Interpreting tractable versus intractable reciprocal sentences

Bott, O.; Schlotterbeck, F.; Szymanik, J.

Publication date
2011
Document Version
Final published version
Published in
Proceedings of the Ninth International Conference on Computational Semantics

Interpreting tractable versus intractable reciprocal sentences

Oliver Bott[a], Fabian Schlotterbeck[a] & Jakub Szymanik[b]
SFB 833, University of Tübingen[a], University of Stockholm[b]
oliver.bott@uni-tuebingen.de
fabian.schlotterbeck@uni-tuebingen.de
jakub.szymanik@gmail.com

Abstract

In three experiments, we investigated the computational complexity of German reciprocal sentences with different quantificational antecedents. Building upon the *tractable cognition thesis* (van Rooij, 2008) and its application to the verification of quantifiers (Szymanik, 2010) we predicted complexity differences among these sentences. Reciprocals with *all*-antecedents are expected to preferably receive a strong interpretation (Dalrymple et al., 1998), but reciprocals with proportional or numerical quantifier antecedents should be interpreted weakly. Experiment 1, where participants completed pictures according to their preferred interpretation, provides evidence for these predictions. Experiment 2 was a picture verification task. The results show that the strong interpretation was in fact possible for tractable *all but one*-reciprocals, but not for *exactly n*. The last experiment manipulated monotonicity of the quantifier antecedents.

Formal semantics hasn't paid much attention to issues of computational complexity when the meaning of an expression is derived. However, when it comes to semantic processing in humans (and computers) with limited processing resources, computational tractability becomes one of the most important constraints a cognitively realistic semantics must face. Two consequences come to mind immediately. If there is a choice between algorithms, we should choose tractable ones over intractable ones. And secondly, meanings which cannot be effectively computed shouldn't be posited for natural language expressions. In this paper we present three psycholinguistic experiments investigating the latter aspect.

Following traditions in computer science, a number of cognitive scientists have defined computational tractability as polynomial-time-computability (for an overview see van Rooij, 2008) leading to the *P-Cognition Hypothesis* (PCH): cognitive capacities are limited to those functions that can be computed in polynomial time. These functions are input-output functions in the sense of Marr (1982)'s first level. One objection against the PCH is that computational complexity is defined in terms of limit behavior as the input increases. In practice, however, the input may be rather small. van Rooij (2008) points out that the input size can be parametrized turning a problem that is intractable for a large input size into a tractable one for small inputs. We manipulated the input size in an experiment to test this more refined version of the PCH.

An interesting test case for the PCH are quantified sentences containing reciprocal expressions of the form *Q of the As R each other*. Consider (1-a) – (1-c).

(1) a. Most of the dots are connected to each other.
b. Four of the dots are connected to each other.
c. All dots are connected to each other.

It has been commonly observed that such sentences are highly ambiguous (see eg. Dalrymple et al., 1998). For instance, under its logically strongest interpretation (1-a) is true iff given n dots there is a subset of more than $\frac{n}{2}$ dots which are pairwise connected. But there are weaker readings of reciprocity, too, i.e. connectedness by a path (a continuous path runs through Q of the dots) or – even weaker – Q of the dots are interconnected, but no path has to connect them all. Following Dalrymple et al. (1998) we call these reciprocal meanings *strong*, *intermediate*, and *weak*, respectively. As for verification, Szymanik (2010) has shown that the various meanings assigned to reciprocals with quantified antecedents differ drastically in their computational complexity. In particular, the strong meanings of reciprocal sentences with proportional and counting[1] quantifiers in their antecedents are intractable, i.e. the verification problem for those readings is NP-complete. This is due to the combinatorial explosion in identifying the relevant completely-connected subsets for these two types of quantifiers (cf. CLIQUE problem, see Garey and Johnson (1979, problem GT19)) which does not emerge with *all*. However, intermediate and weak interpretations are PTIME computable. For example, going through all the elements in the model, thereby listing all the paths, and then evaluating the paths against the quantifier in the antecedent solves the problem in ploynomial time. The PCH thus allows us to derive the following predictions. A strong interpretation should be impossible for sentences (1-a) and (1-b), but possible for the tractable sentence (1-c). Therefore, Szymanik (2010) suggests that if the processor initially tries to establish a strong interpretation, there should be a change in the meanings of sentences (1-a) and (1-b) to one of the weaker interpretations.

