Scalar Items in Embedded Position: An Experimental Revisit

Fabian Schlotterbeck, Michael Franke and Petra Augurzky

Abstract

• What is needed to complement the picture is data from a verification experiment that is as independent as possible of potentially conflating properties of the pictorial material and that is nonetheless sensitive enough to detect local readings even if they are dispreferred.

Contents

1 Introduction

The existential quantifier *some* is usually assumed to receive a semantic interpretation similar to logical ∃, so that the sentence *Some boys cried* is literally true in a situation where all boys cried. But it is also usually considered to be a scalar item in that its use invites comparison with (at least) the semantically stronger universal quantifier *all* (c.f. Horn, 1972; Gazdar, 1979; Atlas and Levinson, 1981). This comparison can lead to an upper-bounding 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. ⇝ 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) (see Geurts, 2010, for recent overview), is that (1b) is a pragmatic inference, a so-called *quantity implicature*, derived by an abductive inference as the best explanation of why informed, knowledgable and cooperative speakers would utter (1a) when they could also utter the semantically stronger and relevant (1c).

This paper deals with the interpretation of two types of sentences, where the scalar item *some* occurs in the scope of other logical operators. In AS-sentences (short for ALL-SOME) as in (2) the scalar item *some* is embedded under universal quantifier *all*. In ES-sentences (short for EXACTLYONE-SOME) like (3) *some* takes scope under the non-monotonic quantifier *exactly one*.[a] According to current pragmatic theory, there are at least three relevant 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, according to Gricean intuition, we enrich utterances of (2) and (3) with the negation of alternative utterances of the corresponding sentences (4) and (5) where *some* is replaced by *all*; 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.

a. explain what non-monotonic means? here or later?

(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 relative abundance of theoretically conceivable readings, two interlocked empirical questions arise:

Q1: which of these readings are available to naïve subjects; and

Q2: which of the available readings are preferred.[1]

Addressing these questions empirically is relevant because they lie at the heart of the current debate about the exact location and nature of the interface between semantics and pragmatics. The controversy concerns the question to what extent the upper-bounding inference from *some* to *some but not all* is conventionalized within the compositional computation of semantic values. For that matter, it is particularly relevant to assess the availability and relative preference of local readings.

A number of empirical studies have already examined the availability of readings for AS- and ES-sentences (e.g. Geurts and Pouscoulous, 2009; Clifton and Dube, 2010; Chemla and Spector, 2011). However, results have not been as clear-cut as one might have hoped for: for instance, the available empirical evidence is inconclusive as to whether local readings exist (see also van Tiel, 2012, for related discussion), and previous studies did not explicitly address all preference relations between readings relevant to distinguish between theoretical positions. In the following, we will argue that these heterogeneous results were partly due to the use of standard procedures such as picture-verification paradigm, which might not be sufficiently sensitive for the phenomenon under investigation, for instance, due to possible confounding effects of the pictures used. Moreover, previous studies have only accumulated limited evidence pertaining to the second question that may help decide between theoretical positions, namely which of the attested readings subjects prefer. Finally, previous studies presented target sentences visually. But as it is often argued that focal stress on an embedded scalar item can favor a local reading (e.g. Horn, 2006; Geurts, 2009; Chemla and Spector, 2011; Geurts, 2010; van Tiel, 2012), it seems important to present target sentences auditorily, so as to be able to control for effects of "silent prosody" that subjects may apply when reading a sentence see Bader 1998, (Fodor, 1998) **b**.

b. provide references

In reaction to this situation, we therefore presented sentence materials auditorily and also employed a different kind of visual presentation of the pictorial material. This

[1] It may ultimately be impossible to keep unavailability and strong dispreference strictly apart. For our purposes this is not important. Every time we speak of availability, the so-inclined reader may think of non-negligible preference.

way we obtained behavioral data that is both indicative of the reading subjects assumed and independent of the potential pictorial and prosodic confounds. Adding to previous studies, we made sure that our method is able to reveal information about reading preferences by relating our targets to ambiguous sentences with a well-known preference on available readings.

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 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

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 case 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 given 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.[c] Notice that other arrangements of arrows might equally well serve as examples for the various situations.

[c] improve graphics

ES-sentences. The situation 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).

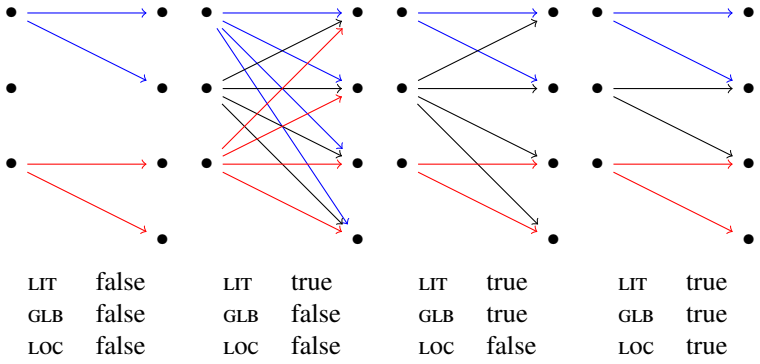
situation	truth value		
	LIT	GLB	LOC
false	0	0	0
literal	1	0	0
weak	1	1	0
strong	1	1	1

(a) AS-sentences

situation	truth value		
	LIT	GLB	LOC
false	0	0	0
literal	1	0	0
local	0	0	1
all	1	1	1

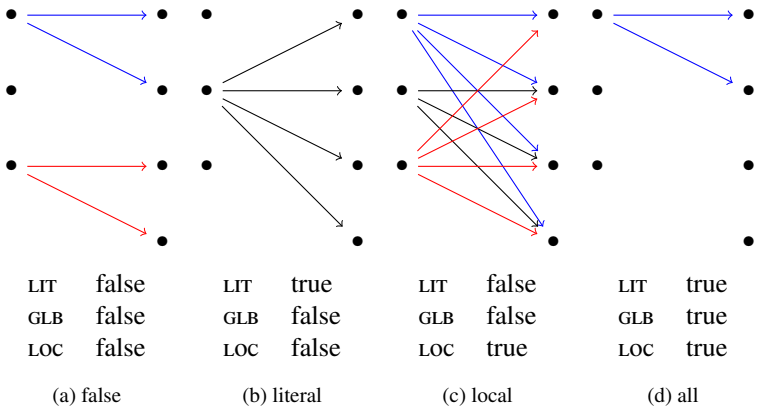
(b) ES-sentences

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



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

Figure 1: Distinguishing scenarios for AS-sentences



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

Figure 2: Distinguishing scenarios for ES-sentences

(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-linear:

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

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 possible 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 and named following Chemla and Spector (2011). Examples for each situation are given in Figure 2.

3 Theories and predictions

We consider three main theoretical positions which make different predictions about the readings of AS- and ES-sentences (c.f. Horn, 2006; Geurts, 2010; Sauerland, 2012,

for overview). We will refer to these here as traditionalism, conventionalism and grammaticalism and treat each one in turn. Since there is some leeway in assessing the predictions for each of these positions (depending on which of several reasonable additional assumptions we should adopt), we will distinguish different varieties of each. For convenience, the predictions of each (variety of each) position are also summarized in Table 2 at the end of this section.[d]

d. micha says: I'll leave all the varieties in here until we know which ones we need and which ones we don't need

3.1 Traditionalism

We refer to traditionalism as traditionalism because of its conservative stance towards Grice's original theory of conversational implicatures (Grice, 1975). Many author's have defended traditionalist positions in this sense. Of the more recent literature, we would consider as traditionalist, among others, contributions by Spector (2006), Sauerland (2004), Russell (2006), Schulz and van Rooij (2006), Geurts (2010) or Franke (2011).

According to Grice, 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 helpful information from the speaker to the hearer.

Consequently, the Gricean recipe (c.f. Geurts, 2010) for deriving a simple scalar inference like that in (1b) from an utterance of (1a) is as follows: 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,e]

e. do we need to enlarge on epistemic implicatures?

(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 Gricean recipe applies also to AS- and ES-sentences and derives the global reading in a straightforward way. Consequently, traditionalism predicts that both literal and global readings are available: literal readings, because these form the starting point of pragmatic reasoning; global readings because these may be arrived at by the Gricean recipe.

Which of these readings, if any, does traditionalism predict to be preferred? This depends on whether the auxiliary assumptions necessary to derive global readings by the Gricean recipe are plausibly met in the particular case of utterance of AS- and ES-sentences. These auxiliary assumptions include relevance of the extra information provided in the global reading, mutual awareness that the stronger alternative has been a speaker option, the speaker's competence about the issue, etc. Normally, traditionalist

[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), as this is not crucially relevant for the issues we are interested in here.

accounts would assume that these extra assumptions are met. In that case, traditionalism would predict that global readings should be preferred over literal readings. On the other hand, it might also be hypothesized that, for example, a competence assumption is harder to justify for AS- and ES-sentences in general than for simpler sentences such as (1), because of the additional quantificational element: it might be less clear that the speaker knows exactly how many students solved how many problems, than that the speaker knows exactly how many problems, e.g., Hans solved. In that case, or if any other assumption of the Gricean recipe cannot be maintained, traditionalism would predict that the literal reading would be preferred over the global one. But that means that there are at least two varieties of traditionalism that, depending on which additional assumptions we would make, yield slightly different predictions: the strong variety of traditionalism maintains that the auxiliary assumptions of the Gricean recipe hold usually/strongly/unless-completely-intenable and therefore predicts that global readings are preferred over literal ones; the weak variety of traditionalism holds that the auxiliary assumptions are more fragile and predicts that literal readings are preferred over the global ones.[3] **f**

f. acknowledge comment by Philippe Schlenker here;possibly expand

Does traditionalism also predict local readings to be available? The answer is slightly different for AS- and the ES-sentences. For ES-sentences, traditionalism does not predict that a local reading is available, at least not as a quantity implicature (c.f. Geurts and Pouscoulous, 2009; Chemla and Spector, 2011). This is because traditionalism assumes that quantity implicatures are pragmatic enrichments of the literal meaning of an utterance, obtained by conjoining the literal meaning with a suitable set of negated alternatives. But since the literal and the local reading of ES-sentences are logically independent, there is no way that local readings can be derived in a traditionalist manner as a quantity implicature: if X and Y are logically independent propositions, then there is no proposition Z such that X would be equivalent to $Y \wedge \neg Z$. Traditionalism often does concede that (something like) local readings can occur if scalar items are marked with *special accentuation*, albeit then as a signal of a different pragmatic process (e.g. Horn, 2006; Geurts, 2009, 2010). We will come back to this issue in Section 4.4.

