Underspecified representations of scope ambiguity?

Janina Radó & Oliver Bott

SFB 833, University of Tübingen

1 Introduction

Underspecified representations are commonly used to model semantic ambiguity, and scope ambiguity in particular. They offer an elegant way to capture the interpretation possibilities in a single compact representation, thus avoiding a combinatorial explosion of readings as the number of operators increases (see [3] for a critical discussion). This property has repeatedly been claimed to make them cognitively plausible: since only one representation is constructed, scope-ambiguous sentences can be interpreted in a fast and efficient way.

There is indeed evidence that perceivers only specify the interpretation of the sentence as much as necessary for a particular task (e.g. pronouns are not automatically resolved, cf. [6], [11]). A similar claim has been made with respect to quantifier interpretation: scope is only resolved when necessary, e.g. when disambiguation is encountered, and may remain completely underspecified in some cases (see [11] for an example). Sanford and Sturt use the lack of scope preferences in some constructions to argue for underspecified scope representations.

It is important to note that the connection between underspecification and shallow processing is only indirect. Comprehenders often do not compute a representation that is detailed, complete, and accurate with respect to the input, but only one that is 'good enough' (see [4]). They process only as deeply as necessary under the circumstances and may end up with only a partial representation if that is sufficient for the task at hand rather than constructing a single connected representation (which may or may not be fully specified) for the complete sentence. Underspecified representations, on the other hand, are complete compact representations encoding all interpretation possibilities. They thus require deep processing but may not result in a specified representation (e.g. of scope), if there is no disambiguating information.

Intuitively, some scope readings are easier to get than others. This is also supported by psycholinguistic findings; for instance, [8] report a preference for the wide–scope interpretation of the second quantifier in the inverse linking construction in (1):

(1) George has a photograph of every admiral.

Such graded preferences have been captured in underspecification accounts by postulating weighted dominance constraints (e.g. [7]), which make some representations cheaper and easier to derive than others. The question now is whether

the different readings are computed (and ranked) automatically, or only when the task e.g., a decision task requires it, as underspecification theory would have it. As it turns out, existing evidence does not allow us to answer this question (cf. [2]). For instance, [5] examined eye movements while perceivers read sentences like (2) (see also [10]) in order to find evidence for early disambiguation of quantifier scope.

(2) a. The celebrity gave a reporter from the newspaper every in-depth interview, but the reporter(s) was/were not very interested.
b. The celebrity gave every reporter from the newspaper an in-depth interview, but the interview(s) was/were not very interesting.

They report longer *total* reading times at the second quantifier in (2-a) than in (2-b), and interpret it as evidence for a scope conflict arising at the second quantifier in (2-a): linear order and grammatical role favor the wide-scope existential reading, whereas inherent properties of the quantifier support the wide-scope universal interpretation. In (2-b) all factors point to the wide-scope existential reading, thus processing is easier. This is reflected in off-line ratings as well: the $\exists\forall$ reading is accepted less in (2-a) than in (2-b). Crucially, however, both the reading time effects and the scope preferences are late effects; the scope conflict in (2-a) does not have any measurable influence on first-pass reading times. The results thus do not exclude the possibility that perceivers only computed a fully specified scope interpretation when they read the disambiguating information in the second clause.

To test the cognitive plausibility of scope underspecification we need to use a task that guarantees deep processing and look for evidence for scope interaction before disambiguating information is present. We designed a combined self-paced reading/verification experiment that we think fits the bill. Participants first read doubly quantified sentences and then had to decide whether the sentence matched a scope disambiguating card display. The sentences in the reading task exhibited a scope conflict between the inherent properties of the quantifiers and the construction they occurred in. Crucially, disambiguation was not available during the reading phase, so any reading time effects could be attributed to on-line resolution of scope relations. Moreover, participants were instructed to be accommodating and accept the picture if the corresponding reading was available at all. This way we hoped to encourage them to maintain an underspecified interpretation as long as possible.

2 Methods

Materials: We tested German inverse linking constructions and manipulated the quantifier in the embedded position. One type of complex quantifiers was built of the determiners *genau ein* (*exactly one*) and *alle* (*all*) as illustrated in the sample item in (3-a), whereas the other type embedded distributive *jeder* (*each*) as in (3-b). The asterisks indicate segmentation in self-paced reading.

