Incremental Truth Value Judgments

Oliver Bott & Fabian Schlotterbeck

Abstract. We propose a new method for experimental research in semantics and pragmatics. This method, the incremental truth value judgment task, allows us to test whether a particular meaning can be accessed during online comprehension. We report a case study on quantifier scope ambiguity in German subject-before-object and object-before-subject sentences. Experiment 1 provides offline evidence that inverse scope is generally dispreferred but possible in both construction types. In Experiment 2 we used the new method to show that during online processing inverse scope is only available in object-before-subject sentences. Experiment 3 demonstrates that the method is particularly sensitive to typicality biases. If we pay attention to this limitation, it yields perfectly unbiased results.

1. Introduction

Semantic and pragmatic theories are concerned with the possible meanings of a given sentence: they predict which readings are available and whether one is preferred to another. To test a semantic theory it is thus important to collect valid empirical data concerning the status of the proposed interpretations. From a psycholinguistic perspective a question of particular interest is how the final interpretation comes about. During processing there may be intermediate steps where the interpretation is rather different from the final one, as for instance in sentences with a semantic garden-path. To study semantic processing we would therefore like to probe all potential interpretations of the ambiguous input while it is being processed. In the present paper we propose a new method which allows us to track interpretation preferences as the sentence unfolds.

With already existing online methods it is often not easy to tell when a particular reading, especially a dispreferred one, becomes available. Consider a hypothetical experiment intended to establish that one of two potential readings is initially preferred, but that the processor can readily shift to the other interpretation. We could, for instance, follow the common practice in research on garden-path sentences and measure reading times of a disambiguation towards the dispreferred reading. We would then expect to find processing difficulty when the disambiguation is encountered. In addition, in the final interpretation the sentence should be judged fully accepta-
ble. This could be tested in a reading time study with end-of-sentence acceptability judgments. At first sight, difficulty at the disambiguating region as well as the ultimate availability of the dispreferred reading would look like the desired result. At closer inspection, however, it cannot tell us which of the following alternatives is correct. The dispreferred reading could be immediately computed at the disambiguation. Alternatively, it could be only arrived at at a later point, that is after having read the sentence during the judgment phase. When the dispreferred interpretation arises thus remains unclear. In general, we may distinguish three stages at which a reading may be generated: (i) during the processing of the ambiguous region eg. as one of several differently ranked alternatives, (ii) at the disambiguation, eg. during reanalysis or (iii) offline or post-interpretively, when providing a judgment. Following Caplan and Waters (1999) we would like to make a distinction between the extraction of meaning from a linguistic signal, which they called “interpretive processing”, and in the use of that meaning to accomplish other tasks, such as reasoning, planning actions, and other functions, which they called “post-interpretive processing”. These two types of processing are closely related to Fodor’s (1983) distinction between modular and central systems.

The new method proposed here enables us to decide whether an interpretation becomes available already during reading or is constructed only post-interpretively. It combines two well-established experimental paradigms: the truth value judgment task and incremental stops-making-sense judgments. We will demonstrate the method in a case study on quantifier scope, a perennial issue in semantics where intuitions are often far from clear. In particular, we were interested in whether and when inverse scoping becomes available during processing doubly quantified German sentences. We will present evidence that in object-before-subject sentences the inverse reading becomes available already during reading while for subject-initial sentences the inverse reading is only available post-interpretively.

2. Introducing the new method

2.1. The Truth-Value Judgment (TVJ) Task

In order to measure interpretation preferences as they evolve during incremental processing of semantic ambiguities we took the truth value judgment task (henceforth TVJ task, see Crain & Thornton 1998 p. 209)
and turned it into an online method. The TVJ task has been successfully applied to a wide range of semantic and pragmatic phenomena. In this task participants read or hear a sentence and have to verify or falsify it with respect to a given context. The context is often supplied in the form of a picture, but can also be purely linguistic. Thus, the experimenter has control over two variables, the linguistic stimulus and the context.

The TVJ task is better suited to gather information about dispreferred interpretations than its alternatives, what we may call selection and production tasks. By “selection tasks” we refer to tasks where the experimenter only controls the sentence and provides a whole set of disambiguating contexts from which the participant can choose. “Production tasks” work the opposite way. A disambiguating context is presented and the linguistic expression has to be chosen that fits it best. What is problematic with both of these types of tasks is the possibility of a winner-take-all effect. The preferred reading/expression may attract the participant’s attention causing a dispreferred option to be overlooked and, therefore, wrongly appear to be impossible. A winner-take-all effect is unlikely to occur in a TVJ task because, on each trial, the sentence appears with only one context. A dispreferred reading can therefore be evaluated without competition with the preferred alternative (see Crain & Thornton 1998 for a discussion). Note that the methodological considerations regarding selection and production tasks are not limited to offline methods but also extend to online paradigms. For instance, the Visual World Paradigm (Cooper 1974; Tanenhaus et al. 1995) must be considered an online selection task since participants are looking at a set of referents associated with the different readings. Again, not looking at a referent associated with a certain interpretation can either mean (i) that a certain reading is not available or (ii) that the reading is principally available but the presence of a preferred alternative prevented participants from looking at the dispreferred alternative.

A second advantage of the TVJ task is that it allows the experimenter to decide whether a reading with a very low acceptance rate is available or not. To disentangle highly dispreferred from unavailable readings we simply need to compare acceptance of dispreferred readings to a baseline error rate, for instance, acceptance in a disambiguated construction. We will elaborate on this point in more detail when we discuss our experiments.

A potential problem with the TVJ task is that a given context may induce a bias to interpret the sentence in a way that fits it. For instance, 80% acceptance of a sentence in a particular context does not imply that the corresponding meaning would be predominant without supportive context.
High acceptance rates just indicate that an interpretation can be generated with relative ease. The difference in acceptance rates between two conditions does, however, represent the relative preferences. It is thus crucial to test a sentence with all its potential meanings to be able to properly interpret the relative differences.

