The Relevance of Cost to the Computation of Inferences

Adi Behar Medrano, Tel Aviv University, beharmedrano@mail.tau.ac.il

April 28, 2022

1 Background

1.1 Scalar Implicatures

- A sentence such as "John ate some of the cookies" implies that John ate some but not all of the cookies. Such an inference is referred to as a scalar implicature (SI).

(1) John ate some of the cookies.
SI: John didn't eat all of the cookies $\sim$
$\neg$[John ate all of the cookies]

- Some accounts of SIs have attributed this inference to the negation of an alternative utterance. In sentence (1), the attested SI corresponds to the negation of the alternative utterance "John ate all of the cookies".

- Not every utterance can be considered a potential alternative. If we assume "John ate some but not all of the cookies" is a potential alternative utterance to (1), we will end up with the unattested inference that John ate all the cookies.

(2) John ate some of the cookies.
No SI: John ate all of the cookies $\sim$
$\neg$[John ate some but not all of the cookies]

- It is clear that there should be a way to differentiate between utterances in order to determine which is considered an alternative and which is not.

- Alternatives S_1 and S_2 to an utterance S are *symmetric* if S is equivalent to the disjunction of S_1 and S_2 and the conjunction of S_1 and S_2 is a contradiction. (This is a somewhat simplified version of symmetry. For a general definition see Katzir (2014)).

- Following this definition, the alternatives corresponding to the inferences above are symmetric. The presence of an SI corresponding to one but not the other is referred to as the *symmetry problem* (Kroch 1972; Horn 2000 and Fox & Katzir 2011).

- A theory of SIs that is able to achieve the attested SIs in the case of symmetric alternatives, is said to break symmetry. In order to break symmetry, there must be some way to systematically differentiate among potential alternatives. This raises two questions:

  I. **How are alternatives differentiated? Can they be differentiated by invoking the notion of cost to the speaker? How will such cost be measured?**

II. **Are there independent reasons to assume that the set of potential alternatives is restricted or can we allow for all sentences to serve as possible alternatives?**

- To anticipate: we will see that differentiation by assuming simple costs is not possible, and that not restricting alternatives leads to undesirable predictions.

1.2 Proposed Solutions

- Horn (1972) and Gazdar (1979) restrict the set of alternatives by assuming lexically encoded scales. In (1) "all" is a scalar alternative to "some", but "some but not all" is not.

- The structural approach argued for in Katzir (2007) and Fox & Katzir (2011) generates a similar restriction constructively. Alternatives that are structurally more complex are absent from the alternative set since they are not generated by the alternative generation mechanism. In (1) and (2) the alternative including "all" can be generated structurally while the "some but not all" alternative is not generated.

- The structural approach invokes the notion of complexity to differentiate between alternatives. This complexity is stated in terms of the syntactic structure.

1.3 The RSA model and general reasoning

- A different answer to question **I** and **II** has been proposed in the framework of Iterated Rationality Models (IRM). Such models include Iterated Best Response (IBR) (Franke 2009; Franke 2011) and the Rational Speech Act (RSA) (Frank & Goodman 2012; Goodman & Stuhlmüller 2013; Franke & Jäger 2014 and Bergen et al. 2016, among others).

- These models are used to account for various pragmatic inferences, notably, scalar implicatures and exhaustive inferences associated with pitch accent (Bergen and Goodman 2015). IRMs derive these inferences by assuming recursive reasoning steps between speaker and listener.

- While IRMs are not inherently committed to assuming a specific source for the set of alternatives, some IRMs assume there are no formal restrictions on alternatives and that IRMs are able to break symmetry even under this assumption (Bergen et al. 2016).

- Under this approach alternatives are differentiated based on the notion of cost. Unattested alternatives are longer, and therefore require more (phonological) effort on the side of the speaker. Therefore, they are penalized and are sometimes dispreferred in the course of SI computation.

1.4 The computation of inferences in the RSA

- Below is a description of a type of IRM model that I refer to as the baseline RSA model. The IRM and RSA literature includes many variants to this baseline model (Degen et al. 2015; Spector 2017; Bergen et al. 2016 and Franke & Bergen 2020, among others). To my knowledge, these variants do not influence the results achieved by the model and result in the same predictions in the cases I will discuss. However, this will need to be checked more thoroughly.

- What all these models share is the idea that speakers and listeners are rational agents that infer meaning based on Bayesian reasoning.

- In the following description, I present what happens in the case where there are only two messages: *some* and *all*. The case where *some but not all* is included in the set of messages is addressed in the next section.

Intuitive description of the model:
Imagine a sequence of listeners and hearers, each slightly more sophisticated than the one before it.