(3) a. Genau* ein Tier* auf* allen* Karten* ist* ein Affe.
Exactly one animal on all cards is a monkey.
b. Genau* ein Tier* auf* jeder* Karte* ist* ein Affe.
Exactly one animal on each card is a monkey.

Condition (3-a) exemplifies a case of scope conflict between quantifiers in the same DP: the inverse linking construction strongly biases towards inverse scope, but since the non-distributive *alle* (*all*) prefers not to outscope the first quantifier we expected competition between readings in the *all* conditions. This is different in (3-b) with *jeder* (*each*). Here, both factors point in the same direction therefore *each* should be interpreted with wide scope without any difficulty. As for the final interpretation, scope conflict should lead to fewer inverse interpretations in the *all* than in the *each* conditions.

In addition, we also tested doubly quantified sentences in which the quantifiers appeared in a non-embedded configuration.

(4) a. Genau* ein Affe* ist* auf* allen* Karten* zu* finden.
Exactly one monkey is on all cards to find.
b. Genau* ein Affe* ist* auf* jeder* Karte* zu* finden.
Exactly one monkey is on each card to find.

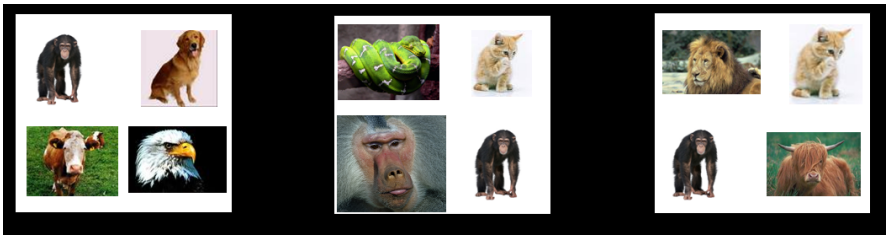
Intuitively, the conditions (4) show the same contrast as (3), thus we expect the same pattern of interpretations in both construction types. In addition, comparing the DP- and the sentence conditions we can test when scope computation takes place if the initial interpretation is not underspecified. We will elaborate this point in the Predictions.

Each of the doubly quantified conditions (3) and (4) was paired with two disambiguating card displays yielding eight doubly quantified conditions in a 2 x 2 x 2 (*construction* x *quantifier* x *picture*) factorial design. The linear ∃!∀ card displays had the same object on all three cards, but the second card contained an additional object of the same category. Figure 1a is an example. The inverse ∀∃! card displays had exactly one object of the relevant kind on each card, but different ones (cf. Figure 1b). Again, the second card provided the disambiguating information.

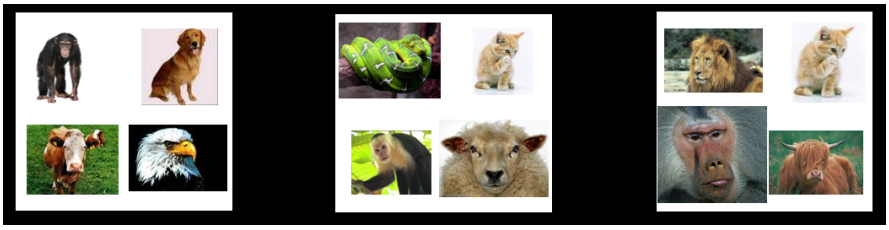
To control for potential differences in lexical processing between *all* and *each* we included the controls in (5). The controls were always paired with a card display that had the same object on all three cards (cf. Figure 1c).

(5) a. Auf* allen* Karten* ist* ein Schimpanse.
On all cards is a chimpanzee.
b. Auf* jeder* Karte* ist* ein Schimpanse.
On each card is a chimpanzee.

