Probabilistic Pragmatic Modelling of Elliptical Corrections Using First Order Logic Semantic Representations

October 31, 2019

Marvin Koss, Universität Heidelberg, Institut für Computerlinguistik

Abstract

Previous work on modelling pragmatic intuitions [Grice(1975)] using the Rational Speech Act (RSA) framework has proven to adequately capture the derivation of pragmatic meaning from literal utterance meaning and context in varying applications, such as hyperbole understanding [Kao and Goodman(2014)] and reference games [Frank and Goodman(2012)]. In the former work, listeners determine which question under discussion a speaker is adressing with an utterance and infer the speakers affect. The present work[1] applies an analogous inferential process to compositional semantics by studying ellipses in a first oder logic representation. The probabilistic inference at the core of RSA is laid ontop of first order logic derivations.

1 Introduction

In everyday conversation disagreements, participants often correct a previous utterance anaphorically by uttering a correction that is to be substituted for the equivalent constituent in the corrected utterance. Consider the following example:

[1]The project's code is available on github.

Knowledge § 14-309.8. [Bingo] [...] sessions must not exceed a period of five hours each per session.

A: *"The defendant played at the bingo event for 7 hours straight."*

B: *"At the mayfair!"*

B manages to inform A that their superlative was nonsensical as there is only one enlightened person because A can replace the modifier *"the most liked"* with *"the only"*. Previous work has shown that this speaker preference for less complex utterances can be modeled by associating higher production cost with more complex utterances according to the Gricean maxime of quantity [Grice(1975)].
In this paper I apply this kind of substitution to a Neo-Davidsonian semantic representation [Parsons(1994)] which I show can be used to adequately reason pragmatically about the semantic roles participating in an event and even disambiguate between Questions Under Discussion (QUDs, as used in e.g. [Kao and Goodman(2014)]).
I am going to base my elaborations mainly on one central example that draws on the classic example used by [Parsons(1994)]:

$$\exists e \exists x [stab(Brutus, Caesar, e) \wedge AG(e, Brutus) \wedge PAT(e, Caesar)$$

$$\wedge LOC(e, Forum) \wedge knife(x) \wedge INST(e, x)]$$

To be read as: "Brutus stabs Caesar in the forum with a knife." This example illustrates the utility in expanding upon the Davidsonian representation, which first established that events should be passed as arguments to predicates like other discourse referents while disjoining the semantic roles of events as discourse referents to semantic role predicates.

This disjunctive form is practical because it both delivers a semantic form that is interpretable by First Order Logic (FOL) engines for model evaluation of the dialogue participants and lists the constituents of the discussed event discjunctively, so that individual elements can have an associated utterance cost and can be omitted as well as substituted.
This builds on the assumption I make of an identity between this modular semantic representation and the constituents of the phrases uttered in dialogue and held in mind, respectively. Consequently, at this level of analysis the constituents represented by the event expression (e.g. *stab(x,y,e)*) and

the role expressions (e.g. *LOC(e,Forum)*) are the building blocks and no more fine-grained substitution can be taken into consideration by speaker or listener in this model.

1.1 A Working Example

Given a court setting, where two parties engage in dialogue to discuss an event and their common communicative goal is to resolve the QUD q from the set of QUDs Q possibly addressed by the speaker, consider the following exemplary preconditions and common ground shared world knowledge swk agreed on by both interlocutors. Here, the Dialogue is initiated by the listener L with the given statement g to be corrected by the speaker S with utterance u based on S's belief b. Note that the focus is on the choice and interpretation of the correction u; L is identified with the person initiating the dialogue but is modeled as the RSA listener for the interpretation of u[2]:

Q:

I Should Brutus be executed?

II Should Brutus be lashed?

swk:

1 Anyone who kills someone else within Rome should be executed.

2 Anyone who uses a knife should be lashed.

3 Stabbing someone kills them.

4 The Forum is in Rome.

5 The Rubicon is not in Rome.

S: b Brutus stabbed Caesar at the Rubicon valiantly with a sword.

L: g *"Brutus stabbed Caesar in the Forum sneakily with a knife."*

S: u *"At the Rubicon!"*

[2]I introduce the example as natural language; in the model FOL forms are assumed to be given

In this example, S and L satisfy the items in Q differently (S, Brutus' defense, answers both Questions negatively while L, the prosecutor, thinks Brutus is guilty on both accounts.). The choice of which constituent is uttered by S should inform L about S's belief about the event, which QUD S is referring to and that this QUD is evaluated differently in S's belief. This means that if S makes no utterance, L should infer that the QUD S cares about is satisfied to the same end in S's b as in L's g.

In this work, I model the choice as well as the interpretation of the correction made here. For this I presume the existence of FOL forms for the QUD, swk, g, b. Frame semantics in the style of [Fillmore et al.(2006)] are used and a larger role inventory than the mock inventory used in the code is imaginable, but low granularity in role specificity is desirable for a higher probability of having the the same roles assigned across different phrases.