1. At its basis, there is a naive listener that interprets messages using the lexicon.
   - The naive listener will interpret a message including *some* as corresponding to the world $\forall$, and a world where $\exists\neg\forall$ and assigns each such world the probability 0.5.
   - The naive listener will interpret a message including *all* as corresponding to the world $\forall$ and assigns this world the probability 1.
2. At the next step, the speaker is less naive and knows what is the world state. The speaker knows that she speaks to a naive listener, so she attempts to a select a good message.
   - In the case of the world state $\forall$, the speaker will prefer to transmit the unambiguous message including *all.*
   - In the case of the world state $\exists\neg\forall$, the speaker will prefer to transmit the true message *some*.
3. Over that, we have a more pragmatic listener that interprets messages with the knowledge of the speaker's considerations.
   - Upon hearing *some* the listener infers that the world state is not likely to be $\forall$. Because he knows that if that were the case, the listener would have said *all.*
   - Thus, the listener interprets the message with *some* as corresponding to the world $\exists\neg\forall$ and assigns this world a probability greater than 0.5.
4. And so on...

- An (simplified) illustration of the probability assignments is given below:

(3) Naive listener's probability assignment (*some* / *all*):

$P(w\|u)$	*some*	*all*
$\exists\neg\forall$	0.5	0
$\forall$	0.5	1

(4) Pragmatic Speaker's probability assignment (*some* / *all*):

$P(u\|w)$	*some*	*all*
$\exists\neg\forall$	1	0
$\forall$	0	1

(5) Pragmatic Listener's probability assignment (*some* / *all*):

$P(w\|u)$	*some*	*all*
$\exists\neg\forall$	1	0
$\forall$	0	1

- The world is slightly more complicated than the intuitive story above. Sometimes we need to consider that the probability of being in one world is different than being in another world. We also need to consider that some messages require more effort on the side of the speaker.

- A formal depiction of the model including probabilities and costs is given below:

- The model includes a set of possible world states (worlds for short). Each such world is marked by w ; a probability distribution P over world states representing the speaker's beliefs (known to the speaker); a set of messages. Each such message is marked by u and has a cost $c(u)$.

- Inferences are a result of multiple reasoning steps:

- At the first stage, the literal listener L_0 interprets messages by assigning worlds a probability distribution. Probabilities are determined by the prior probability of the world that corresponds to the message's lexical meaning.

- The meaning of a message $[\![u]\!]$ is identified with the set of worlds in which it is true. $P([\![u]\!])$ is the probability of all such worlds.

(6) $$L_0(w|u) = L_0(w|[\![u]\!]) = \left\{ \begin{array}{ll} 0, & \text{if } w \notin [\![u]\!] \\ \frac{P(w)}{P([\![u]\!])}, & \text{if } w \in [\![u]\!] \end{array} \right\}$$

- This probability corresponds to the relative likelihood given to every world-state upon hearing a message.

- The 1st-level pragmatic speaker S_1 considers the usefulness of messages using a utility function. This function assigns a utility value to a message-world pair, based on how the literal listener interprets u and the cost of the message $c(u)$.

- The speaker balances two competing factors; selecting the message that will most likely result in the correct interpretation and the physiological effort associated with uttering a message.

- $\log(L_0(w|u))$ represents the informativity of different messages. As such, this utility function encapsulates the tradeoff between informativity and cost.

(7) $$U_1(u|w) = \log(L_0(w|u)) - c(u)$$

- This utility value is used by the 1st-level pragmatic speaker to assign messages u a probability[1].

(8) $$S_1(u|w) = \frac{\exp\lambda(U_1(u,w))}{\sum_{u'}\exp(\lambda U_1(u',w)}$$

- This probability corresponds to the relative usefulness of every message assuming a world-state.

[1] λ is some positive real number that correspond the degree of the rationality of the speaker; how likely he is to select the optimal message.

- The 1st-level pragmatic listener now updates her world-state probabilities using Bayes' rule.

(9) $L_1(w|u) = \frac{P(w)S_1(u|w)}{\sum_{w'} P(w')S_1(u|w')}$

- The listener now assigns the probability distribution not by considering the prior probabilities on world states. She now considers what is the likelihood of a message being uttered by a pragmatic speaker assuming a certain world-state.
- This allows for the listener to interpret messages in a more sophisticated manner. Messages that are considered more useful by the pragmatic speaker in a certain world-state, are now interpreted as corresponding to that world-state by the listener.
- Reasoning steps at the following steps are conducted similarly.

(10) $U_{n+1}(u|w) = \log(L_n(w|u)) - c(u)$

(11) $S_{n+1}(u|w) = \frac{\exp\lambda(U_{n+1}(u,w))}{\sum_{u'} \exp(\lambda U_{n+1}(u',w)}$

(12) $L_n(w|u) = \frac{P(w)S_n(u|w)}{\sum_{w'} P(w')S_n(u|w')}$

- To summarise, the utility function corresponds to the utility and cost of a message. The speaker assigns messages probabilities based on this utility function. The listener assigns world-states probabilities based on the likelihood of the uttered message being used to convey that world state.

