The role of scalar diversity and question under discussion in deriving implicatures with embedded scales[1]

Adina Camelia BLEOTU — *University of Bucharest*
Anton BENZ — *Leibniz-Centre General Linguistics (ZAS), Berlin*

Abstract. We investigate experimentally the role of scalar diversity and question under discussion for implicature rates of sentences with multiple scalar terms such as *Some meals are adequate* with embedding scale ⟨*all, some*⟩ and embedded scale ⟨*good, adequate*⟩. These sentences can trigger different types of scalar implicatures. We modified the inference task by van Tiel et al. (2016) and tested the 43 scales studied by them in a position embedded under *some* and *possible*. We were particularly interested in whether implicatures involving embedded scales can be boosted if made relevant by *Questions under Discussion* (QUDs). Our results showed that all tested types of implicatures are sensitive to QUDs. Most interestingly, the contrast between *bounded* and *unbounded* scales, which was a strong predictor in previous studies, no longer correlates with rates of implicatures once a QUD is added. We argue that our findings support a version of the Alternatives-based Account (the Contextual Alternatives and Scalar Distinctness Account) where contextual availability of alternatives is more important than lexical availability, and where, additionally, the (lexical/contextual) distinctness of the scales plays a role.

Keywords: experimental pragmatics, scalar diversity, embedded implicatures, questions under discussion.

1. Introduction

1.1. On scales and scalar implicatures

Scalar implicatures represent inferences that we draw in conversation when conversational maxims have not been observed (Grice, 1989). According to Grice (1989), in a context where a speaker knows that all of the roses in the garden are red, producing the sentence in (1a) instead of the sentence in (1b) is pragmatically underinformative, as is producing (2a) instead of (2b) in a context where the speaker knows it is hot outside: the speaker has failed to abide by the Maxim of Quantity, flouting the submaxim 'Make your contribution as informative as required'.

(1) a. Some roses are red.
b. All roses are red.

(2) a. It is warm outside today.
b. It is hot outside today.

[1] This research was supported by a DFG project *SIGames2: Experimental Game Theory and Scalar Implicatures: Investigating variation in context and scale type* led by Dr. Anton Benz (Grant Nr. BE 4348/4-2). In 2022, Adina Camelia Bleotu was a postdoc in the project. We thank the audience at *Sinn und Bedeutung* for feedback on this work, as well as Shirley Orr for her useful comments and suggestions. Additionally, we would like to thank the semantics and pragmatics researchers at ZAS Berlin for interesting discussions on preliminary versions of this work.

The notion of **scale** is particular important to understand the failure in informativeness. According to Horn (1972), a scale represents a range of items ordered in terms of informational strength. Languages showcase an impressive number of scales: the quantifier scale ⟨*all, some*⟩, the numeral scale ⟨*..., two, one*⟩, modal scales such as ⟨*necessarily, possibly*⟩, ⟨*must, may*⟩, connectives such as ⟨*and, or*⟩, adverbs such as ⟨*always, often, sometimes*⟩, degree adjectives such as ⟨*hot, warm*⟩, or degree verbs such as ⟨*know, believe*⟩ or ⟨*love, like*⟩. Scales involve at least two terms: a strong scalar term like *all* and a weak scalar term like *some*, such that the utterance employing the strong scalar term, i.e. S(all) or S(hot) entails the utterance employing the weak scalar term, i.e. S(some) or S(warm), but not the other way round. While both the strong scalar term and the weak scalar term express the same property, for instance, warmness, they express it to a different degree (Kennedy and McNally, 2005). Importantly, there must be some distance between the lower bounds of the two scalar terms, otherwise the two terms could be considered synonyms (see recently Orr et al., 2024). When a speaker produces (1a) instead of (1b) or (2a) instead of (2b) in a situation optimally described by (1b) or (2b) , they are failing to make their communicated utterance adequately informative because they are employing the weak scalar term instead of the strong scalar one.

1.2. Do implicature rates vary with scale type?

1.2.1. Implicatures with one scale

A question that has been the focus of many studies has been whether the rate of implicatures varies with the type of scale and in what way. While the most investigated scale has been the ⟨*all, some*⟩ scale starting with Noveck (2001); Pouscoulous et al. (2007); Foppolo et al. (2012); Bleotu (2021), other scales such as the modal scale, the numerical scale, disjunction or ad-hoc implicatures have also been the object of linguistic scrutiny (Noveck, 2001; Papafragou and Tantalou, 2004a; Huang and Snedeker, 2009; Bleotu et al., 2021a, 2022b, 2023; Tieu et al., 2017) It has thus been shown that the rate at which weak scalar items give rise to scalar implicatures is not uniform across scale types (van Tiel et al., 2016; Kuppevelt, 1996; Zondervan et al., 2008; Degen, 2013; Degen and Tanenhaus, 2015; Cummins and Rohde, 2015; Yang et al., 2018; Ronai and Xiang, 2020).

In an influential study, van Tiel et al. (2016) investigated 43 different scales with an inferencing task. For instance, for the scale ⟨*good, adequate*⟩, participants had to read an utterance and give a ‘Yes’ or ‘No’ to the question in (3):

(3) John says: *The food is adequate.*
Would you infer from this that, according to John, the food is adequate?

If they answered ‘Yes’, it was inferred that the participant strengthened *adequate* to *adequate but not good*.

1.2.2. Implicatures with multiple scales. Local implicatures.

In two experiments, we tested van Tiel et al. (2016)'s 43 scales when embedded under *some* and *It is possible that*, probing into the rates of various types of implicatures (see Table 1), including embedded/local implicatures, i.e. implicatures with the scales embedded under other scales, such as those in (4).

(4) Mary says: *Some meals are adequate.*
Would you infer from this that, according to Mary, some meals are adequate but not good?

There has been a long debate whether local implicatures can occur when scales are embedded under other scalar items (Geurts and Pouscoulous, 2009; Clifton and Dube, 2010; Chemla and Spector, 2011; Bill et al., 2021; Bleotu et al., 2022b). Solving this debate has been regarded in the literature as a way to better understand how implicatures are derived. Assuming local implicatures share the same derivation mechanism with implicatures derived with one single scale, the grammatical account (Chierchia, 2004; Chierchia et al., 2012) predicts the existence of local implicatures via exhaustification, a mechanism by which a weak scalar term is strengthened to the negation of its stronger alternative scalar term. In contrast to the grammatical account, the pragmatic-Gricean account (Grice, 1989; Horn, 1972) predicts that participants should derive no local implicatures in principle, given that Gricean reasoning applies to whole utterances not parts of utterances.[2] Experimental evidence was thus crucial in settling the debate straight. Geurts (2009) argued on the basis of various experimental methods (inference task, verification tasks) that local implicatures are very rare in both upward entailing and downward entailing contexts, and consequently, they argued in favour of pragmatic account for implicature derivation. Subsequently, using a picture selection task, Clifton and Dube (2010) showed that participants would often pick both pictures corresponding to local implicatures and global implicatures, thus arguing that local implicatures are in fact possible. Additionally, by means of a rating task, Chemla and Spector (2011) showed that adults do derive local implicatures for a sentence such as (5):

(5) Every letter is connected to some of its circles.