In an attempt to explain variations in the literal meaning of the reciprocal expressions Dalrymple et al. (1998) proposed the *Strong Meaning Hypothesis* (SMH). According to the SMH, the reading associated with the reciprocal is the strongest available reading which is consistent with the properties of the reciprocal relation and the relevant information supplied by the context. Consider (2-a) to (2-c).

(2) a. All members of parliament refer to each other indirectly.
b. All Boston pitchers sat alongside each other.
c. All pirates were staring at each other in surprise.

The interpretation of reciprocity differs among those sentences. Sentence (2-a) implies that each parliament member refers to each of the other parliament members indirectly. In other words, the strong interpretation seems to be the most natural reading. This is different in (2-b) and (2-c) which receive intermediate and weak interpretations, respectively. Here the predicates *sit alongside* and *stare at* arguably constrain the meaning. Observations like these lend intuitive support to the SMH. Kerem et al. (2010) modified the SMH and provided experimental evidence that comprehenders are biased towards

[1] It is natural to assume that people have one quantifier concept *Exactly* k, for every natural number k, rather than the infinite set of concepts *Exactly* 1, *Exactly* 2, and so on. Mathematically, we can account for this idea by introducing the counting quantifier $\mathsf{C}^{=A}$ saying that the number of elements satisfying some property is equal to the cardinality of the set A. The idea here is that determiners like *Exactly* k express a relation between the number of elements satisfying a certain property and the cardinality of some prototypical set A (see Szymanik (2010) for more discussion).

the most typical interpretation of the reciprocal relation. Thus, the reciprocal relation seems to constrain the meaning. Neither the original SMH nor Kerem et al. (2010)'s account leads us to expect that the three quantifiers in (1-a) – (1-c) should differ with respect to how they constrain reciprocal meanings. With 'neutral' predicates like *to be connected by lines* the SMH predicts an overall preference for the strong interpretation in all three sentences. A property that should matter, though, is the monotonicity of the quantificational antecedent. Since monotone decreasing quantifiers have the exact opposite entailment pattern as increasing ones, the SMH leads us to expect that preferences should be reversed in monotone decreasing quantificational antecedents.

We tested the PCH and the SMH in three experiments. In the first we surveyed the default interpretation of reciprocal sentences with quantificational antecedents like (1-a) – (1-c) by having participants complete dot pictures. The second experiment tested the availability of strong and intermediate interpretations in a picture verification task using clearly disambiguated pictures where, in addition, the input size was manipulated. The last experiment compared upward increasing and decreasing quantifiers.

Experiment 1: what is the preferred interpretation?

According to the SMH, sentences like (3-a) are preferably interpreted with their strong meaning in (3-b).

(3) a. All/Most/Four of the dots are connected to each other.

b. $\exists X \subseteq DOTS[Q(DOTS, X) \wedge \forall x, y \in X(x \neq y \rightarrow connect(x, y))]$,
where Q is ALL, $MOST$ or $FOUR$.

The PCH, on the other hand, predicts differences between the three quantifiers. While the strong meaning of *reciprocal all* can be checked in polynomial time, verifying the strong interpretation of *reciprocal most* and *reciprocal four* is NP-hard[2]. By contrast, the weaker readings are computable in polynomial time for all three types of quantifiers. It is thus expected that the choice of Q should affect the preference for strong vs. intermediate/weak interpretations. Bringing the SMH and the PCH together we get the following predictions: *reciprocal all* should receive a strong reading, but *reciprocal most/four* should receive an intermediate or weak one.

Method

These predictions were tested in a paper-and-pencil questionnaire. 23 German native speakers (mean age 24.3 years; 10 female) received a series of sentences, each paired with a picture of unconnected dots. Their task was to connect the dots in such a way that the resulting picture matched their interpretation of the sentence. We tested German sentences in the following three conditions (*all* vs. *most* vs. *four*).

(4) Alle / Die meisten / Vier Punkte sind miteinander verbunden.
All / The most / Four dots are with-one-other connected.
All / Most / Four dots are connected with each other.