1.5 Breaking symmetry in the RSA using costs

- Suppose now that we don't restrict the alternatives and that we do have *some but not all* as an alternative.
- In the case of example (1), we now have the messages *all*, *some* and *some but not all*, and we need to reason about the same two states as before, $\exists\neg\forall$ and $\forall$.
- We will assume that the cost of the messages with *some* and *all* is 0 and that the cost of the message including *some but not all* is some positive value $c > 0$.
- Reminder: costs reflect the physiological effort expended in uttering a message. We are interested in costs because we wish to know if they can be used in order to differentiate between alternatives.
- Speakers will tend to minimize their use of the longer message *some but not all*. This message true in some states, but it is costlier.
- In our example, we are interested in the conditions in which the speaker will select the shorter message *some* when she means to express $\{w_{\exists\neg\forall}, w_{\forall}\}$ and also the conditions in which the listener will interpret this message as corresponding to this world state. We will show this for the 1st-level pragmatic speaker S_1 and listener L_1.
- We can compute the conditions in which the speaker selects the shorter message, when $S_1(\textit{some}|w_{\exists\neg\forall}) > S_1(\textit{some but not all}|w_{\exists\neg\forall})$.
- This happens when $c > -\log(P(w_{\exists\neg\forall}))$.

- We will return to this point later in section 3.3, but for now, we will assume the cost to the speaker and the prior probability for $w_{\exists\neg\forall}$ are such that the inequality $c > -\log(P(w_{\exists\neg\forall}))$ holds.

- On the side of listener, exhaustivity is expressed by the change in probability assignment. When $L_1(w_{\exists\neg\forall}| \mathit{some}) > P(w_{\exists\neg\forall})$.

- This happens when $c + \log(P(w_{\exists\neg\forall})) > \log(P(w_{\forall}))$.

- If these conditions on costs and priors are met then an utterance with *some* will be interpreted as conveying *some but not all*.

- Bergen et al. (2016) propose that cost to the speaker is determined by the length of the utterance, measured in number of words or syllables. We will refer to this cost as "flat" costs, since it does not relate to the structural complexity of an utterance.

- Under this assumption, the RSA model can differentiate between the alternatives that result in attested and unattested SIs in (13) below:

(13) John ate some of the cookies.
SI: John didn't eat all of the cookies $\sim$ ¬[John ate all of the cookies]
No SI: John ate all of the cookies $\sim$ ¬[John ate some but not all of the cookies]

- By assuming this cost difference, the utility function assigns an overall lower utility to the costly message in this case. This low utility corresponds to a lower probability to be used by the speaker. We saw this for the 1st-level pragmatic speaker and this probability only diminishes at every reasoning step.

The intuition behind the computation:
In the case of non-costly messages, the fact the speaker did not use a certain message is informative and is taken to imply the message is false. In the case of costly messages we can't follow the same reasoning, because there is the possibility that the speaker did not use a message, not because it is false, but due to its cost.

- So from the message *some* we can still reason that the message *all* is false (since if it were true it would have been a better message than *some* and would have been used), but we cannot reason that the message *some but not all* is false (since even if it were true, its cost might prevent it from being a better message than *some*).

2 Differentiation between alternatives assuming flat costs

- The idea that we can differentiate between alternatives based on their length is suggested in non-probabilistic theories as well (going back to Grice (1975)).

- This proposal faces some challenges. In the case of indirect implicatures, where a simpler and less costly alternative does not generate an SI (Atlas & Levinson 1981; Harnish 1976; Hirschberg 1985; Horn 1989 and Sauerland 2004).

(14) Kai didn't have all of the peas last night.
SI: Kai had some of the peas last night. $\sim$
¬[Kai didn't have some of the peas last night]

No SI: Kai didn't have any of the peas last night ⇝
¬[Kai had some of the peas last night]

- The structural approach has the resources needed to avoid generating the unattested alternative (Katzir 2007; Fox & Katzir 2011 and Trinh & Haida 2015). The structural approach assumes that alternatives result from a restricted set of operations on the syntactic structure. These operations generate the alternative corresponding to the attested SI, and avoid generating the alternative corresponding to the unattested SI.
- If we assume, as in the RSA approach that both alternatives are available and that they are differentiated by flat costs, we will find it difficult to break symmetry in this case.
- Moreover, in some cases, we do negate long and costly alternative. In (15) below (from Matsumoto 1995), the attested SI corresponds to the negation of a long contextual alternative.

(15) It was warm yesterday, and it is a little bit more than warm today.
SI: ¬[It was a little bit more than warm yesterday]

- The example in (15) is not unique to this context. This is a general problem for approaches that rely on naive costs, and will arise in any situation that involves a costly alternative provided by the discourse context.
- In the the following example from Katzir (2007), a longer and costlier alternative including *just some* result in an SI.

