Embedded Scalars, Preferred Readings & Prosody: An Experimental Revisit

Michael Franke, Fabian Schlotterbeck & Petra Augurzky
Seminar für Sprachwissenschaft & SFB 833
Eberhard Karls Universität Tübingen

Abstract

The scalar item *some* is widely assumed to receive a meaning enrichment to *some but not all* if it occurs in matrix position. The question under which circumstances this enrichment can occur in certain embedded positions plays an important role in deciding how to delineate semantics and pragmatics. We present new experimental data that bear on this theoretical issue. In distinction to previous experimental approaches, we presented sentence material auditorily in order to explicitly control prosodic markedness of the scalar item. Moreover, our experiment sheds light on the relative preferences or salience of candidate readings. The presented data turn out to be unexpected under a traditional Gricean view, but also challenge the idea of disambiguation by logical strength in grammaticalist approaches.

1 Introduction

The existential quantifier *some* is usually assumed to receive a semantic interpretation similar to logical $\exists$, so that (1a) is literally true even when Hans solved all of the problems. But use of *some* is also usually considered to invite comparison with (at least) the semantically stronger universal quantifier *all* (c.f. Horn, 1972; Gazdar, 1979; Atlas and Levinson, 1981). This scalar comparison can lead to an upper-bound meaning enrichment, e.g., when an utterance of (1a) is taken to invite the inference in (1b).

(1) a. Hans solved some of the problems.
    b. $\rightsquigarrow$ Hans solved some but not all of the problems.
    c. Hans solved all of the problems.

The classical explanation of this inference, following the pioneering work of Grice (1975), is that (1b) is a pragmatic inference, a so-called *quantity implicature*, derived by an abductive inference to the best explanation of why informed, knowledgable and cooperative speakers would utter (1a) when they could also utter the semantically stronger and relevant (1c) (see Geurts, 2010, for a recent overview).

Whereas the existence of such implicatures is rather uncontroversial for plain occurrences of *some*, it is still an issue of debate whether comparable enrichments are

also found in so-called "embedded" cases, where the scalar item *some* occurs in the scope of other logical operators. This paper deals with two such cases. In AS-sentences (mnemonic for *all ... some ...*) as in (2) the scalar item *some* is embedded under the universal quantifier *all*. In ES-sentences (mnemonic for *exactly one ... some ...*) like (3) *some* takes scope under the non-monotonic quantifier *exactly one*. Current theories, have proposed at least three candidate readings for AS- and ES-sentences: (i) a *literal reading* like in (2a) and (3a) where *some* has only its literal meaning; (ii) a *global reading* like in (2b) and (3b) where the meaning of (2) and (3) is enriched with the negation of the corresponding sentences (4) and (5) respectively; and also (iii) a *local reading* like in (2c) and (3c) where *some* is interpreted as *some but not all* in the scope of the embedding quantifier.

(2) **All** of the students read **some** of the papers. (AS)

   a. **All** of the students read **some and maybe all** of the papers. (AS-LIT)

   b. **All** of the students read **some and maybe all** and it's not the case that **all** of the students read **all** of the papers. (AS-GLB)

   c. **All** of the students read **some but not all** of the papers. (AS-LOC)

(3) **Exactly one** of the students read **some** of the papers. (ES)

   a. **Exactly one** of the students read **some and maybe all** of the papers. (ES-LIT)

   b. **Exactly one** of the students read **some and maybe all** and it's not the case that **exactly one** of the students read **all** of the papers. (ES-GLB)

   c. **Exactly one** of the students read **some but not all** of the papers. (ES-LOC)

(4) **All** of the students read **all** of the papers.

(5) **Exactly one** of the students read **all** of the papers.

Given this plurality of hypothesized readings, the question arises: how *salient* are these theoretically conceivable readings compared to each other, or, put differently, what is the *preference order* on hypothesized readings (if they are detectable by empirical means at all)? Addressing this question empirically is relevant because it lies at the heart of a current debate about the nature of quantity inferences. The controversy is usually seen as one between two main competing theoretical positions, which we call *traditionalism* and *grammaticalism* and take to be general ideas behind large classes of approaches to quantity implicatures (c.f. Chemla and Singh, 2014, for related discussion with a focus on deriving processing theories).

Results from previous empirical studies have been discussed controversially (e.g. Geurts and Pouscoulous, 2009; Clifton and Dube, 2010; Chemla and Spector, 2011; Benz and Gotzner, 2014). This is because of several reasons. Firstly, and most importantly, previous studies presented target sentences visually. However, traditionalists often acknowledge the availability of local readings for prosodically marked utterances (e.g. Horn, 2006; Geurts, 2009; Chemla and Spector, 2011; Geurts, 2010; van Tiel, to appear; Geurts and van Tiel, 2013). Under this view, evidence for the availability

of local readings (see Clifton and Dube, 2010; Chemla and Spector, 2011) may be an effect of "silent prosody" (see e.g. Bader, 1998; Fodor, 1998). We therefore address this position as the *prosodic markedness hypothesis* explicitly. Secondly, as van Tiel (to appear) and Geurts and van Tiel (2013) argue, the alleged evidence for the availability of local readings of AS-sentences, might well be confounded with, in particular, "typicality effects" (see Section 4 for discussion). Finally, previous studies have only accumulated limited evidence pertaining to the relative salience of candidate readings of sentences like (2) and (3) (see Section 4 for discussion).

To deal with these methodological concerns, this paper presents results from an *incremental verification task*. In order to address the problem of silent prosody, we presented sentence materials auditorily and explicitly manipulated contrastive stress on embedded scalar items. Moreover, to address worries about pictorial "typicality effects," parts of a to-be-evaluated picture were hidden and only revealed piece-wise at the participants' request until they felt able to give a binary truth-value judgement (see Conroy, 2008). In order to obtain information about salience of readings, we use both a classical regression analysis and a Bayesian statistical model to analyze the data from this novel task, and validate this method by including control conditions testing ambiguous sentences with a well-known preference on available readings.

The paper is structured as follows. Section 2 elaborates on the three kinds of relevant readings for our target sentences. Section 3 works out testable *ex ante* predictions of *traditionalism* and *grammaticalism*. Section 4 reviews related experimental studies to the extent necessary to motivate our own approach. Section 5 describes our experimental design, analyses and results. Finally, Section 6 discusses our method and interpretation of the results in a broader theoretical context.

2 Get to know your readings

Three readings are *prima facie* conceivable for the AS- and ES-sentences in (2) and (3). These are logically dependent in intricate ways.

AS-sentences. An AS-sentence like (2), repeated below, has a literal reading (LIT) as in (2a), a global reading (GLB) as paraphrased in (2b) and a local reading (LOC) as in (2c).

(2) **All** of the students read **some** of the papers.

a. **All** of the students read **some and maybe all** of the papers. (AS-LIT)

b. **All** of the students read **some and maybe all** and it's not the case that **all** of the students read **all** of the papers. (AS-GLB)

c. **All** of the students read **some but not all** of the papers. (AS-LOC)

These readings stand in a strict entailment relation: the local reading asymmetrically entails the global reading, which asymmetrically entails the literal reading:

(6) LOC $\subset$ GLB $\subset$ LIT

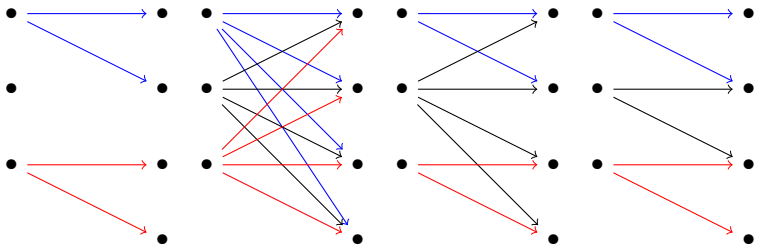
situation	reading		
	LIT	GLB	LOC
false	0	0	0
literal	1	0	0
weak	1	1	0
strong	1	1	1

(a) AS-sentences

situation	reading		
	LIT	GLB	LOC
false	0	0	0
literal	1	0	0
local	0	0	1
all	1	1	1

(b) ES-sentences

Table 1: Possible truth-value distributions for readings of AS- and ES-sentences



(a) false (b) literal (c) weak (d) strong

Figure 1: Examples of distinguishing situations for AS-sentences. The dots on the left represent students, the dots on the right papers. An arrow from left to right represents that a student has read a paper.

GLB entails LIT because, in general, global readings are defined as the conjunction of the literal reading and the negated (relevant/feasible) alternative(s) of the to-be-interpreted utterance. This entailment is asymmetric, because the information that not all of the students read all of the papers is not entailed by the literal reading. To see that LOC $\subset$ GLB, notice that the case where all of the students read some but not all of the papers is a special case of the class of situations where all of the students read some (and maybe all), while not all of the students read all of the papers.

Given these entailment relations, there are four kinds of situations, the names for which we borrow from Chemla and Spector (2011), that we can distinguish based on different truth values for our candidate readings. These are given in Table 1a. Examples of these kinds of situations are shown in Figure 1, where the dots on the right of each diagram represent students, the dots on the left represent papers and an arrow from a student to a paper indicates that the student read the paper.

ES-sentences. The case for ES-sentences like (3) is similar but a little more complicated because the embedding quantifier is non-monotonic. Again, we consider a literal reading as in (3a), repeated below, a global reading as in (3b) and a local reading as in (3c).

(3) **Exactly one** of the students read **some** of the papers.

   a. **Exactly one** of the students read **some and maybe all** of the papers. (ES-LIT)

   b. **Exactly one** of the students read **some and maybe all** and it's not the case that **exactly one** of the students read **all** of the papers. (ES-GLB)

   c. **Exactly one** of the students read **some but not all** of the papers. (ES-LOC)

Entailment relations in this case are non-linearly ordered:

(7) LOC $\supset$ GLB $\subset$ LIT

By definition of global readings, GLB entails LIT. This entailment is asymmetric because the extra information that it is not the case that exactly one student read all of the papers is not entailed by the literal reading (3a). However, unlike for AS-sentences, LOC is not stronger than GLB, but asymmetrically entailed by the latter. To see this, notice that the global reading is equivalent to:

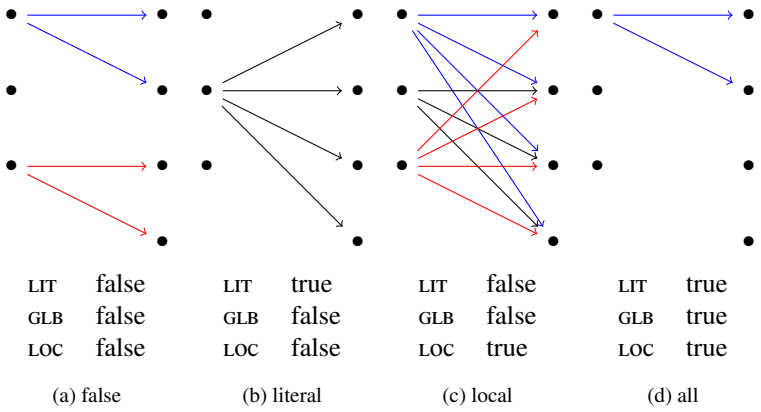
(3b′) **Exactly one** of the students read **some but not all** and **everybody else** read **none** of the papers.

Finally, LOC and LIT are logically independent: all combinations of truth-values for LOC and LIT are possible.

Given these entailment relations, there are again four different situations corresponding to the four possible distributions of truth values for candidate readings. These are given in Table 1b on page 4 and named following Chemla and Spector (2011). Examples for each situation are given in Figure 2.

3 Theories and predictions

There are two main theoretical positions which appear to make different predictions about the readings of AS- and ES-sentences (see Horn, 2006; Geurts, 2010; Sauerland, 2012; Chemla and Singh, 2014, for overview). We will refer to these as *traditionalism* and *grammaticalism* and treat each in turn. Both approaches are addressed here as "core theories" (terminology from Chemla and Singh): two main ideas behind several approaches to quantity implicatures. While traditionalism qua "core theory" makes some non-trivial predictions about salience of readings of AS- and ES-sentences, grammaticalism does not. To derive from grammaticalism a set of falsifiable predictions about outcomes of a concrete behavioral experiment like ours *ex ante*, i.e., before having seen any data from that particular experiment, additional "auxiliary assumptions" are necessary. Here, we focus on one such "auxiliary assumption," for reasons to be



(a) false (b) literal (c) local (d) all

Figure 2: Examples of distinguishing situations for ES-sentences

spelled out below. Section 6 will revisit grammaticalism as a "core theory" *ex post*, i.e., in the light of our data.

Traditionalists frequently concede that local readings are licensed if prosodically marked. We address this *prosodic markedness hypothesis* separately in Section 3.3, alongside a brief recap of the relevant experimental data.

3.1 Traditionalism

We use "traditionalism" as a vague umbrella term for approaches that build on Grice's (1975) original notion that conversational implicatures, of which quantity implicatures are a special case, are to be thought of as rationalizations of speaker behavior. Central in this reasoning is the assumption that the speaker's behavior is efficient (if not optimal) and goal-oriented. Usually, the assumed goal of conversation is the cooperative exchange of relevant information from speaker to hearer.[1]

The traditionalist view gives rise to what Geurts (2010) calls the *Gricean recipe* for deriving (1b) from (1a): if the issue whether Hans solved only some or all of the problems is relevant, then a cooperative and knowledgable speaker would utter (1c) if in a position to do so; hence, one of the most natural explanations of why such a speaker has not uttered (1c), but only (1a) is that she is uncertain of whether (1c) is true; but on the assumption that she is knowledgeable (competent, opinionated, informed ...) it follows that (1b) should in fact be true.[2]

(1) a. Hans solved **some** of the problems.

b. $\rightsquigarrow$ Hans solved **some but not all** of the problems.

[1] Of the more recent literature, we would consider as traditionalist, among many others, contributions by Spector (2006), Sauerland (2004), Russell (2006), Schulz and van Rooij (2006), Geurts (2010), Franke (2011), or Goodman and Stuhlmüller (2013).

[2] We are glossing here somewhat swiftly over the more nuanced details of the derivation of implicatures targeting the speaker's epistemic state (e.g. Gazdar, 1979; Soames, 1982). More on this below.

c. Hans solved **all** of the problems.

Traditionalism's predictions about general availability of reading depend on which alternative utterances we consider. For AS-sentences, traditionalism can derive the global reading by assuming (uncontroversially) that (4) is an alternative. It can derive the local reading by assuming (more controversially) that (8) is a relevant alternative (e.g. Sauerland, 2004).[3]

(8) **Some** of the students solved **all** of the problems.

For ES-sentences, traditionalism can derive the global reading if we assume (controversially) that (5) is a relevant alternative. Since local readings are logically independent of literal readings, there is no way that traditionalism can derive them from the Gricean recipe, no matter what alternatives we consider.

Traditionalism's predictions about preferences among available readings depend on the reliability of contextual information, such as speaker knowledgeability. Specifically, considering a speaker knowledgeable results in a strong non-epistemic implicature that a given alternative is false. By contrast, considering a speaker as ignorant results in a literal reading or a weak epistemic implicature (that the speaker does not know whether the alternative is true). In the context of a given abstract experiment that does not manipulate speaker expertise (like ours), it is then *prima facie* compatible with a traditionalist position to expect literal or implicature readings to be more prominent (see Goodman and Stuhlmüller, 2013, for experimental data on this).

In sum, a traditionalist core theory is compatible with a range of observations about the availability and relative salience of implicature readings. Most importantly for the subsequent discussion, traditionalism does not predict local readings for ES-sentences and can favor implicature readings over literal ones or vice versa.

3.2 Grammaticalism

The central idea behind grammaticalist approaches is that quantity implicatures are derived by application of exhaustivity operators at varying scope-sites during compositional computation of a sentence's truth-value (see Chierchia, 2006; Magri, 2011; Sauerland, 2012; Chierchia et al., 2012). For our purposes, it is enough to assume that $\text{Exh}(\cdot)$ is a poly-typed function that enriches an expression X, which need not be a full proposition, based on a set $\text{Alt}(X)$ of (suitable, relevant) alternatives to X that yields an enriched meaning of the form:

(9) $\text{Exh}(X, \text{Alt}(X)) = X \bigwedge_{A \in \text{Alt}(X)} \neg A$.

A grammaticalist core theory predicts that all three readings of AS- and ES-sentences are in principle available: literal readings arise if no exhaustification operator is applied; global readings arise if the exhaustification operator takes sentence-wide scope as in (10); and local readings arise if the exhaustification operator takes scope under the respective quantifiers as in (11).

(10) a. Exh(**All** of the students solved **some** of the problems).

[3] It is more controversial that (8) is an alternative to (2), because the former does not entail the latter.

b. Exh(**Exactly one** of the students solved **some** of the problems).

(11) a. **All** of the students Exh(solved **some** of the problems).

b. **Exactly one** of the students Exh(solved **some** of the problems).

It is a topic of active current debate as to how a grammaticalist approach should be supplemented with a general and well-motivated *disambiguation criterion* so as to rule out unattested readings obtainable by embedded exhaustification and to make concrete predictions about preferences over readings made available. Several approaches have been suggested in the literature (c.f. Fox, 2007; Magri, 2009; Chemla and Singh, 2014).

