# 15. The application of experimental methods in semantics



Abstract

*The purpose of this paper is twofold. On the methodological side, we shall attempt to show that even relatively simple and accessible experimental methods can yield significant insights into semantic issues. At the same time, we argue that experimental evidence, both the type collected in simple questionnaires and measures of on-line processing, can inform semantic theories. The specific case that we address here concerns the investigation of quantifier scope. In this area, where judgments are often subtle and controversial, the gradient data that psycholinguistic experiments provide can be a useful tool to distinguish between competing approaches, as we demonstrate with a case study. Furthermore, we describe how a modification of existing experimental methods can be used to test predictions of underspecification theories. The programme of research we outline here is not intended to be a prescriptive set of instructions for researchers, telling them what they should do; rather it is intended to illustrate some problems an experimental semanticist may encounter but also the profit*

*of this enterprise.*

## 1. Introduction

A wide range of data types and sources are used in the field of semantics, as is demonstrated by the related article 12 (Krifka) *Varieties of semantic evidence* in this volume. The aim of this article is to show with an example research study series what sort of questions can be addressed with experimental tools and suggest that these methods can deliver valuable data which is relevant to basic assumptions in semantics. This text also attempts to address the constraints on and limits to such an approach. These are both methodological and theoretical: it has long been recognized that links between empirical measures and theoretical constructs require careful argumentation to establish.

The authors therefore have two aims: one related to experimental methodologies and the other to do with the value of processing data. They first seek to show that even relatively simple and accessible experimental methods can yield significant insights into semantic issues. They second wish to illustrate that experimental evidence such as that gathered in their eye-tracking study has the potential to inform semantic theory.

Semanticists have of course always sought confirmatory evidence to support their analyses. There is, on the one hand, fairly extensive use of computational techniques and corpus data in the field, and a growing body of experimental work on semantic processing, language acquisition, and pragmatics, but in the area of theoretical and formal semantics the experimental methods are less frequently employed.

Now there are good reasons for this. There are inherent factors related to the accessibility of the relevant measures why controlled data gathering techniques are still

somewhat less frequent in this field than in some others. We shall discuss what these reasons are and demonstrate with a case study what constraints they place on empirical studies, particularly experimental studies. The example research program that we shall report is thus not simply a recipe for others for what should be done, rather it is an illustration of the difficulties involved, which aims to explore some of the boundaries of what is accessible to experimental studies.

The specific case that we address here concerns the investigation of quantifier scope, a perennial issue in semantics. Previous attempts to account for the complex data patterns to be found in natural languages have met with the difficulty that the causal factors and preferences need first to be identified before a realistic model can be developed. This requires as an initial step the capture and measurement of the relevant effects and their interactions, which is no trivial task.

The next section lays out a range of reasons why semanticists do not routinely seek to test the empirical bases of their theories with simple experiments. Section 3 reports the series of empirical investigations on quantifier scope carried out by Bott and Radó in on-going research. Section 4 lays out some of the theoretical background and importance of these studies for current theory (the underspecification debate). The final section takes as a starting point Bott and Radó (2009) to suggest how some of the problems noted in section 3 may be overcome with a more sophisticated experimental procedure.

## 2. The stumbling blocks

As Manfred Krifka notes in his neighbouring article 12 (Krifka) *Varieties of semantic evidence*, a major problem with investigating meaning is that we cannot yet fully define

what it is. This is indeed a root cause of difficulty, but here we shall attempt to illustrate in more practical detail what effects this has on attempts to conduct experiments in this field.

## 2.1. Specifying meaning without using language

The essential feature distinguishing experiment procedure is control. In language experiments we may distinguish three (sets of) variables: linguistic form, context, and meaning. In the typical experiment we will keep two of them constant and systematically vary the other. Much semantic research concerns the systematic interdependence of form, context, and meaning. These issues can be investigated for example by:

a) keeping form and context constant, manipulating meaning systematically, and measuring the *felicity* of the outcome (in judgements, or reaction times, or processing effort), or

b) manipulating (at least one of) form and context, and measuring perceived meaning.

The first requires the experimenter to *manipulate* meaning as a variable, which entails expressing meaning in a form other than language, (pictures, situation descriptions, etc); the second requires the experimenter to *measure* perceived meaning, which again normally demands reference to meanings captured in non-linguistic form. But precisely this expression of tightly constrained meaning in non-linguistic form is very difficult.

To show how this factor affects studies in semantics disproportionately, it is worth noting how this makes controlled studies in semantics more challenging than in syntax. Work in experimental syntax is often interested in addressing precisely those effects of form change which are *independent* of meaning. The variable meaning can thus be held

constant, but this does not require it to be exactly specified. It often does not much matter exactly what interpretation subjects assign to the example structures as long as it is the same for all of them. Thus only the syntactic analysis need be controlled, not the meaning that this analysis gives rise to. This makes empirical studies in syntax much less difficult than those in semantics.

## 2.2. The boundaries of form, context, and meaning

A further problem of exact studies concerning meaning is that the three variables are not always clearly distinguished, in part because they systematically covary, but also in part because linguists do not always agree about the boundaries. This is particularly visible when we seek to identify where an anomaly lies. Views have changed over time in linguistics about the nature and location of ill-formedness (e.g. the discussion of the status of *I am lurking in a culvert* in Ross 1970) but the fundamental ambiguity is still with us. For example, Weskott & Fanselow (2009) give the following examples and judgements of syntactic and semantic well-formedness: (1a) is syntactically ill-formed (*), (1b) is semantically ill-formed (#), and (1c) is ill-formed on both accounts (*#).

(1) a. *Die Suppe wurde gegen versalzen.
the soup was against oversalted

b. #Der Zug wurde gekaut.
the train was chewed

c. *#Das Eis wurde seit entzündet.
the ice was since inflamed

Our own judgements suggest that the structures in (1-a) and (1-c) have no acceptable syntactic analysis, and therefore no semantic analysis can be constructed -- they are thus

both syntactically and semantically ill-formed. Crucially, the semantic anomaly is dependent upon the syntactic problem; the lack of a recognizable compositional interpretation is a result of the lack of a possible structural analysis. We would therefore regard these examples as primarily syntactically unacceptable. This contrasts with (1-b), which we regard as well-formed on both parameters, being merely implausible, except in a small child's playroom, where a train being chewed is an entirely normal situation (cf. Hahne & Friederici 2002).