In total, this yielded ten conditions and we constructed 60 experimental items (ie. sentence/picture sets) in each of them. Additionally, we created 80 fillers which served several purposes. First, we made sure that all sorts of quantifiers would appear in the experiment. Second, the fillers varied in structure and in



(a) ∃!∀ card display



(b) ∀∃! card display

(c) control conditions

Fig. 1. Fully uncovered card displays for the sample item. Figure 1a is only compatible with wide scope of *exactly one* (∃!). Figure 1b disambiguates towards wide scope of the universal quantifier. Figure 1c is the card display presented with the control conditions. In the experiment, card displays were uncovered card by card from left to right.

a number of cases presented crucial information at the sentence final segment (eg. *A drill can be found on all cards in their upper half*). Third, a number of fillers had pictures that required accommodating interpretations, ie. a total of three objects for *einige* (*a few*). Finally, 40 of the fillers were clearly false to ensure that on a reasonable proportion of trials participants had to reject the card display. We took the experimental items and the fillers and constructed 10 lists according to a latin square.

Predictions: Assuming underspecification with weighted constraints, the quantifier manipulation should have no effect on the reading times. We do expect differences at the disambiguating card (card 2), however, since disambiguation should force scope resolution.

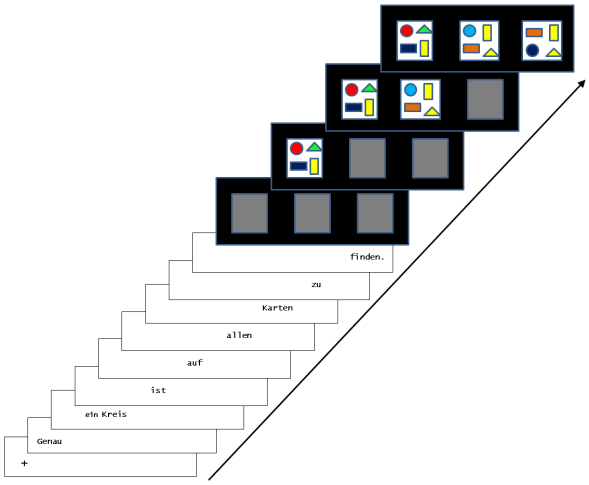


Fig. 2. Sample trial (‘*Exactly one circle can be found on all cards*’)

By contrast if scope is interpreted automatically, i.e. even without disambiguation, then the scope conflict should lead to processing difficulty already before the second card. In fact, we may find processing effects at the earliest point where scope computation may take place, i.e. at the second quantifier. We will dub this *immediate full interpretation*.

Finally, it is conceivable that semantic interpretation lags behind syntactic processing. At least some semantic interpretation processes require domains larger than individual words [1]. The same may hold for quantifier scope as well, especially as thematic roles seem to be one factor in determining the scope needs of a quantifier [9]. Then full scope interpretation would only be expected to take place once the verb and all of its arguments have been received. We will call this the *minimal domain* account. Under this view scope interpretation should be possible at the second quantifier in the sentence conditions in (4), but scope would remain underspecified until the lexical verb in the DP conditions in (3). Note that this still assumes the automatic computation of scope relations as soon as all relevant information is available.

To sum up, all three approaches predict a quantifier effect at the disambiguating card. In addition, under *immediate full interpretation* and the *minimal domain* approach we also expect reading time differences either at the second quantifier, or at least when the verb and all its arguments have been received. Thus the pattern of data in the inverse linking conditions will be crucial to distinguish between these two accounts.

Procedure and Participants: Participants' task was to first read a sentence phrase by phrase employing non-cumulative self-paced reading with a moving window display. Segments outside the presentation window were covered by replacing all characters including punctuation marks with empty spaces. This was done to ensure that participants would not know how much material was yet to come. After reading the final segment the sentence disappeared from the screen and a layout of cards was presented. Participants had to uncover the cards one by one and for each card decide whether the sentence is already true or false or whether more information is needed and in that case move on to the next card. A sample trial is illustrated in Figure 2.

An experimental session started with written instructions followed by a practice session of 10 trials. In the instructions, we emphasized that whenever a picture was compatible with a sentence (even in case the interpretation is hard to get) participants should judge "yes, true". Then followed the experiment in a single block with an individually randomized presentation order for each participant. An experimental session took approximately 30 minutes. 40 participants (mean age 25.9; 31 female) from Tübingen University were paid 5 Euro for their participation.