However, their results were criticized by van Tiel (2014) who argued that typicality plays an important part in picture-selection. Nevertheless, local implicatures have been shown to occur at ceiling if supported by a pragmatic task. In an interactive game–theoretic reward task set-up which satisfies Grice's conversational requirements for implicature generation (a specific purpose of the conversational exchange), Gotzner et al. (2018) showed that adults can draw local implicatures to a very high degree. Recent research by Bill et al. (2021) found that, when deriving implicatures, English adults preferred global implicatures over local implicatures, while children preferred local implicatures. Moreover, a recent study by Bleotu et al. (2022b) employing a Shadow Play Paradigm, building on Bleotu et al. (2021b, c) found that, when deriving implicatures, both Romanian children and adults preferred global implicatures and derived almost no local implicatures. These findings keep the debate about local implicatures alive. As in

[2] Nevertheless, if one assumes that local implicatures are derived via a different mechanism than global implicatures, such as in virtue of a special stress pattern (Geurts and van Tiel, 2013) or in special pragmatic contexts (Geurts and Pouscoulous, 2009), then local implicatures could be expected.

the case of un–embedded weak scalar terms, this research has almost exclusively concentrated on the ⟨*all, some*⟩ and ⟨*and, or*⟩ scales (see also van Tiel, 2014; Crnič et al., 2015; Benz and Gotzner, 2017; Gotzner et al., 2018; Franke et al., 2017; Bill et al., 2021). Exceptions are Geurts and Pouscoulous (2009) who also tested for implicatures of *some* embedded under *think*, *want*, and *has to*, and Bleotu et al. (2022b) who studied *some* embedded under ⟨*certain, possible*⟩. This research showed that the rates with which local implicatures occur depend on the type of verb or operator under which *some* is embedded. To the best of our knowledge, no study has looked at different types of scales in the embedded position.

In our current experiments, we investigate different types of implicatures involving the embedded scale (see Table 1), once in a setting where they are not supported by a question introducing the Question under Discussion, i.e. the QUD (Experiment 1), and once in one with QUD support (Experiment 2). The rationale was that if local implicatures are not present in pragmatically unsupported contexts, then we should not see an effect of scalar diversity in Experiment 1. Moreover, if implicatures depend on the activation of alternatives, then activating the alternatives by a QUD should increase the rates of local implicatures in Experiment 2. We were also interested to what extent different scales are sensitive to QUDs, and if these can be predicted by grammatical features, in a similar fashion to van Tiel et al. (2016).

2. Research questions

2.1. Implicature rates and scalar diversity

A first question (Q1) we ask is whether implicature rates vary with implicature type. Given that previous studies show that participants generally tend to derive fewer local implicatures than global implicatures, we would expect to see a similar overall pattern in Experiment 1 and, possibly, in Experiment 2.

2.2. Predictors of implicature rates

A second question (Q2) is what predicts rates of implicature for different scales, i.e. scalar diversity. While we are nevertheless aware that other studies have considered factors such as homogeneity, local enrichment in Sun et al. (2018) or question availability in Ronai and Xiang (2020), we here considered the factors discussed by van Tiel et al. (2016): the availability of the lexical scales and the distinctness of scale-mates. The availability of lexical scales was evaluated by van Tiel et al. (2016) through association strength, grammatical class, frequency and semantic relatedness. We briefly define each of these subfactors. Association strength represents the strength of association between the scalar expression used in the speaker's utterance. van Tiel et al. (2016) hypothesized that the greater the association strength between the weak and the strong scalar terms, the more available the scale should be. Association strength was measured by van Tiel et al. (2016) through a cloze task, either in a neutral version containing pronouns (*he/she*), or in a non-neutral version containing nouns (e.g., *this student*). The neutral version of the cloze is exemplified in (6):

(6) In the following you will see 43 sentences. In every sentence, one word will be highlighted, like this:

She is angry.

Which words could have occurred instead of the highlighted one? Some of the alternatives that may come to mind are *beautiful, happy, married*, and so on. We ask you to tell us the first three alternative words that occur to you when you read these sentences.

Association strength was calculated by van Tiel et al. (2016) based on whether participants mentioned a stronger scale in their answers (in the lenient coding).

Grammatical class refers to whether the scale under consideration belongs to an open class or a closed class. For instance, the closed class can be exemplified by quantifiers and modals. van Tiel et al. (2016) hypothesized that, given that the search space of alternatives is much smaller for closed grammatical classes than for open ones, scales belonging to closed classes should be more available.

van Tiel et al. (2016) also considered the frequency of the strong scalar term compared to the weaker one. van Tiel et al. (2016)'s hypothesis was the the more frequent the strong term relative to the weaker one, the more available the scale consisting of both members. After extracting the frequencies of the scalar expressions in the materials from the Corpus of Contemporary American English (Davies 2008), van Tiel et al. (2016) calculated the relative frequency by dividing the frequency of the stronger scalar term by the frequency of the weaker one, and logarithmising the outcome.

van Tiel et al. (2016) also looked at semantic relatedness, i.e. the relatedness of the scale-mates, measured by how often a strong scalar term and a weak scalar term occur in similar linguistic environments. The expectation was that, if the two scale-mates are more likely to co-occur with the same words, the scale would be more available. To measure semantic relatedness, they used Latent Semantic Analysis (Landauer and Dumais, 1997; Landauer et al., 1998), which constructs a matrix with words from a corpus as rows and columns and computes a value in the interval [0, 1] that denotes the extent to which the words at issue occur with the same words.

Importantly, van Tiel et al. (2016) found that no measure of lexical availability showed any correlation with rates of implicatures in their experiment. Consequently, we expect that they should also not correlate with implicature rates of embedded scales in complex sentences.

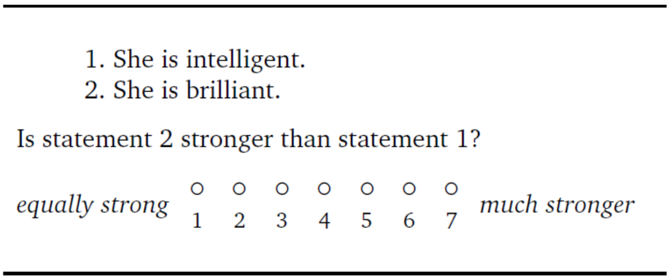
We also investigated the role of the distinctness of the scale-mates, evaluated through semantic distance and boundedness. For both factors, we adopted the same measurements/decisions used by (van Tiel et al., 2016: see also Zevakhina 2012).[3] Semantic distance, the distance

[3] While in our current paper, we have adopted van Tiel et al. (2016)'s measurements/judgments, it is worth mentioning that more recent studies such as Orr et al. (2024) have tried to improve the manner in which semantic distance and boundedness are measured. With respect to semantic distance, for instance, Orr et al. (2024) replaced the question *Is statement 2 stronger than Statement 1* with (i):

(i) Is statement 2 interchangeable with statement 1?

With respect to boundedness, as an alternative to an intuitive definition, Orr et al. (2024) proposes the use of the comparative as a test for boundedness, as in (ii).

(ii) John says: *The assistant is brilliant.* In principle, is it possible for someone, for example, an assistant, to between the bounds of the weak and the strong scalar term, was measured by van Tiel et al. (2016) through ratings of statements containing strong/weak scale-mates such as exemplified in Figure 1.



1. She is intelligent.
2. She is brilliant.

Is statement 2 stronger than statement 1?

equally strong ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 *much stronger*

Figure 1: Example of experimental item from the Semantic Distance Task in van Tiel et al. (2016)

van Tiel et al. (2016) hypothesized that more semantically distinct scale-mates would correlate with higher implicature rates, and his findings supported this hypothesis. We also entertain a similar hypothesis for the different types of implicatures in our experiment.

Regarding boundedness, bounded scales represent scales where the stronger scalar term denotes an endpoint (e.g. *free* in ⟨*cheap, free*⟩), in contrast to unbounded scales like ⟨*content, happy*⟩, which refer to intervals. van Tiel et al. (2016) establishes whether a scale is bounded or unbounded on an intuitive basis. van Tiel et al. (2016) hypothesized that bounded scales would give rise to more implicatures, and, indeed, this was found to be the case both in their experiment, as well as in Sun et al. (2018). Based on van Tiel et al. (2016)'s findings, we also expect to find a correlation between boundedness and higher implicature rates of various types.

While our general expectation is that van Tiel et al. (2016)'s findings should carry over to multiple types of implicatures, it might be that this will be more apparent for global implicatures rather than local ones, if participants struggle with the mechanism of deriving local implicatures.