### 2.3. Plausibility

Such examples highlight another problem in manipulating meaning as an experimental variable: the human demand to make sense of linguistic forms. We associate possible meanings with things that we can accept as being true or plausible. So 'the third-floor appartment reappeared today', which is both syntactically and semantically flawless, will cause irrelevant experimental effects since subjects will find it difficult to fit the meaning into their mental model of the world. Zhou & Gao (2009) for example argue that participants interpret *Every robber robbed a bank* in the surface scope reading because it is more *plausible* that each robber robbed a different bank.

This links in to a wider discussion of the role of plausibility as a factor in semantic processing and as a filter on possible readings. Zhou & Gao (2009) claim that such doubly quantified sentences are ambiguous in Mandarin, since their experimental evidence suggests that both interpretations are built up in parallel, but one reading is subsequently filtered out by plausibility, which accounts for the contrary judgements in work on semantic theory (e.g. Huang 1982, Aoun & Li 1989).

## 2.4. Meaning as a complex measure

The meaning of a structure is not fixed or unique, even when linguistic, social, and discourse context are fixed. First, a single expression may have multiple readings, which compete for dominance. Often a specific relevant reading of a structure needs to be forced in an experiment. Some readings of theoretical interest may be quite inaccessible, though nevertheless real. This raises the issue of expert knowledge, which again contrasts with the situation in syntax. Syntactic well-formedness judgements are generally available and accessible to any native speaker and require no expertise. On the other hand, it can require specialist knowledge to 'get' some readings since the access to variant readings is usually via different analyses. This is a crucial point in semantics, since it reduces the likelihood that the intuitions of the naïve native speaker can be the final arbiter in this field, as they can reasonably be argued to be in syntax (Chomsky 1965). A fine example of this is from Hobbs & Schieber (1987):

(2) Two representatives of three companies saw most samples.

They claim that this sentence is five-ways ambiguous. Park (1995) however denies the existence of one of these readings (*three* > *most* > *two*). It is doubtful whether this question is solvable by asking naïve informants.

Even within a given analysis of a construction, the meaning may not be fully determined. Aspects of meaning are left unspecified, which means that two different perceivers can interpret a single structure in different ways. This too requires great care and attention to detail when designing experiments which aim to be exact.

## 2.5. The observer's paradox

A frequent aim in semantic experiments is to discover how subjects interpret linguistic

input under normal conditions. A constant problem is how experimenters can access this information, because whatever additional task we instruct the subjects to carry out renders the conditions abnormal. For example, if we ask them to choose which one of a pair of pictures illustrates the interpretation that they have gathered, or even if we just observe their eye movements, the very presence of two pictures is likely to make them more aware that more than one interpretation is possible, thus biasing the results. Even a single picture can alter or trigger the accessibility of a reading.

### 2.6. Inherent meaning and inferred meaning

One last linguistic distinction which we should note here is that between the inherent meaning of an expression ("what is said") and the inferred meaning of a given utterance of an expression. This distinction is fundamental in the division of research into meaning into separate fields, but it is in practice very difficult to apply in experimental work, since naïve informants do not naturally differentiate the two. The recent 'literal Lucy' approach of Larson et al. (2010) is a promising solution to this problem; in this paradigm participants must report how 'literal Lucy', who only ever perceives the narrowly inherent meaning of utterances and makes no inferences, would understand example sentences. This distinction is particularly important when an experimental design requires a disambiguation, and extreme care must be taken that its content is not only inferred. For example, in (3), it is implicated that every rugby player broke one of their own fingers, but this is not necessarily the case. This example cannot thus offer watertight disambiguation.

(3) Every rugby player broke a finger.

*Implication:* Every rugby player broke one of their own fingers.

## 2.7. Experimental measures and the object of theory

As a rule, semantic theory makes no predictions about semantic processing. Instead it concerns itself with the final stable interpretation which is achieved after a whole linguistic expression, usually at the sentence level, has been processed and all reanalyses, for example as a result of garden paths, have been resolved. It fundamentally concerns the stative, holistic result of the processing of an expression, indeed many theoretical approaches regard meaning as only coming about in a full sentence (cf. article 8 (Meier-Oeser) *Emergence of linguistic semantics*). But the processing of a sentence is made up of many steps which are incremental and which interact strongly with each other, partly predicting, partly parsing input as it arrives, partly confirming or revising previous analyses. Much of the experimental evidence available to us provides direct evidence only of these processing steps.

It thus follows that for many semantics practioners much of the empirical evidence which we can gather concerns at best our *predictions* about what the sentence is going to mean, not really aspects of its actual meaning. The time course of our arriving at a particular reading, whether it be remote or readily accessible, has no direct implications for the theory, since this makes no predictions about processing speed (cf. Phillips & Wagers 2007). One aim of this article is to show that experimental techniques can deliver data which can contribute to theory building.

## 2.8. Categorical predictions and gradient data

Predictions of semantic theories typically concern the *availability* of particular interpretations. Experiments deliver more fine-grained data that reflect the relative

preferences among the interpretations. Mapping these gradient data onto the categorical predictions, that is, drawing the line between still available and impossible readings is a non-trivial task. At the same time, the ability to distinguish preferences among the "intermediate" interpretations may be highly relevant for testing predictions concerning readings that fall between the clearly available and the clearly impossible.

### 2.9. Outlook

In the remainder of this paper we will discuss two ways in which systematically collected experimental data can contribute to semantic theorizing. We will use quantifier scope as an example of a phenomenon where results of psycholinguistic experiments can make significant contributions to the theoretical discussions. We will not attempt to review here the considerable psycholinguistic literature on the processing of quantifiers (for a comprehensive survey cf. article 103 (Frazier) *Meaning in psycholinguistics*). Instead we will concentrate on a small set of studies that show the usefulness of end-of-sentence judgements in establishing the available interpretations of quantified sentences. Then we will sketch an experiment to address aspects of the unfolding interpretation of quantifier scope which are of interest to theoretical semanticists as well.