2 The Rational Speech Act model

In the rational speech act (RSA) model first introduced in [Frank and Goodman(2012)], the literal listener's (L_0) decision for interpreting an utterance is modelled as:

$$P_{L_0}(m|u) \propto [[u]](m) * P(m) \tag{1}$$

L_0 bases their guess for the meaning m of the speaker utterance u only on the literal meaning $[[u]](m)$ of the utterance while weighing in their prior belief about the probability of m. The domain of the semantic meaning function $[[u]](m)$ is boolean.

Pragmatic speakers choose an utterance by maximizing the probability that L_0 interprets it correctly:

$$P_{S_1}(u|m) \propto P_{L_0}(m|u) * e^{-C(u)}, \tag{2}$$

where the exponential comes from a softmax. The cost $C(u)$ of the utterance has been interpreted and used differently in works on the RSA model ([Kao and Goodman(2014)],[Monroe and Potts(2015)],[Kao et al.(2014)Kao, Bergen, and Goodman]), but generally represents "the psychological cost of an utterance, potentially determined by factors such as the utterance's frequency, availability, and complexity" [Kao and Goodman(2014)]. In this paper I use the cost to model the aspect of complexity, by assuming that

speakers adhere to the Gricean maxime of quantity and may reduce their utterance to what is relevant to the QUD.
Finally, the meaning of S_1's utterance is inferred by a pragmatic listener L_1 using Bayes rule, again weighing in the prior belief about the probability of m:

$$P_{L_1}(m|u) \propto P_{S_1}(u|m) * P(m) \tag{3}$$

2.1 Adapting RSA for correction

In the following I will use the notation from 1. I modify this base RSA model along the lines of [Kao and Goodman(2014)]. The state s of the world reasoned about in the base RSA is now the speaker belief b about the discussed event.

For computational tractability (see Experiments) some concessions have to be made: *swk* should be small in number and complexity of expressions. Furthermore, the listener only considers a finite amount of possible beliefs the speaker may have. The set of quds should also be small.

To reason about assigned semantic roles from the set of possible roles R, let $f\colon P \to R$, $p \mapsto assigned\_roles(p)$ be the function mapping Neo-Davidsonian phrases p from the set of possible Phrases P to the set of roles that are assigned in p. A phrase p is itself a set of constituents c and $p(\{c_1, ..., c_n\}) \in P$ the function that constructs phrases from constituents.

The cost of a phrase is defined as the sum of the cost of its constituents.

The literal listener L_0 reasons about b given the utterance u, weighing in their prior probability over B:

$$P_{L_0}(b|u) \propto [[corrected(u; g)]]_{swk}(b) \cdot P(b) \tag{4}$$

Internally, L_0 substitutes in the elliptical u in the belief it gave before g to obtain $corrected(u; g)$ as in section 1 and runs logical inference to find if b holds in $[[u_{corrected}]]_{swk}$:

$$corrected(u; g) = p(\{c \in g \cup u \mid (c \in g \text{ and } f(c) \notin f(u) \text{ or } (c \in u)\}) \tag{5}$$

The speaker considers utterances from the union of the power sets of the roles assigned differently in each possible belief the speaker may have than in the given statement g:

$$U = \cup_{b \in B} 2^{\{c \in b \mid (c \in f(b) \setminus f(g)) \text{or } (f(c) \in f(g) \text{ and } f_g^{-1}(f(c)) \neq c)\}}, \text{ where } f(b), f(g) \subseteq R \tag{6}$$

Note that the power sets contain the null utterance, which means S_1 considers not interrupting at all (see 1).

My speaker wants to communicate about a specific QUD q and considers its own value for their belief via logical inference to find if it holds in $[[q]]_{swk}$ based on swk (1. For all $u \in U$ the speaker projects the listener's interpretation of the QUD q if they substitute in u in g:

$$P_{S_1}(u|b,q) \propto \left[\sum_{b'} \delta_{[[q]]_{swk}(b')=[[q]]_{swk}(b)} P_{L_0}(b'|u) \cdot e^{-C(u)} \right]^{\alpha} \forall u \in U, \tag{7}$$

where α is an optimality or temperature parameter controlling the sharpness of P_{S_1}. The pragmatic listener finally takes the speakers QUD q into account and constructs a speaker in their head after calculating the belief prior over B and QUD prior over Q:

$$P_{L_1}(s|u) \propto \sum_{b \in B} P(b) \cdot P(q) \cdot P_{S_1}(u|b,q) \tag{8}$$

As is common across prior work on RSA modelling ([Kao et al.(2014)Kao, Bergen, and Goodman], [Frank and Goodman(2012)]), I do not take the recursive reasoning process between speaker and listener further than recursive depth 1, as this is usually sufficient to appropriately model human communicative behaviour [Kao and Goodman(2014)].

3 Experiments

3.1 Probabilistic Modelling

The RSA model is a probabilistic model of language. The Python module Pyro [Bingham(2018)] was used to model the probabilistic inferential process, while NLTK [Loper and Bird(2002)] was used for logical inference. The

Pyro RSA example for hyperbole was used as a guideline.

