Foreword

This is a collection of papers presented at the Formal and Experimental Pragmatics Workshop that was part of the 16th European Summer School for Logic, Language and Information held in Tübingen, Germany, in August 2014. The workshop aimed to provide a forum for the presentation of cutting-edge research in data-oriented formal pragmatics. Seven contributed talks and two alternates were complemented by three guest lectures by invited speakers Judith Degen (Stanford), Bart Geurts (University of Nijmegen) and Roger Levy (UC San Diego).

We would like to thank the organizers of ESSLLI and the local staff for their support and hard work. We appreciate deeply the time and effort that the members of our programme committee have invested. Lastly, we also would like to thank the presenters and the audience for making formal and experimental pragmatics such a vibrating and exciting field to work in.

Judith Degen
Michael Franke
Noah D. Goodman
Programme Committee

Anton Benz (ZAS, Berlin)
Oliver Bott (University of Tübingen)
Sarah Brown-Schmidt (University of Illinois)
Michael C. Frank (Stanford)
Bart Geurts (University of Nijmegen)
Dan Grodner (Swarthmore College)
Daphna Heller (University of Toronto)
Gerhard Jäger (University of Tübingen)
Dan Lassiter (Stanford)
Roger Levy (UC San Diego)
Chris Potts (Stanford)
Jacopo Romoli (University of Ulster)
Uli Sauerland (ZAS, Berlin)
Kristen Syrett (Rutgers University)
Bob van Tiel (University of Bielefeld)
# Table of Contents

**Embedded implicatures revisited: Issues with the Truth-Value Judgment Paradigm**  
Anton Benz & Nicole Gotzner .................................................. 1

**Exhaustive inferences and additive presuppositions: The interplay of focus operators and contrastive intonation**  
Nicole Gotzner & Katharina Spalek ........................................... 7

**Convergence and divergence between word learning and pragmatic inferences**  
Napolean Katsos & Elspeth Wilson ........................................... 14

**Mandatory implicatures in Gricean pragmatics**  
Sven Lauer ............................................................................... 21

**The processing cost of interpreting superlative modifiers and modals**  
Yaron McNabb & Doris Penka ................................................. 29

**Between Intersective and ‘Split’ Interpretations of Predicate Conjunction: The Role of Typicality**  
Eva Poortman ............................................................................ 36

**Markedness, Frequency and Lexical Change in Unstable Environments**  
Gerhard Schaden ......................................................................... 43

**The production of partial answers and ad-hoc inferences: evidence from semi-spontaneous speech**  
John Michael Tomlinson & Camilo Rodriguez-Ronderos ................. 51
Embedded implicatures revisited: Issues with the Truth-Value Judgment Paradigm

Anton Benz (benz@zas.gwz-berlin.de)
Centre for General Linguistics, Schützenstrasse 18
10117 Berlin, Germany

Nicole Gotzner (nicole.gotzner@googlemail.com)
Humboldt-Universität zu Berlin
Department of German Studies and Linguistics, Dorotheenstr. 24
10117 Berlin, Germany

Abstract

A seminal paper by Geurts and Pouscoulous (2009) has claimed clear experimental evidence against local implicatures and superiority of the truth value judgement paradigm over inferencing tasks. We present three experiments that challenge these claims. We show that due to inherent methodological problems, the study by Geurts & Pouscoulous, contrary to appearance, provides neither evidence for nor against localism.

Keywords: embedded implicatures, truth value judgements, localism, globalism.

Introduction

One of the most important debates in pragmatics concerns the question of whether implicatures can arise locally or are computed after completion of semantic derivation. Consider a sentence like 1, in which the scalar item some is embedded under the quantifier all.

(1) All kids ate some of their cookies.

Pragmatic theories make different predictions concerning the implicatures of embedded quantifiers. A possible interpretation of 1 is that all kids ate some but not all of their cookies. According to localist theories (Chierchia, 2004; Chierchia, Fox, & Spector, 2012), a quantifier like some can trigger an implicature under embedding, leading to the interpretation just described. Globalist theories (Sauerland, 2004; Geurts, 2009), on the other hand, claim that a reading with an embedded implicatures does not arise, since strengthening applies at the sentence level (globally). On a global reading, 1 is interpreted as meaning that all kids ate some (and possibly all) cookies.

It is clear to everyone involved that without objective experimental basis the debate about the pragmatic effects of embedded scalar expressions cannot be decided. Reasoning about scalar expressions is subject to a number of constraints. Of particular importance is the condition that the contrast between scalar alternatives $A(e_i)$ must be relevant. Cancellability of implicatures and dependency on contextual relevance can cause problems for experimental studies. If implicated meaning is not relevant, or if it is even negated by contextual expectation, then the experimental results may not reflect the true extent to which the test group may normally arrive at an implicature.

A study by Geurts and Pouscoulous (2009) (henceforth G&P) tested implicatures of complex sentences experimentally and claimed overwhelming experimental evidence against strong localism. In view of the influence of this study, we set out to replicate and test G&P’s experiments. In this paper we present the results of three experiments that followed up on G&P with the same experimental methodology. We found that, contrary to appearance, the data of G&P provide evidence neither for nor against localism.

Previous experimental studies

A large body of experimental research on implicatures has emerged over the past decade. However, the majority of experiments concentrate on questions of acquisition (Noveck, 2001; Papafragou & Musolino, 2003; Katsos & Bishop, 2011), or the time course of implicature processing (Noveck & Posada, 2003; Grodner, Klein, Carbayo, & Tanenhaus, 2010; Tomlinson, Bailey, & Bott, 2013) and the question whether they are generated by default or triggered in context (Breheny, Katsos, & Williams, 2006). There has only been a small number of studies on implicature of complex sentences (Chemla, 2009; Geurts & Pouscoulous, 2009; Clifton Jr & Dube, 2010; Chemla & Spector, 2011). In the following, we provide an overview of the relevant previous experiments.

Among the first to test embedded implicatures experimentally were Geurts and Pouscoulous (2009). As noted before, their paper is widely considered a final refutation of (strong) localism. In their study, they tested local implicatures triggered by the quantifier some embedded under different quantifiers. As a second methodological point, G&P claimed to have demonstrated the superiority of the truth-value judgement task over inferencing tasks in testing implicatures.

In their truth-value judgement paradigm, participants were exposed to pairs of pictures with geometrical shapes and corresponding sentences. They had to judge whether the sentences were true or false in the situation shown in the picture. For example, participants judged the statement ‘All the squares are connected with some of the circles’ in a picture.

1See (Sauerland, 2012) for an overview of different positions.

2As witnessed, for example, by the fact that it received three immediate replies in Semantics and Pragmatics: (Sauerland, 2010; Ip-polito, 2010; Clifton Jr & Dube, 2010).
All the squares are connected with some of the circles.

☐ true  ☐ false

Figure 1: Example stimulus from Geurts and Pouscoulous (2009)

For (2a) in Figure 1, the logic of the TVJT was that participants should respond with false if they had drawn the embedded implicature that the squares were connected to some but not all circles, as predicted by localist accounts.

G&P found no indication of local implicatures in the truth-value judgement task (e.g., in the example all participants responded with true). As the experimental results show no sign of local implicatures, G&P concluded that strong localism is refuted. This view was shared by many commentators.

A Methodological Remark

So, what do we test when we test for embedded implicatures? Ultimately, we want to find out the composition of the population of pragmatic language users. The complete population divides into those who interpret semantically and those who interpret pragmatically, by which we mean that they follow some interpretation strategy based on Gricean maxims. The localist / globalist debate is about the composition of the pragmatic population solely. Globalists claim that they are identical to the set of language users who employ a globalist interpretation strategy, and localists claim that they are identical to the users who employ a localist strategy. This decomposition is shown in Figure 2.

The test subjects will, in general, be drawn from a mixed population consisting of semantic interpreters (Sem) and pragmatic interpreters (Prag). The claims of globalists and localists are conditional in nature. Globalists claim that, if test persons belong to Prag, then they interpret globalistically (Prag ⊆ Glob). Localists make the respective conditional claim stating that Prag ⊆ Loc. None of the discussed theories make any claim about semantic interpreters, nor do they claim that semantic interpreters do not exist. Hence, in order to falsify the respective positions experimentally, it has first to be
shown that the test persons are indeed pragmatic interpreters. Exactly this is missing in G&P’s study: there is no test that would show that their subjects are drawing implicatures at all. The results of the study are consistent with the assumption that subjects interpret semantically. This argumentative gap alone would render G&P’s arguments inconclusive. What we are going to do in our first two experiments is to show that, in the set-up of G&P, subjects overwhelmingly interpret semantically. This means that G&P’s results say nothing about the composition of Prag, and therefore about the localist / globalist controversy.

A crucial fact which seems to be unnoticed is that a global interpretation cannot be distinguished from a semantic interpretation in the experimental setup. That is, a true response could reflect that participants pursue a global reading or that they simply judge based on semantic meaning. Our experimental research, therefore, follows up on the sentence-picture-verification paradigm developed by G&P with the overall goal to verify how substantial the proportion of test subjects is who follow a pragmatic interpretation strategy.

**Experiment 1: Unembedded *some* as a Control**

In the first experiment, we added a control item with an unembedded implicature to assess the proportion of pragmatic and semantic responders. We presented the test sentence *Some of the squares are connected to circles* in conjunction with a picture on which all squares were connected to circles. Further, we tested all other conditions and experimental items from the original experiment. As mentioned before, the experimental results of G&P are consistent with the null hypothesis that participants evaluate purely on the basis of semantic meaning (see also Figure 3). If G&P’s paradigm is sensitive to implicatures, then this item should receive a high percentage of false answers.

**Participants**

40 Participants with U.S. IP addresses were recruited via the Amazon Mechanical Turk platform and screened for native language. Only participants with English as a native language were included in the final analysis. A total of 37 participants (16 male, 21 female) with a mean age of 37.6 were analysed. Participants received 15 cents for compensation.

**Materials and Procedure**

Each experimental survey started with demographic questions concerning age, gender and educational level. The main part of the experiment instructed participants to judge whether a given sentence was *true* or *false* in the situation represented by a preceding picture. Participants were supposed to look at the pictures carefully but to give the response that first came into their minds (all instructions are found in the supplementary materials). We used exactly the same conditions and experimental materials as G&P to guarantee full comparability with their experiments. In particular, we kept the items in which test sentences are downward entailing, or non–monotonic (neither downward nor upward entailing), although they are not critical to our investigation. Three of the experimental items were predicted to be judged *true* (all, more than one and exactly two) on a globalist account whereas they should be judged *false* when participants generate a local implicature (see Fig. 3 in the results section).

The negative contexts *not all* and *not more than one* are predicted as *false* on all accounts. We further added a control item with the sentence *Some of the squares are connected to circles*. In this non-embedded occurrence of *some*, a high percentage of *false* responses should be observed.

Each participant judged one item per condition only, totalling in 7 experimental items per subject. In addition, we constructed 6 filler items that separated each experimental item in the formal experiment. The filler items resembled the experimental items but did not contain quantifiers in the test sentences. Two survey versions with pseudo-randomized orders of test and filler items were created. A given participant only took one of the survey versions.

**Results and Discussion**

Figure 3 summarizes the predictions by the semantic, localist and globalist accounts in conjunction with the results of G&P and the results of Experiment 1. As can be seen from Figure 3, the results for the original items are in agreement with G&P’s. However, only 38% judged the control item with the unembedded case of *some* *false* (62% *true*). This item fits best with the semantic hypothesis, although it deviates from all three hypotheses, the semantic, localist, and globalist. The difference in the amount of true answers and the hypothesized response pattern (100 % for semantic and 0 % of true answers for globalist and localist hypothesis) was significant as shown by Fisher’s exact test (p’s <.0001).

The percentages for the unembedded control item are similar to that reported by G&P for their own control experi-
ment (Experiment 2, 34%), which used different experimental items. Other experiments with comparable designs report higher percentages, e.g., Noveck (2001, Table 4) reports 59%, Noveck and Posada (2003) report 63%, and Papafragou and Musolino (2003) 92.5% in a modified TVJT as percentages of adults who reject ‘some’ sentences in conditions in which their implicatures are not satisfied.

The low number of scalar implicature (henceforth, SI) readings for the unembedded control item in G&P’s paradigm indicates that generally in this paradigm only few participants pursue a pragmatic interpretation. It should therefore not be surprising that these numbers are even lower in the embedded conditions. Hence, the lack of implicatures in the embedded cases might be a simple null finding due to lack of task sensitivity.

Summarising our results, we can say that, on the one hand, the results of Experiment 1 replicate G&P’s original study, on the other hand, the low percentage of SI responses in the unembedded ‘some’ condition implies that the TVJT in G&P’s paradigm is not very sensitive to implicatures overall.

However, there remain 38% who judged the unembedded ‘some’ control item false. In order to have an estimate of how strongly the scalar reading influences the experimental results, we performed a control experiment, Experiment 2, in which test subjects were explicitly instructed to interpret literally. If the percentage differences between the two experiments are small, then it follows that the composition of test subjects in both experiments is similar, and hence, that subjects in Experiment 1 were predominantly semantic responders.

**Experiment 2: Pragmatic or Literal Meaning?**

It seems that participants draw few implicatures generally in the sentence-picture verification paradigm by G&P. Hence, it is possible that subjects evaluate the sentences on what they believe is their semantic meaning. The second experiment tested how subjects judge the sentences when explicitly instructed to concentrate on the truth of sentences, and when, in addition, warned not to be influenced by how they would understand them in a normal conversation.

**Participants**

38 participants (20 male, 18 female) with a mean age of 34.5 were included in the final analyses. Two participants were excluded because they did not answer the practice item correctly (see next section). Participants received 15 cents for compensation.

**Materials and Procedure**

We used the same materials as in Experiment 1 but told participants in the introduction to base their judgements on literal meaning of the sentences. In order to strengthen their awareness, we gave them the example sentence ‘The squares or the circle are blue’ and told them that this sentence is normally understood such that either the squares or the circles are blue but not both. We then told them that, according to its literal meaning, both the squares and circles could be blue and instructed them to base their subsequent judgements on exactly this type of meaning.

In order to make sure that participants had read and understood the instructions, we included a practice item that displayed a picture with blue squares and circles and asked whether the sentence ‘The squares or the circle are blue’ was true in the picture. 38 participants gave the expected response and the 2 that responded with false were excluded from further analyses.

**Results and Discussion**

The results of the second experiment are shown in Figure 4 in conjunction with the results of Experiment 1. As can be seen, still 32% judged the critical control item false (68% true). We carried out a series of Fisher’s exact tests comparing the results of Experiment 1 and 2. To account for multiple testing the alpha-level was adjusted to .007 by means of a Bonferroni correction. All pairwise comparisons revealed p’s > .26 (see Fig. 4).

<table>
<thead>
<tr>
<th>Condition</th>
<th>Exp 1</th>
<th>Exp 2</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>.92</td>
<td>.92</td>
<td>1</td>
</tr>
<tr>
<td>more than one</td>
<td>.95</td>
<td>.89</td>
<td>.67</td>
</tr>
<tr>
<td>exactly two 1</td>
<td>.95</td>
<td>.92</td>
<td>1</td>
</tr>
<tr>
<td>exactly two 2</td>
<td>.14</td>
<td>.05</td>
<td>.26</td>
</tr>
<tr>
<td>not all</td>
<td>.05</td>
<td>.08</td>
<td>1</td>
</tr>
<tr>
<td>not more than one</td>
<td>.17</td>
<td>.13</td>
<td>.75</td>
</tr>
<tr>
<td>some (control)</td>
<td>.62</td>
<td>.68</td>
<td>.63</td>
</tr>
</tbody>
</table>

Although we explicitly instructed participants in this experiment to evaluate the literal meaning of the sentences, and although the practice item demonstrated that they understood the instructions, the results did not differ significantly from that of the first experiment. This indicates that in TVJT’s of the G&P paradigm, test persons judge items according to what they believe to be their semantic meaning. We conjecture that the geometrical nature of G&P’s test items in com-
bination with the TVJTs lowered the relevance of implicatures to such an extent that test persons were led to consider literal meaning only.

It remains to address the problem that the results are not consistent with the assumption of test persons judging truth values according to the compositional semantic meaning in the sense of, for example, Heim and Kratzer (1998). The 32% false answers in the unembedded some condition are significantly different from the 0% predicted by compositional semantics. Hence, the previous qualification that subjects judge items according to what they believe to be their semantic meaning. A possible explanation can be the influence of information structure. The deviation from pure semantic interpretation could be explained, for example, along the lines of (Kuppevelt, 1996) who sees implicatures as semantic consequences of information structure. In particular, van Kuppevelt claimed that scalar implicatures are triggered only if the scalar term appears in the focus / comment part of a sentence. For example, in the question–answer pair (3), the implicature ‘not more than 14’ is inferred from the fact that 14 is in focus (comment) which is defined by the background question.

(3) A: ‘How many children does Nigel have?’
B: ‘Nigel has 14 children.’

In contrast, if the question had been ‘Who has 14 children?’, the implicature would not have been triggered. If van Kuppevelt is right, this could explain why test subjects consider the implicature from some to not all in the control condition part of semantic meaning. However, this issue of the interplay between information structure and implicatures is still not very well understood and needs more studies, both theoretical and experimental.

**General Discussion and Conclusions**

The study by G&P proved highly influential. It was widely seen as an experimental refutation of strong localism, and as a proof of the superiority of TVJTs over inferencing tasks for testing implicatures. However, the claims of the authors need to be taken cautiously. The first experiment reported here showed that participants rarely follow the strengthened meaning in the paradigm, even for unembedded some. In conjunction with Experiment 2, the results indicate that participants in the G&P setup evaluate test sentences semantically.

It might be argued that in Exp. 1, even though the proportion of semantic interpreters is high, there is a significant difference between the false answers to the all condition and the some control condition, and that this is telling against local implicatures. However, there are other possible confounders, as e.g. sentence complexity or information structure, which could explain the difference so that, without further evidence, no conclusions can be drawn from this observation.

In order to confirm our conjecture that the difference in percentages between the control some and the all condition is not related to the localism / globalism issue, we conducted a third experiment with additional test sentences (e.g., Some of the squares are connected to red or blue circles) for which globalism makes stronger predictions than localism. We added three conditions, in all of which literal interpreters should answer true. In two conditions both localism and globalism predict a false answer, and in the third condition only globalism predicts it. Together with the all condition of G&P’s study, where only localism predicts a false answers, we had four critical conditions with complex sentences. We do not present the results in detail here, but to summarize most participants accepted the critical sentences as true. Furthermore, there was no significant difference between the condition where only localism predicts a false answer, and the condition where only globalism predicts it. These results confirmed our assumption that the differences between false answers in the unembedded some condition and the all condition provides no argument against localism.

Overall, these results show that G&P’s experiment cannot be adduced as evidence for or against the existence of local or global implicature. This demonstrates the importance of methodological considerations in the field of experimental pragmatics. It seems that the right kind of experimental paradigm for investigating embedded implicatures is still awaited for.

The problem is not only an issue of experimental design. An aspect which has received little attention so far is the inability of theoretical approaches to describe the actual experimental situation and to predict its experimental outcome. For example, relevance of information, expectations of test subjects, and questions under discussion are extraneous parameters for both the localist as well as the globalist approaches.

The past few years have seen the emergence of new theoretical approaches which may help to widen the debate from the localist / globalist issue to a more general discussion about context dependence. Among them, we can name Asher (2013) who accounts for embedded implicature in the context of discourse interpretation, Franke (2009) and Benz (2012) who derive them from game theoretical models, and Sauerland (2012) who provides new arguments in favour of localism. In Experiment 2 we have also seen evidence for the interaction between implicature and information structure, which may provide support for theories along the lines of Kuppevelt (1996). We hope that our study contributes to renewed interest in experimental investigations on embedded implicature, and helps to instigate a debate about methodological issues.

**References**


investigation into the role of context in generating pragmatic inferences. *Cognition, 100*, 434-463.


Exhaustive inferences and additive presuppositions: The interplay of focus operators and contrastive intonation

Nicole Gotzner (nicole.gotzner@hu-berlin.de)
Department of German Language and Linguistics, Humboldt-Universität zu Berlin
Unter den Linden 6, 10099 Berlin, Germany

Katharina Spalek (katharina.spalek@staff.hu-berlin.de)
Department of German Language and Linguistics, Humboldt-Universität zu Berlin
Unter den Linden 6, 10099 Berlin, Germany

Abstract
The research reported here investigates the interplay of the overt focus operators only and also with contrastive intonation. Our experiment consisted of two parts: a naturalness rating and a delayed truth value judgement task. The results of the rating task indicate that focus particles are perceived as equally natural combined with an H* accent and an L+H* accent. The truth value judgement task showed that L+H* accents reinforce exhaustive inferences to the same extent as the overt operator only. In the case of also, listeners infer an additive presupposition even if such a presupposition is not explicitly introduced in the context. Overall, the results indicate that once an overt focus operator is present in an utterance, L+H* accents do not contribute any further information. In utterances without overt focus operators, on the other hand, L+H* accents act like an only operator.

Keywords: Focus operators; contrastive intonation; exhaustive inferences; presuppositions.

Introduction
Successful communication involves drawing inferences about implicitly conveyed content of a message. For example, sentence (1) is usually interpreted exhaustively such that the speaker liked Tübingen but did not like Konstanz, even though this information is not explicitly stated. In cases such as (1), a so-called exhaustive implicature arises through pragmatic enrichment of meaning (see for example Groenendijk & Stokhof, 1984; Van Rooij & Schulz, 2004).

(1) Context: Last year, I went to Tübingen and Konstanz.
I liked Tübingen.

What is implicitly conveyed by the utterance above is part of the lexical semantics of only, that is, its conventional meaning. For example, a sentence like 'I only liked Tübingen' asserts that the speaker did not like Konstanz (e.g., König, 1991). With overt only, the exhaustive inference always arises and it cannot be cancelled. In the case of an exhaustive implicature, on the other hand, the speaker could continue the discourse by stating 'In fact, I liked Konstanz too' cancelling the exhaustive inference. Nevertheless, we could think of the utterance in (1) as containing a silent only operator, accounting for the fact that a similar exhaustive meaning arises as with overt only (Groenendijk & Stokhof, 1984; Van Rooij & Schulz, 2004; Chierchia, 2004).

Such exhaustive inferences are reinforced by contrastive intonation (e.g., Chevallier et al., 2008, see also the next section). According to Pierrehumbert and Hirschberg (1990), certain intonational contours are associated with a contrastive interpretation.1 Pierrehumbert and Hirschberg (1990) distinguish between a complex L+H* accent consisting of a low initial target followed by steep rise and a simplex H* accent with a single high pitch accent. The authors assume that L+H* accents trigger a contrastive inference while H* accents have a non-contrastive meaning.2

In the present study, we compare the exhaustive inferences triggered by contrastive intonation, non-contrastive intonation, the overt focus operator only and a combination of only with either type of intonation. One possible hypothesis is that particles like only require a contrastive accent (see for example Jacobs, 1983; Dimitrova, 2012). Alternatively, contrastive intonation might not add any further information once an overt operator is present in an utterance. In other words, contrastive intonation might act like a covert only operator (Groenendijk & Stokhof, 1984; Van Rooij & Schulz, 2004; Chierchia, 2004).

We further compare exhaustive inferences to the additive presupposition associated with also. Particles like also presuppose that an alternative holds and add the assertion that the proposition is true for the focused element (König, 1991). For example, the sentence 'I also liked Konstanz' presupposes that the speaker liked Tübingen in addition to Konstanz. also is considered a 'hard' presupposition trigger in that it always triggers a presupposition, which in most contexts cannot be easily accommodated (Abusch, 2002). Since also is an additive particle, contrastive intonation should not affect speaker's inferences. Another possibility could be that also is incompatible with a contrastive accent. In that case, utterances that combine also with an L+H* accent might be judged as less felicitous than utterances with also and H* accent.

Previous research
Focus particles and contrastive intonation In order to determine how focus particles interact with contrastive intonation, it is important to find out with what accent type focus

1We use the terms contrastivity and exhaustivity synonymously here without intending any theoretical implications. In our examples contrastive and exhaustive inferences coincide (e.g., A and not B is equivalent to A and nothing else).

2Note, however, that there is an open theoretical debate whether those two accent types represent distinct categories or are variants of the same accent (see for example Krahmer & Swerts, 2001)
particles most naturally combine with.\textsuperscript{3} Two previous studies provide some insights into this question. Sudhoff (2010) conducted a production and a perception experiment to investigate whether focus particles induce contrastive focus or rather interact with a given focus background structure. The results showed that focus particles were produced with both variants, either contrastive or non-contrastive pitch accents (independent of whether
only, even or also was used). In the perception study, participants had to indicate whether a contrast was present or not. The results revealed that the contrast judgements were independent of whether and which specific focus particle was used. Therefore, Sudhoff (2010) suggested that the phonetic realization of a focused element is dependent on the focus background structure of the linguistic context and not on the presence or absence of a focus particle.