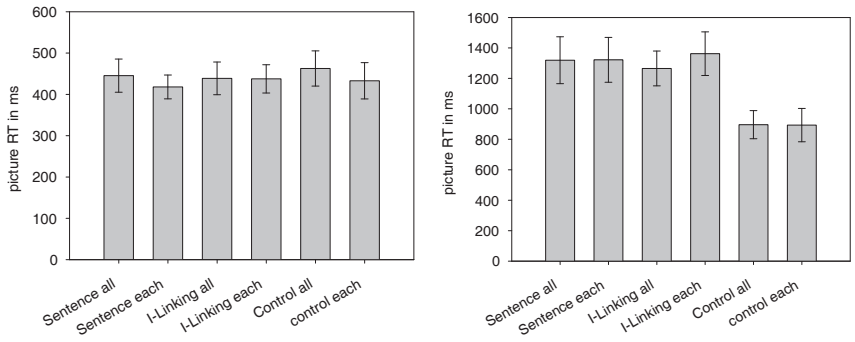
3 Results and Discussion

Judgments: Table 1 presents the mean judgments in the ten conditions. Across the board, the linear reading was only accepted 17.2% of the time whereas the inverse conditions were on average accepted 85.1% of the time. However, acceptance in the linear conditions was 9.8% higher than in the false fillers which were only accepted 7.4% of the time indicating that linear scope, even though highly dispreferred, was still possible. Besides this very strong general preference for inverse scope, the distribution of readings showed a clear contrast between *all* and *each*. The inverse conditions were accepted 77.5% in the *all* sentence condition, 90.4% in the *each* sentence condition, 81.2% in the *all* DP condition and 91.3% in the *each* DP condition. We analyzed the judgments in the inverse linking and sentence constructions in a logit mixed effects model analysis including the fixed effects of *construction*, *quantifier* and *reading* and the random intercepts of participants and items. The analysis revealed significant fixed effects of *reading* ($z = -13.48$; $p < .01$) and of *quantifier* ($z = 3.15$; $p < .01$). Besides a marginally reliable interaction between *quantifier* and *reading* ($z = -1.67$; $p = .09$) no other effects reached significance (all $p \geq .29$). To further break down this interaction we computed separate logit mixed effects model analyses for the linear and the inverse scope conditions. While the linear scope conditions didn't differ significantly from each other (all $p \geq .23$), the analysis of the inverse conditions revealed a significant effect of *quantifier* ($z = 3.44$; $p < .01$). Taken together, the judgments provide evidence for a clear *all*/*each* contrast. Whereas *each* very strongly biases towards inverse scope, the bias is weaker with *all*. On the other hand, the complete lack of effects involving the factor *construction* suggests that the two constructions had comparable scope distributions.

	Universal Quantifier	
	all	*each*
Inverse Linking ∃!∀	15.4 (2.3)	19.2 (2.5)
Inverse Linking ∀∃!	81.2 (2.5)	91.3 (1.8)
Sentence ∃!∀	16.3 (2.4)	17.5 (2.5)
Sentence ∀∃!	77.5 (2.7)	90.4 (1.9)
Control	95.8 (1.3)	97.1 (1.1)

Table 1. Mean proportions of "yes, true" judgments in percent (+ SE of the mean).

Verification Stage: Besides the judgments we analyzed two further dependent measures in the verification stage: the decision point, i.e. the particular card at which participants made their decision and the RTs of "yes, go on" button presses. The analysis of decision points revealed that participants decided at the earliest possible point. Most of the rejections in the linear card displays were launched directly from the second, disambiguating card without ever inspecting the last card. In the linear sentence *all* condition 130 out of a total of 201 rejections (64.7%) came from the disambiguating card, the corresponding *each* condition had 116 out of 198 rejections (58.6%), the inverse linking *all* condition had 121 out of 203 (59.6%) and the inverse linking *each* condition 129 out of 194 (66.5%). In the inverse card displays the situation was different. Here in 92.6% of the trials participants uncovered all three cards before they provided their judgment which in the majority of all cases was "yes, true". Thus, after processing the second card a verification strategy was clearly in place.