The respective prior probabilities $P(b)$, $P(u)$, $P(q)$ are distributed uniformly over B, U, Q. Distributions based on utterance complexity can be used to model cost here instead of in an extra cost term, as in prior workprob-langorg.

The elementary constituent cost is set uniformly to 1, while the null utterance costs 0.5 to produce.

3.2 Running time

To implement probabilistic inference, Pyro runs Markov Chain Monte Carlo (MCMC, see e.g. [Sutton et al.(1992)Sutton, Barto, and Williams]) over stochastic sampling calls. As at the top level I have 5 nested samples (L_0,L_1 belief prior, S_1 utterance prior, S_1 QUD projection, L_1 sampling of S_1) it is important to keep these search spaces small. The largest of them are B, which is not sensible to reduce, and U, which can optionally be reduced to a subset of itself depending only on the speakers actual belief[3].

The time complexity of the outermost L_1 MCMC can be reduced to be proportional to the product $t \cdot |B|^2 \cdot |U| \cdot |Q|$ by caching sample function calls (where t is the time for one logical derivation), so the largest impact on performance seems to come from the slow logical NLTK derivation t which grows in non-deterministic polynomial time with the size of swk. It is therefore sensible to keep swk small. An execution of the full program using 256 GB of RAM takes about 2 days. With the reduced utterance prior, it runs in about 6 hours.

3.2.1 Plots

Plots of all three distributions can be found in the appendix for alpha values of 1 and 10.

3.2.2 Notebook

The project's github repository contains commented object oriented code, but has an annotated functionally oriented copy in the Jupyter notebook

[3]smoke test in the notebook

RSA-corrections.ipynb closely inspired by the structure of RSA-hyperbole notebook.

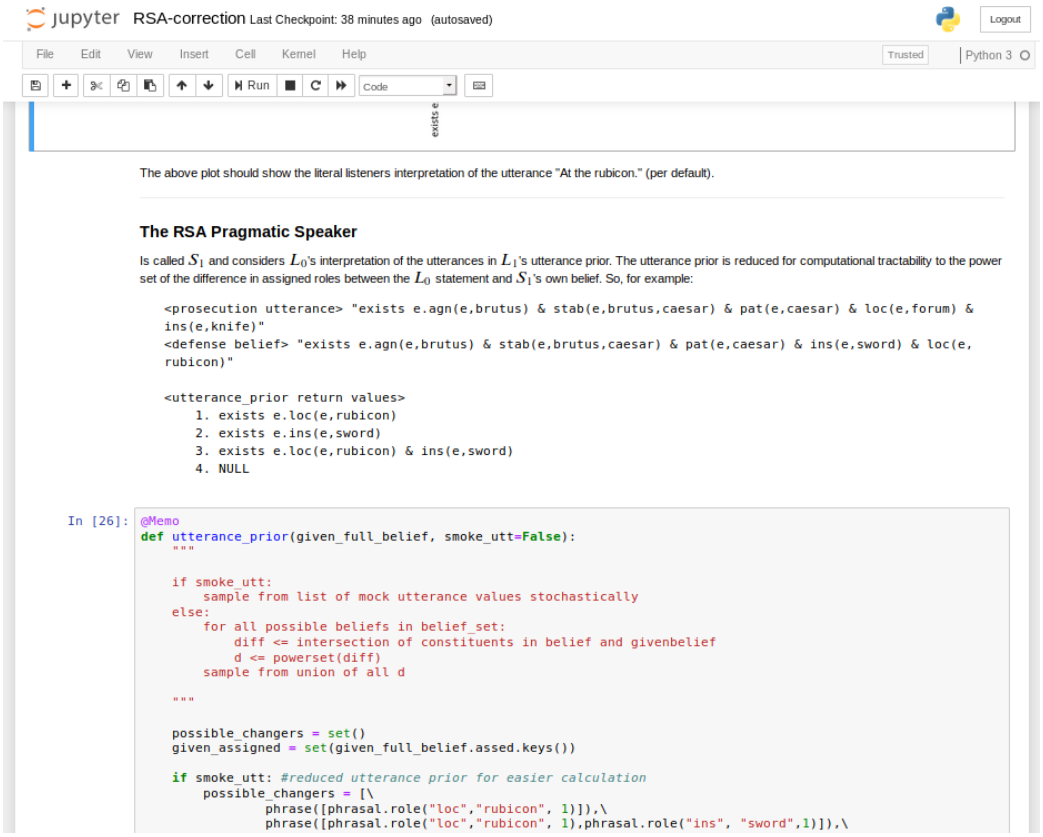

Figure 1: Jupyter Notebook Excerpt

4 Conclusions

A classical linguistic perspective was successfully united with modern probabilistic notions of inference to model intuitions for a natural language phenomenon. The setting remains contrived, but the author sees this work as merely an instance of more general work to come in the vein of probabilistic inference on top of logical derivations; most importantly the computational efficiency needs to be brought down to make this tractable.

5 Future Directions

Bridging the gap between natural language and first order logic input data would be needed to make the model applicable to real world tasks. A combinatory categorical grammar (CCG) treatment of RSA exists in [Scontras et al.(????)]; as λ-calculus parsers [Martínez-Gómez et al.(2016)Martínez-Gómez, Mineshima, Miyao, and Bekki] as well as annotated natural language data [Hockenmaier and Steedman(????)] exist for CCG, using CCG could be an interesting avenue toward natural language processing.