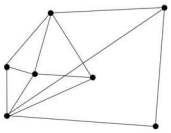
All-sentences were always paired with a picture consisting of four dots, whereas *most* and *four* had pictures with seven dots. Each participant completed five pictures for each quantifier. For this purpose, we

[2] See footnote 1.

drew 15 pictures with randomly distributed dots. In addition, we included 48 filler sentences. Half of them clearly required a complete (sub)graph, just like the experimental sentences in their strong interpretation. The other half were only consistent with a path. We constructed four pseudorandomized orders, making sure that two adjacent items were separated by at least two fillers and each condition was as often preceded by a complete graph filler as it was by a path filler. The latter was done to prevent participants from being biased towards either strong or intermediate interpretations in any of the conditions.

The completed pictures were labeled with respect to the chosen interpretation taking both truth conditions and scalar implicatures into account[3]. A picture was judged to show a strong meaning if the truth conditions in (3-b) were met and no implicatures of Q were violated. It was classified as intermediate if a (sub)graph of appropriate size was connected by a continuous path, but there was no complete graph connecting these nodes. Finally, a picture was labeled *weak* if Q nodes all had some connections, but there was no path connecting them all. Since we didn't find any weak readings, we will just consider the strong and intermediate readings in the analysis. Cases that did not correspond to any of these readings were coded as mistakes. Here is an example:

(5) Most of the dots are connected to each other.



Since the strong meaning of (5) requires at least four dots to form a complete subgraph, (5) is clearly false in this reading. The intermediate or weak reading is ruled out pragmatically, since all dots are connected by a continuous path. We checked whether participants obeyed pragmatic principles by analyzing sentences in the condition with *four*. In this condition participants (except for six cases) never connected more than four dots suggesting that they paid attention to implicatures.

Results

The proportions of strong meanings in the three conditions were analyzed using logit mixed effects model analyses (see eg. Jäger (2009)) with *quantifier* as a fixed effect and participants and items as random effects. We computed three pairwise comparisons: *all* vs. *most*, *all* vs. *four* and *most* vs. *four*. In all of these analyses, we only included the correct pictures.

Participants chose strong meanings in the *all*-condition 47.0% of the time, 22.9% in the *most*-condition and 17.4% in the *four*-condition. The logit mixed effects model analyses revealed a significant difference between *all* and *most* ($estimate = -1.82$; $z = -3.99$; $p < .01$) and between *all* and *four* ($estimate = -3.16$; $z = -5.51$; $p < .01$), but only a marginally significant difference between *four* and *most* ($estimate = .80$; $z = 1.65$; $p = .10$).

The error rates differed between conditions. Participants did not make a single mistake in the *all*-condition. In the *four*-condition 94.8% of the answers were correct. In the *most*-condition the proportion of correct pictures dropped down to 83.5%. Two pairwise comparisons using Fisher's exact test revealed

[3]Implicatures were only an issue in the four- and the most-conditions, but not in the all-condition.

a significant difference between *all* and *four* ($p < .05$) and a significant difference between *four* and *most* ($p < .01$).

Discussion

The results provide evidence against the SMH. Participants overwhelmingly drew pictures which do not satisfy a strong reading. In the *all* condition our data provide evidence for a real ambiguity between the strong and the intermediate interpretation. This is unexpected under the SMH; if the predicate *to be connected* is neutral, a strong interpretation should be favored. For the quantifiers *most* and *four*, the results provide even stronger evidence against the SMH. In these two conditions intermediate readings were clearly preferred over strong ones which were hardly, if at all, available.