On the other hand, as for AS-sentences, there is a traditionalist way of deriving the local reading, namely by assuming that not only (4) is an alternative to (2), but also the sentence in (8).

(2) All of the students solved some of the problems.

(4) All of the students solved all of the problems.

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

Clearly, if we conjoin a literal reading of (2) with the negation of (8) we obtain exactly the local reading. (Notice that the negation of (8) entails the negation of (4), so it makes no difference (not) to add it in this case.) Consequently, traditionalism does predict that the local reading of AS-sentences is available if (8) is an available alternative. We should

[3] We mention for clarity that although weak traditionalism does not predict the global reading to be preferred, it might still predict an *epistemically weak implicature*, similar to the global reading, that the speaker is uncertain whether the stronger alternative is true. Whether it does predict that depends on which auxiliary assumptions are assumed to be met and which are not.

therefore introduce another distinction: restricted traditionalism assumes that (8) is not available and therefore predicts that the local reading for AS-sentences is not available; in contrast, unrestricted traditionalism assumes that (8) is available and so is the local reading.

Still, even unrestricted traditionalism would not predict that the local reading is preferred over the global reading. This is because the derivation of the local reading via (8) hinges on yet another auxiliary assumption, namely the availability of the alternative in (8), which seems much less obvious and therefore presumably is less readily available than the alternative in (4). Consequently, we derive the following predictions about preferences for unrestricted traditionalism: weak unrestricted traditionalism prefers the literal over the global over the local reading, while strong unrestricted traditionalism prefers the global over the local over the literal reading (see also Table 2).

3.2 Conventionalism

Lexical conventionalism is a position that has been defended for instance by Levinson (2000) and Chierchia (2004).[4] The general idea is that if the upper-bounding inference from *some* to *some but not all* occurs frequently, then it would be inefficient to execute the Gricean reasoning that traditionalists propose over and over again. Conventionalism therefore assumes that scalar inferences have become routinized, hard-wired, so much so that there are two lexical entries for a scalar item like *some*, one with meaning *some and maybe all* and one with the meaning *some but not all*. Levinson (2000) proposed that the latter is the preferred default meaning, a position which we will refer to as defaultism.[5] **g** For completeness sake, we will call the opposite view that the literal meanings are *ceteris paribus* preferred literalism. We are not aware that anyone has defended this option, but it pays to include this option, because, as we will see later on, this possible position is the best direct explanation of our experimental data.

g. Petra says: empirical stuff in footnote irrelevant; I agree, let's move it to discussion section

Both varieties of conventionalism make interesting predictions about available readings for AS- and ES-sentences. Unaided, conventionalism does not predict that global readings are available. The lexically stored upper-bound for scalar *some* only allows to derive literal and local readings. Moreover, defaultism predicts that local readings are preferred over literal ones. Contrary to that, our dummy position of literalism predicts that literal readings are preferred over local ones.

3.3 Grammaticalism

The third and final kind of theoretical position we consider is grammaticalism. Recently, a lot of evidence based on intuitive judgements and more theoretical considerations has been discussed in favor of this position (c.f. Chierchia, 2006; Fox, 2007;

[4] See also Sauerland (2012) who argues quite favorably for a pragmatically informed conventionalism, although eventually dismissing it in favor of grammaticalism.

[5] Levinson's defaultism has strong empirical evidence against it, because several processing studies show that the computation of scalar inferences appears to be time- and effort-consuming, when they *do* arise, not when they *don't*, as defaultism would have it (c.f. Breheny et al., 2006; Breheny and Katsos, 2008; De Neys and Schaeken, 2007). But there is also competing evidence that scalar inferences are immediate, low-level processes (Grodner et al., 2010).

Magri, 2011; Sauerland, 2012; Chierchia et al., forthcoming; Chierchia, forthcoming). Grammaticalism approaches quantity implicatures by postulating a silent operator that can be variably applied during compositional computation of a sentence's truth-value (Chierchia, 2006), if necessary multiple times (Fox, 2007). This silent operator is variably referred to as O(·) or Exh(·), because it is assumed to be similar in effect to the meaning of particle *only* or of the mechanism of *exhaustive interpretation* (Groenendijk and Stokhof, 1984; von Stechow and Zimmermann, 1984; van Rooij and Schulz, 2004; Schulz and van Rooij, 2006; Fox, 2007). For our purposes, it is enough to note that Exh(·) is a poly-typed function that enriches an expression X, which crucially need not be a full proposition, based on a set Alt(X) of (suitable, relevant) alternatives to X that yields an enriched meaning of the form:[6]

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

As this operator can apply at various scope sites, the grammatical approach predicts that all three readings of AS- and ES-sentences are 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).
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).

The grammatical approach, as described so far, is rather flexible in that it makes many readings available. It is so flexible, in fact, that it must be constrained in some way or other to shield the approach against overgeneration. The easiest examples where this is crucial are occurrences of scalar items in downward-entailing environments, such as in:

(12) It's not the case that Hans solve some of the problems.

An assertion of (15) normally would be taken to convey (13a), not the disjunctive meaning in (13b).

(13) a. Hans solves none of the problems.
b. Hans solved none or all of the problems

The latter, however, is a reading that grammaticalism predicts to be available by letting the Exh(·)-operator take scope under negation, as in:

(14) It's not the case that Hans solved Exh(some of the problems)

[6]More sophisticated formulations of exhaustive interpretation have been proposed (e.g. Schulz, 2005; Schulz and van Rooij, 2006; Spector, 2006; Fox, 2007) but this simple formulation is sufficient for the purposes of this paper. Also, we gloss here over non-trivial details in the way this operation is to be specified exactly within a compositional semantics (c.f. Chierchia, 2006).

Indeed, this reading is not unattested, for witness a continuation of (12) as in:

(15) It's not the case that Hans solve *some* of the problems. He solved *all* of them.

But this reading seems to be marked, in that it requires exceptional contextual circumstances and a non-standard intonational stress.

To account for the markedness of certain examples, grammaticalism is therefore aided by a principle that specifies which readings are to be preferred if several readings are generated by optional applications of Exh(·) in the compositional computation of semantic values. The principle that grammaticalists adhere to (c.f. Fox and Spector, 2009; Chierchia et al., forthcoming; Chierchia, forthcoming) is the strongest meaning hypothesis of Dalrymple et al. (1998). Roughly speaking, if a given sentence has several conceivable readings, the strongest meaning hypothesis selects for the strongest of these. Chierchia et al. (forthcoming) discuss two concrete variants of the strongest meaning hypothesis: fix a sentence with propositional content S with candidate readings $C(S)$, where $C(S)$ contains S and all readings derivable from inserting Exh(·) at suitable scope sites;[7] then define *ceteris paribus* preferences among members of $X, Y \in C(S)$ as either

(16) X is preferred to Y iff $X \subset Y$ (Chierchia et al., forthcoming, (104))

or

(17) X is preferred to Y iff $X \subset Y$ and X and Y are readings obtained from applications of Exh(·) at exactly the same scope sites, except for one.
(Chierchia et al., forthcoming, (105))

Since the former is entailed by the latter, we will speak of weak grammaticalism and strong grammaticalism, depending on whether (16) or (17) is used.

These two varieties of grammaticalism give rise to different predictions about preferred readings. The preference for readings that weak grammaticalism predicts mirrors the entailment relations in (6) and (7): for AS-sentences weak grammaticalism predicts that local readings are preferred over global readings which in turn are preferred over literal readings; for ES-sentences weak grammaticalism predicts that global readings are most preferred, while literal and local readings are not ranked with respect to preference. Strong grammaticalism, on the other hand, predicts that AS-sentences preferably get either a local or a global reading rather than a literal reading, but does not rank these former two with respect to each other because they differ in more than one application of the Exh(·) operator. For ES-sentences, then, strong grammaticalism does not predict *any* preference relations between candidate readings at all.

Taking stock, traditionalism, conventionalism and grammaticalism predict different patterns of availability and preference for AS- and ES-sentences. Most notably, traditionalism does not predict local readings for ES-sentences to be available, while conventionalism does not predict any global readings to be available. Grammaticalism predicts that all readings are available, but, given the strongest meaning hypothesis, also comes

[7] Strictly speaking, we should compare *parses* of sentences with Exh(·) at various scope sites with respect to the readings that these parses give rise to (c.f. Chierchia et al., forthcoming), but we may ignore this detail here for ease of exposition.

	preference & availability	
theoretical position	AS	ES
traditionalism		
weak restricted	LIT > GLB	LIT > GLB
weak unrestricted	LIT > GLB > LOC	LIT > GLB
strong restricted	GLB > LIT	GLB > LIT
strong unrestricted	GLB > LOC > LIT	GLB > LIT
conventionalism		
defaultism	LOC > LIT	LOC > LIT
literalism	LIT > LOC	LIT > LOC
grammaticalism		
weak	LOC > GLB > LIT	GLB > LIT, LOC
strong	GLB, LOC > LIT	GLB, LOC, LIT

Table 2: Predictions of the relevant theoretical positions. Listed are the predicted preference relations. If a reading is not listed, it is predicted to be unavailable.

with a clear indication of preference: literal meanings are dispreferred. All predictions are summarized in Table 2.