Pyro is designed with inference on neural models in mind. It would be sensible to extend to recursive dialogical reasoning between Agents parameterized by neural networks to handle natural language.
One possible task that could be derived from the data reasoned about in this paper would be to have Speaker and Listener keep track of a joint *swk* FOL knowledge base to which they softly attend as one such agent does in e.g. [Eric and Manning(2017)].

A Plots

Plots are made for the working example with normal temperature ($\alpha = 1$) and for a sharper distribution. Finally, P_{S_1} and P_{L_1} are visualized for a different value of q as well as b and g. The author would have liked to have more plots of L_1, but as it takes upward of 2 days to debug, more plots may follow in the github repository

A.1 Working Example

A.1.1 $\alpha = 1.0$

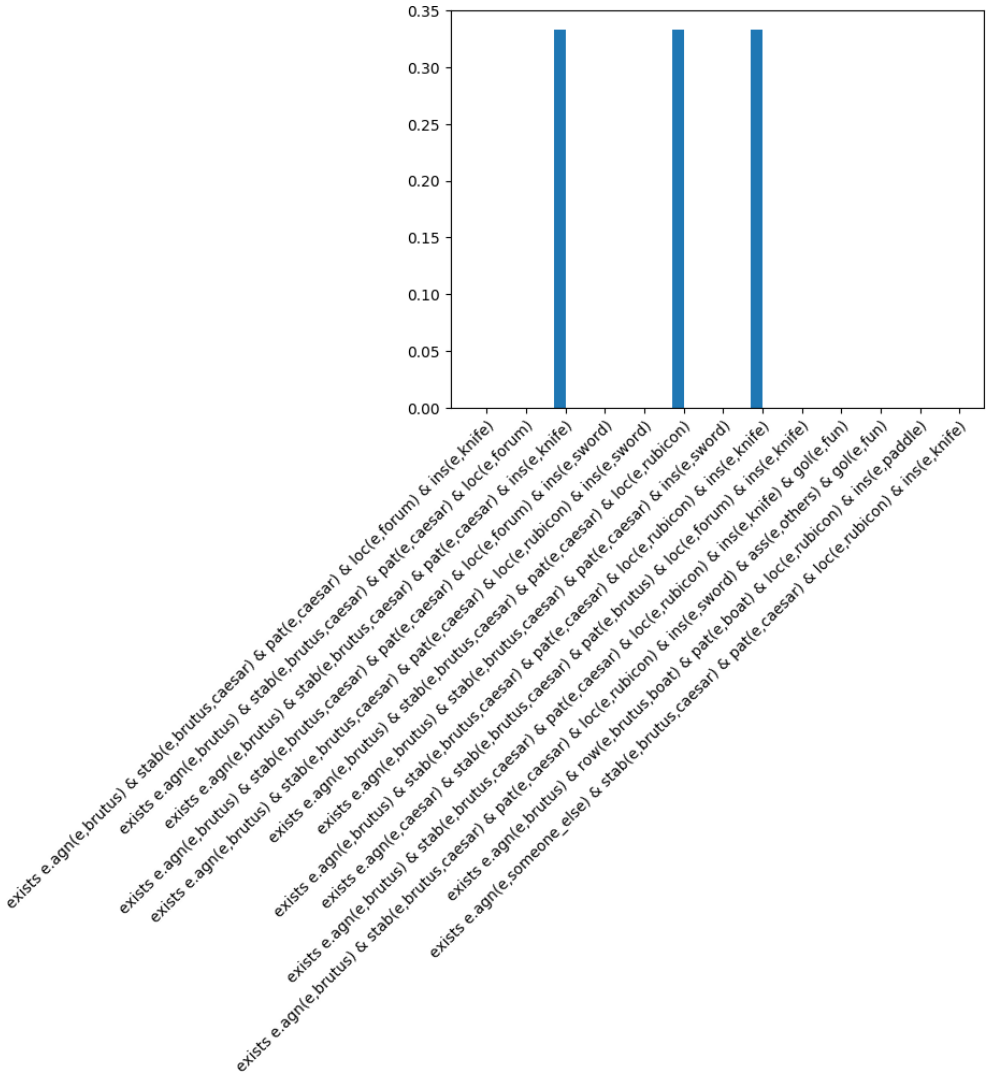

Figure 2: Literal Listener Distribution $P_{L_0}(B)$ for the statement *"At the Rubicon!"* In the working example, L_0 only places probability mass on beliefs containing *"at the Rubicon"*.