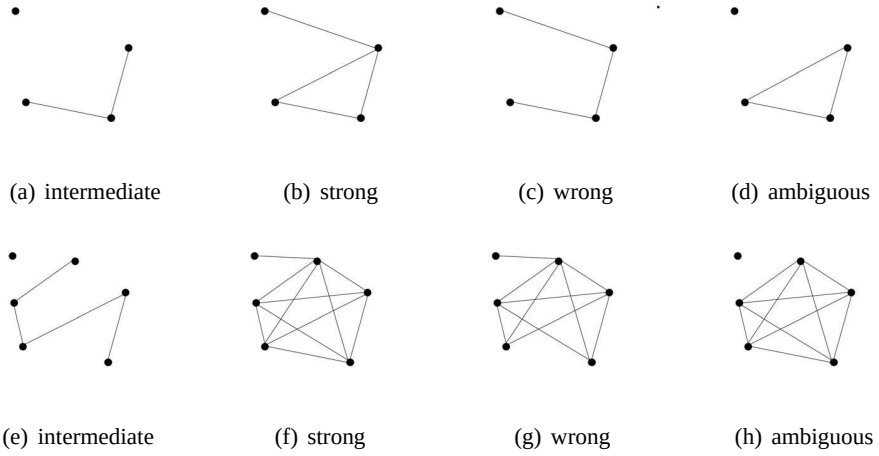
The PCH, on the other hand, receives initial support by our findings, in particular by the observed difference in the proportion of strong interpretations between *reciprocal all*, *reciprocal most* and *reciprocal four*. The error rates provide further support for the PCH. *Most* and *four* led to more errors than *all* did. This can be accounted for if we assume that participants sometimes tried to compute a strong interpretation but due to the complexity of the task failed to do so. To clarify whether this explanation is on the right track we clearly need real-time data on the interpretation process. This has to be left to future research. Another open question is whether the strong readings of *reciprocal most* and *reciprocal four* are just dispreferred or completely unavailable. This cannot be decided on the basis of the current experiment. What is needed instead is a task which allows us to determine whether a particular reading is possible or not.

Experiment 2: which readings are available?

The second experiment employed a picture verification task using clearly disambiguating pictures for strong vs. intermediate readings. Unfortunately, the quantifiers we used in the last experiment are all upward monotone in their right argument and therefore their strong interpretation implies the intermediate reading. Hence, even if the diagrams supporting the strong reading were judged to be true, we still wouldn't know which interpretation subjects had in mind. Luckily, in sentences that contain non-monotone quantifiers neither reading entails the other. We therefore chose the quantifiers *all but one*, *most* and *exactly n* in (6). *All but one* and *exactly four* are clearly non-monotone. For *most*, if we take the implicature *most, but not all* into account, it is possible to construct strong pictures in a way that the other readings are ruled out pragmatically. Crucially, the strong reading of *all but one* is still PTIME computable, although it is more complex than *all*. For instance, for verifying a model of size n, only the n subsets of size $n-1$ have to be considered. By contrast, verifying the strong meaning of (6-b,c) is intractable.

(6) a. Alle Punkte bis auf einen sind miteinander verbunden.
All dots except for one are with-one-another connected.

b. Die meisten Punkte sind miteinander verbunden.
The most dots are with-one-another connected.



(a) intermediate (b) strong (c) wrong (d) ambiguous

(e) intermediate (f) strong (g) wrong (h) ambiguous

Figure 1: Diagrams used in Exp. 2

c. Genau drei Punkte sind miteinander verbunden.
Exactly three dots are with-one-another connected.

We paired these sentences with diagrams disambiguating towards the intermediate or strong reading. Sample diagrams are depicted in Figure 1(a) and 1(b). For strong pictures, the PCH predicts lower acceptance rates for (6-b,c) than for (6-a). In order to find out whether the strong readings of (6-b,c) are dispreferred or completely unavailable we also paired them with false control diagrams (see Figure1(c)). The wrong pictures differed from the strong ones in that a single line was removed from the completely connected subset. If the strong reading is available for these two sentences at all, we expect more positive judgments following a strong diagram than following a false control. Furthermore, we included ambiguous diagrams as an upper bound for the intermediate pictures (cf. Figure 1(d)). If the strong meaning should conflict with an intermediate picture, we would expect more positive responses following an ambiguous diagram than following an intermediate diagram.