Dimitrova (2012) used the ERP technique to investigate how sentences with the dutch particle ‘alleen’ (‘only’) are processed in contexts with and without contrastive accenting. She found a positive ERP component around 200-500 ms for contrastive accents compared to utterances without an accent. The particle only modulated the processing of accented constituents (compared with sentences without only): the accent positivity was delayed in sentences with only and additional early anterior negativities and late left anterior positivities were triggered. According to Dimitrova, the results suggest that sentences with only trigger additional processing costs, reflecting the expectation of an accent or the fact that only may require a contrastive accent. To summarize, it is currently unclear whether focus particles require a contrastive accent or are equally felicitous with contrastive and non-contrastive accents.

**Exhaustive inferences and additive presuppositions** Regarding the associated meaning of focus particles and contrastive accents, a study by Onea and Beaver (2011) showed that participants draw more exhaustivity inferences in sentences with only compared to simple intonational focus in German (but no exact information about the accent type is given). Contrastive intonation seems to reinforce such exhaustive inferences. For example, a study by Chevallier et al. (2008) showed that participants are more likely to interpret disjunctive or exclusively when this word receives contrastive stress.

Fraundorf, Watson, and Benjamin (2010) provided evidence that L+H\textsuperscript{*} accents benefit long-term representations of exhaustive inferences in sentences without any lexical implicature triggers. They compared H\textsuperscript{*} and L+H\textsuperscript{*} pitch accents in a delayed recognition memory task involving a truth value judgement. The results indicated that the L+H\textsuperscript{*} accent increased the number of correct rejections of statements about a mentioned alternative. This indicates that the L+H\textsuperscript{*} accent triggered an exhaustive inference, which participants remembered even a day later.

There is also a growing body of research investigating the processing of exhaustive inferences, additive presuppositions and the role of contrastive intonation in online processing. Tomlinson, Bailey, and Bott (2013) showed that contrastive accents facilitate the speed with which listeners arrive at an exhaustive interpretation. A recent visual world study by Schwarz (in press) compared the exhaustive assertion triggered by only to the additive presupposition of also. The eygaze patterns indicated that listeners evaluated presuppositions more immediately than asserted content. So, the additive presupposition of also might be integrated earlier into sentence meaning than is the exhaustive assertion associated with only (see also (Kim, 2012) for work comparing only and also in online processing). Schwarz (in press) assumes this result to be most consistent with semantic theories that take presuppositions as a ‘pre-condition’ on the common ground. In summary, processing results may provide important insights into theories of semantic and pragmatic meaning. However, we focus on the end result of processing, that is a speaker’s final interpretation here.

**The present experiment**

In the experiment presented here, we explore the interplay of overt focus operators with contrastive intonation, comparing the inferences triggered by (i) L+H\textsuperscript{*} accents, (ii) H\textsuperscript{*} accents, (iii) the exclusive particle only, (iv) the additive particle also, and (v) a combination of focus operators with either type of intonation.

We pursue two research questions. First, we address the question what accent types focus particles most naturally combine with by using acceptability ratings. Second, we are interested in whether listeners draw the same amount of exhaustive inferences from the truth-functional operator only compared to a contrastive accent and whether a combination of the two leads to additive effects. In the case of also, contrastive intonation should not affect judgements.

Participants in our experiment were presented with auditory discourses in German. In a context sentence, a contrast set with two persons was introduced (e.g., ‘The judge and the witness followed the argument’). In a critical sentence, one of those persons was mentioned again carrying either an H\textsuperscript{*} accent or an L+H\textsuperscript{*} accent without any preceding particle (e.g., ‘The judge believed the defendant’). Four further conditions were created in which the critical sentences started with the particle only or also combined with either the H\textsuperscript{*} or L+H\textsuperscript{*} accent on the following critical noun (e.g., ‘Only the judge believed the defendant’).

After participants were presented with the stimuli, they performed a naturalness rating. Further, we assessed participants’ memory for inferences about the alternative mentioned in the first context sentence (i.e. not the critical noun itself). After a delay of about 1.5 minutes (5 trials) participants were presented with statements about the discourses and had to judge whether these statements were true or false regarding the content of a given story.\textsuperscript{4}

\footnotesize\textsuperscript{3}We thank Duane Watson for a helpful discussion concerning this point.

\footnotesize\textsuperscript{4}We use a delayed truth value judgement task to avoid
If participants interpret the discourses exhaustively, they are expected to indicate that the statement about the mentioned alternative is false (especially in the conditions with only and L+H* accent). In the case of the particle also, we expect participants to infer that the alternative statement is true. Our stimuli do not introduce the presupposition of also explicitly. Nevertheless, we expect listeners to infer this presupposition.

Methods

Participants A total of 24 native speakers of German (18 female and 6 male, mean age 25.25 years, age range 20-31) were recruited from a participant pool at the Institute of Psychology of Humboldt University. In total, 25 participants had participated in the study, but one of the subjects was bilingual (English as first native language) and was therefore excluded from further analyses. Participants were paid seven Euros in compensation. None of them reported any vision or hearing difficulties.

Materials The discourses followed the structure presented in (2). A first context sentence introduced two persons. A second critical sentence mentioned one of the two persons again describing an action. A third sentence continued the narrative. Note that the discourse did not introduce the presupposition associated with also. For example, there was no sentence stating that somebody else (the witness) believed the defendant.

(2) Context sentence:
Der Richter und der Zeuge verfolgten die Beweisführung.
'The judge and witness followed the argument.'

Critical sentence:
(Nur)/(Auch) der Richter/RICHTER glaubte dem Angeklagten.
'(Only)/(Also)5 the judge/JUDGE believed the defendant.'

Continuation sentence:
Er verkündete das Urteil.
'He announced the verdict.'

The experimental design fully-crossed the factors particle condition and accent type resulting in the six conditions summarized in Table 1.

All sentences were recorded in a sound-proof room by the first author of this paper. The critical sentences were spoken in four different versions, with an H* accent, an L+H* accent, the particle only and also. Figure 1 shows the averaged intonational contour of the accented syllable of the critical noun across accent type conditions. Acoustic measurements were conducted revealing that the condition with L+H* accent had a longer duration, a greater intensity and a higher pitch excursion. Details of the acoustic analysis are available from Gotzner, Spalek, and Wartenburger (2013). To create all six versions of an experimental item, we cross-spliced the particles from the originally recorded stimuli into either the sentence with H* accent or L+H*.

In total, there were 60 critical items corresponding to the structure presented in (2). Participants were probed with critical statements concerning the experimental items. For example, participants were asked to judge whether the statement "The witness believed the defendant" was true or false. The critical statements were always about the alternative mentioned in the context sentence. For example, if the critical sentence was "The judge believed the defendant" the statement mentioned the witness.

We further added 20 filler items with the same structure as the experimental items and a set of 40 filler items with mild pragmatic violations so that participants rated acceptable and less acceptable items. An example of an unacceptable filler item is shown in (3). Participants were probed with general statements about the narrative in the filler items (not involv-
ing alternative statements). These general statements could concern any of the three sentences and any of the characters.

(3) Marie und Josi waren im Pferedstall.
‘Marie and Josi were in the horse barn.’
Marie wollte die Pferde striegeln.
‘Marie wanted to groom the horses.’
Sie bereitete immer das Essen vor.
‘She always prepared the food.’

Six experimental lists were created, rotating through the accent and particle conditions. To each list, the 60 filler items were added totalling in 120 items per list. Each participant received a different randomization with at most two filler items appearing in a row and the different experimental conditions being repeated at most twice.

Procedure An on-screen instruction explained the structure of the experiment. The instructions told the participants that they will have to rate the naturalness of auditory stimuli and to remember the content of the stories. They were told to judge how coherent and natural the stories were on a scale from 1 (not at all acceptable) to 7 (very acceptable). We further asked them to indicate whether a statement was true or false regarding the content of a particular story. We told participants that the statements could be implicit in the story. Five practice trials were administered before the experiment started.

The experiment consisted of two paired phases: (a) rating phases and (b) truth value judgement phases. Figure 2 represents the structure of one experimental block. Rating and judgement phases interchanged every five items. After a total of five rating trials, a screen informed participants that the judgement phase would start and they had to judge the given statements as true or false.

Each experimental trial began with a central fixation cross displayed for 500 ms. Then, participants heard an item over headphones. Subsequently, they saw another fixation cross for 500 ms immediately followed by a scale from 1 to 7. Participants had to use a left and right button to browse through the array of numbers. By pressing a third button, they confirmed the selected number. Participants had a time window of up to 8000 ms for the rating. One presentation block lasted about 1.5 minutes.

In the judgement phase, a fixation cross was displayed for 500 ms and then a blank screen was shown for 100 ms, immediately followed by the statement. A time window of up to 10000 ms was allowed for the judgement. The statement was coloured green so that participants could easily identify the judgement phase. The sequence of statements corresponded to the order of presentation of the auditory stimuli. After 5 judgement trials, a screen announced the start of the next presentation/rating block.

In total, there were 24 short blocks. The entire experiment lasted about 45 minutes, including the post experimental questionnaire. In the questionnaire, none of the participants seemed to have been aware of the purpose of the experiment.

Results

Naturalness rating

Table 2 shows the mean ratings across conditions. The average rating of the unacceptable filler items was 2.5.

<table>
<thead>
<tr>
<th>Accent/Particle</th>
<th>only</th>
<th>also</th>
<th>bare</th>
</tr>
</thead>
<tbody>
<tr>
<td>H*</td>
<td>5.9</td>
<td>5.7</td>
<td>5.8</td>
</tr>
<tr>
<td>L+H*</td>
<td>5.9</td>
<td>5.7</td>
<td>5.9</td>
</tr>
</tbody>
</table>

We computed a mixed model with the rating scores including the factors particle condition, accent type, an interaction of these two factors and random effects for subjects and items. The bare H* condition served as reference level. The model did not reveal any significant differences across conditions (all p’s >.3).

The results of the mixed model are summarized in Table 3. The lack of significant differences indicates that the combination of the different particles with either an H* or an L+H* accent was perceived as equally felicitous. Further, note that the discourse with also was felicitous even though the additive presupposition was not explicitly introduced.

Table 3: Summary of mixed effects model for naturalness ratings (n = 1920, log-likelihood = -3256)

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>5.84936</td>
<td>0.2032</td>
<td>28.783</td>
<td>0.0001</td>
</tr>
<tr>
<td>L+H*</td>
<td>0.06731</td>
<td>0.22169</td>
<td>0.304</td>
<td>0.7468</td>
</tr>
<tr>
<td>only-L+H*</td>
<td>-0.20353</td>
<td>0.2177</td>
<td>-0.935</td>
<td>0.3072</td>
</tr>
<tr>
<td>also-L+H*</td>
<td>-0.10234</td>
<td>0.2177</td>
<td>-0.47</td>
<td>0.6016</td>
</tr>
<tr>
<td>only-L+H*:H*</td>
<td>0.18429</td>
<td>0.3107</td>
<td>0.593</td>
<td>0.5132</td>
</tr>
<tr>
<td>also-L+H*:H*</td>
<td>-0.08997</td>
<td>0.3107</td>
<td>-0.29</td>
<td>0.7556</td>
</tr>
</tbody>
</table>
Truth value judgement

The mean percentages of TRUE answers from the judgement task are found in Table 4. Figure 3 shows an interaction plot of the results.

Table 4: Mean percent of TRUE responses across conditions

<table>
<thead>
<tr>
<th>Particle/Accent</th>
<th>only</th>
<th>also</th>
<th>bare</th>
</tr>
</thead>
<tbody>
<tr>
<td>H*</td>
<td>18.3 %</td>
<td>74.6 %</td>
<td>35.4 %</td>
</tr>
<tr>
<td>L+H*</td>
<td>13.7 %</td>
<td>74.2 %</td>
<td>21.3 %</td>
</tr>
</tbody>
</table>

Figure 3: Mean percent of TRUE answers across conditions (circle = also, square = bare, triangle = only).

The responses were analyzed with logit mixed effects models using the bare H* condition as reference level. The results showed that participants gave significantly more FALSE responses in the condition with L+H* than H* accent ($\beta = -.80$, SE = .30, z = -2.64, p < .01) as well as comparing only and the H* condition ($\beta = -1.00$, SE = .30, z = -3.25, p < .01). There were significantly more TRUE responses in the condition with also than with H* accent ($\beta = 1.89$, SE = .30, z = 6.33, p < .0001). There was also a marginal interaction of also with accent type ($\beta = 0.80$, SE = .43, z = 1.85, p = .065). The results of the mixed model are detailed in Table 5.

A post hoc Tukey test revealed that there was no effect of accent type in the only and also particle conditions. That is, responses did not differ in the H*-only condition and L+H*-only condition (p = .9). Similarly, there was no difference between the H+ also and L+H*+also condition (p = .99). Moreover, neither the difference between only-L+H* and the bare L+H* accent nor between only-H* and the bare L+H* was significant (p = .98 and p = .5, respectively).

We further evaluated whether the bare H* condition differed from chance level. A t test showed that participants gave reliably more FALSE responses than expected by chance (t = -3.39, p < 0.01).

Discussion

The rating data presented here indicate that the use of a particular focus particle was perceived as equally natural with an H* pitch accent and an L+H* accent on the focused element (and the different bare accent types were also rated the same). This finding is in line with the study by Sudhoff (2010) and it suggests that focus particles do not necessarily require a contrastive pitch accent.

An objection might be that the lack of acceptability differences indicates that participants did not perceive the difference between our intended accent conditions. Such an objection is ruled out by data from the truth value judgement task: Participants computed more exhaustive inferences in the condition with L+H* compared to H*, indicating that the L+H* accent was perceived as more contrastive than the H* accent.

The judgement data replicate the finding by Fraundorf et al. (2010) that participants compute and encode more exhaustive inferences when exposed to contrastive accents compared to non-contrastive accents. Interestingly, participants were equally able to derive such an inference in the L+H* condition and the explicit only condition and the effects of accenting and focus particles were not additive (i.e., only-L+H* did not differ from only-H*). This finding indicates that the L+H* has the same function as the overt only operator. We are not assuming that L+H* accents and the explicit only operators do not differ on any level. However, it seems that the exhaustive inference triggered is equally strong in both cases. Therefore, one might argue that the L+H* accent acts as an only operator. This finding is line with theories assuming that implicatures are triggered by a silent only/exh operator (Groenendijk & Stokhof, 1984; Van Rooij & Schulz, 2004; Chierchia, 2004).6

6Note, however, that we are not claiming that the data are incompatible with the Gricean view, since we did not pit theories of implicature against each other.
In this respect, it is interesting to note that participants gave more FALSE responses in the H* condition than is expected by chance level. This suggests that a similar process might have been triggered with the H* accent and that this process was reinforced by the L+H* accent. Even without the L+H* accent, one could infer that the stronger statement does not hold (e.g., The judge and the witness believed the defendant). The L+H* accent increases the likelihood of the exhaustive implicature to occur and possibly the speed with which listeners arrive at the inference (see for example Tomlinson et al., 2013). Regarding the function of pitch accents, such a result might be more compatible with a theory that views the L+H* accent as the more contrastive variant of the H* accent (see for example Krahmer & Swerts, 2001) rather than the categorical view proposed by Pierrehumbert and Hirschberg (1990).

Our study further revealed that listeners infer the presupposition of also even if it is not contextually introduced. Also is considered as a hard presupposition trigger (e.g., Abusch, 2002). This means that (i) the presupposition of also is always triggered and (ii) that it is not easily accommodated (as compared with for example the presupposition of the definite article). The structures we used did not introduce the presupposition of also. Nevertheless, participants were willing to infer this presupposition. This is evident in the high amount of TRUE responses in the condition with also as well as the high naturalness ratings. In fact, ratings were as high as for the most neutral condition with the H* accent.

This finding indicates that, at least in certain contexts, presupposition accommodation of hard triggers might be easier than previously thought or at least not impossible. We propose that this result is due to the structural position in which also occurred in our sentences. Krifka (1993) notes that in sentence initial position (contrastive topic position), it is possible to accommodate an implicit question like ‘What happened?’ We assume that since also was not embedded in the critical sentence and because a possible antecedent was close in the context sentence, presupposition accommodation was possible. We predict that if one tested a sentence with also in object position (e.g., ‘Der Richter glaubte auch dem Angeklagten’) in the same experimental paradigm acceptability and the amount of TRUE responses would decrease profoundly. In future research, we aim to examine the impact of information structure and the structural position of also on presupposition accommodation.

Conclusions

The study reported here provides insights into the impact of information structure, in particular contrastive intonation, on exhaustive inferences. First, it indicates that intonation has an important impact on inferencing (in conjunction with other recent studies). Second, the study indicated that L+H* accents act like implicit only operators and were found to be equally effective as overt only operators. If an overt focus operator was present, contrastive accents, however, did not seem to contribute any further information. A third interesting finding is that, in certain structural positions, the presupposition of also can be accommodated.

Acknowledgments

For helpful comments we are grateful to Stephen Crain, Filippo Domeneschi, Cory Bill, Peng Zhou, Jacopo Romoli, Jesse Harris and Brian Leahy. This work was supported by the DFG as part of the collaborative research centre SFB 632 Information Structure.

References


Schwarz, F. (in press). Presuppositions vs. asserted content in online processing.


Tomlinson, J., Bailey, T., & Bott, L. (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of Memory and Language, 69*(1), 18–35.

Convergence and divergence between word learning and pragmatic inferencing

Napoleon Katsos (nk248@cam.ac.uk)
Department of Theoretical and Applied Linguistics
University of Cambridge
9 West Rd, Cambridge, CB3 9DP

Elspeth Wilson (ep321@cam.ac.uk)
Department of Theoretical and Applied Linguistics
University of Cambridge
9 West Rd, Cambridge, CB3 9DP

Abstract

Word learning by exclusion and pragmatic inferencing of implicature have typically been studied separately. However, there is a growing assumption that they are in fact similar processes, and therefore it can be expected that children might be competent in implicature younger than previously thought. We motivate this assumption with some theoretical observations and by drawing on empirical studies, and suggest that, indeed, pragmatic maxims can be applied by children in word learning contexts. Thus young competence in implicature is to be expected. We moderate this claim with observations of divergence across the word learning and implicature inferences, in how they are modelled by Gricean theory, and finally consider what pragmatic theory would predict for development, proposing that some word learning and implicature inferences may be easier than others and therefore acquired or applied earlier.

Keywords: Implicature; word learning by exclusion; pragmatic inferencing; language acquisition.

Introduction

Until recently the two extensive literatures on word learning and on implicature have remained largely separate. But there is now a dawning realisation, and even growing assumption, that this separation may be an artefact rather than a meaningful distinction (e.g., Barner, Brooks, & Bale, 2011; Katsos, in prep; Katsos & Bishop, 2011; Stiller, Goodman, & Frank, under review).

In the word learning literature, much debate has centered on word learning by exclusion, or ‘mutual exclusivity’, starting with Markman & Wachtel’s (1988) description of the phenomenon. In a typical experiment, the child must disambiguate the referent of a novel word from two potential referents. They are presented with two objects, one familiar and one unfamiliar, and then instructed, for example, ‘show me the blicket’ or questioned ‘which one’s the blicket?’. Many studies have found that young children and babies choose the unfamiliar object for the novel label at above chance levels, which is typically accounted for by reasoning that excludes the familiar object on the basis of constraints or pragmatic inference (e.g., Bion, Borovsky, & Fernald, 2013; Diesendruck, 2005; Diesendruck, Carmel, & Markson, 2010; Diesendruck & Markson, 2001; Jaswal, 2010; Markman & Wachtel, 1988; Markman, Wasow, & Hansen, 2003; Scofield & Behrend, 2007; Zosh, Brinster, & Halberda, 2013).

Experimental pragmatically, meanwhile, have been concerned with the nature and acquisition of pragmatic inferences, particularly quantity implicature, and its subtype, scalar implicature, as well as manner and relevance implicatures (e.g., Barner et al., 2011; Guasti et al., 2005; Noveck, 2001; Papafragou & Tantalou, 2004; Schulze, 2013; Tribushina, 2012). In this typical example of a scalar implicature, the hearer reasons that had all the friends been at the party, the speaker would have said so:

A: Were all your friends at the party?
B: Some were.

+>Some but not all of A’s friends were at the party.

If we consider how word learning and implicature might be related, there are three options:

1. There is no similarity at all.

2. There is a superficial similarity, but no underlying common process.

3. There is a similarity; they share the same processes.

In this paper we argue that there is indeed an interesting similarity in terms of processes (option 3), though this must be moderated by some observed differences.

Specifically, we make two suggestions. Firstly, word learning by exclusion is just one of the multiple pragmatic strategies available to infants for word learning. If pragmatic word learning and deriving implicatures are similar processes, we would expect to find convergence in their appearance over development. Secondly, there are differences between these types of inference, which suggest differences in acquisition. We use these observations to predict an order of acquisition for different word learning and pragmatic inferences, in the simplest case. This leads to further suggestions and predictions for experimental work to test these hypotheses, with the aim of filling in the picture of early pragmatic development in the transition from non-verbal to verbal communication.

We support our theoretical observations with evidence from experimental studies, and throughout assume a Gricean perspective, while also drawing on neo-Gricean theories (e.g., Levinson, 2000).
Convergence and divergence across inferences

Convergence

There are, broadly, two views on the phenomenon of word learning by exclusion: it is explained either in terms of constraints (e.g., Jaswal, 2010; Markman & Wachtel, 1988), or with pragmatic reasoning about the speaker’s intended meaning (e.g., Clark, 2009). In principle, the two views are compatible, with both strategies potentially available depending on context or developmental stage (as Lewis & Frank, 2013, suggest from a formal perspective), and there is experimental support for both (e.g., Diesendruck, 2005, and Grassmann, Stracke, & Tomasello, 2009, for the pragmatic view; Jaswal, 2010, for constraints). Here we focus on the pragmatic view, and raise the possibility that word learning converges with other pragmatic processes, like implicature derivation. Three lines of evidence support this claim.

A theoretical motivation

Pragmatic formulations of word learning by exclusion are usually based on Clark’s (2009) Principles of Contrast & Conventionality: an expectation that for certain meanings, a conventional form is used in the language community (Bloom, 2002; Clark, 2009; Diesendruck, 2005; Diesendruck et al., 2010). ‘If the speaker does not use the conventional form for the known referent, they must intend the other, novel referent’, it is reasoned. This can equally be understood in Gricean terms. Clark’s Principles involve reasoning about form, which suggests that word learning by exclusion is akin to Gricean manner inferences (Grice, 1991), because they also involve reasoning about the form of an expression (as Marchena, Eigsti, Worek, Ono, & Snedeker, 2011, and Brosseau-Liard & Hall, 2011, suggest). In the following example, the hearer contrasts the more complex construction ‘made the car stop’ with the alternative that could have been said, ‘stopped the car’, and infers that the speaker intended to communicate stopped the car in the normal way, they would have said so in the normal way:

\[
\text{He made the car stop.} \\
\rightarrow \text{He made the car stop not in the normal way.}
\]

If we take ‘obscure’, to mean ‘non-conventional’, then avoiding obscurity is the same as sticking to convention. Word learning by exclusion is an application of the manner maxim to the task of word learning. In the situation with a known and unknown object, in which a novel word would be an obscure way of referring to a known object for the child, the child, assuming the speaker’s co-operativity and conventionality, infers that the speaker is referring instead to the novel object in the usual way.

This leads to the idea that word learning by exclusion is not a special case, but just one pragmatic strategy – one that employs the maxim of manner – among several used in word learning. Children could equally learn words by applying quantity and relevance maxims, as Frank & Goodman (under review) and Akhtar (2002), respectively, propose. In Frank & Goodman’s study, 3-year-olds were presented with a picture of two dinosaurs, one with one unfamiliar item like a headband, and another with a headband and another unfamiliar item, a bandana. The experimenter pointed to the dinosaur with two items and said “here is a dinosaur with a dax”. To test their mapping of the novel word, the children were then presented with two more dinosaurs, one with a headband and one with a bandana, and asked, “which is the dinosaur with the dax?”. Children more often chose the unique item, the bandana, presumably on the basis of informativeness, or the maxim of quantity: to identify this referent (the dinosaur), the speaker should say as much but no more than is necessary; if the speaker intended to refer to the headband, the speaker would have not said enough to be informative because the other dinosaur also has a headband, therefore the speaker must be referring to the bandana. In Akhtar’s (2002) study, 2- and 3-year-olds heard a novel object described as ‘a wuggy one’ in one of two conditions: the preceding discourse had concerned either the shapes of novel objects, or their textures. Perhaps unsurprisingly, the children were more likely to infer that the novel adjective, wuggy referred to shape than to texture in the shape condition, and vice versa. Presumably, they could reason along such lines as: “to say something relevant in this situation, the speaker would say something about shape, so wuggy must be about the object’s shape”. See Table 1 (‘word learning – relevance’ and ‘word learning – quantity’) for a more thorough sketch of these inferences.