To our knowledge, only one of these proposed disambiguation mechanisms is principally able to give predictions about potential outcomes of our experiment *ex ante*, i.e., before having seen any data of that experiment. The *strongest meaning hypothesis* of Dalrymple et al. (1998) involves a disambiguation mechanism by comparison of logical strengths. Supplementing grammaticalism with a strength-based selection criterion has been suggested, in one form or another, in many contributions to this paradigm (e.g. Fox and Spector, 2009; Chierchia et al., 2012; Chierchia, 2013; Spector, 2014). At the same time, the strongest meaning hypothesis has theoretical significance beyond the debate about embedded implicatures and has been frequently called upon for disambiguation purposes in other domains, including reciprocals (Dalrymple et al., 1998), plural predication (Winter, 2001) or vagueness (Cobreros et al., 2012).

Most straightforwardly, disambiguation by logical strength predicts that preferences for readings would just mirror the entailment relations in (6) and (7): for AS-sentences, grammaticalism predicts that local readings are preferred over global readings which in turn are preferred over literal readings; for ES-sentences grammaticalism predicts that global readings are most preferred, while literal and local readings are not ranked with respect to preference.[4]

3.3 The Prosodic Markedness Hypothesis

Although traditionalism does not predict local readings for ES-sentences, many authors who adhere to a traditionalist position propose that local readings can be available if the embedded scalar item is prosodically marked (e.g. Horn, 2006; Geurts, 2009, 2010; Geurts and van Tiel, 2013). This is taken to be continuous with other cases where pragmatic enrichments can take scope under logical operators if the relevant focal accent is present (12).

(12) This chili is not SPICY, it's SPICY.

The *prosodic markedness hypothesis*, as we will call it here, is then an addition to the traditionalist's explanatory repertoire. Under this extra assumption, local readings are a marked phenomenon, and not the product of the same process as run-of-the-mill scalar implicatures. If supplemented with the prosodic markedness hypothesis, traditionalists

[4]Chierchia et al. (2012) consider another variant of disambiguation based on strongest meanings in which parses that differ in more than one absence/presence of silent exhaustification are incomparable. We ignore this for simplicity, because it matters little for the conclusions drawn from our experimental data.

would therefore predict that local readings are available for ES-sentences, but only if prosodically marked by contrastive stress.

To date, empirical investigations of this assumption are still scarce. One notable exception is a pilot study reported in Frazier (2008), in which participants were instructed to read embedded and non-embedded versions of English AS-sentences. In the embedded sentences (e.g. *All of the students wrote some of the official memos*), *some* was either capitalized or non-capitalized. Prior to the experiment, participants were informed that capitalizing corresponded to contrastive stress. The task consisted in a forced-choice questionnaire with two given alternative paraphrases (such as *All of the students wrote some but not all of the official memos* vs. *All of the students wrote at least some of the official memos*). Interestingly, strengthened interpretations occurred rather often across all sentences in this task (59 %), but were not affected by capitalizing.[5] Frazier concludes that accentuation did not affect the silent reading data, but raises the possibility that prosodic effects may have been present at least in a subset of the items (see Footnote 1 in Frazier (2008), p. 330).

Another study relevant for the present considerations comes from Schwarz et al. (2008), examining the effects of contrastive accentuation of *or* in non-embedded sentences like *Mary will invite Fred or/OR Sam to barbecue*. After listening to these sentences, two alternative paraphrases were presented to the participants (e.g. *She will invite Fred or Sam or possibly both* vs. *She will invite Fred or Sam but not both*). The authors found a significant impact of overtly realized contrastive accents on implicature inferences: Besides a general advantage for non-strengthened readings, focal accents significantly increased the strengthened interpretation (from 16.4 % to 28.6 %). These results are compatible with the "focus strengthening hypothesis" proposed by the authors.

To sum up so far, though an effect of contrastive accentuation has been claimed to be essential for traditionalism for triggering local implicatures, experimental evidence is currently mixed, suggesting potential effects of presentation mode (visual vs. auditory), item-specific characteristics, as well as the specific construction under investigation (e.g., embedded cases vs. non-embedded cases such as in Schwarz et al. (2008)). It is for these reasons that an experimental investigation into the availability of readings of AS- and ES-sentences should present sentential material auditorily and explicitly address the prosodic markedness hypothesis.

4 Previous studies

In order to test the divergent predictions of the relevant theoretical positions, a number of empirical studies have been carried out. We will focus on the works of Geurts and Pouscoulous (2009), Clifton and Dube (2010) and Chemla and Spector (2011) (c.f. Benz and Gotzner, 2014, for related discussion). Unfortunately, these studies do not provide conclusive evidence about the availability and preference of relevant readings and do not speak to prosodic markedness effects.

[5] Note that, unfortunately, the paper does not provide percentages for the embedded conditions alone, but only the means across sentence types.

4.1 Geurts and Pouscoulous (2009)

Geurts and Pouscoulous (2009) conducted a picture-verification task experiment (their Experiment 3) to find out whether local readings of AS- and ES-sentences are available.[6] Subjects were presented with pictures similar to those in Figure 1c and 2c (pages 4 and 6) where the local reading gets a different truth-value than the literal and the global reading. For AS-sentences, the local reading is false for the critical picture in Figure 1c whereas the literal and global readings are true; for ES-sentences the local reading is true for the critical picture in Figure 2c, whereas the literal and global readings are false.

The results of Geurts and Pouscoulous were strikingly unambiguous: there were *no* responses indicative of a local reading; *all* of the subjects judged AS-sentences true in a situation like in Figure 1c and *all* of the subjects judged ES-sentences false in situations like 2c.

These results were criticized on theoretical grounds (e.g., Sauerland (2010), but see Ippolito (2010) for support), as well as based on empirical observations (Clifton and Dube, 2010; Chemla and Spector, 2011). In the following, we will review these empirical studies briefly, as they provide the background for our own experiment.

4.2 Clifton and Dube (2010)

In a reply to Geurts and Pouscoulous's study, Clifton and Dube (2010) raised the question whether the use of a picture-verification paradigm might have been infelicitous for testing the availability of strong readings, at least in the case of AS-sentences. Asking whether a sentence fits a picture might have created a bias for accepting sentences also on a weaker and probably dispreferred reading. Clifton and Dube's study was therefore aimed at finding out about a potential preference relation between local and literal readings. To this end, Clifton and Dube developed a picture-choice task where subjects were presented with an AS-sentence and a pair of pictures, hence introducing the option to choose between different alternatives. Subjects were asked to "indicate which shape is best described by the sentence" and could choose either picture, or options 'both' and 'neither.' There were two versions of this experiment, differing in which kind of picture pairs were presented on critical trials. In version 1, the picture pair consisted of the weak and strong situations in Figures 1c and 1d. The response percentages observed by Clifton and Dube were:

weak	strong	both	neither
3	39	57	1

That the majority answer is "both" could be taken as evidence that the literal reading is the preferred one. But the almost 40% of choices for the strong situation, so Clifton and Dube argue, is indicative of the availability of the local reading. In version 2 of the

[6] We will focus here on a subset of the conditions tested by Geurts and Pouscoulous (2009). Actually, these authors did not use ES-sentences, but sentences where scalar *some* was embedded under non-monotonic quantifier *exactly two* (with appropriate pictures, of course). This case is a little more complex, but we will gloss over this here, treating their data, as if it had been obtained for ES-sentences.

experiment, the picture pair consisted of the literal and weak situations in Figures 1b and 1c. In this case, response percentages were:

weak	literal	both	neither
28	6	50	17

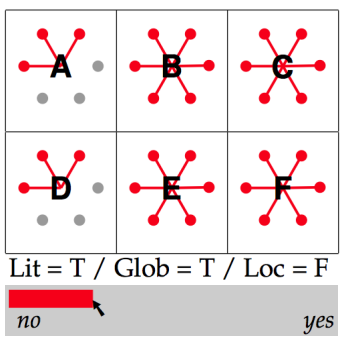
Again, the majority response "both" might speak for a preference for the literal reading, but, as Clifton and Dube propose, the 17% of "neither" answers in this case again suggest that the local reading is available. In sum, Clifton and Dube take these results to contradict Geurts and Pouscoulous's findings. Local readings are, after all, attested if subjects are given a choice as to which situation they consider most fitting for an AS-sentence.

The diverging results of Geurts and Pouscoulous and Clifton and Dube seem to indicate that participants' choices for readings of AS-sentences might have been affected by the specific experimental paradigm used (but see Geurts and van Tiel, 2013, for critique of the latter design). Unfortunately, Clifton and Dube (2010) did not test both universal and non-monotonic quantifiers within the same experiment. Additionally, it would be informative to also probe into the availability of global readings. The study of Chemla and Spector (2011) did both of that.

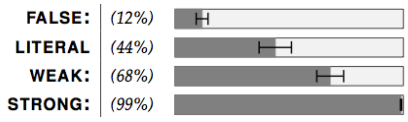
4.3 Chemla and Spector (2011)

Chemla and Spector (2011) also took issue with Geurts and Pouscoulous's design, arguing that, firstly, the pictorial material used in Geurts and Pouscoulous's study was unduly difficult; that, secondly, these pictures also may have failed to make the local reading sufficiently relevant; and that, thirdly, the restriction to a categorial choice (whether the sentence fits the picture or not) may induce a bias against non-preferred readings in cases where candidate readings stand in entailment relations (see Sauerland, 2010, for this latter criticism). To meet these potential problems, Chemla and Spector (2011) presented subjects with pictorial material like that in Figure 3a, which was assumed to be easier to assess and better at highlighting the relevance of the local readings. Albeit in a different format, the pictures used by Chemla and Spector were instantiations of the situation types in Figures 1 and 2. Additionally, subjects were asked, not for categorial judgements, but for *graded judgements*: subjects could freely click on a scale, as shown in Figure 3a, to indicate how much they considered a picture fitting for a given sentence (see Chemla, 2009, for more on this method).

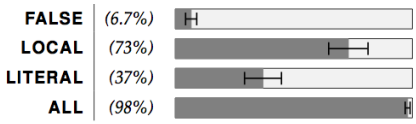
Chemla and Spector hypothesized that the degree to which a sentence is rated acceptable is proportional to the number of available true readings. Observed averaged clicking positions are shown in Figures 3b and 3c. According to Chemla and Spector, the crucial piece of evidence for the availability of local readings for AS-sentences is that these sentences yielded higher graded acceptability scores for the strong situation than for the weak situation (although these differ only with respect to the truth value of the local reading). Evidence for the availability of the local reading for ES-sentences comes from the difference between the local and the literal situation. Strikingly, ES-sentences received an average 73% degree of acceptability in the local situation although the literal and global readings are false in this case.



(a) Example of the AS-weak condition of Chemla and Spector (2011)



(b) Results AS-condition



(c) Results ES-condition

Figure 3: Example trial and results of Chemla and Spector's (2011) study. Test sentences for pictures like on the left where: "Every letter is connected to some of its circles."

4.4 Reflection

Summing up, three experimental studies have presented diverging pieces of evidence. Differences in results might be due to differences in experimental design, in particular due to the type of elicited judgement and the possibility of conflating pictorial effects. Also, we should consider potential effects of "silent prosody."

Judgement Type. We could draw a distinction between categorical truth-value judgments and other potentially more sensitive measures. The former are possibly not sensitive enough to reveal dispreferred readings with small samples, but the latter may be. We could then hypothesize along with Clifton and Dube and Chemla and Spector that local readings are available, but so dispreferred that they do not affect categorical truth-value judgments.

On the other hand, Crain and Thornton (1998) argue in favor of truth-value judgments as a means of detecting dispreferred readings. Consequently, traditionalists could reply that the putative evidence for local readings in allegedly more sensitive tasks might reflect something other than (strongly) truth-relevant speaker-intended meaning enrichments. Hence no case can be made for lexicalism or grammaticalism on the basis of these data (see Geurts and van Tiel, 2013; van Tiel, 2014, to appear, for arguments along these lines).

To resolve this issue, we would ideally like to have a design that elicits categorical truth-value judgements, but is still possibly sensitive enough to detect dispreferred readings.

Pictorial Effects. It is often possible to present different numbers of links between icons in pictures like in Figures 1 and 2, without changing the truth-values of relevant readings. Indeed, the pictorial material used in previous studies differed in this respect, and that might explain some of the differences in results. If that is so, then we cannot

ascribe the effects reported by Clifton and Dube (2010) and Chemla and Spector (2011) to the availability of local readings with certainty. Chemla and Spector (2011) also acknowledge the possibility that the *typicality* of a picture with respect to some meaning may affect graded truth-value judgments. Chemla and Spector show that graded judgments of AS-sentences differ significantly for different pictures that agree on the truth value of relevant readings but differ with respect to the amount of connections between icons. Chemla and Spector suggest that typicality of the pictorial material can account for these differences, but submit that this does not explain away the high acceptance of pictures like Figure 1c.[7]

The latter point is disputed by van Tiel (to appear), who demonstrates that a huge chunk of variance in the responses to AS-sentences can be explained as typicality effects, dispensing the need to appeal to the distribution of different readings. In support of this idea, van Tiel (to appear) elicited what he calls the typicality structures associated with the quantifiers *all* and *some* (as done also by Degen and Tanenhaus, 2011) and predicted the judgments of AS-sentences obtained by Chemla and Spector based on these data. Thereby, he obtained an excellent model fit.

We are generally sympathetic towards van Tiel's innovative line of reinterpretation of the data, but note, as Chemla and Spector (2011) already observed, that his typicality-based explanation does not extend to ES-sentences in an obvious way. Moreover, typicality itself is not a satisfactory primitive in a putative explanation of the use of quantifiers, but rather something that needs explanation itself (c.f. Cummins, 2014, for critical reflection on typicality-based explanations). Indeed, it is possible to explain typicality judgements as the outcome of Gricean speaker preferences for informative descriptions, just like those that rationalize quantity implicatures (Franke, 2014). For these reasons, we believe that typicality and other pictorial effects need to be taken seriously in the experimental design, but do not necessarily give a satisfactory account of the observed variation all by themselves. In other words, if we take worries about pictorial effects seriously, we need to reconsider conclusions from previous studies in the light of data from a task that minimizes pictorial effects as much as possible.

"Silent prosody." Several studies suggest that numerous factors might influence accent placement and prosodic phrasing even while reading, amongst them default accentuation, constituent length, rhythmic phenomena, and individual variation (e.g. Bader, 1998; Fodor, 1998; Steinhauer and Friederici, 2001; Fodor, 2002; Augurzky, 2008; Breen et al., 2011; Kentner, 2012). If we adopt the prosodic markedness hypothesis described in Section 3.3, as many traditionalists do, we predict that the availability of a local reading hinges on the realization of contrastive stress on the scalar item (e.g. Horn, 2006; Geurts, 2009, 2010; van Tiel, to appear; Geurts and van Tiel, 2013). But if prosody can have this role, it is necessary to control for silent accent placement. Ideally, therefore, we should present sentences auditorily and systematically manipulate

[7] It is a matter of controversy what "typicality" actually is. Intuitively, a sentence like *Some of the 10 marbles are white* is more "typical" or "natural" in a situation with 4 white marbles than in a situation with 8 or 9 white marbles. These intuitions have been tested and reported as "typicality" or "naturalness" data for a number of quantifiers (Degen and Tanenhaus, 2011; van Tiel, to appear, 2014; Degen and Tanenhaus, to appear). It remains controversial, however, what exactly it is that is measured and labelled "typicality" in these tasks (c.f. Cummins, 2014; Franke, 2014).

contrastive stress.

Upshot. This leaves us with the following desiderata. (i) In order to test the availability and preferences of different readings of AS- and ES-sentences, we need a way to unambiguously map responses, ideally categorical, to specific readings. (ii) We want to minimize possible pictorial effects. (iii) We would like to obtain information about the relative preferences of the different readings, ideally by comparison to some other, "base-line" condition. (iv) We should try to avoid issues of "silent prosody" and actively explore the role of prosodic markedness.

5 An Incremental Verification Task

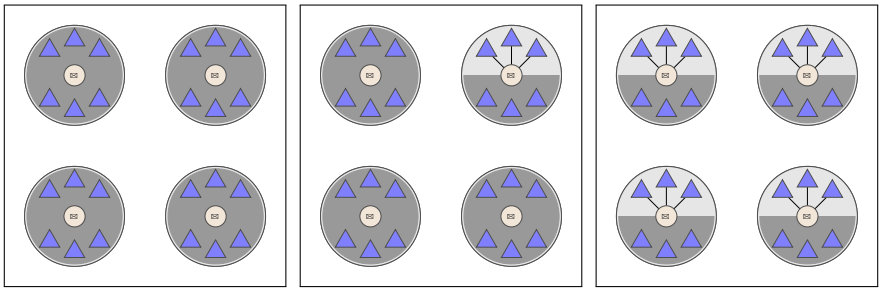
5.1 Design

General motivation. To deal with desiderata (i) and (ii), we used an *incremental verification task* (IVT), which is a modified version of picture verification (see Conroy, 2008). The general idea is that subjects are requested to judge sentence material with respect to pictures that do not necessarily contain all the information relevant for judging a certain reading true or false. In that case, participants could demand that more information be revealed. Participants were instructed to make a truth-value judgment as soon as they were able to do so. When they did, the trial ended. Motivated by desideratum (iii), we included "preference-related control" conditions. The sentence materials were presented auditorily and we manipulated the prosodic realization to cater for desideratum (iv).