2.3. Question Under Discussion and (local) Implicature Rates

A third question we address is whether explicit questions introducing the Question Under Discussion lead to a boost in (local) implicature rates. Previous research has shown that the Question Under Discussion does lead to an increase in implicatures in utterances containing a single weak scalar item in both adult and child language (Degen, 2013; Zondervan et al., 2008; Yang et al., 2018; Ronai and Xiang, 2020; Papafragou and Tantalou, 2004b; Skordos and Papafragou, 2016). This has been demonstrated for both explicit Questions Under Discussion (Zondervan et al., 2008; Yang et al., 2018; Ronai and Xiang, 2021, 2020) and implicit ones accommodated via a story (Degen, 2013; Guasti et al., 2005) or through various cues (Skordos and Papafragou, 2016). Importantly, the QUD makes the stronger alternative contextually relevant, and it often makes use of the stronger scale-mate. This can be explained within an Alternatives-Based

be even more brilliant?

Account of implicatures (Barner et al., 2011; Tieu et al., 2017), where implicatures depend on the activation of alternatives, and explicit access to the stronger alternatives makes implicature derivation easier.

(7) Sue: *Is the movie excellent?*
Mary: *It is good.*
Would you conclude from this that Mary thinks the movie is not excellent? Yes/No (Ronai and Xiang, 2021)

(8) *Are all shapes blue?*
Some shapes are blue. (Ronai and Xiang, 2020)

Importantly, access to the stronger alternatives increases not only adults' but also children's ability to derive implicatures (Guasti et al., 2005; Foppolo et al., 2012; Skordos and Papafragou, 2016). However, what seems to matter even more than the presence of the stronger alternative is the contextual relevance contributed by the Question Under Discussion: children are able to derive implicatures to a high degree in a context approximating naturalistic conversation (Papafragou and Tantalou, 2004b) or in situations where the stronger alternative becomes relevant (Skordos and Papafragou, 2016):

(9) Experimenter: Did you color the stars? Elephant: I colored some.

While most previous research focused on implicatures with utterances containing a single weak scalar item, recent studies have also started looking at the effect of QUD on implicatures in utterances containing two scalar terms. (Gotzner et al., 2018) showed that, in an interactive game-theoretic reward task set-up satisfying Grice's conversational requirements for implicature generation (i.e., a talk exchange with a specific purpose/direction), adults showed high rates of local implicatures. However, recent findings from (Bleotu et al., 2022a) seem to suggest that the QUD may sometimes increase global implicature rates only to a limited extent. (Bleotu et al., 2022a) probed into the role of a scalar question introducing a QUD upon Romanian adults' and children's interpretation of utterances such as those in (10) embedding a scalar term belonging to the scale ⟨*all, some*⟩ under a scalar term belonging to the scale ⟨*certain, possible*⟩.

(10) Poate că unii câini sunt albaștri.
maybe that some dogs are blue
'It is possible that some dogs are blue.'

In one experiment, Experiment 1, the question involved the ⟨*certain, possible*⟩ scale, and, in another experiment, Experiment 2, the question involved the ⟨*all, some*⟩ scale (see (11)).

(11) a. ⟨*certain, possible*⟩ QUD
The wizard asks: *Is it possible or certain that there are blue dogs in the spotlight?*
b. ⟨*all, some*⟩ QUD
The wizard asks: *Are some or all of the dogs in the spotlight blue?*

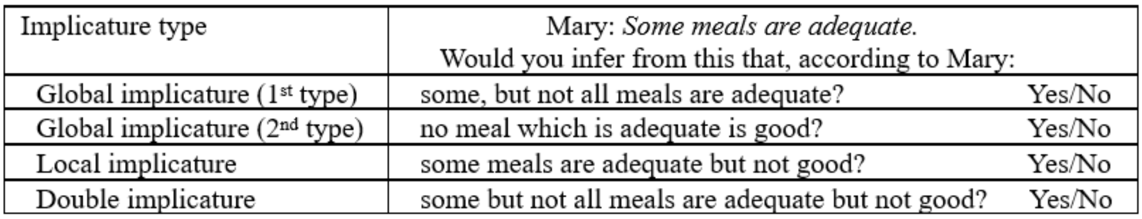
While the two experiments were expected to lead to increases in different implicature rates, both adults and children derived more global implicatures of the type *It is not certain that some dogs are blue* ($\mathrm{GI}_{NotCertainSome}$) in the ⟨*certain, possible*⟩ QUD experiment than in the ⟨*all, some*⟩ QUD one. Nevertheless, there was a QUD effect upon implicature rates.

Thus, there is reason to expect that an explicit QUD might lead to an increase in implicature rates.

3. Experiment 1

3.1. Aim

In Experiment 1, we extend (van Tiel et al., 2016)'s investigation to multiple scalar sentences, targeting a richer array of implicatures: global implicatures, local implicatures, and double implicatures, i.e., implicatures strengthening both scales (Table 1).



Implicature type	Mary: *Some meals are adequate.* Would you infer from this that, according to Mary:
Global implicature (1st type)	some, but not all meals are adequate? Yes/No
Global implicature (2nd type)	no meal which is adequate is good? Yes/No
Local implicature	some meals are adequate but not good? Yes/No
Double implicature	some but not all meals are adequate but not good? Yes/No

Figure 2: Example of an item in Experiment 1

We ask which implicature types participants derive more and inquire into the best predictors for rates of different implicatures (the availability of the lexical scales or the distinctness of the scale-mates).

3.2. Participants

We tested 60 American English native speakers recruited via Prolific.

3.3. Predictions

Based on the previous findings in the literature related to generally lower rates of local implicatures compared to global implicatures (Geurts and Pouscoulous, 2009; Clifton and Dube, 2010; Chemla and Spector, 2011; Bill et al., 2021; Bleotu et al., 2022b), we expect to find lower rates of local implicatures and double implicatures compared to global implicatures.

Based on the findings in (van Tiel et al., 2016), we expect the distinctness of the scale-mates to explain scalar diversity best.

3.4. Materials and Methodology

We employed a similar inference task to that in (van Tiel et al., 2016). We embedded the 43 scalar terms in (van Tiel et al., 2016) under *some* and *possible*. For each sentence, participants answered four randomized questions targeting four implicature types (see 2).

While our task was overall quite similar to (van Tiel et al., 2016), we made some important modifications to the presentation of weak and strong scalar items compared to (van Tiel et al., 2016). Their study employed instructions which used a negated scalar term, as in (12):

(12) Mary says: *This meal is adequate.*
Would you infer from this that, according to Mary, the meal is not good?

However, (Benz et al., 2018) have shown that an utterance containing a negated strong scalar item can sometimes give rise to negative strengthening interpretations of negated adjectives, such that *not good* is interpreted as 'totally bad' rather than as 'adequate'. In order to avoid such an interpretation, we constructed our statements by also mentioning the weak scalar term before the negated strong scalar term (see (13) and 2)[4].

(13) Mary says: *Some meals are adequate.*
Would you infer from this that, according to Mary, some meals are adequate but not good?

We combined the test items with seven attention checks containing antonyms (*clean-dirty*) and unrelated properties (*sleepy-rich*).

3.5. Results

We find that participants derive different types of implicatures at different rates: global implicatures involving the 1st scale at a rate of 94.47%, followed by local implicatures at a rate of 68.78%, followed by double implicatures at a rate of 67.59%, followed by global implicatures involving the 2nd scale at a rate of 28.87%. To exemplify, the rates for the different implicature types are represented graphically in Figures 3, 4, 5, and 6.

The scalar terms *some* and *possible* give rise to similar global implicatures with the 1st scalar item. Overall, we notice considerable variation in rates of implicature types for different lexical scales.

