Quantifying quantifier scope: A cross-methodological comparison

Oliver Bott & Janina Radó
Good linguistic theory can only be developed on the basis of solid data. Linguists have long recognized that it is not sufficient to trust the researcher’s intuitions – broader data sets need to be considered to make valid generalizations. In some cases the relevant data can be found in corpora. However, often the best solution is to systematically collect judgments from naive speakers, for instance by constructing questionnaires with multiple examples of the construction type under consideration. Subjects (naive informants) then rate these items on a (usually 5 or 7-point) scale or relative to a reference item (magnitude estimation). Such questionnaires have increasingly been used in syntax to buttress theoretical arguments with statistically quantifiable results.

Gathering quantifiable data may be advantageous in semantics as well, as semantic theories also tend to turn on subtle distinctions. Unfortunately, though, the questionnaires used in syntax cannot simply be “imported” into semantics, since the questions asked in the two fields are quite different. In judging syntactic wellformedness, subjects need to decide “Is the sentence grammatical/acceptable?”, which can be done on the basis of the sentence alone. In semantics, however, what we usually want to find out is how many (and what) readings a given sentence has. Thus we would need subjects to examine pairs of sentences and judge “Is the second sentence a good paraphrase of the first one?” To do that, subjects would have to compute the available meaning(s) for sentence 1, interpret sentence 2 and compare the two – a hopelessly complex task for naive informants. Clearly, a more suitable method is needed if we are to collect judgments from untrained native speakers.

In this paper we will develop some criteria for identifying methods that may be appropriate for collecting semantic judgments, and demonstrate the use of these criteria by comparing three candidate methods\(^1\).

1. **Finding suitable methods**

1.1. Sensitivity of the methods

Ease of use for untrained subjects is just one of the criteria a suitable method will have to fulfill. It is equally important to make certain that the method itself does not introduce any bias in the answers that subjects provide, and that it is capable of detecting subtle differences in judgment across construction types. We will start with the last criterion, and come back to the other two
In order to see whether the methods under consideration are sensitive to subtle differences, we need to test them on constructions where such differences are expected. For our comparison we have selected German versions of doubly-quantified sentences like (1):

(1) Everyone loves someone.

Sentences of this type typically permit two scope readings, depending on whether the universal or the existential quantifier is interpreted as having wide scope (the $\forall \exists$- and the $\exists \forall$-reading, respectively). The relative preference for these readings depends on a number of factors, such as the choice of quantifiers (for instance each vs. every), their linear order, and the syntactic role of the quantified expressions (see Beghelli and Stowell 2002, Pafel 2005, Tunstall 1998, among others). German is particularly interesting in this respect, as it allows the manipulation of linear (or hierarchical) order independently of syntactic role. By varying linear order and quantifier type, we generated a spectrum ranging from fully ambiguous to practically unambiguous sentences. Comparing preference judgments for sentences from different positions on the spectrum, we can test the sensitivity of our methods.

Doubly-quantified sentences are interesting for another reason as well: the available readings reported in the literature are sometimes quite controversial. For instance, (1) has been claimed both to be fully ambiguous (May 1977, 1985; Hornstein 1984; Higginbotham 1985), and to only allow the $\forall \exists$-reading (Reinhart 1976, 1983; Hornstein 1995; Beghelli and Stowell 2002). It thus seems desirable to obtain more solid evidence based on naive speakers’ judgments.

1.2. Disambiguation

As we are interested in assessing the relative availability of meanings, we need to pair potentially ambiguous sentences with appropriate disambiguation. The three methods we examine share the property that the disambiguating information is provided as a kind of context in which to evaluate the sentence. This is intended to simplify the subjects’ task, as no explicit comparison of meanings is necessary, only the judgment of how well the sentence fits the context. The candidate methods differ in how the context is provided: using linguistic means, that is, a preceding question, or visual means: set di-
agrams or natural-looking scenarios. We will describe the methods in detail in section 2, but first let us turn to some psycholinguistic motivation for the present study.