(16) John was required to do *some* of the homework yesterday, and he was required to do *just some* of the homework today.
SI: ¬[John was required to do all of the homework yesterday]
SI: ¬[John was required to do *just some* of the homework yesterday]

- An additional example from Trinh & Haida (2015), where the SI corresponds to the negation of a long and costlier alternative:

(17) Bill went for a run and didn't smoke. John (only) went for a run.
SI: ¬[John went for a run and didn't smoke]

- If we return to the question **I** it is clear that it is not possible to differentiate between alternatives based on flat costs. Flat costs do not predict the SI pattern when long contextual alternatives are available and when short alternatives are absent.

3 Unrestricted alternatives

- We now turn to the second question in **II** regarding the restriction on the set of alternatives.
- The IRM approach to SIs stated in Bergen et al. (2016) does not restrict the set of alternatives. Whether an inference is computed or not is determined by the cost and informativity of a message.
- In this respect, the IRM approach differs from other accounts in the literature. Neo-Gricean accounts and the structural approach restrict the set of alternatives while the IRM approach does not.

- Since this approach considers costly alternatives in the inference process it is possible that in some cases costly alternatives will not be ignored.
- In sections 3.1 and 3.2 we will discuss specific examples in which it has been argued that costly alternatives are used to derive the correct inferences. The first is the exhaustive interpretation of pitch accent and the second the interpretation of fragment answers. We will see that both of these accounts face major empirical problems. In section 3.3 we will discuss the predictions of assuming an unrestricted set of alternatives even in simple cases of SIs. This assumption will lead to unattested anti-exhausitivity inferences.

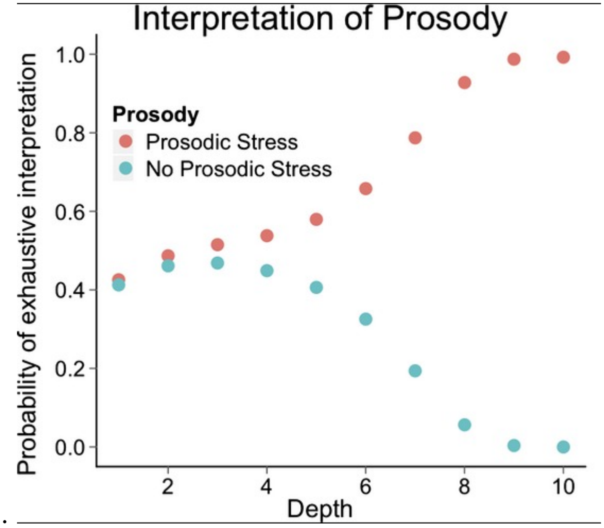
3.1 Pitch Accent

- The following discussion of an account of pitch accent serves as an indirect argument against considering costly alternatives in the inference process.
- A central case where costly alternatives have been argued to be part of the inference process exists in the literature in Bergen & Goodman (2015)'s analysis of pitch accent.
- In (18) below, the given answer indicates an exhaustive interpretation. We interpret this answer as stating that John introduced Mary to Bill and John did not introduce anyone else to Bill (accented material is indicated by capital letters).

(18) Q: Who did John introduce to Bill at the party?
A: John introduced MARY to Bill.

- This is a general property of using pitch accent in English.
- Bergen & Goodman (2015) assume that the exhaustive interpretation of sentences like (18) is the result of iterated reasoning.
- The modification of the RSA suggest in Bergen & Goodman (2015) includes the possibility of noise disrupting the speaker's message. This possibility motivates accenting in cases where the speaker intends to convey an exhaustive message.
- For instance, a possible context for (18) can include three individuals: Mary, Bill and John. If the speaker knows that only Mary was introduced to Bill, then the speaker knows that John was not introduced to Bill. In order to remain truthful and cooperative, the speaker does not want the listener to mistake "Mary" for "John". By making "Mary" acoustically prominent, the speaker minimizes the probability that noise will corrupt the message and that the listener will hear a false message.
- The listener is aware of this consideration by the speaker. When the listener hears an accented utterance, the listener can infer that the speaker intended to make the accented constituent prominent.
- Through iterated Bayesian reasoning, the listener concludes that the speaker intended to convey that only Mary was introduced to Bill.
- Accenting of constituents involves physiological effort, so speakers minimize their use of this prosodic device. Under this account, listeners are aware of this limitation on accent. In the absence of pitch accent, listeners are less likely to infer an exhaustive inference.

Figure 1: The relationship between accent and exhaustive interpretation, from Bergen & Goodman (2015). The graph shows the relationship between the number of reasoning steps (x axis) and exhaustive interpretation (y axis)



- The listener knows that if the sentence had an exhaustive interpretation, then the speaker would have used a special intonation. Because the speaker did not place pitch accent, the listener can infer the speaker does not have exhaustive knowledge along any dimension.