## 3. Off-line evidence for scope interpretation

Semantic theories are typically based on introspective judgements of a handful of theoreticians. The judgements concern available readings of a sentence, possibly ranked as to how easily available these readings are. Not surprisingly, judgements of this sort are subtle and often controversial. For instance, the sentence *Everyone loves*

*someone* has been alternately considered to only allow the wide-scope universal reading (e.g. Hornstein 1995; Beghelli & Stowell 1997) or to be fully ambiguous (May 1977, 1985; Hornstein 1984; Higginbotham 1985). Example (2) above illustrates the same point. Park (1995) and Hobbs & Shieber (1987) disagree about the number of available readings.

The data problem has been known for a long time. Studies as early as Ioup (1975) and VanLehn (1978) tried to consider the intuitions of naïve speakers in developing an empirically motivated theory. However, it has been clear from the beginning that "obvious" tasks such as paraphrasing a presumably ambiguous doubly-quantified sentence or asking informants to choose a (preferred) paraphrase is rather complex and that linguistically untrained participants may not be able to carry them out reliably.

Another purely linguistic task has been problematic for a different reason. Researchers have tried to combine the quantified sentence with a disambiguating continuation, as in (4).

(4) Every kid climbed a tree.

(a) The tree was full of apples.

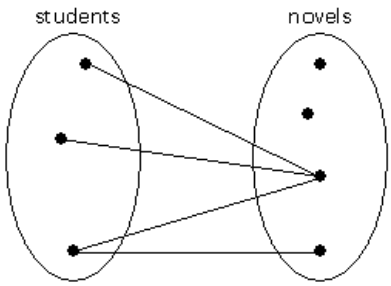
(b) The trees were full of apples.

Disambiguation of this type was used by Gillen (1991), Kurtzman & MacDonald (1993), Tunstall (1998) and Filik, Paterson & Liversedge (2004), for instance. Here the plural continuation is only acceptable if multiple trees are instantiated, that is, the wide-scope universal interpretation is chosen, whereas the singular continuation is intended to only fit the wide-scope existential interpretation. Unfortunately the singular continuation fails to disambiguate the sentence, as Tunstall (1998) points out: *the tree* (4b) can easily be taken to mean *the tree the kid climbed*, thus making it compatible

with the wide-scope universal interpretation as well (see also Bott & Radó 2007 and article 103 (Frazier) *Meaning in psycholinguistics*).

Problems of these kinds have prompted researchers to look for non-linguistic methods of disambiguation. Gillen (1991) used, among other methods, simple pictures resembling set diagrams. In her experiments subjects either drew diagrams to represent the meaning of quantified sentences, chose the diagram that corresponded to the (preferred) reading or judged how well the situation depicted in the diagram fitted the sentence. Bott & Radó (2007) tested a somewhat modified form of the last of these methods using diagrams like those in Figure 15.1. to see whether they constitute a reliable mode of disambiguation that naïve informants can use easily. They found that participants consistently delivered the expected judgements both for scopally unambiguous quantified sentences (i.e. sentences where one scope reading was excluded due to an intervening clause boundary) and for ambiguous quantified sentences where expected preferences could be determined based on theoretical considerations and corpus studies. These results show that there is no a priori reason to exclude the judgements of non-linguist informants from consideration.

A) exactly one > each




B) each > exactly one

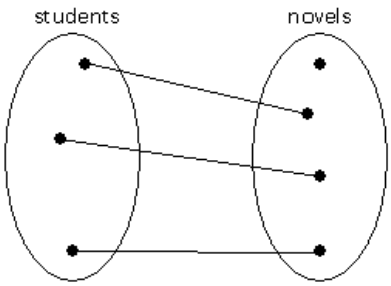



Figure 15.1: *DISAMBIGUATING DIAGRAMS FOR THE SENTENCE Exactly one novel was read by each student.*

For informative experiments, however, we need to be able to derive testable hypotheses based on existing semantic proposals. Although semantic theories are not formulated to make predictions about processing, it is still possible to identify areas where different approaches lead to different predictions concerning the judgement of particular constructions. The interpretation of quantifiers provides an example here as well.

One possible way of classifying theories of quantifier scope has to do with the way different factors are supposed to affect the scope properties of quantifiers. In configurational models such as Reinhart (1976, 1978, 1983, 1995) and Beghelli & Stowell (1997), quantifiers move to/are interpreted in different structural positions. A quantifier higher in the (syntactic) tree will always outscope lower ones. The absolute position in the tree is irrelevant; what matters is the position relative to the other quantifier(s). While earlier proposals only considered syntactic properties of quantifiers, Beghelli and Stowell also include semantic factors in the hierarchy of quantifier positions. Taking *distributivity* as an example, assuming that a +dist quantifier is interpreted in Spec,QP which is the highest position available for quantifiers, Q1 will

outscope Q2 if only Q1 is +dist, regardless of what other properties Q1 or Q2 may have. An effect of other factors will only become apparent if neither of the quantifiers is +dist. By contrast, the basic assumption in multi-factor theories of quantifier scope is that each factor has a certain amount of influence on quantifier scope regardless of the presence or absence of other factors (cf. Ioup 1975; Kurtzman & MacDonald 1993; Kuno 1991 and Pafel 2005). The effects of different factors can be combined, resulting in greater or lesser preference for a particular interpretation. Theories differ in whether one of the readings disappears when it is below some threshold, or whether sentences with multiple quantifiers are always necessarily ambiguous.