We expected factors involving the 2nd scale to be correlated with a higher rate of local implicatures, 2nd global implicatures, and double implicatures. We ran multiple correlation tests between each type of implicature and each predictor in van Tiel et al. (2016). Similarly to van Tiel et al. (2016)'s findings about implicatures with weak scalar terms in utterances involving one single scale, we found that local implicatures and double implicatures were impacted by the distinctness of the scale-mates of the 2nd scale, as can be seen in Figure 7.

[4] Similarly, in a recent study, (Orr et al., 2024) also addressed this potential difficulty, changing the materials in (van Tiel et al., 2016) by modifying the strong scalar term by means of *possibly*, as in (i):

(i) John says: *The assistant is intelligent.*
Would you infer from this that, possibly, according to John, he is possibly brilliant?

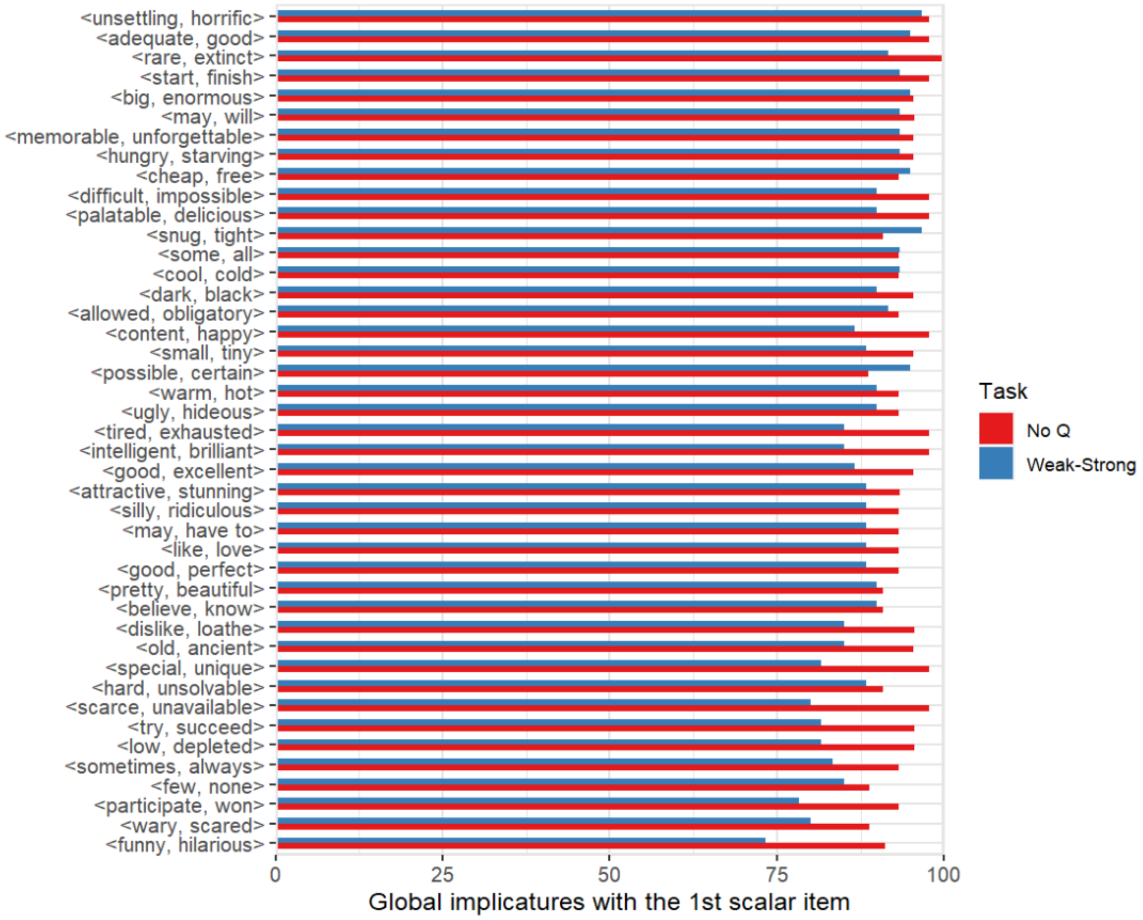


Figure 3: Global implicature rates involving the 1st scale

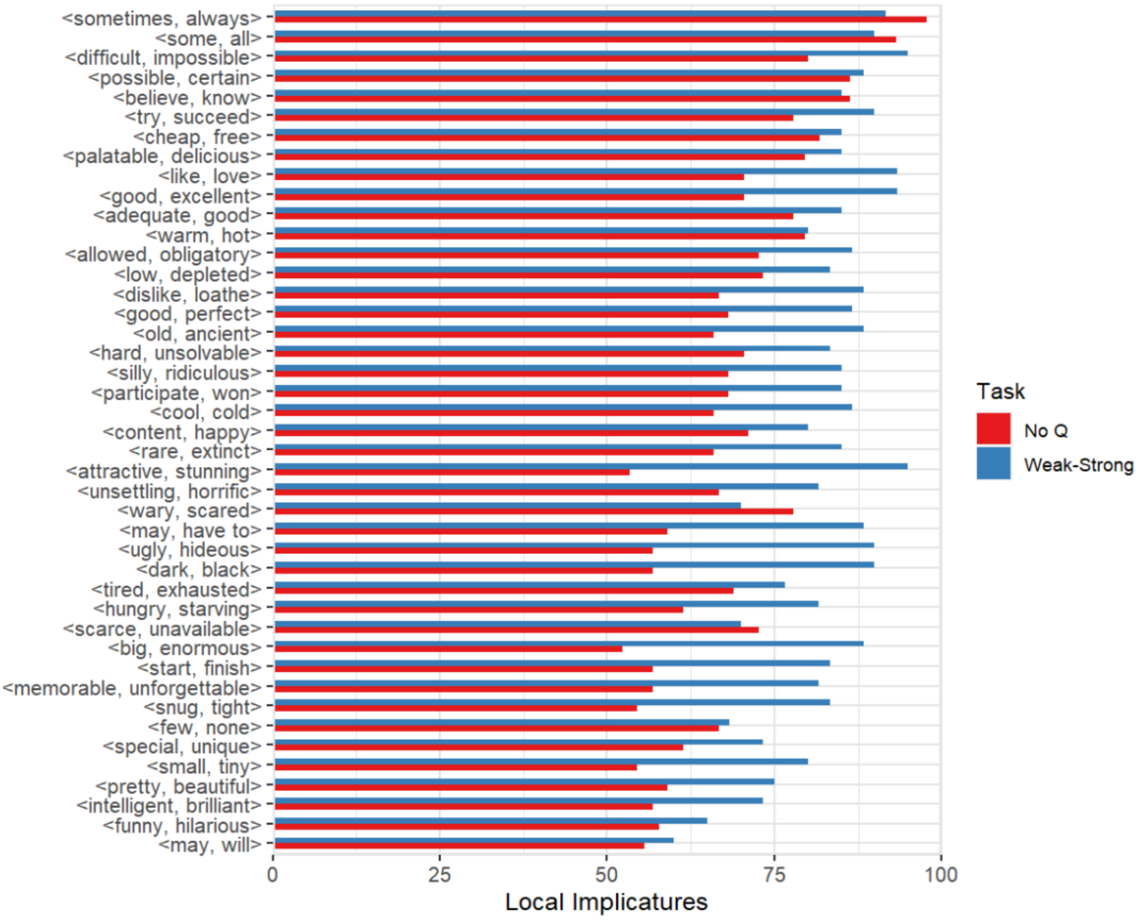


Figure 4: Local implicature rates in our experiments

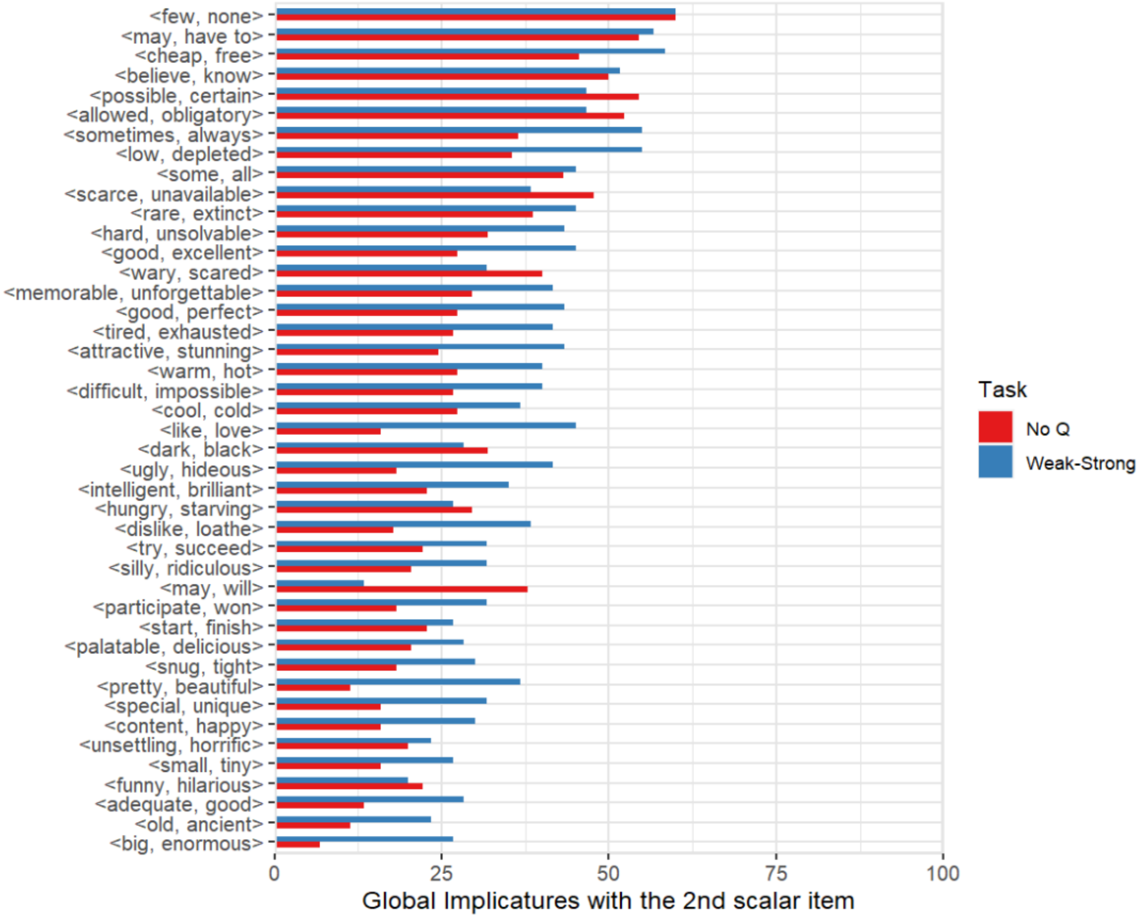


Figure 5: Global implicature rates involving the 2nd scale

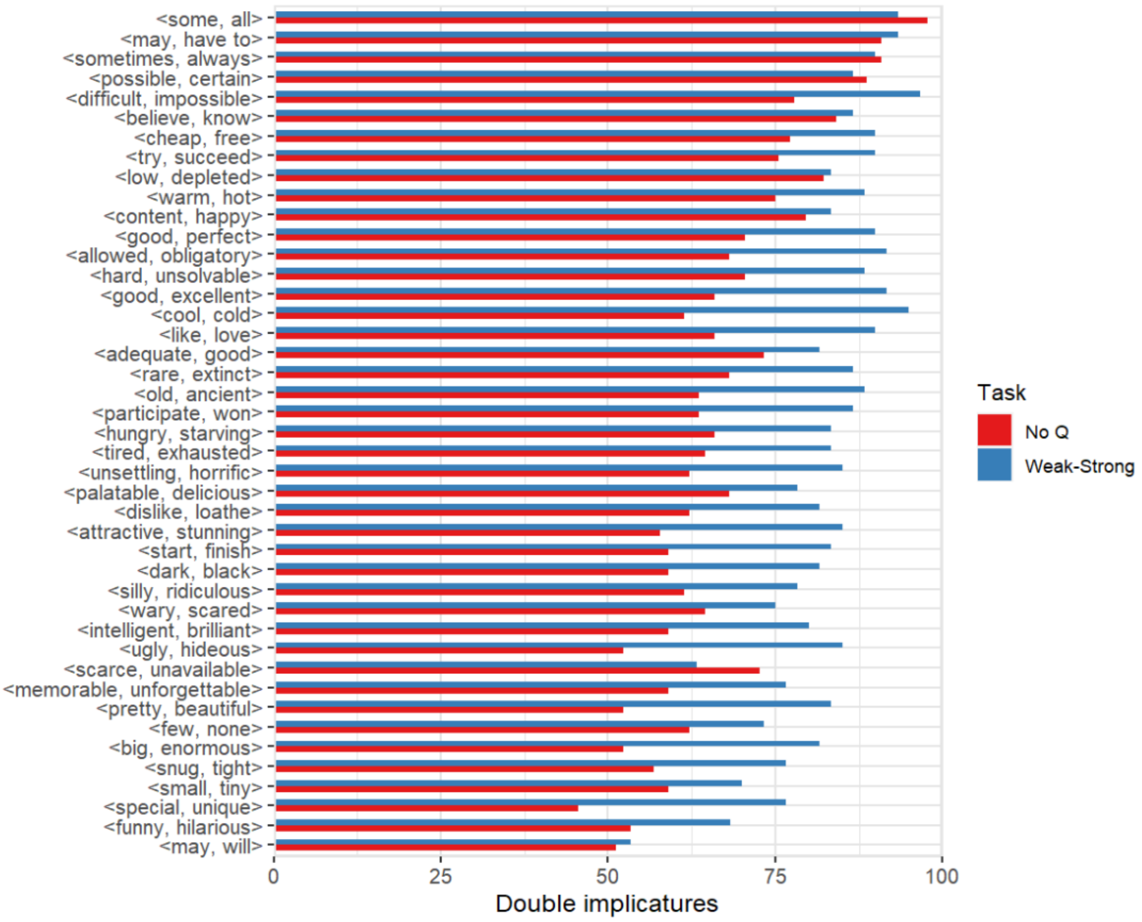
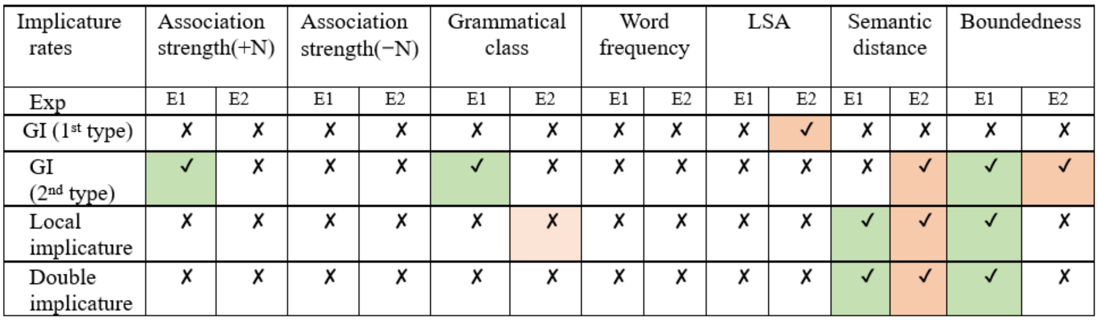