No matter what kind of task is used we have to be aware of a potential problem which is very rarely considered in psycholinguistic research on semantics. Often, meaning is only indirectly assessed by measuring the fit with a particular disambiguating context. Meaning is standardly characterized by the whole set of situations in which it is true and a given context is clearly only one instance of these. Situations may of course differ in how typical a representation they are of a particular meaning. If we test atypical or unnecessarily complex situations, it is possible that a meaning might be wrongly taken to be unavailable. So, ideally, we should always choose the most typical situation as the disambiguating context. We will come back to these issues when we discuss Experiment 3.

However, the classical TVJ task can only inform us about the final interpretation. To find out whether a reading becomes available during the ongoing processing of a sentence we need an online method. For this purpose, we created an online version of the TVJ task, the incremental truth value judgment (ITVJ) task, which collects truth-value judgments incrementally for each new piece of the incoming sentence.

2.2. The Incremental Truth Value Judgment (ITVJ) Task

The ITVJ task is procedurally very similar to the stops making-sense judgment task (Tanenhaus, Boland, Garnsey & Carlson 1989). Participants have to make successive TVJs while reading a sentence. In contrast to the stops-making-sense judgment task, participants in the ITVJ task first inspect a disambiguating context (eg. a picture). Then the context disappears and a sentence is displayed self-paced in an incremental fashion (eg. via moving window presentation). At each segment participants have to decide whether, up to this point, the sentence still fits the context. By pressing the “yes, it fits, go on” button they get to the next segment. If they press the “no, can’t be true” button the presentation of the sentence is aborted.

As in the stops-making-sense judgment task, two dependent variables are analyzed for each segment: cumulative rejections, ie. how often the trial has been aborted up to this point and reaction times of “yes, go on” judgments. To illustrate how cumulative rejections are computed consider an
example data set with three sentence segments and ten participants. At the first segment all participants choose “yes”. At the second segment five participants choose “yes”, while the other five abort the trial. At the last segment three out of the five remaining participants choose “yes”. This would yield the following cumulative rejection rates: region 1 – 0%, region 2 – 50%, region 3 – 70%.

The rationale behind the task is the following. If the interpretation conveyed by the disambiguating context is inconsistent with the incoming sentence, more rejections are expected than when it is consistent. Furthermore, if one reading is preferred to another one fewer rejections and shorter reaction times for positive judgments are expected for the preferred than for the dispreferred reading.

As in the original TVJ task, the experimenter controls both the linguistic construction and the disambiguating context. The two tasks differ, however, with respect to the information they provide. On the one hand, the ITVJ task provides more information than the original TVJ task, because it allows us to track the availability of a reading during incremental interpretation. At the same time, some information may be lost or skewed in the ITVJ task (compared to the TVJ task). Take, for instance, the sentence ‘All kids are at school or they are at home’ in a context where some of the kids are at home. Participants might abort the trial when processing school because locally, up to this point, the sentence is inconsistent with the context. It only turns out to be consistent after encountering the second disjunct. Therefore, researchers using the ITVJ task have to bear in mind that non-monotonic updates of the semantic representation may yield unexpected effects.

Moreover, because in the ITVJ task participants have to keep a context in mind and compare it to the linguistic input, language external factors like complexity of the contexts or working memory limitations may influence the judgments. This clearly limits the applicability of the task to a smaller range of phenomena compared to the original TVJ task.

In contrast to other online methods, readers are forced to fully interpret every incoming piece where under normal circumstances they might have engaged in shallow processing. It is therefore important to stress that ITVJ task experiments are not intended as a substitute for other online methods, but to provide additional information not available otherwise.

Conroy (2008) introduced a superficially similar method, the incremental verification task, and investigated interpretation preferences of scopally ambiguous sentences in children and adults. In her experiments participants
first heard a potentially scope ambiguous sentence and then uncovered a complex picture one piece at a time. They had three options: (i) uncover more of the picture, (ii) abort the trial by pressing “true” or (iii) pressing “false”. Conroy used the method to investigate whether scope is already determined during reading or only computed at the verification stage. While her task is intended to find out whether a particular semantic phenomenon, namely scope, is left underspecified, the ITVJ task is intended to reveal which readings are generated online when the processor is driven to compute them. Clearly, the methods complement each other and should be applied to the same phenomena.

3. Introducing the case study

3.1. The phenomenon

Scope Ambiguity has been at the core of semantic theory at least since Montague (1973). The literature is so huge that we cannot go into details here (for a recent overview see Szabolcsi 2010). Instead, we will arbitrarily pick three theories which make fundamentally different predictions. During the course of the paper we will demonstrate that the disagreement about the empirical facts may have to do with how “deeply” doubly quantified sentences are processed when judging whether a reading is available or not. First we will look at the constructions tested in our case study.

(1) is a doubly quantified German sentence with the two determiners jeden dieser (‘each of these’, ∀) and genau einer (‘exactly one’, ∃!). Note that in this sentence the object each of these students appears in topicalized position and precedes the subject exactly one teacher. The standard assumption is that sentences like (1) have the two readings illustrated by the logical formulae in (2) and (3)\textsuperscript{1}.

\begin{align*}
(1) & \quad \text{Jeden dieser Schüler lobte genau einen Lehrer.} \\
& \quad \text{‘Each of these students was praised by exactly one teacher.’}
\end{align*}

\begin{align*}
(2) & \quad \forall x \ [ \text{Student}(x) \rightarrow \exists ! y \ [ \text{Teacher}(y) \land \text{Praise}(y, x)]] \\
(3) & \quad \exists ! y \ [ \text{Teacher}(y) \land \forall x \ [ \text{Student}(x) \rightarrow \text{Praise}(y, x)]]
\end{align*}
We refer to the reading where the linearly first quantifier scopes over the second as the surface scope interpretation which is illustrated in (2). The inverse reading is an interpretation where the second quantifier takes scope over the first, as in (3).