Let us assume that the two scope-relevant factors we are interested in are distributivity and discourse-binding, the latter indicated by the partitive NP *one of these N*, see (6). Crossing these factors yields four possible combinations: +dist/+d-bound, +dist/-d-bound, -dist/+d-bound, and -dist/-d-bound. In a configurational theory presumably there will be a structural position reserved for discourse-bound phrases. Let us consider the case where this position is lower than that for +dist, but higher than the lowest scope position available for quantifiers. Thus Q1 should outscope Q2 in the first two configurations, Q2 should outscope Q1 in the third, and the last one may in fact be fully scope ambiguous unless some additional factors are at play as well. Moreover, as configurational theories of scope have no mechanism to predict relative strength of scope preference, the first two configurations should show the same size preference for a wide-scope interpretation of Q1. In statistical terms, we expect an interaction: d-binding should have an effect when Q1 is -dist, but not when it is +dist.

In multi-factor theories, on the other hand, the prediction would usually be that the effects of the different factors should add up. That is, the difference in scope bias

between a d-bound and a non-d-bound +dist quantifier should be the same as between a d-bound and a non-d-bound -dist quantifier. A given factor should be able to exert its influence regardless of the other factors present.

Bott and Radó have been testing these predictions in on-going work. In two questionnaire studies subjects read doubly-quantified German sentences and used magnitude estimation to indicate how well disambiguating set diagrams fitted the interpretation of the sentence. Experiment 1 manipulated distributivity and linear order and used materials like (5). Experiment 2 tested the factors distributivity and d-binding using sentences like (6).

(5) a. Genau einen dieser Professoren haben alle Studentinnen verehrt.
Exactly one these professors$_{acc}$ have all female students adored.
*All female students adored exactly one of these professors.*

b. Genau einen dieser Professoren hat jede Studentin verehrt.
Exactly one these professors$_{acc}$ has each female students adored.
*Each female student adored exactly one of these professors.*

c. Alle Studentinnen haben genau einen dieser Professoren verehrt.
All female students have exactly one these professors$_{acc}$ adored.
*All female students adored exactly one of these professors.*

d. Jede Studentin hat genau einen dieser Professoren verehrt.
Each female student has exactly one these professors$_{acc}$ adored.
*Each female student adored exactly one of these professors.*

(6) a. Genau einen Professor haben alle diese Studentinnen verehrt.
Exactly one professor$_{acc}$ have all these female students adored.
*All of these female students adored exactly one professor.*

b. Genau einen dieser Professoren haben alle Studentinnen verehrt.
Exactly one these professors$_{acc}$ have all female students adored.
*All female students adored exactly one of these professors.*

c. Genau einen Professor hat jede dieser Studentinnen verehrt.
Exactly one professor$_{acc}$ has each these female students adored.
*Each of these female students adored exactly one professor.*

d. Genau einen dieser Professoren hat jede Studentin verehrt.
Exactly one these professors$_{acc}$ has each female student adored.
*Each female student adored exactly one of these professors.*

Bott and Radó found clear evidence for the influence of all three factors. The distributive quantifier *jeder* took scope more easily than *alle*, d-binding of a quantifier and linear precedence both resulted in a greater tendency to take wide scope. Crucially, the effects were additive, which is compatible with the predictions of multi-factor theories but unexpected under configurational approaches.

These results show that even simple questionnaire studies can deliver theoretically highly relevant data. This is particularly important in an area like quantifier scope, where the judgements are typically subtle and not always accessible to introspection. Of course the study reported here cannot address all possible questions concerning the interpretation of quantified sentences like those in (5)-(6). It cannot for example clarify whether the processor initially constructs a fully specified representation of quantifier scope or whether it first builds only a underspecified structure which is compatible with both possible readings, an outstanding question of much current interest in semantics. The data that we have presented so far is off-line, in that it measures preferences only at the end of the sentence, when its content has been disambiguated. In section 5 we

present an experimental design which will allow investigating the on-going (on-line) processing of scope ambiguities. In the next section we relate the semantic issue of underspecification to experimental data and predictions for on-line processing.

## 4. Underspecification vs. full interpretation

It is generally agreed that syntactic processing is *incremental* in nature (e.g. van Gompel & Pickering 2007) i.e. a full-fledged syntactic representation is assigned to every incoming word. Whether semantic processing is incremental in the strict sense, is far from beyond dispute and still an empirical question. To formulate hypotheses about the time-course of semantic processing, we will now look at the on-going debate in semantic theory on underspecification in semantic representations.

Underspecified semantic representations are a tool intended to handle the problem of ambiguity. The omission of parts of the semantic information allows one single representation to be compatible with a whole set of different meanings (for an overview of underspecification approaches, see e.g. Pinkal, 1999; articles 24 (Egg) *Semantic underspecification* and 110 (Pinkal & Koller) *Semantics in computational linguistics*). It is thus an economic method of dealing with ambiguity in that it avoids costly reanalysis, used above all in computational applications.

Taking the psycholinguistic perspective, one would predict that constructing underspecified representations in semantically ambiguous regions of a sentence avoids processing difficulties in ambiguous regions and at the point of disambiguation.

Underspecification can be contrasted with an approach that assumes strict incrementality and thus immediate full interpretation even in ambiguous regions. This would predict processing difficulties in cases of disambiguations to non-preferred

readings. A candidate for a semantic processing principle guiding the choice of one specified semantic representation would be a complexity-sensitive one (for example: "Avoid quantifier raising" captured in Tunstall's *Principle of Scope Interpretation* 1998 and Anderson's 2004 *Processing Scope Economy*).

In the psycholinguistic investigation of coercion phenomena, the experimental evidence is interpreted along these lines. Processing difficulties at the point of disambiguation are taken as evidence for full semantic interpretation (see e.g. Piñango, Zurif & Jackendoff 1999; Todorova, Straub, Badecker & Frank 2000) whereas the lack of measurable effects is seen as support for an underspecified semantic representation (see e.g. Pylkkänen & McElree 2006; Pickering, McElree, Frisson, Chen & Traxler 2006).

Analogously, in the processing of quantifier scope ambiguities, experimental evidence for processing difficulties at the point of disambiguation will be interpreted as support for full interpretation. However, this need not be taken as final. If we look at underspecification approaches in semantics, non-semantic factors are mentioned which might explain (and predict) difficulties in processing local scope ambiguities (see article 24 (Egg) *Semantic underspecification*, section 6.4.1.). And these are exactly the factors which are assumed by multi-factor theories to have an impact on quantifier scope: syntactic structure and function, context, and type of quantifier. The relative weighting and interaction of these factors are not made fully explicit, however.

