Auditing the Auditor's Economy

Larry Muhlstein (lmuhlstein@ucsd.edu)
University of California San Diego
Department of Cognitive Science
9500 Gilman Drive
La Jolla, CA 92037 USA

March 22, 2016

Introduction and background

Much of linguistic pragmatics is built upon the idea of contrasting pressures, where the speaker must choose an utterance to minimize her cost of articulation while at the same time maximizing informativity (or minimizing the cost to the listener.) Out of these pressures arise many of the inferences that can be derived by a listener about what a speaker means by what she says. Each of these competing pressures is necessary, as communication would be infeasible without the balance.

To ignore the cost of articulation would result in the speaker being endlessly verbose, as saying more makes it more likely that the listener will correctly interpret your utterance.

To ignore the cost of interpretation would result in the speaker saying nothing, but the listener therefore being left without any information at all.

Zipf (1949) provided the initial insight into this phenomenon, and used it to derive an entire theory of human behavior. He characterizes the trade-off based on the number of words mapped to each meaning in a language or the number of words a speaker would have to know in order to produce language. From this, he derives his famous power law that captures a stable relationship between the length of a word (which may be related to the cost of producing it) and the frequency of occurrence of the word in the language. This property reflects what later became known as communicative efficiency, as it predicts that a language evolves to minimize effort by making the more frequently produced words shorter.

Later, Grice (1975) captured this in his famous maxims of pragmatics from which he derives a theory of how the interpreted meanings of words can go beyond the literal and interact with context.

Horn (1984) returned to the Zipfian duality, and reframed these Gricean maxims into a pair of competing forces–the Q and R principles–which capture Horn's claim that the optimal utterance is both necessary and sufficient. The Q-principle states that you should say as much as is necessary and the R-principle states that you should say no more than is sufficient. The R-principle has an obvious motivation for the speaker, namely that she minimize her own production effort, whereas the Q-principle is instantiated less directly, since it is based on improving the listener's chance of correctly interpreting the speaker's intended meaning. If we view this interaction game-theoretically, we see that the speaker has no direct reason to care about the listener's utility, as it is the listener's and not hers.

However, we do see people communicating, and so we expect that there is a mechanism by which the listener's utility is passed at least partially along to the speaker.

There has been a lot of novel work on communicative efficiency and pragmatic theory lately including work in the vein of Zipf that illustrates how the form of the language is shaped so as to be maximally efficient (at least in terms of n-gram probabilities)(Piantadosi, Tily, & Gibson, 2011, n.d.), a theoretical framework that predicts that language is organized so as to have uniform information density (A. Frank & Jaeger, 2008; T. F. Jaeger, 2010; T. Jaeger & Levy, 2006; Genzel & Charniak, 2002), and formal theories of pragmatics that argue that listeners use rationality and efficiency norms to infer a speaker's intended meaning given the literal form of what she says (M. C. Frank & Goodman, 2012; Jäger, 2008; Franke, 2013).

Most of this literature presupposes the competition between Q and R principles without asking what pressures actually comprise these principles. The exact mechanism that causes the speaker to take the listener's cost of interpretation into account is conspicuously absent from the literature, and has a significant bearing on the nature of these proposals, which have little bite without more precise specification. Identifying the details of this mechanism will be valuable for the construction of both synchronic pragmatic theories of communication and diachronic theories of language change. In this paper, I outline three hypotheses about why speakers might care about listener costs[1] and illustrate their implications for theories of pragmatics.

Mechanisms

Mechanism 1

The first mechanism is the most prima facie plausible from a game theoretic perspective, and reflects a common economic mechanism by which the costs to another agent are internalized in one's own value function.

One of the most straightforward hypothesis classes to account for cooperative behavior is that of repeated games. The insight here is that a speaker will not always be the speaker, but will soon be the listener instead. In order to improve your chances of receiving a reasonably informative utterance when you are a listener, you might produce reasonably informative utterances when you are a speaker. In this way, you are helping to establish an implicit convention that speakers will be informative because speakers not only consider the cost of producing an utterance, but the interpretation cost they might incur if they were the recipient of this utterance.

The details of this hypothesis can take a number of forms:

1. You will be the listener soon

The first and simplest is that which I briefly sketched out above. Basically, if you expect to trade roles with your interlocutor soon, then there is a direct within-interaction pressure to be cooperative, since non-cooperation or being too lazy might be reciprocated imminently by your discourse partner. In order to elicit considerate behavior from your partner, you must be considerate to them.