Following Grassman (2013), the inference for word learning by exclusion can be formulated in terms of a disjunctive syllogism with an embedded modus tollens. A disjunctive syllogism has the form: \( p \) or \( q \), not \( p \), therefore \( q \). A modus tollens is an instance of: if \( p \) then \( q \), not \( p \), therefore not \( q \). Taking the typical experimental example with two objects, one known, say, a banana, and one unknown, the ‘blicket’, the inference proceeds as follows, with (1, 3b, 4) forming a disjunctive syllogism, and (2, 3) the embedded modus tollens:

The speaker S said ‘p’

1. S could be intending to refer to \( p \), the banana, or \( q \), the unknown object
2. If S meant \( q \), S would have said ‘q’, ‘banana’
3. a) S said ‘p’, ‘blicket’, b) so S does not mean \( q \), the banana
4. Therefore, S intends to refer to \( p \), the unknown object
5. Therefore ‘blicket’ refers to \( p \), the unknown object

This is a model of the reasoning only, not a description of conscious, or even unconscious thought, although Halberda (2006) uses eye-tracking to find evidence for a disjunctive syllogism in word learning by exclusion. Strikingly, manner and quantity implicatures, and their corresponding word learning inferences, can be modelled in the same way (we return to relevance implicatures below), as they too involve contrast of what was said with what was not said, and the negation of what was not said, as depicted in Table 1. This is a theoretical indication that the relationship between word learning and implicature deserves further investigation.
### Table 1: Word learning and implicature inferences

<table>
<thead>
<tr>
<th>What was said</th>
<th>Manner</th>
<th>Quantity</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>S said ‘p’</td>
<td>S said ‘p’ instead of more conventional ‘q’</td>
<td>The context is about a topic, q</td>
<td></td>
</tr>
<tr>
<td>Possible alternatives</td>
<td>with meaning of known object</td>
<td>with meaning of q similar to p</td>
<td></td>
</tr>
<tr>
<td>Alternative action</td>
<td>If S meant q, S would have said ‘q’</td>
<td>If S meant to communicate something relevant in this situation, S would have said something about q</td>
<td></td>
</tr>
<tr>
<td>Assume</td>
<td>conventionality</td>
<td>co-operativity</td>
<td>informativeness</td>
</tr>
<tr>
<td>Negation of alternative</td>
<td>S did not say ‘q’ therefore does not mean q (in a conventional way)</td>
<td>S did not say ‘q’ therefore does not mean the more/less informative q</td>
<td>N/A</td>
</tr>
<tr>
<td>Bridging inference</td>
<td>N/A</td>
<td>‘p’ must be about q</td>
<td></td>
</tr>
<tr>
<td>Intended meaning</td>
<td>Therefore S means q in a non-conventional way</td>
<td>Therefore S means p but not q</td>
<td></td>
</tr>
<tr>
<td>Word learning</td>
<td>Therefore ‘p’ means p</td>
<td>Therefore ‘p’ means p</td>
<td></td>
</tr>
</tbody>
</table>

#### Empirical observations: common prerequisites

Word learning and implicature derivation have in common the necessary prerequisites for pragmatic reasoning: children need to be able to share in a joint-attention frame with the speaker; read their intentions; track the common ground and take in contextual cues; and assume that the speaker is co-operative. There is much evidence that these abilities develop around the first birthday (for an overview, see Tomasello, 2003, 2008), and so are available for both word learning and implicature in the early years of language acquisition.

#### Empirical observations: common behaviour

Both word learning and implicature derivation are sensitive to pragmatic manipulations, such as changes in speaker reliability (Diesendruck et al., 2010, and Grodner & Sedivy, 2011, respectively), speaker belief or epistemic state (Carpenter, Call, & Tomasello, 2002; Diesendruck, 2005; Goodman & Stuhlmüller, 2013, for word learning; Breheny, Ferguson, & Katsos, 2013, for implicature), joint attention and common ground (Grassmann et al., 2009, and Tribushinina, 2012), situation or communicative purpose (Grosse, Moll, & Tomasello, 2010; Grosse & Tomasello, 2012; Sرافон, 2009; Verbuk, 2012; and Tomasello, Strosberg, & Akhtar, 1996), and speech-accompanying acts (Grassmann & Tomasello, 2010, for word learning), though see Jaswal (2010) and Brosseau-Liard & Hall (2011) for opposing views.

From these empirical findings and theoretical observations we suggest that just as speakers apply manner, quantity and relevance maxims in implicature derivation, so too are they used by children for word learning.

#### Divergence

However, this conclusion must be moderated by observations of divergence across these inferences. Firstly, we consider differences between word learning and implicature inferences per se, and secondly divergence across manner, quantity and relevance maxims.

#### Divergence between word learning and implicature

We can note some systematic differences between word learning inferences and their corresponding implicature inferences. The end step of word learning reasoning is not an implication that enriches some literal meaning, but the endpoint is the reference, or meaning, of the novel word form. The difference between word learning by exclusion and manner inferences, in particular, probably rests on the greater knowledge required for manner implicatures for identifying the alternative expression that was not used and its stereotypical implication. Also, Davies & Katsos (2010) and Morisseau, Davies & Matthews (2013) point out that for word learning, the need to make the inference is always obvious when a novel word is heard, while this may not be the case for implicature in-
ferences, and likewise the consequence of not inferring the meaning of a novel word may have a severe consequence for communication, whereas not deriving an implicature does not always fully hinder communication.

**Divergence between maxims** Firstly, there is divergence in the type of reasoning (see Table 1). Manner and quantity inferences, and their associated word learning inferences, involve, on the pragmatic account, a disjunctive syllogism with embedded modus tollens – contrast with what was not said and its negation. Relevance inferences, on the other hand, are rather an attempt to bridge the gap between what was said and what was expected, and the implicature incorporates assumptions that are not explicitly said. For example, in this simple case the hearer relates what is said to the question by relying on background assumptions:

A: Would you like some soup?
B: I’m not hungry.

> No, I would not like some soup.

(based on the assumption, from world knowledge, that people want to eat when they are hungry, and not eat when they are not)

Secondly, the inferences differ in whether the *form* or the *informational content* of what is said and what is not said is contrasted (see Table 1 ‘possible alternatives’ and ‘alternative action’). Manner inferences involve comparing, primarily, forms that have the same literal informational content, and word learning by exclusion arguably only involves the forms of the words used (contrasting the novel form with the known form associated with the familiar object). Quantity inferences, meanwhile, involve conceptualisation and contrast of sets of lexical alternatives within the same semantic field, and relevance inferences work at the level of propositions; for both, informational content, rather than form, is key.

Thirdly, there are clearly a number of factors that vary across all the different inferences. World knowledge is essential for relevance inferences, for example:

A: Would you like a cup of coffee / camomile tea?
B: I need to work late tonight.

> Yes, I would / No, I wouldn’t.

Knowing that coffee is a stimulant and camomile soporific is necessary to infer whether the answer here is positive or negative. Similarly, deriving ad hoc scalar implicatures can be dependent on world knowledge:

A: Did you cycle from Cambridge to Brighton?
B: I cycled to London.

> I did not cycle to Brighton.

(Based on the world knowledge that London is closer to Cambridge than Brighton.)

Of course, world knowledge may play a role to varying degrees for all other types of inferences, too.

The degree of difficulty of conceptual knowledge also varies. For example, contrasting terms in the epistemic domain such as *think–know*, are understood and used at age 3 or 4, after verbs of desire (Perner, Sprung, Zauner, & Haider, 2003), while in the quantificational domain, children may understand *all* in an adult-like way aged 2 (Barner, Chow, & Yang, 2009), allowing an implicature with *some*. Children may be able to make scalar inferences in more tangible domains, such as *warm–hot*, still earlier, though such scales are not typical in experimental studies. Finally, inferences may also be more or less dependent on context, according to how salient the contrasting alternatives are in the common ground and preceding discourse (see e.g., Saylor & Sabbagh, 2004 for the affect of salience on word learning by exclusion).

In sum, there are indeed reasons to suggest convergence of pragmatic word learning and implicature inferencing, as is increasingly assumed, and so we would expect similar behaviour under manipulation of pragmatic conditions in comparable studies. However, there are also theoretical differences across inferences, and between pragmatic inferences and word learning *per se*, which lead us to expect divergence in their use by children.

**Acquisitional convergence and divergence**

These largely theoretical similarities and differences can be applied to the empirical question of how and when children make these inferences in word learning and conversation over development. The Gricean model does not demand predictive applications to acquisition, but educated guesses can be made about an order of appearance.

**Convergence**

As we have suggested that pragmatic word learning and implicature derivation are in many ways alike, we would predict a closing of the gap between the ages at which children are observed to learn words pragmatically and derive implicatures. Indeed, while early studies seemed to find that children perform poorly with implicature inferences (e.g., Guasti et al., 2005; Noveck, 2001; Papafragou & Musolino, 2003), with improved, more age-appropriate methodologies which do not require meta-linguistic judgements, the age of above-chance competence, if not adult-like performance, has been significantly revised downwards, to as young as three years (Grosse, Schulze, Noveck, Tomasello, & Katsos, in prep; Schulze, 2013; Scranton, 2009; Stiller et al., under review; Tribushinina, 2012). Further, comparable inferences are observed in non-verbal responses still earlier, at age 2 or younger (Grosse et al., 2010; Tomasello, Carpenter, & Liszkowski, 2007).

**Divergence**

The differences across inferences suggest diverging patterns in acquisition, too. To make any suggestions about the order of appearance in acquisition we need to decide: a) for each feature in the inference, which setting is likely to be easier or harder and b) which of the features carry more or less weight, i.e., how conflict between the different settings is resolved.

We suggest that:

- counterfactual reasoning (manner, quantity) is harder than bridging inferences (relevance) because it involves computation of what is not said, contrast with it, and its negation
Table 2: Features of inferences – shading indicates a difficult setting

<table>
<thead>
<tr>
<th>Reasoning</th>
<th>WL by exclusion</th>
<th>Manner</th>
<th>WL – quantity</th>
<th>Quantity</th>
<th>WL – relevance</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>with informational content</td>
<td></td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>with form</td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>involves embedded modus tollens</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Negation of what is not said</td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Form-reference association</td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Assumption of what is not said</td>
<td></td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td>Y</td>
</tr>
<tr>
<td>Particular world knowledge required</td>
<td></td>
<td>(Y)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Need for inference not necessarily obvious</td>
<td></td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td>Y</td>
</tr>
<tr>
<td>Communication possible without inference</td>
<td></td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td>Y</td>
</tr>
</tbody>
</table>

- reasoning about contrasting forms and informational content (manner) might be harder than reasoning about informational content alone (quantity and relevance)
- reasoning about informational content is harder than reasoning about form (word learning by exclusion)
- the more world knowledge required, the harder the conceptual knowledge required, and the more context independent the inference, the more difficult the inference

From these speculations a general order of expected acquisition for the easiest case emerges:
Word learning by exclusion < Word learning through relevance or informativeness / Relevance implicature < Quantity implicature < Manner implicature

This is apparent looking at Table 2, in which the suggested difficult settings of features appear shaded. No judgement is being made as to whether an implicature or mapping of a novel word to a referent is more complex.

While some studies have looked at quantity and relevance implicatures in children as young as 3 years, there are, to our knowledge, no studies that offer comparable results for different word learning and pragmatic inferences at the critical younger age, 2–5 years. If we conducted such a study, we would expect to find the simplest cases of inferencing appearing in the order suggested above, after pragmatic prerequisites and – for all but relevance – the ability to do some form of counterfactual reasoning are in place, along with a large amount of variability depending on context, conceptual and world knowledge, and the need to make an inference.

A few studies point in this direction, however. Bernicot, Laval & Chaminoud (2007) find that relevance inferences are acquired early and sarcastic ones late in 6–10-year-olds. Antoniou, Grohman, Kambanaros & Katsos (2013) use the same complete-the-story picture-matching task with 6–12-year-olds and find that relevance is easiest and irony hardest, with no difference between manner and quantity in the middle. With younger children aged 3, 4 and 5, Eskritt, Whalen & Lee (2008) find best performance in sensitivity to violations of Gricean maxims for relevance, and poor performance for quantity implicature in 3-year-olds. Again, Siegal et al’s (2010) study on bilingual 3–6-year-olds and Surian, Baron-Cohen & van der Lely’s (1996) study on 6-year-olds find better performance for relevance than quantity. Furthermore, relevance inferences arguably appear even at a pre-verbal stage in pointing (Tomasello et al., 2007), while the earliest reported robust competence with manner implicatures in pointing is at age 3 (Liebal, Carpenter, & Tomasello, 2011).

**Conclusion**

In sum, we suggest that pragmatic word learning and implicature derivation are similar to a great degree in their prerequisites and inferential processes; indeed, word learning strategies such as reasoning by exclusion may be just particular instantiations of pragmatic maxims in a word learning context. This has implications for pragmatic acquisition, specifically motivating an expectation of younger competence in implicature than initially supposed. There are also differences between inferences, as well as between word learning and implicature per se, from which we suggest an acquisitional ordering. We leave open the possibility that there may be divergence or convergence of word learning and implicature over development: as cognitive abilities, linguistic experience and contexts change, so too might learning strategies. Future research might address questions such as: in comparable conditions, are word learning and implicature inferences affected...
in a similar way by manipulation of pragmatic cues? Do manner, quantity and relevance inferences for word learning and implicature appear at different stages of development? How do other factors (e.g., number and salience of alternatives, conceptual and world knowledge) affect performance for word learning and implicature? Our specific conclusions, from the extrapolation of theory to testable hypotheses for acquisition, may be wrong, but we hope that it is nevertheless an interesting proposal as an attempt to spell out what pragmatic theory predicts for acquisition.

Acknowledgments

Elspeth Wilson is funded by an Economic and Social Research Council scholarship.

References


Frank, M. C., & Goodman, N. D. (under review). Inferring word meanings by assuming that speakers are informative.


Huang, Y. T., & Snedeker, J. (2009). Semantic meaning and pragmatic interpretation in 5-year-olds: Evidence from...
real-time spoken language comprehension. Developmental Psychology, 45(6), 1723.


Mandatory implicatures in Gricean pragmatics

Sven Lauer (sven.lauer@uni-konstanz.de)
Zukunftskolleg / Department of Linguistics
Box 216, University of Konstanz, D-78457

Abstract

Gricean implicatures are often viewed as a very weak kind of implication, viz., as optional enrichments of the literal meaning of an utterance. In virtue of being optional, they need not be present on every occasion of utterance, and they can be explicitly denied. Further, if implicatures are viewed as optional enrichments, a false implicature cannot be the reason why an utterance is infelicitous: If an implicature is known to be false, it should simply fail to arise.

I argue that this view is mistaken. There is no reason to think that Gricean reasoning cannot lead to very robust inferences, which are neither optional nor cancelable, and whose falsity can render an utterance infelicitous. Moreover, I point out that a number of explicit recent theories do not exclude such robust implicatures without additional stipulation, and that informal appeal to Gricean reasoning triggering infelicity has been made in the investigation of a number of apparently semantic phenomena. I conclude that implicatures are only weak, optional and cancelable if the pragmatic pressures driving them are (solely) based on Grice's (1975) MAXIM OF QUANTITY.

Introduction

Conventional wisdom in pragmatics has it that if a (potential) implication i of an expression e is a conversational implicature, then i must be optional and cancelable. i is optional if there are contexts in which a sincere utterance of e does not give rise to i, and it is cancelable if i can be explicitly ‘called off’ in a context in which it otherwise would arise.1

This is undoubtedly correct for many conversational implicatures, in particular the best-studied ones, implicatures that are (solely) based on Grice’s (1975) MAXIM OF QUANTITY: (1a) will implicate (1b), or at least the weaker (1c), in many contexts; and (2a) will frequently implicate (2b), but there contexts where these implicatures are absent, and they can be denied effortlessly by asserting the stronger alternative that triggers them, as in (3).

(1) a. John invited some students in his class.
   b. John did not invite all students in his class.
   c. The speaker does not know that John invited all students in his class.

(2) a. John is in Europe.
   b. The speaker does not know where in Europe John is.

(3) a. John invited some students in his class. In fact, he invited all of them.
   b. John is in Europe. In fact, he is in Paris.

Seeing as implications that are conventional in nature (such as entailments, semantic presuppositions, and conventional implicatures) are neither optional nor cancelable, the two properties have been widely used as a diagnostic for implicature-ness. Indeed, Sadock (1978) calls cancelability (which he took to include optionality, cf. n. 1) ‘the best of the tests’. While it is sometimes noted that this diagnostic use is problematic in particular cases (e.g., Sadock, 1978; Geurts, 2010, Ch. 1.5), there has been little discussion of how failures of optionality and cancelability mesh with the Gricean view of how implicatures arise. Grice’s traditional view that conversational implicatures should always be optional and cancelable stands largely unchallenged.

In the following, I argue that mandatory (i.e., non-optional, non-cancelable) conversational implicatures exist and that, despite initial appearances, their existence is compatible with a Gricean conception of pragmatic inference. The plot is as follows: In the next section, I will point out that a certain class of theories of implicatures, the optimization-based ones, already have the potential to predict mandatory implicatures. Then I will discuss a well-known implicature that arguably is of just this kind, the ‘ignorance’ implicature of disjunction. This implicature, though it initially seems like any other QUANTITY implicature, is distinguished from those by the fact that it is unaffected by considerations of relevance. For reasons that will become obvious, I call such implicatures ‘Need a Reason’ (NaR) implicatures. I will then move on to a more general characterization of such implicatures, and outline what features of an optimization-based account allow them to be modeled faithfully. Switching gears somewhat, the rest of the paper is intended to show that something quite like NaR reasoning is frequently appealed to in the explanation of implicatures that are both very robust and that can lead to infelicity, or ‘oddness’, of an utterance if the implicature is known to be false. The upshot of the discussion is that mandatory implicatures are fully expected on a Gricean conception of pragmatics, that they arise in the same way as op-

---

1Traditionally, when an implicature is not explicitly called off, but is absent, this has been called ‘implicit cancelation’ or the implicature has been said to be ‘canceled by the context’. As Hirschberg (1985) pointed out, the term ‘cancelation’ is problematic in general, as it is even more true for ‘implicit cancelation’, where it is much more natural to say that the implicature does not arise in the first place (rather than saying that it arises and is then canceled by the context), and hence we should keep optionality and cancelability apart as separate properties. Here, I focus mainly on optionality, though much of what I say extends to cancelability as well, at least on an appropriate construal of ‘cancelation’ (cf. Lauer, 2013, p. 259ff).
tional ones, and that their existence indeed has been silently presupposed in parts of the semantics literature.

**Optimization-based theories of implicature**

Many recent formalizations of Gricean theory are compatible with the existence of mandatory conversational implicatures, unless this is excluded by additional stipulation. This is true of any theory that construes conversational implicatures as inferences\(^2\) about the speaker’s beliefs and preferences that the hearer draws based on the assumption that the speaker chose his utterance so as to optimally satisfy a set of constraints, preferences, or maxims, given his beliefs. Such optimization-based theories include recent optimality- and game-theoretic ones (Blutner, 2000; Franke, 2009, a.o.) as well as the intention-based approach of (Geurts, 2010) and the dynamic pragmatics of (Lauer, 2013), but exclude generate-and-defeat theories such as that of (Gazdar, 1979).

Here is why such theories potentially predict mandatory, implicatures\(^3\) in the most abstract terms: As theories of implicatures, they will determine a set of contexts \(C_{e \rightarrow i}\) in which an utterance of \(e\) will give rise to the implicature \(i\). At the same time, as optimization-based theories, they will determine a set of contexts \(C_{\text{opt}}(e)\) in which the utterance of \(e\) is optimal, according to the speaker’s beliefs. \(i\) will then be predicted to be mandatory if

\[
C_{\text{opt}}(e) \subseteq C_{e \rightarrow i}
\]

That is, optimization-based theories will predict a (potential) implicature \(i\) of utterances of a form \(e\) to be mandatory if any context in which \(e\) is (believed to be) optimal is a context in which \(i\) arises.

If implicatures indeed were always optional the fact that these theories allow for the existence of non-optimal ones would at most be a mild embarrassment—we would have to slightly amend these theories, perhaps by specifying boundary conditions that exclude the troublesome cases. But I shall argue that this fact is much more interesting: It captures something real about how implicatures behave. The implicatures in question simply happen to be different from the most well-studied cases, pure QUANTITY implicatures like those in (1) and (2).

**The ‘ignorance’ implicature of disjunction**

In the following, I want to show that a familiar implicature is indeed mandatory in the sense just described: the ‘ignorance’ implication of unembedded disjunction. In many contexts, a speaker who utters ‘\(A\ or\ B\)’ will be taken to implicate that he does not know which of \(A\ and\ B\) is true.\(^4\)

At first blush, this looks like a run-off-the-mill QUANTITY implicature: There are logically stronger expressions (\(\text{viz.},\) the two disjuncts) that the speaker could have used, but did not, but which would have provided more relevant information. From this we conclude that the speaker was prevented from uttering these stronger expressions because he did not know them to be true. Construed in this way, the ignorance implicature of disjunction seems just like that of (2a) above, repeated here in (4a) with a context that encourages the implicature:

\[
(4)\quad \text{Ad: Where is John? I need to track him down.} \\
\quad \text{Sp: He is in Europe.} \\
\quad \therefore \text{Sp does not know where in Europe John is.}
\]

(5) is an example involving disjunction that brings out the similarity of the two cases.

\[
(5)\quad \text{Ad: Where is John? I need to track him down.} \\
\quad \text{Sp: He is in Paris or in London.} \\
\quad \therefore \text{Sp does not know that John is in London.} \\
\quad \therefore \text{Sp does not know that John is in Paris.}
\]

**Relevance and context-dependence**

The implicature in (4) is heavily context-dependent, and it is not hard to furnish a context in which it does not arise. All we have to do is to ensure that the additional information provided by stronger assertions is irrelevant in the context, as in (6).

\[
(6)\quad \text{[Somewhere in San Francisco, CA, Ad and Sp are planning a dinner party, talking about who they should invite.]} \\
\quad \text{Ad: Is John in town?} \\
\quad \text{Sp: No, he is in Europe.}
\]

Given that John’s precise location is irrelevant if he is not in town, Sp’s utterance in (6) does not implicate that he does not know where in Europe John is. Things are very different with (5). When we put the sentence in the same context, it retains its implicature:

\[
(7)\quad \text{[Context as in (6)]} \\
\quad \text{Ad: Is John in town?} \\
\quad \text{Sp: No, he is in Paris or in London.} \\
\quad \therefore \text{Sp does not know that John is in London.} \\
\quad \therefore \text{Sp does not know that John is in Paris.}
\]

Sp’s utterance in (7) implicates that Sp does not know in which city John is, even though this information is irrelevant in the context of utterance.

\(^2\)Throughout, I use ‘implicature’ to refer to an inference drawn by the audience. This is at variance with the use advocated by philosophers such as Bach (2006), who insist that implicatures are intended inferences (which hence are present even if the inference is not drawn, and absent if the inference is drawn, but was not intended). I take this to be a terminological disagreement, and invite readers who share Bach’s view to mentally replace every instance of ‘(conversational) implicature’ throughout with ‘pragmatic inference’.

\(^3\)From here on out, I drop the modifier ‘conversational’, as conventional implicatures will not play a role in what is too follow.

\(^4\)There is, of course, another much-discussed putative implicature of (unembedded) disjunctions, the exclusivity implicature ‘\(\text{not}\ A\ and\ B\)’, which arguably behaves like a run-off-the-mill pure QUANTITY implicature (if it is an implicature at all). I hence have nothing to say about it here.
One implicature or several?

A slight complication: Of course, none of the sentences in (4)–(7) necessarily implicates that $Sp$ does not know where John is. The sentences may also implicate that he is unwilling to share this information, or that while he is willing to share it with $Ad$, he does not want to reveal it to potential overhearers, etc. In all cases, $Sp$ has some reason to not give more information.

We can describe this in two ways. Either we say that the sentences in (6) and (7) give rise to several potential alternative implicatures (ignorance, non-cooperativity, politeness, etc.) or we say that the sentences give rise to single, underspecified implicature, which can be paraphrased as ‘the speaker had a reason to not give more information’, where the reason can be lack of knowledge, a desire to withhold knowledge, politeness, etc. The second way of talking allows us to leave unmodified the claim that unembedded disjunctions trigger an obligatory implicature: The addressee must infer that there was such a reason for using the disjunction in (7), while he did not have to infer this in the case of (6).

