Conceptual alternatives*

Brian Buccola† · Manuel Križ‡ · Emmanuel Chemla‡

†Language, Logic and Cognition Center, The Hebrew University of Jerusalem
‡Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS)
Département d'Études Cognitives, École Normale Supérieure, PSL Research University

November 15, 2016

Abstract

Competition between sentences is at the core of numerous phenomena in natural language semantics. For instance, the competition between sentences such as *Some of the shapes are red* and *All the shapes are red* is taken to explain why an utterance of the former implicates that the latter is false. It is generally assumed that such competition is regulated by the relative syntactic simplicity of the competing utterances. Using theoretical and experimental tools devised to study human non-linguistic, conceptual abilities, we provide evidence for a new perspective: the rules that govern competition in natural language may be better and more deeply understood as rooted in relative *conceptual* simplicity, and thus are inherited from non-linguistic domains. We also provide evidence that the various construals of numerals preferentially attested in human language may likewise be rooted in non-linguistic, conceptual preferences.

Keywords: concepts; language of thought; formal semantics; linguistic competition

1 Formal semantics for concepts

The tools used in formal linguistics are in principle appropriate to study languages in a broad sense — that is, systems of signs with a syntax (some strings of signs are in the language, and some are not) and a semantics (each string in the language can be said to match or not match with a given state of the world). This definition encompasses not only human oral and sign languages, but also logical and artificial languages, from

We would like to thank Denis Bonnay, Alexandre Cremers, Paul Egré, Raj Singh, Benjamin Spector, Philippe Schlenker, and Paul Smolensky. The research leading to this work was supported by the Israel Science Foundation (ISF 1926/14), by the German-Israeli Foundation for Scientific Research (GIF 2353), by the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 313610, and by the Agence Nationale de la Recherche (ANR-10-IDEX-0001-02 PSL and ANR-10-LABX-0087 IEC).

which these tools originated (Tarski 1937; Montague 1970), and, as recently and forcefully defended by Schlenker et al. (2016), animal vocalizations.

The current study is concerned with the language of thought. Various scholars have been interested in exploring the building blocks of thought and reasoning systems — for adults, children, and even non-human animals — and despite the lack of a visible syntax, they have shown how a study of the semantics is possible (Carey 2009; Tomasello and Call 1997). In an ambitious, recent paper, Piantadosi et al. (2016) propose to model the language of thought as a formal language whose syntax is governed by a probabilistic grammar that assigns weights to different logical elements, such as conjunction, negation, etc. They presented several series of objects to participants, asking them to identify what rules a particular subset of these objects follow, such as 'the object is red or square'. The rule here serves as the 'sentence' in the language of thought. In principle, each rule can be expressed in many different ways, depending on what the logical primitives of the language of thought are taken to be, e.g. whether 'or' is a primitive but not 'if and only if', or vice versa, and so on. They therefore use this rule discovery task to determine which rules are hypothesized most readily by participants, thereby providing an estimation of which primitives are the most likely candidates of the language of thought.

We capitalize on these advances and propose to investigate as a case study a specific connection that can be drawn between linguistic studies in the conceptual domain (as described above) and in the domain of natural languages.

2 Linguistic or conceptual alternatives and the symmetry problem

A sentence like (1a) is often understood to imply that not all the relevant shapes are red, (1b). The standard account of this inference is that it arises from a competition between (1a) and the *alternative* sentence in (1c): since the alternative is more informative (because it is logically stronger), it is preferable for a speaker to use (1c) instead of (1a) in a situation where both are true; therefore, since (1c) was not used, it is presumed to be false.[1] The negation of (1c) yields exactly the observed inference in (1b) that we were trying to explain. However, such a theory needs to also explain why (1c) is the alternative to (1a) and not, for instance, (1d). If the latter were an alternative, then the predicted inference would be the negation of (1d), and the sentence would overall convey that all the shapes are red, which is the opposite of what we observe.

(1) a. Some of the shapes are red.
b. Not all the shapes are red. (observed inference)
c. All the shapes are red. (actual alternative)
d. Some but not all of the shapes are red. (possible alternative)

This issue is a version of what is known as the *symmetry problem* (Fox 2007): if one cannot determine which is the active alternative between (1c) and (1d), then the competition theory does not make useful predictions. Several solutions can be offered to

[1] This description is the standard Gricean, or neo-Gricean, description of the phenomenon (Grice 1975, 1989; Gamut 1991; Sauerland 2004). Recent attempts have proposed alternative perspectives on the topic, but all converge on the use of competition among sentences as a starting point (Chemla and Singh 2014).

explain why (1c) is an alternative and (1d) is not. One solution involves considering the complexity of the alternatives, in syntactic terms (Katzir 2007). In the example above, it is transparent that the active alternative (1c) is less complex than the inactive alternative (1d). This could explain why, if either of these had to be an alternative, it would be the former, therefore solving the symmetry problem.

