How to provide exactly one interpretation for every sentence, or what eye movements reveal about quantifier scope

Oliver Bott & Janina Radó
SFB 441, University of Tübingen

1. Introduction

A sentence with two quantifiers is potentially scope-ambiguous: the second quantifier may take narrow or wide scope with respect to the first one. Whether both interpretations are discernible, though, depends on factors like the type of the quantifier, its syntactic function, and word order. For instance, *all* tends not to take wide scope. When it appears in the object position, following a subject quantifier, the only perceived reading may be the one corresponding to the linear order of the quantifiers, as in (1a)

(1) a. Some airline serves all continents.
   b. Some airline serves each continent.

However, replacing *all* with *each* as in (1b) makes a second, inverse scope reading available as well. This is because *each*, which is distributive, demands wide scope.

If our primary interest is to determine what factors influence quantifier scope and how strong these factors are, it may be sufficient to examine the relative preference among the readings of a multiply-quantified sentence. In some cases, though, it is also important to find out how we arrive at those readings. Is a quantified expression interpreted immediately? If it is then we expect the computation of relative scope to begin as soon as the second quantifier is encountered. Alternatively, the interpretation of quantifiers could be delayed until some interpretation domain (e.g. the clause) has been processed completely.

Although both the process of quantifier interpretation and the final preference are highly relevant for semantic and psycholinguistic theories of
quantifiers, existing studies examine these questions separately. The offline measures that are used to establish the preferred reading cannot assess when and how the reported reading was computed. The time course of interpretation is investigated in online experiments, typically using reading time measures. These experiments are supposed to reveal whether there is initial commitment to one reading; if this is the case then a continuation incompatible with that reading should cause measurable difficulty in processing. For this method it is thus necessary to disambiguate the sentences towards one or the other scope interpretation and compare the reading times at the point of disambiguation. A frequently used disambiguation method is shown in (2).

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

We have argued elsewhere (Bott and Radó 2007) that this type of disambiguation is not sufficient: The singular continuation in (2a), which is intended to only allow the wide-scope existential reading (all children climbed one and the same tree) is compatible with the wide-scope universal reading as well, roughly as "the tree that the child climbed". What this means is that the reading times for (2a) may include both scope interpretations, or possibly an underspecified representation. Thus a different method of disambiguation is necessary to be able to interpret the results. However, there is also a more general problem with this type of online experiment, namely a possible distortion of the preferences on earlier, ambiguous parts of the sentence. If the disambiguation is successful, we can't tell anymore how big the preference for one reading had been – the indication of difficulty only shows which reading was preferred at the point where the disambiguating material was encountered.

Thus online and offline results contribute different aspects to the picture of quantifier scope interpretation, but the pieces cannot be fitted together easily. Even if the same set of materials is tested both in an offline questionnaire and in a reading-time experiment, it is still not possible to map a set of reading times for a given item to the reported readings for that item. This is a problem not only for psycholinguistic theories but also for semantic analyses of scope: it seems that the only thing we can test is whether the predicted reading is in fact the preferred one – we cannot measure or compare the size of preference in different constructions.
In this paper we propose a method to overcome this problem. We describe
a way of measuring reading times and determining the final interpretation
without imposing a particular interpretation on the subjects. That way we
should be able to tell both which reading is preferred, and when readers
converge on that reading. The method we will present here is a version of
the visual world paradigm (Tanenhaus et al. 1995). Participants read scope-
ambiguous sentences of the sort given in (3)\(^1\):

\begin{enumerate}
\item \textit{Genau ein Tier auf jedem Bild sollst du nennen!}
\begin{itemize}
\itemexactly one animal on each picture should you name
\item ‘Name an animal in each field.’
\end{itemize}
\item \textit{Genau ein Tier auf allen Bildern sollst du nennen!}
\begin{itemize}
\itemexactly one animal on all pictures should you name
\item ‘Name an animal in all fields.’
\end{itemize}
\end{enumerate}

They inspect computer displays in order to provide an answer. The displays
are constructed in such a way as to be compatible both with a wide-scope
universal and a wide-scope existential reading of the sentence. Eye-
movements monitored during reading reveal whether there is any difference
in the way the constructions with \textit{all} vs. \textit{each} are interpreted; the final
answer participants provide show which interpretation they chose.

2. Experiment

2.1 Inverse linking

The experiment we report here was conducted in German. We tested
sentences like (3) above, which exemplify the phenomenon called inverse
linking. In inverse linking constructions the quantifier embedded in a PP
inside an NP prefers to take wide scope (May and Bale 2006). In our
examples this corresponds to a highly salient wide-scope universal
interpretation.