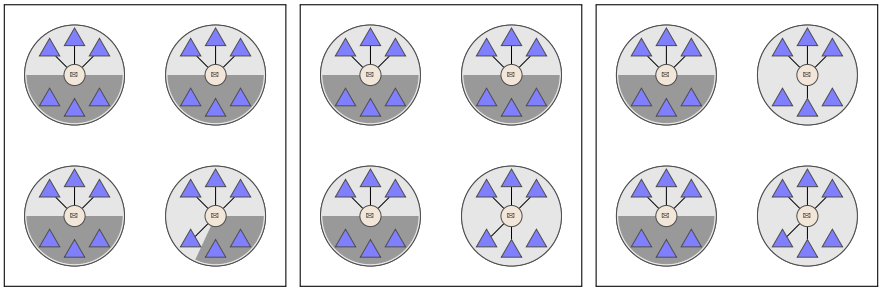
Target Conditions. We present sequences of pictures depicting a set of four identical central elements (e.g., letters), which could be connected to surrounding elements (e.g., triangles), as shown in Figures 4–7. Initially, any potential connections between central and surrounding elements were covered by dark gray color (see Figure 4a). Sentences to be judged were presented auditorily, and participants were asked to uncover the picture until they felt able to give a truth value judgment. Three options were available for participants at each step: (i) judge the sentence as true, (ii) judge the sentence as false, (iii) demand more information (unavailable when the picture was fully uncovered). Trials ended when a truth-value judgement was made. Importantly, each of the three potentially available readings corresponded to a specific step in the uncovering process where the truth value of that reading (and only of that reading) could be assessed for the first time. We refer to this step in the sequence as the *critical position* of a given reading. The critical position and the corresponding truth-value judgment differed between AS- and ES-sentences, as described presently. (This is, partly, because of the logical dependencies between readings, and, partly, also to rule out positional biases.)

Consider the German AS-sentence in (13), included in our study.

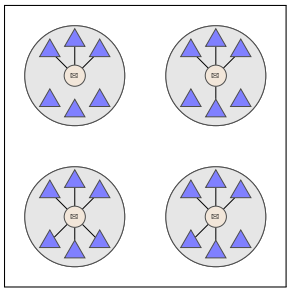
(13) **Alle** diese Briefe sind mit **einigen** ihrer Dreiecke verbunden.
All these letters are with some their triangles connected.



(a) Step 1 (b) Step 2 (c) Step 3 (LIT true)

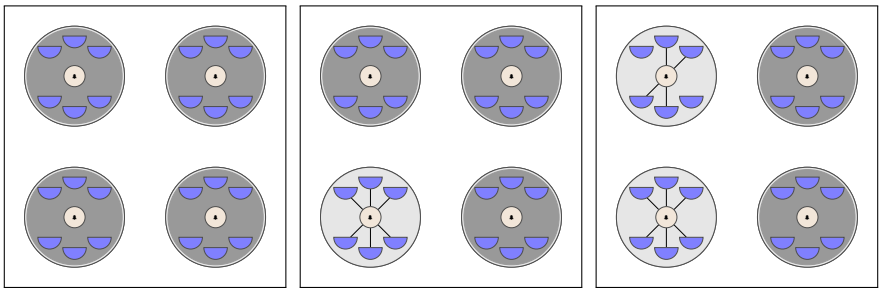


(d) Step 4 (e) Step 5 (GLB true) (f) Step 6

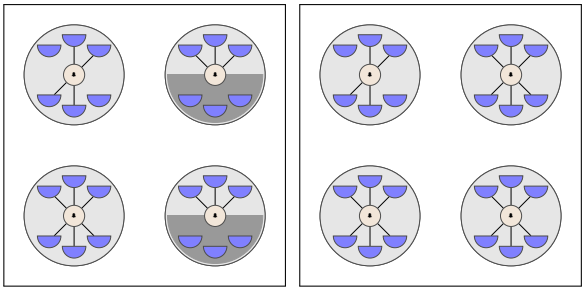


(g) Step 7 (LOC false)

Figure 4: Example picture sequence for AS-sentences.



(a) Step 1 (b) Step 2 (GLB false) (c) Step 3 (LIT false)



(d) Step 4 (e) Step 5 (LOC true)

Figure 5: Example sequence for ES-sentences.

All of these letters are connected to some of their triangles.

Figure 4 shows the corresponding picture sequence. The critical positions are on step 3, 5 and 7. Pictures that were interspersed between these positions served as spillover-pictures. These were used in order to counter potentially delayed judgments and they additionally served as distractor items. Prior to step 3 there is not enough information to judge any candiate reading true or false. The situation at step 3 in Figure 4c is true under a literal reading, while the local and global readings cannot be evaluated yet. On step 5 in Figure 4e the literal reading is still true, but now also the global reading can be confirmed. Finally, decisions concerning the local reading are possible as soon as all connections have been uncovered at step 7 in Figure 4g. Note that the situation at step 7 is *in*compatible with a local reading. This enables us to separate local readings, which would yield *false* judgements at this position, from literal and global readings, which would yield *true* judgements. In sum, *true* or *false* answers on particular positions in the incrementally revealed picture sequence can be mapped uniquely to candidate readings. All other *true* or *false* answers were counted as errors.

Due to the non-linear entailment order of readings described in Section 2, ES-sentences require a slightly different sequence of unfolding, which yields a different order in which truth-value judgments can be made. For a sentence like (14), we obtain an unambiguous mapping between truth judgements and readings with the sequence in 5.

(14) **Genau** eine der Glocken ist mit **einigen** ihrer Halbkreise verbunden.
Exactly one of-the bells is with some its semicircles connected.
Exactly one bell is connected to some of its semicircless.

The critical positions are steps 2, 3 and 5. At step 2 in Figure 5b, a global reading would yield a *false* judgement. The literal reading can be evaluated at step 3 in Figure 5c, where it would trigger a *false* judgement. Finally, confirming 5e by a *true* response indicates a local reading. Note that global and literal readings again differ from local readings with respect to their truth values.

Positional controls. To make sure that subjects understood the task, i.e., gave truth-value judgements at the first possible position in a sequence, we included a set of control conditions. These also controlled for response biases. For each of the three readings in the AS- and ES-conditions, we constructed an unambiguous control sentence as in (15) and (16), requiring the same judgement at the identical position in the same sequence used for the respective targets. For instance, example (15a) requires a *true*-response analogous to the AS-sentence in (13) under its literal reading at step 3 in Figure 4c, (15b) corresponds to the AS-sentence under its global reading and (15c) corresponds to its local reading. With regard to the ES-sentences, controls like (16a) correspond to the global reading, (16b) to the literal and (16c) to the local reading. If, independently of sentence meaning, there was any bias to respond in a certain way at any point in the sequences, such as to preferably unravel the whole picture, this should affect control sentences to the same degree as it affects target conditions.

(15) a. Alle Briefe sind mit mindestens drei ihrer Dreiecke verbunden.
All letters are with at-least three their triangles connected.
All letters are connected to at least three of their triangles.

b. Mindestens ein Brief ist mit genau fünf seiner Dreiecke verbunden.
At-least one letter is with exactly five his triangles connected.
At least one letter is connected with exactly five of its triangles.

c. Jeder Brief ist mit mindestens vier seiner Dreiecke verbunden.
Every letter is with at-least four his triangles connected.
Every letter is connected to at least four of its triangles.

(16) a. Alle Glocken sind mit weniger als vier ihrer Halbkreise verbunden.
All bells are with fewer than four their semicircles connected.
All bells are connected with fewer than four of their semicircles.

b. Alle Glocken sind mit allen ihren Halbkreisen verbunden.
All bells are with all their semicircles connected.
All bells are connected to all of their semicircles.

c. Mindestens drei Glocken sind mit allen ihren Halbkreisen verbunden.
At-least three bells are with all their semicircles connected.
At least three bells are connectd with all of their semicircles.

Preference-related controls. In order to test whether the order of critical positions for different readings within a sequence had an influence on responses, we included a second type of control conditions. These controls also tested whether the incremental verification task could possibly detect at least ordinal preference relations among multiple candidate readings. Note that for our target sentences, the logical entailment relations between readings always require local readings to be evaluated at the end of a sequence. It could be that subjects never reach this point, because they give truth-value judgements earlier, thereby ending the trial. In that case it would be unclear whether these decisions had been affected by a general unavailability of local readings or by the fact that one of the earlier presented readings is the preferred one. By including globally ambiguous structures with a known preference over available readings, we thus tested whether participants in our task occasionally choose dispreferred readings, even if these were available only at a critical position following the preferred readings. Finally, these conditions also controlled whether prosodic information can, in principle, shift answer patterns in the present task.

We therefore included sentences with *attachment ambiguities* as in (17), which have been shown to exhibit interpretive preferences and can be disambiguated by prosodic information.[8]

(17) Der Brief ist mit Kreisen und Vierecken mit Sonnen verbunden.
The letter is with circles and squares with suns connected.
The letter is connected with circles and squares with suns.

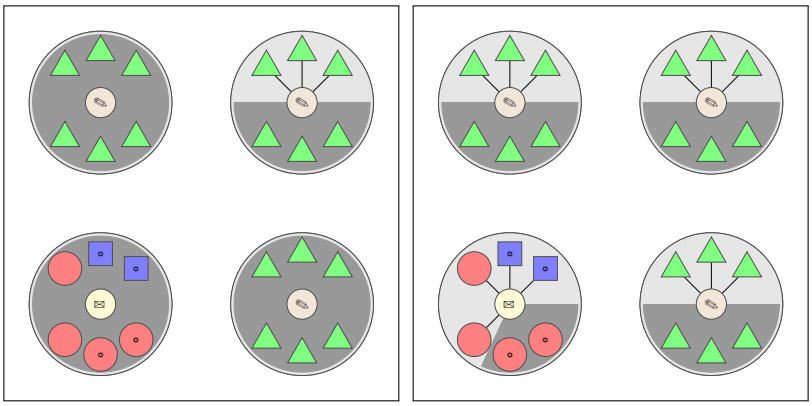
a. The letter is connected with squares containing suns, and it is also connected with circles. (LC)

b. The letter is connected with circles and squares, both of which are containing suns. (EC)

In attachment ambiguities involving post-nominal modification, an adjunct like a relative clause or a prepositional phrase (PP) can be attached to one of two preceding hosts. For instance, in (17), the PP *with suns* can be attached to the preceding noun *squares*, resulting in the so-called *late-closure* (LC) reading (17a). Alternatively, the PP can be attached to the whole conjunctive NP *circles and squares*, corresponding to the *early-closure* (EC) reading (17b). LC-readings have been generally observed to be preferred over EC-readings (e.g., Fodor, 1998 for an overview of different languages). Crucially, an LC-preference for PP-attachment structures has also been attested for German (Konieczny and Hemforth, 2000).

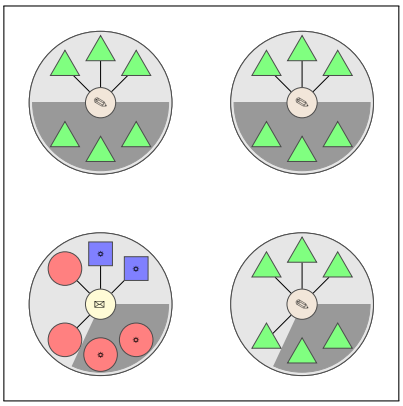
For each sentence, two sequences were designed: one in which the preferred LC-reading could be judged first (Figure 6), and another in which the dispreferred EC-reading could be judged first (Figure 7). We also tested structurally disambiguated versions like (18), one for each sentence type, on each sequence type for proper comparison.

[8] It is possible that attachment ambiguities are different from the "pragmatic ambiguity" in AS- and ES-sentences, so that information about the former is not ideally informative about the latter. We will address this worry in Section 6, where we argue that this does not affect the main conclusion that we would like to draw from our data eventually.

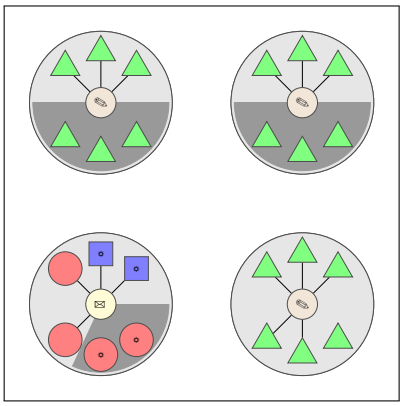


(a) Step 2

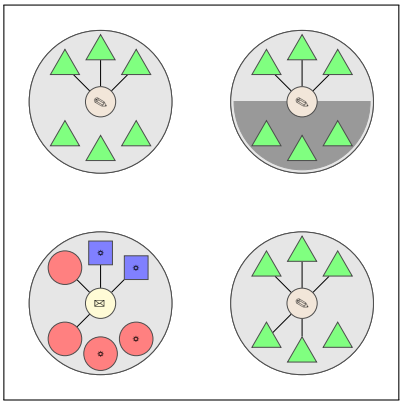
(b) Step 3 (LC true)



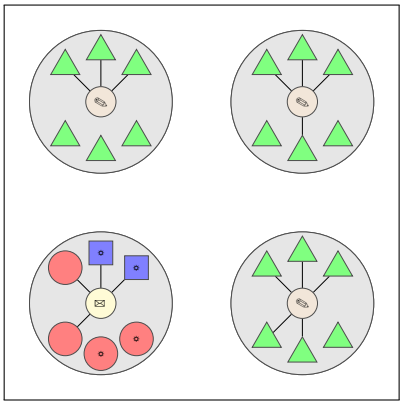
(c) Step 4



(d) Step 5

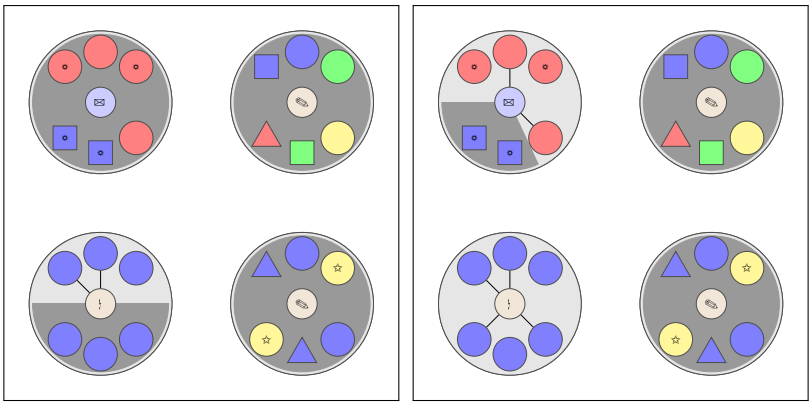


(e) Step 6 (EC false)

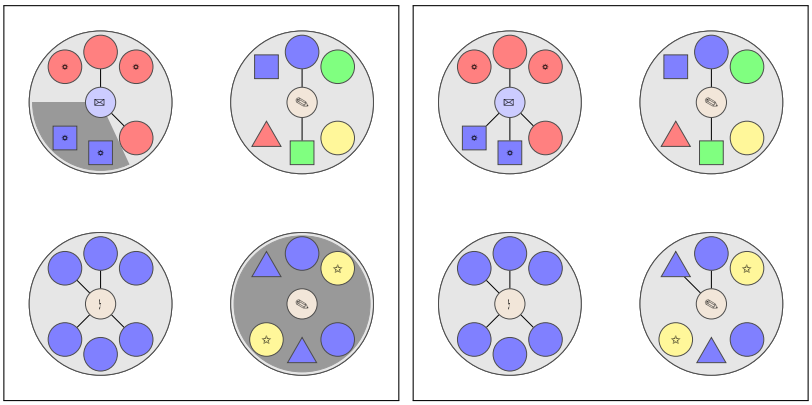


(f) Step 7

Figure 6: Example of an LC-sequence for sentence (17) where the LC-reading (17a) can be judged first. The first step where all connections are covered is omitted.



(a) Step 2 (b) Step 3 (EC false)



(c) Step 4 (d) Step 5 (LC true)

Figure 7: Example of an EC-sequence for sentence (17) where the EC-reading (17b) can be judged first. The first step where all connections are covered is omitted.

(18) Der Brief ist mit Kreisen, die Sonnen beinhalten, und Vierecken
The letter is with circles, which suns contain and squares
verbunden.
connected.
The letter is connected with circles containing suns, and with squares.