- In this case the speaker and listener consider costlier alternatives to an utterance, i.e the accented alternative.

- If successful, this account is a potential argument in favor of the inclusion of costly alternatives. However, I will argue that the account suffers from several serious problems which undermine this potential argument.

3.1.1 Undergeneration

- Pitch accent does not always result in exhaustive inference. In particular kinds of questions, named mention-some questions (Groenendijk & Stokhof 1984), pitch accent is obligatory even if corruption results in a true message. The answer given in (19) is felicitous in a context where all individuals can be given as possible answers.

  (19) Context: Both Bob and Mary can babysit Bill.
  Q: Who can babysit Bill?
  A: MARY can.

- Bergen & Goodman (2015) predict that the constituent "Mary" will not be accented in this case because corruption (to "Bob") will not entail a mismatch with the speaker's world state.

3.1.2 Overgeneration

- Bergen & Goodman (2015) predict that an entire F-marked constituent should be accented. Since this constituent is the source of possible corruption, we expect the entire phrase to be prosodically prominent.

- However, it has been noted in literature that in large constituents, only a sub-constituent is accented (Chomsky & Halle (1968); Selkirk (1986))

- For instance, in the following sentence (20) accent falls on the word "banana", but this prosody admits exhaustive interpretation over other constituents.

(20) Q: What did John have for breakfast?

a. A: This morning, John (only) ate a green BANANA.

b. # A: This morning, John (only) ATE A GREEN BANANA.

- In the context in (20) accenting on BANANA is used to imply that John only ate a green banana and did not eat a red apple or drink coffee. Although the speaker knows that John did not eat or drink anything else and corruption of every word will lead to falsehood.

- Moreover, the answer in (20b) is odd. This is a problem for Bergen & Goodman (2015) since they draw a direct relationship between accented material, corruption by noise and exhaustive interpretation. The dimension over which the speaker has exhaustive knowledge of is the accented material.

- A possible reply by Bergen and Goodman that can account for the infelicity of (20b) and the inferences associated with (20a) is a condition on utterance cost. If pitch accent is costly, then accenting many words as in (22b) could be too costly for the speaker. But, in order to communicate their exhaustive knowledge they accent the sub constituent BANANA.

- However, it is not clear how this strategy will help the inclusion of the large constituent in (21) as an alternative, but exclude large constituents in (23):

(21) Q: What did John have for breakfast?
# A: This morning, John (only) ate a GREEN banana.

- In (21) we only consider alternatives to the adjective "green" and not the larger constituents "green banana" and "ate a green banana".

- This problem is not shared by non-IRM accounts of pitch accent. Pitch accent is taken to depend on the presence of a special syntactic feature (F-marking), along with various governing principles (see Jackendoff 1972 and Schwarzschild 1999).

- To conclude, Bergen & Goodman (2015)'s account of exhaustive inferences associated with pitch accent fails to derive the right predictions in many cases.

- In order to derive the correct inferences, this account assumes that potential alternatives are not restricted. In some cases, it makes crucial use of more complex alternatives in order to generate inferences. But, this account does not offer a way to predict where pitch accent should be placed, and does not derive the right interpretation when accenting occurs.

- Bergen & Goodman (2015)'s account of pitch accent is the main account which relies on costlier alternatives in order to derive pragmatic inferences. This account is unable to successfully account for the empirical data. Bergen & Goodman (2015)'s other assumptions are mostly sensible so the assumption of costlier alternatives is a central suspect.

3.2 Fragment Answers

- Widening the scope beyond the relevance of costs to the computation of exhaustive inferences, Bergen & Goodman (2015) have argued that speaker costs can be used to account for the acceptability of certain fragment answers.

- Fragment answers, as in (22) and (23) are sensitive to surrounding discourse. For example, the fragment "the cookies" is acceptable if we are discussing what John ate, as in (22), but not if we are discussing whether he ate, as in (23).

  (22) Q: What did John eat?
  A: The cookies.

  (23) Q: Did John eat?
  # A: The cookies.

- It is clear that the use of fragment answers is sensitive to contextual information, but the linguistic literature also notes that fragment answers must meet some syntactic and semantic conditions. For example, fragments are taken to correspond (syntactically, semantically, or both) to the Wh-phrase in a licensing question (von Stechow 1990).

- Bergen & Goodman (2015) have proposed that the use and felicity of fragment answers can be accounted for by conversational principles. Grammar does not restrict fragment answers and it is a condition on recoverability that determines their use. Any utterance (grammatical or ungrammatical) can be used as a potential fragment answer.

- In order to account for the interpretation and felicity of fragment answers, Bergen & Goodman (2015) make use of similar general reasoning principles. They formalize the following intuition;

- Since "The cookies" is not a complete sentence, in (22) it will be considered ungrammatical by the listener. The listener assumes the speaker is cooperative and rational and attempts to salvage the intended message. This leads the listener to reason about grammatical sources for this fragment. This reasoning takes the form of considering the probability of different changes and edits to the original sentences. The listener identifies the likeliest source to be "John ate the cookies".