Secondly, as mentioned in the introduction we wanted to investigate whether availability of the strong reading in sentences with counting or proportional quantifiers depends on the size of the model. The strong meaning of (6-b,c) may be easy to verify in small universes, but not in larger ones. To test this possibility we manipulated the number of dots. Small models always contained four dots and large models six dots. We chose small models only consisting of four dots because this is the smallest number for which the strong meaning can be distinguished from the intermediate interpretation, so we could be sure that the task would be doable at all[4]. For the more complex six-dot pictures we presented sentences with *exactly five* instead of *exactly three*. Example diagrams are given in Figure 1[5]. In total, this yielded 24 conditions according to a 3 (*quantifier*) $\times$ 4 (*picture type*) $\times$ 2 (*size*) factorial design.

[4]We had the intuitive impression that pictures with ten dots were already far too complex to be evaluated by naive informants.

[5]The wrong pictures with six dots were slightly different for *most*. In these diagrams, all dots were connected by lines, but there was no subset containing four or more elements forming a complete graph.

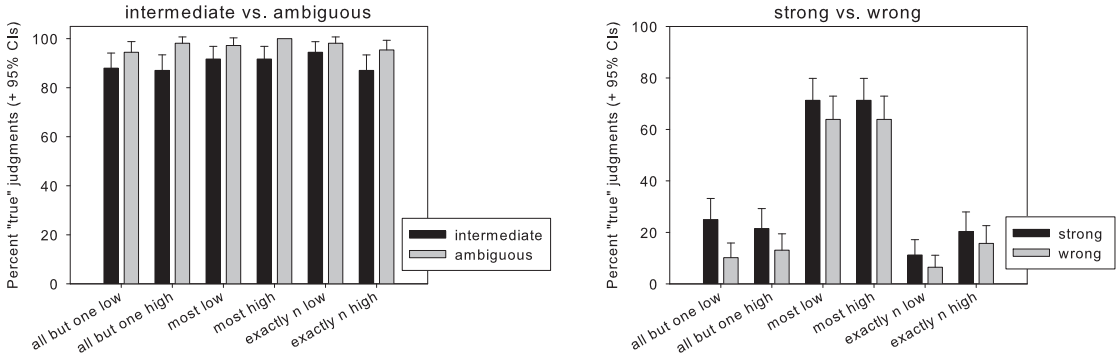


Figure 2: Mean judgments in Exp. 2 (*low* = pictures with 4 dots; *high* = pictures with 6 dots)

Method

Each participant provided three judgments per condition yielding a total of 72 experimental trials. We added 54 filler trials (20 false/34 true) and the 12 monotonicity trials from Experiment 3.

36 German native speakers (mean age 26.9 years; 23 female) read reciprocal quantified sentences on a computer screen in a self-paced fashion. When they finished reading the sentence, it disappeared from the screen and a dot picture was presented for which a truth value judgment had to be provided within a time limit of 10s[6]. Participants received feedback about how fast they had responded. This was done to trigger the first interpretation they had in mind. We collected judgments and judgment times, but because of space limitations will only report the former. The experiment started with a practice session of 10 trials, followed by the experiment with 138 trials in an individually randomized order. An experimental session lasted approximately 15 minutes.

Results

Two kinds of analyses were conducted on the proportion of 'true' judgments. The upper bound analyses concerned the default status of the intermediate interpretation by comparing intermediate picture conditions with ambiguous conditions. Lower bound analyses aimed at clarifying the status of the strong interpretation by comparing strong picture conditions with wrong conditions. The mean judgments of both analyses are presented in Figure 2.

Upper bound analysis: A logit mixed effects model analysis including *quantifier*, *reading* (*ambiguous* vs. *intermediate*), *complexity* and their interactions as fixed effects and participants and items as random effects only revealed a significant main effect of *reading* ($estimate = -2.37$; $z = -2.88$; $p < .01$). This main effect was due to an across-the-board preference (7.3% on average) of ambiguous pictures to pictures disambiguating towards an intermediate interpretation.