4 Previous studies

In order to disentangle the divergent predictions of these opposing theoretical positions, a number of empirical studies have been carried out. In this section, we will focus on the influential studies by Geurts and Pouscoulous (2009), Clifton and Dube (2010) and Chemla and Spector (2011). Unfortunately, as we argue here, the conjoined evidence from all of these studies is inconclusive as to the availability and preference of relevant readings.

4.1 Geurts and Pouscoulous (2009)

Geurts and Pouscoulous (2009) conducted a picture-verification task to find out whether local readings of AS- and ES-sentences are available.[8] Subjects were presented with pictures like those in Figure 1c and 2c where the local reading gets a different truth-value from 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.[h]

h. Petra wants more detail in the description of this study; I'm not sure that this is needed...?

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

[8] We will focus here on a subset of the conditions tested by Geurts and Pouscoulous (2009). Moreover, note that 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 was obtained for ES-sentences.

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 supporting evidence), as well as based on empirical observations (Clifton and Dube, 2010; Chemla and Spector, 2011). In the following, we will review these empirical studies in some detail, as they constitute an important background for our own experiment. We will first consider a comment by Clifton and Dube (2010), which explicitly distinguishes between the availability of a reading and its preference.

4.2 Clifton and Dube (2010)

In a reply to Geurts and Pouscoulous's study Clifton and Dube 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:

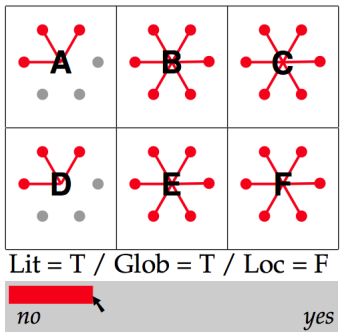
weak	strong	both	neither
3	39	57	1

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

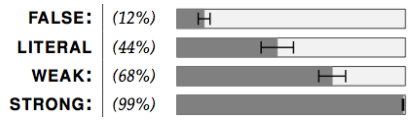
weak	literal	both	neither
28	6	50	17

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. Taken together, 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.

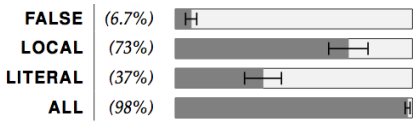
The diverging results of Geurts and Pouscoulous and Clifton and Dube demonstrate that participants' choices for AS-readings might be strongly affected by the specific experimental paradigm used. If task demands are only felicitous for a literal reading,



(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.

the *absence* of choices of local readings cannot be accounted for by assuming an unavailability of these readings *per se*.[i] However, at present it is less clear whether these results can also be generalized to further situations. For instance, subjects rejected the ES-sentences in Geurts and Poscoulous' original study, despite the fact that they could have accepted them based on the local reading. It is thus necessary to 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.

i. micha does not understand this sentence

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 of not) may induce a bias against non-preferred readings in cases where candidate readings stand in entailment relations (c.f. 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 (c.f. 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.

4.4 Reflection: Categorical Judgements & Typicality

Summing up so far, three previous studies have presented diverging pieces of evidence concerning the status of putative local readings. In their picture verification experiment Geurts and Pouscoulous (2009) did not observe any truth value judgments that would indicate local readings. On the basis of this observation, they argue that these readings are simply not available. On the other hand, Clifton and Dube and Chemla and Spector argue that the findings of Geurts and Pouscoulous may be due to a bias in their experimental design and employed different methods intended to be sensitive enough to reveal effects of local readings: either by a choice of the situation that better befits a given sentence (Clifton and Dube), or by a graded truth-value judgment (Chemla and Spector). Both studies found putative evidence in support of local readings.

To account for the diverging results we could draw a distinction between *categorical* truth-value judgments and other more sensitive measures. Firstly, it could be hypothesized 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 (see, however, Crain and Thornton (1998) for arguments in favour of truth-value judgments as a means of detecting highly dispreferred readings). Secondly it could also be argued, in line with traditionalism, that the apparent difference in sensitivity of the different judgment types can be explained by the stipulation that the results reported by Clifton and Dube and Chemla and Spector reflect something other than (strongly) truth-relevant speaker-intended meaning enrichments after all, so that no case can be made for lexicalism or grammaticalism on the basis of these data (c.f. van Tiel, 2012, for arguments along these lines).

Note, however, that the type of judgment is not the only crucial difference between the methods of Clifton and Dube and Chemla and Spector on the one hand and Geurts and Pouscoulous (2009) on the other. These studies also differed with respect to the *types* of situations (as distinguished in Figures 1 and 2) that were shown to the participants during the course of the experiment. The presence or absence of competing situations may have led to skewed distributions of judgments for situations like in Figure 1c. Such effects might emerge simply because the type of sentence-picture combinations shown during the experiment affects the standard of judgment but could – maybe more elegantly – also be explained by findings from Raffray and Pickering (2010) suggesting that exposing participants to models may prime corresponding logical forms and thereby shift reading preferences. Taking these possibilities into account, it remains unclear whether the effects reported by Clifton and Dube and Chemla and Spector would also show up using ordinary picture verification, in the first place.

For these reasons, we cannot ascribe the effects reported by Clifton and Dube and Chemla and Spector to the availability of local readings with certainty. Chemla and Spector (2011) themselves acknowledge the possibility that the *typicality* of a picture

with respect to some meaning may affect graded truth-value judgments. Strengthening this possibility, they show that graded judgments of AS-sentences differ significantly for different instantiations of the situation types from Figure 1, that vary in the amount of connections without changing the assignment of truth values to candidate readings. 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. The latter point is disputed by van Tiel (2012), who demonstrates that a huge chunk of variance in the data of Clifton and Dube and Chemla and Spector on AS-sentences can be explained as typicality effects, dispensing the need to appeal to the distribution of different readings. van Tiel (2012) elicited what he calls the typicality structures associated with the quantifiers *all* and *some* 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 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.

To conclude, although considerable progress has been made in understanding which factors have to be controlled for in order to asses the status of local readings, the evidence about availability, preference and nature of local readings is still inconclusive. What is needed to complement the picture is data that is as independent as possible of potentially conflating properties of the pictorial material and that is nonetheless sensitive enough to detect local readings even if they are dispreferred. Given the logical dependencies of the relevant readings this seems almost impossible to achieve. As we will argue below we do, however, believe it is.

On top of this, it would be desirable to obtain further information about the relative preferences among all putative readings, so as to be able to decide between theoretical positions, as outlined in Section 3. Neither of the previous studies gives us that, either because they simply have not investigated all the relevant comparisons (Geurts and Pouscoulous (2009), Clifton and Dube (2010)), or because they cannot safely derive the required information for danger of conflating preferences for readings with preferences for pictures (Chemla and Spector (2011)).

Finally, the above-mentioned studies did not control for possible conflating effects of "silent prosody". A growing number of evidence suggests that numerous factors might influence accent placement and prosodic phrasing even while reading, amongst them default accentuation, constituent length, rhythmic phenomena, and individual variation (Augurzky, 2008; Bader, 1998; Fodor, 2000, 2001; Kentner, 2012, Steinhauer et. al., 2001, see also the discussion in Frazier, 2009).[j] As it has been suggested in the theoretical literature that the availability of a local reading might hinge on the realization of contrastive stress on the scalar item, (e.g. Horn, 2006; Geurts, 2009, 2010; van Tiel, 2012), it is necessary to control for potential variation due to potential differences in silent accent placement. We therefore presented sentences auditorily and addressed the frequently proposed claim that local readings are triggered by contrastive stress.[k]

j. references

k. relate to Schwarz et al. (2008) and others?

5 Pilot Study

To start off, we conducted a pilot study investigating whether the diverging evidence reviewed above can be explained by considerations on how sensitive the different kinds of measures are. This information was critical to our main experiment below. Because the main experiment employed categorical truth-value judgments out of methodological necessity, we wanted to test the sensitivity of this measure independently. In addition, the question is interesting in its own right since it has implications about the nature of the ambiguity at hand, namely whether it is due to truth-relevant meaning enrichment.

In particular, we tested whether the critical effects Chemla and Spector (2011) found in their graded truth-value judgment task do also show up in a categorical truth value judgment task. In order to test this we simply replicated Chemla and Spector's experiment using a categorical truth-value judgment task. If, as has been suggested, categorical truth-value judgments are not sensitive enough to detect a difference of acceptance between situations like 1c and 1d given AS-sentences or between 2c and 2d given ES-sentences, these should be judged identical in a truth value judgment task.

In addition to the conditions from Chemla and Spector (2011) we also included another type of sentence. This type of sentence, EA-sentences, contained the scalar item *all* embedded under *exactly one*. Here, a rather atypical kind of global implicature would be arrived at by replacing *all* with its weaker alternative *some*. We tested whether this kind of implicature is detectable using a truth-value judgment task. In the discussion of the present experiment we will, however, mainly focus on the replication of Chemla and Spector (2011).

5.1 Materials

Test sentences were German cleft-constructions as in (18)–(20). Sentences like (18) are AS-sentences, (19) are ES- and (20) EA-sentences. We decided to use cleft constructions in order to fix the relative scope of the quantifiers involved. This was ensured via clause-boundedness. In total, 48 experimental items were constructed in all three conditions yielding a total of 144 sentences.

(18) Für jedes dieser Kinder gilt: es mag einige dieser Speisen.
For every of these children it holds it likes some of these dishes.
For all of these children it holds that it likes some of these dishes.

(19) Für genau eines dieser Kinder gilt: es mag einige dieser Speisen.
For exactly one of these children it holds it likes some of these dishes.
For exactly one of these children it holds that it likes some of these dishes.