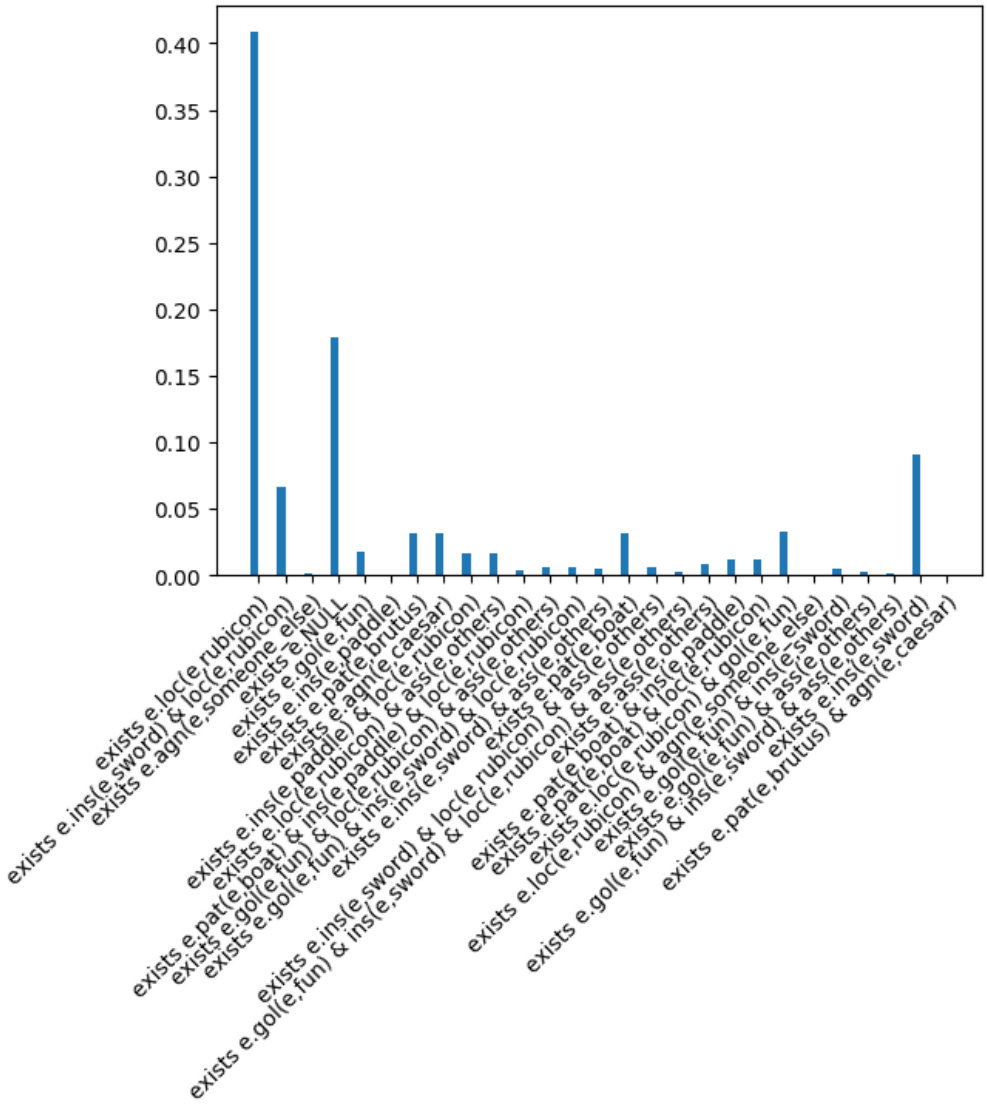

Figure 3: Pragmatic Speaker Distribution. In the working example, the speaker correctly places higher probability on *" at the rubicon "* utterances; the utterance cost plays in as expected: The Null Utterance gets more probability mass than *"with a sword"*