1.3. Do we need a methodological comparison?

The practice of using experimental methods to assess scope interpretations is in fact not new. Both linguistic context and diagrams have been employed to study scope preferences (Gillen 1991; Jackson and Lewis 2005). Is there any need for a new offline method, or is there anything else a methodological comparison can tell us? We think there is. First, the linguistic context used in a number of psycholinguistic studies turns out not to provide adequate disambiguation. As an example let us take the discourses used in Kurtzman and MacDonald (1993):

(2) Every boy climbed a tree.
(3) a. The tree was full of apples.
   b. The trees were full of apples.

While (3b) is felicitous only if the wide scope universal interpretation of (2) is taken, (3a) is consistent with both scope readings\(^3\). Thus if subjects accept (3a) as a good continuation we can’t tell what reading(s) they computed. This is a serious problem: it may turn out that in most cases when subjects select the singular continuation they in fact have the other reading in mind. Yet different variants of this type of disambiguation have been widely used both in online studies of processing and to assess offline preferences.

Furthermore, in order to test whether the (linguistic or visual) contexts we consider here fully disambiguate scope ambiguous sentences, we need to compare them under standardized conditions. This is made obvious by some of the conflicting results in the psycholinguistic literature. For instance, using Kurtzman and MacDonald-style disambiguation, Filik et al. (2004) found a significant effect of quantifier type (*every* vs. *a*) both offline and online, while Anderson (2004) detected no such influence. The discrepancy may have arisen from the different constructions and different tasks used in the two sets of experiments. Thus if we want to identify suitable methods to assess scope readings we must make the methods themselves the object of our study.
2. Pretesting the methods

For our methodological comparison we constructed 24 test items with a universal and an existential quantifier taking care to keep them equally plausible under both scope interpretations. Our intuitions concerning plausibility were further confirmed in a norming study.\textsuperscript{4}

To ensure a fair comparison, an effort was made to keep the test sentences (and the distractors) maximally similar across the candidate methods. Testing of the methods took place in two steps. The purpose of the first stage was to make sure that the methods under investigation are suitable for scope disambiguation in the first place. We then compared the same methods on the basis of their performance on ambiguous quantified sentences.

Apart from the different means of disambiguation, which constituted the point of comparison, the design and procedure used in the pretest were the same in all three methods. We first describe the general features, then turn to the particulars of the individual methods.

2.1. Materials and design

The aim of the pretest was to find out whether the chosen methods provide appropriate disambiguation. For this purpose we used constructions where the universal and existential quantifiers were placed in different clauses in a sentence, making it scopally unambiguous. The quantifiers appeared in two possible orders, as in (4):

\begin{enumerate}
  \item \textit{Für genau einen Professor gilt, dass jede Studentin ihn verehrt.}  
  \textit{Exactly one professor is such that every student adores him.}  
  \item \textit{Für jede Studentin gilt, dass sie genau einen Professor verehrt.}  
  \textit{Every student is such that she adores exactly one professor.}
\end{enumerate}

As the glosses indicate, in this construction the linear order of the quantifiers determines their scope order, due to the clause boundary between them. The 24 items all appeared with the two quantifier orders in (4). For each sentence we constructed two possible disambiguating contexts in each method under investigation, one for the $\forall$-$\exists$- and one for the $\exists$-$\forall$-interpretation. Each sentence version was paired with each context version, yielding a total of four conditions per sentence.
Four counter-balanced lists were created, each containing six items per condition in such a way that each item appeared only once on a list and across lists, all items appeared in all four conditions. Thirty-six distractor sentences were constructed as well to keep subjects from guessing the aim of the study, and to check whether they had followed the instructions. These sentences also included two quantifiers, representing a range of quantifier types, some of which were negated. A clause boundary between the two quantifiers makes these sentences unambiguous as well. The distractors were paired with disambiguating contexts and were included in all lists.

2.2. Subjects and procedure

Subjects were recruited at the University of Tübingen and were tested individually. They received written instructions followed by the questionnaire containing the 24 test items and 36 fillers in an individually randomized order. They were asked to provide yes-no answers to the question “Does the sentence fit the context?” A total of 68 subjects participated in the pretest (question-answer pairs: 24, set diagrams: 24, scenarios: 20). They were all native speakers of German and naive to the purpose of the study. No subject completed more than one questionnaire. Testing took approximately 20 minutes, subjects were paid €5.