Testing Effects of prosody. Whereas contrastive accents are usually assumed to be realized by a special contour in English (i.e. L+H*, see Pierrehumbert and Hirschberg, 1990), the exact phonological classification of contrast in German is still under debate (see Uhmann, 1991; Féry, 1993; Grabe, 1998; Toepel, 2006; Sudhoff, 2010). Still, all of these approaches have in common that from an acoustic perceptive, prosodic prominence of contrastively accented constituents is realized by

(i) a higher F0 maximum or a higher pitch range (difference between minimal and maximal F0 values), and

(ii) a longer duration of the accented element as opposed to its non-accented counterpart.

Our sentences were read by a phonetically trained female native speaker of German familiar with the concept of contrastive focus. Both AS-sentences and ES-sentences were recorded in two versions each. In the accented version, a contrastive stress was placed on *einige*. The second version had neutral prosody. If accentuation is the driving force for local readings, we would expect a higher proportion of local readings in the former than in the latter version of the sentences.

Preference-related controls also differed prosodically. Modifier attachment ambiguities have been shown to be sensitive to differences in prosodic phrasing. Though there has been discussion as to whether speakers reliably *produce* disambiguating cues in such constructions (e.g. Allbritton et al., 1996; Schafer et al., 2000b; Snedeker and Trueswell, 2003; Kraljik and Brennan, 2005), overt prosodic boundaries are generally acknowledged to guide interpretive processes of attachment ambiguities in *comprehension* (e.g. Beach, 1991; Kjelgaard and Speer, 1999; Steinhauer et al., 1999; Schafer et al., 2000a; Clifton et al., 2002; Augurzky, 2006). Specifically, it has been shown that prosodically separating a modifier from the directly preceding material supports an EC-reading, whereas separating the two potential attachment sites is supportive of an LC-reading (e.g. Clifton et al., 2002; Fodor, 2002; Jun, 2003). Thus, a phrase boundary between *squares* and *with* in (17) supports an EC-reading, whereas a prosodic phrase boundary following *circles* supports an LC-reading.

To test whether prosodic information could generally have an effect on answer patterns in our task, we presented sentences in both of these prosodic variants. In addition, in order to test prosody-independent preferences, we also included a neutral version without any pronounced boundary. Further, we also presented unambiguous sentences corresponding to the EC-reading. As in the case of the AS- and ES-sentences, the unambiguous counterparts served as a baseline to control for response biases that are independent of sentence meaning.

5.2 Methods

5.2.1 Procedure

Each experimental trial proceeded as described above. The sentence was presented using active PC loudspeakers and the picture materials were presented on a computer screen. Participants responded using the keyboard. By pressing one of four buttons they could (i) listen to the sentence (up to three times), (ii) request additional pictorial information, (iii) respond "yes, fits" or (iv) "no, does not fit" (German: "passt" or "passt nicht"). At the beginning of an experimental session participants received written instructions on their task. Then they completed an interactive training consisting of 15 trials in which they were familiarized with the task. After the training the actual experiment started. Participants were tested individually in a silent room. Each session was divided into two blocks and participants were told that they could take breaks between blocks. In total, an experimental session took about one hour. At the end of the session participants received 8 euros compensation.

5.2.2 Materials

Target sentences and positional controls. We constructed a set of 15 items for German AS- and ES-sentences, respectively, analogous to the examples in (13) and (14). Sentences were introduced by a subject DP, including the quantifiers *alle diese* ("all of these") or *genau eine(r) der* ("exactly one of these"), as well as the head noun which denoted different icons, such as letters or bells. Subject DPs were followed by an auxiliary, and a PP containing *einige*, a possessive pronoun, and a noun denoting a geometrical object, e.g., *einigen seiner Dreiecke/Quadrate* ("some of its triangles/squares"). The possessive pronoun was used in order to fix relative quantifier scope to surface scope, thus ensuring that *einige* has embedded status. The last word was the main verb *verbunden* ("connected"). Each sentence was recorded in a stressed and an unstressed version of scalar *einige*. As described above, three control sentences were created for each experimental item, corresponding to the critical positions in the course of uncovering the accompanying sequences. The positional controls were recorded with neutral prosody. Items were evenly distributed across 5 lists using a Latin square design.

Preference-related control sentences. For controlling preferences, 30 sentences that were ambiguous between a LC and an EC-reading, as well as 30 of their disambiguated counterparts were constructed as described above. In these sentences, subject DPs were always denoting icons and were followed by an auxiliary. In the ambiguous sentences, the auxiliary was followed by two *with*-PPs (German: "mit"). The first of these PPs consisted in a coordination of two nouns denoting geometrical objects, whereas the nouns in the second PP denoted icons. In contrast to these sentences, the unambiguous sentences contained only one PP with a conjunction the first part of which was modified by a relative clause.

Acoustic properties. For all target sentences (AS and ES), the determiner *einige* was produced with a contrastive pitch accent as well as with a neutral accent. A set of 15 ex-

perimental items was recorded for each condition (AS vs. ES, *accented* vs. *unaccented*), resulting in a total number of 60 target sentences.

In contrast to the accent manipulation, the ambiguous preference-related controls differed with respect to prosodic phrasing. Prosodic phrase boundaries in German are realized by a rise in F0 as well as by a durational increase on the final part of the constituent preceding the boundary (prefinal lengthening) plus an optional pause (e.g. Vaissière, 1983; Féry, 1993). Boundaries for these control sentences were either realized at the position separating the second PP from the preceding material (*late boundary*, corresponding to an EC-reading) or directly preceding the second conjunct in the first PP (*early boundary*, corresponding to an LC-reading). As the prosodic realization of the targets involved the comparison between an accented and a neutral variant, we also included a third version of the ambiguous preference-related controls without any pronounced boundaries. For each prosodic variant, a set of 30 items was read, yielding 90 preference-related controls.

Altogether a total number of 300 sentences consisting of 60 target sentences, 90 ambiguous preference-related controls, 30 unambiguous preference-related controls, 90 positional control sentences and 30 unrelated fillers was recorded. The session was recorded in an acoustically shielded booth (44.1 kHz sampling rate, 16 bit amplitude resolution).

Before entering the judgment task, experimental items and preference-related controls were analyzed with respect to their acoustic properties. As both accented elements as well as prosodic boundaries were expected to differ with respect to their F0 and/or durational properties, we calculated durational values as well as difference values between minimal and maximal F0 for each word. Since targets slightly differed with respect to the total number of words as well as with respect to certain lexical properties (i.e., *seinen* vs. *ihren*, we considered the following analysis regions:

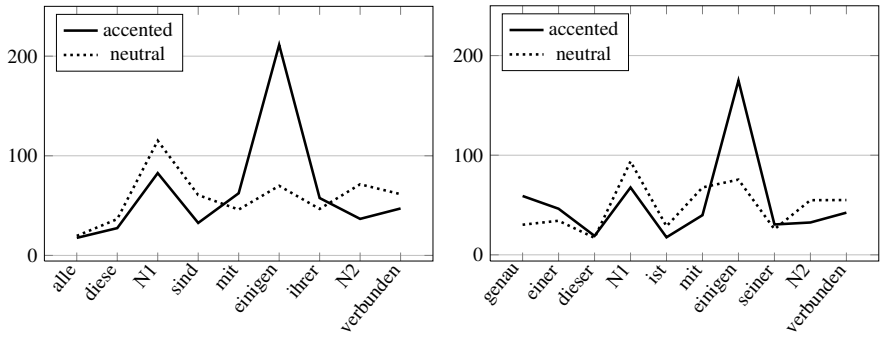
(19) a. $|_{R1}$ Alle $|_{R2}$ diese $|_{R3}$ NP_1 $|_{R4}$ sind $|_{R5}$ mit $|_{R6}$ einigen $|_{R7}$ ihrer $|_{R8}$ NP_2 $|_{R9}$ verbunden.

b. $|_{R1}$ Genau einer $|_{R2}$ der $|_{R3}$ NP_1 $|_{R4}$ ist $|_{R5}$ mit $|_{R6}$ einigen $|_{R7}$ seiner $|_{R8}$ NP_2 $|_{R9}$ verbunden.

Note that differences between Regions 1, 2, 4 and 7 can be expected due to lexical differences between the AS and ES-conditions. As preference-related controls did not differ with respect to their lexical properties, we carried out word-by-word analyses for these conditions. Durational values included the respective word plus any following silent interval. We did not include the disambiguated fillers in these analyses as they involved very different sentence types (i.e., constructions involving prepositional phrases vs. relative clauses).

(20) $|_{R1}$ D $|_{R2}$ NP_1 $|_{R3}$ ist $|_{R4}$ mit $|_{R5}$ NP_2 $|_{R6}$ und $|_{R7}$ NP_3 $|_{R8}$ mit $|_{R9}$ NP_4 $|_{R10}$ verbunden.

For the durational analyses, constituents were automatically labeled by the *Aligner* tool (Rapp, 1998), and the obtained values were manually corrected afterwards. For the targets, two-factorial ANOVAs with the factors QUANTIFIER (*all* vs. *exactly one*) and



(a) AS-sentences (b) ES-sentences

Figure 8: Differences between minimal and maximal F0 values (in Hz) for the single regions for AS-sentences (a) and ES-sentences (b).

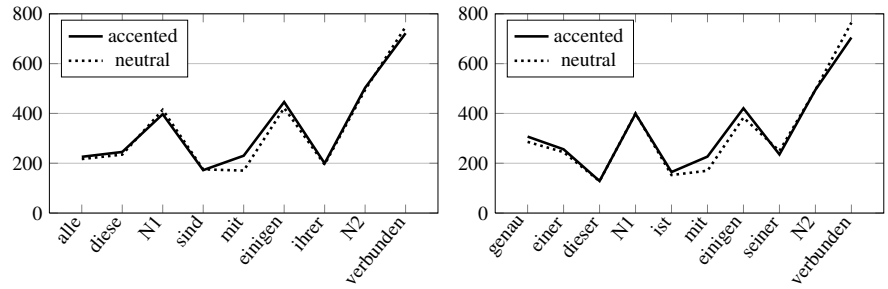
PROSODY (accented vs. unaccented) were carried out. For the preference-related controls, we carried out one-factorial ANOVAs with the factor PROSODY (early boundary, late boundary, neutral prosody). F0 values were extracted by means of special Praat scripts (`http://www.fon.hum.uva.nl/praat/`). For the present analyses, differences between minimal and maximal F0 values for each region or word were calculated. Again, two-factorial ANOVAs were carried out for statistical comparison of the target sentences, and one-factorial ANOVAs were carried out for preference-related controls.

Targets. Differences between maximal and minimal F0 values for each of the single words in the sentence are depicted in Figure 8. As expected, accented determiners showed a larger F0 range compared to unaccented versions, an effect which was confirmed by statistical analyses (effects on the determiner: PROSODY: F=94.8; p<.001); Interaction of QUANTIFIER and PROSODY: F=13.7; p<.01). The observed interaction indicates that these differences are even more pronounced in the AS-condition.

Figure 9 shows the durational values for each of the single regions in the sentence. Though descriptively small, the duration of accented determiners was significantly increased as opposed to their non-accented counterparts (effects on the determiner: QUANTIFIER: F=31.4; p<.001; PROSODY: F=23.8; p<.001). A list of the statistical results of each sentential region is provided in Appendix A (F0 values: Table 6; durational values: Table 7).

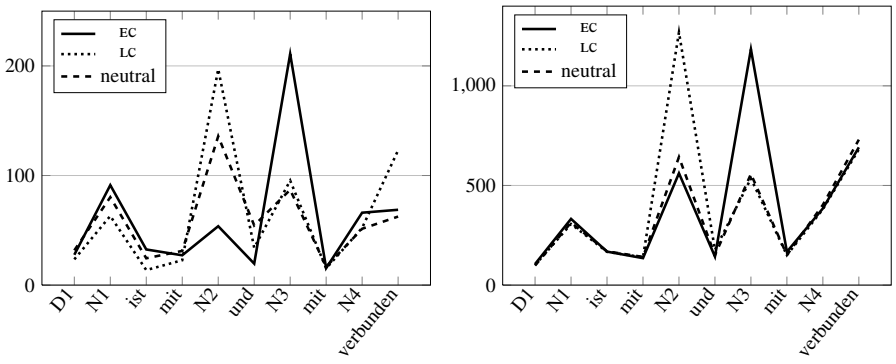
Preference-related controls. Differences between maximal and minimal F0 values for each of the single words in the sentence as well as durational values are depicted in Figure 10. As is evident, the largest durational and F0 differences were realized at the boundary regions (i.e. Region 5 and Region 7).

Statistical analyses for F0 on Region 5 reveal that all prosodic realizations (*neutral* vs. (*early boundary* vs. (*late boundary*)) differ significantly from each other (effect of PROSODY: F=129.3; p<.001; all single comparisons p<.001). On Region 7,



(a) AS-sentences (b) ES-sentences

Figure 9: Durational values (in ms) for the single regions for AS-sentences (a) and ES-sentences (b).



(a) F0 values (b) Duration

Figure 10: Differences between minimal and maximal F0 values (in Hz) (a) as well as durational values (in ms)(b) for the single regions .

F0 analyses again show a main effect of PROSODY: F=132.5; p<.001. However, the comparison between neutral prosody and a *late boundary* did not reach significance (p>.2). Finally, durational analyses reveal that all prosodic realizations (*neutral* vs. (*early boundary* vs. (*late boundary*)) differ significantly from each other (Region 5: effect of PROSODY: F=819.5; p<.001; all single comparisons p<.001; Region 7: effect of PROSODY: F=1920.3; p<.001; all single comparisons p<.001.) A list of the statistical results of each sentential region is provided in Appendix A (F0 values: Table 8; durational values: Table 9).

In sum, our speaker reliably produced (i) differences in accent realization for the target sentences and (ii) the expected boundary realizations for the preference-related controls. Whereas accented elements clearly differed from their unaccented counterpart by showing an increase in duration and F0 range, prosodic boundaries for our preference-related controls were realized by pre-final lengthening (i.e., an increase in

F0 and duration at the position preceding the boundary).

Pictures. The visual sequences accompanying each sentence consisted of pictures like in Figures 4–6. In each picture, four alternative sub-scenarios were presented, in which an icon was surrounded by six geometrical shapes. Depending on the sentence material, shapes and icons differed across sentence types and conditions. For example, in our target conditions, 4 identical icons were surrounded by geometrical shapes that were all of the same type, whereas various icons and surrounding elements were used in each sequence of the preference-related control structures (see Section 5.1 for more detailed exposition). Pictures for unrelated fillers had critical positions at random uncovering stages, including cases where no uncovering was needed for a truth-value judgement.

Lists. Experimental items and preference-related controls were evenly distributed across lists. For the AS- and ES-sentences together with their respective controls, five lists were used employing a Latin square design. Picture sentence-pairs for the preference-related controls were spread across eight lists. Each target list was combined with each list from the preference-related controls, thus yielding to a total of forty lists. The thirty unrelated filler items were included in each of these resulting lists.

5.2.3 Participants.

Forty native speakers of German took part in our study, none of whom had any prior exposure to logic or formal semantics. We excluded two participants due to insufficient performance on controls ($\leq 60\%$ correct answers).

5.3 Results

In the following, we will report the results for target conditions and preference-related controls separately. After providing the descriptive results, we present statistical analyses using log-linear models. We complement these with a Bayesian analysis that is sensitive to the sequential nature of the incremental verification task and probes specifically into the preference relations between candidate readings.

Target conditions and positional controls. Performance on positional controls was overwhelmingly correct (90% correct responses on average with a variance of 0.012) indicating that participants understood the task and were able to give correct truth-value judgments at the adequate point in a sequence. In general, they were not affected by inessential features of the picture materials or general response biases.

Judgements obtained for the four target conditions are listed in Table 2. We coded these answers as *literal*, *global* or *local* if they were as expected under one of these readings and as *error* if not. The majority of answers are indicative of a literal reading (AS-neutral: 90.4%, AS-accented: 86.8%, ES-neutral: 68.4%, ES-accented: 68.4%). Only two answers (0.4%) across all critical conditions fall into the category indicative of global readings. Judgements indicative of local readings, on the other hand, are

		truth value at step							classification			
condition		1	2	3	4	5	6	7	LIT	LOC	GLB	err
AS-ntr	T	1	1	**103**	2	**0**	0	2	103	3	0	8
	F	0	0	0	0	2	0	**3**				
AS-acc	T	0	1	**99**	2	**2**	0	1	99	8	2	5
	F	0	0	1	0	0	0	**8**				
ES-ntr	T	0	11	3	1	**14**	-	-	78	14	0	22
	F	0	**0**	**78**	5	2	-	-				
ES-acc	T	0	10	3	0	**16**	-	-	78	16	0	20
	F	0	**0**	**78**	5	2	-	-				

Table 2: Observed truth-value judgments in critical conditions. The left-hand part of the table shows the number of *true* and *false* responses obtained at each step in the sequence. The right-hand part shows the count for our imposed classification scheme. The answer counts indicative of relevant readings are given in bold-face on the left-hand side.

more frequent (AS-neutral: 2.6%, AS-accented: 7.0%, ES-neutral: 12.3%, ES-accented: 14.0%).