(20) Für genau eines dieser Kinder gilt: es mochte jede dieser Speisen.
For exactly one of these children it holds: it likes some of these dishes.
Exactly one of these children it holds that it likes some of these dishes.

The sentences were paired with pictures like in Figures 1, 2, . For ES- and AS-sentences there were four types of situations just as in the study of Chemla and Spector

(2011). Sentences in the AS-condition were paired with situations as in 1b, 1a, 1c, 1d. Sentences in in the ES-condition were paired with 2b,2b,2c and 2d. The EA-sentences were paired with pictures like in 5.1. Figure **??** is only true under the literal reading (*literal* diagrams) while Figure **??** is also true with respect to the global implicature (*global* diagrams) denying the weaker scalar alternative containing *einige* (some). Diagrams like those in Figures **??** and **??** were not compatible with the EA-sentences. Figure **??** is incompatible because there are too few connections. Figure **??** is incompatible because there are too many connections. We used two types of incompatible pictures only because of technical reasons.

In total, this yielded twelve conditions and 576 pictures. The conditions were distributed over twelve lists according to a latin square design ensuring that in each list every item was only present in one condition. In addition, to the target sentences, 201 (?) distractors were included in each list. Of these *n* were intended to yield a "yes true" response, *m* were false.

5.2 Procedure

The present pilot study was conducted as part of a dual task experiment employing picture verification and a sternberg task Sternberg, 1966. In some trials participants had to solve a memory task in addition to the picture verification task. In the *high load* condition participants had to remember six letters while performing the picture verification. In the *low load* condition they had to remember one letter and in the *no load* condition they didn't have to remember anything while they were performing the picture verification. In the *memory only* conditions they only had to perform a memory task (i.e. remember one or six letters) but no picture verification. All the critical items of the present experiment were presented in the *no load* condition. Only distractor items were presented in one of the dual task conditions.

In each trial participants were first asked to indicate readiness by pressing a button. In the *no load* condition, after they had pressed the button an asterisk was flashed on the screen six times within $7.2s$. Then a sentence and a picture were presented simultaneously. Within $12s$ participants had to judge whether the sentence was true with respect to the picture. They indicated their judgment by pressing one of two buttons. After they had provided their judgment the next trial started automatically. If they did not respond within twelve seconds, the next trial was started. Judgments and reaction times were logged.[9]

The experiment started with a practice session consisting of 25 trials divided into three blocks (1st block: memory task only, ten trials; 2nd block: picture verification only, five trials; 3rd block: dual task, ten trials). Then one list consisting of 249 trials divided into three blocks was presented. An experimental session took about 75

[9] In the *low load* and *high load* condition a trial looked similar to a trial in the *no load* condition except for two things. Firstly, in the dual task conditions instead of asterisks letters were flashed on the screen. Secondly, after finishing the picture verification task a question appeared on the screen probing for a letter. Participants had to indicate by pressing a button whether they thought the letter had been present in the array from the beginning of the trial. Except for one difference, the *memory only* trials were identical to the dual task trials. In *memory only* trials the picture verification part was replaced with a blank screen. That is, in the middle of such trials the screen stayed blank for $5s$.

minutes. During the whole experiment no feedback was provided to the participants.

5.3 Participants

60 native German speakers (40 female, mean age: 24, ranging between 20 and 40) from the University of Tübingen participated in the study. Participants were naïve to the purpose of the study. They were paid €10 compensation.

5.4 Results

Overall participants were able to solve the task. In the distractor items the picture verification was done accurately in 88% (sd : 33%) of the cases. All participants were above 70% and significantly better than chance. On average it took participants 4989ms (sd : 2176ms) to come up with their answer in the filler trials. We will now consider results for the three sentence conditions in turn.

The mean judgments we obtained for AS-sentences are depicted in Figure **??**. The *strong* diagrams were judged to be true in 90% of the cases, *weak* diagrams received "yes" judgments in 79% , *literal* diagrams in 72% and *false* in 0.9% of the trials. A logit mixed effects model with binomial link function containing a fixed effect of *diagram* as well as random effects of items and participants revealed a clear effect of *diagram*. In particular *false*, *literal* and *weak* diagrams all received fewer "yes" judgments than *strong* diagrams (all $p < .001$). Another logit mixed effects model only modeling data for *literal* and *weak* diagrams also revealed an reliable effect of *diagram* (p<.05). Finally, a model only considering *false* and *literal* diagrams also revealed a significant effect of *diagram* (p<.001).

The responses for ES-sentences are summarized in Figure **??**. Unsurprisingly, these sentences were accepted as truthful description of *false* diagrams in only 0.8% of all cases. The *local* diagrams received "yes" responses in 20%, *literal* in 71% and *all* in 93% of the time. A logit mixed effects model with *diagram* as fixed effect as well as item and participant as random effects revealed a significant effect of *diagram*. The *all*, *literal* and *local* diagrams were all accepted more often than the *false* diagram (all $p < .001$). A simpler model only considering *local* and *literal* diagrams revealed a reliable effect of *diagram* (p<.001) as did a model only considering *literal* and *all* diagrams (p<.001).

The mean responses for the EA-sentences are shown in table **??**. The observed difference between the global and literal diagrams is significant ($t = -3.8259$, $df = 322.375$, $p < 0.001$).

AS-sentences			
literal	global	false1	false2
98.3	90.3	7.1	3.3

5.5 Discussion

Concerning the AS- and ES-sentences the observed distribution of judgments were rather similar to the results from Chemla and Spector. Qualitatively, the only notable differ-

ence lies in the rather low acceptance of situations like 1b in the present experiment as compared to Chemla and Spector's original study. Crucially, similar to the findings of Chemla and Spector, we found that situation 1d was more often accepted than situation 1d and that situation 2c was accepted more often than situation 2a. This suggests that local readings are available and are even detectable via categorical truth-value judgments.

However, we have to be careful with our interpretation of the results. Since the picture materials differd drastically in the critical comparisons, we cannot exclude pictorial effects. As reviewed above van Tiel did even show that with regards to the AS-sentences the differences in acceptance reported by C&S and therefore presumably also the replicated results presented here can be explained by typicality of the pictures with respect to the sentence meaning.

That truth is not the only determining factor affecting truth-value judgments also becomes evident from the distribution of judgments we obtained for the EA-sentences. The observed difference between **??** and **??** is a curious finding at first sight. Firstly, the literal reading is true in both situations. Secondly, from the point of view of implicature calculation, the only thing that could plausibly explain an acceptability difference between these two situations is the presence of a global implicature, resulting from the negation of the corresponding ES-sentence.[10] However, this putative global implicature would be true in the situation **??**, but false in the situation **??**. Consequently, the observed difference in acceptability cannot be explained simply in terms of the hypothesis of Chemla and Spector (2011) that acceptability increases with the number of available true readings. Rather this finding, just as van Tiel's study, suggests that some feature of the pictorial material that is independent of the putative readings under scrutiny does affect truth-value judgments, even categorical ones.

We conclude, that effects of pictorial complexity and/or typicality are real and affect even categorical judgments. For this reason the following experiment was set up to investigate the availability of local readings while minimizing pictorial effects such as typicality or complexity.

6 An Incremental Verification Task

6.1 Design

General Idea. In order to test the availability and preferences of different readings of AS- and ES-sentences in German, we were looking for a way to unambiguously map responses from a picture verification task to specific readings. We were particularly interested in the question whether local readings would be observed at all, and whether their presence or absence would be modulated by a contrastive accent on the scalar item. Moreover, we were interested in obtaining information about the relative preferences of the different readings. In addition, we wanted to minimize effects of graphical

[10] Notice that (**??**) is the alternative to the corresponding ES-sentence needed for computing the putative global implicature. But these sentences are logically independent. So, it may be that the corresponding ES-sentence is an alternative also to sentences like (**??**). It may also be that being an alternative is not a symmetric notion. Be that as it may, our argument does not hinge on whether (**??**) does or does not actually have a global implicature of the alleged kind.

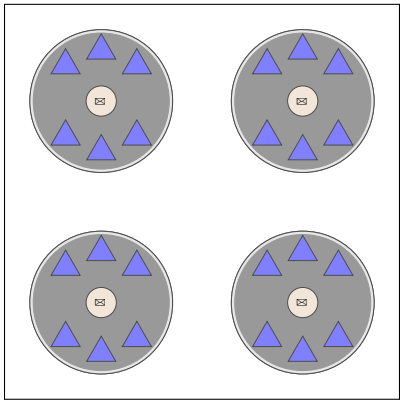


Figure 4: Initial picture presented to participants where all possible connections from each letter to its surrounding triangles are covered.

differences between the picture materials. To achieve this, we designed an experiment employing a modified version of picture verification, namely the incremental verification task (IVT, see Conroy ???)[l]. The general idea is that subjects are requested to judge sentence material based on pictures that do not necessarily contain all the information necessary to judge a certain reading true or false. In that case, participants can demand that more information is revealed. Participants are instructed to make a truth-value judgment as soon as they are able to.[m]

l. reference

m. need to mention somewhere that trial ends with each TV judgment

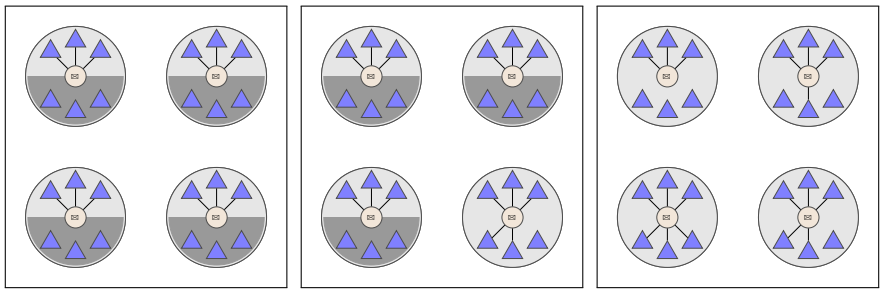
Target Conditions. In the present study, participants were presented pictures depicting a set of four identical central elements (e.g., letters), which could be connected to surrounding elements (e.g., triangles). Initially, any potential connections between central and surrounding elements were covered by dark gray color (see Figure 4). Sentences to be judged were presented auditorily, and participants were asked to listen to a given sentence and then uncover the potential links step by step until they felt able to give a truth value judgment. Three options were available for participants at each step: (i), demand more information, (ii), judge the sentence as true, and (iii), judge the sentence as false. 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, which we will refer to as *critical position*. The location in the sequence and the corresponding truth-value judgment differed between AS- and ES-sentences, as described presently. (This is, again, because of the logical dependencies between readings.)