For the full picture, it would be necessary to examine not only the point of disambiguation but also the ambiguous part of the input, for it is there that the effects of these factors might be identified. Underspecification is normally only temporary, however, and a full interpretation will presumably be constructed at some stage. This might be recognizable for example in behavioural measures, but the precise predictions

of underspecification theory are not always clear. For example, it might be assumed that even representations which are never fully specified by the input signal (or context) do receive more specific interpretations at some later stage. This of course raises the question what domains of interpretation are relevant here (sentence boundary, utterance, ...). In the next section we present experimental work which may offer a starting point for the empirical investigation of such issues.

## 5. On-line evidence for representation of scope

Given the underspecification view, relative scope should remain underspecified as long as neither interpretation is forced. Indeed there should not even be any preference for one reading. The results of the questionnaire studies reported in Section 3 already indicate that this view cannot be right: A particular combination of factors was found to systematically support a certain reading. Furthermore it is unlikely that the task itself introduced a preference towards one interpretation -- although the diagram representing the wide-scope existential reading was somewhat more complex, this did not seem to interfere with participants' performance. The observed preferences must thus be due to the experimental manipulation. That is, even if all possible interpretations are available up to the point where disambiguating information arrives, there must be some inherent ranking of the various scope-determining factors that results in certain interpretations being more activated than others.

Off-line results such as those discussed above are thus equally compatible with two different explanations; one where quantifier scope is fully determined (at least) by the end of the sentence, and another one where several (presumably all combinatorially possible) interpretations are available but weighted differently. A different

methodology is needed to find out whether there is any psycholinguistic support for an underspecified view of quantifier scope.

As it turns out, the currently existing results of on-line studies cannot distinguish the two alternatives, either. In on-line experiments a scope-ambiguous initial clause is followed by a second one that is only compatible with one scope reading. An indication of difficulty during the processing of the second sentence is typically taken as evidence that the disambiguation is incompatible with the (sole) interpretation that had been entertained up to that point. However, there is another way to look at such effects. When the disambiguation is encountered, the underspecified representation needs to be enriched to allow only one reading and exclude all others. It is conceivable that updating the representation may require more or less effort depending on the ultimate interpretation that is required.

This situation poses a dilemma for researchers investigating the interpretation of quantifier scope. If explicit disambiguation is provided we can only test how easily the required reading is available -- the results don't tell us what other reading(s) may have been constructed. Without explicit disambiguation, however, reading time (or other) data cannot be interpreted, since we do not know what reading(s) the participants had in mind.

Bott & Radó (2009) approached this problem using eye-tracking while participants read ambiguous sentences and then asking them to report the interpretation they computed. Although the results they got are only partly relevant for the underspecification debate, we will describe the experiment in some detail, since it provides a good starting point for a more conclusive investigation. We will then sketch a modification of the method that makes it possible to avoid some problems with the original study.

The scope-ambiguous sentences in Bott and Radó's study were instructions like those in (7):

(7) a. Genau ein Tier auf jedem Bild sollst du nennen!
Exactly one animal on each picture should you name!
*Name exactly one animal from each picture!*

b. Genau ein Tier auf allen Bildern sollst du nennen!
Exactly one animal on all pictures should you name!
*Name exactly one animal from all pictures!*

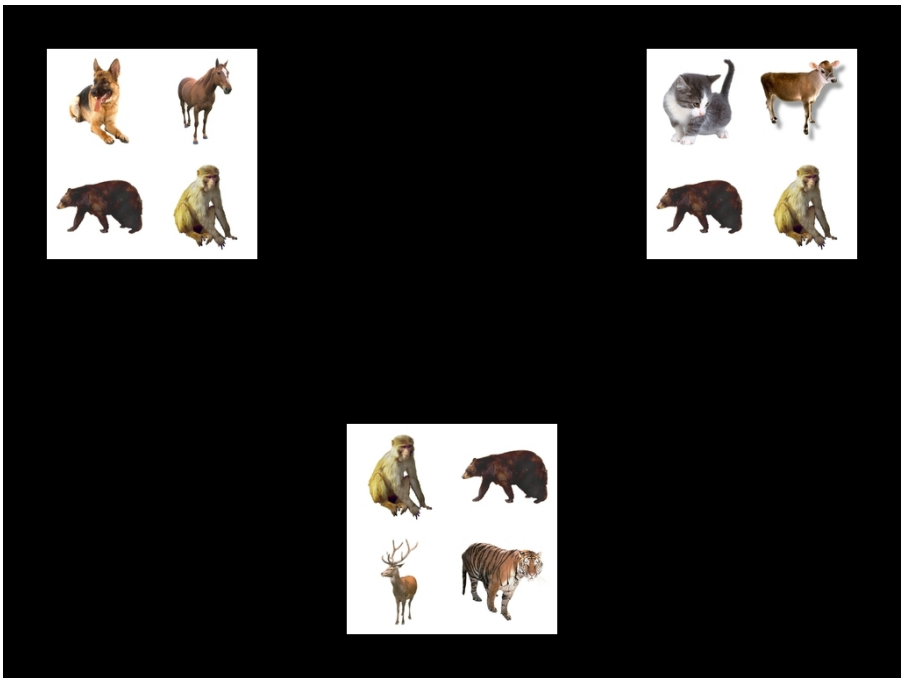


Figure 15.2: *Display following inverse linking constructions.*

The first quantifier was always the indefinite *genau ein* "exactly one". Q2 was either distributive (*jeder*) or not (*alle*). In one set of control conditions Q1 was replaced by a definite NP (*das Tier* "the animal"). In another set of control conditions the two possible interpretations of (7) (one animal that is present in all fields vs. a possibly

different animal from each field on a display) were expressed by scope-unambiguous quantified sentences, as in (8).