[1] Alternatively, being "informative" to the listener

2. You will see this person again

The next case is only a slight extension of the idea that you will be taking a turn imminently. Instead of knowing that you will immediately switch roles with your interlocutor, this more distal mechanism suggests only that you are aware that you might interact with this person again and that they will remember your behavior and reciprocate accordingly. The key difference here is to emphasize that interlocutor-based repeated games pressures need not be limited proximally to an interaction, but can extend into your longer term social relationship with this person.

3. There are over-hearers

Repeated games pressures are not limited to repeated interactions with the same person. There might also be considerations about those who observe this interaction but do not directly participate. These over-hearers, while not directly affected by your relative prolixity choices as a speaker, may associate the choices you make with your general character. If they use these observations to build a model of your behavior, your choices as a speaker may not only affect the future behavior of your direct interlocutor, but the behavior of those who observe it. If these observers see you failing to provide sufficient information to your interlocutor, they might be more hesitant to interact with you in the future (or at least to interact cooperatively), since they may expect you to repeat these same over-terse speaker choices in interactions with them.

4. There is a reputation mechanism more generally

This final and most complex mechanism captures the idea of the formation of societal and cultural conventions. If there exists any mechanism by which your cooperation or lack thereof could become known to others beyond your immediate circle, then there exists a pressure for you to be cooperative as a speaker. This reputation property plays a significant role in ethical theory, economic theory, philosophy of language(Lewis, 1969), and even business and engineering contexts such as Yelp, Ebay, AirBNB, etc. Reputation mechanisms extend the above mechanisms by allowing for agents to communicate about the behaviors of others. In this way, agents who have not directly participated in an interaction with you or even directly observed an interaction with you can learn about your general behavior. Though the specific communication channels vary, this class of repeated games pressures allows for the possibility that any action you take might be known to any other agent. Because of this, it may become generally worthwhile for you as a speaker to be informative lest your reputation of being particularly brief interfere with the behavior of others towards you when you take the role of a listener. The transitivity of this mechanism might cause all speakers in a culture to be informative due to the convention-forming pressures of reciprocal behavior and reputation formation.

How the speaker might use these repeated game mechanisms

These repeated game mechanisms create the possibility that a speaker might be informative simply because he expects reciprocal social behavior from agents arbitrarily distant from your interactions. However, we have not yet specified exactly how a speaker might take these repeated games pressures into account. There are two general classes of mechanisms that come to mind here:

1. Online consideration: For each action they choose, a speaker might compute a probability distribution over the likelihood of each of their actions having various effects on the kinds of reciprocal speaking they might receive from various agents. The form of this computation looks like.... Here, the speaker might choose to say less when they expect their choice to have a minimal reciprocal effect, but more when they predict higher future consequences. Such a mechanism would predict that, in an experiment where participants act as speakers giving instructions to a listener, a speaker who is told that they will have a more extended interaction with their interlocutor would produce more informative and costly utterances than one who is told that they will not take a turn as a listener.

2. Culturally reified convention: Instead of computing this cost online, a speaker might simply assume for any given interaction that there is a cost to being uninformative. This is less optimal than the online computation in terms of the trade-off between utterance cost and long-term costs (as it usually involves higher utterance costs), but it involves significantly less computation and therefore may be preferable to the online mechanism that might involve high cognitive demands. If this were the case, we would expect the speaker's degree of informativity and prolixity not to vary when they do or do not expect to have a repeated interaction with an agent.

Mechanism 2

The second mechanism claims that the speaker cares about the result of the local communication. This is most akin to what could be called the Wittgensteinian hypothesis. By this I mean that it takes seriously the idea that language is for doing things(Wittgenstein, 1953) and therefore caches the listener-based utility of an utterance out in terms of the importance of the interlocutor having the correct interpretation and the probability of them having the correct interpretation. This concept can be summarized mathematically in the following equation $U_L(a) = P(i \in I|a)V(i \in I)$ where $U_L(a)$ is the listener-centric utility of the utterance, $P()$ is the probability, $V()$ is the value of the listener interpreting the utterance in a particular manner, $a$is an action or utterance that the speaker takes, I is the set of possible interpretations, and i is a particular interpretation.

With this model, the full utility of a speaker decision (combining the auditor and speaker economies) is captured simply by adding the cost of the action $a$to the listener-centric utility $U_L(a)$ of that action and is summarized by $U(a) = U_L(a) + C(a)$ where $C(a)$ is the speaker-centric production cost of the utterance/action a.

