Cognitive Psychology 102 (2018) 105–126

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Some inferences still take time: Prosody, predictability, and the speed of scalar implicatures

Yi Ting Huang[a,*], Jesse Snedeker[b]

[a] Department of Hearing and Speech Sciences, University of Maryland College Park, United States
[b] Department of Psychology, Harvard University, United States

ARTICLE INFO

Keywords:
Scalar implicatures
Quantifiers
Prosody
Prediction
Semantics
Pragmatics

ABSTRACT

Experimental pragmatics has gained many insights from understanding how people use weak scalar terms (like *some*) to infer that a stronger alternative (like *all*) is false. Early studies found that comprehenders initially interpret *some* without an upper bound, but later results suggest that this inference is sometimes immediate (e.g., Grodner, Klein, Carbary, & Tanenhaus, 2010). The present paper explores whether rapid inferencing depends on the prosody (i.e., *summa* rather than *some of*) or predictability of referring expressions (e.g., consistently using *some* to describe subsets). Eye-tracking experiments examined looks to subsets (2-of-4 socks) and total sets (3-of-3 soccer balls) following *some* and found early preferences for subsets in predictable contexts but not in less predictable contexts (Experiment 1 and 2). In contrast, there was no reliable prosody effect on inferencing. Changes in predictability did not affect judgments of the naturalness of *some*, when a discourse context was available (Experiment 3). However, predictable contexts reduced variability in speakers' descriptions of subsets and total sets (Experiment 4). Together, these results demonstrate that scalar inferences are often delayed during comprehension, but reference restriction is rapid when set descriptions can be formulated beforehand.

1. Introduction

Contemporary theories of language distinguish between the linguistically encoded meaning of an utterance (*semantics*) and how this meaning is enriched by the context, world knowledge, and speaker goals (*pragmatics*). The division between semantics and pragmatics sheds light on the stability and flexibility of language use during communication, but their boundary can often be unclear and counterintuitive. Take for instance, the dialogue in (1):

(1) Reporter: Will you answer our questions during the press conference?
Politician: I will answer some of them.

Here, we interpret the politician's statement to mean that she will answer one or more of the questions posed to her, but that she will certainly not answer all of them. This intuition is so strong that it is tempting to assume that the meaning of *some* necessarily excludes *all*. Yet, exchanges like (2) demonstrate that this is not the case. Unlike the politician, the movie star uses *some* to be compatible with *all*.

* Corresponding author.
E-mail address: ythuang1@umd.edu (Y.T. Huang).

https://doi.org/10.1016/j.cogpsych.2018.01.004
Accepted 30 January 2018
Available online 06 March 2018

(2) Reporter: Did you address some of the rumors in your book?
Movie star: Of course I addressed some of them. In fact, I addressed all of them.

Yet, others uses of *some* are stubbornly resistant to context information. Exchanges like (3) are infelicitous because they lead to a contradiction between *some* and *none*.

(3) Mob Boss: Did you answer any of the officer's questions?
Flunky: ∗Don't worry! I answered some of them. In fact, I answered none of them!

This shifting pattern of interpretation reflects the distinction between semantically encoded meaning and pragmatic enrichment (Gadzar, 1979; Horn, 1972, 1989). The semantic meaning of *some* is lower bounded: It picks out any amount greater than the minimum value on the quantity scale (i.e., any value greater than *none*, or if the plural is used any value greater than *one*). This semantically encoded content cannot be cancelled. In contrast, *some* excludes *all* by way of an enrichment of this basic meaning. This pragmatic inference is based on listeners' expectation that speakers will be as informative as required but not more informative than is required (Grice, 1975). Thus, it is dependent on conversational goals and beliefs about speakers' knowledge. If the politician in (1) had intended to spill every secret, she could have said (4) instead.

(4) Politician: I will answer all of your questions.