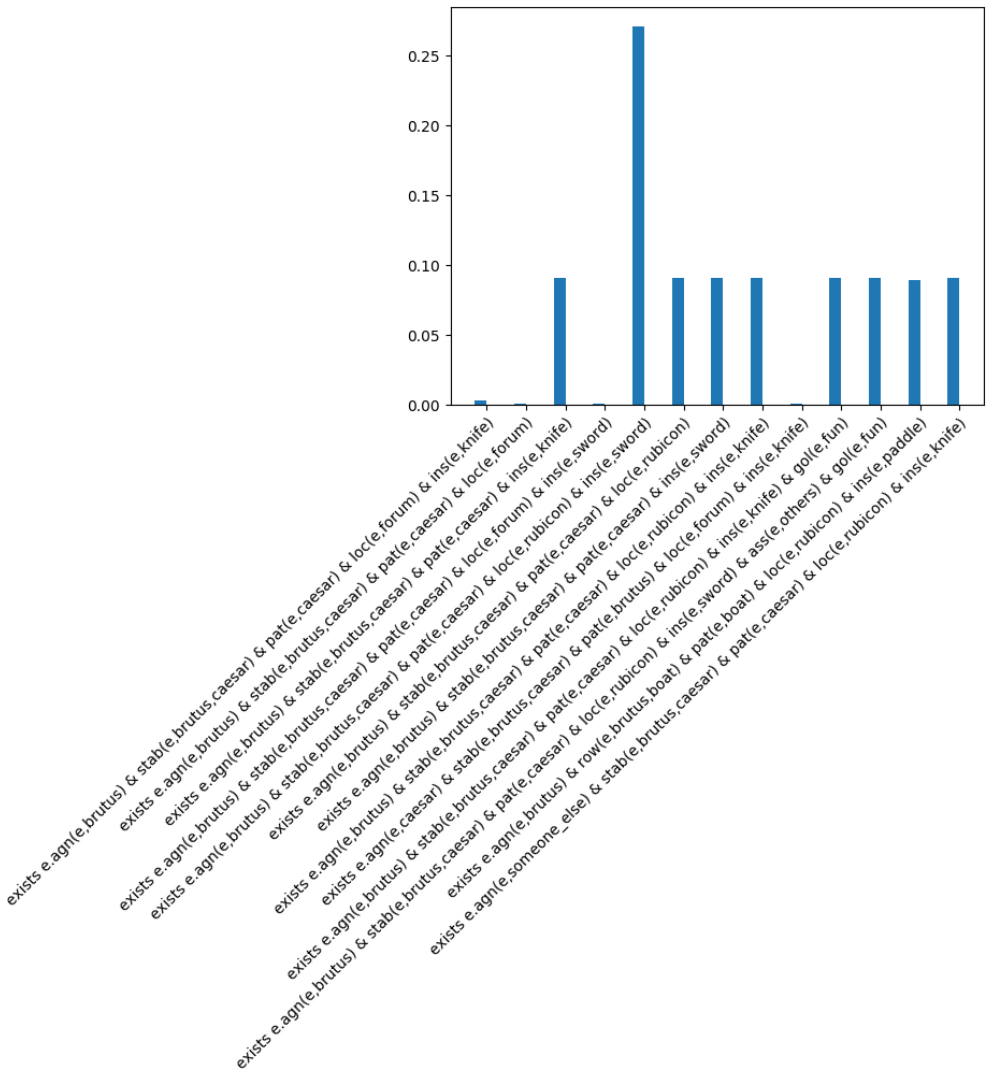

Figure 4: Prag Listener Distribution $P_{L_1}(B)$ for the statement *"At the Rubicon!"* In the working example, L_1 correctly infers S_1's belief containing *"at the Rubicon"*.

A.1.2 $\alpha = 10$

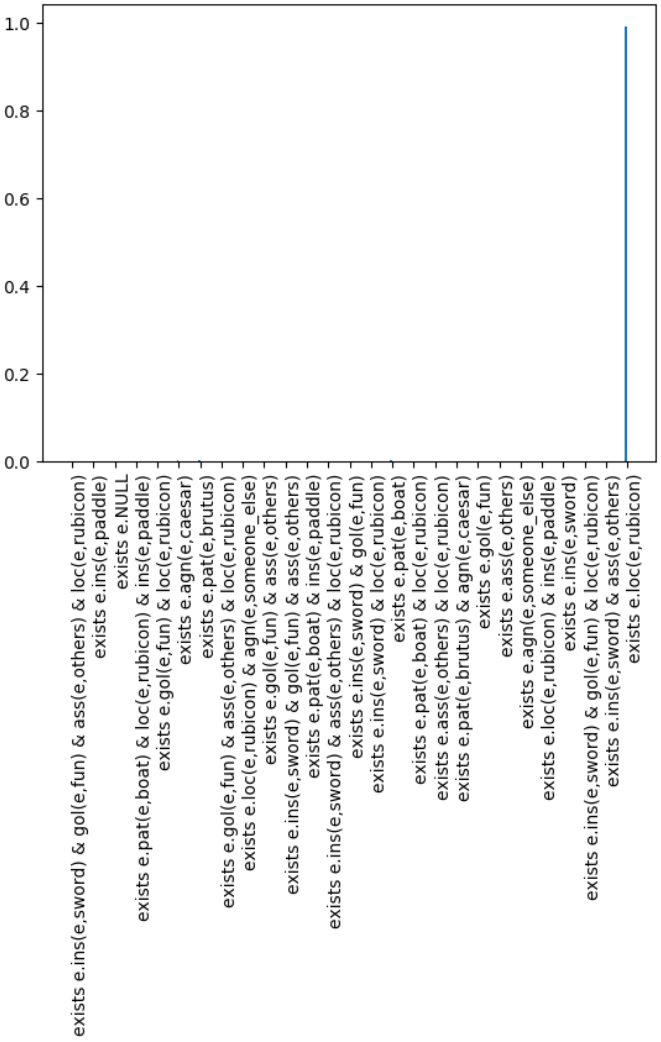

Figure 5: Pragmatic Speaker Distribution with high temperature.