Consider the AS-sentence in (21).

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

Figure 5 shows the three corresponding critical positions for this sentence in the order of the uncovering process. First, the situation in Figure 5a becomes available, which is

true under a literal reading, while the local and global readings cannot be evaluated yet. Second, in Figure 5b, the literal reading is still true, but now the global reading can be additionally confirmed by a *true* response, as all the connections of one of the letters are now uncovered, and it is now visible that one of the letters is not connected to all of its triangles. Finally, decisions concerning the local reading are possible as soon as all connections have been uncovered. In this case, the critical position shown in Figure 5c is incompatible with a local reading: One of the central elements is linked to all of its surrounding elements, thus yielding to *false* answers if participants indeed adopted the local reading. Note that associating the final position with *false* answers allows us to exclude both literal and global readings, which correspond to a *true* answer at this position. In sum, *true* or *false* answers on particular positions in the incrementally revealed picture sequence can be mapped uniquely to candidate reading. All other *true* or *false* answer were counted as errors. The mapping between responses and readings is summarized in Table 3a.



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

Figure 5: Example sequence for AS-sentences

	more info	true	false
Step 1	-	LIT	error
Step 2	-	GLB	error
Step 3	n.a.	error	LOC

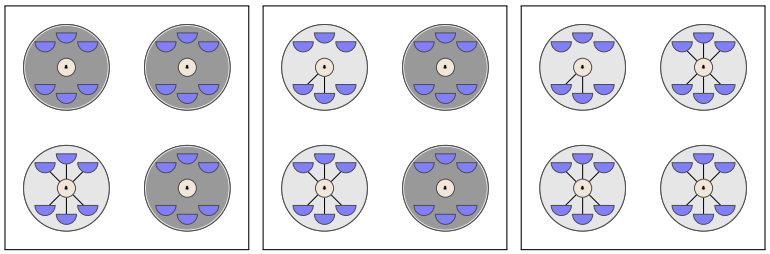
(a) AS-Sentences

	more info	true	false
Step 1	-	error	GLB
Step 2	-	error	LIT
Step 3	n.a.	LOC	error

(b) ES-Sentences

Table 3: Mapping from response types to readings. Notice that on steps 3 the option "more info" is not available because the whole picture has been revealed.

ES-sentences, due to the non-linear relationship of candidate readings described in Section 2, require a slightly different sequence of unfolding which yields a different order in which truth-value judgments can be made. For a sentence like (22) we obtain an unambiguous mapping between truth values and readings with respect to the sequence in 6.



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

Figure 6: Example sequence for ES-sentences

(22) 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.

At the first critical position shown in Figure 6c, the global reading corresponds to a *false* response, as one of the bells is already linked to all of its semicircles. The literal reading can be evaluated in Figure 6b, where it would trigger a *false* response, as there are two bells that are connected to at least some of their surrounding elements. Finally, confirming 6c by a *true* response corresponds to the local reading, as one of the bells is linked to some but not all of its semicircles, while all of the other bells are linked to all of them. Note that again, global and literal readings differ from local readings with respect to their truth values. Mappings between responses and readings are summarized in Table 3b.

Positional controls. In order to make sure that subjects understood the task, i.e., gave truth-value judgements neither sooner not later than at the first possible position in a sequence, we included a set of control conditions. These also controlled for response biases that are independent of the sentence meaning, but only depend on the experimental procedure or on the picture materials. For each of the three readings in the AS- and ES-conditions, we constructed an unambiguous control sentence as in (23) and (24), requiring the same judgement at the identical position in the same sequence used for the respective targets. For instance, example (23a) requires a *true*-response analogous to the AS-sentence in (21) under its literal reading, (23b) corresponds to the AS-sentence under its global reading and (23c) corresponds to its local reading. With regard to the ES-sentences, controls like (24a) correspond to the global reading, (24b) to the literal and (24c) 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.

(23) 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.

(24) 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 control for artificial effects of presentation order on the availability of readings, we included a second type of control conditions. In particular, we tested (i), whether the order of given readings affects preferences for ambiguous sentences and (ii), whether prosodic information can, in principle, shift reading preferences in the present task. 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 thus in principle be possible that no subject ever reaches this point, because they gave truth-value judgements earlier, thereby ending the trial. In that case it would be unclear if these decisions have been affected by a general inavailability 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 intended to test whether participants in our task occasionally choose dispreferred readings, even if they were available only at a critical position following the preferred readings.

We therefore included globally ambiguous *late-closure structures* as in (25), which have been repeatedly shown to exhibit interpretive preferences.

(25) 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)

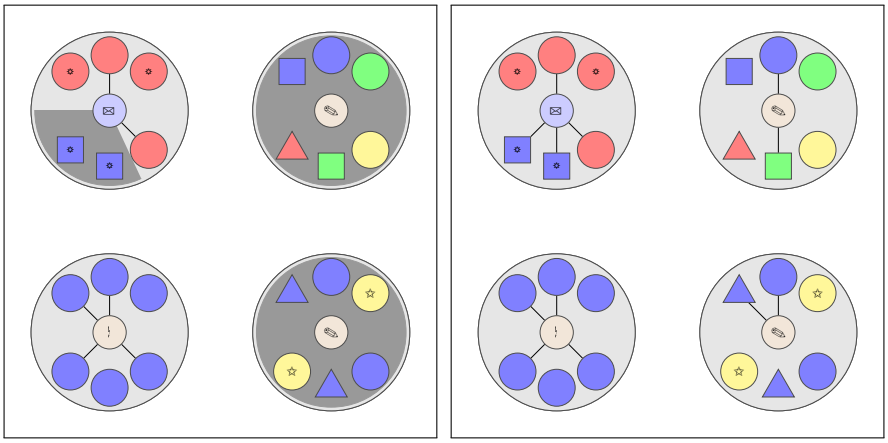
(26) Der Brief ist mit Kreisen, die Sonnen beinhalten und Vierecken
The letter is with circles, which suns contain and squares
verbunden.
connected.
The letter is connected with circles containing suns, and with squares.

In late-closure sentences, a modifier can be attached to one of two preceding hosts. More specifically, in (25), the prepositional phrase *with suns* can be attached to the preceding noun *squares*, resulting in the so-called *late-closure* or *low-attachment* reading ((25a), see Frazier, 1987, for details)[n]. Alternatively, the PP can be attached to the whole conjoined NP *circles and squares*, corresponding to the *early-closure* or *high-attachment* reading (25b). Generally, the late-closure reading is preferred over the early closure reading for sentences comparable to those in (25) (REFS)[o].

n. references

o. references

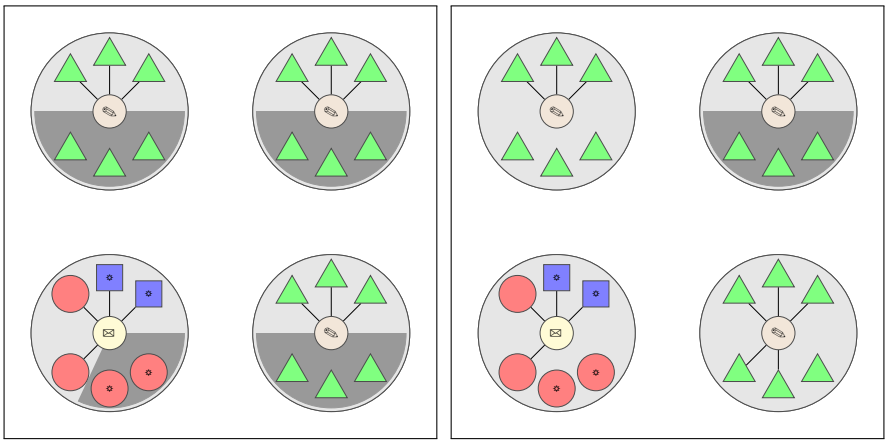
For each sentence, two sequences were designed. In the first, the critical position associated with the dispreferred early-closure reading preceded the critical position for the preferred late-closure reading. In this case, a partly covered picture as in Figure 7a preceded a picture as in 7b. If participants adopted an early-closure reading as in



(a) Step 1 (b) Step 2

Figure 7: Critical steps of a sequence for sentence (25) where the early-closure reading (25b) can be judged first.

(25b), a *false*-judgment would be expected at the position illustrated in 7a, as none of the circles connected to the letter contain any suns. Under the late-closure reading in (25b), in contrast, a *true*-response would be expected no sooner than in the step shown in Figure 7b. The order of possible judgements is reversed in the second kind of sequence including pictures like 8a followed by pictures like 8b. Here, the late-closure reading in (25a) can be judged first requiring a *true*-response on 8a. The early-closure reading in (25b) can only be judged later in the sequence. When we reach 8b a *false*-response is expected under an early-closure/high-attachment reading.



(a) Step 1 (b) Step 2

Figure 8: Critical steps of a sequence for sentence (25) where the late-closure reading (25a) can be judged first.

Testing Effects of prosody. [p]

The availability of local readings has often been claimed to be affected by the presence or absence of a focal accent on the scalar item. By contrastively accenting the scalar, potential alternatives are highlighted, resulting in more implicatures being computed. Some approaches have even claimed that local readings exclusively emerge when the scalar item is accented and are otherwise generally unavailable (Refs).