Despite the apparent ease of the solution in this particular case, it is important to entertain the possibility that another source of complexity is at play. The previous considerations assume that alternatives are well-formed sentences themselves, but if this is the correct view, then one would have to show that for *every* possible way of expressing (1d), there is a simpler way of expressing (1c). The system proposed by Bergen et al. (2016) in fact has the potential to deal with multiple alternatives with the same meaning and could be taken as a better starting point. But one may also shortcut the issue entirely and consider the possibility that the competition between alternatives occurs at a non-linguistic level — which may well be approximated by what linguists call Logical Form, a potentially intermediate level of theorization/representation between language and our description of concepts (see, e.g., Chomsky 1981; May 1985; Hornstein 1995, and the references therein) — so that it would be the complexity of their meanings that matters. Exploring such a route is adventurous, especially given the appeal and simplicity of a view according to which such phenomena occur because of the competition between plain utterances, in the Gricean tradition (Grice 1975, 1989). There are, however, some substantive reasons to suspect that alternatives are not full linguistic citizens.

First, some cases have been exhibited where a phenomenon is best explained by the existence of a particular alternative, but where the phenomenon occurs even in languages in which the relevant alternative is not (straightforwardly) expressible. For instance, in many languages, the competition between *all* and *both* can explain why *all* typically triggers the inference (or comes with the requirement) that the domain it quantifies over has more than two elements (Percus 2006; Sauerland 2008). This explains the oddness of a sentence such as *He broke all of his arms*. But similar facts apply to French *tous* ('all'), even though there is no lexicalized form of *both* in this language (Chemla 2007). Similarly, in German, a sentence such as *Kein Student hat alle Probleme gelöst* ('No student solved all of the problems') triggers the inference that 'some students solved some of the problems'. This inference would be easily explained if the original sentence were in competition with one where *all* is replaced with an existential quantifier (as in the English *No student solved any of the problems*), but there is no appropriate quantifier in German, and the most natural way to express the needed alternative would be as follows: *Kein Student hat auch nur (irgend)ein Problem gelöst* ('No student solved *even one of* the problems'). This needed alternative is syntactically more complex than the original sentence and may therefore be ruled out of the set of candidate alternatives, at least according to the most current view about creating alternatives, described in Katzir 2007. Hence, in German and in French, some attested inferences are not explained if we constrain alternatives to compete at the linguistic level.[2]

[2] A similar argument, mentioned to us by Philippe Schlenker (p.c.), can be constructed for English: *There were some students from my class at the party* implies that not all the students from my class were at the party, and yet the alternative that one would need to negate to yield this inference, viz. **There were all students*

Second, we note that the very same construction also creates some difficulty in English, albeit a different difficulty. If *some* and *all* are alternatives to each other, then *No student solved all of the problems* should be in competition with *No student solved some of the problems*. But *some*, as a word, has special properties. In particular, it is a so-called positive polarity item, meaning that it cannot be interpreted in the scope of a negative operator like *none*. The putative alternative is thus interpreted as meaning that some of the problems were solved by no student. If negated, this alternative would predict the inference that each problem was solved by at least one student. This inference is unattested.

All of the problems above — the French and German undergeneration problems and the English overgeneration problem — arise only if one assumes that competition occurs at the linguistic level. If, instead, competition is between an existential and a universal 'primitive', for instance, then the German and English problems vanish.

Independently of the strength of the arguments above, exploring the possibility that competition occurs at a non-linguistic level has the potential to offer a deeper form of explanation. The reason why (1d) appears to be more complex than (1c) is that 'all' is lexicalized while 'some but not all' (henceforth, 'SBNA') is not lexicalized (and in fact, there appears to be no evidence of a lexicalized 'SBNA' across languages of the world; however, it is worth noting that the search has focused on other quantifiers, probably because 'SBNA' is an obvious case rather than the opposite; see, e.g., von Fintel and Matthewson 2008; Katzir and Singh 2013). But why is that so? Lexicalization may very well be determined in part by some notion of the intrinsic complexity of a meaning (see, e.g., the revival of such discussions from Horn 1972, 1989). Hence, syntactic complexity considerations may actually be a mere consequence of meaning complexity, and thus exploring the latter path could have more explanatory value.

3 Experiment

The goal of this experiment is to establish preferences between expressions, or rather between their counterparts at the level of thought (i.e. in a hypothesized language of thought). Using again the example above, we would like to know whether one can find an intrinsic preference for 'all' over 'some but not all', which would help solve the symmetry problem exposed above (as well as the lexicalization facts, potentially). To do so, we used an implicit rule discovery task, very much inspired by Piantadosi et al. 2016 (as well as predecessors in the 1960s, e.g. Haygood and Bourne 1965; King 1966; Bourne 1970). We adjusted the task, however, so as to be in a position to draw conclusions about pairwise differences between potential alternatives (while Piantadosi et al. had the wider-reaching ambition to evaluate as a whole the functional lexicon of the language of thought). The first two pairwise comparisons involve the 'some'/'all' scale of alternatives. Specifically, we will investigate whether 'all' is preferred over 'some but not all', thus explaining the paradigm in (2) (repeated below from above). We will also explore whether 'no' is preferred to 'some but not all', thus explaining the parallel paradigm in (3).

from my class at the party, is ungrammatical. Of course, a sentence like *All students from my class were at the party* is grammatical, and negating it would yield the observed inference, but this sentence cannot be straightforwardly derived as an alternative (to the sentence under discussion) on the theory of Katzir 2007.