(a) fully covered cards (b) first card

Fig. 3. Mean RTs of "yes, go on" button presses in ms (+ 95% confidence intervals; computed by participants) in the six sentence conditions. Figure 3a shows the RTs of the fully covered card layout. Figure 3b shows the RTs of the first card.

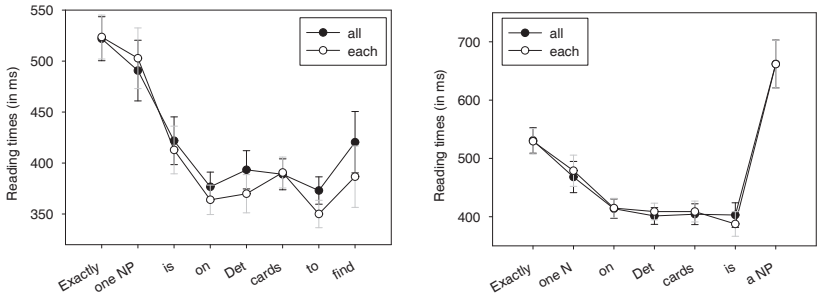
The analysis of RTs of "yes, go on" button presses provides us with a measure of how difficult a particular card was to evaluate. We corrected RTs for outliers by excluding values above 5s. RTs up to the second card were analyzed with mixed effects models including the fixed effects of *construction* (three levels: *sentence* vs. *inverse linking* vs. *control*) and *determiner* (two levels: *all* vs. *each*) and the random intercepts of participants and items. The picture RTs of the fully covered card display and the first card are depicted in Figure 3[1]. While there were no reliable differences between the conditions at the fully covered card display (LME: all $t < 1$), the picture RTs of the (identical) first card clearly differed between conditions. Participants spent a mean RT of 1305ms on the first card in the doubly quantified conditions compared to only 890ms in the control conditions. Whereas the former conditions required evaluation of *exactly one*, i.e. categorization of all four objects and counting, the latter only required detection of the object in question. This clear-cut difference led to a significant contrast between the controls and the doubly quantified constructions (LME: $estimate = -387.00$, $t = -7.70$, $p < .01$). This was the only effect that turned out to be reliable. The contrast between the control condition and the doubly quantified conditions indicates that participants had already chosen the correct verification strategy before they uncovered the first card. The results of the verification stage are fully compatible with *immediate full interpretation* and the *minimal domain* account, but are unexpected in underspecification accounts where comprehenders delay interpretation until they encounter a disambiguation.

Reading Stage: We analyzed the reading times in the following six conditions: inverse linking constructions with *each* vs. *all*, sentence constructions with *each* vs. *all* and control conditions with *each* vs. *all*. To correct for outliers, we trimmed reading times by replacing values below 200ms by a value of 200ms and values above 2000ms by a value of 2000. This correction affected 3.65% of the data.

Figure 4 shows the mean reading times in the three construction types. In the sentence conditions there was a clear contrast between *all* and *each*: the determiner *all* was read on average 23.5ms more slowly than *each*. In contrast, neither the inverse linking nor the control conditions revealed similar differences between determiners (mean difference *jeder* minus *alle* in *inverse linking*: -7.6ms; *control*: 0.7ms). We analyzed the RTs of the determiner *all* vs. *each* in the inverse linking and the sentence constructions in repeated measures ANOVAs. The observed differences led to a significant main effect of *construction* ($F_1(1, 39) = 10.25$, $p < .01$; $F_2(1, 59) = 8.14$, $p < .01$) and a significant interaction between *construction* and *determiner* ($F_1(1, 39) = 6.17$, $p < .05$; $F_2(1, 59) = 6.27$, $p < .05$). To further break down the interaction we computed pairwise comparisons in each construction type. In the sentence conditions *all* took reliably longer to read than *each* ($t_1(39) = 2.45$, $p < .05$; $t_2(59) = 2.93$, $p < .01$). Except for the