2.3. Question-answer pairs

In the first method we examined, preceding questions provided a disambiguating context for the target sentence. We selected this method as it is a fairly intuitive means of disambiguation, often employed by linguists to determine possible readings. A further advantage of using question-answer pairs is that only the linguistic modality is needed for making judgments, which may aid in keeping the task simple.

The questions contained a universal and an existential quantifier separated by a clause boundary, just like in the target sentences. This resulted in the quantifier in the matrix clause taking scope over the one in the embedded clause. Two types of contexts were created (for the $\forall\exists$- and the $\exists\forall$-reading) by placing either the universal or the existential quantifier in the matrix clause. An example is given in (5):
Each version of the context question was paired with each version of the target sentence, repeated in (6), preceded by *Ja, stimmt* “Yes, that’s right”.

(6)  

a. Für genau einen Professor gilt, dass jede Studentin ihn verehrt.  
'Exactly one professor is such that every student adores him.'  
b. Für jede Studentin gilt, dass sie genau einen Professor verehrt.  
'Every student is such that she adores exactly one professor.'

Subjects were asked to provide yes-no answers to the question “Does the sentence match the question?”. As the example shows, the pairings that were expected to be good fits given the scope relations ((5a-6a) and (5b-6b)) are also the ones where the order of the quantifiers in the question is the same as in the answer. This raises the concern that higher acceptance rates in the matching conditions might be due to a preference for word order parallelism, rather than to scope interpretation. To rule out this possibility, in the second test phase we tested similar question-answer pairs where the quantified expressions were replaced with definite NPs with *diese/r* “this” like in (7) and (8) and found no evidence for a word order effect.

(7)  

a. Kann man von diesem Professor sagen, dass diese Studentin ihn verehrt?  
b. Kann man von dieser Studentin sagen, dass sie diesen Professor verehrt?

(8)  

a. Diesen Professor hat diese Studentin verehrt.  
b. Diese Studentin hat diesen Professor verehrt.
2.3.1. Results and discussion

The leftmost panel in Figure 1 shows the percent “yes” answers. The conditions where the context question and the target sentence allowed the same scope reading were overwhelmingly judged as “matching” (∃∀-sentences: 98.6%, ∀∃-sentences: 97.9%) whereas conditions where they were in conflict were for the most part judged non-matching (∃∀: 38.2%, ∀∃: 34.9%). Repeated measures ANOVAs\(^5\) revealed no main effect (all \(F < 1\)), but, as expected, a significant interaction was found between sentence and context (\(F_1(1, 23) = 95.98; p < 0.01; F_2(1, 23) = 267.539; p < 0.01\)).

These results show that overall context questions can disambiguate quantifier scope in a target sentence quite well, and are suitable for use with naive subjects. Still, the relatively high percent of false “yes” answers for non-matching pairs raises the question how reliable the judgments gathered by means of question-answer sequences really are. We will come back to this point in section 2.6.

2.4. Set diagrams

Another mode of disambiguation that psycholinguists have employed to study quantifier scope is set diagrams like those in (9) below (see Gillen 1991; Jackson and Lewis 2005). This method seems attractive since the diagrams are quite simple and easy to work with.

(9) a. Für genau einen Professor gilt, dass jede Studentin ihn verehrt.  
‘Exactly one professor is such that every student adores him.’
b. Für jede Studentin gilt, dass sie genau einen Professor verehrt.
'Every student is such that she adores exactly one professor.'

In the $\forall\exists$-diagram (Figure 2), each element of the “student” set depicted on the right is connected to some element of the “professor” set (indicating the “adore” relation) – different “professors” connected to different “students”, consistent with the wide scope universal reading. In the $\exists\forall$-diagram, on the other hand, there is one single professor that is admired by all students. To fully disambiguate the diagram another instance of the relation was included as well, one between a different professor and one of the students. In this form the diagram is no longer consistent with a $\forall\exists$-reading where all students admire exactly one professor, who happens to be the same one in each case. This extra line apparently made the $\exists\forall$-diagrams more complex, reflected in comments we received from the subjects. However, as the results below indicate, this did not have a negative effect on subjects’ performance.