- The speaker can now take advantage of the hearer's reasoning and use the shorter yet ungrammatical fragment "the cookies", knowing that the listener will correctly recover "John ate the cookies", thus conveying the intended meaning while avoiding some of the effort.

- The intuition above reduces acceptability to recoverability. This condition is not sufficient and predicts unacceptable fragment answers.

- For example, if in (22) there are two possible replies: "John ate the cookies" and "John ate a green banana". Then the former answer could be identified by the determiner "the". Thus, it is predicted to be an acceptable fragment. The latter answer can be identified by the determiner "a" or "green" or "a green".

- Although these fragment answers allow for recoverability they are clearly infelicitious. Larger fragments such as "John green banana" are also infelicitous even though they allow for (possibly greater) recoverability.

- It seems that a syntactic condition on fragment answers is necessary in order to derive the correct result.

- This account of fragment answers takes a very permissive approach on possible utterances. Not only do all grammatical sentences serve as possible alternatives, almost every part of speech can be used in order to derive an interpretation.

- In question **II** we wondered whether we should restrict the set of potential alternatives. If we assume that this set is unrestricted, it is difficult to prevent these types of phenomena from emerging. This example serves as indirect against a permissive approach to alternatives.

3.3 Informative alternatives

- As we saw in section 1.5, the RSA model can assume that the set of alternatives includes costly alternatives. The alternatives are not of part of the inference process as long as their cost surpasses a certain probabilistic threshold. The threshold in the case of *some but not all* is repeated below:

  (24) $c > -\log(P(w_{\exists\neg\forall}))$

- This threshold can be stated in intuitive terms, as corresponding the informativity of the message. Costly messages can be part of the computation process, as long as they are informative enough.

- In our characterization of the utility function, if a world-state is very unlikely, it is more useful to convey that world-state precisely even if it requires more effort.

- In these types of scenarios, the model predicts that if a shorter message is used, the listener infers that the costly useful message is false.

- This type of SIs is unattested. In (25) below we observe the following SI pattern:

  (25) Q: Did Bob eat cake or ice-cream?
  A: Bob ate cake.

  a. SI: Bob did not eat ice-cream. $\sim$
  ¬[Bob and cake and ice-cream.]

  b. No SI: Bob ate cake and ice cream. $\sim$
  ¬[Bob ate cake and not ice-cream.]

- Similarly to the example in (2), the longer symmetric alternative in (25b) results in an unattested SI. Differentiation between these alternatives can be done assuming costs. The unattested SI corresponds to a costlier alternative than the attested SI.

- However, we can consider a context where the prior probability of that Bob ate ice-cream is very high. In such a context, when a literal listener hears the message expressed by the answer in (25), she will assign a very high probability to the world-state where Bob ate both cake and ice-cream.

- So the message "Bob ate cake" is almost as good as the message "Bob ate cake and ice-cream" in terms of informativity, and is better in terms of costs.

- Thus, in the case of high-prior probability, this model predicts the type of inferences as in (25b), referred to as **anti-exhaustivity** inferences.

- Cremers et al. (2022) and Schreiber & Onea (2021) have shown experimentally that anti-exhaustivity inferences do not emerge even when prior beliefs are very skewed.

- To conclude, the inclusion of costly alternatives in the inference process results in unattested anti-exhaustivity inferences. This conclusion bears on question **II**. If we do not restrict the set of potential alternatives and include costly alternatives in that set, we will end up with an unattested pattern of SIs even in simple cases.

4 Conclusion

- I discussed two central questions regarding differentiation and restriction on the set of alternatives in the cases of SI computation. Questions **I** and **II** are repeated below.

  I. **How are alternatives differentiated? Can they be differentiated by invoking the notion of cost to the speaker? How will such cost be measured?**

  II. **Are there independent reasons to assume that the set of potential alternatives is restricted or can we allow for all sentences to serve as possible alternatives?**

- I considered the possibility, that has been previously argued for, that costs to the speaker can provide a reply to both of these questions and showed that the cost assumptions made by these types of accounts are naive and do not result in the desired differentiation in many cases. Alternatives can be differentiated by structural considerations and not by costs.

- I have also considered the usefulness of speaker costs for reasoning in the cases of pitch accent and fragment answers. In these cases, costs to the speaker alone were not enough to account for their grammatical distribution.

- I showed that the unrestrictiveness assumption leads to undesirable empirical predictions: anti-exhaustivity effects in the case of SIs. This discussion leads to the conclusion that the set of alternatives is restricted.

References

Atlas, Jay David, & Stephen C. Levinson. 1981. It-clefts, informativeness and logical form. In *Radical pragmatics*, ed. Cole P., 1–62. Academic Press.