This Wittgensteinian mechanism is the simplest of the three, as the speaker cares about the listener's interpretation only because he cares about the result of the communication. The value of a particular interpretation is given by the degree to which this interpretation satisfies the goals of the speaker. For example, if the speaker asks the listener "can you bring me a coffee?" there is a lot of uncertainty about the kind of coffee (with skim milk, 2%, half and half, sugar, or cinnamon; light roast or French roast; from Starbucks, McDonalds, or Bird Rock, etc.) that the listener might bring back. The speaker might prefer the French roast with half and half from Bird Rock, but might not mind the other options too much, or they might be very picky and find all other interpretations unacceptable. These preferences are captured in the value function and modulate how likely the speaker is to be terse vs. prolix, as a speaker who will only accept one type of coffee is more likely to go through the additional effort to ask "can you bring me a French Roast from Bird Rock with half and half and no sugar," than the speaker who simply desires the effect of the caffeine.

Though the simplicity and pragmatism of this approach might make it prima facie most plausible, it is worth noting that this interpretation is in stark contrast to those of Zipf and Horn. Zipf and Horn insist that the listener's economy has to do with cost of interpretation or with the listener's effort, while this mechanism deals only with the value of an interpretation to the speaker and the probability of the listener understanding each interpretation. In this way the listener's cost is entirely absent. While this does satisfy our requirement about giving the speaker reason to be verbose, it does so without considering a listener's economy at all.

Mechanism 3

The third mechanism claims that the speaker considers, not just the local value of the listener's interpretation, but the affect of her utterance choice on the entire subsequent discourse. There are three components of the subsequent discourse that I hypothesize may affect the speaker's choice of utterance: sequential cost of misinterpretation, probability of misinterpretation detection, and cost of repair.

1. Sequential cost of misinterpretation

The cost of misinterpretation on subsequent interpretations simply acknowledges the fact that a misinterpretation of one utterance may have a causal effect on the listener's interpretation of ensuing utterances. For example, if a speaker asks "does your friend have a bicycle that I could borrow?" and you interpret the referent of "your friend" as John when the speaker intended you to interpret it as Kevin, then the speaker's subsequent utterance, "his bike was just stolen," would also likely be misinterpreted. This possibility of multiple misinterpretations would increase the cost of a speaker being too brief and would therefore serve to extend the effects of mechanism 2 further into a discourse.

2. Misinterpretation detection

The main limit to the detrimental effect of subsequent misinterpretations is the fact that, in certain circumstances, a speaker might detect the listener's misinterpretation[2]. This issue of misinterpretation detection lies in a surprisingly vacant region of the literature, which is due, I suspect, to the general tendency for pragmatics both to model isolated utterances and to assume veridical comprehension. Misinterpretation detection occurs when the speaker receives information back from the listener that is highly unlikely to have been generated if the listener had interpreted the speaker's original utterance as the speaker intended. This can be seen in the example interaction:

Speaker: "Isn't this guy crazy?!"

Listener: "I know, he believes that we should pay people just to exist."

Speaker: "You mean Bernie Sanders? I was talking about Donald Trump. He just made a phallic euphemism on live TV..."

[2]Of course there also exists a limit due to the fact that this information will not be permanently relevant. This relevance limit, though it plays a role in the cost structure of misinterpretation, does not provide robustness, since core facts may remain relevant for a long while.

Here the speaker intends to refer to Donald Trump with the phrase "this guy" and likely believes that this is possible because she assumes that the listener has liberal sentiments. The listener, instead having a conservative bent, interprets the speaker's referent as Bernie Sanders, and elaborates on why he (believes he) agrees with the speaker. Upon hearing this reply from the listener that could not possibly be consistent with her intended interpretation, the speaker infers that the listener interpreted her as referring to Bernie Sanders and therefore detects the misinterpretation.

While a fully specified theory of misinterpretation is beyond the scope of this paper (and will be articulated shortly in a subsequent paper), this concept may play a significant role in a speaker's production decisions. For instance, a speaker might produce a particularly brief utterance if she believes that she will have a high chance of quickly detecting any misinterpretations that the listener might make. If she believes that detection of misinterpretations will be difficult or will take a long while, then she might be more verbose and therefore reduce the probability of misinterpretation in the first place.