(8) a. Name exactly one animal that is found on all pictures.

b. From each picture name exactly one animal.

In each experimental trial participants first read one of these instruction sentences and their eye-movements were monitored. Then the instruction sentence disappeared and a picture display as in Figure 15.2. replaced it. Participants inspected this and had to provide an answer within four seconds. Displays were constructed to be compatible with both possible readings: a wide-scope universal one where different animals can be selected from each field, as well as a wide-scope existential one where a particular animal appeared in all fields (e.g. the monkey in Figure 15.2.). To make the quantifier *exactly one* felicitous, the critical displays always allowed two potential answers for the wide-scope existential interpretation.

The scope-ambiguous instructions were so-called inverse linking constructions, in which the two quantifiers are contained within one NP. It has been assumed (e.g. May & Bale 2006) that in inverse linking constructions the linearly second quantifier preferentially takes scope over the first. The purpose of the study was to test this prediction and to investigate to what extent the distributivity manipulation is able to modulate it. Based on earlier results (Bott & Radó 2007) it was assumed that *jeder* would prefer wide scope, which should further enhance the preference for the inverse reading. When *alle* occurred as Q2, there should be a conflict between the preferences inherent to the construction and those arising from the particular quantifiers.

The experimental setup made it possible to look at both the process of computing the relative scope of the quantifiers (eye-movement behavior while reading the instructions)

and at the final interpretation (the answer participants gave) without providing any disambiguation. Thus the answers could be taken to reflect the scope preferences at the end of the sentence, whereas processing difficulty during reading would serve as an indication that scope preferences are computed at a point where no decision is yet required.

The off-line answers showed the expected effects. There was an overall preference for the inverse scope reading, which was significantly stronger with *jeder* than with *alle*. Crucially, the reading time data showed clear evidence of a conflict between the scope factors: there was a significant slow-down at the second quantifier in (7b). The effect was present already in first-pass reading times suggesting that scope preferences were computed immediately. Bott and Radó interpret these results as strong indication that readers regularly disambiguate sentences during normal reading.

However, this conclusion may be too strong. In Bott and Radó's experiment participants had to choose a particular interpretation in order to carry out the instructions (i.e. *name an animal*). Although they did not have to settle on that interpretation while they were reading the instruction, they had to make a decision as to the preferred reading immediately after the end of the sentence. This may have caused them to disambiguate constructions that are typically left ambiguous during normal interpretation.

Moreover, the instructions used in the experiment were highly predictable in structure: they always contained a complex NP with two quantifiers (experimental items), a definite NP1 followed by a quantified NP2 (fillers A), or else an unambiguous sentence with two quantifiers. Although the content of NP1 (animal, vehicle, flag) and distributivity of Q2 was varied, the rest of the instruction was the same: *sollst du nennen*

"you should name". This pattern was easy to recognize and may have resulted in a strategy of starting to compute the scope preferences as soon as the second NP had been received. To rule out this explanation Bott and Radó compared responses provided in the first and the last third of each experimental session and failed to find any indication of strategic behavior. Still the possibility remains that consistent early disambiguation in the experiment resulted from the task of having to choose a reading quickly in order to provide an answer. The ultimate test of underspecification would have to avoid such pressure to disambiguate fast.

We propose a modification of Bott and Radó's experiment that may not only avoid this pressure but actually encourage participants to delay disambiguation. In the proposed experiment participants will have to judge the accuracy of sentences like those in (9):

(9) a. Genau eine geometrische Form auf allen Bildern ist rechteckig.
Exactly one geometrical shape on all pictures is rectangular.
*Exactly one geometrical shape on all pictures is rectangular.*

b. Genau eine geometrische Form auf jedem Bild ist rechteckig.
Exactly one geometrical shape on each picture is rectangular.
*Exactly one geometrical shape on each picture is rectangular.*

A) wide scope existential disambiguation B) wide scope universal disambiguation

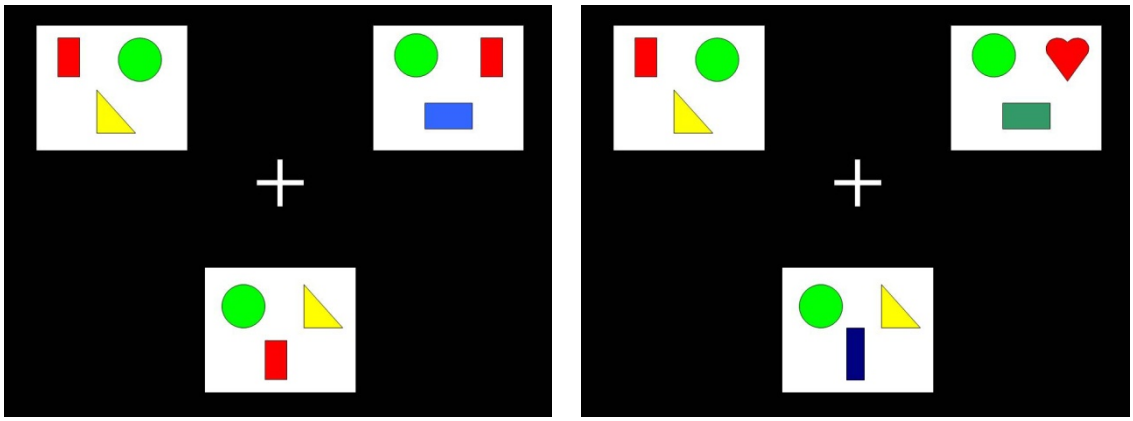


Figure 15.3: *Disambiguating displays in the proposed experiment*

The experiment procedure is as before. The sentences will be paired with unambiguous displays supporting either the wide-scope universal or the wide-scope existential reading (Figure 15.3.). In (9) full processing of the semantic content is not possible until the critical information (*rechteckig*) has been received. Since the display following the sentence is only compatible with one reading which the participant cannot anticipate, they are better off waiting to see which interpretation will be required for the answer. If underspecification is indeed the preferred strategy, there should be no difference in reading times across the different conditions, nor should there be any difficulty in judging any kind of sentence-display pair. Assuming immediate full specification of scope, however, we would expect the same pattern of results as in Bott and Radó's study: slower reading times in (9a) than in (9b) at the second quantifier, as well as slower responses to displays requiring the wide-scope existential interpretation, the latter presumably modulated by *distributivity* of Q2.