This construction was chosen for two reasons: first, the influence of certain
scope factors such as distributivity is well-documented in subject-object
and double-object configurations (cf. Kurtzman and MacDonald 1993,
Tunstall 1998, Filik, Paterson, and Liversedge 2004, for German see Pafel
2005, Bott & Radó 2007), but it has not been investigated in inverse linking
constructions. Second, in this construction both quantifiers are contained in
an NP preceding the verb. This makes it possible to separate the "pure" effect of the quantifiers from the influence of syntactic position or thematic roles.

2.2 Design

The ambiguous experimental conditions always involved a sentence-initial NP consisting of an existential quantifier and a PP containing a universal quantifier. *Genau ein* (exactly one) was chosen as existential quantifier to exclude the possibility of a non-quantificational reading. We compared two types of universal quantifiers: *jeder* (each) and *alle* (all) (cf. (3)). *Jeder* is distributive and has a strong tendency to take wide scope, which should make the inverse scope reading even more salient in (3a). *Alle*, on the other hand, prefers narrow scope, which is expected to conflict with the inverse-scope preference inherent in the construction.

Figure 1: Sample picture used with scope ambiguous sentences and Control B
The sentences were paired with displays consisting of three fields with four pictures each (cf. Figure 1). The wide-scope existential interpretation of (3) requires a particular animal in every field. There were two animals in the display (viz. the bear and the monkey) that satisfied this requirement, to make the use of genau ein (exactly one) felicitous as well\(^2\). The other two pictures in a field were also animals, but different ones across fields. Thus for a wide-scope universal interpretation subjects could name the dog in field 1, the bear in 2, and the tiger in 3, for instance.

In addition to the ambiguous quantifier conditions we also included two types of control conditions. Control B consisted of unambiguous sentences containing two quantifiers, such as those in (4).

(4) Control B:  
\(<a>\text{Von jedem Bild sollst du irgendein Tier nennen!}
\text{ of each picture should you some animal name}
\text{‘From each field name an animal.’}\)
\(<b>\text{Ein Tier, das sich auf allen Bildern befindet, sollst du nennen!}
\text{ an animal that self in all pictures locates should you name}
\text{‘Name an animal that is to be found in all fields.’}\>

These constructions used similar quantifiers (\(\forall\) and \(\exists\)) as the ambiguous quantifier conditions in (3). However, the particular syntactic configuration only allows a linear scope reading. The sentences in Control B were paired with the same displays as the items in (3). Their purpose was to check whether subjects do indeed compute the necessary reading.

Another set of control conditions (Control A) was needed in order to interpret the reading times on the doubly-quantified sentences. In Control A the existential quantifier was replaced with a definite NP. Definite NPs are typically considered non-quantificational (e.g. Strawson 1950, Heim 1991, Glanzberg forthcoming, but see Russell 1905, Neale 1990), thus we did not expect any scope interaction with the universal quantifier in these conditions. That means that any reading time difference between (5a) and (5b) must reflect pure lexical differences between jeder and alle, which must be taken into account in comparing (3a) and (3b). At the same time the answers provided to (5a) and (5b) may be informative concerning the debate about the status of definite NPs (see the references above).
(5) Control A
   a. *Das Tier auf jedem Bild sollst du nennen!*
      the animal on each picture should you name
      ‘Name the animal in each field.’
   b. *Das Tier auf allen Bildern sollst du nennen!*
      the animal on all pictures should you name
      ‘Name the animal in all fields.’

To satisfy the uniqueness presupposition introduced by the definite NP, the displays that appeared with Control A only included one picture corresponding to the NP in the sentence. This picture (in this case, the deer) was the same in all three fields. All other pictures belonged to a different category (cf. Figure 2).

![Figure 2: Sample picture used with Control A](image)

To summarize, the sentence materials included the following six conditions:
(6) a. two quantifiers, 'each', ambiguous
b. two quantifiers, 'all', ambiguous
c. definite NP, 'each'
d. definite NP, 'all'
e. Control B: ∀∃ only
f. Control B: ∃∀ only

An experimental trial consisted of the subject reading a sentence, then inspecting the corresponding display, and finally providing an answer. We expected condition (6a) to be easy to process since the inverse scope bias that is inherent in the inverse-linking construction fits well with jeder's need to take wide scope. Further, the resulting wide-scope universal interpretation should be reflected in a large proportion of wide-scope universal responses. In (6b), however, there should be a conflict between the inverse scope imposed by the construction, and alle's resistance to take wide scope. This should lead to processing difficulty at the point where the conflict becomes apparent (presumably at the second quantifier). Moreover, at least in some cases the conflict should be resolved in favor of the linear scope reading, thus a greater proportion of wide-scope existential responses is expected than in (6a).

2.3 Materials and subjects

72 items were written in six conditions each. Each item was paired with two displays: one used with the quantifier conditions and with Control B, the other used with Control A. The displays prepared for half of the items included pictures (photographs or drawings), the other half consisted of words (e.g. names), letters, or numbers. In addition, 70 fillers were constructed. They included other numerals (two or three) or other quantifiers (both cars, only one animal) instead of exactly one, or other kinds of displays.