Though there seems to be general consensus about these potential prosodic effects, empirical evidence for this assumption is still scarce. One notable exception comes from a pilot study by Frazier, in which participants were instructed to read embedded and non-embedded versions of English ae sentences, in which some was either capitalized or non-capitalized. Prior to the experiment, participants were told that capitalizing corresponded to contrastive stress. Interestingly, strengthened interpretations occurred rather often in the embedded sentences in this task (59 %), but were not affected by capitalizing. ... Another study relevant for the present consideration comes from Schwarz et al., 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, participants were presented two alternative readings (*1. She will invite Fred or Sam or possibly both*, or *2. She will invite Fred or Sam but not both*).

We also wished to make sure that differences in prosody may generally lead to a shift in interpretation preferences in the ITV task. Sentences like (25) are well suited to test this assumption since it is well-known that their interpretational preferences are sensitive to prosody. If an utterance of (25) includes a prosodic phrase boundary between *squares* and *with*, this intonational pattern strongly corresponds to the high-attachment reading. If, instead, a prosodic phrase boundary after *circles* is present, the acceptance of the low-attachment reading is boosted. To test whether prosodic infor-

p. These two paragraphs still need a proper home; Petra opts for “materials” section.

mation is taken into account in the ITV we presented both types of prosodic structures expecting the described boost in acceptance for the low-attachment reading. In addition, we also included a neutral version containing no phrase marking whatsoever. This should receive intermediate acceptance rates. Further, we also presented unambiguous sentences corresponding to the early-closure 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.

6.2 Procedure

6.3 Materials

Target sentences and positional controls We constructed a set of 15 items for AS- and ES-sentences, respectively, analogous to the examples in 21 and 22. 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. Subjects 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* (see section xx below). 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 (see below). 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, 28 sentences that were ambiguous between a late-closure and an early-closure reading, as well as 28 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 *mit*-PPs. The first of these PPs consisted in a coordination of two nouns denoting geometrical objects, whereas 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 A phonetically trained female native speaker of German was instructed to realize two prosodic versions of each target sentence, and three versions of the ambiguous preference-related controls. In addition to these sentences, the unambiguous preference-related controls, the positional controls and a set of unrelated fillers were also recorded. For these constructions, the speaker was instructed to realize prosodic contours as neutral and natural as possible.

For all target sentences (AS and ES), the determiner *einige* was produced with a contrastive pitch accent as well as with a neutral accent. Contrastive accents in German

are realized by an H*L contour (XYZ)[q], thus employing a generally falling pattern. A set of 15 experimental 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 (XYZ)[r]. Boundaries for these control sentences were either realized at the position separating the second PP from the preceding material (*late boundary*, corresponding to a high-attachment reading) or directly preceding the second conjunct in the first PP (*early boundary*, corresponding to a low-attachment 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, in which our speaker produced the sentences without any pronounced boundaries. For each prosodic variant, a set of 30 items was read, yielding 90 preference-related fillers.

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 fillers 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. As 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*, see above (XYZ)[s]), we considered the following analysis regions:

(27) a. $|_{R1}$ Alle $|_{R2}$ dieser $|_{R3}$ 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 fillers 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. Note that 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).

(28) $|_{R1}$ 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 (XYZ),[t] and the obtained values were manually corrected afterwards. For the targets, two-factorial ANOVAs with the factors QUANTIFIER (all vs. exactly one) and PROSODY

q. insert references

r. references: Vaissiere, 1983; Fery, 1993

s. proper reference

t. reference! Rapp, 1998

(accented vs. unaccented) were carried out. For the preference-related fillers, 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 s were carried out for statistical comparison of the target sentences, and one-factorial ANOVAs s were carried out for preference-related controls.

The details of these analyses are found in Appendix A. 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 fillers 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 comparable to those illustrated in Figure 4. In each picture, four alternative subscenarios 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. As described in section xx, grey colour initially covered the links between icons and shapes. As the connections were to be uncovered step by step, different pictures were designed corresponding to different stages in the uncovering process. Note that the single conditions impose distinct requirements on sequences. For instance, inherent entailment relations between AS-sentences and ES-sentences lead to different critical positions at which participants can make their judgments. We will therefore describe the sequences for each condition separately.

For AS-sentences, as well as for the corresponding positional control sentences described in Section 6.1, we designed a set of seven pictures for each trial. In each sequence, the critical position associated with the literal meaning was presented at the third step, the position associated with the global meaning was shown at the fifth step, and the local reading was presented at the final step (see Figure 5 for an illustration of the critical positions). Pictures that were interspersed between these positions served as spillover-pictures. These were used in order to control for potentially delayed judgments and they additionally served as distractor items.

Sequences in the ES conditions (and their respective positional controls) consisted of a set of five pictures with the following critical positions. The critical positions as illustrated in Figure 6 are associated with the second picture (global reading), the third picture (literal reading), as well as with the final picture (local reading).

For the preference-related controls, we used two types of sequences. First, a sequence in which the cririca position for the (preferred) Late Closure reading was presented preceding the critical position for the Early Closure reading (*LC sequence*), and second, a sequence in which the readings were presented in reversed order (*EC se-*

quence. In the LC sequences, seven pictures were presented, with the LC reading given at the third picture, and the EC reading at the sixth picture. EC sequences consisted of five pictures, in which the critical positions were realized at the third picture (EC reading) and at the fifth picture (LC reading).

Finally, we added pictures compatible with unrelated fillers with critical position at random uncovering stages, including cases where no uncovering was needed for a truth-value judgement.

Lists. Experimental items and preference-related fillers 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.

6.4 Participants

41 native speakers of German took part in our study, none of which had any prior exposure to logic or formal semantics. We excluded 3 subjects due to insufficient performance on controls ($\leq 50\%$ correct answers).

- possibly mention ages, backgrounds and sexes?

6.5 Results

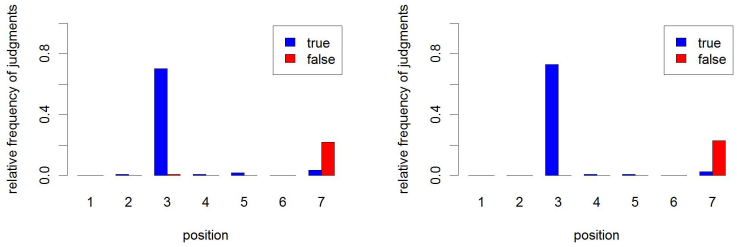
Performance on basic control conditions was high (92% correct responses on average), indicating that subjects understood the task, giving truth-value judgements at the "right" moment in a sequence without being influenced in general by inessential features of the picture material or showing general response biases.

The judgments obtained for the four target conditions are depicted in Figure 9.[u] We coded the judgments as *literal*, *global* or *local* if they were as expected under one of these readings and as *error* if not. The distribution of readings thus coded is presented in Figure 10. Participants mostly gave judgments indicating literal or local readings. Judgments compatible with a global reading were seldom obtained. In the AS-conditions participants made hardly any errors, whereas the amount of errors was considerably larger in the ES-conditions.

In the AS-conditions (see Figure 10a) judgments were consistent with literal readings in 65.0% of the trials with neutral prosody and in 67.5% of the trials with accented prosody.[v] The number of global readings was very low, with 1.6% of the neutral and 0.8% of the accented trials (corresponding to a total of 3 answers across all trials). Judgments indicating local readings were given in 22.0% of both the neutral and accented trials. In the ES-conditions (see Figure 10b) judgments consistent with literal readings were given in 47.2% and 44.7% of the trials with neutral and accented prosody respectively. Global readings were observed in 0.8% of the trials in both the neutral and the accented condition. Judgments indicating local readings were given in 22.8% of the neutral and 27.6% of the accented trials.

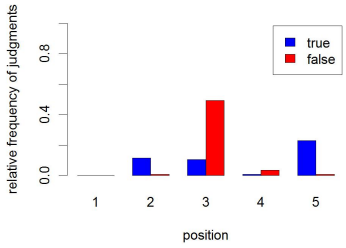
u. we should introduce the actual sequences that we presented so that it is clear what it means to say *true* or *false* on some position or other

v. IMPORTANT: Figure contains different numbers (even if rounded)! Fabian, did you exclude the 3 bad performers when you obtained this data?

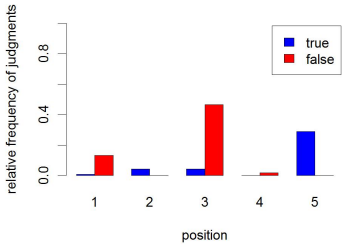


(a) AS-accented

(b) AS-neutral

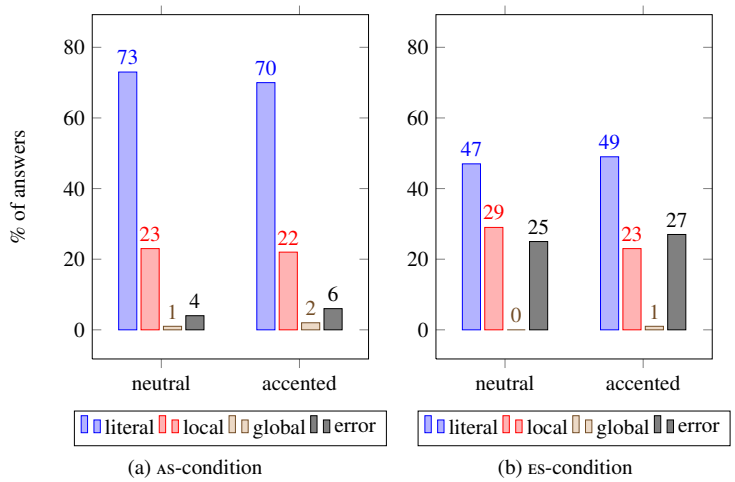


(c) ES-accented



(d) ES-neutral

Figure 9: Judgments for the four target conditions in Experiment 1. Relative frequency of *yes*/*no* judgments is plotted against the position within the trial. In the AS-conditions there were seven positions. On the third position a *yes* judgment corresponded to a literal reading, while on the fith position a *yes* judgment corresponded to a global reading. A *no* judgment on the seventh position was only consistent with a local reading. The ES-conditions had five positions. *No* judgments on positions two and three corresponded to global and literal readings, respectively. A *yes* judgment on the last segment corresponded to a local reading.