Lower bound analyses: We computed a logit mixed effects model analysis including *quantifier*, *truth* (*strong* vs. *wrong*), *complexity* and their interactions as fixed effects and participants and items as random effects. The only reliable effect was the fixed effect of *quantifier* ($estimate = 3.31$; $z = 8.10$; $p < .01$). The effect of *truth* was marginal ($estimate = 0.72$; $z = 1.77$; $p = .07$). As it turned

[6]Participants were very fast. On average they spent 2.5s reading the sentence and 1.8s to provide a judgment.

out, a simpler model taking into account only these two main effects and the random effects accounted for the data with a comparable fit. This was revealed by a comparison of the log-likelihood of the saturated and the simpler model ($\chi^2_{(8)} = 12.36$; $p = .14$). Thus, *complexity* had no significant influence on the judgments. The simple model revealed a significant main effect of *truth* ($estimate = 0.67$; $z = 4.08$; $p < .01$) which was due to 7.9% more 'true' judgments on average in the strong conditions than in the wrong conditions. The main effect of *quantifier* was also significant (*most* vs. *all/exactly*: $estimate = 3.21$; $z = 15.10$; $p < .01$). This was due to more than 60% acceptance for all *most* conditions but much lower acceptance for the other two quantifiers.

We analyzed the data by computing separate logit mixed effect models with fixed effects of *truth*, *complexity* and their interaction for all three quantifiers and simplified the models when a fixed effect failed to contribute to model fit. The best model for *all but one* contained only the fixed effect of *truth* which was reliable ($estimate = 1.04$; $z = 3.47$; $p < .01$), but neither *complexity* nor the interaction enhanced model fit ($\chi^2_{(2)} = 1.04$; $p = .60$). Thus, independently of *complexity* strong pictures were more often accepted than wrong pictures. The same held for *most* (fixed effect of *truth*: $estimate = 0.98$; $z = 2.71$; $p < .01$). *Exactly n* was different in that the fixed effect of *truth* and the interaction didn't matter ($\chi^2_{(2)} = 2.68$; $p = .26$), but *complexity* was significant ($estimate = -0.97$; $z = -2.96$; $p < .01$). This effect was due to more errors in complex pictures than in simpler ones.

Discussion

Overall, the intermediate reading was overwhelmingly preferred to the strong one. However, both the upper bound and the lower bound analyses provide evidence that the strong reading is available to some degree. Both analyses revealed a significant effect of picture type. Intermediate diagrams were less often accepted than the ambiguous diagrams. Moreover, strong diagrams were more often accepted than false ones. Focussing on *all but one* and *exactly n* with respect to the difference between the strong and wrong conditions the pattern looks as predicted by the PCH. The strong reading was possible for tractable *all but one* reciprocals but less so for intractable *exactly n* reciprocals. With *most*, the picture looks different. Even though verification of its strong meaning should be intractable, there was a reliable difference between the strong and wrong conditions. Thus, participants seemed to sometimes choose strong readings. An intractable problem can of course be innocuous under certain circumstances, for instance, when the input size is sufficiently small. The lack of effects of the number of dots manipulation points in this direction. Perhaps even the 'complex' conditions with six dots presented a relatively easy task. This brings us to a parametrized version of the PCH. A hard verification problem may be easy if we include parameters like the size and arrangement of the model. Although far from conclusive, we take our results as pointing in this direction.

Surprisingly, *most* was accepted quite often in the strong and the allegedly wrong conditions. The high acceptance rates in the latter indicate that participants were canceling the implicature of *most* and interpreting it as the upward monotone *more than half*. This also explains the high acceptance of the strong *most* conditions which were, without implicature, consistent with an intermediate interpretation.

Experiment 3: monotone increasing vs. decreasing antecedents

So far, we have been investigating reciprocal sentences with the upward monotone quantifiers *all*, *most*, *four* (Exp. 1) and the non-monotone quantifiers *all but one* and *exactly n* (Exp. 2). As it looks, only *all* licenses a strong interpretation easily. This finding may follow from the monotonicity plus implicatures. According to Dalrymple et al. (1998)'s SMH strong readings are preferred in sentences with upward monotone quantificational antecedents. For downward monotone quantifiers, on the other hand, intermediate readings should be preferred to strong readings. The reverse preferences are triggered by opposite entailment patterns. In the present experiment we compared upward monotone *more than n* with downward monotone antecedents *fewer than n+2*.