(4) Genau ein Lehrer\textsubscript{nom} lobte jeden dieser Schüler\textsubscript{acc}.

‘Exactly one teacher praised each of these students.’

Example (4) can be used to illustrate that the sentence structure may have an influence on scope preferences: Intuitively, an inverse reading seems to be more difficult to get in (4) than in (1). Semantic frameworks often fail to account for this intuitive difference and assume multiply quantified sentences to uniformly display all combinatorially possible scopal relations (eg. May 1977 concerning scope in English). Unconstrained theories thus predict a sentence with n quantifiers to allow for n! readings and sentences (1) and (4) are predicted to be equally ambiguous.

More constrained theories predict different scopal relations depending on syntactic structure (eg. Reinhart 1983, Frey 1993). Based on c-command relations, Frey (1993), for instance, would predict (1) to be ambiguous, but (4) to only allow surface scope.

The above-mentioned theories do not account for lexical and discourse factors which may also influence quantifier scope. This is different for instance in Pafel’s theory (2005) which takes quantifier scope preferences to be influenced by multiple, syntactic as well as nonsyntactic, factors. His theory predicts that (1) should only have a surface scope interpretation. This is due to the fact that the object quantifier not only linearly precedes the subject quantifier, but is also distributive and discourse bound as an effect of the partitive construction. By contrast, (4) is predicted to be ambiguous with a preference for the surface scope interpretation since scope factors symmetrically apply to both quantifiers. To sum up, the theories make fundamentally different predictions regarding the constructions to be tested.

Psycholinguistic studies of quantifier scope (for an overview see eg. Pylkkänen and McElree 2006) provide evidence that a number of factors influence preferences. What is important for our study is that (i) the linear order of the quantifying expressions and (ii) their grammatical functions have been shown to influence scope. As for word order, surface scope is preferred to inverse scope (eg. Johnson-Laird 1969). Concerning grammat-
Oliver Bott & Fabian Schlotterbeck

atical functions, subjects tend to scope over objects (e.g. Ioup 1975, Kurtzman & McDonald 1993). Applied to sentences (1) and (4), these findings are in line with Frey’s (1993) semantic theory. In (1) linear precedence and grammatical function hierarchy are in conflict, whereas in (4) both factors favor surface scope.

In cases like (4), however, the existing studies do not allow us to determine whether the inverse reading is highly dispreferred or completely unavailable. A recent study by Musolino (2009) illustrates this point. In a TVJ experiment he reported 7.8% acceptance for the inverse scope of three boys are holding two balloons. The low acceptance suggests that this reading isn’t available, but, as Musolino himself points out, “this conclusion would be premature as the very low acceptance rate in this case might simply reflect the fact that [inverse scope] is massively dispreferred” (p. 36).

This brings us to the main research questions of the present case study. Do comprehenders have access to the inverse interpretation during reading a doubly quantified sentence? If so, is the inverse reading only possible in some constructions but not in others? The strongest claim concerning these questions is stated in the Unlimited Scope Ambiguity Hypothesis in (5)3.

(5) Unlimited Scope Ambiguity: During processing a multiply quantified sentence, all combinatorially possible orders of quantifiers constitute possible readings (i.e. a sentence with n quantifiers should allow n! readings).

3.2. The tested constructions

To test this hypothesis we constructed four types of German doubly quantified sentences. A sample item is provided in (6)–(9) with vertical lines indicating segmentation in the ITVJ task. The subject of each sentence always contained the determiner genau ein (‘exactly one’) and the object included jeden dieser (‘each of these’). Sentence conditions (6) and (7) had SVO word order, whereas conditions (8) and (9) were object topicalized OVS sentences. For each word order, we compared potentially ambiguous sentences (6)/(8) with clearly unambiguous cases (7)/(9) that only have surface scope. The latter were disambiguated by making the quantifiers clause-bounded. These unambiguous controls served as baseline comparison for the potentially ambiguous conditions. The ambiguous conditions and the unambiguous controls were kept identical from the verb region.
until the end of the sentence to be able to compare them both with respect to rejection rates and reading times in the ITVJ task.

(6) *Ambiguous, SVO*

*Genau ein Lehrer lobte jeden dieser Schüler*…

exactly one teacher\(\text{nom}\) praised each of these students\(\text{acc}\) …

‘Exactly one teacher praised each of these students…’

(7) *Unambiguous, SVO*

*Für genau einen Lehrer gilt: er lobte jeden dieser Schüler*…

for exactly one teacher holds: he praised each of these students…

‘For exactly one teacher holds: he praised each of these students…’

(8) *Ambiguous, OVS*

*Jeden dieser Schüler lobte genau ein Lehrer*…

each of these students\(\text{acc}\) praised exactly one teacher\(\text{nom}\) …

‘Each of these students was praised by exactly one teacher…’

(9) *Unambiguous, OVS*

*Für jeden dieser Schüler gilt: ihn lobte genau ein Lehrer*…

for each of these students holds: he praised exactly one teacher…

‘For each of these students holds: he was praised by exactly one teacher…’

(10) *…voller Wohlwollen.*

‘…full of goodwill.’

One may ask whether the unambiguous controls really exclude the inverse interpretations. A possible counterargument that was pointed out to us by an anonymous reviewer is that (9) can be continued with an elaboration of a specific referent as in …nämlich Lehrer Lutz (‘namely teacher Lutz’). We take this to be a matter which is orthogonal to scope. The linear interpretation of (9), i.e. (3), is compatible with a situation in which the same teacher praised all the students. The scopal facts only become clear when we test the sentences in contexts which clearly disambiguate towards either scope reading.
3.3. Disambiguating Contexts

We paired the sentence materials with two types of set diagrams. Sample diagrams are shown in Figure 1. The first type of diagram (∀∃!-diagrams; Fig 1a) was a model satisfying the ∀∃!-reading but was inconsistent with the ∃∀∀-reading. The second type, ∃∀∀-diagrams (Fig 1b), depicted a model only satisfying the ∃∀∀-reading but not the ∀∃!-reading. Both sets and the verbal predicate were labeled in the diagrams. In a cross-methodological comparison Bott and Radó (2007) showed that disambiguation via set diagrams of this sort yield highly reliable and valid results compared to two other methods of disambiguation.