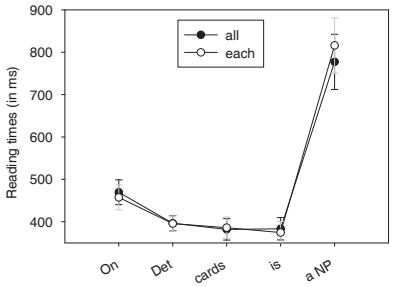
[1] Note that RTs of the second and third cards could not be properly analyzed because of almost no data in the linear conditions and systematic differences between cards in the different conditions.

highly predictable following segment *cards* ($t_{1/2} < 1$) this effect persisted until the end of the sentence (*to*: $t_1(39) = 3.36$, $p < .01$; $t_2(59) = 2.81$, $p < .01$; *find*: $t_1(39) = 2.20$, $p < .05$; $t_2(59) = 2.22$, $p < .05$). This was different in the inverse linking conditions where *all* and *each* didn't differ at any region from the second quantifier until the end of the sentence (all $t_{1/2} < 1.5$) [2]. The same held for the control conditions where *all* and *each* didn't differ anywhere in the sentence (all $t_{1/2} < 1.2$). The lack of difference between *all* and *each* in the control conditions shows that the difficulty in the sentence conditions cannot be attributed to lexical factors. Instead, it must be due to interference between quantifiers.



(a) sentence constructions

(b) inverse linking constructions



(c) control conditions

Fig. 4. Mean reading times in ms (+ 95% confidence intervals; computed by participants) as a function of determiner.

[2] The *minimal domain* account would predict scope conflict at the point when the predication is complete. Finding no indication of scope conflict at the end of the sentence is therefore somewhat unexpected under this account. We suspect, however, that a delayed effect may have been covered by sentence wrap-up.

4 Conclusions

To sum up, the experiment provides evidence for quantifier interpretation before disambiguation is encountered. Even though the participants were instructed to delay scope interpretation as long as possible, they computed scope relations already during reading. Our findings are thus clearly inconsistent with semantic underspecification accounts. The observed scope conflict might indicate that readers selected one interpretation right away. However, it is also possible that they computed both interpretations and ranked them according to the scope factors we have discussed. The results do not allow us to decide between these alternatives. We have shown, however, that a complete minimal domain is required for scope resolution. Our data thus call for introducing the notion of processing domain into cognitively realistic models of semantic interpretation. The results suggest that interpretation processes within a processing domain differ from those that take place at the domain boundary, and that the former may not be fully specified yet.

References

1. O. Bott (to appear). The processing domain of aspectual interpretation. In B. Arsenijevic, B. Gehrke & R. Marín (eds.): *Subatomic semantics of event predicates*. Springer.
2. O. Bott, S. Featherston, J. Radó & B. Stolterfoht (2011). The application of experimental methods in semantics. In C. Maienborn, K. von Heusinger & P. Portner (eds.): *Semantics: An international handbook*. De Gruyter.
3. C. Ebert (2005). Formal investigations of underspecified representations. PhD thesis. King's College: London.
4. F. Ferreira and N.D. Patson (2007). The 'Good Enough' approach to language comprehension. *Language and Linguistics Compass*, 1 (1-2).
5. R. Filik, K.B. Paterson & S.P. Liversedge (2004). Processing doubly quantified sentences: Evidence from eye movements. *Psychonomic Bulletin & Review*. 11 (5).
6. S.B. Greene, G. McKoon & R. Ratcliff (1992). Pronoun resolution and discourse models. *JEP: LMC*, 18.
7. A. Koller, M. Regneri & S. Thater (2008). Regular tree grammars as formalism for scope underspecification. Proceedings of ACL-08.
8. H.S. Kurtzman & M.C. MacDonald (1993). Resolution of quantifier scope ambiguities. *Cognition*, 48.
9. J. Pafel (2005). *Quantifier scope in German*. John Benjamins.
10. K.B. Paterson, R. Filik & S.P. Liversedge (2008). Competition during the processing of quantifier scope ambiguities: Evidence from eye movements during reading. *QJEP*, 61(3).
11. A.J. Sanford & P. Sturt (2002). Depth of processing in language comprehension: Not noticing the evidence. *TCS*, 6 (9).