2.4.1. Results and discussion

Results of the set diagram pretest are given in the middle panel in Figure 1. Again, diagram-sentence pairs indicating the same scope reading were rated as matching ($\exists\forall$-sentences: 95.1%, $\forall\exists$-sentences: 88.9%) while conflicting conditions were overwhelmingly rated as non-matching ($\exists\forall$-sentences: 6.3%, $\forall\exists$-sentences: 14.6%). ANOVAs revealed a significant interaction of sentence and context ($F_1(1, 23) = 267.349; p < 0.01; F_2(1, 23) = 1103.530; p < 0.01$), but no significant main effects.

Thus it appears that set diagrams are easily used by naive subjects and are capable of disambiguating the scope readings. Clear disambiguation of the $\exists\forall$-reading is particularly important in light of the problems in psycholin-
guistic experiments discussed above. Further, the non-matching conditions were judged correctly just as often as their matching counterparts, in contrast to the question-answer pairs (see section 2.3).

2.5. Natural-looking scenarios

We also tested another kind of visual context, viz. severely simplified representations of two children playing with geometrical figures in distinct but partially overlapping “corners”, as in Figures 4-5 below. In this method the test sentences were also slightly modified to fit the scenarios:

(10) a. Für genau ein Dreieck gilt, dass jedes Kind es in seiner Ecke hat.
    ’Exactly one triangle is such that every child has it in his corner.’

b. Für jedes Kind gilt, dass es genau ein Dreieck in seiner Ecke hat.
    ’Every child is such that he has exactly one triangle in his corner.’

One advantage of these scenarios is that they depict relatively natural situations. Moreover, the possibility of lexical content influencing plausibility is eliminated, since it is not any more likely that a child has a particular geometrical object than any other state of affairs.

2.5.1. Results and discussion

The distribution of “yes”-answers is displayed in the rightmost panel in Figure 1. Again, there is a clear difference in the percentage of “yes” answers
between the matching (∃∀: 81.9%, ∀∃: 86.8%) and non-matching conditions (∃∀: 20.8%, ∀∃: 15.3%). In ANOVAs, this resulted in a significant interaction between sentence and context (\(F_1(1,23) = 116.943; p < 0.01; F_2(1,23) = 201.813; p < 0.01\)), while the main effects didn’t reach significance. This shows that the diagrams were only consistent with the reading they were intended to depict.

In sum, the pretests show that all three methods of disambiguation were successful in indicating the intended scope readings. However, a closer inspection of the results reveals sizeable differences among the candidate methods, especially in the case of false “yes” answers. This raises the question whether some of our methods may have performed “better” than the others.

2.6. Reliability of the three methods

Rating methods have often been considered unsuitable for scientific purposes. However, the application of test theoretic standards in psychology has shown that methods involving judgments do in fact produce highly consistent and valid results\(^7\). We treated our pretest results as a kind of psychological test in order to ask the following questions: to what extent are our methods able to provide consistent data? Are there any substantial differences between the methods in terms of consistency/reliability? To answer these questions we have to determine how consistent the participants were contingent on the applied method\(^8\). This can be done by computing the interrater reliability.

Intraclass correlation coefficients are statistics for measuring homogeneity, not only for pairs of measurements but for larger sets of measurements as well (for an overview see McGraw and Wong 1996; Wirtz and Caspar 2002). Perfect agreement among participants would be reflected by an ICC coefficient of 1, while total disagreement would yield a coefficient of 0. Although there are no guidelines concerning the interpretation of ICC coefficients (see the discussion in Wirtz and Caspar 2002), tests yielding values of 0.7 or greater are often considered to be sufficiently reliable.

In the pretests each participant saw six items in each condition according to a latin square design. We computed the relative frequency of “yes”-answers for each subject in each condition (ranging from zero to six out of six answers). Using these we computed two-way mixed effects intraclass correlation coefficients (interaction absent) for each method based on absolute agreement among participants. For single raters we obtained ICCs of 0.69
for question-answer pairs, 0.85 for set diagrams and 0.78 for scenarios. Although question-answer pairs received the smallest ICC score, the numerical differences between the candidate methods were not significant. Furthermore, for average raters all three methods produced coefficients greater than 0.95, showing that experimental results obtained by any of our candidate methods and using a large sample of participants should yield highly consistent semantic judgment data.