In order to keep participants from developing strategies, we varied the depicted situations in both types of diagrams. The ∀∃!-diagrams consisted in situations with one-to-one mappings between the two sets, but majorily were situations as in (1b) which are intuitively less typical models of a ∀∃!-reading. The ∃∀∀-diagrams always had additional lines to rule out an ∀∃!-interpretation, but we varied the exact number of additional lines between items. We will come back to processing consequences of typical vs. atypical models in Experiment 3.

Crossing the two picture types with the four construction types yielded eight conditions in a $2 \times 2 \times 2$ (word order $\times$ ambiguity $\times$ reading) within-design. We constructed 32 experimental sentence-picture pairs in eight conditions each and distributed them over eight lists in a latin square design. The same 73 fillers (31 true and 42 false sentence-diagram pairs) were added to each list.

![Figure 1. Disambiguating set diagrams used in Experiment 1 and 2 (labels translated from German into English).]
3.4. Predictions

The *Unlimited Scope Ambiguity Hypothesis* makes the following predictions. The unambiguous controls should only be compatible with diagrams disambiguating towards surface scope. The ambiguous constructions should be equally compatible with contexts disambiguating towards surface and inverse scope. This pattern should be the same in the OVS- and in the SVO-conditions.

Frey (1993) and Pafel (2005) are more restrictive. Frey predicts the ambiguous OVS condition in (8) to be compatible with both surface scope and inverse scope diagrams, whereas the ambiguous SVO condition in (6) should only be compatible with linear scope. In diagrams disambiguating towards inverse scope, the purportedly ambiguous SVO construction (6) should thus be rejected as often as unambiguous controls. Pafel, by contrast, predicts exactly the opposite pattern. While the SVO construction in (6) should be ambiguous, the OVS construction in (8) should only allow surface scope.

The semantic theories do not allow us to derive any predictions about the processing of these constructions. We hypothesized, though, that the available readings might be more limited in the ITVJ task than in an ordinary TVJ task. In the TVJ task participants have access to the complete sentence and may thus reformulate it in a way to fit the context. This should be impossible in the ITVJ task where readers have to decide whether the incoming, yet incomplete sentence fits the context. More concretely, we expected only those readings to be available in the ITVJ task which can be arrived at by means of grammatical operations independently required for parsing. It has been claimed that OVS constructions involve more structure than SVO constructions in that they require a trace and movement of the topicalized object (e.g. Gorrell 2000). The presence of the trace, however, opens up the possibility to reconstruct the topicalized object quantifier in its base position leading to inverse scope. The reconstruction process can be either syntactically guided (e.g. von Stechow 1991) or purely semantic (Sternefeld 2001). By contrast, reconstruction is impossible in the SVO construction and inverse scope should thus be impossible online, even though it may be possible offline.
4. Comparing TVJ and ITVJ

4.1. Experiment 1 – offline data (TVJs)

The first experiment was an ordinary TVJ task. The findings of this experiment served as evaluative comparison for the new method. As outlined above ambiguity can arise at different stages during interpretation. While ordinary TVJs reflect all stages, ITVJs are much closer to online processing and, in particular, should not reflect readings that come about post-interpretively, that is only during reasoning about the possible meanings. If the results of an ITVJ task experiment contrast with the results of the present experiment that will allow us to draw conclusions about whether both the OVS and SVO sentences are ambiguous and will shed light on whether scope ambiguity can already be attested during the online processing of doubly quantified sentences.

4.1.1. Method

The experiment was a TVJ task experiment in which participants judged how well the sentence fits the diagram without any time pressure. The experiment was administered over the internet using WebExp 2 (Mayo et al. 2006). Participants were tested in a quiet computer pool. They first received written instructions and completed a practice session with ten trials. Then the experiment followed in a single block. The sentence-diagram pairs were presented in a random order. Acceptability judgments had to be provided on a seven point scale. This was done to give participants the opportunity to indicate imperfect matches by intermediate values. The experiment took about 20 minutes. 48 German native speakers from the faculty of modern languages (mean age: 24.7; range: 20–33; 32 female) participated in the study for payment of 5€. We normalized the data by computing z-scores for each subject.

4.1.2. Results

<table>
<thead>
<tr>
<th></th>
<th>surface scope</th>
<th>inverse scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambiguous SVO</td>
<td>0.42 (0.05)</td>
<td>-0.79 (0.08)</td>
</tr>
<tr>
<td>Unambiguous SVO</td>
<td>0.65 (0.03)</td>
<td>-1.26 (0.06)</td>
</tr>
<tr>
<td>Ambiguous OVS</td>
<td>0.44 (0.06)</td>
<td>-0.58 (0.07)</td>
</tr>
<tr>
<td>Unambiguous OVS</td>
<td>0.57 (0.04)</td>
<td>-0.82 (0.07)</td>
</tr>
</tbody>
</table>
Table 1 presents the mean judgments in the experimental conditions. Surface scope was generally preferred to inverse readings. This preference was modulated by sentence type. It was strongest in the unambiguous SVO sentences followed by the unambiguous OVS construction. In the ambiguous conditions the surface scope preference was weaker, but again it was more pronounced in the SVO than in the OVS sentences.