Figure 6: Double implicature rates



Implicature rates	Association strength(+N)		Association strength(−N)		Grammatical class		Word frequency		LSA		Semantic distance		Boundedness	
Exp	E1	E2	E1	E2	E1	E2	E1	E2	E1	E2	E1	E2	E1	E2
GI (1st type)	✗	✗	✗	✗	✗	✗	✗	✗	✗	🟧 ✓	✗	✗	✗	✗
GI (2nd type)	🟩 ✓	✗	✗	✗	🟩 ✓	✗	✗	✗	✗	✗	✗	🟧 ✓	🟩 ✓	🟧 ✓
Local implicature	✗	✗	✗	✗	✗	🟧 ✗	✗	✗	✗	✗	🟩 ✓	🟧 ✓	🟩 ✓	✗
Double implicature	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	🟩 ✓	🟧 ✓	🟩 ✓	✗

Figure 7: Correlations between different implicature rates and the grammatical factors in van Tiel et al. (2016) in our experiments

4. Experiment 2

4.1. Aim

In the second part of our study, we investigated the role of Question Under Discussion in implicature derivation. As already mentioned, previous research shows that implicatures are derived at higher rates and in a less costly manner if the topic of discussion is supportive of implicatures (Kuppevelt, 1996; Zondervan et al., 2008; Degen, 2013; Degen and Tanenhaus, 2015; Cummins and Rohde, 2015; Yang et al., 2018; Ronai and Xiang, 2021, 2020) or if the QUDs make use of the stronger term of a scale rather than the weaker term (Zondervan et al., 2008; Degen, 2013; Ronai and Xiang, 2021, 2020). However, most of these conclusions have been reached by investigating implicatures with utterances which contain one single scalar item, while the effect of QUD on implicatures in sentences involving multiple scales has received little attention: (Bleotu et al., 2022b) have shown that QUD may boost global implicatures in utterances such as *It is possible that some As are B*. Our second experiment addresses this gap in the literature by looking at how an explicit question containing two scalar items belonging to two scales may affect rates of implicatures.

4.2. Participants

We tested 60 American English native speakers recruited via Prolific.

4.3. Materials and methodology

The experiment investigates whether QUD impacts multiple implicature types for 43 scales embedded under the scales ⟨*all, some*⟩ and ⟨*certain, possible*⟩. The design is similar to Experiment 1 but, taking inspiration from Ronai and Xiang (2020), the sentence giving rising to implicatures now represents an answer to a question introducing the QUD. This question involves the weak scale mate of the 1st scale ⟨*all, some*⟩ and the strong scale mate of the 2nd scale ⟨*good, adequate*⟩.

(14) Bill: Are some meals good?
Mary: *Some meals are adequate.*
Would you infer from this that, according to Mary, some meals are adequate but not good?

4.4. Predictions

We predict that a QUD employing the strong scale mate of the 2nd scale should lead to more implicatures involving the 2nd scale (Local implicatures, Double Implicatures and Global Implicatures involving the 2nd scale) than in Experiment 1. Since the QUD uses the weak member of the 1st scale, we expect no increase for implicatures with the 1st scale.

4.5. Results

In Experiment 2, we found that implicature rates vary with implicature type. Thus, overall, participants derived global implicatures involving the 1st scale at a rate of 88.45%, local implicatures at a rate of 83.14%, double implicatures at a rate of 83.14%, followed by global implicatures involving the 2nd scale at a rate of 37.67%. To exemplify, the rates for the different implicature types are represented graphically in Figures 3, 4, 5, and 6.

Taking the ⟨*all, some*⟩ as a baseline, we conducted an ANOVA with the dependent variable number of Yes implicature answers (coded as 1) and the fixed effects Interpretation (global implicature with the 1st scalar item, local implicature, double implicature, global implicature with the 2nd scalar item) and 2nd Scale. Interpretation demonstrated statistically significant effects, as evidenced by its F value of 971.005 (p <2e-16 ***). Similarly, the 2nd Scale factor exhibited significant effects with an F value of 5.893 (p <2e-16 ***). The interaction between Interpretation and 2nd also showed a statistically significant F value of 1.838 (p = 4.1e-08 ***). While scalar diversity does not go away, 23 scales show no difference in implicature rates (e.g. ⟨*hot, warm*⟩, ⟨*finish, start*⟩).

We then compared the rates of implicatures in Experiment 2 to Experiment 1. We conducted an ANOVA with the dependent variable number of Yes answers (coded as 1) and the fixed effects Task (Experiment 1: no QUD vs. Experiment 2: QUD) and Interpretation (global implicature with the 1st scalar item, local implicature, double implicature, global implicature with the 2nd scalar item). The analysis of variance revealed significant main effects for both Task (F(1, 17930) = 189.66, p <2e-16) and Interpretation (F(3, 17930) = 1686.31, p <2e-16), as well as a highly significant interaction effect between Task and Interpretation (F(3, 17930) = 68.16, p <2e-16). These results suggest that both individual factors and their interaction have a substantial impact on the dependent variable. As expected, posthoc Tukey tests reveal no significant difference in the rates of global implicatures with the 1st scalar item. Moreover, the rates of local implicatures, global implicatures involving the 2nd scale and double implicatures are overall significantly higher in Experiment 2: between the two experiments, there is a substantial difference in the rates of local implicatures, (with a mean difference of 0.1469 (95% CI: [0.1099, 0.1837], p <.001)), global implicatures involving the 2nd scale (with a mean difference of 0.3849 (95% CI: [0.3453, 0.4245], p <0.001)) and double implicatures (with a mean difference of 0.1591 (95% CI: [0.1222, 0.196], p <.001)). However, the rate of global implicatures involving the 1st scale is overall significantly smaller in Experiment 2 (with a mean difference of -0.0599 (95% CI: [-0.0968, -0.0230], p <.001)).

We see lexical scale variation in the rates of local implicatures. An ANOVA with acceptance rates for local implicatures as the dependent variable and the fixed effects Task and 2nd scale reveals significant main effects for both Task (F(1, 4399) = 141.685, p <2e-16) and 2nd Scale (F(42, 4399) = 3.422, p = 8.63e-13), as well as a highly significant interaction effect between Task and 2nd Scale (F(42, 4399) = 2.146, p = 2.66e-05). Posthoc Tukey tests reveal that this significant interaction is due to the scales ⟨*hot, warm*⟩, ⟨*hideous, ugly*⟩, ⟨*black, dark*⟩, ⟨*enormous, big*⟩ and ⟨*stunning, attractive*⟩. Other scales do not manifest significant difference in the rates of local implicatures in Experiment 2 compared to Experiment 1. A similar scalar diversity effect can be seen in the rates of double implicatures. An ANOVA with acceptance rates for double implicatures as the dependent variable and the fixed effects Task and 2nd scale revealed significant main effects for both the Task (F(1, 4398) = 165.645, p <2e-16) and 2nd Scale factors (F(42, 4398) = 4.633, p <2e-16). Additionally, there was a significant interaction effect between Task and 2nd Scale (F(42, 4398) = 1.713, p = 0.00291). The interaction suggests that the effect of Task on the dependent variable may vary across different levels of the 2nd Scale. Posthoc Tukey tests reveal that there is a significant difference between the two experiments for the scales ⟨*beautiful, pretty*⟩, ⟨*cold, cool*⟩ ⟨*unique, special*⟩ and ⟨*ugly, hideous*⟩. In the case of global implicatures with the 2nd scalar item, an ANOVA with acceptance rates as the dependent variable and the fixed effects Task and 2nd scale reveals significant main effects for both the Task (F(1, 4398) = 41.330, p = 1.42e-10) and 2nd Scale factor (F(42, 4398) = 5.675, p <2e-16). Additionally, there was a marginally significant interaction effect between Task and the 2nd Scale (F(42, 4398) = 1.375, p = 0.0547): a significant difference between experiments can be seen for the scales ⟨*will, may*⟩, ⟨*certain, possible*⟩, ⟨*unavailable, scarce*⟩ and ⟨*scared, wary*⟩.