Bergen, Leon, & Noah D. Goodman. 2015. The strategic use of noise in pragmatic reasoning. *Topics in cognitive science* 7.2:336–350.

Bergen, Leon, Roger Levy, & Noah D. Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics* 9.

Chomsky, Noam, & Morris Halle. 1968. *The sound patterns of english*. Harper Row, New York.

Cremers, Alexandre, Ethan G. Wilcox, & Benjamin Spector. 2022. Exhaustivity and anti-exhaustivity in the rsa framework: Testing the effect of prior beliefs. *arXiv preprint* 2202.07023.

Degen, Judit, Michael Henry Tessler, & Noah D. Goodman. 2015. Wonky worlds: Listeners revise world knowledge when utterances are odd. *Cogsci* .

Fox, Danny, & Roni Katzir. 2011. On the characterization of alternatives. *Natural Language Semantics* 19:87–107.

Frank, Michael C., & Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science* 336.6084:998–998.

Franke, Michael. 2009. Signal to act: Game theory in pragmatics. Doctoral dissertation, Universiteit van Amsterdam.

Franke, Michael. 2011. Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics* 4:1–82.

Franke, Michael, & Leon Bergen. 2020. Theory-driven statistical modeling for semantics and pragmatics: A case study on grammatically generated implicature readings. *Language* 96.2:77–96.

Franke, Michael, & Gerhard Jäger. 2014. Pragmatic back-and-forth reasoning. In *Pragmatics, semantics and the case of scalar implicatures*, 170–200. Palgrave Macmillan.

Gazdar, Gerald. 1979. *Pragmatics: Implicature, presupposition, and logical form*. Academic Press, NY.

Goodman, Noah D., & Andreas Stuhlmüller. 2013. Knowledge and implicature: "modeling language understanding as social cognition". *Topics in Cognitive Science* 5:173—–184.

Grice, H. P. 1975. Logic and conversation. syntax and semantics. *Syntax and semantics* 3:41–58.

Groenendijk, Jeroen Antonius Gerardus, & Martin Johan Bastiaan Stokhof. 1984. Studies on the semantics of questions and the pragmatics of answers. Doctoral dissertation, Universiteit van Amsterdam.

Harnish, Robert. 1976. Logical form and implicature. In *An integrated theory of linguistic ability*, ed. T. Bever et al, 313–391. Crowell, New York.

Hirschberg, Julia Bell. 1985. A theory of scalar implicature. Doctoral dissertation, University of Pennsylvania.

Horn, Laurence. 1972. On the semantic properties of the logical operators in english. Doctoral dissertation, UCLA.

Horn, Laurence. 1989. *A natural history of negation*. University of Chicago Press, Chicago, IL.

Horn, Laurence. 2000. From if to iff: Conditional perfection as pragmatic strengthening. *Journal of Pragmatics* 32:289–326.

Jackendoff, Ray S. 1972. Semantic interpretation in generative grammar. .

Katzir, Roni. 2007. Structurally-defined alternatives. *Linguistics and Philosophy* 30.6:669–690.

Katzir, Roni. 2014. On the roles of markedness and contradiction in the use of alternatives. In *Pragmatics, semantics and the case of scalar implicatures*, ed. Pistoia-Reda S., 40–71. Palgrave Macmillan, London.

Kroch, Anthony. 1972. Lexical and inferred meanings for some time adverbs. Technical Report 104, MIT, Cambridge, MA.

Matsumoto, Yo. 1995. The conversational condition on horn scales. *Linguistics and Philosophy* 18:21–60.

Sauerland, Uli. 2004. Scalar implicatures in complex sentences. *Linguistics and Philosophy* 27:367–391.

Schreiber, Alexander, & Edgar Onea. 2021. Are narrow focus exhaustivity inferences bayesian inferences? *Frontiers in Psychology* 12.

Schwarzschild, Roger. 1999. Givenness, avoidf and other constraints on the placement of accent. *Natural Language Semantics* 7:141–177.

Selkirk, Elisabeth O. 1986. *Phonology and syntax: the relationship between sound and structure*. MIT press.

Spector, Benjamin. 2017. The pragmatics of plural predication: Homogeneity and non-maximality within the rational speech act model. In *Proceedings of the 21st Amsterdam Colloquium*, ed. F. Roelofsen Cremers, T. van Gessel, 435.

von Stechow, Arnim. 1990. Focusing and backgrounding operators. In *Discourse particles*, ed. W. Abraham, 37–84. Amsterdam: John Benjamins.

Trinh, Tue, & Andreas Haida. 2015. Constraining the derivation of alternatives. *Natural Language Semantics* 23:249–270.

5 Appendix: breaking symmetry in the RSA

- Let us consider the example in (1), and assume, for simplicity just two worlds, $w_\forall$ (in which *John ate all the apples* and *John ate some of the apples* are true, but it is not true that *John ate some but not all of the apples*), and $w_{\exists\neg\forall}$ (in which *John ate some but not all of the apples* and *John ate some of the apples* are true but it is not true that *John ate all of the apples*).