Since she did not use this obvious alternative, listeners can infer that there must be questions that she will not address. By adding an upper bound to *some*, this inference (often called a *scalar implicature*) excludes referents that are compatible with the maximum value on the quantity scale (*all*). Critically, since it is distinct from the semantics of *some*, listeners can still make sense of statements when the inference is cancelled or never calculated, as in (2).[1]

Psycholinguistic studies of scalar implicature have focused on how these two meanings emerge during comprehension. The earliest studies measured response times for judgments of underinformative sentences like *Some elephants are mammals* (Bott & Noveck, 2004; De Neys & Schaeken, 2007; Noveck & Posada, 2003; Rips, 1975). Responding false to these statements indicates the listener has made the scalar implicature, while responding true suggests that she has not. Bott and Noveck (2004) found that participants who judged the statements to be false took longer than those who judged them to be true. This suggests that scalar implicatures are not calculated immediately during comprehension, but instead require time to compute (see Bott, Bailey, & Grodner, 2012 for related work using speed-accuracy tradeoff method, and Tomlinson, Bailey, & Bott, 2013 for work using a mouse-tracking paradigm).

In work using the visual-world paradigm, we found further evidence that scalar implicatures are made after semantic analysis is well under way (see Huang & Snedeker, 2009, henceforth HS, and Huang & Snedeker, 2011). Participants were presented with instructions like *Point to the girl that has some of the socks* while their eye movements were measured to displays featuring a girl with a subset of one item (e.g., 2-of-4 socks) and a second girl with a total set of another item (e.g., 3-of-3 soccer balls). Critically, there was a period of potential ambiguity from the onset of the quantifier to the disambiguation of the final noun (e.g., *-ks*) where the semantics of the quantifier was compatible with both characters. If participants rapidly calculate scalar implicatures, this ambiguity could be resolved since only one of the girls has a proper subset of items. However, after the onset of the quantifier, we found that participants looked equally often at both the subset and total set, leading to slower reference resolution for *some* compared to unambiguous terms like *all*, *two*, and *three*. In fact, evidence of a scalar implicature (as indexed by a reliable preference for the subset compared to the total set) did not emerge until 800 ms after quantifier onset. These results demonstrate that under certain conditions, there is a measurable lag between initial semantic processing and the generation of a pragmatic inference.

The delay observed in HS is consistent with most of the existing research, including many studies that are cited as evidence for rapid calculation of scalar implicatures. For example, Breheny et al. (2006) used a reading paradigm where scalar phrases (e.g., *some of his relatives*) were followed by anaphors that referred back to the excluded complement set (e.g., *the rest*). They found that reading times at the anaphor were shorter in contexts which encouraged the implicature, suggesting that the upper-bounding inference had been completed by the time the anaphor was encountered. However, this study introduces approximately 2000 ms between the onset of the scalar term and the appearance of the anaphor, thus these findings are consistent with a theory where semantic analysis of the scalar term occurs prior to the upper-bounding implicature. Extended time lags are also present in Bergen and Grodner's (2012) reading study (roughly 1800–2400 ms) and Nieuwland et al.'s (2010) ERP study (roughly 1300–1700 ms). In fact, studies that

[1] The exact definition of *scalar implicature* depends on your theory of the phenomenon. Under some accounts, the implicature in (1) results from the insertion of an operator which negates alternatives (like *all*) and can be embedded in semantic structure (see Chierchia, 2004; Chierchia, Fox, & Spector, 2012). On these accounts, all the studies that we will be discussing involve the same enrichment process, and thus are all scalar implicatures. However, in other theories, including the classical Gricean account, scalar implicatures are inferences based on entire speech acts and thus embedded implicatures are impossible (see Geurts, 2009; Breheny, Ferguson, & Katsos, 2013). On these accounts, the utterances in the present study – as well as Huang and Snedeker (2009, 2011) and Grodner, Klein, Carbary, & Tanenhaus, 2010 – are not scalar implicatures because they are embedded in a definite description. Nevertheless, we will be calling them scalar implicatures because: (1) this framing is consistent with the prior studies under discussion; (2) embedded (local) implicatures do occur (see Chemla & Spector, 2011), thus we favor theories that explain them; and (3) psycholinguistic studies suggest that processing patterns in definite descriptions are the same as those found in standard, upward-entailing contexts (c.f. Huang & Snedeker, 2009 to Panizza, Chierchia, Huang, & Snedeker, 2009 or Grodner et al., 2010 to Breheny et al., 2013), thus it is parsimonious to treat these as two examples of a single phenomenon. Those who disagree are free to replace the term *scalar implicature* with their preferred alternative throughout (e.g., *the inference formerly known as scalar implicature*).