(2) a. Some of the shapes are red.
b. Not all the shapes are red. (observed inference)
c. All the shapes are red. (actual alternative)
d. Some but not all of the shapes are red. (possible alternative)

(3) a. Not all the shapes are red.
b. Some shapes are red. (observed inference)
c. No shapes are red. (actual alternative)
d. Some but not all of the shapes are red. (possible alternative)

In addition, we will investigate a potential preference between 'at least *n*' and 'at most *n*' when it comes to the interpretation of numerical expressions. Although numerals typically have an 'exactly' reading, they often receive an 'at least' reading instead (the classic reference is Horn 1972; see also Spector 2013 for a recent survey of the semantic issues regarding numerals). For example, (4a) is normally understood to imply that Mary didn't solve more than three problems, (4b); or, put differently, in most contexts, (4a) is intuitively equivalent in meaning to (4c), with *exactly*. However, (5a), on its most natural reading, does not imply that Mary must not solve more than three problems, (5b); thus, (5a) and (5d), with *exactly*, intuitively express different meanings, and (5a) is instead intuitively equivalent in meaning to (5c), with *at least* instead of *exactly*.

(4) a. Mary solved three problems.
b. Mary didn't solve more than three problems. (observed inference)
c. Mary solved exactly three problems. (suitable paraphrase)
d. Mary solved at least three problems. (*not* a suitable paraphrase)

(5) a. Mary must solve three problems (in order to pass).
b. Mary must not solve more than three problems. (*not* an observed inference)
c. Mary must solve at least three problems. (suitable paraphrase)
d. Mary must solve exactly three problems. (*not* a suitable paraphrase)

Furthermore, in certain contexts, even the numeral in (4a) can be construed as having an 'at least' reading, as illustrated by the natural dialog in (6).

(6) Context: Everyone who solved three or more problems passed. John is interested in knowing whether Mary passed.
a. John: Did Mary solve three problems?
b. Bill: Yes, Mary solved three problems — in fact, she solved all five problems.

Crucially, however, numerals never seem to have genuine 'at most' readings.[3] For

[3] We say 'genuine' because there certainly are cases where a numeral *n* is interpreted as equivalent to 'at most *n*', as in the sentence *If you have three children, you do not qualify for tax exemptions*. However, as Breheny (2008) observes, this reading can be explained as the combination of an 'exactly' interpretation of *three* plus knowledge of how tax exemptions work: the literal meaning of the sentence asserts only that if you have exactly three children, you do not qualify; but together with the knowledge that if you do not qualify with *n* children, then you do not qualify with $n-1$ children, the sentence ends up being contextually equivalent to *If you have at most three children, you do not qualify for tax exemptions*. See also Spector 2013 for further arguments against the view that numerals can have 'at most' readings.

example, there is no context in which (4a) can ever be equivalent in meaning to *Mary solved at most three problems*. The decidedly unnatural dialog in (7) serves to illustrate this point: neither John's question nor Bill's answer can be interpreted in an 'at most' fashion (Bill's answer is simply self-contradictory).

(7) Context: Everyone who solved three or fewer problems failed. John is interested in knowing whether Mary failed.

a. John: Did Mary solve three problems?

b. Bill: Yes, Mary solved three problems — in fact, she solved only one problem.

There are several possible reasons why languages would work this way with regard to numerical interpretation. First, 'at least n' may be simpler for purely logical reasons. There is some plausibility to this thesis, since under certain conditions, it is possible to express an 'at least' reading with exact numerals, while it may require a larger set of elementary operations to express an 'at most' reading. To see this, assume that we have the possibility of quantifying over sets. We can then express the statement 'there is a set with exactly n members which all have property P'.[4] Importantly, this does not exclude the possibility of there being, in addition, other individuals that have property P. In fact, if more than n individuals have property P, then the statement 'there is a set with exactly n members which all have property P' is necessarily true: just pick any n individuals which have property P, and they constitute such a set. Thus, even though the set in question is said to have *exactly n* members, the overall effect with respect to the property P is an 'at least' meaning. An analogous construction of an 'at most' meaning is not possible with these primitives alone.[5] However, just as in the case of 'some' vs. 'some but not all' and of 'or' vs. 'if and only if', such claims necessarily rely on assumptions about primitive logical operations, and we prefer to stay away from such *a priori* considerations: we believe that we cannot know for sure from intuition alone what the relevant primitives are.

Second, then, is the possibility that 'at least' is favored at a more primitive, non-linguistic level, and that this cognitive preference somehow shaped languages. This second possibility is the one we investigate in the present experiment.

3.1 Participants

A group of 45 participants were recruited on Mechanical Turk and paid \$3 each for their participation. IPs were restricted to the United States, and all participants reported to be native speakers of English. We precommitted to excluding participants with a median response time below 500ms on the assumption that they would not have been doing the task, but we found that this applied to none.

[4] In logical notation: $\exists X[|X| = n \wedge \forall x[x \in X \rightarrow x \in P]]$.

[5] Of course, if we add negation to the inventory (and restrict ourselves to the natural numbers), then 'at most n' would be expressible as 'there is no set with exactly $n+1$ members which all have property P'. Or, in logical notation: $\neg\exists X[|X| = n+1 \wedge \forall x[x \in X \rightarrow x \in P]]$. This is quite generally the case because 'at least' and 'at most' are simply negations of one another.