Six presentation lists were created according to a Latin square design. Two pseudo-random orders were generated, making sure that adjacent experimental items belonged to different conditions. A filler was inserted between any two items. The same pseudo-random orders were used in all presentation lists. Twelve subjects were tested with the first order and eighteen with the second.
Thirty subjects participated in the experiment for a payment of 10 euros. They were all native German speakers and had normal or corrected vision. Eight additional participants had to be excluded from the analysis due to calibration problems (N=3) or error rates higher than 20% (N=5).

2.4 Apparatus

A tower-mounted Eyelink 1000 eyetracker monitored the gaze location of participant's right eyes. The eyetracker has a spatial resolution of 0.01 degrees of visual angle and samples gaze location every millisecond. Participants viewed the stimuli binocularly on a 19 inch monitor 70 cm from their eyes. A head rest minimized head movements. The experiment was implemented using the SR Research Experiment Builder software and eyetracking data were exported with the SR Research Data Viewer.

2.5 Procedure

Subjects were tested individually. The tracker was calibrated using a 3x3 grid guaranteeing that all fixations were less than 0.5 degrees apart from the calibration stimuli. After calibration was completed, participants read the experimental instructions on the screen. This was followed by a practice session of 10 items. In the experiment, each trial started with a calibration check. The tracker was recalibrated as necessary. Eye-movements were recorded both during reading and while inspecting the displays. The displays were presented in a way that it was impossible for the participants to inspect the whole screen at once. To get information about what’s in a field, they had to fixate on it.
Figure 3 summarizes subsequent steps in a trial. The trial began with the presentation of a screen which served as calibration check with a little black dot in the position where the center of the first word would appear. If no fixation was registered within five seconds, recalibration was enforced. Otherwise a sentence appeared in the center of a white screen. It was printed in black letters with 14 point font size. Four characters corresponded approximately to one degree of visual angle. After reading the sentence participants had to move their eyes to an asterisk in the lower right corner. Fixating the asterisk triggered the presentation of a black screen with a white fixation cross which appeared in central position for 250 ms. This was done to guarantee that subjects were looking at the center of the screen when the display appeared. Then the display was presented and the timer was started. Answers had to be provided orally within four seconds, measured with a voice key. If the participant started answering within the time limit the display remained visible until the end of the trial. If no answer was provided the trial was automatically aborted and a 'too late' message appeared on the screen. Participants started the next trial by pressing a button on a joy-pad.
The answers that the subject provided were recorded by the experimenter on a prepared score sheet. They were categorized as corresponding to the wide-scope existential reading, the wide-scope universal reading, or neither. The experiment lasted 35-60 minutes.

2.6 Results

Comprehension rates were high. On the filler trials, participants provided a correct answer 92.7% (SD 26.0) of the time. All participants scored higher than 80%.

2.6.1 Scope interpretation

The chosen scope interpretation was determined on the basis of the inspection-answer pattern. A trial was categorized as having received a wide-scope existential reading if the participant inspected all three fields and provided a single answer. The reading was coded wide scope universal if the subject responded field-by-field, that is, provided a multiple answer that started before all fields had been inspected. This coding procedure can be applied to scope ambiguous sentences and the unambiguous sentences in Control B as well as to the definite descriptions in Control A. According to these criteria, 18% of all cases couldn't be categorized and were excluded from the analysis.
Figure 4 shows the percentage of wide-scope universal readings in all six conditions. In Control B the reported readings matched the scope of the disambiguated sentences. The $\forall\exists$-condition received 97.2% wide-scope universal interpretations, the $\exists\forall$-condition 2.2% wide-scope universal interpretations.

The scope data of the ambiguous conditions and the definite descriptions in Control A were subjected to 2x2 repeated measures analyses of variance (ANOVAs) with the within factors NP-type (quantifier vs. definite NP) and distributivity (all vs. each) using participants ($F_1$) and items ($F_2$) as random factors.\(^3\)

In the scope ambiguous quantifier sentences, the proportions of wide scope universal readings differed between each and all. The $\forall\exists$ interpretation was more preferred for each (83.4%) than for all (59.6%). One sample t-tests (one tailed) revealed that the latter was significantly above 50% ($t_{1}(29)=1.70; \ p<.05; \ t_{2}(71)=3.45; \ p<.01$). The definite NPs (Control A)
showed a similar contrast: A $\forall \exists$ interpretation was more frequent in constructions with each (24.1%) than with all (11.4%). ANOVAs revealed a significant main effect of NP-type ($F_1(1,29)=87.05; p<.01; F_2(1,71)=682.01; p<.01$) as well as of distributivity ($F_1(1,29)=42.89; p<.01; F_2(1,71)=51.50; p<.01$). The former reflects more inverse scope readings in the ambiguous conditions than in Control A. The latter comes from each taking wide scope more frequently than all. The interaction between NP-type and distributivity was significant by items ($F_1(1,29)=3.71; p=.07; F_2(1,72)=6.49; p<.05$). It is due to a somewhat bigger contrast between all and each in the quantifier conditions than in the definite NP conditions. Nevertheless, the difference between all and each in control A was significant, as a pairwise t-test shows ($t_1(29)=3.79; p<.01; t_2(71)=3.54; p<.01$).