The experiment sketched above would be able to distinguish intermediate positions between the two extremes of complete underspecification and immediate full

interpretation. It is conceivable, for instance, that scope interpretation is only initiated when the perceiver can be reasonably sure that they have received all (or at least sufficient) information. This would correspond to the same reading time effects (and same answering behavior) as predicted under immediate full interpretation, but the effects would be somewhat delayed. Another possibility is an initial underspecification of scope, but the construction of a fully specified interpretation at the boundary of some interpretation domain such as the clause boundary. That would predict a complete lack of reading time effects but answer times showing the same incompatibility effects as under versions of the full interpretation approach.

It is worth emphasizing how this design differs from existing studies. First, it looks at the ambiguous region and not just the disambiguation point. Second, it differs from Filik, Paterson & Liversedge (2004), who also measured reading times in the ambiguous region, but who used the kind of disambiguation that we criticized in section 3.

## 6. Conclusions

In this article we have attempted to show that experimentally obtained data can, in spite of certain complicating and confounding factors, be of relevance to semantic theory and provide both support for and in some cases falsification of its assumptions and constructs. In section 2 we noted that the field of theoretical semantics has made less use of experimental verification of its analyses and assumptions. We have seen that there are some quite good reasons for this and laid out what some of the problematic factors are. While some of these are shared to a greater or lesser degree with other

613 branches of linguistics, some of them are peculiar to semantics or are especially severe
614 in this case.
615 The main part of our paper reports a research programme addressing the issue of
616 relative scope in doubly quantified sentences. We present this work as an example of
617 the ways in which experimental approaches can contribute to the development of
618 theory. They also illustrate some of the practical constraints upon such studies. For
619 example, we have seen that clear disambiguation is not always easy to achieve, in
620 particular, it is difficult to achieve without biasing the interpretational choices of the
621 experiment participant. The use of eye-tracking and fully ambiguous picture displays is
622 a real advance on previous practice (Bott & Radó 2009).
623 Section 3 shows how experimental procedures which are simple enough for non-
624 specialist experimenters can nevertheless yield evidence of value for the development of
625 semantic theories: a carefully constructed and counter-balanced design can produce data
626 of sufficient quality to answer outstanding questions with some degree of finality. In
627 this particular case the configurational account of scope can be seen as failing to
628 account for data that the multi-factor account succeeds in capturing. The unsupported
629 account is demonstrated to need adaptation or development. Experimentation can make
630 the field of theory more dynamic and adaptive; an account which repeatedly fails to
631 capture evidence gathered in controlled studies and which cannot economically be
632 extended to do so will eventually need to be reconsidered.
633 In section 5 we lay out some experimental designs to provide evidence which
634 distinguishes between two accounts (section 4) of the way that perceivers deal with
635 ambiguity in the input signal: Underspecification vs. Full Interpretation. This is an
636 example of how processing data can under certain circumstances provide decisive

evidence which distinguishes between theoretical accounts. While it is often the case that theory does not make any direct predictions about psycholinguistically testable measures of processing, this is not always the case, and it may require the collaboration of psycholinguists and semanticists to make these apparent.

We therefore argue for experimental linguists and semanticists to cooperate more and take more notice of each other's work for their mutual benefit. Semanticists will gain additional ways to falsify theoretical analyses or aspects of them, which can deliver a boost to theory development. This will be possible, because experimenters can tailor experimental methods, tasks, and designs to their specific requirements.

Experimenters for their part will benefit by having the questioning eye of the semanticist look over their experimental materials, which will surely avoid many experiments being carried out whose materials fail to uniquely fulfill the requirements of the design. An example of this is the mode of disambiguation which we discussed in section 3. Further to this, experimenters will doubtless be able to derive more testable predictions from semantic theories, if they discuss the finer workings of these with specialist semanticists. We might mention here the example of semantic underspecification: can we find evidence for its psychological reality? Further questions might be: if some feature of an expression remains underdetermined by the input, how long can the representation remain underspecified? Is it possible for a final representation of a discourse to have unspecified features and nevertheless be fully meaningful?

We conclude, therefore, that controlled experimentation can provide a further source of evidence for semantics. This data can under certain circumstances give a more detailed picture of the states of affairs which theories aim to account for. This additional

661 evidence could be the catalyst for some advances in semantic theory and explanation, in
662 the same way that it has in syntactic theory.

## 7. References


Anderson, Catherine 2004. *The Structure and Real-time Comprehension of Quantifier Scope Ambiguity*. Ph.D. dissertation. Northwestern University.

Aoun, Joseph & Yen-hui Audrey Li 1989. Scope and constituency. *Linguistic Inquiry* 16, 623–637.

Beghelli, Filippo & Tim Stowell 1997. Distributivity and negation: The syntax of *each* and *every*. In: A. Szabolcsi (ed.). *Ways of Scope Taking*. Dordrecht: Kluwer, 71–107.

Bott, Oliver & Janina Radó 2007. Quantifying quantifier scope. In: S. Featherston & W. Sternefeld (eds.). *Roots. Linguistics in Search of its Evidential Base*. Berlin/New York: Walther de Gruyter, 53–74.

Bott, Oliver & Janina Radó 2009. How to provide exactly one interpretation for every sentence, or what eye movements reveal about quantifier scope. In: S. Winkler & S. Featherston (eds.). *The Fruits of Empirical Linguistics, Volume 1: Process.* Berlin/New York: Walther de Gruyter, 25–46.

Chomsky, Noam 1965. *Aspects of the Theory of Syntax*. Cambridge, Mass.: The MIT Press.

Filik, Ruth, Kevin B. Paterson & Simon P. Liversedge 2004. Processing doubly

quantified sentences: Evidence from eye movements. *Psychonomic Bulletin & Review* 11(5), 953–959.