A repeated measures ANOVA revealed a reliable main effect of reading ($F_1(1,47) = 183.70, p < .01; F_2(1,35) = 1144.25, p < .01$) reflecting the surface scope preference. There was also a significant main effect of order ($F_1(1,47) = 8.38, p < .01; F_2(1,35) = 23.45, p < .01$) and a marginal main effect of ambiguity ($F_1(1,47) = 4.58, p < .05; F_2(1,35) = 3.74, p = .06$). The former was due to SVO sentences being rated better than OVS sentences and the latter to higher ratings for ambiguous than for unambiguous sentences. As it turned out, there was also an interaction of order and reading ($F_1(1,47) = 6.88, p < .05; F_1(1,35) = 18.26, p < .01$) which was due to higher ratings in the inverse OVS conditions (mean z-score: $-0.70$) than in the inverse SVO conditions (mean z-score: $-1.03$). Most importantly, the analysis revealed a significant interaction of reading and ambiguity ($F_1(1,47) = 55.28, p < .01; F_2(1,35) = 43.51, p < .01$) as well as a marginal three-way interaction of order, reading and ambiguity ($F_1(1,47) = 3.75, p = .59; F_2(1,35) = 2.90, p = .99$). The two-way interaction reflects stronger surface scope preferences in the unambiguous than in the ambiguous conditions. The three-way interaction was due to the fact that this pattern was more clear-cut in the SVO than in the OVS sentences.

In separate analyses of the SVO conditions only, we found a reliable main effect of reading ($F_1(1,47) = 474.45, p < .01; F_2(1,35) = 956.23, p < .01$) reflecting the general surface scope preference. Furthermore, the main effect of ambiguity ($F_1(1,47) = 4.14, p < .05; F_2(1,35) = 3.90, p = .06$) was reliable by participants, but marginal by items. The main effect was due to higher ratings in the ambiguous conditions than in the unambiguous conditions. Crucially, the interaction of reading and ambiguity was also significant ($F_1(1,47) = 44.03, p < .01; F_2(1,35) = 33.16, p < .01$). This was because the surface scope preference was stronger in the unambiguous than in the ambiguous conditions. Pairwise comparisons revealed significant differences both in the inverse ($t_1(47) = -11.10, p < .01; t_2(35) = -14.78, p < .01$) and in the surface scope conditions ($t_1(47) = 5.47, p < .01; t_2(35) = 6.85, p < .01$).
ANOVAs analyzing only the OVS conditions revealed a reliable main effect of reading. It was due to the fact that surface scope diagrams were rated better than inverse scope diagrams \((F_1(1, 47) = 165.57, p < .01; \ F_2(1, 35) = 345.19, p < .01)\). We also found that the surface scope preference was reliably stronger in the unambiguous than in the ambiguous conditions leading to a significant interaction of reading and ambiguity \((F_1(1, 47) = 10.24, p < .01; \ F_2(1, 35) = 8.15, p < .01)\). The interaction was somewhat weaker than in the SVO conditions. In pairwise comparisons both the inverse \((t_1(47) = –5.29, p < .01; \ t_2(35) = –4.65, p < .01)\) and the surface scope conditions \((t_1(47) = 3.41, p < .01; \ t_2(35) = 3.43, p < .01)\) differed significantly from one another.

4.1.3. Discussion

In the present experiment both ambiguous OVS and SVO conditions allowed an inverse interpretation. The inverse readings were above the baseline controls across both construction types. At first sight, this provides evidence in favor of the Unlimited Scope Ambiguity Hypothesis. What remains unexplained, however, is the preference for surface scope. What is clearly needed is a gradient notion of scope which can also account for graded preferences.

In contrast to Frey’s (1993) predictions about doubly quantified SVO sentences, the ambiguous SVO construction was compatible with inverse scope. The SVO sentences seemed to be judged even “more ambiguous” than the OVS sentences.

Our findings are not compatible with Pafel’s (2005) theory, either. Contrary to the predictions derived from his model, the OVS construction was ambiguous. With respect to the ambiguous SVO construction, though, the findings nicely fit the graded preferences predicted by this model.

4.2. Experiment 2 – online data (ITVJs)

The findings of the first experiment show that readers are able to compute inverse interpretations when the sentence and the diagram are both present and there is no time limit. We cannot decide, however, whether readers were able to do so during reading or only at a later stage. To gain access to the intermediate stages we conducted an ITVJ version of the experiment. This might prove insightful in two respects. On the methodologi-
In the current study, we aim to provide a nuanced account of the cognitive processes involved in truth value judgment tasks, focusing on the relative merits and implications of the ordinary Truth Value Judgments (TVJ) and the Incremental Truth Value Judgments (ITVJ) tasks. On the empirical side, diverging results between the ordinary TVJ and the ITVJ task will demonstrate that the two tasks in fact allow us to tap into different stages of sentence interpretation. On the semantic side, different results in the two tasks are relevant for the debate about quantifier scope ambiguities. To anticipate the findings, we will suggest that the fundamentally divergent intuitions reported in the literature can be reconciled once we take into account the cognitive effort which theoreticians might have put into their judgments.

4.2.1. Participants and Procedure

40 native German speakers (mean age 25.6, range 20–67, 32 female) from Tübingen University participated in the study. They were naïve to the purpose of the study. Each participant received 8 € compensation. The ITVJ task experiment began with a practice session of ten trials. Feedback was provided only during the practice session. Three experimental blocks followed. Presentation order was randomized between and within blocks. Participants were tested individually. An experimental session took approximately 30 minutes.

4.2.2. Statistical Analysis

Cumulative rejections on the verb region and the region containing the second quantifier were analyzed using logit mixed effects models with fixed effects of word order, ambiguity, reading and their interactions as well as the random intercepts of items and participants (cf. Jäger 2008).

RTs of “yes, go on” judgments on the second quantifier region were analyzed in the conditions with diagrams disambiguating towards surface scope. We limited the analysis to these cases because the other conditions received “no” responses on the majority of trials. We trimmed RTs longer than 3500ms or below 100ms. RTs were analyzed in a linear mixed effects model with fixed effects of word order and disambiguation, their interaction and random intercepts of items and participants (cf. Baayen, Davidson and Bates 2008). We limited the statistical analysis to regions up to the point when one of the experimental conditions was massively aborted4.
4.2.3. Results

Mean cumulative rejection rates are presented segment by segment for SVO sentences in Figure 2 and for OVS sentences in Figure 3. Prior to the region containing the second quantifier all rejection rates were below 8%.