2.6.2 Reading Times

Three regions of interest were defined for the purposes of the analysis. The pre-target region consisted of the first DP (genau ein Tier vs. das Tier). The post-target region contained the last three words of the sentence (sollst du nennen). 98.4% of all regions were fixated during the first pass. Fixations shorter than 80 ms were eliminated from the data set. This excluded less than 0.5% of the data.

Four standard eye-movement measures were computed for all three regions. We will define them as follows: First pass reading time is the sum of all fixations from first entering a region until leaving it. Total reading time is the sum of all fixations on a region. Regressions-out is the proportion of times a regression was launched from a region. Finally, regressions-in is the proportion of times a regression was made back into the region. While first-pass reading time and regression out are taken to reflect immediate processing, total time and regression in are taken to reflect late processes like reanalysis.

Reading time data were subjected to 2x2 repeated measures ANOVAs with the within factors NP-type and distributivity. In addition, we analyzed reading times contingent on the reported reading using linear mixed effects models (for an introduction see Baayen, Davidson, and Bates (2008) and the references therein).
Table 1: Reading time data reporting means (+ SDs)

<table>
<thead>
<tr>
<th>Condition</th>
<th>Measure</th>
<th>Pretarget Region</th>
<th>Target Region</th>
<th>Posttarget Region</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Pretarget</td>
<td>Target</td>
<td>Posttarget</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Region</td>
<td>Region</td>
<td>Region</td>
</tr>
<tr>
<td><strong>Q-All</strong></td>
<td>First-Pass</td>
<td>725 ms (349)</td>
<td>514 ms (262)</td>
<td>476 ms (361)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1194 ms (843)</td>
<td>964 ms (793)</td>
<td>708 ms (671)</td>
</tr>
<tr>
<td></td>
<td>Regression Out</td>
<td>9.2% (15.1)</td>
<td>29.0% (27.4)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Regression In</td>
<td>14.6% (15.0)</td>
<td>9.9% (15.2)</td>
<td></td>
</tr>
<tr>
<td><strong>Q-Each</strong></td>
<td>First-Pass</td>
<td>754 ms (400)</td>
<td>446 ms (202)</td>
<td>461 ms (318)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1142 ms (855)</td>
<td>743 ms (629)</td>
<td>652 ms (526)</td>
</tr>
<tr>
<td></td>
<td>Regression Out</td>
<td>9.6% (16.9)</td>
<td>25.3% (27.1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Regression In</td>
<td>12.1% (14.1)</td>
<td>11.6% (18.1)</td>
<td></td>
</tr>
<tr>
<td><strong>Def-All</strong></td>
<td>First-Pass</td>
<td>524 ms (291)</td>
<td>488 ms (220)</td>
<td>475 ms (322)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>785 ms (503)</td>
<td>809 ms (523)</td>
<td>654 ms (573)</td>
</tr>
<tr>
<td></td>
<td>Regression Out</td>
<td>8.7% (13.2)</td>
<td>27.4% (26.8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Regression In</td>
<td>17.5% (19.7)</td>
<td>12.9% (17.5)</td>
<td></td>
</tr>
<tr>
<td><strong>Def-Each</strong></td>
<td>First-Pass</td>
<td>510 ms (330)</td>
<td>497 ms (215)</td>
<td>495 ms (363)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>811 ms (614)</td>
<td>824 ms (558)</td>
<td>698 ms (533)</td>
</tr>
<tr>
<td></td>
<td>Regression Out</td>
<td>7.5% (12.9)</td>
<td>25.8% (25.5)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Regression In</td>
<td>17.3% (19.2)</td>
<td>11.2% (15.9)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Reading time data reporting means (+ SDs)

Table 1 summarizes the reading time data of the quantifier and definite NP conditions for all three regions. Total and first pass reading times are depicted in Figure 5.