Additionally, an ANOVA with with acceptance rates for implicatures as the dependent variable and the fixed effects Task, 1st scale and Interpretation reveals significant main effects for both the 1st Scale Factor (F(1, 4481) = 137.571, p <2e-16) and Task(F(1, 4481) = 4.371, p = 0.036615), as well as a significant interaction between Task and the 1st Scale Factor (F(1, 4481)= 12.958, p = 0.0003). In Experiment 2, participants tend to derive a similar rate of local implicatures, as well a similar rate of double implicatures with scales embedded under ⟨*certain, possible*⟩ and under ⟨*all, some*⟩, whereas in Experiment 1, local and double implicature rates tend to be lower for scales embedded under ⟨*all, some*⟩ than for ⟨*certain, possible*⟩. Interestingly, global implicatures involving the first scale tend to be quite high for scales embedded under either *some* and *possible*. Global implicatures involving the second scale tend to be derived at lower rates for scales embedded under either *some* and *possible* in both experiments.

Regarding the predictors of scalar diversity in van Tiel et al. (2016), we find that the rates of implicatures with the 2nd scale item correlate more with semantic distance than with boundedness or other factors (see Figure 7). The addition of the QUD thus results in an important difference concerning the relation between predictors and implicatures rates compared to Experiment 1.

5. Discussion

With respect to Q1, the question regarding the extent to which various implicatures types are derived across different scales, our study has shown that rates of local implicatures, double implicatures and global implicatures involving the 2nd scale vary with scalar diversity in multiple scalar item utterances. Overall, there seems to be a general preference to derive global implicatures with the 1st scalar item, followed by local and double implicatures, and a general dispreference for global implicatures with the 2nd scalar item. The high rates of global implicatures with the 1st scalar item compared to the lower rates of other types of implicatures suggest that the order of appearance of scalar items matters: the scalar item which appears first gives rise to more implicatures than the scalar item which appears second, regardless of scale type. However, we do find non-negligeable rates of local implicatures with the 2nd scalar item, as well as double implicatures (higher than 50%). These results go against a gricean view which assumes that local implicatures cannot be derived given that the mechanisms of deriving implicatures target whole utterances. Instead, they suggest that it is possible to derive implicatures in embedded contexts. This is further supported by the existence of double implicatures, where both weak scalar terms are strengthened to the negation of their stronger alternatives. Nevertheless, the first scalar item seems to be privileged with respect to the second, which may be taken to suggest either that the mechanisms of deriving implicatures with the first vs second scalar item are different (pragmatic vs grammatical, for instance) [5], or simply that the first position is more accessible or available to participants.

Interestingly, we find that participants tend to generally derive more local and double implicatures with scales embedded under *possible* than under *some*. This goes against the findings of Bleotu et al. (2022b), who found that participants derived very few local implicatures under *possible*. It is unclear why this contrast arises, but in the current experiments, when deriving local implicatures, participants may treat *possible* as a *think* predicate, which they could potentially even ignore. This matter is in need of further exploration.

Additionally, as an answer to our second research question (Q2), we find that, in both experiments, implicature rates for different lexical scales correlate with semantic distance: the more semantically distinct the scale-mates of the 2nd scalar item are, the more local implicatures and double implicatures we find. The availability of lexical scales had no effect. Thus, the findings of van Tiel et al. (2016) seem to carry over to implicatures with utterances containing multiple scalar terms. The absence of a correlation between lexical availability of scales and rates of implicatures with the 2nd scalar item does not seem to support an Alternatives-Based Account where implicatures depend on lexical availability. Instead, the correlation between scalar distinctness and higher implicature rates suggests that a theory of implicature is needed which takes into account the contrast between the two scale-mates. We shall refer to such an account as the *Scalar Distinctness Account* of implicatures.

Finally, regarding the third question (Q3), addressing the role of QUD on implicature derivation, we find that local implicatures, double implicatures and global implicatures involving the 2nd scale are also sensitive to a complex QUD which employs the weak scalar term of the 1st scale and the strong scalar term of the 2nd scale. The findings of Experiment 2 support the

[5] The considerable rates of double implicatures, comparable overall to local implicature rates, suggest that exhaustification can apply locally, to parts of utterances, thus supporting the Grammatical account.

idea that access to the stronger alternative of the 2nd scale boosts implicatures involving the 2nd scale. The results are thus in line with the Alternatives-Based approach and research on alternatives for single scale utterances (Gotzner and Romoli 2022; Tieu et al. 2016; Skordos and Papafragou 2016). However, it is worth mentioning that, while in Experiment 1, the lexical availability of the 2nd scale was not a predictor of derivation of implicatures employing the 2nd scale, in Experiment 2, the contextual discourse availability of alternatives seems to impact implicatures more than their lexical availability. Our results thus highlight that there is an noteworthy difference between the lexical availability of alternatives and their contextual availability: implicatures seem to depend on how easy it is for participants to retrieve a stronger alternative in a given context rather than in general.

Another important observation we can make is that the QUD seems to reduce scalar diversity to a significant extent: most of the scales show high rates of implicatures employing the 2nd scalar term. Moreover, context reduces the effect of boundedness on implicature-derivation, possibly because the strong scale-mate of the 2nd scale acts as an upper bound. This is also expected in a theory which assumes that implicature derivation depends on the discourse availability of the scale. Once a stronger alternative is made available in the discourse by means of a question containing the weak scale mate of the 1st scale and the strong scale mate of the 2nd scale, participants no longer need to go through the effort of retrieving the strong scale mate of the 2nd scale, they will simply strenghten the embedded term by negating the upper bound and thus deriving an implicature.

The QUD findings complement the findings related to the predictors of scalar diversity, suggesting that an explanatory theory of implicature derivation should consider (at least) two components: (i) scalar distinctness, and (ii) contextual availability of the scale in the discourse. We thus embrace a specific version of the Alternatives-Based Account, which we refer to as **the Contextual Alternatives and Scalar Distinctness Account**. Overall, participants tend to derive more implicatures when they are aware of a (lexical/contextual) contrast between the two scale-mates, and when the stronger scale-mates is made available in the discourse context, but not when the scale is generally more lexically available to them.

6. Conclusion

In the current paper, we have extended van Tiel et al. (2016)'s inference task to investigate various implicature types (global, local and double) in utterances embedding scalar terms belonging to multiple scales under *some* and *possible*. We noticed an overall pattern: global implicatures involving the 1st scale tend to be derived at higher rates than implicatures involving the 2nd scale (local and double implicatures or global implicatures involving the 2nd scale). We showed that all the types of implicatures we tested increase in the presence of an explicit question introducing the QUD. Moreover, while in the absence of a QUD, implicatures involving the 2nd scale are correlated with semantic distance and with boundedness, once a QUD is added, boundedness no longer predicts implicature rates. We have suggested that this can be taken to support a version of the Alternatives-based Account (**the Contextual Alternatives and Scalar Distinctness Account**) where contextual availability of alternatives is more important than lexical availability, and where, additionally, the (lexical/contextual) distinctness of the scales matters. We are currently extending our investigation to other types of QUD, further manipulating the strength of the scalar terms.

References

Barner, D., N. Brooks, and A. Bale (2011). Accessing the unsaid: the role of scalar alternatives in children's pragmatic inference. *Cognition 118*, 84–93.

Benz, A., A. Bombi, and N. Gotzner (2018). Scalar diversity and negative strengthening. *ZAS Papers in Linguistics 60*, 191–203.

Benz, A. and N. Gotzner (2017). Embedded disjunctions and the best response paradigm. In R. Truswell, C. Cummins, C. Heycock, B. Rabern, and H. Rohde (Eds.), *Proceedings of Sinn und Bedeutung 21*, University of Edinburgh.

Bill, C., E. Pagliarini, J. Romoli, L. Tieu, and S. Crain (2021). Children's interpretation of sentences containing multiple scalar terms. *Journal of Semantics 38*(4), 601–637.