Nothing of substance would change if we adopted the other way of talking instead. While there no longer would be a single obligatory implicature, it still would be a fact worthy of explanation that (7) necessarily must have one of the ‘alternative implicatures’, while (6) need not have any of them. In the following, I will continue to say that there is one underspecified implicature, which is optional in (6), but mandatory in (7).

Mandatory vs. ‘generalized’ implicatures

So, in contrast with run-off-the-mill QUANTITY implicatures, an utterance of (unembedded) ‘A or B’ will always trigger the implicature that the speaker had a reason to avoid an utterance of ‘A’ and ‘B’—that is, this implicature arises mandatorily. Why would $Ad$ give rise to such a non-optional implicature? The reason for this not deep or mysterious, but it is instructive to examine it in detail. Eckardt (2007) succinctly explains what is going on:

“In using a disjunction, the speaker necessarily has to mention two properties which are usually more specific. These properties are presented as salient and relevant. The simpler sentences are salient alternative utterances in context. The hearer hence will look for a reason why the speaker chose a more complex expression in order to give less information.”

Eckardt mentions two properties of disjunctions that feed into the robustness of the implicature: (i) the alternative expressions that asymmetrically entail the uttered sentence are made salient by the very utterance itself; (ii) the alternative utterances are shorter and less complex than the uttered expression.

Disjunctions arguably share property (i) with expressions that are sometimes said to trigger ‘generalized’ QUANTITY implicatures (Grice, 1975; Levinson, 2000, a.o., e.g. some, whose potential not all implicature intuitively arises more regularly than ‘particularized’ QUANTITY implicatures such as the one in (4). Indeed, on a Gricean account of these implicatures, the assumption that these items make their alternatives (all, in the case of some) salient by default arguably is all that is needed to explain why these implicatures arise more frequently than ‘particularized’ ones.

However, even though such ‘generalized’ QUANTITY implicatures may arise more frequently than ‘particularized’ ones, they are clearly optional and cancelable, and they are sensitive to the relevance of the additional information conveyed by the stronger alternative, just as the ‘particularized’ implicature in (4). Levinson (2000)’s examples in (8) and (9) illustrate this: In response to the question in (8a), (9) strongly favors the implicature that the speaker does not know (or want to reveal, etc.) that all their documents are forgeries, but the same is not true for the same sentence in response to (8b).

(8) a. Are all of their documents forgeries?
   b. Is there any evidence against them?

(9) Some of their documents are forgeries.

---

5 There is an apparent exception to this claim that, on reflection, proves the rule (I am grateful to an anonymous reviewer by reminding me of this issue by constructing a similar example involving Maximize Presupposition (see below) and hypothesizing that similar cases exist for disjunctions. Paul Portner also mentioned such cases to me in conversation): The implicature(s) can be absent if the disjunction was mentioned in the near-by context:

(i) $Ad$: If John only were in London or in Paris!
   $Sp$: But John is in London or in Paris. He landed in Heathrow this morning.

On reflection, this fits into the generalization drawn above: The reason why the speaker opted for the disjunction over the disjuncts, in this case, is discourse-based. $Sp$ opted for the disjunction to create a parallelism with $Ad$’s utterance. Note that, much like with other instances of such ‘echic’ utterance choices (such as Horn’s ([1985]) metalinguistic negation), this kind of use depends on (near-)identity with a previously uttered expression. (ii) is decidedly odd.

(ii) $Ad$: If John only were in Europe!
   $Sp$: ??But John is in London or in Paris. He landed in Heathrow this morning.

6 Eckardt (2007) is mainly concerned with embedded uses of disjunctions, but, as she points out in passing, her considerations apply to unembedded uses as well.

7 That is to say that a natural (and minimal) way to capture the idea that ‘scales’ like (some, all) are ‘lexicalized’ or ‘convention- alized’ to some degree (Horn, 1972) is to say utterances that contain an item of these scales automatically make alternative utterances containing their scale-mates salient. Not all Gricean proponents of such ‘lexicalized scales’ think of them in this way, but it is arguably the dominant conception among proponents of ‘contextualists’ who deny that ‘generalized’ implicatures arise by default (in particular among proponents of optimization-based theories, such as Geurts, 2010; Franke, 2009; Degen, 2013; Lauer, 2013).

8 In recent years, a host of experimental studies has shown that these ‘generalized’ implicatures do not in fact arise as often as introspection on isolated examples may suggest (see Degen, 2013, and references therein).

9 Degen (2013, chapter 3.6) provides experimental validation for this intuitive contrast, showing that the rate at which addressees draw the some --- not all implicature varies dramatically with the relevance of the stronger alternative.
‘Generalized’ scalar implicatures hence pattern with ‘particularized’ ones in that they are sensitive to relevance considerations. The implicature of disjunction is different, in that it arises even when the additional information provided by the stronger alternatives is not relevant. This is due, so Eckardt, to the fact that the alternatives in the case of disjunction are shorter and less complex than the uttered sentence. That is why, even when the speaker does not aim to convey the additional information, he must have a reason for choosing the dispreferred (longer, more complex) form over the alternatives. Consequently, the implicature that there is such a reason arises in every context in which an unembedded disjunction is used. For obvious reasons, I propose to call such implicatures Need A Reason (NaR) implicatures.

Generalizing NaR implicatures
Selfish preferences and cooperativity

Obviously, NaR implicatures will only be predicted to be truly mandatory if the preference for economical expression that drives them can be assumed to be present in every context of use. This is why we would not expect such mandatory implicatures on an Ur-Gricean conception, according to which the pragmatic pressures that shape pragmatic reasoning are dependent on an presumption of cooperativity, for such an assumption may always be lifted. A preference for shorter, less complex expressions, however, can also be construed as a selfish preference, in which case it is plausible to assume that a speaker would adhere by it even when he is not being fully cooperative. This meshes well with a trend in recent theories of implicature (in particular, the optimization-based ones), which make no appeal to Maxims of conversation that are motivated by a Cooperative Principle. In game-theoretic accounts like that of Franke (2009) and the dynamic pragmatics of Lauer (2013), the role of Grice’s Maxims is played, instead, by the more general notion of speaker preferences. Preferences corresponding to Grice’s Maxims (such as a preference for conveying an appropriate amount of information), in this setting, are simply preferences that speakers have in many contexts, in particular those where the preferences of speaker and hearer are well-aligned. But not all preferences that enter into pragmatic reasoning need to be of this kind, the preferences in question may also be motivated entirely by selfish concerns.¹⁰

Ceteris paribus preferences and NaR reasoning

In such a setting, it is crucial to represent preferences in the right way. It is natural to assume that speakers have a selfish preference for economical expression that is operative in every context. Such a preference, however, cannot be absolute. It applies only if everything else is equal. That is, a preference for economical expression should be subordinate to more substantive preferences, such as a preference for conveying information. That is why unembedded conjunctions do not trigger NaR implicatures: Even though the preference for economical expression favors ‘A’ over ‘A and B’, the speaker’s choice of the dispreferred form can easily be motivated by his desire to convey both A and B.

The mentioned accounts both employ a representation of such ceteris paribus preferences. Franke (2009, Ch. 2.2.4) and other game-theoretic accounts do this by assuming that messages (utterances) have costs that get subtracted from the sender’s (speaker’s) response utility which represents his other preferences. These costs, crucially, are constrained to be nominal: it is ensured that the costs are small enough that they only can break ties in the response utilities. Similarly, Lauer (2013) represents preferences by means of preference structures (Condoravdi & Lauer, 2011, 2012), which are sets of individual preferences that are ranked in terms of their ‘importance’. This ranking is taken into account in action choice in a ‘lexicographic’ manner, i.e., lower-ranked preferences make a difference only when higher-ranked preferences do not suffice to make a decision. The result, in both cases, is that a preference for shorter expressions can be assumed to be universally present, even though it can (and frequently will) be defeated by more important concerns.

Crucially, if speakers generally prefer one form e over another e’, all else being equal, then, whenever a speaker utters the dispreferred form e’, it must be that not everything else is equal. That is, there must be a higher-ranked preference that defeats the preference for e. This is how NaR implicatures are predicted to be mandatory: Whenever the speaker uses a dispreferred form, the addressee must infer the existence of such a higher-ranked preference, if he is to maintain his assumption that the speaker chose a form that he deemed ‘optimal’.

We hence can characterize the conditions under which NaR implicatures arise, as follows: Supposing the speaker faces a choice between two forms e and e’, (i) there is a preference favoring e over e’ that applies across contexts; (ii) any utterance of the dispreferred form e’ will make the preferred form e salient; and (iii) the asserted content of the dispreferred form e’ does not asymmetrically entail that of the preferred form e.

Summary

I have argued that the ‘ignorance’ implicature of disjunction is mandatory in the sense that, whenever a speaker utters an unembedded disjunction, the hearer must infer a reason why he did not assert one of the disjuncts. In the following sections, I want to show that this is not an isolated quirk of utterances involving disjunctions, by showing that the same style of NaR reasoning arguably underlies a number of recent appeals to pragmatic reasoning of a Gricean sort in order to explain implications that are so robust that, if they are known to be false, this renders an utterance infelicitous. The guiding idea is that if the speaker’s contravention of a (ceteris paribus) preference cannot be justified in context, the utterance will appear infelicitous, as the addressee cannot make sense of the speaker’s choice of a non-optimal form.

¹⁰An independent advantage of such a conception is that the accounts in question directly extend to pragmatic reasoning in cases of open conflict of interest (cf. Franke, Jager, & Rooij, 2012).
More NaR reasoning: Explaining infelicities

Magri (2009) proposes an attractive, implicature-based analysis of the infelicity of temporal modification with so-called ‘individual-level’ predicates, like ‘# John is of noble birth this month.’ However, he argues that the implicature in question cannot be a Gricean conversational implicature, because:

“Within the Gricean theory, scalar implicatures are pragmatic inferences. Hence, they have a weak status: they are optional, cancelable, and suspendable. Thus, it is not at all clear why the mismatching implicature is kept in place and an utterance of [a sentence] deemed odd, rather than the implicature cancelled or suspended or never computed, and thus the utterance rescued.”

I have cast doubt on Magri’s premiss that pragmatic inference are always optional, cancelable and suspendable. But it is worth noting that the premiss is certainly widely accepted among pragmaticists. It is frequently asserted, without qualification, in influential textbooks, both classic and recent, and optionality and cancelability are frequently used as fairly definitive tests of implicature-hood (cf. Lauer, 2013, Ch. 9).

Nonetheless, if what I said in the foregoing sections is right, we expect to find cases where Gricean implicatures are so robust that they render an utterance infelicitous if they are (known to be) false. This section is points out instances where such implicatures have indeed been hypothesized, though in all cases, the status of the involved principles or maxims is left somewhat unclear. My main purpose here is to demonstrate that if we construe the hypothesized principles as ceteris paribus preferences that are universally present, the pragmatic reasoning in question can be seen to be of a very familiar, Gricean sort.

High-negation polar questions

High-negation polar questions (HPNQs) like (10) have an ‘epistemic bias’ implication: A speaker who utters (10) implies that he believed or at least suspected until recently that John drinks.

(10) $Sp : \neg \text{John drinks?}$
\[ \neg \neg Sp \text{ believed or at least expected that John drinks.} \]

This implication is not optional, it arises whenever a HPNQ is used. It cannot be canceled. And HPNQs are generally felt to be odd/infelicitous if uttered in a context in which this implication cannot be accommodated. That is why they are frequently strange as out-of-the-blue utterances, and, as AnderBois (2011) points out, cannot serve as ‘speculative questions’ which are “designed to instigate thought and/or discussion without necessarily being answered or answerable” (Gunlogson, 2003).

(11) a. #Doesn’t God exist?
\[ (\text{cf. Does God exist?}) \]
b. #Didn’t Oswald act alone?
\[ (\text{cf. Did Oswald act alone?}) \]

And yet, recent accounts of HPNQs (Romero & Han, 2004; AnderBois, 2011) treat the epistemic-bias implication as a conversational implicature. This is an attractive move, as it is quite unclear how the presence of negation could lead to a conventional implication about the speaker’s prior beliefs. At the same time, making this assumption requires that the putative implicature is mandatory, so as to explain how it can lead to infelicity.

Romero and Han: A new economy principle. Romero and Han assume that HPNQs contain a VERUM operator, akin to the one proposed by Höhle (1992). The details of their analysis need not concern us here. What is important is that the presence of this operator is supposed to turn the utterance of a question into a particular kind of speech act, a ‘meta-conversational move’. Further, they propose that such ‘meta-conversational’ moves are subject to the economy constraint in (12).

(12) **Principle of Economy**: Do not use a meta-conversational move unless necessary (to resolve epistemic conflict or to ensure Quality).

The obvious question is what kind of principle (12) is. At the outset of their paper (n. 1), Romero and Han call it a ‘non-violable conversational principle’. This, together with the explanatory work the principle is supposed to do, that we should view (12) as a normative constraint on felicitous language use, akin to a rule in a game. This would make it easy to explain why an utterance is perceived to be infelicitous if the epistemic-bias implication is known to be false. At the same time, it would make the epistemic bias-implication something quite different from run-off-the-mill conversational implicatures, which are supposedly not driven by normative rules.

Romero and Han’s label ‘economy constraint’ suggests a different conception: Take the principle to capture a ceteris paribus preference against meta-conversational moves. As a consequence we can simplify the content of the putative preference:

(13) **Economy preference**: Avoid meta-conversational moves!

With this adjustment, the epistemic bias implication becomes a NaR implicature in the sense introduced above, at least if we can assume that a HPNQ makes salient the corresponding positive interrogative (which does not make a meta-conversational move). We directly obtain the result that the implication is universally present, and hence can explain why infelicity results when the implication cannot be accommodated, as then there is no way to construe the utterance as ‘optimal’.

AnderBois: Brevity again Though AnderBois’ account of HPNQs is based on quite different semantic assumptions, his explanation for how the epistemic-bias implication arises is essentially of the same shape (emphasis mine):

“A speaker who utters a [HPNQ] could have instead
chosen the corresponding [positive question] which makes a richer contribution to the discourse [. . .].

[. . .] To paraphrase, then, the HiNegQ conveys something like ‘Just tell me whether p holds, especially if the answer is negative.’ The first part of this paraphrase is the literal semantics we have ascribed to verum focus. The second part arises from the addressee’s pragmatic reasoning about why the speaker avoided the simpler [positive question].’

In essence, then, AnderBois assumes, just as Romero and Han do, that a speaker who opts for a HPNQ over a positive question must have a reason to do so. As the final sentence of the quotation indicates, he assumes that this is due to the fact that the positive question is structurally simpler and shorter. If this kind of reasoning is supposed to explain the non-optional epistemic-bias implication, we must assume that the preference for a shorter expression is universally present, and needs to be defeated on every occasion of utterance of a HPNQ. That is, an AnderBois’ account, the epistemic-bias implication is a NaR implicature.

Summary The informal specification of pragmatic reasoning that derives the epistemic bias implication on both Romero and Han’s and AnderBois’ account either straightforwardly is (AnderBois) or can be conceived as (Romero and Han) NaR reasoning. It hence provides an illustration of how such reasoning has been informally appealed to in order to explain very robust implications as implicatures. In turn, the NaR perspective taken here shows that these implicatures, despite their robustness, can be explained without much ado by optimization-based theories of implicatures.

Maximize Presupposition
The principle Maximize Presupposition! (MP, so named by Sauerland (2003, 2008), originally suggested by Heim (1991)) has been claimed to be responsible for a variety of implications which are now commonly referred to as ‘antipresuppositions’.11 An typical example is in (14).

(14) Mary thinks that John has a girlfriend.

\[ \neg \phi \] does not believe that John does has a girlfriend.

The idea, roughly, is that speakers are presumed to follow a principle like (15). The inference in (14) can then be derived in the usual way, as arising from the competition with forms that would have lead to a stronger presupposition.

(15) Maximize Presupposition! (MP)

Presuppose as much as possible!

While there has been a lively debate about how the content of (15) should be made more precise (Sauerland, 2008; Percus, 2006; Chemla, 2008; Singh, 2011; Leahy, 2011), little has been said about what the status of the principle is supposed to be, though it is frequently said that the principle bears a certain similarity to Grice’s MAXIM OF QUANTITY. The implicit assumption appears to be that MP functions in largely the same way. At the same time the principle is also supposed to explain infelicities like the following (indeed, this was Heim (1991)’s reason for suggesting the principle):

(16) #John interviewed a father of the victim.
cf. John interviewed the father of the victim.

(17) #Mary broke all her arms.
cf. Mary broke (both) her arms.

It may seem tempting, then, to understand MP as a normative constraint of language use, or a conventional felicity constraint. This seems to be what Singh (2011) has in mind when writes (p. 149, emphasis mine): “[MP] is a principle of language use that forces speakers to sometimes use a sentence \( \phi \) rather than a competing alternative \( \psi \).”

On this conception, MP is quite different from Gricean Maxims, which are not supposed to be regulative rules that specify what a speaker is or is not allowed to do. Instead, they are expectations about how speakers will behave, based on general assumptions about cooperative behavior.

And yet, many of the cited authors talk of MP as if it were essentially another Gricean Maxim, and appear to think that assuming such a Maxim is sufficient for explaining the infelicity of (16)–(17). And again, the considerations from the previous sections allow us to make sense of this. All we need to assume that speakers have a ceteris paribus preference for expressions that are associated with stronger semantic presuppositions, and that the items that trigger antipresuppositions as a matter of course automatically make certain alternative expressions salient.12 If this MP preference can be assumed to be active in all contexts, we obtain an explanation for the infelicities rather straightforwardly. The felicitous alternatives given in (16) and (17) differ from the infelicitous sentences only in having presuppositions that will usually be already part of the common ground (in virtue of the fact that it can generally be presumed that humans have at most one father and at most two arms). But then, uttering the sentence with the weaker presupposition will have exactly the same context-change effect as the one with the stronger presupposition. Consequently, it most contexts, it will be impossible to justify the speaker’s choice of a form that is dispreferred by the MP-preference.

It is noteworthy that, on this conception, MP is not terribly similar to the MAXIM OF QUANTITY at all. Instead, it is rather like a submaxim of the MAXIM OF MANNER. I think this is appropriate (if it is appropriate to assume MP is Maxim-like at all), for two reasons. Firstly, even though MP does not make direct reference to the morpho-syntactic form


12The recent formulations of MP in Percus (2006); Sauerland (2008); Chemla (2008) already make reference to ‘lexically specified alternatives’. As with ‘generalized’ conversational implicatures, we can assume that such lexical specification has the result of automatically making the alternatives salient.
of the utterance, it does make reference to conventionally-determined linguistic features of the utterance—in all recent formulations, it is restricted to lexically-triggered semantic presuppositions. Secondly, and more importantly, QUANTITY, but not MANNER, is always bounded by relevance considerations. QUANTITY does not require that the speaker provide as much information as he possibly can, but rather that the speaker provide as much information as is necessary or relevant. But MP, if it is to explain the infelicities in (16)–(17) must be insensitive to considerations of relevance. Otherwise, the sentences should be fine in contexts where the exact number of arms (or fathers) a person has are not relevant.

My main point here is to show that researchers that take MP to be akin to Gricean Maxims—and hence, take antipresuppositions to be essentially implicatures—must implicitly assume that the reasoning involved is essentially NaR reasoning. This does not show, of course, that we should regard MP as a Maxim-like principle, rather than as a normative rule that regulates language use. There is good reason to think, however, that if MP is real at all, construing it as a Maxim-like principle (or rather, a ceteris paribus preference) is the way to go. Here is why: Such a construal allows us to make sense of cases where the principle apparently is suspended. I am grateful to an anonymous reviewer for providing a nice example of such a case.

(18) [Context: The constitution of Phantasia stipulates: “A son of the king has to be present at the opening of the parliament”. It is known to everyone that the current king has exactly one son, who is in attendance. Running through the regulations during the opening ceremony, the speaker of parliament declares:]

As the constitution demands, all MPs have pledged their allegiance to the king, a son of the king is present, ...

The felicity of the utterance would be puzzling on a construal of MAXIMIZE PRESUPPOSITION that takes it to be a normative principle of language use (or a bona fide grammatical constraint). On the construal of MP as a MANNER-like preference, it is not puzzling at all: In this case, the use of the (non-presupposing) indefinite instead of the (presupposing) definite is motivated, not because it makes a difference to the common ground (which it could not, as the common ground entails that there is exactly one son), but rather by the parallelism with the text of the law. That is, in this case (and others like it), we can simply assume that the preference for maximizing presuppositions is defeated by a preference for staying as close to the text of the law as possible.

Summary

This section has made the case that recent accounts of putatively pragmatic inferences (viz., the epistemic bias implication of HPNQs and antipresuppositions) either can be straightforwardly seen as instances of NaR reasoning or must be construed as such if they are to meet their explanatory targets while still counting as essentially Gricean. A question that is left open by the considerations offered here is why agents should have the ceteris paribus preferences the reasoning relies on. A preference for shorter or less complex expressions perhaps makes obvious sense as an economy principle, but the same is not true for a preference against ‘metaconversational moves’ or a preference for forms that contain lexical items that trigger stronger semantic presuppositions. We would like to have a better understanding about how these preferences are motivated, or, at the very least, some independent verification (outside of the phenomena we are seeking to explain) that speakers indeed have these preferences. However, this question—which has not been addressed at all, to my knowledge, in the rather extensive literature on MAXIMIZE PRESUPPOSITION—arises just the same if the principles involved do not capture ceteris paribus preferences that drive NaR reasoning, but instead are inviolable normative constraints on language use.

My main point remains unaffected by this explanatory lacuna. I did not set out to argue that the principles in question are the right ones to account for bias implications and antipresuppositions. Rather, I wanted to show that the researchers that have endorsed these principles can (and perhaps should) be viewed as appealing to the same kind of NaR reasoning that is responsible for the mandatory implicature of unembedded disjunctions; and that hence the reasoning they informally characterize is readily captured by recent formal theories of Gricean pragmatics.

Conclusion

I have argued that there are robust, indeed mandatory, inferences that are not just pragmatic in a vague sense, but neatly fit into the Gricean fold: They arise in exactly the same way as classical examples of implicatures, and are readily predicted by existing theories of pragmatic inference. On the one hand, there are inferences that, at first blush, appear to simply be QUANTITY implicatures, but which can be shown to behave differently from canonical examples in that they survive in contexts where QUANTITY implicatures are absent. The ‘ignorance’ implication of disjunction is of this kind. Moreover, the same style of reasoning—NaR reasoning—can be shown to be operative in recent, pragmatic explanations of the bias implication of HPNQs, as well as in analyses employing Heim’s MAXIMIZE PRESUPPOSITION. In these cases NaR reasoning leads to inferences that are robust enough to explain the infelicity of utterances, something that is impossible on the traditional conception of Gricean pragmatics, according to which pragmatic inferences are always optional and cancelable.

Now that we have an understanding of how such robust implicatures arise, we can turn around and ask why the most well-studied instances of conversational implicatures—

13Because of this, I think that Schlenker (2012)’s valiant attempt to reduce MP to QUANTITY is ultimately doomed to failure. I cannot argue the point here, though, as space restrictions prevent me from doing justice to the details of Schlenker’s proposal.
pure QUANTITY implicatures—are never mandatory. In the present perspective, this is not because they arise in a different way, but because these implicatures are driven by a pragmatic pressure that is, by necessity, sensitive to relevance—QUANTITY does not favor a stronger alternative if the additional information it provides is irrelevant in the context of use. Consequently, for any given QUANTITY implicature, there will be many contexts in which it does not arise. Mandatory implicatures are different only in that they are driven by pragmatic pressures that are not sensitive to relevance considerations in the same way.

Acknowledgments

Large parts of this contribution have antecedents in Lauer (2013), and hence I am grateful to the many people who discussed aspects of this work with me over the years, in particular to Chris Potts, Paul Kiparsky and especially Cleo Condoravdi, who also provided many helpful comments on an earlier draft. All remaining errors and inaccuracies are, of course, mine and mine alone.

Part of this research was kindly supported by the EU FP7 Marie Curie Zukunftskolleg Incoming Fellowship Programme, University of Konstanz (grant no. 291784). This support is gratefully acknowledged.

References


Horn, L. R. (1972). On the semantic properties of logical operators in English. Doctoral dissertation, University of California, CA.