A)

B)

Figure 5: Panel A depicts total reading times for all three regions. Panel B shows first pass reading times. Error bars indicate 95% confidence intervals.
In the pretarget region, the only difference among conditions was that the much shorter definite descriptions were read faster than quantifying expressions. This difference was reflected by a significant main effect of NP-type both in total times ($F_1(1,29)=49.59; p<.01; F_2(1,71)=115.45; p<.01$) and in first pass times ($F_1(1,29)=156.05; p<.01; F_2(1,71)=168.89; p<.01$). The *all* and *each* conditions didn't differ either in total times ($F_{1/2}<1$) or in first pass times ($F_{1/2}<1$). Moreover, the interaction between NP-type and distributivity wasn't reliable either in total times ($F_{1/2}<1.2$) or in first pass times ($F_{1/2}<1.7$). ANOVAs on regressions-in revealed a significant main effect of NP-type ($F_1(1,29)=13.17; p<.01; F_2(1,71)=26.07; p<.01$). This isn't surprising because the much shorter definite NPs were fixated less often than quantifying NPs resulting in an apparently higher proportion of regressions. No other effects reached significance ($F_{1/2}<3.6; F_{1/2}<2.6$).

In the target region, in the quantifier conditions *all* was read slower than *each*. However, the definite NP conditions show the opposite pattern: at least numerically, *all* was faster than *each*. Statistically, this was reflected in a significant main effect both in total times ($F_1(1,29)=8.84; p<.01; F_2(1,71)=8.53; p<.01$) and in first pass times ($F_1(1,29)=5.71; p<.05; F_2(1,71)=6.25; p<.05$) and a significant interaction between NP-type and distributivity both in total times ($F_1(1,29)=19.75; p<.01; F_2(1,71)=12.75; p<.01$) and in first pass times ($F_1(1,29)=14.79; p<.01; F_2(1,71)=12.67; p<.01$). The main effect of NP-type didn't reach significance (total times: $F_{1/2}<2$; first pass times: $F_{1/2}<2$). ANOVAs performed on regressions-in and regressions-out didn't reveal any significant differences.

At the posttarget region, ANOVAs didn't reveal any significant effects (all $F_{1/2}<2.5$). In regressions out, we analyzed the difference between the quantifier conditions with a paired t-test. The *all* condition lead to a higher proportion of regressions out which was marginal by subjects and significant by items ($t_{1}(29)=1.92; p=.065; t_{2}(71)=2.03; p<.05$).

2.6.3 Contingent Reading Times

Total and first pass reading times contingent on the reported reading are depicted in Figure 6.
Figure 6: Panel A depicts total reading times contingent on the scope reading for the scope ambiguous conditions. Panel B shows contingent first pass reading times. Error bars indicate standard errors of the mean.

The graph shows that the quantifier *all* condition was read more slowly than the quantifier *each* condition irrespective of the reading. But there was also a contrast among the two readings of the quantifier *all* condition in the total times. *All* was read more slowly when a linear scope interpretation was computed than under inverse scope. For statistical analyses, we used linear mixed effect models to analyze reading times on the target region. Linear mixed effects models incorporate both fixed and random effects and provide a flexible method to deal with missing data. Participants and items were treated as random effects while quantifier, scope and length of the second quantifier measured in number of characters were treated as fixed effects. We computed the following four models only including reading times from trials for which the reading could be determined according to the criteria mentioned above. Each model was computed separately for total reading times and first pass times.

Model A) Analyzing the complete set of ambiguous conditions:

\[ y_{ij} = \mu + \text{Scope}_{ij} \beta_1 + \text{Quantifier}_{ij} \beta_2 + \text{Length}_{ij} \beta_3 + S \beta_4 + W \beta_5 + \epsilon_{ij} \]

Model B) Analyzing the two scope readings separately in the ambiguous 'all' condition:
$y_{ij} = \mu + \text{Scope}_{ij} \cdot \beta_1 + S_s + W_w + \epsilon_{ij}$

Model C) Comparing the ambiguous 'all' and 'each' condition only under inverse scope:

$y_{ij} = \mu + \text{Quantifier}_{ij} \cdot \beta_2 + \text{Length}_{ij} \cdot \beta_3 + S_s + W_w + \epsilon_{ij}$

Model D) Control comparison of DefQ conditions only analyzing linear scope:

$y_{ij} = \mu + \text{Quantifier}_{ij} \cdot \beta_2 + \text{Length}_{ij} \cdot \beta_3 + S_s + W_w + \epsilon_{ij}$

The logic behind this analysis is similar to the one used in computing ordinary ANOVA and paired t-tests. After computing the global analysis (Model A), we systematically conducted simple comparisons (Models B and C). Finally, we compared the effects of Model C with the definites of Control A in their preferred scope interpretation (Model D).

Consider the global analysis in Model A. The vector $y_{ij}$ represents the reading time of subject $i$ for item $j$ and is modeled by the sum of a constant term $\mu$, the fixed effect of scope modeled by the coefficient $\beta_1$, the fixed effect of quantifier $\beta_2$, the fixed effect of length $\beta_3$ and the random effects of subject, where $s_i$ is a random intercept for each subject and the random effect of item, where $w_j$ is a random intercept for each item. Finally, $\epsilon_{ij}$ is the vector of residual errors and models the per-observation noise. When a mixed-effects model is fitted to the data, its set of estimated parameters includes the coefficients for the fixed effects. For these, we computed t-statistics and estimated p-values based on Markov chain Monte Carlo sampling from the posterior distribution of the parameters (running 10000 simulations for each model) as suggested by Baayen et al. (2008). Linear mixed effects models were computed using the lme4 library and the MCMC package of R.