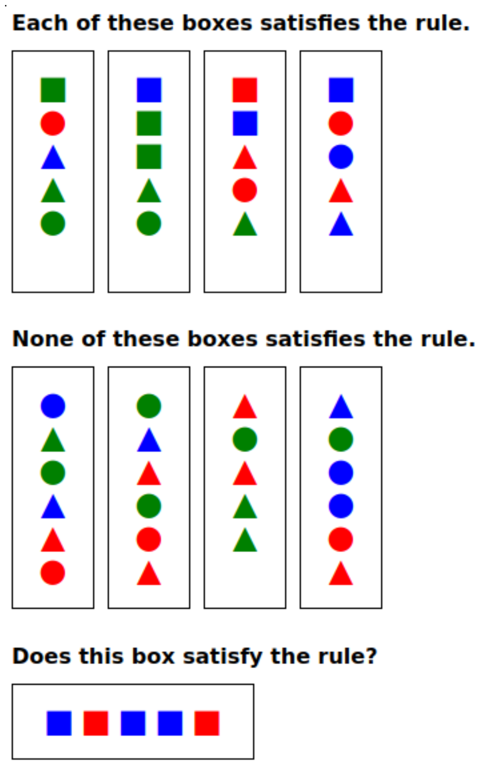


Figure 1: Example stimulus of the condition QUANTIFIERS=(some, SBNA), TYPE=TARGET, where the relevant property is 'square'. The positive boxes all satisfy the two rules 'some (or all) of the objects in the box are square' and 'some but not all of the objects in the box are square', while the negative boxes satisfy neither of those rules. The participants' task was infer a rule that the positive boxes satisfy and then to indicate whether the additional box at the bottom satisfies that rule or not by clicking either 'Satisfies the rule' or 'Doesn't satisfy the rule'.

3.2 Design and task

Figure 1 presents an example of a stimulus that was shown to participants. Their task was to infer a rule that certain boxes of objects satisfy. They were presented with a group of boxes that they were told satisfy the rule (henceforth, *positive* boxes), and another group of boxes that they were told do not satisfy the rule (henceforth, *negative* boxes). At the bottom of the screen was another box, and the participants' task was to indicate whether they thought that this final box satisfied the rule, by clicking either 'Satisfies the rule' or 'Doesn't satisfy the rule'.

The experimental design involved two fully crossed factors: QUANTIFIERS and TYPE. The levels of the QUANTIFIERS factor are based on pairs of quantifiers, one of which is logically weaker than (entailed by) the other. In association with a particular property (random factor), a given quantifier determines a rule. For instance, in association with the property 'red' (see below for specific details about the stimuli), the quantifier 'all' determines a rule that can be stated as 'All the objects in the box are red'. Two quantifiers (in association with a given property) determine two rules which can be used to characterize the positive and negative boxes of items: positive boxes satisfied both rules, while negative boxes satisfied neither rule. The 8 levels of the QUANTIFIERS factor were: (some, all), (some, SBNA), (not all, no), (not all, SBNA), (at least 3, exactly 3), (at most 3, exactly 3), (at least 4, exactly 4), (at most 4, exactly 4). For the sake of the analysis, the last four were collapsed into two of the form (at least *n*, exactly *n*) and (at most *n*, exactly *n*).

The TYPE factor, together with the two rules determined by the QUANTIFIERS factor, determined the final test box. It varied according to whether the final test box satisfied both of the rules determined by the QUANTIFIERS level (in the YES type), none of these rules (in the NO type), or exactly one of them (in the TARGET type). As we will explain in further detail in the analysis, the YES and NO types serve as baselines to decide whether participants inferred a rule coherent with our goal, while the TARGET type helped us evaluate which of these two rules they favor: a positive response would correspond to the weaker rule (e.g. 'some'), a negative response to the stronger rule (e.g. 'SBNA').

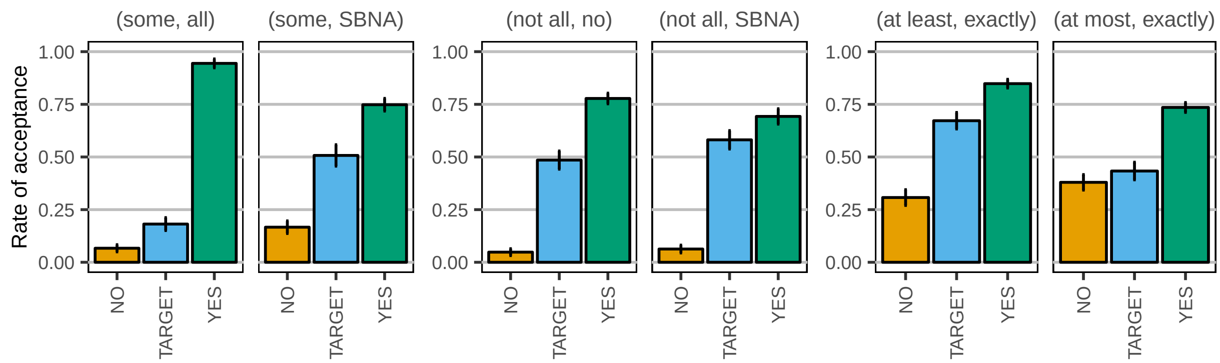
The quantifiers in a pair thus always stand in an entailment relation, which allows for all the kinds of boxes needed to create a given stimulus. But we are also able to compare two quantifiers which do not stand in a logical entailment relation by comparing each of them to a third quantifier that they both entail and obtaining a measure of how strongly each of them is preferred (or dispreferred) with respect to that common reference point. For example, if we learn that 'all' is strongly preferred to 'some', while 'SBNA' is only weakly preferred to 'some', then this indirectly indicates that 'all' is preferred to 'SBNA' (precisely the preference that would solve the symmetry problem).