The processing cost of interpreting superlative modifiers and modals

Yaron McNabb (y.mcnabb@uu.nl)
Utrecht Institute of Linguistics OTS
Trans 10, Utrecht, 3512 JK, The Netherlands

Doris Penka (doris.penka@uni-konstanz.de)
Zukunftskolleg and Department of Linguistics
Box 216, 78457 Konstanz, Germany

Abstract

Superlative modifiers like at least and at most pose several challenges to formal semantic and pragmatic analyses. A particular challenge is accounting for the ignorance inferences they give rise to, and whether to attribute these inferences to the lexical semantics, the semantic combinators, pragmatic implicature, or the interaction thereof. We conducted a self-paced reading study in order to adjudicate between the various analyses proposed for superlative modifiers, taking their interaction with deontic modals as a test case. In the experiment, we aimed to determine which superlative modifier-deontic modal combinations are correlated with which readings as well as the time-course of the interpretation of these expressions in order to better adjudicate between competing analyses. We found that some superlative-modifier combinations have a clear response pattern coupled with faster processing time and argue that this is an indication that they give rise to an authoritative reading that comprehenders favoured. When such reading wasn’t available, participants resorted to a repair mechanism to generate an authoritative reading that is not compositionally available, which came at a processing cost.

Keywords: superlative modifiers; modified numerals; ignorance inferences; experimental pragmatics.

Superlative modifiers and ignorance inferences

Superlative modifiers like at least and at most have received a lot of attention recently in the semantics and pragmatics literature. A particular challenge for formal analyses comes from the fact that these expressions give rise to ignorance inferences (Geurts & Nouwen, 2007; Nouwen, 2010). For example, at least 50 minutes in (1) implies that the speaker is unsure about the baking time and for all she knows, the cake could have baked for exactly 50 minutes or longer.

(1) The cake baked for at least 50 minutes.

A number of analyses have been proposed to account for the ignorance inferences of superlative modifiers, ranging from analyses that attribute speaker ignorance to the lexical semantics to accounts deriving speaker ignorance as a pragmatic implicature. A crucial test case for these analyses is provided by the pattern of interactions of superlative modifiers and modals. As first observed by Geurts and Nouwen (2007), ignorance inferences can be suppressed in certain combinations of superlative modifiers and deontic modals. For example, when at least co-occurs with a necessity modal, as in (2), the so called authoritative reading is possible where 50 minutes specifies the lower bound of the range of allowed values, i.e. 50 minutes and longer baking times are allowable, but not baking times shorter than 50 minutes.

(2) The cake has to be baked for at least 50 minutes.

The existing analyses of superlative modifiers, which will be discussed in more detail in the following section, make different predictions regarding which combinations of superlative modifiers and modals are able to suppress ignorance inferences and what the available readings are in terms of the lower and upper bound of the range of permissible values.

We conducted a self-paced reading study to determine which superlative modifier-deontic modal combinations are correlated with which readings as well as the time-course of the interpretation of these expressions in order to better adjudicate between competing analyses.

Analyses of superlative modifiers

This section provides a brief overview of existing accounts of ignorance inferences arising with superlative modifiers focusing in particular on the predictions these analyses make regarding the interaction of superlative modifiers and deontic modals.

Ignorance inferences as lexical entailments (Geurts & Nouwen, 2007)

In Geurts and Nouwen’s analysis, ignorance inferences are hardwired into the lexical meaning of superlative modifiers. According to their analysis, at least n A are B means that the speaker is certain that there is a set of n As that are B and considers it possible that there is a larger set of As that are B.

At most n A are B means that the speaker considers it possible that there is a set of n As that are B and is certain that there is no larger set of As that are B.

Regarding the interaction with modals, Geurts and Nouwen assume a rule of modal concord, which strips off the layer of epistemic modality just in case the primary epistemic operator in the lexical entry of the superlative modifier (epistemic necessity for at least, epistemic possibility for at most) matches the modal force of the modal. This predicts that authoritative readings not conveying speaker ignorance are available if at least is combined with a necessity modal (cf. 3a) and at most with a possibility modal (cf. 6a). As modal concord is assumed to be optional, the speaker insecurity reading reading is also predicted to be possible in these cases (cf. 3b and 6b). In the other two combinations – at least plus possibility and at most plus necessity modal – given that the epistemic modal in the superlative modifier and the deontic modal do not correspond in their modal force, modal
concord is not possible and thus only the speaker insecurity reading is available (cf. 5b and 4b).

**Nouwen (2010)**

Nouwen (2010) derives ignorance inferences from a covert epistemic possibility modal embedded under the superlative modifier. He proposes that superlative modifiers are degree operators indicating minima (for at least) or maxima (for at most). The proposal builds on two additional assumptions. The first is that numerals and measure phrases are generally ambiguous between a lower- and a double-bounded meaning. Nouwen’s second assumption is that linguistic expressions compete: If a certain meaning can be expressed by two or more expressions differing in their complexity, the simpler expression is preferred and more complex expressions are blocked. The components of the analysis, taken together, predict that in many cases superlative modifiers cannot be used because the resulting sentences either express a contradiction or a meaning that is equivalent to the sentence with a bare numeral and thus blocked. To rescue such sentences, Nouwen (2010) argues that a covert epistemic possibility modal can be inserted in the scope of the superlative modifier. If the speaker is unsure about the exact value, i.e. the value varies across the worlds epistemically accessible to the speaker, the superlative modifier applies to a degree property denoting a range of values. This results in non-contradictory truth-conditions, which are not expressed by the bare numeral and thus not blocked.

As a possibility modal rescues a sentence with a superlative modifier, there is no need to insert an additional covert epistemic one in cases with an overt possibility modal. Nouwen’s account thus predicts that authoritative readings always arise if at least and at most take scope over a deontic possibility modal (cf. 4a and 6a). (The narrow scope readings are either contradictory or blocked by the bare numeral.)

When combined with necessity modals, Nouwen’s analysis predicts that neither at least nor at most expresses sensible truth-conditions, because the narrow as well as the wide scope readings are either contradictory or blocked. But we can assume that these combinations too can be rescued by inserting a covert epistemic possibility modal in the scope of the superlative modifier and above the deontic necessity modal, resulting in the speaker insecurity reading (cf. 5b). Nouwen (2010) moreover proposes that a necessity modal is interpreted as a possibility modal when minimality is at stake, such that at least plus necessity modal comes out equivalent to at least plus possibility modal and thus has the authoritative reading (cf. 5a).

**Ignorance inferences as quantity implicatures**

Another line of research, pioneered by Büring (2008) and taken up by Schwarz (2011, 2013) and Kennedy (2013), derives ignorance implications of superlative modifiers as pragmatic inferences, more precisely as quantity implicatures in a neo-Gricean fashion. While the different proposals differ in the details, the key idea is that utterances with superlative modifiers are obligatorily considered against alternative, more informative utterances. In case of unembedded occurrences of superlative modifiers, the scalar alternatives are symmetric, i.e. they cannot simultaneously be false while the assertion is true. In this case ignorance implications rather than scalar implicatures are generated, similarly to the pragmatic mechanism that gives rise to ignorance inferences in disjunction (see Sauerland, 2004).

These pragmatic accounts also predict that ignorance inferences can be obviated when superlative modifiers are combined with necessity modals. When at least and at most are interpreted in the scope of a necessity modal, the scalar alternatives are not symmetric, and consequently scalar implicatures rather than ignorance implications are generated, giving rise authoritative readings (cf. 3a, 5a). In addition, speaker insecurity readings (cf. 3b, 5b) are available from an LF where if at least or at most takes wide scope over a necessity modal. In these cases the scalar alternatives are symmetric leading to ignorance implications (cf. 3b and 5b).

For combinations with possibility modals, the neo-Gricean approach predicts obligatory ignorance inferences for both at least and at most (cf. 4b and 6b), because the narrow as well as the wide scope readings lead to symmetric scalar alternatives and thus to ignorance implications.1

**Coppock and Brochhagen (2013)**

Coppock and Brochhagen (2013) also take a pragmatic approach to the ignorance inferences arising with superlative modifiers, but cast their analysis in a different framework, namely Inquisitive Semantics. They analyze superlative modifiers as expressions denoting sets of alternatives (“possibilities” in Inquisitive Semantics parlance) that are ranked at least as high (for at least) or at most as high (for at most) according to some pragmatic ranking. Speaker ignorance is attributed to a Maxim of Interactive Sincerity, according to which a speaker should only utter a sentence denoting a set of alternatives if her information state is consistent with those alternatives.

Regarding the interaction with modals, we again have to consider the wide and narrow scope configurations. The configuration where a superlative modifier takes wide scope over a modal denotes a set of alternatives, just as unembedded cases of superlative modifiers, and thus gives rise to ignorance implicatures (cf. 3b, 5b, 4b and 6b).

If superlative modifiers take narrow scope under a modal, no ignorance inferences arise due to Existential Closure, which applies in the scope of modals and whose function is to gather all the alternatives into a single proposition corre-

---

1An obvious way to extend the pragmatic account would be to build on the fact that disjunction in combination with possibility modals leads to free choice inferences, which would go beyond the neo-Gricean approach (see Fox (2007) among others). If we assume that the Büring-Schwarz-Kennedy-account can be extended along the line of free choice, we would expect that both at least and at most give rise to authoritative readings under possibility modals. The predictions of this extended version of the analysis would then be equivalent to the ones discussed for the account of Coppock and Brochhagen (2013).
sponding to the disjunction of all these alternatives. Coppock and Brochhagen (2013) therefore predict that for each superlative modifier-modal combination, both a reading with and without speaker ignorance is possible. In the scope of a necessity modal, at least and at most specify the lower and upper bound of the deontic range, respectively (cf. 3a and 5a). For at most n in the scope of a possibility modal, Coppock and Brochhagen (2013) argue that the resulting reading, which says that values up to n are permissible, is strengthened by an exhaustivity implicature, according to which values higher than n are not permissible (cf. 6a). The same reasoning should apply to at least n in the scope of a possibility modal: The truth conditions derived from this structure specify that n and higher numbers are permissible and are subsequently strengthened by an implicature to the effect that lower numbers are not permissible (cf. 6b). Note that this strengthening by implicature effectively makes at least + ♦ equivalent to at least + □, and at most + ♦ equivalent to at most + □.

**Summary of predictions**

For the following discussion, it will be useful to summarize and graphically illustrate the readings predicted to be available by the different analyses. The straight line signifies the range of permissible paper lengths, which we will also call the deontic range, the shaded area (marked with forward slashes) signifies the epistemic range, i.e. the range of values that for all the speaker knows might or might nor be permissible.

(3) □ + at least n:
   a. G&N, N, B/S/K, C&B
   b. G&N, B/S/K, C&B

(4) ♦ + at least n:
   a. N, C&B
   b. G&N, B/S/K, C&B

(5) □ + at most n:
   a. B/S/K, C&B
   b. G&N, N, B/S/K, C&B

(6) ♦ + at most n:
   a. G&N, N, C&B
   b. G&N, B/S/K, C&B

**Experimental study**

**Research question**

We see that all the analyses discussed here make clear predictions regarding (i) which combinations of superlative modifiers and modals can suppress ignorance inferences and give rise to the authoritative reading and which only have a speaker insecurity reading, and (ii) whether the respective reading is in terms of upper or lower bound of permissible values. As the discussion in the previous section made clear, the different analyses vary considerably regarding their predictions. The aim of the experiment we report on here is to determine which readings predicted by the various analyses are in fact borne out and whether the time-course of the detected interpretations could shed light on the semantic and pragmatic complexity of the inferences required to arrive at the attested interpretations.

**Methods**

We conducted an incremental self-paced reading experiment, in which 40 German speakers (27 Female, Mean Age: 24.5) read scenarios like the following. The context (7) introduced two interlocutors, where Speaker A asked Speaker B for information. The context left open whether Speaker B had the relevant knowledge and was presented sentence for sentence. Speaker B then provided the requested information in the form of an utterance (8), which included a necessity (□, e.g., muss ‘must’) or possibility (♦, e.g., darf ‘can’) deontic modal and a superlative modifier (mindestens ‘at least’ or höchstens ‘at most’). The utterance was then followed by a description sentence (9), in which the number was either lower (UNDER CONDITION) or higher (OVER CONDITION) than the one used in B’s utterance. The utterance and description sentences were introduced on the screen incrementally as with the context, but unlike the context, they were introduced region by region and not sentence by sentence, where each region was a constituent (see 8-9 for illustration.) Then, participants were asked whether the description was in accordance with the utterance.

(7) CONTEXT: (German)

John möchte einen Kuchen backen. | Deshalb fragt er seine Mutter nach dem Rezept für seinen Lieblingskuchen. | Nachdem er alle Schritte befolgt hat, schiebt er den Kuchen in den Ofen. | Da er nicht möchte, dass der Kuchen verbrennt oder roh ist, fragt er seine jüngere Schwester Lisa, wie lange er den Kuchen backen soll. | Sie sagt ihm: |

CONTEXT: (English translation)

John wants to bake a cake. | So he asks his mother for the recipe of his favourite cake. | After he follows all the instructions, he puts the cake in the oven. | As he doesn’t want the cake to be under- or over-baked, he asks his younger sister Lisa how long he should bake the cake for. | She tells him: |

(8) UTTERANCE:
The task presented here is adapted from an offline study of superlative modifiers and deontic modals in English (McNabb & Penka, 2014).

The purpose of the online task was two-fold: First, we wanted to see whether some of the superlative modifier-modal combinations are more difficult to interpret and thus lead to processing difficulty. We expect such an effect to manifest itself in two main regions: the first one being the superlative modifier region in the utterance (8) and any spill-over effects in the following regions, and the second one being the EVALUATION REGION, in which a precise value is specified in the description sentence (47 / 53 minutes in (9)) and where we hypothesize participants arrived at a decision.

Second, we aimed to determine what the preferred readings were for the various superlative modifier-modal combinations in terms of upper and lower bound of permissible values. To see how our task allows us to determine the preferred readings, consider the two readings predicted by the various analyses for □ + at least illustrated in (3). Under the authoritative reading in (3a) only higher values than $n$ would be allowable when at least $n$ is combined with a necessity modal. Therefore, we expect to get No responses in the Under condition and Yes responses in the Over condition. Under the speaker insecurity reading in (3b), the speaker is unsure about the minimally-required number and thinks that the lower bound of the deontic range might be $n$ or more. Therefore, again, we expect to get No responses in the Under condition. But since the speaker only considers it possible, but is not certain, that $n$ or higher numbers are permissible, both Yes and No responses in the Over condition are compatible with this reading.

**Results**

In presenting the results of our study, we start with the preferred readings in terms of lower and upper of permissible values. We found that three of the superlative modifier-modal combinations led to clear interpretations, namely at most with either □ or ◇ and □ + at least. As shown in Table 1, in the □ + at least combination, the vast majority of participants accepted the description in the Over Condition and rejected it in the Under Condition. This means that the number was interpreted as denoting the lower bound. The number in the □ + at most combination was interpreted as specifying the upper bound, shown by the lower Yes rates in the Over Condition. In the ◇ + at most combination, the number is interpreted as the upper bound, shown by the fact that most participants rejected the description of in the Over Condition and accepted it in the Under Condition.

The pattern we found for ◇ + at least is less clear cut. Participants tended to choose the lower-bound reading, as shown by the fact that participants said Yes in 85.89% of the cases in the Over condition. But in the under condition, in a reliable minority (23.36%) of the cases, values lower than the ones specified by the modified numeral (e.g. 47 minutes in (9)) were accepted (W = 16297, p < 0.01), which is incompatible with the lower-bound reading.

The pattern we found for □ + at least is less clear cut. Participants tended to choose the lower-bound reading, as shown by the fact that participants said Yes in 85.89% of the cases in the Over condition. But in the under condition, in a reliable minority (23.36%) of the cases, values lower than the ones specified by the modified numeral (e.g. 47 minutes in (9)) were accepted (W = 16297, p < 0.01), which is incompatible with the lower-bound reading.

**Table 1: Means of “Yes” Responses**

<table>
<thead>
<tr>
<th>Discrepancy</th>
<th>□ + at</th>
<th>◇ + at</th>
<th>□ + at</th>
<th>♦ + at</th>
</tr>
</thead>
<tbody>
<tr>
<td>Over</td>
<td>94.56%</td>
<td>85.89%</td>
<td>5.13%</td>
<td>1.67%</td>
</tr>
<tr>
<td>Under</td>
<td>2.87%</td>
<td>23.36%</td>
<td>89.83%</td>
<td>93.57%</td>
</tr>
</tbody>
</table>

As shown in Figure 1, the reading times in the utterance were significantly longer for at least following ◇ and at most following □ both for the superlative modifier region and region 4. There was no main effect of superlative modifier or modal but there was an interaction between the two (Superlative modifier region: $F_{WWithin}(1,1548) = 5.36$, p < 0.05; Region 4: $F_{Between}(1,36) = 7.14$, p < 0.05; $F_{Within}(1,1503) = 25.21$, p < 0.01).

We have found a few significant differences in the reading times of the evaluation region of the description sentence. An ANOVA of the evaluation region shows no main effect of superlative modifier but a significant interaction between superlative modifier and modal (F(1,1472)Within = 6.68, p < 0.01). Comparing specific conditions with a sufficient number of observations, we have found the arriving at a No answer in the under condition—that is, ruling out values lower than the numeral modified by at least, thereby interpreting it as specifying the lower bound—in the ◇ + at least condition took significantly longer than arriving at the same answer in the □ + at least condition (W = 14968, p < 0.05). In addition, arriving at a Yes answer in the under condition—that is, not interpreting the modified numeral as specifying the lower bound—in the □ + at most condition took significantly longer than arriving at the same response in the ◇ + at most (W = 15350.5, p < 0.01). This difference suggests that suggesting that □ + at most is more difficult to interpret than ◇ + at most despite the clear response pattern shown in Table 1. No other comparisons reached significance.
Table 2: Means reading times of the evaluation region in the description sentence

<table>
<thead>
<tr>
<th>Condition Reply</th>
<th>□ + at least</th>
<th>◇ + at least</th>
<th>□ + at most</th>
<th>◇ + at most</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Over</strong> Yes</td>
<td>930.64</td>
<td>1135.21</td>
<td>784.87</td>
<td>725.23</td>
</tr>
<tr>
<td>No</td>
<td>584.20</td>
<td>987.10</td>
<td>950.99</td>
<td>1008.57</td>
</tr>
<tr>
<td><strong>Under</strong></td>
<td>685.54</td>
<td>1239.54</td>
<td>1207.14</td>
<td>1095.43</td>
</tr>
<tr>
<td>No</td>
<td>938.37</td>
<td>1204.68</td>
<td>814.01</td>
<td>781.88</td>
</tr>
</tbody>
</table>

**Discussion**

Before we considered the results in light of the predictions the various analyses make regarding the interpretation of superlative modifiers embedded under modals, a few remarks regarding what the results teach us about the inferences comprehenders make regarding the speaker’s epistemic state and the preference for pragmatic strength and informativity.

Recall that the contexts presented in the target items left open whether Speaker B had the relevant knowledge that would enable her to make a precise statement about the value in question. It is nevertheless likely that participants assumed that the speaker was informed about the topic the utterance pertained to. For instance, in the example stimulus in (7), that John’s sister should know how long the cake should be baked for may be inferred from the fact that he asked her for such information. Comprehenders’ likely assumption that the speaker is in the position to provide the information in the utterance is related to the concept of epistemic authority in psychology and sociology, whereby individuals attribute high confidence to information provided by a source they identify as epistemic authority, consequently often assimilating it to the common ground as uncontested truth (Kruglanski, 1989). What is relevant to this study is that even in contexts in which the knowledge of the speaker was underdetermined, it could be that participants inferred that the speaker did in fact have sufficient information and thus participants would favour the authoritative reading.

The observation that our task might have biased participants towards authoritative readings is important for the interpretation of our results in two respects. We assume that authoritative readings are faster and easier to compute as well as preferred when a competing, speaker-uncertainty reading is also available. We therefore expect that in those cases where a modal-superlative modifier combination gives rise to both the authoritative and uncertainty reading, participants would prefer the authoritative reading and ignore the speaker uncertainty reading (for the purpose of the task used in this experiment), although the uncertainty reading is a theoretically possible and coherent reading. Moreover, we interpret significantly prolonged reading times of a certain type of superlative modifiers following a certain type of modal in comparison with the reading times of the same type of superlative modifiers following a different type of modal as an indication that the authoritative reading is unavailable for this combination.

As we turn to comparing the results of our study with the predictions of the different analyses, let us start with the processing costs, namely the reading times of the superlative modifier in the utterance, depending on which modal it followed. Recall that we found that two combinations, ◇ + at least and □ + at most, show significantly longer reading times in the utterance than the other two combinations. This is in line with Geurts and Nouwen’s (2007) analysis, who argue that □ + at least and ◇ + at most go together naturally in the sense that they are able to express an authoritative reading, while for ◇ + at least and □ + at most only the speaker insecurity reading is available. Under the analysis of Nouwen (2010), under which superlative modifiers go well together
with possibility modals, it is unexpected that \( \Diamond + \textit{at least} \) is harder to process. Under Neo-Gricean accounts of ignorance inferences of superlative modifiers (Büring, 2008; Schwarz, 2011, 2013; Kennedy, 2013), in contrast, we would expect that superlative modifiers preferably combine with necessity modals to yield an authoritative reading. This is at odds with our finding that \( \textit{at most} \) is harder to process when it combines with \( \square \) than when it combines with \( \Diamond \). Finally, Coppock and Brochhagen (2013) predict that for all of the combinations, both the authoritative and the speaker insecurity reading are available, which should make all combinations equally natural and comparable in processing, contrary to our results.

In order to compare the readings in terms of upper and lower bound of permissible values predicted by the various analyses with our findings, it is helpful to measure the results against the predicted readings summarized in (3)-(6) against our results in Table 1.\(^3\)

\begin{align}
(10) \quad & \square + \textit{at least} n:
\begin{align*}
a. & \quad \text{G&N, N, B/S/K, C&B} \\
b. & \quad \text{G&N, B/S/K, C&B} \\
c. & \quad 2.87\% < 94.56\%
\end{align*}
\end{align}

\begin{align}
(11) \quad & \Diamond + \textit{at least} n:
\begin{align*}
a. & \quad \text{G&N, N, C&B} \\
b. & \quad \text{G&N, B/S/K, C&B} \\
c. & \quad 23.36\% < 85.89\%
\end{align*}
\end{align}

\begin{align}
(12) \quad & \square + \textit{at most} n:
\begin{align*}
a. & \quad \text{B/S/K, C&B} \\
b. & \quad \text{G&N, N, B/S/K, C&B} \\
c. & \quad 89.83\% < 5.13\%
\end{align*}
\end{align}

\begin{align}
(13) \quad & \Diamond + \textit{at most} n:
\begin{align*}
a. & \quad \text{G&N, N, C&B} \\
b. & \quad \text{G&N, B/S/K, C&B} \\
c. & \quad 93.57\% < 1.67\%
\end{align*}
\end{align}

Starting with the combination \( \square + \textit{at least} \), for which we found a clear response pattern, it turns out that the predictions of all four analyses are compatible with our results. For \( \square + \textit{at least} \), greater values than \( n \) were accepted in 94.56\% of the cases, while lower values were rejected in 97.13\% of the cases cf. (10c), as expected under the authoritative reading (10a), which according to all of the analyses is predict to be available. The additional, weaker reading in (10b) may be available as well, but as discussed above, we hypothesize that when participants are faced with a choice between a strong and a weak reading, they will tend to choose the strong one.

Turning next to \( \Diamond + \textit{at most} \) in (13), we found that only smaller values than \( n \) were accepted. This result too is compatible with the predictions of all of the analyses. Although the Neo-Gricean approach (Büring, 2008; Schwarz, 2011, 2013; Kennedy, 2013) only predicts the speaker insecurity reading (13b) to be available for this combination, participants accepting lower values in the vast majority of the cases is compatible with this reading. As explained before, under this reading, the speaker is not sure whether values in the epistemic range are permissible or not. Since participants were forced to decide whether the number in the description sentence was in accordance with the utterance or not and didn’t have the option to hedge their response, we take the high rate for Yes responses to indicate that values in the epistemic range were interpreted as permitted values in spite of the possibility that the speaker may not be entirely certain whether these values are permitted.

The remaining two combinations, \( \Diamond + \textit{at least} \) and \( \square + \textit{at most} \) are more interesting, as here the different analyses make contrasting predictions.

For \( \Diamond + \textit{at least} \) we found that greater values than \( n \) were accepted in 85.89\% of the cases, but smaller values were still accepted in about a quarter of the cases. This pattern is not compatible with the predictions of any of the analyses. If the authoritative reading (11a) had been consistently available, as predicted by Geurts and Nouwen (2007) and Coppock and Brochhagen (2013), this should have been the dominant reading and lower values should have been rejected. But if only the speaker insecurity reading (11b) had been available, lower values should have been consistently accepted. So this mixed pattern suggests that in about a quarter of the cases, participants got the speaker insecurity reading, while in the majority of cases they got the authoritative reading.