3. Comparing the methods

All three candidate methods proved to be reliable when used with scope-unambiguous sentences. This qualifies all of them for the real test case. Are they also suited to detect subtle scope preferences in potentially ambiguous sentences? This is a question about the validity of the methods. We will investigate two aspects of validity. Construct validity indicates whether the results the methods produce correspond to the consensus in the theoretical literature on quantifier scope, whereas criterion-related validity shows how well the results mirror actual scope preferences in ambiguous sentences found in corpora. We will first introduce these two criteria, then discuss the validity of each of our methods.

3.1. Construct validity: hypotheses from the literature

Although some theories on quantifier scope state that a sentence containing two or more quantifiers has all combinatorily possible scope readings (for example May 1977, 1985), two factors have repeatedly been claimed to limit the scope potential of multiply quantified sentences. First, it is commonly assumed that a linear reading is preferred over an inverse reading. Second, distributive quantifiers like each have a stronger tendency to take wide scope than non-distributive quantifiers like all, see for instance Beghelli and Stowell (2002). In (11) both of these features are manipulated.

(11) a. Genau einen dieser Professoren haben alle Studentinnen verehrt.
    exactly one of-these professors have all students adored
    'All students adored exactly one of these professors.'

   b. Genau einen dieser Professoren hat jede Studentin verehrt.
    exactly one of-these professors has every student adored
'Every student adored exactly one of these professors.'
c. Alle Studentinnen haben genau einen dieser Professoren verehrt.
'All students adored exactly one of these professors.'
d. Jede Studentin hat genau einen dieser Professoren verehrt.
'Every student adored exactly one of these professors.'

Pafel (2005) develops a theory on quantifier scope in German with a broad empirical coverage. In this account scope preferences are computed on the basis of an additive linear model of a number of individually weighted scope-determining factors including linear order and distributivity. For doubly quantified sentences like the ones in (11) Pafel’s theory makes the predictions in Figure 6.¹⁰

For every combination of the manipulated factors linear order and distributivity each quantifier received a scope value. The difference between the scope values of the quantifiers within a sentence reflects the scope preferences for that sentence: the \( \exists ... \forall \) cases are predicted to be fairly scope ambiguous, while the \( \forall \) ... \( \exists \) and the \( \forall \) ... \( \forall \) cases should exhibit a preference for the wide scope universal reading, and the \( \exists \) ... \( \forall \) cases should prefer a wide scope existential interpretation. Furthermore, the influence of the two factors linear precedence and distributivity is predicted to be independent, which should give rise to purely additive effects.
3.2. Criterion-related validity: results from a corpus study

We conducted a corpus study on all public written corpora of German in the Cosmas Corpus. Our aim was to extract constructions that are maximally similar to the sentences in (11). We queried sentences without any embedding containing two quantifiers: *ein* as direct object and *jede* or *alle* as subject of a simple transitive verb, where one of the quantifiers occurred in sentence-initial position. The maximum distance between the quantifiers was four intervening words. To rule out any indefinite use of *ein* we excluded those occurrences where it could be phonetically reduced to *n*. Most importantly, we chose only sentences where the context clearly indicated the intended scope.

The distribution of readings obtained in the corpus study is shown in Figure 7. Although sparseness of data in some of the conditions made a quantitative analysis impossible the results of the corpus study fit quite nicely with the predictions of Pafel’s theory.

![Figure 7. Distribution of scope readings: Corpus study](image)

Constructions with a universal quantifier (= subject) preceding the existential quantifier (= direct object), which were predicted to preferentially have a wide scope universal reading, are in fact used mainly with that interpretation. This tendency is stronger for distributive *jede* than non-distributive *alle* which also had some wide scope existential uses. The six occurrences of *ein* ... *alle* all received a wide scope existential interpretation. Only the results for the *ein* ... *jede*-cases are somewhat surprising: although predicted to be fully scope-ambiguous, they were preferably used with a wide scope univer-
sal interpretation. Given the very few corpus instances of this construction this apparent difference cannot be interpreted. Taking together, the theoretical predictions and the results from the corpus study provide a consistent background against which we can validate the three disambiguation methods.

3.3. Design and procedure

As in the pretest, we made an effort to keep the materials and procedure maximally similar across the methods. The test sentences were modified versions of the items used in the pretest in that the two quantifiers were clause-mates, to make both scope readings possible. In addition to varying the linear order of the quantifiers, we also added the factor *distributivity*, yielding the four conditions given in (11) above. They were paired with the same two versions of the disambiguating context that we had used in the pretest, yielding a total of eight conditions. Fillers from the pretest were modified the same way as the test items.