We paired diagrams like Figure 1(f) vs. Figure 1(e) with the two sentences in (7) according to a 2 (*monotonicity*) $\times$ 2 (*truth*) factorial design. The diagrams of the first type were identical to the strong pictures of the last experiment. With monotone increasing quantifiers they were ambiguous between strong and intermediate interpretations while in the monotone decreasing cases they disambiguated towards a strong interpretation. The second type of pictures disambiguated towards weak readings in monotone increasing quantifiers, but were ambiguous for monotone decreasing quantificational antecedents. On the basis of the first two experiments we expected high acceptance of both picture types with monotone increasing quantifiers, but much lower acceptance rates for (7-b) with strong than with ambiguous pictures. We constructed six items and collected three judgments from each participant in each condition. The experiment was run together with Experiment 2 using the same method.

(7) a. Mehr als vier Punkte sind miteinander verbunden.
More than four dots are with-one-another connected.

b. Weniger als sechs Punkte sind miteinander verbunden.
Fewer than six dots are with-one-another connected.

Results and Discussion

As expected, *upward monotone* antecedents were accepted in both picture types (ambiguous 98.1%; intermediate 92.5%). A logit mixed effect model analysis revealed no significant difference between the picture types ($estimate = 1.53$; $z = 1.60$; $p = .11$). This was completely different in sentences with monotone decreasing antecedents where strong pictures were only accepted in 13.0% of all trials while ambiguous pictures were accepted 92.6% of the time. This asymmetric distribution provides clear evidence that the predicate *be connected to each other* induced a bias towards the intermediate reading. Thus, although intended to be neutral we apparently chose a predicate that is far from optimal.

Conclusions

We have presented evidence that the kind of quantificational antecedent influences the amount of ambiguity displayed by reciprocal sentences. For example, in Exp. 1 only *all* reciprocals were fully ambiguous. Furthermore, comparing tractable reciprocals with antecedents *all* and *all but one* to intractable reciprocals with *n* and *exactly n* we found support for the predictions of the PCH. In reciprocals with *all* and *all*

but one strong readings were possible whereas *exactly n* blocked a strong interpretation. As for *most* the results are somewhat mixed. In Exp. 1 the strong reading was hardly available, but Exp. 2 showed that although dispreferred it is nevertheless possible.

At first sight, our findings provide evidence against the SMH. Strong interpretations were not the default in Exp. 1 and for the monotone increasing quantifiers in Exp. 3 weak interpretations were just as acceptable as the ambiguous pictures. However, contrary to our initial assumptions *be connected* doesn't seem to be neutral but seems to bias towards an intermediate interpretation. This may have to do with the transitivity of the relation. If two dots are only indirectly connected, it seems impossible to say that they are *not connected*, yet possible to say they are *not directly connected*. A next step, therefore, will be to apply the design of Exp. 2 to other predicates like *to know someone*, a relation that is clearly not transitive.

Another route to pursue is increasing the size of the models. A particularly strong test for the PCH would be to increase the model size up to a point where the acceptance rate for the strong reading of proportional quantifiers drops to the level of wrong pictures and see whether tractable antecedents still exhibit their strong interpretation. Exp. 2 was a first step in that direction but the size of the models was obviously still too small.

To conclude, we hope to have shown that relatively innocent looking reciprocal sentences with quantificational antecedents are an interesting test case for considerations of tractability in verification. More generally, within this domain research can be applied to a number of different constructions (for instance, branching quantifiers), so claims about computational complexity can be validated extending the test case investigated in the present study.

References

Dalrymple, M., M. Kanazawa, Y. Kim, S. McHombo, and S. Peters (1998). Reciprocal expressions and the concept of reciprocity. *Linguistics and Philosophy 21*(2), 159–210.

Garey, M. and D. Johnson (1979). *Computers and Intractability*. San Francisco: W.H. Freeman and Co.

Jäger, F. (2009). Categorical data analysis: away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language 59*(4), 434–446.

Kerem, N., N. Friedmann, and Y. Winter (2010). Typicality effects and the logic of reciprocity. In *Proceedings of SALT XIX*.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.

Szymanik, J. (2010). Computational complexity of polyadic lifts of generalized quantifiers in natural language. *Linguistics and Philosophy*. DOI: 10.1007/s10988-010-9076-z.

van Rooij, I. (2008). The tractable cognition hypothesis. *Cognitive Science 32*, 939–984.