For \( \square + \textit{at most} \), our results are compatible with the predictions of the pragmatic accounts, the Neo-Gricean approaches (Büring, 2008; Schwarz, 2011, 2013; Kennedy, 2013) on the one hand and Coppock and Brochhagen (2013) on the other. These analyses predict an authoritative reading where the modified numeral specifies the upper bound if the deontic range (12a). This is the reading participants preferred, as shown by the fact that lower values were accepted in about 90\% of the cases, higher values were rejected in the majority of the cases. If only the speaker insecurity reading (12b) were available, as predicted by Geurts and Nouwen (2007) and Nouwen (2010), lower values should have been consistently accepted.

So how can our data best be explained in light of the available theoretical analyses? Since there are two groups of superlative modifier-modal combinations differing in the processing costs they incur, the first conclusion is that two combinations, namely \( \square + \textit{at least} \) and \( \Diamond + \textit{at most} \), go together more naturally than the other two combinations (\( \Diamond + \textit{at least} \) and \( \square + \textit{at most} \)). If we take into account that our task may have bias participants towards authoritative readings, the lower processing costs we observed for \( \square + \textit{at least} \) and \( \Diamond + \textit{at least} \)
at most can be interpreted as indicating that the authoritative reading is available for these combinations, but not for the other two. This is in line with the analysis of Geurts and Nouwen (2007), but not any of the others.

We further hypothesize that the higher processing costs we observed for ◊ + at least and □ + at most might be due to some kind of repair strategy, to which participants resorted in order to derive authoritative readings for combinations for which these readings are not compositionally available. Since the authoritative reading wasn’t available for these superlative modifier-modal combinations, participants may have reanalyzed the modal in order to derive a reading, namely an authoritative reading, which would assist in making a clear-cut decision. In the case of ◊ + at least and □ + at most, the only possible compositional reading is one in which all values are potentially allowed, some within the deontic range and some within the epistemic range (that is, the speaker cannot rule out any values, modulo pragmatic restrictions involving relevance). This unrestricted reading may have been felt to be at odds with the speaker’s utterance, which included two expressions that normally communicate restriction, namely deontic modals and superlative modifiers. Participants might have therefore decided that the reading conveying speaker ignorance was not felicitous and opted to reanalyze the modal to arrive at an authoritative reading. This could also explain why we got mixed results for ◊ + at least: It seems that in the majority of cases, participants opted for the strong reading that necessitated modal reanalysis, but in the minority of cases, participants nevertheless opted for the compositional reading conveying speaker ignorance. This does not seem to carry over to the combination □ + at most, for which we also found higher processing costs but a more clear-cut pattern of preferred readings.

Acknowledgments

We thank audiences at the universities of Konstanz, Tübingen, and Chicago, ZAS Berlin, and at CUNY 2014. We have especially benefited from conversations with Oliver Bott, Irene Heim, Itamar Francez, Chris Kennedy, Manfred Krifka, Rick Nouwen, Fabian Schlotterbeck, and Maribel Romero. Ali Eglimez, Vasiliki Erotokritou, and Bianca Veser were very helpful with creating, editing and translating stimuli and running experiments. We also gratefully acknowledge funding from the DFG and the Zukunftskolleg at the University of Konstanz.

References


Kennedy, C. (2013). A “de-Fregean” semantics for modified and unmodified numerals. (University of Chicago)


Schwarz, B. (2011). Remarks on class B numeral modifiers. (Handout of a talk at the workshop Indefinites and Beyond, Universität Göttingen, November 19, 2011)

Between Intersective and ‘Split’ Interpretations of Predicate Conjunction: The Role of Typicality

Eva B. Poortman (e.b.poortman@uu.nl)
Utrecht Institute of Linguistics OTS, Utrecht University
Trans 10, 3512 JK, Utrecht, The Netherlands

Abstract
This paper aims to account for the different interpretations of plural sentences with conjoined predicates. Such sentences are sometimes interpreted strictly intersectively, sometimes strictly non-intersectively (or ‘split’), and sometimes both interpretations appear to be allowed. We propose that the logical interpretation of these sentences is systematically affected by typicality effects of the complex predicate in the sentence. We show with a set of experiments that a) the acceptability of a non-intersective interpretation can be expressed in terms of a continuum and b) each acceptability proportion is predicted by the typicality of the two conjoined predicates applying simultaneously. This way, we specify at least one of the relevant pragmatic considerations that determine the interpretation of a plural sentence with conjunctive predicates. Moreover, these results stress the importance of conceptual structure of predicates for compositional semantics.

Keywords: conjunction; plurality; plural predication; typicality effects; concepts; reasoning.

Introduction
Plural sentences with conjunctive predicates as in (1) and (2) are considered to be true if and only if every boy that is referred to is in the intersection of the two sets that are denoted by the conjoined verbs. In other words, sentence (1) is true iff each boy is both sitting and reading, and sentence (2) is true iff each boy is both waving and smiling.

(1) The boys are sitting and reading
(2) The boys are waving and smiling

We arrive at such interpretations by applying the well-known boolean analysis of conjunction, according to which it behaves as set-theoretic intersection (Keenan & Faltz, 1985; Partee & Rooth, 1983), and combining it with a distributivity operator (Link, 1983). Importantly, such an analysis assumes that the way we reason about these natural language sentences is independent of lexical information – thus assuming a clear division between compositional and lexical semantics. As a result, the logical interpretations of sentences like (3) and (4) are expected to be derived in a similar way as those of (1) and (2), with the difference between the sentences only being a matter of word meaning.

(3) The boys are sitting and standing
(4) The boys are sitting and cooking

In this paper, we report experiments that show that sentences (3) and (4) in fact receive weaker logical interpretations than sentences (1) and (2). Sentence (3) is generally interpreted such that a subset of the boys is sitting and the rest of the boys are standing – we will call this a ‘split’ interpretation (Heycock & Zamparelli, 2005). Sentence (4) also allows such a ‘split’ interpretation, but crucially to a lesser extent than sentence (3). Understanding such acceptability patterns calls for a systematic investigation of the lexical information in the sentence, as this appears to be inseparable from a proper analysis of conjunction. We show that there is in fact a continuum of acceptability values for non-intersective or ‘split’ interpretations of sentences with conjunctive predicates, and we account for this continuum with a principle that predicts how language users apply predicates to plural subjects based on typicality structure of the complex predicate.

Many previous works have described non-intersective interpretations of plural sentences (e.g. Križka, 1990; Heycock & Zamparelli, 2005; Winter, 2001). Winter (2001) tries to systematically account for the different interpretations, in which properties are distributed differently over individual entities. He does so by extending the Strongest Meaning Hypothesis (SMH) that was put forward by Dalrymple et al. (1998). The SMH aims to resolve ambiguity that is caused by lexical and other contextual information, specifically in the area of reciprocals. Dalrymple et al. claim that the varying logical interpretations of structurally similar reciprocal sentences as in (5) and (6) are captured by their principle.

(5) The boys know each other
(6) The boys are following each other

For each occurrence of the reciprocal, this principle selects as its interpretation the strongest meaning (from an inventory of six possible meanings) that is consistent with context. For example, if we assume three boys, then sentence (5) receives a strong interpretation in which every boy knows every other boy, since there are no contextual restrictions on the amount of possible “knowing-relations”. All weaker meanings are consequently disallowed for (5). By contrast, sentence (6) most likely means something weaker than every boy following every other boy.

---

1 Throughout the paper, when using the term “interpretation” we simply refer to the situations that support a truthful usage of a sentence.
According to the SMH, we weaken the meaning of the sentence as far as context pushes us to. For this example, that meaning is most likely one where boy 1 follows boy 2, and boy 2 follows boy 3. This is the strongest candidate meaning that does not contradict our knowledge about following people.

Winter (2001) proposes that a maximality principle like the SMH is not construction-specific to plural sentences with reciprocals. He rephrases the SMH into a general principle of plural predication, such that any complex plural predicate with a meaning that is derived from one or more singular predicates using universal quantification is interpreted using the logically strongest truth conditions that are not contradicted by known properties of the singular predicate(s) (Winter, 2001). The contrast between minimal pairs like (1) and (3) is then captured in the following way. Again, the SMH selects the logically strongest possible candidate meaning for each sentence. When a strong interpretation (intersective conjunction) is consistent with properties of the predicates, then this is the attested meaning of the sentence – an example is sentence (1). On the other hand, when such a strong interpretation is inconsistent with these properties, the interpretation is weakened. We see this in sentence (3): An intersective interpretation in which all boys are in the intersection of the set of sitting individuals and the set of standing individuals contradicts what we know about ‘sitting’ and ‘standing’. Thus, sentence (3) receives a ‘split’ interpretation, which is the strongest interpretation that does not contradict this knowledge.

In the current paper, we argue that the predictions made by the SMH can be too strong. Consider sentences (7) and (8), which are of a similar nature to sentence (4) above.

(7) The men are lying down and drinking
(8) The men are waving and drawing

If Winter’s extended SMH is correct in assuming that non- intersective interpretations are only available when intersective interpretations are strictly ruled out by the predicates, then these sentences would only allow an intersective interpretation. As we will show in the current paper, non-intersective, ‘split’ interpretations are readily available to many speakers for sentences like (4), (7) and (8), even though the predicates do not strictly exclude an intersective one. For example, it may be exceptional, but it is possible for a person to sit and cook simultaneously. Several previous works have recognized a similar problem for the SMH concerning reciprocal sentences that receive weaker interpretations than predicted (e.g. Winter 2001; Philip, 2000; Kerem, Friedmann & Winter, 2009; Struiksma et al., submitted). Struiksma et al. (submitted), for example, showed that a sentence like The boys are pinching each other in the case of three boys is judged as true in a situation where each boy pinches only one other boy, despite the fact that a stronger interpretation is not excluded by properties of the predicate pinch.

These examples point to a fundamental issue with the proposal at hand. The SMH, both in its original and extended form, assumes that the interpretation of these sentences is only sensitive to so called “definitional” aspects of the meaning of predicate concepts. In other words, it only takes into account whether particular denotations of predicates are possible or impossible, i.e. whether they are an instance of that predicate concept or not. In the case of predicate conjunction, that means that the hypothesis only looks at whether intersective conjunction is possible or not, given the predicates at hand. Such sharp distinctions appear to be insufficient in accounting for the interpretation patterns that we observe. Alternatively, one can take into account typicality effects in categorization. The notion of typicality refers to the phenomenon that human subjects are able to grade different instances of a concept with respect to their representativeness of a given category. Since the 1970’s, a range of psychological studies has shown for one-place predicates that subjects consistently rank some instances of a concept as more typical than others, and that such rankings correlate with other measures of typicality such as categorization speed and error rate (e.g. Rosch, 1973; Smith, Shoben & Rips, 1974; Rosch & Mervis, 1975).

Taking into account typicality effects on reasoning was first proposed as a solution for reciprocal sentences, in the shape of the Maximal Typicality Hypothesis (Kerem et al., 2009; Struiksma et al., submitted). This hypothesis assumes that typicality effects also exist for binary predicate concepts (like pinch), and that these systematically affect the logical interpretation of the reciprocal expression that they combine with. In this paper, we extend the same logic to plural sentences with predicate conjunction. We claim that typicality effects also affect interpretation in plural sentences where two predicate concepts are conjoined. The proposal works as follows. For a complex concept like lying down and drinking, a situation where one is both lying down and drinking at the same time is an instance of that concept, but most likely an atypical one, at least to a certain extent. Consequently, when such a complex predicate combines with a plural, this affects the degree to which language users diverge from an intersective interpretation. Crucially, we predict that both the notions of typicality and acceptability can be expressed in terms of a continuum – allowing for more subtle distinctions than the SMH.

To summarize, our proposal predicts the following: a) that there is a continuum of typicality effects for complex predicates made up of two conjoined singular predicates, b) that there is a continuum of acceptability values for a particular interpretation of a plural sentence with those complex predicates and c) that the values on both continuums correlate – indicating that typicality for concepts in isolation systematically affects interpretation of sentences containing those concepts. We conducted two behavioral experiments and a correlation analysis to test these predictions.
Experimental Investigation

This section reports on pretests, two experiments and a correlation analysis. Experiment 1 checked the acceptability of plural predicate conjunction sentences of the form *The A are X and Y* (where $A$ is a plural noun and $X$ and $Y$ are predicates) in a non-intersective, ‘split’ situation. Experiment 2 measured typicality effects for complex predicate concepts in isolation. Materials for the experiments were constructed based on pretests that were conducted in order to include a wide range of typicality values in the actual experiments.

Pretests: Constructing Materials

The aim of the first pretest was to very informally gather as many Dutch verb combinations $X$ and $Y$ as possible, especially atypical ones. We provided 8 subjects with sets of two pairs in which $X$ was identical: one very natural pair, and one pair that is physically impossible to apply simultaneously, e.g. *sitting and reading* and *sitting and standing*. We then asked them to provide as many verbs that they could come up with that combine with $X$ (i.e. *sitting in this case*) that led to a possible but atypical, uncommon or strange combination. These pairs, combined with more natural pairs that we came up with, led to a list of 91 verb combinations in total.

In the second pretest, we had 29 different subjects rate these 91 pairs for compatibility. For each pair, subjects were asked to rate how odd\(^2\) they would consider it if both verbs applied to one person at the same time. Oddness was rated on a 6-point scale, where 1 meant “not odd at all” and 6 meant “physically impossible”. We mentioned explicitly that 5 thus meant “very odd, but physically possible”, in order to distinguish large atypicality from impossibility.

Results of this pretest showed great variability in ratings between verb pairs. We first selected 12 sets of verb pairs with identical $X$ that showed greatest variability within the set. From each of these 12 sets, we selected three verb pairs for the experiments: the verb pair that was rated lowest on the oddness scale (compatible pairs like *sitting and reading*), the verb pair that was rated highest (incompatible ones like *sitting and standing*), and a verb pair that was rated in between, at a mean of 4 points\(^3\) (atypical pairs like *sitting and cooking*). An overview of the 36 verb pairs that constituted the final material (translated from Dutch) is given in table 1. Creating the three groups (with labels ‘compatible’, ‘incompatible’ and ‘atypical’) was done purely to ensure variability, and for the sake of clarity we will refer to the three groups when discussing set-up and results of experiments 1 and 2. Note however that they are not meaningful in the final correlation analysis of all data points.

Experiment 1: Interpretation of Plural Predicate Conjunction Sentences

This experiment checked the acceptability of 36 plural sentences with two conjoined verbs in a ‘split’ situation. Each sentence was of the form *The A are X and Y* (where $A$ is a plural noun and $X$ and $Y$ are verbal predicates).

Participants A total of 33 students from Utrecht University (28 female, age $M = 21$) participated for monetary compensation. All participants were native speakers of Dutch without dyslexia. Prior to the experiment all participants signed an informed consent form.

Materials The material consisted of two versions of a truth-value judgment task, each containing 18 unique test items plus 18 filler items that were the same across versions. Each test item contained a plural predicate conjunction sentence in Dutch (*The A are X and Y*) and a drawing depicting four individuals in a non-intersective, ‘split’ interpretation of that sentence: predicate $X$ applied only to persons 1 and 2, predicate $Y$ applied only to persons 3 and 4. Half of the pictures depicted male individuals, and the other half depicted female individuals. An example of a test item drawing is given in figure 1.

![Figure 1: Example of a test item drawing of experiment 1.](image)

In each version of the experiment, one third of the test items contained sentences with verb pairs that were considered compatible $X$ and $Y$ in the second pretest (e.g. *The men are sitting and reading*), one third contained sentences with verb pairs that were considered incompatible $X$ and $Y$ (e.g. *The men are sitting and standing*) and one third contained sentences with pairs that were considered atypical $X$ and $Y$ (e.g. *The men are sitting and cooking*). The same drawings

---

\(^2\) Phrasing the question negatively by asking “how odd” subjects would rate a situation was done because directly asking for “how typical” they would judge a situation turned out to be ambiguous in Dutch. Some subjects interpreted the word *typical* to mean “atypical”, whereas asking for oddness is unambiguous.

\(^3\) Additional inclusion criteria included that each verb should be expressed by one word only, ratings for verb pairs should have smallest variation, and atypical verb pairs should have no “6” point ratings.

\(^4\) We measured acceptability of the sentence given a ‘split’ situation because sentences with incompatible pairs cannot be depicted any other way, and we wished to keep all factors in the comparison between pairs equal.
were used for sentences with compatible and incompatible pairs with identical X (e.g. The men are sitting and standing and The men are sitting and reading). To ensure that subjects never saw the same drawing twice (such as the one in figure 1), one of these sentences occurred in version 1 and the other occurred in version 2. The atypical items were divided over the two versions, resulting in two experiments with 6 sentences with compatible pairs, 6 sentences with incompatible pairs and 6 sentences with atypical pairs each, accompanied by 18 unique drawings. The items of each version are represented by light and dark cells in table 1.

Filler items contained similar drawings with four people, but a different type of accompanying sentence. The accompanying sentences in the filler items were either sentences with quantifiers (Some boys are X) or sentences mentioning specific individuals in the picture (Boys A, B and C are X). Half of the filler items were expected to be judged true, and half of them were expected to be judged false. Both versions of the experiment contained the same filler items.

The order of items was pseudo-randomized using Mix software (Van Casteren & Davis, 2006), with the following restrictions: items containing the same verb were at least six items apart; there were at most two test items immediately following each other, and at most two filler items immediately following each other; similar test items (in terms of compatible/incompatible/atypical) or similar filler items (in terms of quantifier/specific individuals) never immediately followed each other. Finally, we constructed two orders of each version, with the second one having reversed order of items.

Procedure Each participant completed one version of the experiment. The task was presented in a sound-proof booth on a PC using Open Sesame software (Mathôt, Schreij & Theeuwes, 2012). Prior to entering the sound-proof booth, each participant received verbal instructions explaining the experimental set-up. Further, more detailed instructions were given on the PC monitor.

After being instructed, each participant completed three practice trials. Subsequently, they were given the opportunity to ask for clarifications, if necessary. All verbs used in the practice session did not appear in the actual experiment. The experiment itself consisted of the 36 items described above. Drawing and sentence were presented in the center of a white screen. Participants were instructed to indicate as soon as possible whether they judged the sentence to be true or false given the situation in the drawing by pressing the left or right button with their dominant hand.

Coding and analysis Responses were coded “1” when participants judged a sentence to be true for a given drawing, and “0” when they judged a sentence to be false. This way we computed the proportion of acceptability of a sentence for a given drawing. We performed a repeated measures ANOVA with Compatibility as the within-subjects factor (with 3 levels: compatible, atypical, and incompatible). Post-hoc Bonferroni corrected multiple comparisons were performed in order to analyze differences between different Compatibility levels in detail.

Results Table 2 provides an overview of the data. It shows the mean acceptability of sentences (in proportions) for the three levels of Compatibility that were tested, for all versions taken together. Overall, mean acceptability per verb pair ranged from .24 to 1.

Table 2: Mean acceptability ratings (in proportions).

<table>
<thead>
<tr>
<th>Compatibility type</th>
<th>Mean acceptability (st. dev.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compatible</td>
<td>.54 (.31)</td>
</tr>
<tr>
<td>Atypical</td>
<td>.78 (.31)</td>
</tr>
<tr>
<td>Incompatible</td>
<td>.84 (.32)</td>
</tr>
</tbody>
</table>

5 A repeated measures ANOVA with Version as between-subjects factor was also performed, but showed no effect of Version ($F (3, 29) = .47, p = .71$). We thus collapsed the versions for the analysis.
A repeated measures ANOVA revealed that there was a main effect of Compatibility ($F(1,36, 43.49) = 37.41, p < .001$). This means that the mean proportions of acceptability for the three Compatibility levels are not equal. Pairwise comparisons show that all three levels differ significantly from each other in acceptability: the acceptability of sentences with compatible predicates differs from the acceptability of sentences with incompatible predicates ($p < .001$); the acceptability of sentences with compatible predicates differs from the acceptability of sentences with atypical predicates ($p < .001$); and the acceptability of sentences with incompatible predicates differs from the acceptability of sentences with atypical predicates ($p < .05$).

**Experiment 2: Typicality Effects for Predicate Pairs**

This experiment checked typicality effects for the 36 complex concepts that were used in sentences of experiment 1. We measured the typicality of one particular instance of each complex predicate, namely the one in which both predicates apply simultaneously to one individual.

**Participants** The same 33 students from Utrecht University from experiment 1 participated in this experiment. Each subject completed the interpretation experiment first, before proceeding with the typicality experiment. Also, in between experiments they took part in a third, unrelated experiment.

**Materials** The materials consisted of a questionnaire containing 36 statements about one person involved in two actions simultaneously. Half of the statements were about males and half of them were about females (matching the gender of persons in the pictures of experiment 1). Each statement contained a singular object (a male or a female) and two conjoined predicates (e.g. *The man is sitting and reading*). The 36 pairs of verbs were the same as the ones used in sentences of experiment 1, thus one third of the pairs were considered compatible in the second pretest (e.g. *sitting and reading*), one third were considered incompatible (e.g. *sitting and standing*), and one third were considered atypical (e.g. *sitting and cooking*).

The order of items was pseudo-randomized using Mix software (Van Casteren & Davis, 2006), with the restriction that at most two items of the same type (in terms of compatible/incompatible/atypical) immediately followed each other.

Finally, we constructed four different orders of the questionnaire: two versions that started with the statements about males (with the second one having reversed order within males and females statements), and two versions that started with the statements about females (with the second one having reversed order within males and females statements).

**Procedure** Each participant received one of the questionnaires on paper, in a sound-proof booth. They were instructed to rate how odd they would consider it if both verbs applied to the given person at the same time. Oddness was rated on a 6-point scale, where 1 meant “not odd at all” and 6 meant “physically impossible”. We mentioned explicitly that 5 thus meant “very odd, but physically possible”, in order to distinguish large atypicality from impossibility.

**Coding and analysis** Responses were coded “1” through “6” corresponding to the participant’s judgment. This way we computed the atypicality rating of two verb pairs applying simultaneously. We performed a repeated measures ANOVA with Compatibility as the within-subjects factor (with 3 levels: compatible, atypical, and incompatible). Post-hoc Bonferroni corrected multiple comparisons were performed in order to analyze differences between different Compatibility levels in detail.

**Results** Table 3 provides an overview of the data. It shows the mean atypicality rating for the three levels of Compatibility that were tested, for all versions taken together. Overall, mean ratings per verb pair ranged from 1.03 to 5.94.

<table>
<thead>
<tr>
<th>Compatibility type</th>
<th>Mean atypicality (st. dev.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compatible</td>
<td>1.35 (.29)</td>
</tr>
<tr>
<td>Atypical</td>
<td>3.82 (.58)</td>
</tr>
<tr>
<td>Incompatible</td>
<td>5.66 (.38)</td>
</tr>
</tbody>
</table>

A repeated measures ANOVA revealed that there was again a main effect of Compatibility ($F(1.95, 62.45) = 1187.02, p < .001$). This means that the mean atypicality ratings for the three Compatibility levels are not equal. Pairwise comparisons show that all three levels differ significantly from each other: the atypicality of supposed compatible pairs applying simultaneously differs from the atypicality of supposed incompatible pairs applying simultaneously ($p < .001$); similarly for the atypicality of compatible vs. atypical pairs ($p < .001$); and the atypicality of incompatible vs. atypical pairs ($p < .001$). This means that the three groups that we selected based on the pretest were confirmed in the typicality experiment (with different subjects and a subset of the stimuli).

**Correlation Between Interpretation and Typicality**

The crucial test for our proposal is the relationship between interpretation and typicality. In order to account for the degree to which non-intersective interpretations of sentences are available given two particular conjoined predicates, we

---

6 Again, phrasing the question negatively by asking “how odd” subjects would rate a situation was done because directly asking for “how typical” they would judge a situation turned out to be ambiguous in Dutch. Some subjects interpreted the word *typical* to mean “atypical”, whereas asking for oddness is unambiguous.
need to check whether this correlates with the degree to which those two predicates applying simultaneously is atypical for the concept. In order to check this, we performed a one-sided correlation analysis between all the results of experiment 1 and those of experiment 2 (figure 2). The result was a positive correlation between mean proportion acceptability of a sentence in a non-intersective interpretation and mean atypicality rating of a predicate pair applying simultaneously ($r = .66, n = 36, p < .001$).

![Figure 2: Relation between mean acceptability (in proportions) of sentences in a ‘split’ situation and mean atypicality rating of predicate pairs applying simultaneously.]