<table>
<thead>
<tr>
<th>Model</th>
<th>Fixed Effect</th>
<th>Est. Coefficient</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$ (total)</td>
<td>Scope</td>
<td>243.0 ms</td>
<td>3.31</td>
<td>0.001**</td>
</tr>
<tr>
<td></td>
<td>Quantifier</td>
<td>161.9 ms</td>
<td>2.66</td>
<td>0.008**</td>
</tr>
<tr>
<td></td>
<td>Length</td>
<td>12.1 ms</td>
<td>0.95</td>
<td>0.343</td>
</tr>
<tr>
<td>$A$ (first pass)</td>
<td>Scope</td>
<td>40.8 ms</td>
<td>1.74</td>
<td>0.082</td>
</tr>
<tr>
<td></td>
<td>Quantifier</td>
<td>57.3 ms</td>
<td>2.97</td>
<td>0.003**</td>
</tr>
<tr>
<td></td>
<td>Length</td>
<td>4.4 ms</td>
<td>1.09</td>
<td>0.276</td>
</tr>
</tbody>
</table>
In our analysis we started with the global Model A and then analyzed the contrasts in Models B and C using only the relevant "conditions", that is subsets of the data used in model A. The results are summarized in Table 2. Applying Model A to total reading times, the fixed effects of both scope and quantifier turned out to be significant. Processing all was more difficult than processing each and computing a linear reading was more difficult than computing an inverse reading. There was no significant effect of length. In Model A using first pass times as the dependent variable, the fixed effect of quantifier was significant but there was no effect of either scope or length. In total times, both of the contrasts investigated with Models B and C turned out to be significant. In Model B, the significant effect of scope shows that linear scope was processed more slowly than inverse scope. In Model C, which keeps the scope reading constant (inverse scope), all was processed more slowly than each. In first pass times only the contrast in model C was significant.

To examine the lexical differences between all and each we also analyzed reading times of Control A keeping the reading constant. In these conditions, participants overwhelmingly chose linear scope, so we compared all vs. each with linear scope. Descriptive statistics revealed that the definite all condition was faster in total reading times than the definite each condition (852 vs. 877 ms), which is the opposite of the contrast found in the quantifier conditions. In first pass reading times the definite all condition (496 ms) was numerically also faster than the each condition (511 ms). The fixed effect of quantifier was not significant in Model D either in the total times or in the first pass analysis. Only length turned out to be a reliable predictor.

<table>
<thead>
<tr>
<th></th>
<th>Scope</th>
<th>Length</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>234.6 ms</td>
<td>2.42</td>
<td>0.016*</td>
<td></td>
</tr>
<tr>
<td>B (first pass)</td>
<td>43.4 ms</td>
<td>1.35</td>
<td>0.178</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>170.4 ms</td>
<td>2.44</td>
<td>0.015*</td>
<td></td>
</tr>
<tr>
<td>C (first pass)</td>
<td>16.6 ms</td>
<td>1.07</td>
<td>0.283</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>63.52 ms</td>
<td>2.78</td>
<td>0.006**</td>
<td></td>
</tr>
<tr>
<td>D (first pass)</td>
<td>5.40 ms</td>
<td>1.08</td>
<td>0.281</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Results of linear fixed effects models modelling contingent reading times
2.7 Discussion

As expected, there was an overwhelming preference for inverse scope in the quantifier conditions. This preference was modulated by the type of quantifier: the condition with *all* resulted in inverse scope readings significantly less often than quantifier-*each*. This is consistent with reports in the literature of *all* having less of a tendency to take wide scope than *each*. The novelty of our results lies in demonstrating the effect of quantifier type in inverse linking constructions as well. Furthermore, our results make it apparent that the effect of the construction is stronger than scope preferences of the quantifiers: reading times on ambiguous quantifier-*all* sentences that received a linear scope interpretation were still slower than in those trials where the ambiguous quantifier-*all* sentences were interpreted with inverse scope, as the results of the analyses in model B show.