In this design, the different scope potentials of *jede* and *alle* predicted by Pafel (2005) should be reflected by an interaction between *distributivity* and *context*: thus for instance the *ein ... jede* cases are expected to be more compatible with the $\exists\forall$-interpretation, and less compatible with the $\forall\exists$-interpretation, than the *ein ... alle* cases. Similarly, the order of quantifiers is predicted to show its influence on scope in an interaction between *order of quantifiers* and *context*: the *jede ... ein* order should be more compatible with the $\forall\exists$-reading, and less compatible with the $\exists\forall$-reading, than the *ein ... jede* order. Finally, since in Pafel’s theory these effects are independent, there should be no three-way interaction of *distributivity*, *order*, and *context* (see also section 3.1 and Figure 6).

The simple yes-no answers we collected in the pretest are not sufficient to capture the subtle scope differences we expected in this part of the study. Therefore we decided to use magnitude estimation, where each test item is compared to a reference item (constant throughout the test). Subjects provide judgments on an individual open-ended scale, expressing as many distinctions among the items as they feel necessary (see Bard et al. (1996) for details of the method).

For all tests we will report z-transformed ratings, computed on the basis of all judgments for experimental items plus filler trials. Positive values indicate “yes”-judgments, negative values indicate “no”. The resulting scores were
subjected to three-way repeated measures ANOVAs with order of quantifiers, distributivity (jede vs. alle) and disambiguating context (∃∀ vs. ∀∃) as within-factors.

Subjects were tested individually. They first received detailed written instructions and then completed a short practice session to familiarize themselves with the magnitude estimation technique. The actual test session consisted of 60 items (24 test items and 36 distractors), and took approximately 30 minutes to complete.

Subjects were recruited in the same way as for the pretest. They were all native speakers of German and naive to the purpose of the study. A total of 96 subjects participated in the tests (question-answer pairs: 32, set diagrams: 24, scenarios: 40) for a payment of €5. No subject completed more than one questionnaire or had participated in any of the pretests.

3.4. Question-Answer pairs

In this test one version of the scope-disambiguated questions in (5) was paired with one of the four item versions given in (11) above, preceded by Ja, stimmt “Yes, that’s right”.

The distribution of readings in the question-answer test is shown in Figure 8. While the ein ... jede-cases were considered scope ambiguous, the jede ... ein- and alle ... ein-cases preferably received a wide scope universal interpretation, and the ein ... alle-cases a wide scope existential interpretation.

![Figure 8. Question-answer pairs: Results](image_url)
ANOVARs revealed significant interactions between *order* and *context* \((F_1(1,31) = 14.176; p < 0.01; F_2(1,23) = 36.674; p < 0.01)\) and between *distributivity* and *context* \((F_1(1,31) = 10.787; p < 0.01; F_2(1,23) = 12.032; p < 0.01)\). The former is due to a general preference for linear scope readings, the latter to distributive *jede* taking wide scope more easily than *alle*. Crucially, the three-way interaction between *order*, *distributivity* and *context* was far from significant \((Fs < 1)\). No other effects were significant.

Thus question-answer pairs appear to be sensitive to subtle distinctions in scope preferences. Both manipulated factors showed an effect in the expected direction. Furthermore, the effects were purely additive, as predicted by Pafel’s theory. In terms of validity, question-answer pairs thus elicit good semantic judgements, at least as long as standards used in psycholinguistics (lots of items, randomized presentation, the use of distractors etc.) are adhered to.

### 3.5. Set diagrams

For this test the four sentence versions in (11) were paired with diagrams of the type shown in Figures 2-3 above.

The distribution of readings is depicted in Figure 9. Again, ANOVARs revealed significant interactions between *order* and *context* \((F_1(1,55) = 20.705; p < 0.01; F_2(1,23) = 41.729; p < 0.01)\) and between *distributivity* and *context* \((F_1(1,55) = 80.059; p < 0.01; F_2(1,23) = 54.146; p < 0.01)\). These interactions are due to two independent factors: First, the leftmost quantifier receives wide scope and second, distributive quantifiers receive wide scope more easily than non-distributive quantifiers. There was only one other significant effect, the interaction between *order* and *distributivity* \((F_1(1,55) = 5.398; p < 0.05; F_2(1,23) = 4.722; p < 0.05)\). This interaction results from the *jede* ... *ein*-conditions receiving higher scores on average than all other constructions.