3. Cost of Repair

Not only is there a cost to the speaker of the listener misinterpreting her utterance, but upon detection of a misinterpretation, there will be an additional cost to repairing this utterance. The most straightforward component of this cost is the cost to produce the additional utterances required to defease the listener's misinterpretation and to provide the listener with the correct interpretation. This "amelioration cost" is actually negligible compared with the cost of having had a discourse that turned out to be irrelevant. This "backtracking cost[3]" reflects the amount of effort that had been expended by the speaker in having a discourse whereby the listener was not correctly understanding the speaker's intended meaning. In cases where it takes a long time for the speaker to detect a listener's interpretation, this backtracking cost may be quite significant. Of course, in realistic scenarios, such misaligned discourse branches might still have had some value to the speaker, but this value is not likely to be as great as that which would have been obtained had the interlocutors explored the intended discourse branch from the get-go. The full cost of correction is the amount of stuff that the speaker said that she would not have had to say if the listener had initially interpreted her utterance as the speaker intended.

Combining the components

So far we have three separate components of mechanism 3, the sequential cost of misinterpretation, misinterpretation detection, and misinterpretation correction. In order to have a coherent hypothesized mechanism, these three components need to be combined into one process that affects the speaker's production decisions. In order to do this, we will define the counterfactual amortized cost or "COC" of an utterance. The COC is motivated by the idea that the only way to compare the costs of an utterance over a discourse is to compare the expected costs of the discourses that would have occurred for each utterance that the speaker might have chosen. This game-tree like method allows us to take into account, not just the local effects of an utterance choice, but the downstream ones as well. Since the speaker does not have full information of course about the discourse that will follow

[3]I call it "backtracking" due to its conceptual resemblance to an agent exploring a maze who, when she encounters a dead end, realizes that the effort required to traverse that path was for naught and may have been avoided by purchasing a better map.

from each utterance choice, we represent these choices instead in terms of a probability distribution over possible discourses given each possible speaker's utterance choice. A full computation of this distribution also requires the speaker to compute all of his future choices in the resultant discourses in order to get the probabilities of the full listener-speaker pairs. This is modeled by assigning a recursive definition such that the speaker explores the tree of counterfactually possible discourses to its full depth (or cutting it off at some limit and using a heuristic to evaluate the cost of the bottom state). Once the tree is explored, we assume that the speaker makes the best possible choice at each junction and therefore ends up, at the root, with costs reflecting the probabilities of each possible discourse path given an optimal choice at each branch (given the current epistemic state). [4]

To get the value of a state, the speaker needs to take into account the two types of costs we have discussed so far: the sequential cost of misinterpretation, and the cost of repair. It turns out that the choice to define these separately was made such that the costs described by the former are all the "Wittgensteinian" costs discussed in the section on mechanism 2, whereby the cost of a misinterpretation is the total value to the speaker of the actions that the listener performs. The amortized Wittgensteinian cost of a decision is the total value to the speaker of the actions that the listener takes along each counterfactual discourse path, weighted by the probability of that discourse path given the speaker's utterance choice. The costs of repair, however, are entirely defined in terms of the amount of production effort that the speaker has had to incur. The amortized costs of repair is similarly defined as the cost to the speaker of producing all of the utterances she produces in a given discourse path weighted by the probability of that discourse path given the speaker's utterance choice. Therefore, the total expected value of an utterance choice is given by the sum of the amortized Wittgensteinian cost and the amortized cost of repair for that utterance. This measure, when filtered through the limits of the speakers knowledge (which are already encapsulated in the probability distribution over discourse paths, the limits of the speaker's computation, and the actual costs and values of the speaker, yields the speaker's optimal utterance choice if they were making use of mechanism 3.

A final benefit of mechanism 3 is that it is not only amortized, but takes the idea of communicative interaction more seriously. It does not simply compute the probability of a listener getting a certain interpretation as current formal pragmatic theories do, but also considers the idea that the speaker will be able to make observations about the listener's interpretation online. The extreme sensitivity of the predictions of this model, to the degree to which speakers make interaction inferences about the listener's interpretations suggests the need to design experiments to test the degree to which speakers do perform these inferences. If speakers do this regularly, then our contemporary theories of pragmatics and of communicative efficiency should be revised to take interactive inference into account.

Discussion

While the three mechanisms above were presented separately, this does not imply that these mechanisms are mutually exclusive. Each of them is related to the others and can also exist simultaneously with them. Mechanism 1 is fully compatible with mechanisms 2 and 3 in that a given speaker might take into account a certain degree of repeated games concerns

[4] Of course, such a computation is heavily involved and will not actually be computed by the speaker, but it makes sense to define the optimal procedure and then to adjust it later for cognitive and computational resource bounds.

as well as the more direct costs associated with the other two mechanisms. Mechanism 2 is a special case of mechanism 3 in that mechanism 3, when the horizon is limited to only the direct effects of the utterance at hand, reduces mechanism 2. This mutual compatibility does not, however, take away the interesting effects of the separate mechanism hypotheses. If it does turn out that a speaker utilizes components of each mechanism, the degree to which she incorporates each, and whether this degree is dependent on additional factors remains an empirical question.