The number of responses that fell out of our classification scheme differs substantially between AS- and ES-conditions. There are 7.0% / 4.4% unclassifiable judgements in the neutral / accented AS-conditions, but 19.3% / 17.6% in the neutral / accented ES-conditions respectively. A look at the left-hand side of Table 2 reveals that there is some systematicity in 'error' answers. Most answers that we classified as 'errors' have a plausible explanation: random errors, a handful of mistakes resulting from confusing buttons for *true* and *false* and a few 'spill-over errors' where participants gave responses one position too late. There is one notable exception, though. A substantial number of *true* responses were given at the second position of ES-sequences (9.7% / 8.8%) which is shown in Figure 5b. Although this position in the sequence is the critical position to reveal a global reading, the latter would be indicated by a *false* response. Given the low number of mistakes attributable to confusion of buttons for *true* and *false* elsewhere, this seems to be a more systematic pattern. It appears as if subjects read ES-sentences as if the non-monotonic quantifier *exactly one* was a monotonic *at least one*, without pragmatic enrichment of *some*. 16 out of the total 21 answers in this category come from 5 subjects, while the remaining 5 were single answers by 5 different subjects. This suggests that at least for some subjects non-monotonic quantifiers were problematic. (We will come back to this intriguing observation later in Section 5.4, where we suggest that a non-monotonic reading of *exactly one* might also explain the high number of *true* answers in step 5 of the ES-conditions.)

We fitted generalized linear models to the count data in Table 2, using factors SENTENCETYPE with levels AS and ES, ACCENT with levels *neutral* and *accented*, and TRIAL with levels 1 through 3 (coding whether it was the first, second or third time that a

coefficient	estimate	std. error	z-value	Pr(> $\|z\|$)
INTERCEPT	0.916	0.258	3.49	< 0.001
READING.*literal*	2.600	0.268	9.716	< 0.001
READING.*local*	-0.310	0.397	-0.781	0.435
SENTENCETYPE.ES	1.030	0.301	3.423	< 0.001
READING.*literal*:SENTENCETYPE.ES	-1.288	0.319	-4.036	< 0.001
READING.*local*:SENTENCETYPE.ES	-0.026	0.463	-0.0579	0.955

Table 3: Coefficients of the "best" log-linear model for the count data in Table 2.

participant saw a critical condition during the experiment) to predict the dependent factor READING with levels *literal*, *local*, and *other*. The latter lumps *global* and *error* responses together, because otherwise we would not have enough cell counts to apply log-linear regression. We determined the "best" model of the data by a gradient search over hierarchical models in terms of AICs, starting from the saturated model. The best model takes main factors READING and SENTENCETYPE into account, as well as the interaction term READING:CONDITION ($\chi^2 = 17.19$, $df = 42$, $p = 0.99$, $AIC = 158.01$ (compared to $AIC = 208.25$ of the saturated model)). Crucially, factors TRIAL and ACCENT were dropped in the best model. This suggests that response patterns only depend on the type of the sentence, but that there were no learning or fatigation effects and that accentuation did not influence answer patterns significantly (at this level of abstraction).

Inspection of the coefficients of the "best" model in Table 3 suggests that the distinction between levels *local* and *other* in factor READING might not be necessary: factor levels *local* do not cause a significant shift in counts, given reference level *false*. In order to test whether there is support for the postulation of local readings, we therefore compared the previous "best" model to a model that only differs from the former in that it subsumes the level *local* of factor READING under level *other* as well (e.g. Crawley, 2007, Chapter 15). Model comparison reveals that there is no significant improvement in explanatory power ($Pr(|\chi|) = 0.2689$) of the more complex model that includes the level *local* (residual df=30, residual deviance = 10.497) over the simpler model without this level (residual df=32, residual deviance = 13.124, AIC=157.268).

This latter analysis suggests that our data does not provide strong evidence for the maintenance of belief in local readings in our setting. The regression analysis tells us that the number of errors and the number of local readings were similar in all conditions. But the distribution of answers in Table 2 shows that small error counts occur at many different steps, whereas we do see a concentrated rise in answers at the critical position for local readings. Also, so far we have not taken into account the sequential structure of our task. This is where preference-related controls are relevant.

Preference-related controls. Positional answers for our preference-related control conditions are shown in Table 4.[9] Judgments were coded as EC or as LC if they were as

[9]We removed 19 data points for the EC-conditions because of a coding mistake that presented subjects with incongruent sentence-picture pairs. Including these and classifying them as *errors* does not change the

		truth value at step							classification		
condition		1	2	3	4	5	6	7	LC	EC	err
LC-ntr	T	0	0	**87**	1	0	9	0	87	32	13
	F	0	0	3	0	0	**32**	0			
LC-LC-cue	T	0	0	**108**	1	1	7	0	108	16	10
	F	0	0	1	0	0	**16**	0			
LC-EC-cue	T	0	0	**51**	4	0	7	0	51	59	25
	F	2	1	7	0	0	**59**	1			
EC-ntr	T	0	0	0	0	**92**	-	-	92	22	13
	F	1	0	**22**	6	6	-	-			
EC-LC-cue	T	0	0	0	0	**112**	-	-	112	13	4
	F	0	1	**13**	1	2	-	-			
EC-EC-cue	T	0	0	0	0	**65**	-	-	65	31	30
	F	2	0	**31**	8	20	-	-			

Table 4: Observed truth-value judgments in preference-related controls.

expected under one of these readings, and as *error* if they were not. As expected, most answers classify as LC-readings, fewer as EC-readings. Moreover, prosody does seem to have the expected effect as well: as compared to neutral phrasing, LC-cueing prosody increases the count for LC-readings, while EC-cueing prosody increases the count for EC-readings. If we compare the number of LC- or EC-readings for each prosodic variant across sequences, we see that there are always more answers indicative of the relevant readings when that reading can be judged later in the sequence. This suggests a general tendency to answer later in the sequence, rather than at the earliest possible step.

There are only very few answers that appear to be random mistakes. Some answers in the 'error' category are plausibly 'spill-over errors', where participants gave answers one step too late. But on top of that, a substantial number of 'error' answers occurs on the critical positions. For example, in the LC-sequence there are 23 *true* answers at step 6 where *false* answers would be taken as indicative of EC-readings. In the EC-sequence there are 28 *false* answers at step 5 where *true* answers would be taken as indicative of LC-readings. Finally, prosodic cues for dispreferred EC-readings seem to have led to more mistakes than other prosodic patterns.

As for the critical conditions, we computed logistic regression models including factors READING with levels LC, EC and *error*, ACCENT with levels *neutral*, LC-*cue* and EC-*cue*, TRIAL with levels 1 to 3, and PICTURE with levels LC and EC for the picture sequences. Starting search from the saturated model along a gradient of decreasing AICs, we find that the "best" model relies on main effects of READING, ACCENT and PICTURE and two two-way interactions READING:PICTURE and READING:ACCENT ($\chi^2 = 18.34$, $df = 42$,

qualitative results of the reported analyses.

coefficient	estimate	std. error	z-value	Pr(> $\|z\|$)
INTERCEPT	2.119	0.162	13.079	< 0.001
READING.*error*	-0.343	0.259	-1.328	0.184
READING.LC	0.526	0.202	2.607	< 0.01
PICTURE.LC	0.497	0.170	2.929	< 0.01
ACCENT.LC	-1.054	0.242	-4.354	< 0.001
ACCENT.*neutral*	-0.112	0.180	-0.626	0.531
READING.*error*:PICTURE.LC	-0.521	0.280	-1.87	0.062
READING.LC:PICTURE.LC	-0.583	0.195	-2.997	< 0.01
READING.*error*:ACCENT.LC	-0.103	0.422	-0.245	0.807
READING.LC:ACCENT.LC	1.657	0.279	5.943	< 0.001
READING.*error*:ACCENT.*neutral*	0.112	0.299	0.375	0.708
READING.LC:ACCENT.*neutral*	1.065	0.222	4.800	< 0.001

Table 5: Coefficients of the "best" log-linear model for the count data in Table 4.

$p = 0.99$, $AIC = 256.63$ (compared to $AIC = 322.3$ of the saturated model)). Factor TRIAL was dropped from the best model, suggesting that answer patterns are stable over the duration of the experiment. Unlike for the critical conditions, factor ACCENT is retained, and it interacts with READING, suggesting that prosodic cues significantly influenced the dominant reading.

Closer examination of the coefficients of the fitted model, shown in Table 5, reveals that there were significantly more EC-readings in the LC-condition than in the EC-condition, and more LC-readings in the EC-condition than in the LC-condition. This is evident from the significant *negative* deviation from the baseline of EC-readings in the interaction coefficient for READING.LC:PICTURE.LC. The log-linear model therefore indicates that there is a bias for late responses in a sequence. Given this, the number of local responses in critical conditions might be increased by a sequential bias to answer late. The following Bayesian analysis further probes into this matter in order to establish how strongly the sequence-effect might have emphasized local readings and, from there, whether we should uphold belief in local readings.

5.3.1 Probabilistic salience competition.

The previous regression-based analyses did not take the sequential nature of the task into account. We noted a possible bias for exiting later in a sequence, but ideally we would like to quantify how strong this effect is and what this entails for the likelihood of local readings. The following paragraphs therefore spell out a generative Bayesian model which computes the likelihood of answer patterns for different latent relative saliencies of candidate readings and biases to answer early or late in a sequence. We use the data to estimate the posterior likelihood of these latent parameters to gain insight into the likely ordering relations over the strength of latent salience of candidate readings.

The model. We present a model that aims to capture the competition between candidate readings, all of which have a given latent level of relative salience, in the incremental verification task, where some readings can be judged before others. Take the AS-conditions with its three candidate readings. The relative strength between these is given by a probability vector $\mathbf{p}^{\text{as-n}} = \left\langle p^{\text{as-n}}_{\text{lit}}, p^{\text{as-n}}_{\text{loc}}, p^{\text{as-n}}_{\text{glb}} \right\rangle$. We think of $p^{\text{as-n}}_{\text{lit}}$, for instance, as the salience of the literal reading for AS-sentences with neutral prosody, relative to the global and local reading. If there were no potential effects of sequentiality and no errors (or other readings that we classify as errors), the vector $\mathbf{p}^{\text{as-n}}$ would be our prediction of observed frequencies of answer patterns. We look at four such three-placed vectors for the critical conditions, one for each target sentence-prosody pair. For the preference-related control conditions, we look at three vectors $p^k = \left\langle p^{\text{k}}_{\text{LC}}, p^{\text{k}}_{\text{EC}} \right\rangle$, where $k \in \{\text{ntr}, \text{LC-cue}, \text{EC-cue}\}$. These vectors give the relative salience of LC- and EC-readings for different types of prosody. The salience of readings is independent of the pictorial sequence. If there is a tendency to judge sentences earlier in the sequence, rather than later, that has to be attributed to the sequential bias.

To add potential effects of sequentiality, consider bias factor $q \in [0; 1]$. This bias is a global parameter held constant over all conditions, because we think of it as a general tendency to answer early or late in the incremental verification task. The bias factor captures whether there is a tendency to choose readings that appear earlier ($q > .5$) or later ($q < .5$) in the sequence. In the AS-sequences, the literal reading can be judged first, then the local and then the global reading. So we will assume that p_{lit} is multiplied by q, that p_{loc} is multiplied by $(1 - q)q$ and that p_{glb} is multiplied by $(1 - q)^2$, when determining the eventual weights of readings.

Finally, we also allow for mistakes. To keep things simple (no differential consideration for spill-overs, button mix-ups etc.), we assume that there is a fixed error rate $e^k \in [0; 1]$ for each relevant condition k. We allow for different error rates for different conditions, because we want to allow for the possibility that some sentence types (e.g., non-monotonic quantifiers) or some kind of prosody (e.g., EC-cues) might be harder to process. A decision can be wrong but incidentally coincide with an answer that is indicative of a relevant reading. Error rates multiply in proportion to the number of steps in a sequence: more choice, more chance to be wrong.

Taking this together we calculate, for each pair k of target sentence type and prosody type, a probability vector $\mathbf{t}^k = \left\langle t^k_{\text{lit}}, t^k_{\text{loc}}, t^k_{\text{glb}}, t^k_{\text{err}} \right\rangle$ with which we expect an answer to be classified as literal, local, global or as error. For example, for AS-sentences with neutral prosody, where literal can be judged before global before local, we get:

$$t^{\text{as-n}}_{\text{lit}} \propto q\ p^{\text{as-n}}_{\text{lit}} + e^{\text{as-n}} \qquad t^{\text{as-n}}_{\text{loc}} \propto (1 - q)^2\ p^{\text{as-n}}_{\text{loc}} + e^{\text{as-n}}$$
$$t^{\text{as-n}}_{\text{glb}} \propto (1 - q)q\ p^{\text{as-n}}_{\text{glb}} + e^{\text{as-n}} \qquad t^{\text{as-n}}_{\text{err}} \propto 12e^{\text{as-n}}$$

Two remarks. The vector $\mathbf{t}^k$ should be normalized eventually, which is why '$\propto$' is used above instead of '='. Also, the probability of error $t^{\text{as-n}}_{\text{err}}$ comes from the observation that there are 15 positions in total in an AS-sequence at which subjects can make an erroneous choice, but we have accounted for three of these already as adding to the count of answers indicative of relevant readings.

Similarly, we compute a target probability $\mathbf{t}^k = \left\langle t^k_{\text{late}}, t^k_{\text{early}}, t^k_{\text{err}} \right\rangle$ for each preference-related control sentence under prosody k, taking into account which reading can be

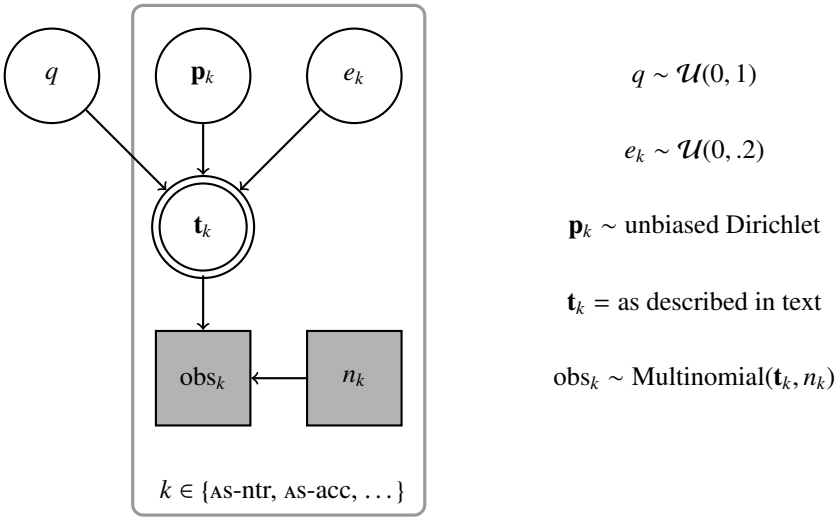


Figure 11: Probabilistic graphical model. All nodes represent variables, arrows represent functional relations between them. Using conventions of Lee and Wagenmakers (to appear), the higher a node, the more latent it is. Square nodes represent discrete-valued variables, circular nodes represent continuous-valued variables. Gray nodes show the observable variables. Values at double-lined nodes are derived deterministically. The variables in the box are repeated for each of the k relevant cases.

judged first and the length of the respective sequences.

The full probabilistic model is visualized in Figure 11. As indicated there, we assume largely uninformative priors. Any positional bias $q \in [0; 1]$ is assumed to be equally likely. The error rates for each condition should be small, so we sample uniformly from interval $[0, .2]$. The biases $\mathbf{p}_k$ are also determined uniformly at random, by sampling from a Dirichlet distribution with equal weights on all dimensions.

Model fitting. We used JAGS (Plummer, 2003) to estimate the posteriors over paramater values given our data. Results reported here are based on two chains of 10.000 MCMC samples from the joint posterior distribution, obtained after an initial burn-in of 10.000 steps. The latter guaranteed convergence.

To check whether the model yields sensible results we first look at the preference-related control conditions. Estimates for the (marginalized) posteriors over relative salience levels for target-related controls are plotted in Figure 12. The results are exactly as we would expect them to be. Given our data, we should believe that the LC-reading is most prominent. The level of prominence varies with accentuation. Under prosody that we hypothesized would favor the LC-reading, the contrast is most pronounced. Under accentuation that we hypothesized favors the EC-reading we are no longer justified in believing with certainty that the LC-reading is preferred (by compar-

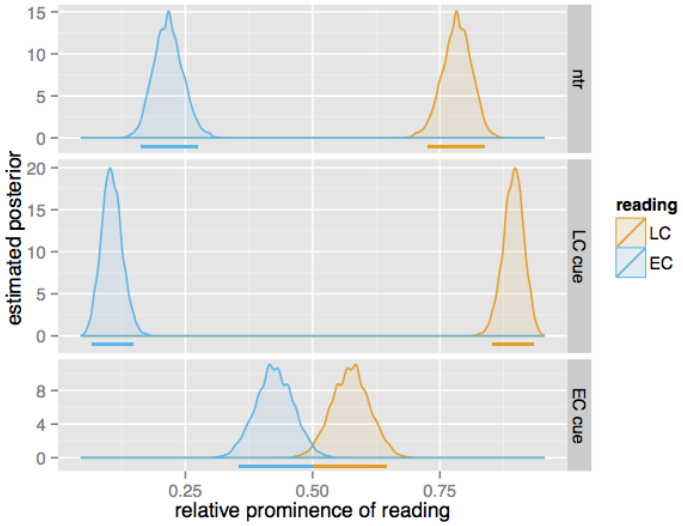


Figure 12: Estimated posteriors for the salience of LC- and EC-readings for different prosodic cues. Lines under curves indicate 95% HDIs (see Footnote 10).

ison of 95% HDIs, see below).[10]

Estimates of posteriors over salience of readings in the critical conditions are given in Figure 13. In the ES-conditions, literal readings are by far the most salient, followed by local, followed by global readings. There does not appear to be a significant effect of prosody (by visual comparison of 95% HDIs). For AS-sentences, the same preference order holds in tendency, but we cannot assert with full confidence that local readings are attested or that they are preferred over global readings (see below).

