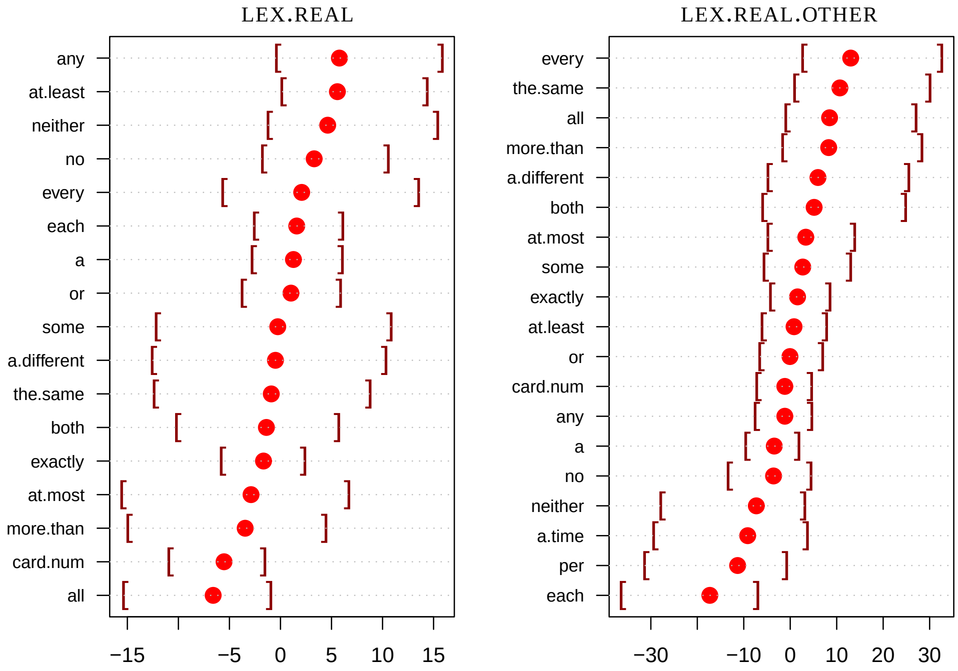
RANDOM EFFECTS		mean	std.dev.
	σ	5.37	2.48
	τ	9.65	4.32
FIXED EFFECTS		mean	std.dev.
	INTERCEPT	4.77	3.13
	LIN.ORD-**last**	-7.22	2.64
	GRAM.FUN-**O**	-2.67	1.07

[8] We are indebted to Robert Daland for emphasizing this point.

[9] 3 chains, 3.5×10^6 iterations per chain, 1×10^6 burnin, 2500 thinning.



The two figures below provide the mean and the central 95% CRIs for the random effects.



5 Conclusion

The three main findings of our investigation are as follows. First, we confirmed the results in the previous literature that linear order and grammatical function have an effect on scope-taking preferences. Second, we discovered that lexical effects on scoping preferences are at least as important as linear order or grammatical function. Third, the relational aspect of these lexical effects is also important: LEX.REAL.OTHER is at least as good a predictor of scope as LEX.REAL.

These findings provide a new kind of empirical support for *relational* theories of quantification that derive scopal behavior by focusing on the way in which one quantifier affects the context of interpretation for another quantifier, e.g., (in)dependence logic or dynamic plural logic. The notion of interpretation context formalized in these logics is inherently relational because it focuses on context change, i.e., on the way in which an expression sets up the context of interpretation for a subsequent expression. But syntactic scoping mechanisms that focus on *hierarchies* of (classes of) quantifiers, e.g., Beghelli & Stowell (1997), are also supported.[10]

The present investigation opens the way towards a broader research program of identifying scoping-behavior patterns that should ultimately enable us to group quantifiers into classes

[10] We are indebted to Lucas Champollion and Jakub Dotlacil for emphasizing this point.

depending on the type of scopal behavior they exhibit. Identifying such classes could provide an empirical basis for semantic theories that assign different kinds of semantic representations to these classes and / or for psycholinguistic theories that hypothesize different processing strategies for different classes.

Finally, this research also opens the way towards examining the typology and cross-linguistic variation of quantifier *systems* in addition to the ‘micro’ typology of individual (sub-classes of) quantifiers.

References

Anderson, C. (2004) The Structure and Real-time Comprehension of Quantifier Scope Ambiguity. PhD thesis, Department of Linguistics, Northwestern University.

Beghelli, F. & T. Stowell. (1997). Distributivity and Negation: The Syntax of Each and Every. In *Ways of Scope Taking*, A. Szabolcsi (ed.), Springer, 71-107.

Fodor, J.D. (1982). The Mental Representation of Quantifiers. In *Processes, Beliefs, and Questions*, S. Peters & E. Saarinen. Dordrecht: D. Reidel, 129-164.

Gelman, A. & J. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gillen, K. (1991) The Comprehension of Doubly Quantified Sentences. PhD Thesis, University of Durham.

Higgins, D. & Sadock, J. (2003). A Machine Learning Approach to Modeling Scope Preferences. *Computational Linguistics* 29(1): 73-96.

Ioup, G. (1975). Some Universals for Quantifier Scope. In J. Kimball, editor, *Syntax and Semantics*, volume 4, pages 37-58. Academic Press, New York.

Kruschke, J. (2011). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press, Elsevier.

Kurtzman, H. & MacDonald, M. (1993). Resolution of quantifier scope ambiguities. *Cognition*, 48:243-279.

Matthewson, L. (2004) On the Methodology of Semantic Fieldwork. *International Journal of American Linguistics*, 70(4): 369-415.

Micham, D., J. Catlin, N VanDerven & K. Loveland (1980). Lexical and Structural Cues in Quantifier Scope Relations. *Journal of Psycholinguistics Research*, 9: 367-377.

R Language (2011). *R: A Language and Environment for Statistical Computing*, R Development Core Team, R Foundation for Statistical Computing, `http://www.R-project.org/`.

Saba, W. & Corriveau, J-P (2001). Plausible reasoning and the resolution of quantifier scope ambiguities. Studia Logica, 67: 271-289.

Srinivasan, P. & Yates, A. (2009). Quantifier scope disambiguation using extracted pragmatic knowledge: Preliminary results. In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Tunstall, S. (1998) The interpretation of quantifiers: semantics and processing. PhD Thesis, University of Massachusetts Amherst.

Van Lehn, K. (1978). Determining the scope of English quantifiers. Technical Report AI-TR-483, AI Lab, MIT.