From the second quantifier onwards there were clear differences in cumulative rejection rates with an overwhelming preference for surface scope diagrams. At the second quantifier, the unambiguous SVO conditions had been rejected more often after inverse diagrams than after surface scope diagrams (64.4% vs. 5%). The pattern was the same but numerically even more pronounced in the ambiguous SVO conditions with 72.5% cumulative rejections after inverse diagrams and 4.4% after surface scope diagrams. The preference for surface scope readings was less clear-cut in the OVS conditions. At the second quantifier, the unambiguous OVS construction had been rejected 64.4% of the time after an inverse and 30.6% after a surface scope diagram. The surface scope preference was even less pronounced in the ambiguous OVS condition. Here the second quantifier region received 56.9% cumulative rejections after an inverse and 36.3% after a surface scope diagram.

Figure 2. Cumulative rejection rates (SVO sentences) in Experiment 2.

The fact that OVS sentences were rejected more often than SVO sentences at the second quantifier led to a reliable effect of word order (esti-
mate = –2.81; z = –6.3; p < .01) in the logit mixed effects model analysis. In addition, inverse diagrams were rejected more often than surface scope diagrams. This resulted in a reliable effect of reading (estimate = 1.04; z = –4.1; p < .01). The interactions of word order and reading (estimate = 3.68; z = 7.1; p < .001) and of ambiguity and reading were also significant (estimate = –.7; z = –1.9; p = .05). We will come back to these effects when we discuss the effects of the models analyzing the two word orders separately. The three-way interaction of word order, reading and ambiguity was marginally significant (estimate = 1.32; z = 1.8; p = .07). Thus, the amount of ambiguity differed between SVO and OVS constructions.

To further break down the three-way interaction, we computed separate logit mixed effects analyses for the SVO and of the OVS conditions. The analysis of SVO conditions revealed a significant effect of reading (estimate = 4.80; z = 9.96; p < .001) reflecting the preference for surface scope. Neither the effect of ambiguity nor the interaction of reading and ambiguity were reliable (|z| < 1). Numerically, the difference between surface scope and inverse scope conditions was even bigger in the purportedly ambiguous cases than in the disambiguated controls.

In the OVS analysis, the effect of reading was again significant (estimate = –1.07; z = –4.21; p < .001) showing an across-the-board preference for surface scope. The interaction of reading and ambiguity was also
significant (estimate = –.72; \( z = -1.97; p < .05 \)). It was due to smaller differences between surface scope and inverse diagrams in the ambiguous than in the unambiguous conditions. A one-sided pairwise comparison revealed that the difference between the ambiguous inverse scope condition and its corresponding control was significant (estimate = –.52; \( z = -1.8; p < .05 \)) and the difference between the two surface scope conditions was marginal (estimate = –.44; \( z = -1.53; p = .06 \)). As predicted, readers were thus able to compute inverse scope in the ambiguous OVS constructions while reading the sentence.

More evidence that scope interaction only took place in the OVS sentences comes from the analyses of RTs (depicted in Figure 4). Up to the verb region the surface scope conditions did not differ from each other (\(|t| < .5\)). On the second quantifier region there were clear differences in RTs. While the unambiguous OVS condition had a mean RT of 1615ms, the ambiguous OVS condition had a mean RT of 2420ms. Unambiguous and ambiguous SVO conditions had mean RTs of 1167ms and 1290ms, respectively. The statistical analysis revealed a reliable effect of word order (estimate = –1709; \( t = -5.01; p < .01 \)) and of ambiguity (estimate = –1324; \( t = -3.67; p < .01 \)). These main effects reflect the fact that on the second quantifier region it took participants longer to respond “yes” in the OVS than in

\[ \text{Figure 4. Judgment RTs of “yes” judgments in Experiment 2 (linear scope conditions only).} \]
the SVO condition and in ambiguous than in unambiguous sentences, respectively. Crucially, the interaction of word order and ambiguity was also reliable (estimate = 615; t = 2.91; p < .01). The interaction is due to the fact that in the OVS conditions the second quantifier had longer RTs in the ambiguous than in the unambiguous condition, whereas RTs were indistinguishable between the SVO conditions. A pairwise comparison analyzing only the OVS conditions revealed a reliable effect of ambiguity (estimate = –639; t = –3.05; p < .01). Another comparison analyzing only the SVO conditions did not reveal a reliable effect (t = –1.01; p = .32). The pattern was the same at the subsequent segments.

4.2.4. Discussion

There were interesting differences between TVJs and ITVJs. For SVO sentences, the new method did not yield any indication of ambiguity. The purportedly ambiguous SVO construction was indistinguishable from the unambiguous controls both in rejection rates and in judgment times. The findings thus provide evidence against the Unlimited Scope Ambiguity Hypothesis. They instead support Frey’s account (1993) who predicted that doubly quantified German SVO sentences would only have surface scope.

Nevertheless, we were able to detect ambiguity in the OVS constructions. In this respect, the data pattern fits the results of the previous experiment. The surface scope preference was not as strong as in the unambiguous control. Finding this interaction in the ITVJ task, too, indicates that the inverse interpretation was accessible already during reading. Further evidence for this claim comes from RTs. We found longer RTs for “yes, go on” button presses in the ambiguous OVS construction than in the disambiguated control. This may be due to solving a scope conflict in the ambiguous construction.

The similarities and differences between the two methods fit our processing assumptions. In Experiment 2 the inverse reading was only available in the object topicalized construction, where reconstruction is independently assumed, but not in SVO constructions which lack this possibility.