Posteriors over error rates are conceptually less interesting, and we will skip them here. The posterior over the sequential bias parameter is shown in Figure 14. Since the 95% is clearly entirely below .5, we have reason for believing, given the model and our data, in a bias for late responses.

Model validation. As a crude sanity check that our model is able to predict the observed data, we gathered 10.000 random values from the posterior predictive distribution and found a highly significant correlation between these and the observed data ($R^2 = 0.987$, $p < 0.001$).

Hypothesis testing. We are chiefly interested in two questions: (i) which readings are available? and (ii) what is the preference order over readings?

[10] A 95% Highest Density Interval (HDI) is a convex region of values over which 95% of the distribution's probability density is distributed, such that no value outside of that region has higher probability density than any point within (Kruschke, 2011). Intuitively speaking, the 95% HDI is the set of values which we may believe to be true with some confidence; what falls outside this region is what can be excluded safely.

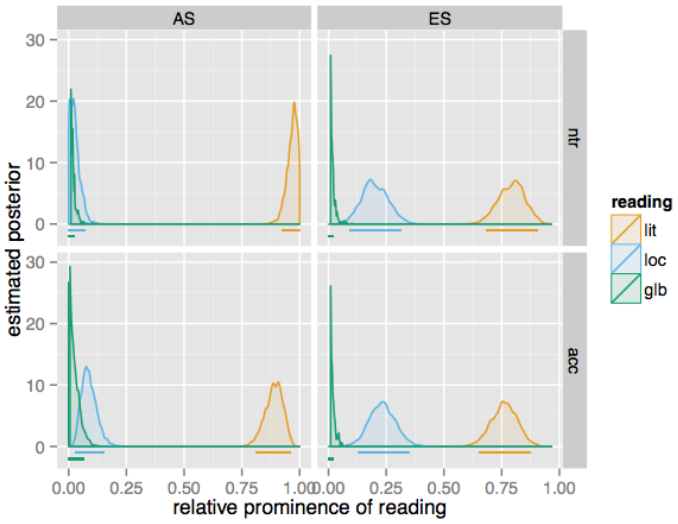


Figure 13: Posteriors over relative salience of target readings in critical conditions.

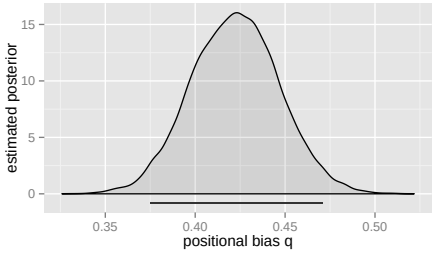


Figure 14: Estimated Posteriors over global sequential bias parameter q.

The first question can be answered by looking at the HDIs of the posteriors of the candidate readings. We would like to check the hypothesis that a given reading is unavailable, i.e., its activation value in the probabilistic model is equal to 0, or close to zero for practical purposes (Region of Practical Equivalence, ROPE). We reject the "null-hypothesis" that the true value is 0 if its ROPE lies entirely outside the 95% HDI; we accept the "null-hypothesis" if the ROPE lies entirely within the 95% HDI (Kruschke, 2011, Ch. 12). The following table lists approximations of the 95% HDIs for each condition and reading:

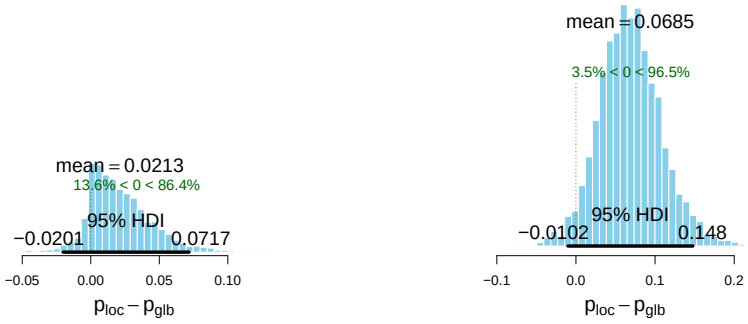
	lit	loc	glb
AS-ntr	$(0.924, 1)$	$(0, 0.068)$	$(0, 0.022)$
AS-acc	$(0.816, 0.964)$	$(0.028, 0.152)$	$(0, 0.066)$
ES-ntr	$(0.665, 0.873)$	$(0.117, 0.323)$	$(0, 0.024)$
ES-acc	$(0.634, 0.851)$	$(0.139, 0.356)$	$(0, 0.022)$

We should thus accept the hypothesis that there are no global readings at all, and no local readings for AS-sentence under neutral prosody. There is support for a belief in local readings in the AS-acc condition up to a ROPE of $[0; 0.028)$.

The preference relations between readings can be checked in a similar way. We look at the posterior beliefs about the *differences* of activation strengths and ask whether these posterior beliefs allow us to safely conclude that differences are bigger than zero or not. It is obvious enough that literal readings are always the most preferred, and that local readings are preferred over global readings for ES-sentences. Figure 15 shows the posteriors over differences for the only non-trivial cases, namely those between local and global readings for AS-sentences. Since the 95% HDIs include 0, the data provides no ground for confidence that, given the model, local readings of AS-sentences are preferred over global ones, despite the obvious tendency. In sum, with the exception of remaining uncertainty about local readings of AS-sentences, the model and the data suggest that global readings are absent entirely, while literal readings are preferred over local readings.

5.4 Discussion

Summary & Interpretation. High scores of success in control conditions show that participants understood the incremental verification task and were able to give judgements at the right point in a picture sequence. Results for the preference-related controls show that our design is able to detect multiple readings and preferences among these. Crucially, the amount of response types for LC- and EC-readings reflected the known preference for the former, with a mild bias for readings that can be judged later in the sequence. Also, the incremental verification task was generally sensitive to prosody. However, our regression analysis showed that significant prosodic effects were restricted to preference-related controls. Contrastively, the Bayesian analysis showed that accenting scalar *some* generally led to more local readings, but that effect was not significant. Local readings occurred in ES-sentences, independently of prosody, but required prosodic marking in AS-sentences. Our analyses support the impression that inspection of the raw data gives, namely that the general qualitative



(a) neutral

(b) accented

Figure 15: Posteriors over differences between salience levels for local and global readings of AS-sentences. As zero is contained in the 95% HDI for both neutral and accented variants, there is no justification for assuming a difference in salience.

pattern of reading preferences is the same for AS- and ES-conditions: literal readings are preferred over local readings which are preferred over global readings. More concretely, our analyses suggest the following preference patterns (bracketed readings are not supported by our data and the previous analyses):

(21) a. AS-ntr: LIT (> LOC ∼ GLB)

b. AS-acc: LIT > LOC (∼ GLB)

c. ES-ntr: LIT > LOC (> GLB)

d. ES-acc: LIT > LOC (> GLB)

Let's consider whether traditionalism or grammaticalism are compatible with these preference relations.

Traditionalism easily explains the absence of global readings by holding that the extra assumptions needed to derive global readings (e.g., a speaker competence assumption) are not available or only allow for epistemically weak implicatures that would show as responses indicative of literal readings in our task. In support of this view, traditionalists could argue that the task's rather high processing demands lead to a suppression of global readings, in line with the findings of De Neys and Schaeken (2007) that working memory load negatively affects the number of scalar implicature responses in sentence verification tasks. The finding of local readings in AS-sentences with contrastive stress is easily explained by adopting the prosodic markedness hypothesis. What traditionalism does not predict is the high number of answers indicative of local readings in the ES-conditions. Resorting to the prosodic markedness hypothesis does not help, because local readings were observed for accented as well as for unaccented ES-sentences.

Grammaticalism, on the other hand, fails to predict the pattern in (21) entirely if it equates preference for readings with logical strength. Strength-based disambiguation predicts that literal readings of AS-sentences are the least preferred, contrary to (21). Again, we need to bear the caveat of Section 3.2 in mind that we are only evaluating grammaticalism with respect to a particular disambiguation criterion here. The verdict, however, seems to be clear. If (21) shows the actual preference orders over readings, then meaning disambiguation in terms of logical strength is generally on the wrong track, because no account that merely looks at the logical strengths of readings can plausibly predict the preference pattern in (21) for both AS- and ES-sentences at the same time.

In sum, the present findings appear to challenge traditionalism as a "core theory", and to challenge grammaticalism under the "auxiliary assumption" that reading preferences mirror logical strength. Given this, we should ask what plausible amendements or alternative auxiliary assumptions would enable either position to accommodate the data *ex post*. This is what we do in the next section, where we also reflect critically on our design and our interpretation of the data.

6 Reflection

6.1 Alternative explanations within the core theories

Section 3 introduced traditionalism and grammaticalism as two competing core theories, whose concrete empirical predictions depend on additional assumptions. Here, we should finally ask more generally: on the supposition that our design and the offered interpretation of our data are sound, what would it take to accommodate the observed data under the core theories?

Grammaticalism. Consider grammaticalism first. Our data contradicts the notion that preferences follow logical strength. But, as noted in Section 3, grammaticalism could be supplemented by a conceptually different disambiguation criterion.[11] Selecting readings in terms of how well they answer the contextually given question under discussion, as suggested by Fox (2007); Gualmini et al. (2008), is an option. But it is quite unclear what the question under discussion should be that guided judgements in our particular task.

Another possibility of explaining our data within a grammaticalist core theory is to adopt Magri's (2011) proposal that exhaustification operators occur at every relevant scope site while alternative sets as their input may also be empty.[12] For AS- and ES-sentences, we would have to consider the parses in (22) (in simplified notation).

(22) a. Exh_{Alt_1}(**All** *x* are such that Exh_{Alt_2}(*x* is connected to **some** ....))

[11] See also the discussion by Chemla and Spector (2011) on the prospects of pushing strength-based disambiguation in the light of their data, which they also take to imply contradicting evidence on reading preferences.

[12] This interesting line of alternative *post hoc* explanation was first suggest by a reviewer, but worked out slightly differently here to strengthen the reviewer's case.

b. Exh_{Alt_1}(**Exactly one** x is such that Exh_{Alt_2}(x is connected to **some** ....))

The alternative sets $Alt_{1,2}$ are plausibly either both empty or both non-empty. In the former case, we get the literal reading. In the latter case, we should assume that alternatives are just the standard ones that asymmetrically entail the literal reading:

(23) a. For AS-sentences:
Alt_1 = {"All $x \ldots x$ is connected to all ..."}
Alt_2 = {"x is connected to all ..."}

b. For ES-sentences:
$Alt_1 = \emptyset$
Alt_2 = {"x is connected to all ..."}

Notice that Alt_1 is empty for ES-sentences if we restrict attention to the "global alternatives" that asymmetrically entail the to-be-interpreted sentence. Interestingly, the readings that this approach derives are exactly the literal readings (under empty alternatives) and the local readings (under non-empty alternatives). What is left to explain is the preference relation over these readings. Here, the Magri-style approach could resort to considerations of *economy*: reasoning with fewer alternatives is easier than reasoning with more alternatives. This would explain the preference for literal readings and also why local readings are more strongly attested for ES-sentences.

On conceptual grounds we are very much in favor of a disambiguation criterion in terms of economy considerations, but whether the particular account sketched here is empirically successful in other cases as well, must remain to be seen. Two things are worth emphasizing here nonetheless. Firstly, the sketched account is a plausible *post hoc* explanation. It was not, at least to our knowledge, a salient possibility before seeing our data set: other specifications of Alt_1 and Alt_2 could have been made equally plausible *ex ante*. Secondly, what our data refutes is that preferences for readings follow logical strength. This, therefore, does not jeopardize grammaticalism as a core theory, but merely calls for theory-internal revision of its disambiguation criterion, ideally alongside more experimental data.

Traditionalism. Traditionalism, as we described it in Section 3, cannot account for the possibility of local readings of prosodically unmarked ES-sentences, and, perhaps, also the large number of local readings in accented ES-sentences compared to the much lower number in accented AS-sentences. A first option might be to account for alleged local readings as the result of some other phenomenon, such as typicality or contrast (van Tiel, to appear; Geurts and van Tiel, 2013; van Tiel, 2014). Since this ties into a critical reflection on our design, we will enlarge on the possibility of typicality- and contrast-based effects in our setting. Eventually, we argue that contrast-effects alone are unlikely to explain the observed high number of local answers in ES-sentences, but that a "modern traditionalist" explanation of the answer pattern might be available if we allow for the possibility that *exactly one* gets an unexpected "referential interpretation."

As discussed in Section 4, van Tiel (to appear) and Geurts and van Tiel (2013) argue that the distribution of responses to AS-sentences found by Clifton and Dube (2010), as well as Chemla and Spector (2011) could be explained by differences in

how *typical* the presented pictures were for an AS-sentence. Notice that this does not apply in the present situation where the task is to explain away the local readings of ES-sentences, for which no alternative explanation in terms of typicality has been offered. Moreover, to apply typicality-based explanations to the incremental verification task, one would have to reason about the typicality of parts of pictures or to figure in subjects' expectations of pictorial typicality given their uncertainty about parts of the picture that they have not yet seen. On intuitive grounds, we take either extension to be implausible. But even if this line of explanation can be plausibly extended to incrementally revealed picture verification, it would still be unclear why typicality-effects should show so strongly in categorical truth-value judgements, and why there is such a stark contrast between AS- and ES-sentences.

A more promising alternative explanation for data indicative of local readings of ES-sentences is pursued by Geurts and van Tiel (2013) in response to the data reported by Chemla and Spector (2011). The main idea is that visual contrast between an item that is connected to all, and one that is connected to only some but not all of the relevant elements may trigger exceptional local enrichment of *some*. In support of this idea, Geurts and van Tiel show that the strength of agreement to an ES-sentence on a 7-point Likert-scale depends on how strong the relevant visual contrast is. When an item with only some connections is presented alongside two universally connected elements, the mean rate of agreement with a suitable ES-sentence is significantly higher than when only one universally connected element is present. So, maybe our ES-sequences similarly provoked local readings because of an overemphasized visual contrast between *some but not all* and *all*.

The third step of our ES-sequences in Figure 5 is a relevant case of direct contrast between a *some-but-not-all*- and an *all*-situation, but it is a weak contrast, in the sense of Geurts and van Tiel (2013), because there is only one universally connected element. Moreover, our study elicited categorical truth-value judgements. Unlike in Chemla and Spector's (2011) and Geurts and van Tiel's (2013) experiments, we did not record degrees of agreement with a statement. For a purely contrast-based explanation to work, it would have to be made plausible why the (weak) visual contrast in a picture like in Figure 5c alone is enough to overrule a truth-value judgment so as to contradict the semantic meaning (even when there is no prosodic markedness to support such a reinterpretation).

On the other hand, there is a conceivable alternative explanation of the puzzling data for ES-sentences that would also account for the large and systematic error responses. Recall from Section 5.3 and Table 2 on page 27 that a surprisingly high number of *true* answers was given at the second step of the ES-sequence. This answer type does not correspond to any of the three candidate readings that the previous literature has focused on. As suggested earlier, this answer type can be explained under the assumption that *exactly one* gets a reading similar to *there is (at least) one*, while the scalar item *some* would just receive its semantic interpretation. Such a reading of *exactly one*, although surprising, could actually be supported by some theories of numerals and modifiers (e.g. Geurts, 2006; Marty et al., 2014), as the outcome of an "existential closure" type-shifting rule in the sense of Partee (1987) to the effect that *exactly one* is mapped onto an existential reading of the form "there is a group with cardinality exactly one."