3.3 Stimuli

Each item of a given condition was constructed by first randomly selecting the following superficial characteristics: the property required to determine the rules ('red', 'blue', 'green', 'triangle', 'square', or 'circle'), the number of positive boxes (3 or 4), the number of negative boxes (3 or 4), and for each box, the number of objects in that box (5 or 6). The shape and color of the objects in each box were then chosen so that each box would satisfy both, neither, or exactly one of the rules, depending on whether that box was a positive, negative, or test box. To simplify the task of the participants, and in particular the extraction of the relevant property, objects that satisfied the selected property were put together at the beginning (top or left) of their box.

To illustrate, the example in Figure 1 has the following superficial characteristics: the relevant property is 'square', the number of positive boxes is 4, the number of negative boxes is also 4, and each box has 5 or 6 objects. The condition is QUANTIFIERS=(some, SBNA), TYPE=TARGET. Therefore, each box in the positive group satisfies the two rules 'some (or all) objects in the box are square' and 'some but not all objects in the box are square', each box in the negative group satisfies neither of those rules, and the test box satisfies only the stronger rule ('some but not all'). Moreover, for each box, the color of each shape was chosen at random and the number of squares was also chosen at random, albeit within the limit imposed by the condition to be exemplified — that is, whether the box needed to satisfy both, neither, or exactly one of the rules.

We used a Python script (https://osf.io/6qarq/) and a Bash script (https://osf.io/7dskm/) to randomly construct the set of items that all participants saw, including 6 repetitions of each of the 24 combinations of the levels of TYPE (3 levels) and QUANTIFIERS (8 levels). The full set of 144 items was randomized for each participant. This randomized set



(a) SOME paradigm.
Comparison: 'all' vs. 'SBNA'

(b) NOT.ALL paradigm.
Comparison: 'no' vs. 'SBNA'

(c) NUM paradigm.
Comparison: 'at least' vs. 'at most'

Figure 2: Proportion of 'yes' responses for each pair of QUANTIFIERS and each TYPE of item. NO cases made both rules attached to the quantifiers false; YES cases made them both true; and TARGET cases made one of them true and the other false. Error bars represent the standard error of the within-subject mean. Results are grouped by 'paradigm', i.e. pairs of QUANTIFIERS which meaningfully share one member ('some', 'not all', or 'exactly *n*') and are used for the comparison between their other members.

was preceded by 4 practice items (the same ones, in the same order, for each participant). Their nature as practice items was not apparent to the participants, and they all involved a non-numerical QUANTIFIERS factor and a non-target TYPE factor. Our motivation for these choices was to ensure that participants did not hit a target item at the very beginning, and to ensure that participants did not see any discontinuity between the practice trials and subsequent trials. Once the list of items was generated offline, we used the online experimental platform created by Alex Drummond (`http://spellout.net/ibexfarm/`) to implement the task, presenting the (non-practice) items in a fresh random order to each participant.

3.4 Results

All data and an analysis script are available at `https://osf.io/uq3qp/`.

The mean results are presented in Figure 2. First, note that the YES type (green bars) and NO type (orange bars) generate appropriate high and low acceptance rates, respectively (grand means of 79% and 17%), while the TARGET type (blue bars) generates intermediate response rates (average 50%). This shows that participants were paying attention and were able to satisfactorily perform what might seem like a difficult task (of a similar nature to the pattern recognition tasks in IQ tests, in that participants have to identify both the dimensions/properties to which the rule pertains and the rule itself).

Furthermore, observe the secondary but interesting difference between the QUANTIFIERS (some, all) and (not all, no). These two pairs are duals of each other (the first member of one is the negation of the second member of the other). Concretely, this means that the constraints on the positive and negative boxes were simply reversed from one to the other: what served as a potential YES test box for one was a NO test box for the other.

Interestingly, we observe an asymmetry in response rates: the YES bar for (not all, no) is not as high as the NO bar for (some, all) is low. This suggests that participants did not perform the task via judgments of similarity (which of the two groups does the test box resemble the most?) — in which case the asymmetry would be unexpected — but instead engaged in a rule discovery task in which the positive exemplars played a different role from the negative exemplars. Overall, these first considerations suggest that participants solved the task as we were hoping.