Moreover, the reading time results indicate that scope relations were computed immediately: the differences were already present in the first pass reading times, and they did not carry over to the next region. As there was no pressure in the experiment to disambiguate the sentences during reading, this finding suggests that quantifier scope is immediately and fully specified during comprehension. However, a word of caution is in order: Although the sentences themselves required no disambiguation, the task of naming an object or objects did. This task was fully predictable and had to be completed within a limited amount of time. It is thus possible that participants performed some disambiguation during reading that would have been delayed otherwise. Another potential cause for worry could be the fact that the instructions had a highly predictable structure, and in particular, that the final segment was always identical (*sollst du nennen/should you name*). This may have led to subjects effectively skipping this segment and performing end-of-sentence interpretation as soon as they had read the second quantifier. We do not consider this a serious objection since the effect was already present in first-pass reading times. Furthermore, the differences in the regressions-out measure between *all* and *each* in the final region strongly indicate that processing wasn't completed as soon as the second quantifier had been encountered. Thus we contend that the consistent early disambiguation of quantifier scope in our
experiment was not an artefact, but reflects immediate interpretation of quantifier constructions during normal processing.

Our results suggest the following picture of quantifier scope interpretation in inverse linking: The first quantifier is interpreted immediately. When the second quantifier is encountered, it is given wide scope, determined by the properties of the construction. In the case of jeder, this interpretation fits well with the quantifier's need for wide scope, thus the resulting reading is overwhelmingly $\forall \exists$. When the universal quantifier is alle, however, then there is a conflict between the construction and the scope preferences of the quantifier. As the construction is the stronger factor of the two, it is still easier to settle on an inverse-scope reading for the sentence. This is reflected both in faster reading times for alle with inverse scope, and a higher percent of $\forall \exists$ interpretations in this condition. Still, alle's reluctance to take wide scope is strong enough to result in an $\exists \forall$ interpretation roughly 40% of the time.

This view of scope interpretation raises the question what happens when the first NP is a definite NP rather than a quantifier. The analysis in May and Bale (2005) would suggest, for instance, that the quantifier in the PP would still resist an in-situ interpretation and would try to obtain wider scope. Is there evidence for a wide-scope interpretation of the quantifier inside the PP in these cases? We will now take a closer look at the conditions in Control A to answer this question.

3. Definite NPs

The definite NPs were originally included in the study to serve as controls in interpreting the reading time results. However, as briefly mentioned before, they were also interesting in their own right: a comparison of scope preferences in the two conditions in Control A makes it possible to address the question whether definite NPs should be considered quantificational or not. The lack of reading time differences in the definite NP conditions (section 2.6, especially Model D) suggest that there is no scope interaction between the definite NP and the universal quantifier. This is consistent with analysing definite NPs as effectively scopeless (e.g. Glanzberg forthcoming). Alternatively, the definite NP could receive widest scope the same way topic elements are claimed to have scope over the rest of the sentence (e.g. Beghelli and Stowell 199, Endriss 2002). Either of these possibilities is compatible with the embedded quantifier getting wider
scope than its surface position would allow, since it still remains within the scope of the definite NP.

The interpretation of the reported readings reveals another picture. The definite descriptions differed in their scopal behaviour when interacting with alle as compared to jeder. First let us consider what it means to distinguish two readings in the conditions with definite NPs. The reading where the universal quantifier has narrow scope with respect to the definite NP can be paraphrased as "Name an animal that is in each/all field(s)". The wide scope universal reading would be "From each field/From all fields, name an animal". Both of these readings are compatible with the displays we used with Control A, which included only one object corresponding to the definite NP in the sentence, and that object was present in all fields. What distinguishes the two answers is thus not the NP named, but the eye-movement behavior of the participants: The linear scope interpretation requires the inspection of all fields first to make sure that the NP corresponding to the sentence is present in all of them. In case of a wide-scope universal interpretation, however, participants can start answering before all fields have been inspected. In addition to the eye-movement patterns, the latter interpretation was sometimes also suggested by answers naming the same object three times ("deer, deer, deer").

Although there was a very strong overall preference for linear scope in the NP-conditions, each resulted in more wide-scope universal interpretations than all. The difference was somewhat smaller than in the quantifier conditions; nevertheless, it was fully significant. Thus it appears that definite NPs do have scope properties: although they have very wide scope it is still possible to outscope them. This is compatible with a view where they take scope from a position that is higher than the one where the universal quantifier is interpreted in the ambiguous quantifier conditions. Consequently the "ordinary" wide scope interpretation of the universal quantifier still results in narrow scope with respect to the definite NP. However, there must be a second position available for universal quantifiers as well, and it must be higher than the position where the definite NP is interpreted. The low percentage of wide-scope universal interpretations in the definite NP conditions suggest that this position can only be filled under special circumstances.

If obtaining a wide-scope universal interpretation over a definite NP is so difficult, why didn't this difficulty show up in the reading time data? We
can only speculate here. Perhaps with definite NPs the interpretation of the universal quantifier requires two steps to result in an inverse-scope reading. The immediate interpretation of the universal quantifier would then involve assigning it the same scope it has in the ambiguous quantifier conditions as well. This is sufficient to satisfy the scope requirements of the construction. A second step would be needed for the universal quantifier to outscope the definite NP. This step may be related to the need for special "support" for this interpretation, and thus may come relatively late in the interpretation of the quantifier, possibly after the whole sentence has been read.