What happens if *some* is additionally strengthened in some way or other under such a reading of *exactly one*? If *some* is pragmatically strengthened under the scope of *at least one* in a pure traditionalist manner by adding the negation of a sentence type *at least one ... all ...*, we would expect *true* judgements at position 3 of the ES-sequence. Although exactly 3 *true* answers occurred there for each neutral and accented ES-conditions respectively, the main point to notice is that this construal does not explain *true* judgements at position 5, which we classified as local responses. These would therefore still seem indicative of local readings. However, there is a slight variant of the hypothesized existentially closed reading of *exactly one*, that would explain *true* judgments at position 5. This is a reading of *exactly one* as "there is a *unique* group with cardinality exactly one." Although this is admittedly only a vague sketch of a possible line of explanation, it is not entirely implausible that such a uniqueness requirement has participants, in a first step of evaluation, look for a distinguished referent, which they can find no sooner than on position 5 when the whole picture is unravelled. The search for uniqueness would result in a witness for the existential quantifier that appears to take a pragmatically strengthened reading of *some* into account, but that is not necessarily a "local reading" in the sense of grammaticalism. Moreover, although not strictly required to explain the behavioral data, we can furthermore explain how the predicate *is connected to some if its ...* can be pragmatically enriched in the desired fashion, along the lines of a "modern traditionalist" account that makes use of interpretive mechanisms originally developed to account for certain discourse phenomena (Geurts, 2010, Chapter 8.4). The listener would then conclude that the referent in question was not connected to all of its surrounding elements, because otherwise the speaker would have attributed that to the now fixed referent in question. This is parallel, so Geurts suggests, to the reasoning that an utterance of "Smith met a woman" implicates that *the woman referred to* was not Smith's mother and not that Smith didn't meet his mother at all.

To wrap up, we suggested amendements to grammaticalism and traditionalism that could help either position explain the data we observed. Whether these suggestions can be vindicated by further empirical research is a matter that we must leave open. It remains, however, that our data provides interesting challenges to both traditionalism and grammaticalism, without necessarily refuting either core theory as such.

6.2 Critique

The foregoing discussion presupposed that the interpretation of our data, given in Section 5.4, was sound. But there are a number of conceivable objections that cast into doubt that our data suggests the preference relation in (21). We will consider the most pressing ones in the following.

The "choose first match" argument. It is tempting to think that participants generally tended to choose the first exit possible. They might, after all, have simply wanted to save time and be done with the experiment. We considered this a possible confound and it is therefore that we included the preference-related controls. But the objection is already clearly refuted by the data itself. For it would entail that we should see many

more global readings for ES-sentences, which could be judged before the literal readings, and also that we should have seen the reverse pattern in the preference-related controls, where we did observe that the reading that could be judged *later* attracted comparatively more responses.

Incomparability of target conditions and preference-related controls. One could object that it is not feasible to compare results from the preference-related controls to results from our target conditions, because the source of different readings is fundamentally different. In the former we have a syntactic ambiguity, in the latter a set of putative pragmatic enrichments. The former is a case of perceived ambiguity (some of our subjects reported this), the latter arguably is not (see Geurts and Pouscoulous, 2009). So perhaps even if the distribution of responses is indicative of the true preference patterns in preference-related controls, this would not necessarily mean that the distribution of responses in target conditions is indicative of preference relations in the same way.

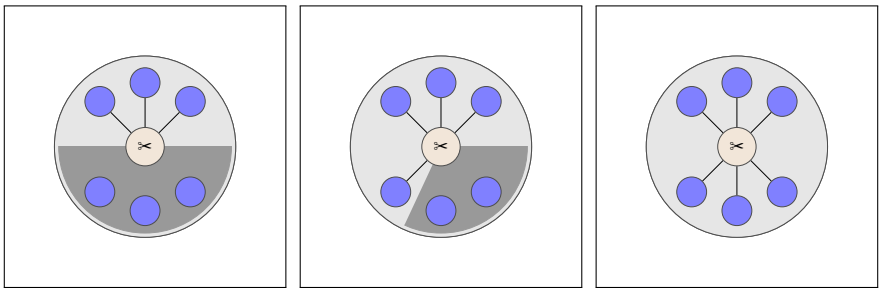
This objection is very serious. It is amplified by an even more general worry, raised as a challenge by a reviewer, who intuits that the incremental verification might altogether be insensitive to scalar implicatures. According to such a view, it could be that, although *some* is typically strengthened in non-embedded conditions, virtually all of our participants would exit at an early position corresponding to the literal reading in a suitable version of our IVT.

To address this worry we ran another stripped-down version of the IVT. We recruited 50 subjects via Amazon's Mechanical Turk and paid them 1 US$ compensation. Similar to our main experiment, the post-experiment started with two training trials with unambiguous sentences, familiarizing our subjects with the task. In the browser-based version, alert boxes pointed out subjects' mistakes when they tried to exit too early, chose the wrong response or tried to unravel the sequence further than strictly necessary. The main part of the experiment consisted of six trials, two of which were critical and four of which were controls. On critical trials, each subject saw different version of sentences like (24), in connection with sequences like in Figure 16.

(24) The scissors are connected to some of the circles.

The first step in the sequence had all potential connections covered, as before. On step 2, the critical sentence could be judged true under a literal reading. On the last step 4, it could be judged false under a pragmatically strengthened implicature reading. Control conditions had unambiguous sentences and served to filter out subjects with insufficient performance. From the total of 100 critical trials (50 subjects times 2 trials each), 66 were literal answers, 27 were pragmatic answers and 7 were 'errors'. If we restrict attention to only those participants that had at least two of the four control conditions correct, we are left with 39 subjects. With these, we find 55 literal, 21 pragmatic and 2 'error' answers. In sum, this strongly suggests that the IVT is not generally insensitive to scalar implicature and does give substantial counts also for implicature readings if these can be judged later than a literal reading.

If we accept that the IVT is sensitive to implicature-based ambiguities, what is left to worry about is that information about response biases (early vs. late) do not carry



(a) Step 2 (LIT true) (b) Step 3 (c) Step 4 (PRAG false)

Figure 16: Example sequence for critical trial in post-experiment.

over from the preference-related control conditions to the target conditions in the main experiment. It could be, after all, that the task triggers subjects to exit late for perceived syntactic ambiguities and to exit early for "pragmatic ambiguities" involving implicature readings. But none of this has any serious impact on the suggested preference relation in (21). For even if there is a sequential bias in the target conditions to exit early, the main effect of such a bias would merely be that we should consider local readings more salient than we do now. In the light of the data from main and post-experiments, it would be far-fetched to uphold that our data indicates that local readings are at least as salient as literal ones. But that means that the suggested interpretation in (21) still holds, and all the problems that we identified for (the considered versions of) traditionalism and grammaticalism apply.

Interpretation of "salience" & the Bayesian model. Another line of criticism that challenges the validity of (21) targets the Bayesian salience-competition model that we used to derive it. To avoid a potential misunderstanding pointed out by a reviewer, it is not the case the our interpretation of the IVT assumes that each subject fixes an interpretation at the beginning of the sequence and applies it, unswayed by anything that is encountered along the sequence of revealed pictures. The Bayesian model is compatible with this interpretation, but it is also compatible with a another one that we think is more plausible, namely that at each choice point during the sequence of pictures, subjects' aggregate probabilistic behavior is described as a function that compares the relative salience of the readings available, together with potential response biases and uniform error rates.

Still, we readily admit that our model is highly simplistic. Several aspects are especially noteworthy here. Firstly, we did not distinguish different types of errors, and did not take into account that when errors occurred earlier in the sequence, later decision points could no longer be reached. This way the model may actually have unduly de-emphasized the salience of readings that can be judged later in a sequence. Another, more complicated model of the likelihood of answer patterns, might therefore have given support to quantitatively different conclusions. Unfortunately, it is not clear how exactly such a more complex generating model should be set-up. Eventually, a

full processing model, designed for data from incremental verification tasks, would be needed. We cannot offer such a model here, but can only offer a first step in that direction. More importantly, though, the only part of (21) that might plausibly change for a more encompassing model along the lines sketched above is that the 3 *false* answers in the AS-conditions with neutral prosody might support the conclusion that the local reading is attested. But this hypothetical effect will certainly be no bigger than marginal, and so no genuinely different conclusions of theoretical relevance should be expected.

7 Conclusions

We tested experimentally the predictions of suitably constrained versions of traditionalism and grammaticalism with respect to the availability of and preferences over three types of conceivable pragmatic enrichments of two types of sentences, in which scalar *some* occurred in the scope of either the monotonic quantifier *all* or the non-monotonic quantifier *exactly one*. To avoid potential confounds of the pictorial material, such as through typicality or contrast (van Tiel, to appear; Geurts and van Tiel, 2013), we employed an incremental verification task (Conroy, 2008), in which subjects were presented with partially covered pictures and asked to uncover the picture sequentially until they felt able to give a categorical truth-value judgements. In order to control for potential confounds of "silent prosody," sentence material was presented auditorily and contrastive stress on embedded *some* was manipulated. This way we were able to also explicitly test the prosodic markedness hypothesis, which is often used to supplement traditionalism. Additionally, we included preference-related control conditions in order to be able to deduce latent preferences for candidate readings from the answer patterns we gathered. To this end, we introduced a generative Bayesian model that gives probabilistic predictions about observable answer patterns in our task, given concrete instances of relative saliences of readings and other parameters. Using the observed data, in particular in connection with the preference-related control conditions, we inferred a posteriori likely values of the latent model parameters.

Backed up by our analyses, we concluded that both traditionalism and grammaticalism do not predict our data under the set of auxiliary assumptions that we provided them with in Section 3. Traditionalism fails, as a core theory, to account for high numbers of responses indicating local readings even for prosodically unmarked ES-sentences. Grammaticalism as a core theory is trivially compatible with our data, because it would be compatible with any observed preference order. But insofar as it builds on disambiguation by logical strength, grammaticalism fails to predict the absence of answers indicative of global readings and the relative abundance of answers indicative of literal readings.

With regard to grammaticalism, the clearest upshot of theoretical importance of this study is that strength-based meaning selection does not seem to be a good disambiguation criterion (c.f. Chemla and Spector, 2011). This point is also relevant in general, because disambiguation based on logical strength has been suggested in other domains as well (e.g. Dalrymple et al., 1998; Winter, 2001; Cobreros et al., 2012). We tentatively suggested an alternative selection principle for grammaticalism in terms of

"parsing economy," but it remains to be seen whether this is empirically successful in other domains as well.

Another important conclusion from our data is that ES-sentences, using non-monotonic quantifier *exactly one*, appeared to be prone to receive an unexpected interpretation. This may be used by traditionalism to explain answers indicative of local readings for ES-sentences. But the point is relevant in general. The possibility of unexpected interpretations of non-monotonic quantifiers should be considered also when interpreting the results of other experimental designs (e.g. Clifton and Dube, 2010; Chemla and Spector, 2011).

We have argued that auditory presentation of sentence material is crucial if prosodic information is hypothesized to influence availability and preferences of readings. We also argued in passing that theoretical positions must be formulated in such a way that they make testable predictions about reading preferences where different potential readings are allowed. We should then try to obtain information about actual preference patterns by explicit comparison with constructions whose preference structures are well-studied, ideally with some model of the data generating process for optimal data analysis.

A The auditory sentence material

	AS		ES			STATISTICS
	acc	**ntr**	**acc**	**ntr**		
R1	17.6	19.6	52.7	31.6		**QUANTIFIER: $F = 14.6$; $p < .01$** PROSODY: $F = 2.5$; p = .14 QUANTIFIER*PROSODY: $F = 3.8$; $p = .07$
R2	27.4	36.4	19.0	17.1		**QUANTIFIER: $F = 17.1$; $p = .001$** PROSODY: $F = 1.9$; $p = .19$ QUANTIFIER*PROSODY: $F = 3.4$; $p = .09$
R3	82.6	115.1	67.6	91.7		**QUANTIFIER: $F = 8.7$; $p < .05$** **PROSODY: $F = 20.5$; $p < .001$** QUANTIFIER*PROSODY: $F = .5$; $p = .50$
R4	32.6	60.6	17.6	28.4		**QUANTIFIER: $F = 16.4$; $p = .001$** **PROSODY: $F = 8.6$; $p < .05$** QUANTIFIER*PROSODY: $F = 1.9$; $p < .19$
R5	62.4	46.0	39.8	65.1		QUANTIFIER: $F = .1$; $p = .80$ PROSODY: $F = .3$; $p < .59$ **QUANTIFIER*PROSODY: $F = 5.2$; $p < .05$**
R6	211.3	69.9	174.3	75.1		QUANTIFIER: $F = 2.9$; $p = .11$ **PROSODY: $F = 94.8$; $p < .001$** **QUANTIFIER*PROSODY: $F = 13.7$; $p < .01$**
R7	57.6	51.7	30.4	24.6		**QUANTIFIER: $F = 9.4$; $p < .01$** PROSODY: $F = .6$; $p = .45$ QUANTIFIER*PROSODY: $F = .1$; $p = .80$
R8	36.7	71.4	32.5	53.5		QUANTIFIER: $F = 3.1$; $p = .1$ **PROSODY: $F = 45.5$; $p < .001$** QUANTIFIER*PROSODY: $F = .4$; $p = .51$
R9	47.1	61.5	42.3	60.7		QUANTIFIER: F = .2; p = .69 PROSODY: F = 3.3; p = .09 QUANTIFIER*PROSODY: F = .1; p = .7

Table 6: Difference between minimal and maximal F0 values in Hz for each of the single words in the target sentences. Region 6 corresponds to the determiner *einigen*.

	AS		ES			STATISTICS
	acc	**ntr**	**acc**	**ntr**		
R1	225.8	216.9	563.0	530.0		**QUANTIFIER: $F = 2251.3$; $p < .001$** **PROSODY: $F = 15.2$; $p < .01$** QUANTIFIER*PROSODY: $F = 3.6$; $p = .09$
R2	245.3	234.7	128.7	128.0		**QUANTIFIER: $F = 525.5$; $p < .001$** PROSODY: $F = 2.5$; $p = .13$ QUANTIFIER*PROSODY: $F = 3.2$; $p = .10$
R3	397.0	414.6	398.6	400.0		QUANTIFIER: $F = .3$; $p = .57$ PROSODY: $F = 1.2$; $p = .29$ QUANTIFIER*PROSODY: $F = 1.5$; $p = .25$
R4	172.3	175.6	164.7	152.7		**QUANTIFIER: $F = 16.1$; $p = .001$** PROSODY: $F = 2.1$; $p = .17$ **QUANTIFIER*PROSODY: $F = 8.7$; $p < .05$**
R5	230.2	170.5	226.7	170.0		QUANTIFIER: $F = .2$; $p = .7$ **PROSODY: $F = 79.0$; $p < .001$** QUANTIFIER*PROSODY: $F = .0$; $p = .90$
R6	445.8	422.0	420.6	383.0		**QUANTIFIER: $F = 31.4$; $p < .001$** **PROSODY: $F = 23.8$; $p < .001$** QUANTIFIER*PROSODY: $F = 1.0$; $p = .34$
R7	199.7	194.7	235.3	249.3		**QUANTIFIER: $F = 19.9$; $p = .001$** PROSODY: $F = .9$; $p = .36$ **QUANTIFIER*PROSODY: $F = 9.6$; $p < .01$**
R8	503.0	494.7	494.6	494.7		QUANTIFIER: $F = .04$; $p = .84$ PROSODY: $F = 2.2$; $p = .17$ QUANTIFIER*PROSODY: $F = 1.7$; $p = .2$
R9	722.0	748.0	704.0	764.7		QUANTIFIER: F = .0; p = .98 **PROSODY: F = 10.9; p < .01** QUANTIFIER*PROSODY: F = 2.9; p = .11

Table 7: Durational values in ms for each of the single regions in the target sentences. Region 6 corresponds to the determiner *einigen*.

	early	late	ntr		STATISTICS: EFFECT OF PROSODY
R1	31.2	23.1	31.7		$F = .9$; $p = .43$
R2	332.3	306.6	313.6		**$F = 7.3$; $p < .01$** **Early vs. late $F = 14.1$; $p = 001$** Early vs. ntr. $F = 1.1$; $p = .30$ **Early vs. late $F = 11.1$; $p < .01$**
R3	168.0	167.0	169.3		**$F = 4.2$; $p < .05$** **Early vs. late $F = 14.9$; $p = .001$** Early vs. ntr. $F = 1.7$; $p = .21$ Early vs. late F 3.7 = ; $p = .07$
R4	135.3	139.7	141.3		$F = 1.3$; $p = .39$
R5	561.3	1268.3	642.6		**$F = 129.3$; $p < .001$** **Early vs. late $F = 314.1$; $p < .001$** **Early vs. ntr. $F = 84.4$; $p < .001$** **Early vs. late $F = 39.2$; $p < .001$**
R6	143.3	176.3	150.0		**$F = 6.7$; $p < .01$** Early vs. late $F = 1.4$ $p = .24$ **Early vs. ntr. $F = 14.8$; $p = .001$** **Early vs. late $F = 5.1$; $p < .05$**
R7	1183.7	535.3	557.3		**$F = 132.5$; $p < .001$** **Early vs. late $F = 145.1$; $p < 001.$** **Early vs. ntr. $F = 303.9$; $p < .001$** Early vs. late $F = 1.5$; $p = .23$
R8	167.0	150.3	151.0		$F = .8$; $p = .47$
R9	384.7	381.0	402.7		$F = 3.1$; $p = .06$
R10	691.3	681.3	730.0		**$F = 25.8$; $p < .001$** **Early vs. late $F = 39.7$; $p < .001$** Early vs. ntr. $F = .2$; $p = .7$ **Early vs. late $F = 60.4$; $p < .001$**