Bleotu, A. C. (2021). Deriving scalar implicatures in Romanian 7-and 9-year-olds. In A. Sevcenco, L. Avram, and V. Tomescu (Eds.), *L1 Acquisition and L2 Learning: The view from Romance*, Chapter 13, pp. 332–353. John Benjamins Publishing Company.

Bleotu, A. C., A. Benz, and N. Gotzner (2021a). Shadow playing with romanian 5-year-olds. epistemic adverbs are a kind of magic! In *Proceedings of ELM 1*, pp. 59–70.

Bleotu, A. C., A. Benz, and N. Gotzner (2021b). Shadow playing with Romanian 5-year-olds. Epistemic adverbs are a kind of magic! In *Experiments in Linguistic Meaning*, Volume 1, pp. 59–70.

Bleotu, A. C., A. Benz, and N. Gotzner (2021c). Where truth and optimality part. Experiments on implicatures with epistemic adverbs. In *Experiments in Linguistic Meaning*, Volume 1, pp. 47–58.

Bleotu, A. C., A. Benz, and N. Gotzner (2022a). Global implicatures and QUD: An experimental investigation. In F. Frau, L. Bischetti, C. Pompei, B. Scalingi, F. Domaneschi, and V. Bambini (Eds.), *Book of Abstracts – XPRAG 2022*.

Bleotu, A. C., A. Benz, and N. Gotzner (2022b). Romanian 5-year-olds derive global but not local implicatures with quantifiers embedded under epistemic adverbs: Evidence from a shadow play paradigm. In *Proceedings of Sinn und Bedeutung*, Volume 26, pp. 149–164.

Bleotu, A. C., R. Ivan, A. Nicolae, G. Bîlbîie, A. Benz, M. Panaitescu, and L. Tieu (2023). Not all complex disjunctions are alike: On inclusive and conjunctive interpretations in child Romanian. In *Proceedings of the Annual Conference of the Cognitive Science Society 45*, pp. 3062–3069.

Chemla, E. and B. Spector (2011). Experimental evidence for embedded scalar implicatures. *Journal of Semantics 28*, 359–400.

Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax / pragmatics interface. In A. Belletti (Ed.), *Structures and Beyond*, pp. 39–103. Oxford: Oxford University Press.

Chierchia, G., D. Fox, and B. Spector (2012). Scalar implicature as a grammatical phenomenon. In C. Maienborn, K. von Heusinger, and P. Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning*, Volume 3, pp. 2297–2331. Berlin: De Gruyter Mouton.

Clifton, Charles, J. and C. Dube (2010). Embedded implicatures observed: A comment on. *Semantics and Pragmatics 3*(7), 1–13.

Crnič, L., E. Chemla, and D. Fox (2015). Scalar implicatures of embedded disjunction. *Natural Language Semantics 23*, 271–305.

Cummins, C. and H. Rohde (2015). Evoking context with contrastive stress: Effects on pragmatic enrichment. *Frontiers in Psychology 6*, 1779.

Degen, J. (2013). *Alternatives in Pragmatic Reasoning*. Ph. D. thesis, University of Rochester.

Degen, J. and M. K. Tanenhaus (2015). Processing scalar implicature: A constraint-based approach. *Cognitive Science 39*(4), 667–710.

Foppolo, F., M. T. Guasti, and G. Chierchia (2012). Scalar implicatures in child language: Give children a chance. *Language Learning and Development 8*, 365–394.

Franke, M., F. Schlotterbeck, and P. Augurzky (2017). Embedded scalars, preferred readings and prosody: An experimental revisit. *Journal of Semantics 34*, 153–199.

Geurts, B. (2009). Scalar implicatures and local pragmatics. *Mind & Language 24(1)*, 51–79.

Geurts, B. and N. Pouscoulous (2009, July). Embedded implicatures?!? *Semantics and Pragmatics 2*(4), 1–34.

Geurts, B. and B. van Tiel (2013). Embedded scalars. *Semantics and Pragmatics 6*(9), 1–37.

Gotzner, N. and J. Romoli (2022). Meaning and alternatives. *Annual Review of Linguistics 8*(1), 213–234.

Gotzner, N., S. Solt, and A. Benz (2018). Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology 9*, 1659.

Grice, H. P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Guasti, M. T., G. Chierchia, S. Crain, F. Foppolo, A. Gualmini, and L. Meroni (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes 20*(5), 667–696.

Horn, L. R. (1972). *On the Semantic Properties of Logical Operators in English*. Doctoral dissertation, University of California, Los Angeles.

Huang, Y. T. and J. Snedeker (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology 1*, 376–415.

Kennedy, C. and L. McNally (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language 81*(2), 345–381.

Kuppevelt, J. v. (1996). Inferring from topics: Scalar implicatures as topic-dependent inferences. *Linguistics and Philosophy 19*(4), 393–443.

Landauer, T. K. and S. T. Dumais (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104*(2), 211–240.

Landauer, T. K., P. W. Foltz, and D. Laham (1998). Introduction to latent semantic analysis. *Discourse Processes 25*(2-3), 259–284.

Noveck, I. (2001). When children are more logical than adults: Investigations of scalar implicature. *Cognition 78*, 165–188.

Orr, S., M. Ariel, and E. Shetreet (2024). Scales and inferences. Unpublished ms.

Papafragou, A. and N. Tantalou (2004a). Children's computation of implicatures. *Language Acquisition 12*, 71–82.

Papafragou, A. and N. Tantalou (2004b). Children's computation of implicatures. *Language Acquisition: A Journal of Developmental Linguistics 12*(1), 71–82.

Pouscoulous, N., I. A. Noveck, G. Politzer, and A. Bastide (2007). A developmental investigation of processing costs in implicature production. *Language Acquisition: A Journal of Developmental Linguistics 14*(4), 347–375.

Ronai, E. and M. Xiang (2020). Pragmatic inferences are qud-sensitive: An experimental study. *Journal of Linguistics 57*(4), 841–870.

Ronai, E. and M. Xiang (2021). Exploring the connection between question under discussion and scalar diversity. *Proceedings of the Linguistic Society of America 6*(1), 649–662.

Skordos, D. and A. Papafragou (2016). Children's derivation of scalar implicatures: Alternatives and relevance. *Cognition 153*, 6–18.

Sun, C., Y. Tian, and R. Breheny (2018). A link between local enrichment and scalar diversity. *Frontiers in Psychology 9*, 2092.

Tieu, L., J. Romoli, P. Zhou, and S. Crain (2016). Children's knowledge of free choice inferences and scalar implicatures. *Journal of Semantics 33*(2), 269–298.

Tieu, L., K. Yatsushiro, A. Cremers, J. Romoli, U. Sauerland, and E. Chemla (2017, 08). On the Role of Alternatives in the Acquisition of Simple and Complex Disjunctions in French and Japanese. *Journal of Semantics 34*(1), 127–152.

van Tiel, B. (2014). *Quantity Matters: Implicatures, Typicality and Truth.* Ph. D. thesis, Radboud Universiteit Nijmegen.

van Tiel, B., E. van Miltenburg, N. Zevakhina, and B. Geurts (2016). Scalar diversity. *Journal of Semantics 33*(1), 107–135.

Yang, X., U. Minai, and R. Fiorentino (2018). Context-sensitivity and individual differences in the derivation of scalar implicature. *Frontiers in Psychology 9*, 1720.

Zevakhina, N. (2012). Strength and similarity of scalar alternatives. In *Proceedings of Sinn und Bedeutung*, Volume 16, pp. 647–658.

Zondervan, A., L. Meroni, and A. Gualmini (2008). Experiments on the role of the question under discussion for ambiguity resolution and implicature computation in adults. In T. Friedman and S. Ito (Eds.), *Proceedings of Semantics and Linguistic Theory (SALT) 18*, pp. 765–777.