What is left to explain now is the lack of a reading time difference between *each* and *all* during the first step of raising the universal quantifier to the (lower) scope position. Recall that in the corresponding ambiguous quantifier conditions *each* was read significantly faster than *all*. We attributed this effect to *all*’s resistance to move to the scope position. If that were the full story, however, then there should be a similar effect in the definite NP conditions as well. We want to propose that the lack of a difference is related to the way (existential) quantifiers (such as *genau ein*) versus definite NPs are interpreted. Suppose the structure of the whole preverbal NP in our sentences is the following:

(7)

```
(scope)    NP
  /\      /
 Q1 PP  Q2
 P
```

We assume that *genau ein* is interpreted in situ, that is, in Q1. However, the universal quantifier (Q2) cannot remain inside the PP and should raise to the position marked as 'scope'. This is relatively easy for *jeder*, leading to short reading times and a large percentage of inverse scope readings in the quantifier-*jeder* condition. *Alle* cannot get past *genau ein* quite so easily, so inverse scope readings are less frequent; when it does reach the scope position, though, interpretation proceeds relatively smoothly. The reading time difference between *each* and *all* comes from those cases where *alle* does not manage to reach the scope position, and has to find some other
position below $Q_1$ where it can be interpreted. As $Q_2$ by assumption does not allow scope interpretation, the position where *alle* ends up in these cases must be just above the PP node.

When $Q_1$ is occupied by a definite NP the situation is different: the definite NP needs very wide scope so it must move to some position higher than the scope position in (7). As $Q_1$ is empty (it only contains a trace), nothing blocks movement of the universal quantifier to the scope position. The resulting interpretation is still one with linear scope, though, since this scope position is lower than the one the definite NP occupies.

Under this modified account, *alle* does not resist wide scope; in fact, it moves to a scope position whenever possible. The apparent preference for narrow scope comes from *alle* being relatively "weak": the scope interpretation it receives is largely determined by other scope-bearing elements within the same sentence. *Jeder* is "stronger", so it can assert its scope needs more easily. Relative scope in a construction is thus the result of the interplay of a number of different factors (Pafel 2005). For instance it should be possible to find some quantifier $Q_1$ that is interpreted in situ the same way we assume for *genau ein*, but that allows movement of *alle* to the scope position more easily. Testing such predictions is beyond the scope of the present paper.

4. Conclusions

We have proposed a method of combining in a single experiment online measures of difficulty during interpretation with offline indications of the resulting interpretation. Individual components of the method have already been used successfully to investigate a number of phenomena in psycholinguistic experimentation. What is novel here is the use of the visual world paradigm to investigate quantifier scope, and its combination with more traditional reading time measures. As the results show, the eye movement data together with the answers participants provide are sufficient to determine the final reading of a scope-ambiguous quantified sentence. This means that no overt disambiguation is necessary to investigate the unfolding interpretation of an ambiguous sentence. The new method thus makes it possible to avoid the unwanted consequences of overt disambiguation of quantifier scope (see section 1) and study quantifier interpretation in a paradigm that is maximally similar to normal interpretation.
References

Baayen, R.H., Davidson, D.J. and Bates, D.M.

Filippo Beghelli and Tim Stowell

Oliver Bott & Janina Radó

Cornelia Endriss

Ruth Filik & Kevin B. Paterson & Simon P. Liversedge

Michael Glanzberg

Irene Heim

Howard S. Kurtzman & Maryellen C. MacDonald

Robert May & Alan Bale
Stephen Neale  

J. Pafel  
2005  *Quantifier scope in German. An investigation into the relation between syntax and semantics*. Amsterdam: Benjamins.

Bertrand Russell  

Peter F. Strawson  

Michael K. Tanenhaus & Michael J. Spivey-Knowlton & Kathleen M. Eberhard & Julie C. Sedivy  
1995  Integration of visual and linguistic information in spoken language comprehension. *Science* 268 (5217): 1632-34.

Susanne L. Tunstall  

---

1 We think that *jeder* is more strongly distributive than English *every*. This is intended to be reflected in translating it as *each*.

2 This becomes clear when a sentence with *exactly one*, for instance, *you may choose exactly one item* is uttered in a situation only containing a single item. In such a situation, the utterance seems to be infelicitous to us.

3 Control B was not included in the analyses of variance, as the impact of the near-perfect answers in the scope-disambiguated conditions was likely to distort the pattern of data in the potentially ambiguous conditions.

4 The *each* condition under inverse scope was left out of the analysis, since there were only very few data points in this condition (N=56).

5 The *all* conditions were approximately 1.6 characters longer than the *each* conditions.