- We will assume that the cost of the messages with *some* and *all* is 0 and that the cost of the message including *some but not all* is some positive value $c > 0$.

- We are interested in the conditions in which the speaker will select the shorter message *some* when she means to express $\{w_{\exists\neg\forall}, w_\forall\}$ and also the conditions in which the listener will interpret this message as corresponding to this world state. We will show this for the 1st-level pragmatic speaker S_1 and listener L_1.

- Utility is computed by the formulas in (7). Since *some* is true in both worlds $w_\forall$ and $w_{\exists\neg\forall}$,$L_0(w|\ \textit{some}) = P(w)$. (The are only two worlds and *some* is true in both of them. As such, *some* is treated as a tautology).

(26) a. $U_1(\textit{some}\ |w_{\exists\neg\forall}) = \log(P(w_{\exists\neg\forall}))$

b. $U_1(\textit{some but not all}\ |w_{\exists\neg\forall}) = \log(\underbrace{L_0(w_{\exists\neg\forall}|\textit{some but not all})}_{1} - c = -c$

- This follows the formula in (4). The utility of the shorter less informative message in (10a) is the informativity value. The utility of the longer, more informative message corresponds to its higher cost.

- The conditions in which the speaker selects the shorter message, when $S_1(\mathit{some}|w_{\exists\neg\forall}) > S_1(\mathit{some\ but\ not\ all}|w_{\exists\neg\forall})$.

(27) a. $S_1(\mathit{some}\ |w_{\exists\neg\forall}) = \frac{\exp(\lambda\log(P(w_{\exists\neg\forall})))}{\exp(\lambda\log(P(w_{\exists\neg\forall}))+\exp(-\lambda c)} = \frac{1}{1+\exp(-\lambda c+\log(P(w_{\exists\neg\forall}))}$

b. $S_1(\mathit{some\ but\ not\ all}\ |w_{\exists\neg\forall}) = \frac{\exp(-\lambda c)}{\exp(\lambda\log(P(w_{\exists\neg\forall}))+\exp(-\lambda c)} = \frac{1}{1+\exp(-\lambda - c - \log(P(w_{\exists\neg\forall}))}$

- The logistic function $\frac{1}{1+e^{-\lambda x}}$ is increasing, thus the following inequalities hold:

(28) a. $S_1(\mathit{some}|w_{\exists\neg\forall}) > S_1(\mathit{some\ but\ not\ all}|w_{\exists\neg\forall})$ **iff**

b. $c + \log(P(w_{\exists\neg\forall})) > -c - \log(P(w_{\exists\neg\forall}))$ **iff**

c. $c > -\log(P(w_{\exists\neg\forall}))$

- On the side of listener, exhaustivity is expressed by the change in probability assignment. When $L_1(w_{\exists\neg\forall}|\ \mathit{some}) > P(w_{\exists\neg\forall})$:

(29) $L_1(w_{\exists\neg\forall}|\mathit{some}) = \frac{P(w_{\exists\neg\forall})S_1(\mathit{some}|w_{\exists\neg\forall})}{P(w_{\exists\neg\forall})S_1(\mathit{some}|w_{\exists\neg\forall})+P(w_\forall)S_1(\mathit{some}|w_\forall)}$

- The probability $P(w_{\exists\neg\forall})$ is some $p \in [0, 1]$, and $P(w_\forall) = 1 - p$.

- So,

(30) $L_1(w_{\exists\neg\forall}|\mathit{some}) > p$ **iff**

a. $\frac{p \cdot S_1(\mathit{some}|w_{\exists\neg\forall})}{p \cdot S_1(\mathit{some}|w_{\exists\neg\forall})+(1-p) \cdot S_1(\mathit{some}|w_\forall)} > p$ **iff**

b. $\frac{S_1(\mathit{some}|w_{\exists\neg\forall})}{p \cdot S_1(\mathit{some}|w_{\exists\neg\forall})+(1-p) \cdot S_1(\mathit{some}|w_\forall)} > 1$ **iff**

c. $S_1(\mathit{some}|w_{\exists\neg\forall}) > p \cdot S_1(\mathit{some}|w_{\exists\neg\forall}) + (1 - p) \cdot S_1(\mathit{some}|w_\forall)$ **iff**

d. $(1 - p) \cdot S_1(\mathit{some}|w_{\exists\neg\forall}) > (1 - p) \cdot S_1(\mathit{some}|w_\forall)$ **iff**

e. $S_1(\mathit{some}|w_{\exists\neg\forall}) > S_1(\mathit{some}|w_\forall)$ **iff**

f. $c + \log(P(w_{\exists\neg\forall})) > \log(P(w_\forall))$