A.1.3 Condensed Utterance Prior

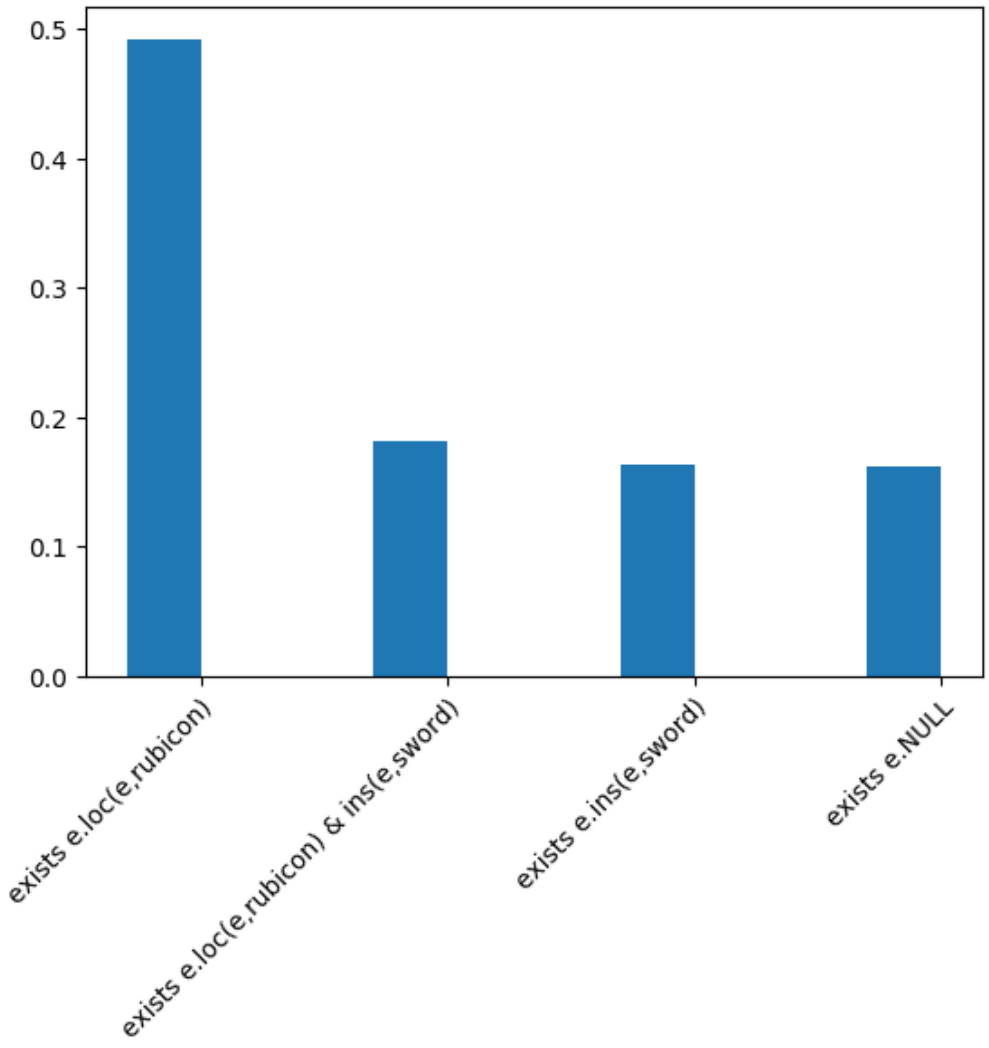

Figure 6: Pragmatic Speaker Distribution for a smaller utterance prior (4 instead of 27 bins).