(a) AS-condition (b) ES-condition

Figure 10: Response type frequencies for target conditions

In order to test wether local readings exist we tested whether the number of local responses in each condition was higher than expected by chance. Local responses had to be given on the last position in each trial. If local readings didn't exist we would expect participants, who reached the last position (ie. who did not abort the trial prior to the last position) to give local judgments not more often than expected by chance. On the last position two responses were possible. One was compatible with a local reading the other was not. Therefore, we would expect 50% local responses by chance. As it turns out local responses were given significantly more often than 50% in all target conditions (all Bonferoni corrected p<.05). Note that this finding cannot be explained by some kind of general response bias on the last position since local readings required a yes-judgment in the ES-conditions but a no-judgment in the AS-conditions.

In order to test whether accentuation or the quantifier has an influence on the distribution of readings, log-liner models were computed (see Schepers 2003). The factors *Reading*, *Accentuation* and *Quantifier* were included in these models. Two variants of these models were computed. In the first, the factor *Item* was added to the above mentioned. In the latter *Participants* was included as a factor. Inclusion of these two factors allowed us to test whether the distribution of readings was identical accros items and participants. We report log-likelihood ratio Chi-squares ($LRCS_1$ and $LRCS_2$), degrees of freedom (df_1, df_2) and significance levels (p_1 and p_2).

Unsurprisingly, there was an effect of *Reading* ($LRCS_1 = 364.77, df_1 = 3, p_1 < .001, LRCS_2 = 364.77, df_2 = 3, p_1 < .001,$) because overall the readings were distributed inhomogeneously (see Figure 11d). The quantifier had an influence on the distribution of readings as revealed by a reliable interaction of *Reading* and *Quantifier* ($LRCS_1 = 49.32, p_1 < .01, LRCS_2 = 32.16, p_1 < .01,$). Looking at the data it seemed unlikely that the quantifier affected the amount of local readings. We suspected that this interaction was due to the higher number of errors and lower number of literal

	ES		
AS	literalist	localist	incons.
literalist	19	2	6
localist	0	7	0
incons.	2	1	1

Table 4: Contingency table showing the number of occurrences of *response types* in the AS- and ES-conditions with neutral intonation. *Literalists* are subjects who gave responses that were classified as either a literal response or an error; *localist* are subjects who gave responses that were classified as either a local response or an error. Subjects who gave at least two responses from different categories (literal, global, local) are classified as *inconsistent*. (No subject's responses where errors exclusively and there where no *globalists*, as defined in parallel fashion.)

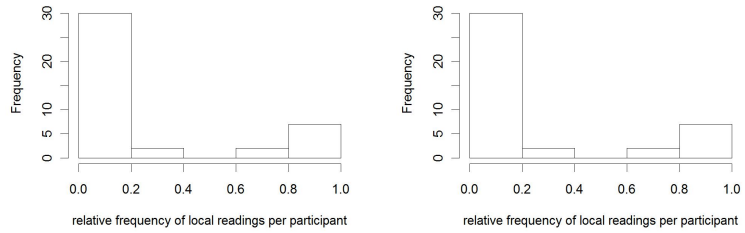
readings in the ES-conditions as compared with the AS-conditions. A difference in the distribution of judgment types between participants was revealed by a reliable interaction of *Reading* and *Participant* ($LRCS_1 = 487.70, p_1 < .01$). We were interested in whether this effect was due to some of the participants being more likely to choose literal readings than others. Finally, there was a three-way interaction of *Participants*, *Construction* and *Reading* ($LRCS_1, = 143.92, df_1 = 117, p_1 < .05$) which could stem from to the fact that some participants were more prone than others to make errors in the ES-conditions.

In order to find out whether the interactions just reported affected the amount of local readings, the log-linear models were computed again with a different coding of the readings. Here, we only considered the amount of local readings versus all other kinds of judgments. Judgments were coded as *local* and *other*. If the reported interactions affect the amount of local readings they should show up again. The expected but irrelevant effect of *Reading* was again significant ($LRCS_1 = 134.55, df_1 = 1, p_1 < .001, LRCS_2 = 144.64, df_2 = 1, p_1 < .001,$). In addition, the interaction of *Participant* and *Reading* ($LRCS_1 = 487.70, p_1 < .01$) as well as the three-way interaction of *Construction*, *Participant* and *Reading* ($LRCS_1 = 487.70, p_1 < .01$) were significant. No other effects were significant. In particular, the interaction of *Construction* and *Reading* was not significant ($LRCS_1 = 2.10, p_1 = .15, LRCS_2 = .74, p_2 = .39$).

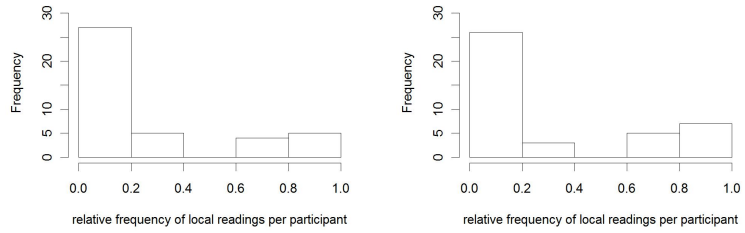
The interaction of *Participant* and *Reading* shows that there are indeed varying preferences for local readings among German speakers (see Table 4). Further examination of the distribution of local readings revealed a clear pattern (see Figure **??**). In all four conditions, about seven out of the 40 participants (16%) were consistently giving judgments that indicate local readings.[w] Further, the relative frequency of local judgments per participant strongly correlated between all four conditions (all $r > .6, p < .001$). Apart from the localists about half of the participants (and 41.5% of all participants) were consistently giving judgments that indicate literal readings. Only a minority of participants exhibited inconsistency. Reading preferences were more clear-cut in the AS-conditions than in the ES-conditions. The number of consistent literalists was lower in the ES-conditions (25.6%) as compared to the AS-conditions

w. please check numbers; this cannot be true!? not accross all four conditions

(57.3%). Also, the number of participants that showed inconsistency was higher in these conditions. However, the number of consistent localists did hardly differ between constructions.



(a) AS-accented (b) AS-neutral

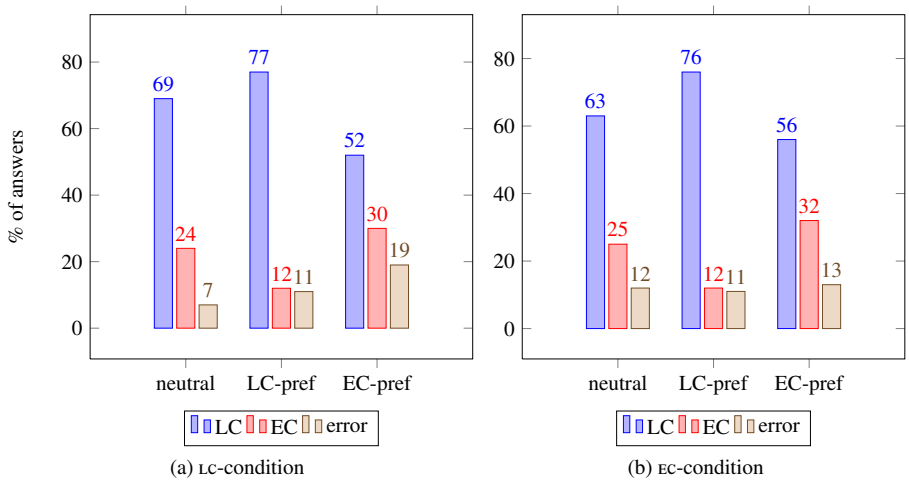


(c) ES-accented (d) ES-neutral

Figure 11: Judgments for the four target conditions

The three way-interaction of *Participant*, *Construction* and *Reading* is due to the fact that for some participants preferences for local over other readings deviated between the two construction types. These deviations were, however, not systematic to any degree. How much local preferences deviated between the two construction types per participant is depicted in Figure **??**. Most participants had exactly the same preferences in the two constructions. However, a few had a stronger and a few others a weaker preference for local readings in the ES- than in the AS-conditions. Since we did not find any systematic patterns here, we speculate that the three-way interaction is due to the ES-construction beeing understood non-standardly by a few participants. This speculation is plausible given the higher number of errors in the ES- as compared to the AS-conditions.

The absence of the interaction between *Construction* and *Participant* indicates that the type of construction does not affect the amount of local readings. We assumed based on the observed distribution of judgments that the type of contsruction affected the amount of literal but not local and global readings. To further test this assumption, we also considered literal and global readings versus other judgments in separate



(a) LC-condition (b) EC-condition

Figure 12: Response type frequences for preference-related controls

analyses. Comparing global readings versus other judgments *Construction* and *Reading* were found to interact ($LRCS_1 = 42.80, df = 1, p < .001$; $LRCS_2 = 22.14, df = 1, p < .001$). Comparing global readings to other judgments no such interaction was obeserved ($LRCS_1 = 1.24, df = 1, p = .29$, $LRCS_1 = .21, df = 1, p = .65$).