This result has implications in two directions. First, by comparing TVJs to ITVJs it is possible to disentangle those readings that are only available post-interpretively from the ones accessible during online interpretation. Secondly, semanticists should pay attention to processing because depending on how deeply a given sentence is processed intuitions may be com-
completely different. If we take depth of processing into account, however, it becomes possible to explain the variation of introspective scope judgments in the semantic literature.

A possible objection to these conclusions could be that due to enhanced task complexity ITVJs are noisier than TVJs or that dispreferred readings are generally less available in the ITVJ task than under normal circumstances. Neither of these alternatives can explain our data in all their particulars. This is because in the ITVJ task readings which were possible in the first experiment were filtered selectively. In the TVJ task the relative difference between the preferred and the dispreferred reading was even smaller in the ambiguous SVO sentences than in the OVS sentences. In the present ITVJ experiment, however, the dispreferred reading was filtered only in the former construction. This is completely unexpected under the assumption of a general shift in preference towards the preferred interpretation.

At first glance two aspects of the data seem problematic, however. The first one is that at the end of the sentence the ambiguous OVS construction under the inverse reading was rejected nearly as often as the corresponding control. Should we really assume that the OVS sentences were ambiguous, then? We replicated the OVS part of the present experiment to see whether the interaction of reading and ambiguity in the OVS conditions was stable across experiments. Again, this interaction turned out to be reliable, this time even up to the last segment of the sentence. A second concern is that the unambiguous OVS construction was falsely rejected in more than 40% of the cases. In Experiment 3, we investigated whether this error might be due to the atypicality of the disambiguating diagrams.

5. Experiment 3: Typicality

Why did participants falsely reject the unambiguous OVS condition almost 40% of the time following a picture compatible with the only possible reading? Should we, nevertheless, rely on a task that gives rise to so many errors? In the present experiment we investigated different kinds of disambiguating situations to find out what might have gone wrong. We will see that results become perfectly unbiased even for the unambiguous OVS surface scope condition if we keep the disambiguations as simple as possible.
Up to now, we have been equating the meaning of an expression with the particular situation used for disambiguation. This may not be entirely correct since a meaning corresponds to a whole set of situations and it is possible that some situations or models are easier to evaluate with respect to a certain meaning than others. Thus, presenting a model that is far from the most typical scenario may lead to more errors than a more typical model. Consider the $\forall \exists!$-diagrams in (1b) again. Reinspecting them reveals two aspects we didn’t pay attention to when we constructed the materials. These might have caused the errors. Both features are illustrated in Figure 1b. The first thing to notice is that the restrictor set contains an additional object (an ‘extra’ teacher), so there is no one-to-one correspondence between the two sets. Secondly, the model includes a branching configuration (ie. the same teacher praising two students). In the present experiment we manipulated these two features in a factorial design while keeping the sentences and their meaning constant.

Interestingly, one of these features has already been extensively studied in the acquisition literature on universal quantification. Since the pioneering work of Inhelder & Piaget (1959) it has been established that children show certain non-adult responses in interpreting sentences containing a universal quantifier. One finding is that children tend to judge sentences like every boy is riding an elephant to be false in a context with each boy riding an elephant and an extra elephant without a rider, but behave like adults when no extra elephant is present in the scene. Thus, children seem to not only restrict the quantificational domain to the set of boys but also falsely consider the whole set of elephants.

In comparison to an ordinary TVJ task, the incremental version arguably puts more cognitive load on the participants. While keeping context in memory they simultaneously have to read and relate the incoming sentence to the picture. Under closer scrutiny, the ITVJ task provides a multiple task setting and it thus seems very likely that it gives rise to errors that wouldn’t occur under ‘normal’ circumstances. So, from a methodological perspective, it would be important to know where the limits of the new method are. As a side-effect of establishing that we will demonstrate that the method can be used to gain insights into the default interpretation that is associated with a particular meaning.
5.1. Method

We tested the 32 experimental sentences from the previous experiment in the unambiguous OVS conditions. For each item, we drew four new $\forall \exists !$-diagrams. Figure 5 depicts sample diagrams. We manipulated the presence of extra (i.e., unconnected) objects in the restrictor set as well as the presence of branching lines (i.e., an element in the set on the left that is connected to two elements on the right) according to a 2x2 factorial design ($\text{extra object}$ ($+/- \text{eo}$) vs. $\text{branching line}$ ($+/- \text{bl}$)). The simplest condition $[-\text{eo}, -\text{bl}]$ were diagrams with neither an extra object nor a branching line. In the $[+\text{eo}, -\text{bl}]$ condition the diagrams included an extra object. The $[-\text{eo}, +\text{bl}]$ diagrams had a branching line but no extra object. The $[+\text{eo}, +\text{bl}]$ diagram had both features. We added the fillers from the first experiments and created counterbalanced lists according to a latin square.

The procedure was the same as in the previous experiment. 40 native German speakers from Tübingen University (mean age 23.9, range 19–35, 31 female) took part in the study for a payment of 8 €.

![Figure 5. $\forall \exists !$ set diagrams in Exp. 3 (eo = extra object; bl = branching line).](image-url)
5.2. Results

Error rates clearly differed between conditions. At the second quantifier region, \([-eo, -bl]\) was rejected only 0.6% of the time, \([+eo, -bl]\) 7.5% of the time, \([-eo, +bl]\) 36.3% and \([+eo, +bl]\) 31.8% of the time. Thus, participants basically made no mistakes in the \([-eo, -bl]\) condition. Errors only occurred when extra objects or branching lines were present. The end of sentence rejection rates followed the same pattern: 1.25%, 11.25%, 44.38% and 45.00% respectively.

We computed a logit mixed effects model analyzing cumulative rejection rates of the second quantifier region with branching line and extra object as well as their interaction as fixed effects and participants and items as random intercepts. The statistical analysis revealed significant main effects of branching line (estimate = 5.00; z = 4.55; p < .01) and of extra object (estimate = 2.68; z = 2.38; p < .05) and a significant interaction between the two factors (estimate = –2.93; z = –2.53; p < .05).