After these preliminary descriptive results, let us move to the analysis of main interest. As a first pass, responses to the TARGET type should indicate whether participants inferred the rule based on the weak quantifier ('yes' responses) or on the strong quantifier ('no' responses). This would be most correct if participants inferred either one of these two rules to the exclusion of any other rule and in the absence of any bias. But, of course, this is not the case, as demonstrated by the fact that responses to YES and NO items are not at ceiling and floor. Responses to these YES and NO items thus provide a measure of the role of independent rules and biases in the ways participants solve the task. We thus ought to normalize the TARGET responses, relative to these natural endpoints, to obtain a cleaner measure of preference between the weak and strong quantifiers involved in a pair. To put it differently, we implement the idea that an absolute preference for the weak quantifier over the strong quantifier should generate a response as high as the response for the YES types, no more, and an absolute preference for the strong quantifier should generate a response as low as the response to the NO types, no less. This eliminates noise created by the fact that general yes/no biases and rules weaker/stronger than both of our two target rules may be entering our results.[6]

Formally, our measure is based on three parameters for each quantifier pair p. First, $\overline{\alpha_p}$ represents our best estimate of the true rate of 'yes' responses in the NO cases, and $\overline{\beta_p}$ represents our best estimate of the true rate of 'yes' responses in the YES cases. In an ideal world, we should find $\overline{\alpha_p} = 0$ and $\overline{\beta_p} = 1$.[7] Critically, $\overline{\gamma_p}$ represents the preference between the weak and the strong quantifier, estimated as the position of the responses to the TARGET types not as an absolute rate, but as a proportion with $\overline{\alpha_p}$ and $\overline{\beta_p}$ as extreme points. In pseudo-formula, this means that the actual response in the TARGET type is

[6] One source of simplification here is that we ignore the possibility that participants might infer a rule that is 'intermediate' between the two members of the pair; for example, 'at least 3' and 'more than half' are intermediate between 'some' and 'all'. Depending on the specific rule participants infer and the particular test box showing, this could alter the rate of 'yes' responses, but not because participants opt for quantifiers *from our pair*. We thus work on the assumption that the two rules we compare are the most salient in their range. However, we also note that our main interpretations are based not only on pairs, but on the comparison between two pairs of quantifiers. We submit that the intermediate rules relevant for these two pairs are similar across the pairs we use, and therefore that their role should be cancelled out in our analysis. For instance, when we compare 'all' and 'SBNA' with the pairs (some, all) and (some, SBNA), we assume that the relevant intermediate rules, which are compatible with the positive and negative examples being what they are, and which could make the test box true, are of the form 'at least x' in both cases.

[7] In the upcoming implementation of this idea using logit models, the parameters will not directly represent rates of 'yes' responses, and they will not range between 0 and 1 but rather between $-\infty$ and $+\infty$. To convey the spirit of the analysis more simply, however, we ignore this technicality in this paragraph. To keep track of it, however, we mark the parameters in this informal description with a line above them, as in $\overline{\alpha_p}$.

predicted by $\overline{\alpha_p} + \overline{\gamma_p} \cdot (\overline{\beta_p} - \overline{\alpha_p})$. Concretely, a maximal preference for the weaker quantifier in p would correspond to $\overline{\gamma_p} = 100\%$, i.e. an expected response rate in the TARGET case as high as in the YES cases: $\overline{\alpha_p} + 100\% \cdot (\overline{\beta_p} - \overline{\alpha_p}) = \overline{\beta_p}$. Conversely, an extreme preference for the stronger quantifier would correspond to $\overline{\gamma_p} = 0\%$, i.e. an expected response rate in the TARGET case as low as in the NO cases: $\overline{\alpha_p} + 0\% \cdot (\overline{\beta_p} - \overline{\alpha_p}) = \overline{\alpha_p}$. Finally, a neutral preference corresponds to $\overline{\gamma_p} = 50\%$ and an expected response rate exactly in between $\overline{\alpha_p}$ and $\overline{\beta_p}$: $\overline{\alpha_p} + 50\% \cdot (\overline{\beta_p} - \overline{\alpha_p}) = (\overline{\alpha_p} + \overline{\beta_p})/2$.

To be precise, we employed logit models that differ slightly in structure from the generalized linear models commonly used in the analysis of binary response data. Their basic form was as follows:

$$Y_{pi} \sim \mathsf{bernoulli}(\mathsf{logit}^{-1}(\pi_{pi}))$$
$$\text{with: } \pi_{pi} = \alpha_p \cdot \textsc{no}_i + \beta_p \cdot \textsc{yes}_i + (\alpha_p + \gamma_p \cdot (\beta_p - \alpha_p)) \cdot \textsc{target}_i.$$

Given a pair of quantifiers $p = (x, y)$, the parameter $\gamma_{(x,y)}$ thus represents the preference for the weaker member over the stronger member. Given two pairs of quantifiers (x, y_1) and (x, y_2) that share a common member x, we can then compare y_1 and y_2 by fitting a model to the data from both pairs and seeing whether there is evidence for a difference between $\gamma_{(x,y_1)}$ and $\gamma_{(x,y_2)}$. If there is one, then one of the y's is preferred to the other, in that it is more distinct from x than the other.

We partitioned our data into three 'paradigms' based on the pairs that shared a member. These were SOME, comprising (some, all) and (some, SBNA); NOT.ALL, comprising (not all, no) and (not all, SBNA); and NUM, comprising (at least n, exactly n) and (at most n, exactly n). On each of these three data sets, we fitted models that included either one or two γ parameters, as justified above, and different varying-effects structures: there was (or was not) a subject intercept u_{0s}, and either no modification of the parameters, parameter modifiers differing by subject, or parameter modifiers differing by both subject and pair. All models were fitted with MCMC methods using STAN through the `rstan` package in R, with STAN's default uniform (improper) prior over α, β, and γ (in the latter case restricted to $[0, 1]$) as well as the standard deviation hyperparameters for the varying subject effects (which themselves were assumed to be normally distributed around 0). Models were evaluated by leave-one-out cross-validation, approximated by Pareto-smoothed importance sampling with the `loo` package (Vehtari et al. 2016) on the basis of 5,000 samples of the likelihood for each data point, drawn after 5,000 burn-in iterations.