Another complication comes from the fact that some of the costs discussed–especially with respect to mechanism 3–seem as if they might be avoidable by simply being more prolix because section 3 suggests that there may be extreme cost to being ambiguous far beyond what has been noted in previous literature. The case is not so simple, however, due to the relative impossibility of being fully explicit and avoiding ambiguity altogether. Oftentimes the speaker's information about the listener's interpretation function is so poor, that she must probe and observe the listener in order to be aware of the listener's misinferences. In such cases, the speaker might not entertain the possibility that a listener might have x belief when producing an utterance, but might still be capable of noticing this possibility when attempting to make sense of a listener's replies. Interestingly, this asymmetry between the production and comprehension hypothesis spaces seems likely to be an artifact of the underlying neural mechanisms responsible for production and comprehension, and cannot be captured by the computational-level (Marr, 1982) theory I am proposing. In order for a pragmatic theory to represent such cognitive nuances, it must have theoretical machinery, not only at the computational level of abstraction, but at the algorithmic level as well. Such a theory, while potentially rich in detail, is not encompassed in the immediate project entertained in this paper.

Additionally, these explorations might also be rich in useful information that would not have been encountered by the speaker had she been more prolix initially. In many cases, this information gleaned by taking a strategy biased towards terseness and the resultant miscommunication detection and correction might allow the speaker to make more efficient decisions in the future. By leaning more heavily on the "explore" side of this conversational explore-exploit trade-off (Sutton & Barto, 1998), the speaker may be choosing to build common ground with her interlocutor and therefore to make future interactions more efficient and more meaningful.

Conclusion

The actual mechanism behind the auditor's economy is not just a philosophical matter, but bears on empirically distinguishable theoretical predictions and therefore on our broader scientific understanding of pragmatic inference.

If mechanism 3 has any component of truth, we will have to start thinking about how costs spread out throughout the discourse and about the mechanisms by which speakers make inferences about their interlocutors' interpretations. This latter pursuit seems most pressing, since all contemporary theories of human communication operate on the assumption of "dead reckoning" whereby speakers must simply assume the listener's interpretation and move on. In almost any naturalistic scenario, there is a non-trivial chance that the listener might misinterpret the speaker, and in fact they frequently do. Despite this high chance of misinterpretation, speakers do not appear to prevent it by extreme verbosity. Instead, they seem to perform these inferences about the listener's interpretation on the fly

and correct misinterpretations as they are detected. Such a mechanism is conspicuously absent from linguistic pragmatic theory, and might help to bear on a number of contemporary issues beyond simple robustness such as why the ambiguity in natural language is ultimately more efficient (the ambiguity in a single utterance is expected to be resolved over the course of a discourse through interactive communicative inference), and why speakers care about listeners' interpretations.

To inquire about the nature of the Zipfian listener's economy is not simply to be picky about details. These details themselves shed vast amounts of light onto the nature of the higher level theory. Different combinations of these mechanisms entail vastly different theories of pragmatics and linguistic communication. In order to make progress towards precise and accurate theories in these domains, we need to determine which combination is actually entertained by speakers. To do this, we should design experiments not just to test high-level information-theoretic hypotheses, but epistemological, sociological, and decision-theoretic ones as well. Such data, when combined with this theoretical aparatus, will help us understand the nature of the pressures that affect speakers' production decisions and therefore the broader dynamics of human communication.

References

Frank, A., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the 30th annual meeting of the cognitive science society* (pp. 933–938).

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*.

Franke, M. (2013). Game theoretic pragmatics. *Philosophy Compass*, *8*(3), 269–284.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 199–206).

Grice, H. (1975). Logic and conversation. *Syntax and Semantics*, *3*, 41–58.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and r-based implicature. *Meaning, form, and use in context: Linguistic applications*, 11–42.

Jaeger, T., & Levy, R. P. (2006). Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems* (pp. 849–856).

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, *61*(1), 23–62.

Jäger, G. (2008). Game theory in semantics and pragmatics. *manuscript, University of Bielefeld*.

Lewis, D. (1969). *Convention: A philosophical study.* John Wiley & Sons.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* New York, NY: Henry Holt and Co.

Piantadosi, S. T., Tily, H., & Gibson, E. (n.d.). Information content versus word length in natural language.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction.* MIT press.

Wittgenstein, L. (1953). *Philosophical Investigations.* Oxford, UK: Blackwell Publishers.

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* addison-wesley press.