Table 8: Difference between minimal and maximal F0 values in Hz for each of the single regions in the preference-related controls. Regions 5 and 7 correspond to the nouns preceding the boundaries.

	early	late	ntr	STATISTICS: EFFECT OF PROSODY
R1	105.7	100.7	98.1	***F* = 5.0; *p* < .05** Early vs. late: $F = 3.4$; $p = .07$ **Early vs. ntr: *F* = 9.1; *p* < .001** Late vs. ntr: $F = 1.7$; $p = .21$
R2	332.3	306.6	313.6	***F* = 28.1; *p* < .001** **Early vs. late: *F* = 70.4; *p* < .001** **Early vs. ntr: *F* = 28.7; *p* < .001** Late vs. ntr: $F = 3.0$; $p = .09$
R3	168.0	167.0	169.3	$F = .2$; $p = .77$
R4	135.3	139.7	141.3	$F = .7$; $p = .52$
R5	561.3	1268.3	642.6	***F* = 819.5; *p* < .001** **Early vs. late: *F* = 1110.4; *p* < .001** **Early vs. ntr: *F* = 94.4; *p* < .001** **Late vs. ntr: *F* = 680.1; *p* < .001**
R6	143.3	176.3	150.0	***F* = 16.1; *p* < .001** **Early vs. late: *F* = 25.3; *p* < .001** Early vs. ntr: $F = 1.1$; $p = .30$ **Late vs. ntr: *F* = 22.4; *p* < .001**
R7	1183.7	535.3	557.3	***F* = 1920.3; *p* < .001** **Early vs. late: *F* = 2251.4; *p* < .001** **Early vs. ntr: *F* = 2108.1; *p* < .001** **Late vs. ntr: *F* = 9.5; *p* < .01**
R8	167.0	150.3	151.0	***F* = 12.8; *p* < .001** **Early vs. late: *F* = 17.7; *p* < .001** **Early vs. ntr: *F* = 15.8; *p* < .001** Late vs. ntr: $F = .0$; $p = .85$
R9	384.7	381.0	402.7	***F* = 11.7; *p* < .001** Early vs. late: $F = .7$; $p = .48$ **Early vs. ntr: *F* = 11.2; *p* < .01** **Late vs. ntr: *F* = 22.7; *p* < .001**
R10	691.3	681.3	730.0	***F* = 10.0; *p* < .001** Early vs. late: $F = .8$; $p = .39$ **Early vs. ntr: *F* = 11.8; *p* < .01** **Late vs. ntr: *F* = 16.8; *p* < .001**

Table 9: Durational values in ms for each of the single regions in the preference-related controls. Regions 5 and 7 correspond to the nouns preceding the boundaries.

References

Allbritton, D.W. et al. (1996). "Reliability of prosodic cues for resolving syntactic ambiguities". In: *Journal of Experimental Psychology: Learning, Memory and Cognition* 22.3, pp. 714–735.

Atlas, Jay David and Stephen Levinson (1981). "It-clefts, Informativeness, and Logical Form". In: *Radical Pragmatics*. Ed. by Peter Cole. Academic Press, pp. 1–61.

Augurzky, P. (2008). "Prosodic phrasing in German sentence production: Optimal length vs. argument structure." In: *Proceedings of the ISCA Tutorial and and Research Workshop on Experimental Linguistics*. Ed. by A. Botinis. University of Athens, pp. 33–36.

Augurzky, Petra (2006). *Attaching Relative Clauses in German: The Role of Implicit and Explicit Prosody in Sentence Processing*. Vol. 77. MPI Series in Human Cognitive and Brain Sciences. Leipzig.

Bader, M. (1998). "Prosodic influences on reading syntactically ambiguous sentences". In: *RReanalysis in Sentence Processing*. Ed. by J.D. Fodor and F. Ferreira. Dordrecht: Kluwer, pp. 1–46.

Beach, C. M. (1991). "The interpretation of prosodic patterns at points of syntactic structure ambiguity. Evidence for cue trading relations". In: *Journal of Memory and Language* 30, pp. 644–633.

Benz, Anton and Nicole Gotzner (2014). "Embedded implicatures revisited: Issues with the Truth-Value Judgment Paradigm". In: *Proceedings of the Formal & Experimental Pragmatics Workshop*. Ed. by Judith Degen et al. Tübingen, pp. 1–6.

Breen, Mara et al. (2011). "Intonational phrasing is constrained by meaning, not balance". In: *Language and Cognitive Processes*, pp. 1532–1562.

Chemla, Emmanuel (2009). "Presuppositions of Quantified Sentences: Experimental Data". In: *Natural Language Semantics* 17.4, pp. 299–340.

Chemla, Emmanuel and Raij Singh (2014). "Remarks on the Experimental Turn in the Study of Scalar Implicature (Part I & II)". In: *Language and Linguistics Compass* 8.9, pp. 373–386, 387–399.

Chemla, Emmanuel and Benjamin Spector (2011). "Experimental Evidence for Embedded Scalar Implicatures". In: *Journal of Semantics* 28, pp. 359–400.

Chierchia, Gennaro (2006). "Broaden Your Views: Implicatures of Domain Widening and the Syntax/Pragmatics Interface". In: *Linguistic Inquiry* 37.4, pp. 535–590.

— (2013). "Free Choice Nominals and Free Choice Disjunction: The Identity Thesis". In: *Alternatives in Semantics*. Ed. by Anamaria Fălăuş. Hampshire: Palgrave Macmillan, pp. 50–87.

Chierchia, Gennaro et al. (2012). "Scalar Implicature as a Grammatical Phenonenon". In: *Semantics. An International Handbook of Natural Language Meaning*. Ed. by Claudia Maienborn et al. Berlin: de Gruyter, pp. 2297–2332.

Clifton, C. et al. (2002). "Informative prosodic boundaries". In: *Language and Speech* 45, pp. 87–114.

Clifton, Charles and Chad Dube (2010). "Embedded Implicatures Observed: A Comment on Geurts and Pouscoulous (2009)". In: *Semantics & Pragmatics* 3.7, pp. 1–13.

Cobreros, Pablo et al. (2012). "Tolerant, Classical, Strict". In: *Journal of Philosophical Logic* 41.2, pp. 347–385.
Conroy, Anastasia Marie (2008). "The role of verification strategies in semantic ambiguity resolution in children and adults". PhD thesis. University of Maryland, College Park.
Crain, S. and R. Thornton (1998). *Investigations in Universal Grammar: A Guide to Experiments in the Acquisition of Syntax and Semantics*. Cambridge, MA: The MIT Press.
Crawley, Michael J. (2007). *The R Book*. West Sussex: Wiley.
Cummins, Chris (2014). "Typicality made familiar: A commentary on Geurts and van Tiel (2013)". In: *Semantics & Pragmatics* 7, pp. 1–15.
Dalrymple, Mary et al. (1998). "Reciprocal Expressions and the Concept of Reciprocity". In: *Linguistics and Philosophy* 21.2, pp. 159–210.
Degen, Judith and Michael K. Tanenhaus (2011). "Making Inferences: The Case of Scalar Implicature Processing". In: *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Ed. by L. Carlson et al., pp. 3299–3304.
— (to appear). "Processing Scalar Implicatures: A Constraint-Based Approach". In: *Cognitive Science*.
Féry, Caroline (1993). *German intonational patterns*. Tübingen: Niemeyer.
Fodor, Janet Dean (1998). "Learning to Parse". In: *Journal of Psycholinguistic Research* 27.2, pp. 285–319.
— (2002). "Prosodic disambiguation in silent reading". In: *Proceedings of the North East Linguistic Society*. Amherst: GSLA, University of Massachusetts, pp. 113–132.
Fox, Danny (2007). "Free Choice and the Theory of Scalar Implicatures". In: *Presupposition and Implicature in Compositional Semantics*. Ed. by Uli Sauerland and Penka Stateva. Hampshire: Palgrave MacMillan, pp. 71–120.
Fox, Danny and Benjamin Spector (2009). "Economy and Embedded Exhaustification". Handout for a talk given at Cornell University.
Franke, Michael (2011). "Quantity Implicatures, Exhaustive Interpretation, and Rational Conversation". In: *Semantics & Pragmatics* 4.1, pp. 1–82.
— (2014). "Typical use of quantifiers: A probabilistic speaker model". In: *Proceedings of CogSci*. Ed. by Paul Bello et al. Austin, TX: Cognitive Science Society, pp. 487–492.
Frazier, Lyn (2008). "Computing Scalar Implicatures". In: *Proceedings of SALT 18*. Ed. by Lyn Frazier et al. Ithaca, NY: Cornell University, pp. 319–339.
Gazdar, Gerald (1979). *Pragmatics: Implicature, Presupposition, and Logical Form*. New York: Academic Press.
Geurts, Bart (2006). "Take 'five'". In: *Non-definiteness and plurality*. Ed. by Svetlana Vogeleer and Liliane Tasmowski. Amsterdam, Philadelphia: Benjamins, pp. 311–329.
— (2009). "Scalar Implicature and Local Pragmatics". In: *Mind and Language* 24.1, pp. 51–79.
— (2010). *Quantity Implicatures*. Cambridge, UK: Cambridge University Press.
Geurts, Bart and Nausicaa Pouscoulous (2009). "Embedded Implicatures?!?" In: *Semantics & Pragmatics* 2.4, pp. 1–34.

Geurts, Bart and Bob van Tiel (2013). "Embedded Scalars". In: *Semantics & Pragmatics* 6.9, pp. 1–37.

Goodman, Noah D. and Andreas Stuhlmüller (2013). "Knowledge and Implicature: Modeling Lanuage Understanding as Social Cognition". In: *Topics in Cognitive Science* 5, pp. 173–184.

Grabe, Esther (1998). *Comparative Intonational Phonology: English and German*. Vol. 7. MPI Series in Psycholinguistics. Wageningen: Ponsen en Looien.

Grice, Paul Herbert (1975). "Logic and Conversation". In: *Syntax and Semantics, Vol. 3, Speech Acts*. Ed. by Peter Cole and Jerry L. Morgan. New York: Academic Press, pp. 41–58.

Gualmini, Andrea et al. (2008). "The Question-Answer Requirement and Scope Assignment". In: *Natural Language Semantics* 16, pp. 205–237.

Horn, Laurence R. (1972). "On the Semantic Properties of Logical Operators in English". PhD thesis. UCLA.

— (2006). "The Border Wars: A Neo-Gricean Perspective". In: *Where Semantics Meets Pragmatics*. Ed. by Klaus von Heusinger. Vol. 16. Current Research in the Semantics/Pragmatics Interface, pp. 21–48.

Ippolito, Michela (2010). "Embedded Implicatures? Remarks on the debate between globalist and localist theories". In: *Semantics & Pragmatics* 3.5, pp. 1–15.

Jun, S. A. (2003). "rosodic phrasing and attachment preferences". In: *Journal of Psycholinguistic Research* 32, pp. 219–248.

Kentner, G. (2012). "Linguistic rhythm guides parsing decisions in written sentence comprehension". In: *Cognition* 123.1, pp. 1–20.

Kjelgaard, M. and S. Speer (1999). "Prosodic facilitation and interference in the resolution of temporary syntactic ambiguity". In: *Journal of Memory and Language* 40, pp. 153–194.

Konieczny, L. and B. Hemforth (2000). "Modifier attachment inGerman: Relative clauses and prepositional phrases." In: *Reading as a perceptual process*. Ed. by A. Kennedy et al. Amsterdam: Elsevier, pp. 517–528.

Kraljik, T. and S.E. Brennan (2005). "Using prosody and optional words to disambiguate utterances: For the speaker or for the addressee?" In: *Cognitive Psychology* 50, pp. 194–231.

Kruschke, John E. (2011). *Doing Bayesian Data Analysis*. Burlington, MA: Academic Press.

Lee, Michael D. and Eric-Jan Wagenmakers (to appear). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.

Magri, Giorgio (2009). "A Theory of Individual-Level Predicates on Blind Mandatory Scalar Implicatures". In: *Natural Language Semantics* 17, pp. 245–297.

— (2011). "Another Argument for Embedded Scalar Implicatures based on Oddness in Downward Entailing Environments". In: *Semantics & Pragmatics* 4.6, pp. 1–31.

Marty, Paul et al. (2014). "Phantom readings: The case of modified numerals". In: *Language, Cognition and Neuroscience*.

De Neys, Wim and Walter Schaeken (2007). "When People Are More Logical Under Cognitive Load: Dual Task Impact on Scalar Implicature". In: *Experimental Psychology* 54.2, pp. 128–133.

Partee, Barbara (1987). “Noun-phrase interpretation and type-shifting principles”. In: *Studies in discourse representation theory and the theory of generalized quantifiers*. Ed. by Jeroen Groenendijk et al. Dordrecht: Reidel, pp. 115–144.

Pierrehumbert, Janet B. and Julia Hirschberg (1990). “The meaning of intonational contours in the interpretation of discourse”. In: *Intentions in Communication*. Ed. by P. Cohan et al. Cambridge: MIT Press, pp. 271–311.

Plummer, Martyn (2003). “JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling”. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Ed. by Kurt Hornik et al.

Rapp, S. (1998). *Automatisierte Erstellung von Korpora für die Prosodieforschung*. Vol. 4. Arbeiten des Instituts für Maschinelle Sprachverarbeitung (AIMS). Stuttgart.

Russell, Benjamin (2006). “Against Grammatical Computation of Scalar Implicatures”. In: *Journal of Semantics* 23.361–382.

Sauerland, Uli (2004). “Scalar Implicatures in Complex Sentences”. In: *Linguistics and Philosophy* 27, pp. 367–391.

— (2010). “Embedded Implicatures and Experimental Constraints: A Reply to Geurts & Pouscoulous and Chemla”. In: *Semantics & Pragmatics* 3.2, pp. 1–13.

— (2012). “The Computation of Scalar Implicatures: Pragmatic, Lexical or Grammatical”. In: *Language and Linguistics Compass* 6.1, pp. 36–49.

Schafer, A. et al. (2000a). “Intonational disambiguation in sentence production and comprehension”. In: *Journal of Psycholinguistic Research* 29, pp. 169–182.

Schafer, A.J. et al. (2000b). “Intonational disambiguation in sentence production and comprehension”. In: *Journal of Psycholinguistic Research* 29.3, pp. 169–182.

Schulz, Katrin and Robert van Rooij (2006). “Pragmatic Meaning and Non-monotonic Reasoning: The Case of Exhaustive Interpretation”. In: *Linguistics and Philosophy* 29, pp. 205–250.

Schwarz, Florian et al. (2008). “Strengthening ‘or’: Effects of Focus and Downward Entailing Contexts on Scalar Implicatures”. Unpublished manuscript; retrived from semanticsarchive March 11 2011.

Snedeker, J. and J.C. Trueswell (2003). “Using Prosody to Avoid Ambiguity: Effects of Speaker Awareness and Referential Context”. In: *Journal of Memory and Language* 48, pp. 103–130.

Soames, Scott (1982). “How Presuppositions are Inherited: A Solution to the Projection Problem”. In: *Linguistic Inquiry* 13.3, pp. 483–545.

Spector, Benjamin (2006). “Scalar Implicatures: Exhaustivity and Gricean Reasoning”. In: *Questions in Dynamic Semantics*. Ed. by Maria Aloni et al. Amsterdam, Singapore: Elsevier, pp. 229–254.

— (2014). “Global Positive Polarity Items and Obligatory Exhaustivity”. In: *Semantics & Pragmatics* 7.11, pp. 1–61.

Steinhauer, K. and A.D. Friederici (2001). “Prosodic boundaries, comma rules, and brain responses: The Closure Positive Shift in ERPs a universal marker for prosodic phrasing in listeners and readers”. In: *Journal of Psycholinguistic Research* 30.1, pp. 267–295.

Steinhauer, K. et al. (1999). “Brain potentials indicate the immediate use of prosodic cues in natural speech processing”. In: *Nature Neuroscience* 2, pp. 191–196.

Sudhoff, Stefan (2010). *Focus Particles in German: Syntax, Prosody, and Information Structure.* Amsterdam: Benjamins.
van Tiel, Bob (2014). "Quantity Matters: Implicatures, Typicality, and Truth". PhD thesis. Radboud Universiteit Nijmegen.
— (to appear). "Embedded Scalars and Typicality". In: *Journal of Semantics*.
Toepel, Ulrike (2006). *Contrastive Topic and Focus Information in Discourse*. Vol. 48. MPI Series in Human Cognitive and Brain Sciences. Leipzig.
Uhmann, Susanne (1991). *Fokusphonologie*. Tübingen: Niemeyer.
Vaissière, J. (1983). "Language-dependent prosodic features". In: *Prosody: Models and Measurements*. Ed. by A. Cutler and D.R. Ladd. Berlin: Springer, pp. 53–66.
Winter, Yoad (2001). "Plural Predication and the Strongest Meaning Hypothesis". In: *Journal of Semantics* 18, pp. 333–365.