Thus this method also allowed us to capture fine distinctions concerning scope ambiguous sentences. Again, the results are fully consistent with our validation criteria.
3.6. Scenarios

The four sentence versions in (12) were paired with diagrams of the type given in Figures 4-5 above.

(12) a. Genau eines dieser Dreiecke haben alle Kinder in ihrer Ecke.  
   exactly one of these triangles have all children in their corners  
   'All children have exactly one of these triangles in their corners.'

b. Genau eines dieser Dreiecke hat jedes Kind in seiner Ecke.  
   exactly one of these triangles has every child in his corner  
   'Every child has exactly one of these triangles in his corner.'

c. Alle Kinder haben genau eines dieser Dreiecke in ihrer Ecke.  
   'All children have exactly one of these triangles in their corners.'

d. Jedes Kind hat genau eines dieser Dreiecke in seiner Ecke.  
   'Every child has exactly one of these triangles in his corner.'

The pattern of results obtained with the scenarios was quite different from the previous ones: there was an across-the-board preference for the wide scope universal interpretation (see Figure 10). In ANOVAs, this preference was reflected by a significant main effect of disambiguation ($F_1(1, 39) = 88.254; p < 0.01; F_2(1, 23) = 97.014; p < 0.01$). No other effects reached significance.
Figure 10. Scenarios: Results

The overwhelming preference for the wide scope universal reading irrespective of the linguistic manipulation is quite puzzling, particularly in light of our pretest results. In the pretest, where we used disambiguated versions of the same sentences and exactly the same diagrams, both readings were equally available. Thus the method seems to show a different behavior when applied to scope ambiguous and to scope disambiguated sentences.

One possible explanation for this strange behavior may have to do with the relative semantic complexity of the two readings. Closer scrutiny reveals that the quantified expression *genau ein* “exactly one x” corresponds to not one, but two quantifiers: like definite NPs, it must be analyzed as $\exists x\phi(x) \land \forall y(\phi(y) \rightarrow y = x)$. Moreover, the same “exactly one x” expression is hidden inside the predicate “have something in one’s corner” as well, which can be paraphrased as “there is exactly one x (x = corner) belonging to y and y has something in x”. When the two “exactly one” quantifiers are adjacent as in (12 c-d), it is possible to simplify them as “exactly one corner with one triangle”. The two scope versions of the sentence could thus be rendered roughly as “Every child is such that there is exactly one corner with one triangle such that it belongs to the child” (the $\forall\exists$-reading) and “There is exactly one triangle such that for every child it is the case that there is exactly one corner belonging to that child such that the triangle is in that corner” (the $\exists\forall$-reading). There is a clear intuitive difference in complexity between the two scope versions; moreover, this difference can be formalized in terms of the arithmetical hierarchy in recursion theory. It is thus possible that the wide
scope existential reading was avoided by the subjects because it is considerably more difficult to process.\textsuperscript{13}

Another possibility has to do with pragmatics. A sentence like “Every child has a triangle in his corner” is typically considered felicitous in a situation where each child is the exclusive owner of a triangle, but not if two children share one. This clearly favors the $\forall\exists$-diagrams as long as the corresponding sentences allow this interpretation. Thus $\forall\exists$-diagrams are expected to receive high ratings even when paired with a sentence with the \textit{genau ein} \ldots \textit{alle} order, where the wide scope existential reading is predicted to be the preferred, but not the only interpretation (recall that this was not the case in the $\exists\forall$-conditions in the pretest). What happens when the same sentence appears together with a $\exists\forall$-diagram? It is conceivable that subjects settle on one interpretation of the sentence first, without examining the scenario. The wide scope existential reading, bolstered by the semantic factors, is rejected in favor of the wide scope universal one, and the latter is the only reading available when the subject inspects the diagram. In this case we would expect the $\exists\forall$-diagram to be rated low, which is exactly what we found.