Gillen, Kathryn 1991. *The Comprehension of Doubly Quantified Sentences*. Ph.D. Dissertation. University of Durham.

van Gompel, Roger P.G., & Martin J. Pickering 2007. Syntactic parsing. In: G. Gaskell (ed.). *The Oxford Handbook of Psycholinguistics*. Oxford: Oxford University Press. 455–504.

Hahne, Anja & Angela D. Friederici 2002. Differential task effects on semantic and syntactic processes as revealed by ERPs. *Cognitive Brain Research* 13, 339–356.

Higginbotham, James 1985. On semantics. *Linguistic Inquiry* 16, 547–594.

Hobbs, Jerry & Stuart M. Shieber 1987. An algorithm for generating quantifier scopings. *Computational Linguistics* 13, 47–63.

Huang, Cheng-Teh James 1982. *Logical Relations in Chinese and the Theory of Grammar*. PhD dissertation. MIT.

Hornstein, Norbert 1984. *Logic as Grammar*. Cambridge, MA: The MIT Press.

Hornstein, Norbert 1995. *Logical Form: From GB to Minimalism*. Oxford: Blackwell.

Ioup, Georgette 1975. *The Treatment of Quantifier Scope in Transformational Grammar*. Ph.D. dissertation. University of New York.

Kuno, Susumu 1991. Remarks on quantifier scope. In: H. Nakajima (ed.). *Current English Linguistics in Japan*. Berlin: Mouton de Gruyter, 261–287.

Kurtzman, Howard S. & Maryellen C. MacDonald 1993. Resolution of quantifier scope ambiguities. *Cognition* 48, 243–279.

Larson, Meredith, Ryan Doran, Yaron McNabb, Rachel Baker, Matthew Berends, Alex Djalali & Gregory Ward 2010. Distinguishing the said from the implicated using a novel experimental paradigm. In: U. Sauerland & K. Yatsushiro (eds.). *Semantics and Pragmatics. From Experiment to Theory*. Houndmills: Palgrave Macmillan.

May, Robert 1977. *The Grammar of Quantification*. Ph.D. dissertation. MIT. Reproduced in 1982. Indiana University Linguistics Club.

May, Robert 1985. *Logical Form: Its Structure and Derivation*. Cambridge, MA: The MIT Press.

May, Robert & Alan Bale 2006. Inverse linking. In: M. Everaert & H. van Riemsdijk (eds.). *Blackwell Companion to Syntax*. Oxford: Blackwell, chap. 36, 639–667.

Pafel, Jürgen 2005. *Quantifier scope in German*. Vol. 84 of *Linguistics Today*. Amsterdam: John Benjamins.

Park, Jong C. 1995. Quantifier scope and constituency. In: H. Uszkoreit (ed.). *Proceedings of the 33rd Annual Meeting of the Association of Computational Linguistics.* Boston, Palo Alto, CA: Morgan Kaufmann, 205–212.

Phillips, Colin & Matthew Wagers 2007. Relating structure and time in linguistics and psycholinguistics. In: M.G. Gaskell (ed.). *The Oxford Handbook of Psycholinguistics*. Oxford: Oxford University Press, 739–756.

Pickering, Martin J., Brian McElree, Steven Frisson, Lilian Chen & Matthew J. Traxler 2006. Underspecification and aspectual coercion. *Discourse Processes* 42(2), 131–155.

Piñango, Maria, Edgar Zurif & Ray Jackendoff 1999. Real-time processing implications of enriched composition at the syntax-semantics interface. *Journal of Psycholinguistic Research* 28, 395–414.

Pinkal, Manfred 1999. On semantic underspecification. In: H. Bunt & R. Muskens (eds.). *Computing Meaning,* Dordrecht: Kluwer, 33–55.

Pylkkänen, Liina & Brian McElree 2006. The syntax-semantics interface: On-line composition of meaning. In: M. A. Gernsbacher & M. Traxler (eds.). *Handbook of Psycholinguistics*, 2. Edition. New York: Elsevier, 537–577.

Reinhart, Tanya 1976. *The Syntactic Domain of Anaphora*. Ph.D. dissertation. MIT.

Reinhart, Tanya 1978. Syntactic domains for semantic rules. In: F. Guenthner & S.J. Schmidt (eds.). *Formal Semantics and Pragmatics for Natural Language*. Dordrecht: Reidel, 107–130.

Reinhart, Tanya 1983. *Anaphora and Semantic Interpretation*. London/Sydney: Croom Helm.

Reinhart, Tanya 1995. *Interface Strategies*. OTS Working Papers in Linguistics.

Todorova, Marina, Kathy Straub, William Badecker & Robert Frank 2000. Aspectual coercion and the online computation of sentential aspect. In: L. R. Gleitman & A. K. Joshi (eds.). *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Mahwah, N.J.: Lawrence Erlbaum Associates, 3–8.

Ross, John R. 1970. On declarative sentences. In: R. Jacobs & P. Rosenbaum (eds). *Readings in English Transformational Grammar*. Waltham, Massachusetts: Ginn, 222–272.

Tunstall, Susanne L. 1998. *The Interpretation of Quantifiers: Semantics and Processing*. PhD. Dissertation. University of Massachussetts Amherst.

Weskott, Thomas & Gisbert Fanselow 2009. Scaling issues in the measurement of linguistic acceptability. In: S. Winkler & S. Featherston (eds.). *The Fruits of Empirical Linguistics, Volume 1: Process.* Berlin/New York: Walther de Gruyter, 229–246.

VanLehn, Kurt A. 1978. *Determining the Scope of English Quantifiers*. Technical Report (AI-TR 483). Artificial Intelligence Laboratory, MIT.

Zhou, Peng & Liqun Gao 2009. Scope processing in Chinese. *Journal of Psycholinguistic Research* 38, 11–24.

*Oliver Bott, Tübingen (Germany)*

*Sam Featherston, Tübingen (Germany)*

*Janina Radó, Tübingen (Germany)*

*Britta Stolterfoht, Tübingen (Germany)*