Discussion

This paper reports on an experimental investigation into the interpretation of plural sentences with predicate conjunction, and its connection to typicality. We proposed that the extent to which non-intersective interpretations are available directly correlates with the atypicality of a situation where the two predicates apply simultaneously. Experiment 1 revealed a continuum of acceptability values of 36 sentences in a non-intersective, ‘split’ situation, ranging from 24% to 100% acceptable. Such a continuum is unexpected under the extended SMH by Winter (2001), which assumes that any given sentence is either true or false in a particular situation, depending on what the context allows. Next, experiment 2 showed that different typicality effects exist with respect to complex predicate concepts that are composed of two singular predicates – similar to the effects that were found repeatedly for one-place predicates (e.g. Rosch, 1973). Specifically, we measured the typicality of a situation in which both predicates apply simultaneously, for a given complex concept. The typicality ratings for 36 pairs ranged over the entire 6-point scale. We proposed that typicality relates to acceptability in such a way that the less typical the situation in experiment 2 is judged to be, the more a non-intersective interpretation is available. Based on a correlation analysis, we can conclude that this prediction was borne out. We take these results to be an indication of conceptual structure of predicates playing a crucial role in sentence interpretation, in line with similar results on reciprocal sentences (Struiksma et al., submitted).

An important next step is to further explore typicality effects for complex predicates. In the current paper, we report an experiment that used one particular typicality measurement with one particular dependent measure, namely the typicality of two simultaneous actions, rated on a scale. One can imagine that in fact the typicality of the opposite situation, i.e. two predicates applying to two separate individuals, or perhaps sequentially to one individual, might also affect the interpretation of a plural sentence with those predicates. Also, it will be good to correlate rating measures with different kinds of dependent measures such as categorization speed or error rate to have a more robust result – similar to the investigations into typicality effects for nouns. However, the fact that even one measure can distinguish different types of verb pairs so clearly, is a promising starting point for this enterprise.

Another related issue is the deeper question of how typicality effects come about: What exactly makes a particular instance of a concept typical? A potential candidate factor is that typicality is formed by prior experiences or likelihood of a situation. An anonymous reviewer, however, pointed out an example like (9).

(9) The boys are unicycling and juggling

The reviewer claims that despite the fact that we probably rarely see a person simultaneously unicycling and juggling, we still interpret the conjunction in sentence (9) intersectively. Such an example points out that typicality is a complex matter that needs to be studied further. The question of what makes something typical does not affect the results described in this paper per se, but knowing what affects typicality would give them more explanatory power, as pointed out by this reviewer.

Another logical next step would be to investigate other cases in language where typicality affects reasoning with logical operators. So far we have seen that understanding both reciprocal sentences and the sentences with conjunction that were investigated in the current paper, is inseparable from the study of concepts. It is highly likely that these are not the only areas in which this is the case.

Conclusion

This paper started from the observation that plural sentences with conjunctive predicates do not always receive the same logical interpretations. Previous work on reciprocal sentences has already taught us that lexical information can influence sentence meaning in systematic ways (e.g. Dalrymple et al., 1998; Kerem et al., 2009; Struiksma et al., submitted). Here we reported on experimental investigation of plural sentences with predicate conjunction, that provided insight into specifically the role of typicality information of predicate concepts. With this result, we add to the line of work that investigates the interface between lexical and compositional semantics, and lead the way towards directions for further research in this area.
Acknowledgments

This work was supported by NWO (The Netherlands Organisation for Scientific Research) VICI grant number 277-80-002, “Between Logic and Common Sense: the Formal Semantics of Words” (2010-15). I am grateful to Hanna de Vries, Marijn Struiksma, Martin Everaert, Choonkyu Lee, Sonia Kasyanenko, Chris Barker, two anonymous reviewers of the “Formal and Experimental Pragmatics” workshop (Tübingen, 2014), and the audiences at the Concept Composition workshop (Utrecht, 2013) and the NYU Semantics Group (New York, 2014) for useful remarks and questions. Special thanks to Yoad Winter for extensive discussions throughout the course of this research. Finally, I am thankful to Nir Kerem for initial ideas that eventually led to the study reported in this paper. The illustration in figure 1 was made by Ruth Noy Shapira.

References


Markedness, Frequency, and Lexical Change in Unstable Environments

Gerhard Schaden (gerhard.schaden@univ-lille3.fr)
Université Lille 3 & CNRS UMR 8163 STL
Domaine Universitaire du Pont de Bois, 59653 Villeneuve d’Ascq

Abstract
This paper proposes an account of homonymy reversals, based on Polya-urn processes. It assumes that words need to be conceptually grounded, and that the presence or absence of a class of referents may destabilize established meaning differences, as well as cause new meanings to emerge.

The emergence of new meanings is conditioned on an asymmetry of frequency between two words, and requires agents with a Theory of Mind. It will be also argued that markedness cannot be explained by frequency alone.

Keywords: Polya-Urn Processes; Reinforcement Learning; Hyponymy-Reversals; Theory of Mind; Markedness.

Introduction
The lexicon of a language is a relatively unstable domain. Often, words shift their meanings along patterns of polysemy, and what once was a derived meaning becomes the basic meaning of a word. For instance, Modern German Bein means ‘leg’, but in earlier periods, its meaning was just like its English cognate ‘bone’.

A particularly puzzling phenomenon in this domain is reversals of hyponymy, as exemplified by English hound vs. dog. In Contemporary English, dog is the hyperonym, referring to the totality of canis lupus familiaris, but in Middle English and earlier, dogge used to be a hyponym of hound, which was then the term denoting canis lupus familiaris. Hound is the cognate of German Hund, which is also the form that appears in all other major Germanic languages, and directly derives from Proto-Indo-European *kwen-/*kun-, meaning ‘dog’ (see Meier-Brügger, Fritz, Krahe, & Mayrhofer, 2002, p. 131). Therefore, English hound and dog have undergone a reversal of hyponymy between the Middle English and the Early Modern English periods.

Crucially, dog and hound did not just swap their respective meanings; a dog used to be a particular, sturdy kind of canis lupus familiaris (see Gąsiorowski (2006, 278) for Medieval textual evidence on this), whereas a hound is today a type of canis lupus familiaris with good sense of smell destined to be used for hunting.

The question is whether such reversals require first a change in lexical meaning such that they become synonyms, and subsequently redifferentiate, or whether the underlying process is different. I will explore the latter option, and I will argue that pragmatics (i.e., ascribing to an agent the intention to communicate a given meaning with a given form) plays at least in some circumstances a crucial role in the differentiation of meaning.

The general aim of this paper is to investigate a subtype of lexical changes in an unstable environment. More precisely, I will investigate the learning of grounded concepts (see, e.g., Steels, 2008) using Polya-urn processes (see, e.g., Skyrms, 2010) in a context where there is nothing to learn.

A second aim is to investigate the relationship between markedness and frequency. Haspelmath (2006, p. 44) writes that “[... ] since frequency of use seems to explain most of the observed phenomena, we do not need a ‘markedness’ concept to understand them.” While I am sympathetic to his general line of argumentation, which tries to show that ‘markedness’ in linguistics is not a unified category, and aims to substitute more tangible notions for markedness, I think that Haspelmath has gone too far. The basic theoretical claim I will defend in this paper is that one cannot get markedness out of pure frequencies, and that an additional ingredient is required:

(1) Markedness = frequency of use + pragmatics

Meaning Inference and Reinforcement Learning of Grounded Concepts
I assume following Steels (2008) that words — at least ‘simple’ words that have obvious and tangible referents like dog or hound — are conceptually grounded, that is, that they are related to their referents in the surrounding environment. In lexical acquisition, a learner’s task is to figure out which kind of entity in his environment a given word refers to. (Lexical) meaning is therefore dependant on the (part of the) world speakers live in.

Support of Meaning Differences and Richness of Environments
When semanticists reason about meaning, they do so generally in abstraction from any particular context, and their terminology reflects this bias. In order to refer unambiguously to the relation between grounded concepts and the environments they occur in — without necessarily making a commitment to the general meaning of a grounded concept —, I will help myself by introducing some definitions.

First of all, (2) introduces a basic relation between meaning and context:

This of course is hardly news; the whole “Wörter und Sachen” movement (see, e.g. Schuchardt, 1912) is based on this idea.
A denotational difference between two words \( a \) and \( b \) is supported in a given situation \( s \) and at a moment \( t \) iff their denotations are not identical (i.e., \([a]^{s,t} \neq [b]^{s,t}\)). If (2) holds, I will also say that the situation \( s \) supports a denotational difference between \( a \) and \( b \), or that \( s \) is sufficiently rich to support a denotational difference between \( a \) and \( b \). In a sense, this is just the inference-side of the familiar phenomenon of interpretation with respect to a model.\(^6\)

Furthermore, if a denotational difference between words \( a \) and \( b \) is not supported in a given situation \( s \), I will say that words \( a \) and \( b \) are denotation-equivalents with respect to \( s \). This notion needs to be distinguished from the concept of denotational synonymy (see, e.g. Cruse, 2000). Denotational synonyms will be denotation-equivalents for any possible context. However, the reverse is not true: denotation-equivalent with respect to a given situation \( s \) need not be denotational synonyms, because there may be a meaning difference (encoded in the mind of speakers) that is supported only in a larger context or situation. Thus, the fact that two words are denotation-equivalent in a given situation does not commit us to a position in which they have the same meaning.

Let us look at an example: my Parisian flat does not support a denotational difference between \( \text{cat} \) and \( \text{pet} \), since the only pet present there happens to be a cat.\(^7\) So, \( \text{pet} \) and \( \text{cat} \) are denotation-equivalents with respect to my flat. However, since all members of the household can access the surrounding environment, where a denotational difference between \( \text{cat} \) and \( \text{pet} \) is supported, my flat is clearly not a linguistically pertinent environment with respect to a possible change in meaning of \( \text{cat} \) and \( \text{pet} \). Yet, in the long run, if in a given region and for several generations a denotational distinction between words remains unsupported, this could cause a confusion or a shift in meaning between the initially distinct words. Therefore, denotation-equivalents might transform into denotational synonyms, since (human) language learners are situated agents, learning conceptually grounded concepts in the environment they happen to be born into.

At this point, I would like to emphasize that the environments we currently live in are probably not very representative of environments in human history.\(^8\) Currently, the semiotic landscape of language-acquiring children (at least in urban areas in the first world) is potentially rich and, importantly, locally determined to a rather small degree. While the average Central or West European today knows zebras, giraffes, penguins and elephants (none of which live natively in the local environment), he would probably have trouble naming a substantive proportion of the animals and plants living in a three-kilometer radius of his home. In traditional societies, people are much more dependent in their survival on a good knowledge of local fauna and flora than we are (cf., e.g. Diamond, 2013).

Therefore, modern societies and the natural languages spoken in them are probably less affected by changes in their local, natural environment than (subgroups of) premodern societies. Generally, the more closed and small-scale a linguistic community is, the more likely it is that chance fluctuations in the environment will have an impact on their language’s lexicon.

**Reinforcement Learning with Polya-Urn Processes**

The tool chosen to investigate lexical meaning change is Polya-urn processes, which will be used to represent learning processes relating words to their meanings.

The basic principle of a Polya-urn is that a ball is drawn at random from the urn, and — if the ball corresponds to an appropriate answer — not only will the ball be put back into the urn, but a given quantity of the same type of ball is added to the urn (the reward). In this way, the probability of providing an appropriate answer in the next turn will increase. From a behaviorist point of view, one could say that the appropriate answer has been reinforced. Such a reinforcement process is illustrated in (3):

(3) a. \( \begin{array}{c|c|c}
\text{URN}_t \\
\hline
\text{white:1} & \text{red:1} \\
\hline
\end{array} \)

b. \( \begin{array}{c|c|c}
\text{URN}_{t+1} \\
\hline
\text{white:2} & \text{red:1} \\
\hline
\end{array} \)

Let us assume that in (3a), the white ball has been drawn from the urn, and that this is the correct answer. As a consequence, \( \text{white} \) has been reinforced in (3b) by adding an additional white ball to the urn. The addition of a ball increases the probability of drawing a white ball from 0.5 in (3a) to 0.6 in (3b).

If there is only one correct answer, and if no errors occur in reinforcing, the probability of drawing the correct response from the urn at random will approach 1 in the limit. The rate of the increase in the probability of the correct answer depends on the initial inclination weights (that is, the quantity of balls present in the urn in the beginning) and the weight of the reward. The higher the reward with respect to the initial endowment of the urn, the faster the increase in the probability of giving a correct answer. This dependence is illustrated in figure 1.

**Simulating Meaning Change With Polya-Urn Processes**

For the simulation, I assume two competing words which are initially denotation-equivalents. Therefore, in a signalling game, both words are systematically appropriate answers, and there is in principle nothing to learn. However, the reinforcement process will continue, and modify the weights.

Additionally, I assume that each word has the same number of weighted submeanings — for instance, something like a *qualia-structure* in the Generative Lexicon, as argued for by

---

\(^6\)The idea of support for a given meaning difference could also be extended to diaphasic or diatopic variation, for instance.

\(^7\)In fact, \([\text{cat}]^{\text{flat(ges)}}.11/05/2014 = [\text{pet}]^{\text{flat(ges)}}.11/05/2014 = \text{Akané}.\)

\(^8\)Kusters (2003, 2008) makes the same point. He showed that differences in community-structure can even affect morphology, and that what one usually thinks of as external factors may have impact on the core of grammar.
Pustejovsky (1995). Remember from the discussion in the introduction that a dog used to be an especially sturdy kind of *canis lupus familiaris*, and that *hounds* today are breeds of *canis lupus familiaris* whose basic purpose is hunting. It happens that in Pustejovsky’s qualia-structure, the difference between Middle English *hound* and *dog* can be understood as a specification of *dog* in the *formal*-quale (i.e., shape) which *hound* lacked; and the difference between Modern English *dog* and *hound* can be seen as a specification of *hound* in the *telic*-quale (i.e., purpose or function) which is absent for *dog*.

Dispensing with a monolithic representation for a word may seem to be a costly and unnecessary complication, but it will turn out to be extremely convenient for the simulation, and there are other factors that favor it. First of all, the fact of having independent submeanings allows us to bridge the difference between lexical and encyclopedic knowledge, which may not be all that useful when working with grounded concepts. Second, linguistically, having meanings that are more structured facilitates dealing with certain issues of compositionality.\(^9\) Finally, there is some evidence from functional dissociation in brain-damaged patients that the parameters of shape and function (these are the crucial ingredients distinguishing *dog* from *hound*) are treated by different regions of the brain (as reported by Bermúdez, 2005, p. 20).

Technically, each word is modeled as a structure (or an object) where the initial inclination-weights \(n\) are the same for each submeaning, and where each submeaning has exactly one type (for instance, regarding the *telic*-quale of the generative lexicon there is exactly one function for the object). All weight is awarded to some submeaning; there is no global weight. At the beginning, we thus have the situation sketched in table 1:

<table>
<thead>
<tr>
<th>Submeaning</th>
<th>Word1</th>
<th>Word2</th>
</tr>
</thead>
<tbody>
<tr>
<td>submeaning1</td>
<td>type-1: (n)</td>
<td>type-1: (n)</td>
</tr>
<tr>
<td>submeaning2</td>
<td>type-1: (n)</td>
<td>type-1: (n)</td>
</tr>
<tr>
<td>submeaning3</td>
<td>type-1: (n)</td>
<td>type-1: (n)</td>
</tr>
<tr>
<td>submeaning4</td>
<td>type-1: (n)</td>
<td>type-1: (n)</td>
</tr>
</tbody>
</table>

Nature chooses at random one submeaning, and the speaker\(^{10}\) chooses one word according to its relative weight for that submeaning. For instance, if Word1 has 10 balls at submeaning1, and Word2 5 balls, and if Nature chooses submeaning1, then the speaker will choose Word1 with probability \(\frac{10}{15}\), and Word2 with probability \(\frac{5}{15}\). Since according to our assumption the words are denotation-equivalents, whatever word is drawn will be reinforced according to some preestablished reward. Therefore, we have 4 independent Polya-urns (one for each submeaning), each containing balls according to the weights of each word at the corresponding submeaning.

Figure 2 illustrates a sample outcome of the reinforcement of the two words. It represents the outcome of 4000 rounds of reinforcement, where the initial inclination weight was 1, and the reinforcement reward also corresponded to 1. One can see that in this particular case, for each submeaning (referred to as “Quale” in figure 2), one word has been reinforced more often than the other.

As illustrated by the leftmost boxplot in figure 3, the outcome in figure 2 is rather typical for this kind of initial inclination weight and reward: the median difference corresponds roughly to 500.\(^{11}\) This means that, for approximately 1000 iterations per submeaning, we will obtain as a median a weight of 750 for one word, and 250 for the other.\(^{12}\) However, there are also cases where the difference is close to 0, that is, where both words have nearly the same weight for a submeaning, and cases where the difference is close to 1000, that is, where one word has (nearly) always been reinforced for a given submeaning.

Figure 3 as a whole shows however that this result depends crucially on the relationship between initial inclination weight and the reinforcement reward: the higher the incli-
exploitation notion (cf. again Haspelmath, 2006), it normally involves the
ful sense. While markedness is a tricky (and often unclear)
case, we do not have semantic markedness in any meaning-
one understands by ‘markedness’. I would argue that in our
unmarked variety for submeaning 3?

alternative for submeanings 1, 2 and 4, whereas Word2 is the
ative, whereas Word1 is the unmarked one? Or would we have
to be more prudent, and say that Word1 is the unmarked al-
tation weight, the lower the absolute difference between the
corresponding submeanings. Whereas the median difference
between the weights of corresponding submeanings with an
inclination weight of 1 amounts to 495.5, the same median
is down at 98 with initial inclination weight of 25, and at 72
with an initial inclination weight of 50.

Let me summarize our findings so far: given low initial in-
clination weights, the assumed circumstances — that is, rein-
forcement learning in contexts where there is nothing to learn
— will produce as a rule strongly unbalanced weights in a
given submeaning. Therefore, we have shown that at least un-
der some circumstances, a purely stochastic process is able to
provide differing frequencies of use for two words, in some-
thing that one can see as a very primitive kind of text.

Coming back to the quote from Haspelmath (2006), equating
textual frequency with markedness, does this mean that we have derived something like markedness in our simula-
tion? For instance, in our sample outcome depicted in figure
2, could we say that Word2 has become the marked alterna-
tive, whereas Word1 is the unmarked one? Or would we have
to be more prudent, and say that Word1 is the unmarked al-
tative for submeanings 1, 2 and 4, whereas Word2 is the
unmarked variety for submeaning 3?

The answers to these questions depend on what exactly
specializations in meaning. What we arrived at in the sim-
ulation is simply a difference in frequency with respect to
given submeanings. For instance, for signalling submeaning
Quale1, an agent would choose Word1 with a probability of
roughly 0.8, and Word2 with a probability of roughly 0.2. But
in the present setup — which is purely stochastic —, both
words continue to have the same lexical meaning, and I do
not think that there is any basis for claiming that either word
has acquired any specialization.

That being said, let us now have a look at factors that might
cause a differentiation in meaning.

A Change in the Environment

What happens if the world changes in a way that might affect
the denotation of the word-pairs? Sticking with Pustejovsky’s
qualia-model, let us assume for instance that a new function
for the denoted object arises, but that the other submeanings
are not affected by this change. For instance, assume that
through the introduction of a new breed and as a consequen-
t of a change in fashion, dogs are used not merely for hunting
as before, but also as lap dogs. Yet, assume that the newly
introduced entities clearly qualify as canis lupus familiaris,
and therefore, fall under the denotation of both hound and
dog.

Technically, the process described above amounts to a mu-
tation in the environment, which causes also a mutation in
one submeaning. As a consequence, we will need to develop
two subtypes in some quale, as in (4) (where k and i denote
weights associated to the different types of the submeanings).

Figure 2: Sample result of reinforcement learning of
denotation-equivalents, after 4000 iterations, with initial in-
clination weights of 1 and a reward of 1.

Figure 3: Absolute Differences Between Corresponding Sub-
meanings With Differing Initial Inclination Weights
Relating Subtypes of Meaning

Once we have concluded that the two meanings are to be put into relation, the question is how exactly we should do that. Up to now, frequencies and random draws have played an important role, and it is desirable not to lose these properties, and to go deterministic merely due to the presence of a second choice. Furthermore, I assume that if there was a strong separation in the old type of the submeaning, this would involve a strong association of one word with the old meaning — which in turn should give incentives to associate the new meaning with the other word, even if both words have identical initial inclination weights for the new type. However, if the weights at the old submeaning are roughly identical, the basis for deciding which word should be associated with which submeaning will be much less clear. These are the basic desiderata that I have tried to incorporate into the algorithm.

The formula I used for weighing off frequencies in case of a submeaning having two subtypes is thus the following:

\[
\text{weight of } M \text{ in } W = \frac{\text{weight of } M \text{ in } W}{\text{weight of } N \text{ in } W + \text{weight of } M \text{ in } V} \]

Assume for the sake of the argument that we want to calculate the weight for the use of Word1 for meaning subtype-1 of some submeaning \( \Sigma \), given an initial case, since it does not require anything that would be different from the procedure used up to now, and merely adds one Polya-urn. Therefore, this should be the base case if one assumes non-intelligent (or: ‘mind-blind’, cf. Baron-Cohen (1997)) agents. Of course, there is nothing that would prevent the new submeaning from acquiring — through stochastic reinforcement — a frequency-pattern that is the inverse of the old submeaning. However, this will not happen systematically, and there is no reason whatsoever why one word should specialize for one submeaning, and the other word for the second.

What would it take then to cause a differentiation in meaning? The crucial dimension is to establish a link between the old type of the submeaning and the new one. Human agents might take into account the fact that before the change of the world, one of the two words was strongly correlated to the (old) subtype of the mutated submeaning, and assume that the hearer will take this into account. This pattern of reasoning is normally assumed in Bi-directional OT (cf. Blutner, 1999) in order to derive markedness patterns in natural language. The predicted outcome of such a process is that using the word that is less correlated with the old type will be preferably used in order to signal the new meaning — if there is an incentive for speaker and hearer to do so, which may be the case if there are differential payoffs for using one entity or the other for the two functions. For instance, hunting with a chihuahua might reduce the chances of finding game, and a mastiff is likely to make a wearisome lap-dog.

It is important to notice that the differentiation in the meanings of the two words with respect to the newly arisen type of submeaning is pragmatically induced, and involves — at least in the version of Bi-directional OT — reasoning about (or a simulation of) what the other participant in the communication would do. In brief, the differentiation process presupposes agents with a Theory of Mind, who exploit the difference in frequencies between the two linguistic forms in order to convey a particular meaning.

<table>
<thead>
<tr>
<th>Submeaning</th>
<th>Word1</th>
<th>Word2</th>
</tr>
</thead>
<tbody>
<tr>
<td>subtype-1</td>
<td>800</td>
<td>200</td>
</tr>
<tr>
<td>subtype-2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

So far, this does not change from what we have experienced before. However, if we calculate the probability of choosing Word2 for the new meaning, the probability is not 0.5 — as would be the case if subtype-2 was independent from subtype-1 —, but rather 0.8 (since we have \( \frac{1/200}{1/200 + 1/800} = 0.8 \)). And if Word2 is chosen and reinforced, this probability rises to 0.8.

Assuming that such a pragmatic reasoning takes place (and it does not matter what exactly the chosen algorithm will be to weigh off the two submeanings), it will only have an effect if the submeaning undergoing change is associated strongly enough with one word. If the frequencies of the two words are
changed world and pragmatics: density estimates for different initials

probability of choosing word1 for meaning1

init:01
init:05
init:10
init:20
init:50

figure 4: variation of initial inclination weights and the impact of pragmatics: clusters around 0.5 indicate free variation; clusters around 0 and 1 indicate a separation of the meanings of the two words.

roughly equal before the change of the world, the pragmatic procedure will fail to achieve the separation of the meaning of the two words. This can once again be shown by varying the different initial inclination weights of the Polya-urns: the higher the initial endowment, the less the differentiation into two clearly separated meanings will be noticeable. Figure 4 illustrates this behavior. I have plotted 1000 simulations involving a mutation of the world for varying initial inclination weights, and the diagram shows density estimates of the probability for choosing Word1 for the subtype1 of the mutated submeaning. If the probability mass clusters around 0 and 1 (with hardly any cases in between), this means that either Word1 will be chosen (nearly) all the time or hardly ever for that particular meaning — which is what happens with low inclination weights. Such a pattern is evidence that the words have acquired a specialization in meaning. However, if the probability mass clusters around 0.5 (as can be seen with the inclination weight of 50), this indicates that there is a tendency to obtain free variation: at 0.5, one or the other word is chosen for one or the other meaning with the same frequency.