To further analyze the effect of branching line we computed pairwise comparisons: Comparing the \([-eo, -bl]\) with the \([-eo, +bl]\) condition revealed a significant effect (estimate = 5.7; z = 4.13; p < .01). A second model comparing only the \([+eo, -bl]\) with the \([+eo, +bl]\) condition also yielded a significant effect (estimate = 2.03; z = 5.35; p < .01). Thus, pictures with a branching line lead to an increase in rejections irrespective of the presence or absence of an extra object.

Two more pairwise comparisons investigated the influence of extra objects. The first was a comparison of the \([-eo, -bl]\) with the \([+eo, -bl]\) condition. It revealed that the presence of extra objects similarly resulted in a significant increase in rejection rate (estimate = 5.19; z = 2.49; p < .05). Another model comparing the \([-eo, +bl]\) with the \([+eo, +bl]\) condition revealed no reliable difference between the two conditions (effect of extra object: estimate = –0.23; z = 0.25; p < .05). Thus, although there was a reliable extra object effect when no branching line was present, there was no effect, if the picture had a branching line.

5.3. Discussion

This experiment demonstrates the limits of ITVJs. We have to be careful when constructing the materials to be tested in the task. Choosing the simplest disambiguating context, ie. the \([-eo, -bl]\) diagrams, yielded perfectly
unbiased results, but, more complex models lead to a massive amount of errors.

For constructing materials to be tested in an ITVJ experiment, we therefore suggest to first consider the most typical model associated with the meaning under investigation. The disambiguating context should then be constructed in a way to adhere to the default model as closely as possible. Presumably, Experiment 2 would have yielded perfectly unbiased results if we had used $\forall \exists!$-diagrams with a one-to-one mapping between the two sets.

6. Conclusions

We have proposed a new method, the ITVJ task, which is intended to track the available interpretations and their relative preferences during the processing of semantic ambiguity. Our case study revealed that the method can be applied to phenomena as difficult and subtle as quantifier scope for which judgments are often far from clear.

When it comes to the computation of a dispreferred reading, we distinguished between the readings that become available online during reading and those that only become accessible later, that is, post-interpretively. The comparison of ITVJs and TVJs in the first two experiments suggests that the inverse readings of SVO and OVS doubly quantified sentences are generated during different processing stages. While the inverse reading was available during reading OVS sentences, it was completely absent in the online processing of SVO sentences. We do not see how we could have arrived at this result using any other method since in both OVS and SVO constructions the inverse reading was massively dispreferred.

Our findings have important implications for semantic theories on quantifier scope that go well beyond the three theories selected for this case study. In the first two experiments we observed a strong preference for surface scope. Only very few of the existing theories on quantifier scope can account for the graded nature of these preferences. In our fairly limited sample, only Pafel’s (2005) model could account for graded preferences. Moreover, when we considered the obtained data in their entirety none of the three approaches was fully successful. Extending the discussion to theories on quantifier interaction in general, to date there is, at least to our knowledge, no model that can fully account for the observed pattern of scope interpretations. A descriptively adequate model is lacking even more
if we want to understand at which processing stages dispreferred scope readings become available. As a first step, we proposed that grammatical mechanisms such as reconstruction can make a dispreferred interpretation accessible during online comprehension.

Finally, we demonstrated that it is crucial to carefully consider the disambiguating contexts to be tested in ITVJ task experiments. We showed that readers are not able to compensate for atypical models. Of course, this opens up another possible application of the task, namely to investigate what the most typical model of a given reading might be.

7. Notes

1. We ignore tense and the internal semantics of the determiners here, and to keep matters simple abbreviate them by $\exists!$ and $\forall$.
2. Which in this case corresponds to the logical form in (3).
3. This is of course a considerable oversimplification. We are completely ignoring issues of absolute scope (eg. quantifiers that are clause-bounded) and other constraints such as variable binding.
4. Statistical analyses on later regions are difficult to interpret because the number of still possible rejections may vary between conditions. This becomes immediately clear in the extreme case. Suppose that early in the sentence one condition is aborted 100%, but in another condition participants keep on reading. If later on the latter condition is aborted, say, 80% of the time we cannot make anything of this because there would be no data to sensibly compare it to. In the following, we will therefore only analyze rejections and judgment times using inferential statistics up to the point where participants massively aborted the trials. For the rest of the sentence we will only report the descriptive statistics.
5. In this replication we tested $\forall\exists!$ diagrams that contained no extra objects or branching lines, cf. Experiment 3. Participants made practically no errors (<5%) with these diagrams, lending further support to the findings of Experiment 3.

References

Bott, Oliver and Radó, Janina

Caplan, David and Waters, Gloria S.

Conroy, Anastasia M.

Cooper, Roger M.

Crain, Stephen and Thornton, Rosalind

Fodor, Jerry A.

Frey, Werner

Gorrell, Paul

Inhelder, Bärbel and Piaget, Jean
1959 The early growth of logic in the child. London: Routledge

Ioup, Georgette

Jaeger, T. Florian
2008 Categorical data analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. Journal of Memory and Language 59: 434–446.

Johnson-Laird, Philip

Kurtzman, Howard S. and MacDonald, Maryellen C.
May, Robert

Mayo, Neil, Corley, Martin and Keller, Frank

Montague, Richard

Musolino, Julien
2009 The logical syntax of number words: Theory, acquisition and processing. Cognition 111:24–45

Pafel, Jürgen
2005 Quantifier scope in German. Amsterdam/Philadelphia: John Benja-mins.

Pyllkänen, Liina and McElree, Brian

Reinhart, Tanya

Sternfeld, Wolfgang

Szabolcsi, Anna


Tanenhaus, Michael K., Boland, Julie E., Garnsey, Susan M. and Carlson, Greg N.

von Stechow, Arnim