4. \textbf{General discussion}

Our methodological comparison turned up two suitable methods for studying quantifier scope: question-answer pairs and set diagrams. They both provided valid and reliable data, although the former did not perform quite as well as the latter. By contrast, our third method did not live up to the expectations: the simple and attractive natural-looking scenarios had to be paired with sentences that are clearly not appropriate for testing naive speakers.

A linguistically relevant result of the present study is providing evidence for subtle distinctions in judgments across the different constructions we tested. This lends support to semantic theories that rely on the interaction of multiple factors to determine scope relations. The methods that did well in the comparison can also be used to assess the influence of other factors on quantifier scope (e.g. intonation). In addition, they can be applied to related phenomena such as the interaction of modals with negation or collective vs. distributive readings.

Moreover, the present comparison has implications for psycholinguistic research. First, our “winner” methods are suitable to determine offline preferences that are needed to interpret online processing results: They can replace
those “disambiguating” contexts which, as we argued in section 1.3, fail to identify the reading subjects have in mind. Of course knowing what reading was eventually computed is crucial for the correct interpretation of processing data as well. The methods that passed our tests can also be employed as comprehension questions in online experiments to assess this reading.

However, the central contribution of the present study goes beyond possible applications of the question-answer pairs or set diagrams. More importantly, we have demonstrated the use test-theoretical standards for finding sound methods to collect data in linguistic or psycholinguistic research.

References


Jackson, Scott and Will Lewis 2005 The relation between prosody and logical scope varies by the operator. Poster presented at the CUNY Conference Tucson, Arizona.


<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Title</th>
<th>Publisher</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oddifreddi, Piergiorgio</td>
<td>1999</td>
<td><em>Classical recursion theory</em></td>
<td>Elsevier, Amsterdam</td>
<td></td>
</tr>
<tr>
<td>Pafel, Jürgen</td>
<td>2005</td>
<td><em>Quantifier scope in German: An investigation into the relation between syntax and semantics</em></td>
<td>Johns Benjamins, Amsterdam</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1983</td>
<td><em>Anaphora and semantic interpretation</em></td>
<td>Croom Helm, London</td>
<td></td>
</tr>
<tr>
<td>Wirtz, Markus and Franz Caspar</td>
<td>2002</td>
<td><em>Beurteiltereinstimmung und Beurteilerreliabilität</em></td>
<td>Hogrefe, Freiburg</td>
<td></td>
</tr>
</tbody>
</table>
Notes

1. Strictly speaking, our discussion in this paper concerns subtypes of the questionnaire method used e.g. in psycholinguistic research. We will nevertheless refer to these subtypes as separate "methods", in order to keep the presentation simple.

2. Or c-command, although we prefer the more theory-neutral linear order.

3. This becomes clear if (3a) is combined with disambiguated paraphrases of (2):

   (i) a. Every boy is such that he climbed a tree. The tree was full of apples.
       b. There was a tree that every boy climbed. The tree was full of apples.

4. All test sentences are given in the Appendix in the scope-ambiguous version, see section 3.3.

5. For all pretests we will report ANOVAs with two within factors: scope in the target sentence (∃∀ vs. ∀∃) and scope in disambiguating context (∃∀ vs. ∀∃).

6. The set denoted by the subject phrase always appeared on the right-hand side of the diagram.

7. Validity will be taken up in section 3.

8. In all pretests the ratings were very consistent across items, indicated by the items analyses. The variance across pretest results is thus mainly due to differences between subjects.

9. We call a reading linear if the order of quantifiers in a sentence corresponds to their scope, and inverse otherwise.

10. For our predictions we used a model without quantitative thresholds, unlike Pafel (2005). This was because in a whole series of experiments on quantifier scope we never found evidence for such thresholds.

11. Note that the actual distribution (with ∀∃:∃∀ = 14:7) does not differ significantly from an assumed perfectly ambiguous one (t(20) = 1.58; p = 0.13).

12. All items are given in the Appendix.

13. The prefixes for the two readings in prenex normal form are ∀x∃y∃z and ∃y∀x∃z, respectively. Since two adjacent quantifiers of the same type can be contracted into a new quantifier over pairs (see Oddifreddi 1999), after permuting the universal and the existential quantifiers we get ∀x∃y,z vs. ∃y∀x∃z, which means a difference of at least one quantifier on the arithmetical hierarchy.