For all three data sets, the best model turned out to be the two-γ model with a subject intercept and parameter modifiers varying by subject, but not by pair:

$$Y_{spi} \sim \mathsf{bernoulli}(\mathsf{logit}^{-1}(\pi_{spi}))$$
$$\begin{aligned}\text{with: } \pi_{spi} = {} & (\alpha_p + u_{\alpha s}) \cdot \textsc{no}_i + (\beta_q + u_{\beta s}) \cdot \textsc{yes}_i \\ & + (\alpha_p + u_{\alpha s} + (\gamma_p + u_{\gamma s})((\beta_p + u_{\beta s}) - (\alpha_p + u_{\alpha s}))) \cdot \textsc{target}_i + u_{0s}.\end{aligned}$$

In order to see whether the preference for two γ-parameters was meaningful, we compared the optimal models to the variants with identical varying effects structure,

	SOME	NOT.ALL	NUM
Δ_{elpd}	–32.0	–3.8	–11.7
$se(\Delta_{elpd})$	7.9	3.1	7.2

Table 1: Differences in estimated log pointwise predictive likelihood and their standard errors for models with one or two γ parameters for sets of quantifier pairs sharing a weaker member. Negative values of Δ_{elpd} indicate evidence in favor of the two-γ model.

but only one (pair-independent) γ-parameter. Differences in estimated log pointwise predictive likelihood and their standard errors for models with one and two γ-parameters are given in Table 1.

This analysis provides evidence for a difference in all paradigms, strongly in the SOME and NUM paradigms and more modestly in the NOT.ALL paradigm.[8] Given the direction of the differences, this corresponds to evidence supporting a preference for 'all' over 'SBNA', for 'no' over 'SBNA', and for 'at least' over 'at most'.

3.5 Discussion

The evidence we found for the preference of 'all' over 'SBNA' and of 'no' over 'SBNA' provides a new solution to the symmetry problem and casts new light on the scalar inferences from 'some' to 'not all', (1), and the dual inference from 'not all' to 'some', (3). We also find that 'at least' is preferred over 'at most'. This second result may be seen either as a natural, reassuring outcome or as a useful fact in the same way as the previous results (see (4) and discussion around it).

The current study is only the first of its kind, and we would like to mention two of its limitations. First, we presented the symmetry problem created by the conflict between an 'all' and a 'SBNA' alternative to 'some'. But the same conflict arises if one considers the potential alternatives 'all' and 'not all'. Note, however, that the 'not all' alternative is not stronger than the utterance itself, and this could make it less preferred *a priori* as an alternative independently of the relative complexity of these alternatives (but see Chemla and Singh 2014 for a review of arguments that alternatives need not be stronger).

Second, one may wonder whether the current paradigm really taps into non-linguistic abilities, or whether its results merely reproduce the complexity facts at the linguistic (lexico-syntactic) level. It would be useful to replicate the current preferences in an even less linguistic setting, either by limiting participants' access to their linguistic abilities with a dual task paradigm (see Norman and Bobrow 1975), or by testing non-linguistic participants who could show the same preference (pre-linguistic infants or non-human animals). Setting up such studies is of course of independent, obvious interest, to try to compare the logical repertoires of different animals, at various linguistic stages, and has been recognized as such for a long time, to say the least.

[8] See Burnham and Anderson 2002, §2.6 for heuristics about differences on the deviance scale. Our Δ_{elpd} values are converted to that scale by multiplying them by –2.

4 Conclusion

Using tools set up to study human non-linguistic, conceptual abilities, this study addresses from a new perspective a certain puzzle concerning the linguistic phenomenon of scalar inference. The results suggest that the rules that govern competitions between linguistic utterances may be better and more deeply understood as rooted in preferences inherited from non-linguistic domains.

References

Bergen, Leon, Roger Levy, and Noah D. Goodman (2016). Pragmatic Reasoning through Semantic Inference. In: *Semantics and Pragmatics* 9. DOI: 10.3765/sp.9.20.

Bourne, Lyle E. (1970). Knowing and Using Concepts. In: *Psychological Review* 77.6, pp. 546–556. DOI: 10.1037/h0030000.

Breheny, Richard (2008). A New Look at the Semantics and Pragmatics of Numerically Quantified Noun Phrases. In: *Journal of Semantics* 25.2, pp. 93–139. DOI: 10.1093/jos/ffm016.

Burnham, Kenneth P. and David R. Anderson (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. New York, NY: Springer-Verlag.

Carey, Susan (2009). *The Origin of Concepts*. New York, NY: Oxford University Press.

Chemla, Emmanuel (2007). French *both*: A Gap in the Theory of Antipresupposition. In: *Snippets* 15. URL: http://www.ledonline.it/snippets/allegati/snippets15001.pdf.

Chemla, Emmanuel and Raj Singh (2014). Remarks on the Experimental Turn in the Study of Scalar Implicature, Part I. In: *Language and Linguistics Compass* 8.9, pp. 373–386. DOI: 10.1111/lnc3.12081.