The reason for this pattern is obvious: the lower the initial inclination weight, the higher the mean difference before the change of the world (see figure 3, page 4). And the higher the difference before the change of the world, the stronger the relative associations with one particular word.

Further Consequences of the Chosen Formula

As figure 4 shows, formula (5) has in some circumstances a bias for separating the meanings of the two subtypes in the mutated quale. I will explain the origin and exact working of this behavior now.

Let us assume a basic configuration with a reward of 1, and a configuration just like in 2, but with Word1 having a weight of 750 at subtype-1, and Word2 a weight of 250 at subtype-1 (that is, the median outcome for an inclination weight of 1), as illustrated in table 3. As the reader may check, the probability of choosing Word1 for subtype-1 will be 0.75, and the probability of choosing Word2 for subtype-2 will equally be 0.75.

Table 3: Sample Case of Submeaning with Two Subtypes.

<table>
<thead>
<tr>
<th>Submeaning k</th>
<th>subtype-1</th>
<th>subtype-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>750</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Reinforcing Word1 for subtype-1 and reinforcing Word2 for subtype-2 will not however have the same consequences: If Word1 is systematically reinforced for 50 rounds for subtype-1, but subtype-2 is never drawn by Nature, the probability for drawing Word1 for subtype-1 will have risen from 0.75 to a little more than 0.76; if Word2 is systematically reinforced for 50 rounds for subtype-2, but subtype-1 is never drawn by Nature, then the probability of drawing Word2 for subtype-2 will have risen from 0.75 to a little more than 0.99. The threshold of 0.99 is already reached after 33 rounds. This effect is illustrated in figure 5, which shows once again that the increase in probability for Word2 is very steep at first, and than flattens out gradually.
So, the bigger the difference in frequency of the two words with respect to the submeaning before the mutation, the smaller the probability that the strongly associated word will ever be reinforced for the new subtype of meaning.

Notice however, that according to formula (5), the first reinforcement of the new submeaning will have drastic consequences, whatever its direction: should in our case Word1 be reinforced for subtype-2 (which would happen with a probability of 0.25), the probability of choosing Word1 for subtype-1 will go down to 0.6.

A last important property of the formula is that it has symmetry built in, and a bias for separation. That is, if Word1 is reinforced for subtype-1, this will alter at the same time, and by the same amount, the probability of choosing Word2 for subtype-2, and vice-versa. Assume that in the scenario outlined in table 3, Word2 is reinforced for subtype-2 to 2, with everything else remaining as is. This reinforcement increases of course the probability of choosing Word2 for subtype-2 from 0.75 to 0.857. But at the same time, it also increases the probability of choosing Word1 for subtype-1 from 0.75 to 0.857.

**A (But Not The) Scenario of Hyponymy Reversal**

Now we have all the ingredients in place to sketch a scenario of how the hyponymy reversal between *dog* and *hound* (or any other two words) might have happened.

A first step would have to be an impoverishment of the environment, rendering the two words denotation-equivalents. Next, over the rounds of reinforcement learning, the submeaning whose mutation would eventually be caused by a change in the environment would have to receive a strongly biased frequency in favor of one particular word. Then, after a second change in the environment, the previously less favored word would become associated with (some aspect of) the newly introduced type of referent through rounds of (pragmatically conditioned) reinforcement learning.

Technically, I have not yet derived the development of a relation of hyponymy, since both forms remain in principle possible for all meanings — although some will become less and less likely to appear for a given form. The solution to this problem will probably need to involve forgetting one subtype (or one submeaning). Exploring the impact and dynamics of forgetting cannot be done in any detail here for want of space (but see Skyrms, 2010, p. 133ff. for a discussion of forgetting in reinforcement learning). Suffice it to say that the smaller the weight of a subtype, the higher its risk of suffering elimination by forgetting. Intuitively, a subtype with little weight is a subtype that has not been encountered often (if at all) in connection with a given word, and all things being equal, it seems reasonable that forgetting affects the rare rather than the very frequent.

We have seen in the case-study above that the probability of using Word2 for the established subtype is low, and declines rapidly. Therefore, it is a good target for forgetting, consequently rendering it impossible for use for the established subtype. The net result would be that Word2 specializes for one subtype of a quale, whereas Word1 continues to be appropriate with both subtypes (although it would be used more often with the established subtype). In this way, Word2 becomes a hyponym of Word1, specifying information for a submeaning that Word1 does not specify.

Summing up, the scenario involves the impoverishment of the environment, causing words whose denotations overlapped (that is, words that had been partial synonyms according to the definition of Cruse, 2000) to become denotational-equivalents. Then, reinforcement learning ensues, followed by a second change in the environment reintroducing diversity, and finally, forgetting.

Before concluding, it needs to be stressed that randomness plays a major role in the simulation. Even with identical external environments, there is no guarantee whatsoever that meanings will shift (or not) in one way or another. This is probably a welcome fact. While I have not found comparative studies of populations of *canis lupus familiaris* in different European countries, we cannot assume that Medieval England (where a hyponomy reversal took place) was very different from Medieval Germany or Denmark (where nothing of that sort occurred). So, we should look for models that make such a shift possible, but not necessary.

**Conclusion, And A Puzzle In Perspective**

In this paper, I have simulated the effect of reinforcement learning of conceptually grounded words in an unstable and delimited environment. The simulation used Polya-urn processes on internally differentiated meanings. Since changes in the extra-linguistic environment were crucial for the proposed scenario, the particular brand of hyponymy reversals explored here is therefore rather an instance of external change.

I have also insisted on the necessity of pragmatic processes, and more precisely, the necessity of agents able to guess what another agent might infer from their signal, for the (re-)differentiation of meaning. Against the position of Haspelmath (2006), I have argued that markedness cannot be derived from frequency alone, in the absence of agents disposing of Theory of Mind to interpret these frequencies.

It is obvious that not all changes of lexical meaning can have their source in a changing environment. For instance, concerning the meaning shift of German *Bein* from ‘bone’ to ‘leg’, it is nearly unimaginable to have a substantial community of Human agents lacking systematically either legs or bones over a prolonged period. Thus, a necessary future direction of research is to explore under what circumstances a differentiation of meaning is achievable when there is some, though only limited, denotational support to the meaning difference of two words.

At least one important subclass of such cases turn out to be stag hunts (for a book-length presentation, see Skyrms, 2004). This is notably the case when the two words at stake are hyponyms. In a stag hunt, players have the option of going for a zero-risk, low-benefit hare, which they will obtain...
irrespective of what the other player does. However, a player might also try to capture the more risky, but more rewarding, stag — which requires however cooperation from the other player. A sample payoff matrix for a game of stag hunt is given in table 4.

Table 4: Sample Stag-Hunt Game.

<table>
<thead>
<tr>
<th></th>
<th>Stag</th>
<th>Hare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stag</td>
<td>2.2</td>
<td>0.1</td>
</tr>
<tr>
<td>Hare</td>
<td>1.0</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Assume now a situation with reinforcement learning like we had before, but where one can use either a hyponym (e.g., *cat*) or its hyperonym (e.g., *pet*), and where the denotational difference is supported (that is, where $n\%$ of pets are cats, and $100-n\%$ are pets, but not cats). Rewards are not equal, but follow table 4. In a game where one has to match either the hyponym or the hyperonym, using the hyperonym is the safe play (and corresponds thus to the hare), since there will be no risk with it.\(^{13}\) The hyponym may yield a greater payoff, but it also involves a greater risk (it corresponds therefore to the stag).

Now comes the puzzling part: it is not clear (at least to the author of these lines) how the structuration of lexicons with hyponymy is sustainable given this fact. As far as I am aware, conditions under which a stag-playing strategy can emerge and persist do not apply in cases of lexical signalling (e.g., locality constraints between players, cf. Skyrms, 2004, p. 15ff.). The fact however that hyponymy is one of the central organizational principles of natural-language lexicons remains empirically obvious.

**Acknowledgements**

I had the opportunity to present a preliminary version of this paper in a workshop on pragmatics at the Université Lille 3, organized by Ilse Depraetere and Raf Salkie, and I would like to thank the organizers and the audience (especially Raf Salkie and Chris Piñón) for their feedback. I would also like to thank the two anonymous reviewers for their comments. Finally, I express my gratitude to Kathleen O'Connor for correcting my English. None of them should be assumed to agree with anything claimed in this paper; all errors and omissions are mine alone.

The simulations were performed with Steel Bank Common Lisp (http://www.sbc1.org); for data analysis and plotting, Gnu R (see R Core Team, 2014) was used. The source code is available upon request from the author.

**References**


Steels, L. (2008). The symbol grounding problem has been solved. So what’s next? In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 223–244). Oxford: Oxford University Press.

\(^{13}\) Or, at least, less risk: where the hyperonym is a false answer, the hyponym will by definition be a false answer as well. The reverse is however not true.
The production of partial answers and ad-hoc inferences: evidence from semi-spontaneous speech.

John Michael Tomlinson, Jr.
Zentrum für Allgemeine Sprachwissenschaft (ZAS)
Schützenstrasse, 18 Berlin 10117, Germany

Camilo Rodríguez Ronderos
Zentrum für Allgemeine Sprachwissenschaft (ZAS)
Schützenstrasse, 18 Berlin 10117, Germany

Abstract

An interactive production-comprehension paradigm investigated the semi-spontaneous production of partial responses to questions such as “Were Anton and Anna at the party?” One member of the dyad asked questions about the attendance of either one or two people at a party and the other member had to communicate the information about attendees on the screen. Overall, speakers took longer to respond to questions when they had limited information about who attended a hypothetical party, however were quicker at communicating non-exhaustive information than exhaustive information when they believed that they had limited information. Speakers also used lexical devices to mark their responses. We discuss our findings and future research directions with our paradigm.

Keywords: implicatures, speech production; experimental pragmatics; dialogue; question-answer pairs.

Introduction

The overarching goal of experimental pragmatics has been to inform theoretical accounts of pragmatic inference by using experimental methods to better understand how listeners process inferences (Noveck & Reboul, 2008). While much has been learned about the comprehension of pragmatic inferences, researchers interested in processing have had to make strong assumptions about the type of linguistic input in conversational speech, from which listeners actually derive and compute implicatures. For example, consider the following question and responses:

(1) Were Mark and Jenny at the party?

(2a) Only Mark was there
(2b) MARK was there. (pitch accent)
(2c) MARk was tHERE (rise fall rise)
(2d) I know that Mark was there.
(…)

Responses (2a) to (2d) show a subset of ways that speakers could partially respond to (1). Gricean pragmatics assumes that speakers would respond partially as a function of their knowledge state and should adhere to the Gricean maxims of Quantity and Quality (Grice, 1975). In this paper, we present a novel interactive production-comprehension paradigm. Our aim is to test if and how speakers alter how they produce partial responses to AB questions. We argue that results from our paradigm better approximate conversational speech and that understanding the interactional nature of implicatures can shed light on the actual type of utterances listeners encounter when deriving inferences.

Processing implicatures

The central debate in experimental pragmatics has primarily revolved around quantity implicatures, specifically scalar implicatures. For example, in an utterance such as “some of the students left early”, listeners would need to interpret the quantifier some against a lexicalized scale {some, all}, in which some could have a lower bound interpretation (some, and possibly all) or an upper bound interpretation (some and not all) (Horn, 1974). While both of these interpretations could be equally likely and equally available for the listener, Levinson (2000) initially claimed that upper bounded interpretations (some, not all) are understood by default, however numerous studies have shown a processing cost associated with such implicatures (Bott & Noveck, 2004, Breheny, Katsos and Williams, 2006, Huang & Snedeker, 2009). Moreover, these findings do not seem to be the by-product of a general processing difficulty associated evaluating alternatives, rather processing costs seem to result from top-down mechanisms needed to compute implicatures (Bott, Bailey, & Grodner, 2012; Tomlinson, Bailey, & Bott, 2013).

Recently, attention has been shifted from simply focusing on the processing implicatures, and have sought more clarity on the interactional nature of implicatures. In particular, probabilistic accounts of implicature correctly point out that not only are listeners confronted with potential alternatives that a speaker could have uttered, rather that listeners should be sensitive to relative frequency of forms that speakers utter in a given discourse context (Degen & Tannehaus, in press). Various rational model accounts (Benz & van Rooij, 2007; Frank 2009; Frank &
Goodman, 2012) not only seek to model processing costs associated with implicatures, but also potential costs that speakers are confronted with in production when considering linguistic alternatives.

While processing implicatures does take time and processing resources, certain cues can aid listeners derive implicatures and reduce processing costs. For example, Grodner, Klein, Cadbury, & Tannenhaus (2010) showed that listeners did not incur processing costs for the phonetically coarticated form for “some of” (summa). This finding is important because it shows the extent to which speakers might be signaling whether the derivation of an implicature is licensed via phonetic detail. Intonation also seems to play a strong role in indicating whether implicatures are derived or not. For example Chevalier, Noveck, Nazir, Bott, & Sperber (2008) showed that readers and listeners are more likely to derive exclusive interpretations for “or” when “or” was capitalized in a written task and received contrastive prosody (L*+H) in a listening task.

Another important variable is the linguistic context in which implicatures are uttered. Zondervan (2010) found that exclusive interpretations of “or” are more likely to be derived when in sentence focus (Question under Discussion). This finding suggests that the prior contextual mention of alternatives will strongly influence how likely and how quickly a given alternative should be considered and eliminated when processing an implicature.

What is less well understood, and what we seek to access, is the extent to which speakers actually encode their utterances with such cues across different communicative contexts. While these properties of the utterance and proceeding context can streamline the processing of inferences, debate surrounds the exact type of information about the speaker that listeners need to evaluate to derive implicatures. Several have argued that processing implicatures requires the epistemic step on the part of the listener (Sauerland, 2005; Breheny, Ferguson, & Katsos, 2013): to derive additional meaning above and beyond the truth proposition, listeners need to assume that speakers have adequate knowledge to make such implications. We investigate if speakers actually modify their utterances, partial answers to questions, as a function of their knowledge state and whether listeners are sensitive to these cues when deriving the implicature.

We present a study investigating how speakers semi-spontaneously produce partial answers to AB questions such as “Were A & B at the party?” in an interactive production-comprehension task. These question-answer pairs provide contextually available alternatives via ad-hoc scales {A and B; A; not B; A, not B; not A, not B}. In our task, one participant asks the recipient whether two people attended a given party and the recipient responds by communicating information available to him or her on a computer screen.

**Interactive Implicature Paradigm**

Research on implicatures has used a variety of tasks to examine how implicatures are processed online, in the hope that online measures can shed light on the type of procedures and mechanisms underlying pragmatic inference. However, most (if not all) experimental materials in these tasks rely on highly controlled experimental items, either written sentences for truth value tasks or laboratory speech for referential disambiguation tasks, e.g. visual world paradigm. We seek to get an overview of the many ways that speakers can express implicatures and whether they consistently change their responses as a function of their knowledge states.

In our task, speakers had to answer questions from an addressee about attendees of two parties: One where the speaker was at the party long enough to see all attendees (complete knowledge) and another where they only stopped by for a short period of time (incomplete knowledge). Several have argued that the derivation and the further processing of implicatures is continent upon the epistemic step (Sauerland, 2005; Breheny, Ferguson, & Katsos, 2013): to derive additional meaning above and beyond the truth proposition, listeners need to assume that speakers have adequate knowledge to make such implications. We tested if speakers actually alter how they formulate their partial responses to AB questions as a function of their knowledge state.

Critically, we can make several predictions regarding how speakers might alter their utterances as a function of their knowledge state. Obviously, speakers can use lexical means to indicated exhaustivity such as using focus particles such as “only”. Likewise, lexical devices such as embedding could be used to mark non-exhaustivity, e.g. syntactic embedding “I know that X came” or illocutionary embedding “I saw Anna at the party”. Next, different types of intonation contours might be expected for exhaustive vs. non-exhaustive responses. A pitch accent (either H* or L* +H) would be licensed for exhaustive answers (“ANNA was at the party”), whereas more dynamic intonation could be associated with non-exhaustive responses, e.g. rise fall rise (Hirschberg & Ward, 1985). Last, we examine the latencies of between mentions of a party attendee in the A’s question and B’s response. While processing costs for implicatures are associated listeners, longer latencies for producing partial reposes by the speaker, either unfilled or filled pauses, indicate that speakers might incur production costs during the production of partial responses.

**Method**

Twenty native German speakers with ages ranging from 18 to 35 participated in the experiment. The paradigm was set up as a parallel production-comprehension experiment, in which dyads sat across from each other separated by a transparent window (see diagram 1). The dyad’s task was to take part in a conversation mediated by written and visual cues in E-Prime: One subject (A) stated questions that appeared on his/her screen, while the other subject (B) subsequently answered these questions according to the visual stimuli that was presented to him/her on his/her screen via E-Prime. A and B were in separate rooms and
could hear each other through headphones connected to a live feed of the respective microphones. The question-answer sequence was recorded via a DAT tape and in E-Prime via the RecordIn function.

At the beginning of the experiment, A and B were given backstories and sufficient time to read through them as well as the opportunity to ask questions. The backstory was as follows: A and B are part of the same circle of friends. During the course of a week, B attended four parties at which several of their friends were also present. A did not go to any of the parties. During their conversation, A asks whether or not certain couples within their circle of friends attended the different parties. B would answer to the best of his or her knowledge. B could report the attendance of the couples in any way he/she choose, there was however one constraint: they could not report who did not attend, except in the case that both did not attend A [X] B [X], e.g. “keiner war da” (neither was there).

The dyads had 4 conversations with 20 questions each, sequentially presented on the computer screen. The 4 conversations were randomized as experimental blocks, representing the four parties that B attended. For half of the parties, B was told that he or she had attended to party for the complete duration and saw every person who attended party (complete knowledge condition). For the other half of the parties, B was told that they only stopped by for a short period of time and saw some of the couples, but couldn’t be sure if other couples came earlier or later (incomplete knowledge condition). A was unaware of this manipulation and was debriefed after the experiment.

The 20 questions-answer pairs consisted of 8 filler items asking about one attendee [A read “Was Lydia at the party” - B saw a picture showing that Lydia or Ben was there (b)], 4 filler items asking about two attendees [A read “Were Lydia & Ben at the party” - B saw a picture showing that both Lydia & Ben were there or that neither was there, 4 control items, and 4 experimental items. The control items and experimental items both had the question “Were Anton & Anna at the party?” and had pictures indicating that one person came, however the other person was either marked as “X” (known not to have come) or “?” (unknown whether this person came). Control items always had “X” for one person across complete and incomplete knowledge conditions. Experimental items were also matched across knowledge conditions, however one person was always marked as “X” in complete knowledge condition and as “?” in the incomplete knowledge condition. The persons who received the various marking were counterbalanced across conditions for order of mention in the question and the type of marking across conditions.

To tease apart the effect of instructions (knowledge condition) and the markings (“X” or “?”), the instructions were mismatched, so that knowledge conditions were either congruent to the relative markings (parties 3 & 4) or incongruent to the markings (parties 1 & 2). Last, we examined if the listener (A) was sensitive to B’s production of partial answers. This was done by having A check off attendees who came and either marking the other person as “X” or “?”. Again, the manipulation was unbeknownst to A and B was aware that A was naive to his or her attendance.

Results
Two dyads were excluded because of a recording error and one participant not following instructions. Also, 42 recordings were lost due to scripting error, hence leaving 213 question and answer pairs across control and experimental conditions. The recordings were then transcribed and analyzed to answer several questions. First, do different knowledge states affect the latency of speakers’ responses to AB questions. To test this, we measure the difference between the onset of party attendee in the question and from the speakers’ mention of the attendee in the partial response. Second, we sought to get an overview of consistent lexical and prosodic markings across the conditions. Last, we examined whether the question asker reliably understood exhaustive answers (“X”) vs. non-exhaustive answers (?) across conditions.

<table>
<thead>
<tr>
<th>Table 1: Referent position x Response</th>
<th>complete knowledge</th>
<th>incomplete knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>preverbal</td>
<td>94</td>
<td>88</td>
</tr>
<tr>
<td>postverbal</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>embedded syntax</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>embedded illocution</td>
<td>3</td>
<td>8</td>
</tr>
</tbody>
</table>

Latency results. While participants could respond in any way they chose, roughly 80% of responses to AB questions were simple responses, in which the attendee (either Proper Noun or pronoun) was in the preverbal, “topic” position,
e.g. Anna was at the party. Other construction types included post-verbal “focal” position, e.g. I saw Anna at the Party, syntactically embedded sentences, e.g. “I know that Anna was there”, and illocutionary embedded utterances (“I had a drink with Anna”).

We restricted our latency analyses to preverbal responses, so that the noun was always in initial (head) position in the response. The latencies were calculated by subtracting the start point of the attendee in the question from the start point of attendee the response.

A matched analysis across control and experimental conditions was originally planned, however could not be conducted due to unequal N from lost recordings. Nonetheless, response latencies were slower for the incomplete knowledge conditions (M= 2854ms, SD = 1313ms) than for complete knowledge conditions (M= 2385ms, SD = 1104ms), t = 3.74, p < 0.001. For the incomplete knowledge condition, there was a significant interaction between condition and instructions, in which response latencies for experimental items (“?”) were quicker than control items (“X”) for matched instructions, whereas experimental items had longer latencies than control items for mismatched instructions, t = 2.64, p < 0.01.

Lexical devices. The presence of several lexical devices were counted across conditions. The focus particle “nur” (only) was used exclusively for exhaustive responses (“X”). Lexical devices such as “zumindest” (at least) and “auf jeden Fall” (in any case) were used almost exclusively for non-exhaustive interpretations. Last, filled pauses (“emmmm”) were used more than twice as much for incomplete knowledge conditions vs. complete knowledge conditions (23 vs. 10), the most being found for the incomplete knowledge condition with matched instructions (17 instances).

Prosodic analyses. Are currently being conducted.

Listener Judgements. Since listeners (A) were naive of the knowledge condition manipulations and the distinction between control and experimental items, we would expect that they should be at chance level for distinguishing between exhaustive (X) and non-exhaustive interpretations (“?”). This was indeed the case except for experimental items in both the mismatched instruction conditions, in which listeners performed over 60%. We are currently analyzing specific judgements with their respective recording to see if there is a relationship between response latency, lexical, or prosodic markings for these judgements.

Discussion
In this paper, we used an interactive production-comprehension paradigm to elicit semi-spontaneous responses to AB questions. Overall, we found that speakers do produce partial responses to AB questions as a function of their knowledge state. We discuss several of these findings below. We also consider several strengths and weakness of our method and argue for its usefulness in future work in experimental pragmatics.

While speakers could express partial responses to AB question in almost anyway excluding negation, they most often opted for simple responses such as “Anna was at the party”. Because of the uniformity of the question and answer pairs, speakers choose these type of responses as the most efficient and optimal way to communicate that only one person out of the pair attended the party. While the rate of simple preverbal constructions did not differ across knowledge condition, speakers did choose to use more syntactic and illocutionary embedding when they had limited knowledge of who attended the party.

We then examined the various production pressures associated with producing partial answers across different knowledge states. Overall, speakers took the longest to respond in the incomplete knowledge condition with matched instructions. The high number of filled pauses further supports that this condition required more effort from speakers to produce a relevant response. However, we observed an interaction between item type (control vs. experimental items) and instruction type (matched vs. mismatched) for the incomplete knowledge condition. Speakers took longer to produce exhaustive responses to AB questions than non-exhaustive responses when they believed that they were only attended the party for a limited period of time. However, when speakers believed that they had complete knowledge of all attendees at the party, speakers took longer to produce non-exhaustive responses than exhaustive responses.

Speakers also used different types of lexical devices depending on their knowledge state. The focus particle “nur” was only used for communicating exhaustive information, whereas constructions such as “auf jeden Fall” and “zumindest” were more common for communicating non-exhaustive information, but were also used for exhaustive information as well. Prosodic analyses are currently being conducted to examine potential differences between pitch accenting and phrasal intonation.
One drawback of our paradigm was not allowing speakers to negate alternatives and this could arguably be considered as unnatural and equally unlikely in natural dialogue. To this point, our paradigm seeks to better approximate the interactional nature of conversational speech, while at the same time maintaining a level of experimental control over both the presentation of experimental conditions and speakers’ productions. Disallowing negation required speakers to find other ways to communicate who did not attend, which in natural speech is quite plausible in cases were politeness norms and convention might be at odds with explicit negation.

Future work seeks to integrate online measures such as eye-tracking during our task. Moreover, our recordings will serve as experimental stimuli for future comprehension studies, which have the advantage of coming from an interactive task as opposed to laboratory speech.

Acknowledgments
This work was funded by the Alexander von Humboldt Foundation in the form of a Post-Doctoral research grant awarded to John Michael Tomlinson, Jr. We would also like to thank Jörg Dryer for this technical assistance.

References