Preference-related controls

- see Figure 12

6.6 Discussion

High scores of success in our general control condition clearly indicate that participants understood the incremental verification task and were in general able to give judgements at the earliest possible point in a picture sequence. Results obtained for the preference-related controls moreover 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, *irrespective* of the order in which each reading could first be judged in a given sequence. Given this, the best and most straightforward explanation of the observed data is the hypothesis that the actual preferences among readings for both AS- and ES-sentences alike are:

(29) LIT > LOC (> GLB)

where it is unclear whether the global readings are entirely unavailable or simply severely dispreferred. Crucially, our data does attest local readings, albeit that they are not preferred across the board but preferred, as it seems, only by a small, but substantial number of consistent localists.

This is bad news for both traditionalism and grammaticalism, as no version of either would predict the pattern in (29). The main problem is that traditionalism, if left on its own, fails to predict local readings, while grammaticalism, when it relies on the strongest meaning hypothesis, fails to predict the preference order altogether. What are the problems exactly? Could either position be aided by innocuous auxiliary assumptions to be compatible with (29)?

It is not a major obstacle for traditionalism to predict that global readings are available. The weak variety of traditionalism holds that the extra assumptions needed to derive global readings are not standardly available, and so traditionalism is compatible with a lack of observed global readings. Alternatively, traditionalists could argue that the task's rather high processing demands lead to a suppression of global implicature 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. Unfortunately, to the extent that global readings are explained away as dispreferred or unavailable, traditionalism also looses the possibility of accounting for local readings of AS-sentences in its unrestricted variety.[11] This is because unrestricted traditionalism explains local and global readings of AS-sentences in essentially the same manner. On top of that, traditionalism cannot explain local readings for ES-sentences at all. In sum, traditionalism alone cannot explain (29), but needs an additional explanation of local readings for *both* AS- *and* ES-sentences.

Grammaticalism, on the other hand, fails to predict the preference relation in (29), because these strictly contradict the strongest meaning hypothesis in either of the two varieties suggested by Chierchia et al. (forthcoming). This shows most clearly in the fact that both versions of the strongest meaning hypothesis predict that local readings of AS-sentences are preferred over literal ones. The problem, therefore, seems to be the selection of readings by the strongest meaning hypothesis. In fact, if (29) is the actual preference over readings, then the strongest meaning hypothesis is also conceptually on the wrong track, because no account that merely looks at the logical strengths of readings can plausibly predict the preference pattern in (29) for both AS- and ES-sentences at the same time. (Try, if you please!) Turning away from comparisons of logical strength, a selection criterion that produces (29) is ready at hand: if we assume that each insertion of Exh(·) operator is *costly* anti-proportional to the depth of its embedding, then we derive a preference of literal readings over local ones over global ones. But whether such an economy principle is sound and in line with further empirical data is an interesting matter of further investigation. What it does not explain in any case, is why some subjects in our study gave local responses *consistently*, while others did not. In sum, in as far as grammaticalism depends on the strongest meaning hypothesis, its predictive success is equally thin as that of traditionalism.

On the face of it, there is a very clear winner among the stylized candidate theories that we introduced in Section 3. Assuming that global readings are not available at all, the preference pattern in (29) is exactly what *literalism* would predict. Remember that literalism was the conventionalist position which posited a lexical ambiguity of scalar

[11] Remember that what we called unrestricted traditionalism can account for local readings of AS-sentences by assuming that a sentence of the form *Some of the X are related to some of the Y* is a feasible alternative.

some with a preference for the semantically weaker meaning *some and maybe all* over the upper-bounded reading *some but not all*.

6.6.1 Bits and Pieces

- The distribution of responses found in previous studies could in part be explained by differences in typicality of the pictures with respect to the possible readings (see van Tiel). In the present case this kind of reasoning would have to be extended to the typicalilty of parts of pictures. On intuitive grounds, we take this kind of reasoning to be implausible. We will now shortly discuss how graphical properties might still affect judgments in the worst case. In the first step of the AS-condition local and global readings cannot be judged, yet. Therefore, graphical properties should not affect judgments if participants have one of these readings in mind. If participants have a literal reading in mind, the worst case scenario is that they refrain from judging the sentence as true at this point due to some, eg. atypical, graphical property of the picture. The situation is similar at the second step. The local reading is still not decidable. Therefore, graphical properties should not play a role if participants have a local reading in mind. A positive judgment might, however, be a delayed judgment of a literal reading. Again, the worst case scenario for our purpose is that participants refrain from answering because of graphical properties. At the last step the local reading requires a negative judgment. Now, we could obtain delayed negative responses from literal or global readings because the graph as a whole is, e.g., logically true but atypical with respect to these readings. These kinds of judgments would be indistinguishable from local responses. Note, however, that we have no reason to believe that typicality has large effects, because, given the results from van Tiel, the picture as a whole is expected to be one of the more typical situations for AS-sentences. Still, we cannot exclude the possibility that graphical properties of the pictures might have the described effect. In the ES-conditions this type of effect is, however, entirely impossible since the literal and global reading is plainly false given the initial parts as well as the whole graph. Summing up, in the worst case, which we consider implausible, graphical effects might lead to an enhanced number of local responses in the AS- as compared to the ES-condition. We have to keep this possibility in mind. We want to stress, however, that the design presented here improves on previous studies with respect to possible graphical effects.

7 Conclusions

A The auditory sentence material

Targets. Table 5 shows the durational values for each of the single regions in the sentence. Crucially, durational values of accented determiners were significantly increased as opposed to their non-accented counterparts. Overall, QUANTIFIER effects were observed, which might be attributed to the above-mentioned lexical differences,

as well as to the fact that ES-structures generally contained fewer words, thus leading to a tendency of a durational decrease for each of the single words.

Differences between maximal and minimal F0 values for each of the single words in the sentence are depicted in Table 6. As is descriptively evident, accented determiners showed a larger F0 range compared to unaccented versions, an effect which is confirmed by statistical analyses. An additional QUANTIFIER *PROSODY interaction indicates that these differences are even more pronounced in the AS-condition. Finally, comparable to the durational analyses, QUANTIFIER effects were observed when conditions exhibited lexical differences.

Finally, differences between maximal and minimal F0 values for each of the single words in the sentence are given in Table 8. Again, the most reliable differences occur at the boundary regions.[x]

x. provide caption for table D

Preference-related fillers. Table 7 shows the durational values of each of the single words in the sentence. As is descriptively evident, the largest durational differences were realized at the boundary regions (i.e. Region 5 and Region 7). Interestingly, small differences between conditions also yielded significance at other positions, suggesting that our speaker produced the different conditions very consistently (i.e., with little variance).

References

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.

Breheny, Richard and Napoleon Katsos (2008). *Experimental Investigations of the Semantics/Pragmatics Interface*. Course Material for ESSLLI.

Breheny, Richard et al. (2006). "Are Generalised Scalar Implicatures Generated by Default? An On-Line Investigation into the Role of Context in Generating Pragmatic Inferences". In: *Cognition* 100.3, pp. 434–463.

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

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

Chierchia, Gennaro (2004). "Scalar Implicatures, Polarity Phenomena and the Syntax/Pragmatics Interface". In: *Structures and Beyond*. Ed. by Adriana Belletti. Oxford University Press, pp. 39–103.

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

— (forthcoming). *FC Nominals and FC Disjunction: The Identity Thesis*. Ed. by Anamaria Fălăuş. Palgrave Macmillan.

Chierchia, Gennaro et al. (forthcoming). "The Grammatical View of Scalar Implicatures and the Relationship between Semantics and Pragmatics". In: *Semantics. An International Handbook of Natural Language Meaning*. Ed. by Claudia Maienborn et al. de Gruyter.

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

	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 *einigen*.

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

	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: ???

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

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.

Dalrymple, Mary et al. (1998). "Reciprocal Expressions and the Concept of Reciprocity". In: *Linguistics and Philosophy* 21.2, pp. 159–210.

Fodor, Janet (1998). "Learning to Parse". In: *Journal of Psycholinguistic Research* 27.2, pp. 285–319.

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.

Gazdar, Gerald (1979). *Pragmatics: Implicature, Presupposition, and Logical Form*. New York: Academic Press.

Geurts, Bart (2009). "Scalar Implicature and Local Pragmatics". In: *Mind and Language* 24.1, pp. 51–79.

— (2010). *Quantity Implicatures*. Cambridge University Press.

Geurts, Bart and Nausicaa Pouscoulous (2009). "Embedded Implicatures?!?" In: *Semantics & Pragmatics* 2.4, pp. 1–34.

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

Grodner, Daniel J. et al. (2010). ""Some," and Possibly All, Scalar Inferences are not Delayed: Evidence for Immediate Pragmatic Enrichment". In: *Cognition* 166, pp. 42–55.

Groenendijk, Jeroen and Martin Stokhof (1984). "Studies in the Semantics of Questions and the Pragmatics of Answers". PhD thesis. Universiteit van Amsterdam.

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.

Levinson, Stephen C. (2000). *Presumptive Meanings. The Theory of Generalized Conversational Implicature*. Cambridge, Massachusetts: MIT Press.

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

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.

Raffray, C.N. and MJ Pickering (2010). "How do people construct logical form during language comprehension?" In: *Psychol Science* 21.8, pp. 1090–1097.

van Rooij, Robert and Katrin Schulz (2004). "Exhaustive Interpretation of Complex Sentences". In: *Journal of Logic, Language and Information* 13, pp. 491–519.

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.

Schulz, Katrin (2005). "A Pragmatic Solution for the Paradox of Free Choice Permission". In: *Synthese* 147, pp. 343–377.

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.

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.

von Stechow, Arnim and Thomas Ede Zimmermann (1984). "Term Answers and Contextual Change". In: *Linguistics* 22.1, pp. 3–40.

Sternberg, S. (1966). "High-speed scanning in human memory". In: *Science* 153, pp. 652–654.

van Tiel, Bob (2012). *Embedded Scalars and Typicality*. Talk held at XPrag 2011, Barcelona.