Chomsky, Noam (1981). *Lectures on Government and Binding: The Pisa Lectures*. Dordrecht, Holland: Foris Publications.

von Fintel, Kai and Lisa Matthewson (2008). Universals in Semantics. In: *The Linguistic Review* 25.1–2, pp. 139–201. DOI: 10.1515/TLIR.2008.004.

Fox, Danny (2007). Free Choice and the Theory of Scalar Implicatures. In: *Presupposition and Implicature in Compositional Semantics*. Ed. by Uli Sauerland and Penka Stateva. Palgrave Studies in Pragmatics, Language and Cognition Series. New York, NY: Palgrave Macmillan. Chap. 4, pp. 71–120. DOI: 10.1057/9780230210752_4.

Gamut, L. T. F. (1991). *Logic, Language, and Meaning*. Vol. 1 (Introduction to Logic). Chicago, IL: University of Chicago Press.

Grice, H. P. (1975). Logic and Conversation. In: *Syntax and Semantics, Volume 3: Speech Acts*. Ed. by Peter Cole and Jerry L. Morgan. New York, NY: Academic Press, pp. 41–58.

Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Haygood, Robert C. and Lyle E. Bourne (1965). Attribute- and Rule-Learning Aspects of Conceptual Behavior. In: *Psychological Review* 72.3, pp. 175–195. DOI: 10.1037/h0021802.

Horn, Laurence R. (1972). On the Semantics of Logical Operators in English. PhD thesis. New Haven, CT: Yale University.

Horn, Laurence R. (1989). *A Natural History of Negation*. Chicago, IL: University of Chicago Press.

Hornstein, Norbert (1995). *Logical Form: From GB to Minimalism*. Cambridge, MA: Blackwell Publishers.

Katzir, Roni (2007). Structurally-Defined Alternatives. In: *Linguistics and Philosophy* 30.6, pp. 669–690. DOI: 10.1007/s10988-008-9029-y.

Katzir, Roni and Raj Singh (2013). Constraints on the Lexicalization of Logical Operators. In: *Linguistics and Philosophy* 36.1, pp. 1–29. DOI: 10.1007/s10988-013-9130-8.

King, William L. (1966). Learning and Utilization of Conjunctive and Disjunctive Classification Rules: A Developmental Study. In: *Journal of Experimental Child Psychology* 4.3, pp. 217–231. DOI: 10.1016/0022-0965(66)90022-1.

May, Robert (1985). *Logical Form: Its Structure and Derivation*. Cambridge, MA: MIT Press.

Montague, Richard (1970). English as a Formal Language. In: *Linguaggi nella Società e nella Tecnica*. Ed. by Bruno Visentini. Milan, Italy: Edizioni di Comunità. Reprinted in Montague 1974, pp. 188–221.

Montague, Richard (1974). *Formal Philosophy: Selected Papers of Richard Montague*. Ed. by Richard Thomason. New Haven, CT: Yale University Press.

Norman, Donald A. and Daniel G. Bobrow (1975). On Data-Limited and Resource-Limited Processes. In: *Cognitive Psychology* 7.1, pp. 44–64. DOI: 10.1016/0010-0285(75)90004-3.

Percus, Orin (2006). Antipresuppositions. In: *Theoretical and Empirical Studies of Reference and Anaphora: Toward the Establishment of Generative Grammar as an Empirical Science*. Ed. by Ayumi Ueyama. Report of the Grant-in-Aid for Scientific Research (B), Project No. 15320052, Japan Society for the Promotion of Science, pp. 52–73.

Piantadosi, Steven T., Joshua B. Tenenbaum, and Noah D. Goodman (2016). The Logical Primitives of Thought: Empirical Foundations for Compositional Cognitive Models. In: *Psychological Review* 123.4, pp. 392–424. DOI: 10.1037/a0039980.

Sauerland, Uli (2004). Scalar Implicatures in Complex Sentences. In: *Linguistics and Philosophy* 27.3, pp. 367–391. DOI: 10.1023/B:LING.0000023378.71748.db.

Sauerland, Uli (2008). Implicated Presuppositions. In: *The Discourse Potential of Underspecified Structures*. Ed. by Anita Steube. Berlin, Germany: Walter de Gruyter, pp. 581–600.

Schlenker, Philippe, Emmanuel Chemla, and Klaus Zuberbühler (2016). What Do Monkey Calls Mean? In: *Trends in Cognitive Sciences*. DOI: 10.1016/j.tics.2016.10.004.

Spector, Benjamin (2013). Bare Numerals and Scalar Implicatures. In: *Language and Linguistics Compass* 7.5, pp. 273–294. DOI: 10.1111/lnc3.12018.

Tarski, Alfred (1937). Über den Begriff der logischen Folgerung. In: *Actes du Congrès International de Philosophie Scientifique, Sorbonne, Paris 1935: Logique*. Vol. 7. Paris, France: Hermann and Company, pp. 1–11. Translated and reprinted as "On the Concept of Logical Consequence" in Tarski 1983, pp. 409–420.

Tarski, Alfred (1983). *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*. Ed. by John Corcoran. Trans. by J. H. Woodger. 2nd ed. Hackett Publishing Company.

Tomasello, Michael and Josep Call (1997). *Primate Cognition*. New York, NY: Oxford University Press.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry (2016). Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC. In: *Statistics and Computing*, pp. 1–20. DOI: 10.1007/s11222-016-9696-4.