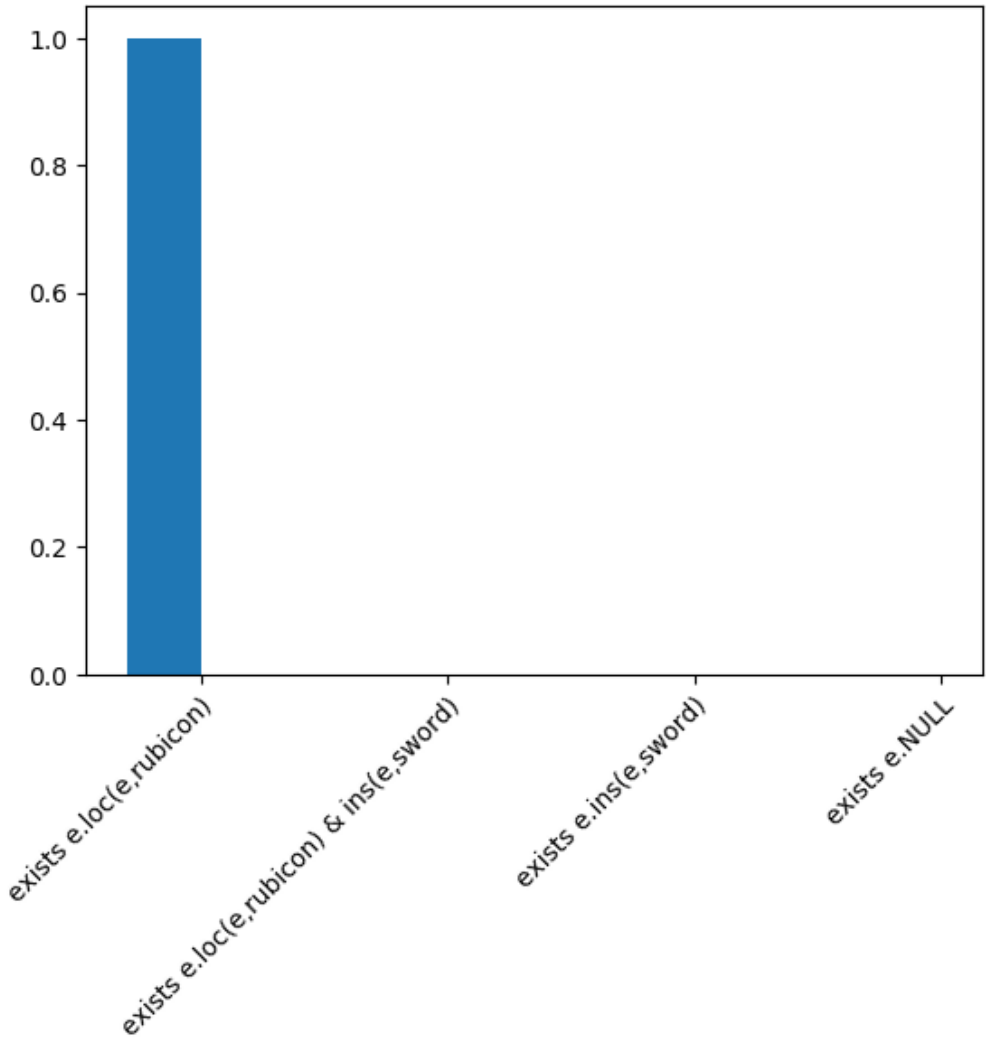

Figure 7: Pragmatic Speaker Distribution for a smaller utterance prior with high temperature.

A.2 Contrasting Example

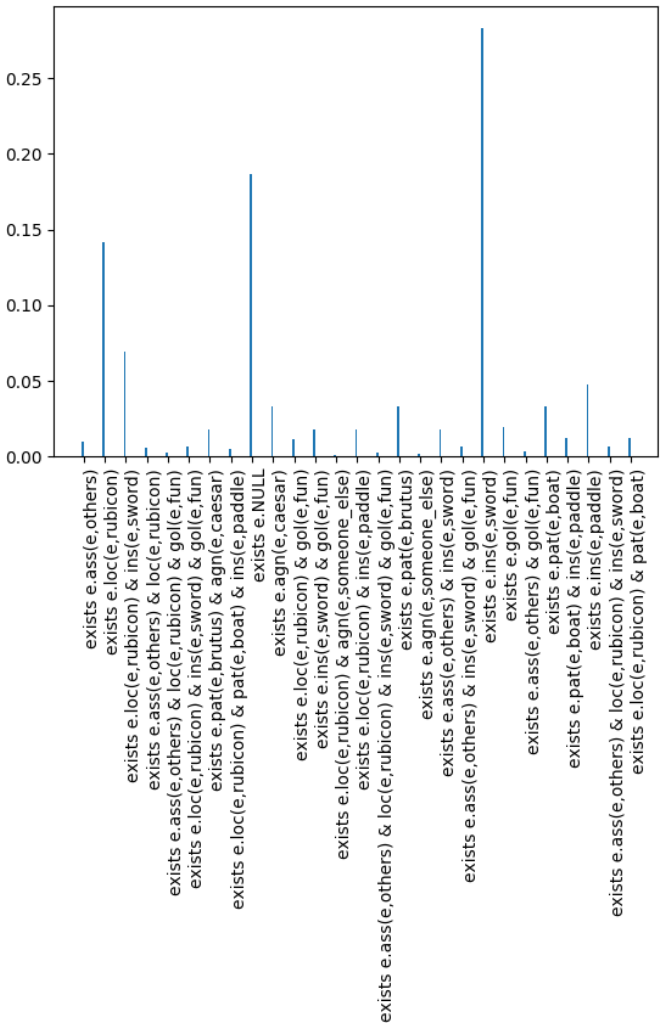

Figure 8: Pragmatic Speaker Distribution for standard L Belief but S_1 Belief: $\exists e \exists x[stab(Brutus, Caesar, e) \wedge AG(e, Brutus) \wedge PAT(e, Caesar) \wedge LOC(e, Forum)$ while attending to the QUD "Should Brutus be lashed?". It is interesting to note that the speaker lies here: They don't know the instrument used and therefore disprove the QUD, then calculate it is best to convince the listener by saying *"with a sword!"*

References

[Grice(1975)] H. P. Grice, Logic and conversation, Syntax and Semantics (1975) 41–58.

[Kao and Goodman(2014)] J. Kao, N. D. Goodman, Nonliteral understanding of number words, Proceedings of the National Academy of Sciences 111 (2014) 12002–12007.

[Frank and Goodman(2012)] M. C. Frank, N. D. Goodman, Predicting pragmatic reasoning in language games, Science 336 (2012) 998–998.

[Parsons(1994)] T. Parsons, Events in the semantics of english. a study in subatomic semantics. number 19 in current studies in linguistics (1994).

[Fillmore et al.(2006)] C. J. Fillmore, et al., Frame semantics, Cognitive linguistics: Basic readings 34 (2006) 373–400.

[Monroe and Potts(2015)] W. Monroe, C. Potts, Learning in the rational speech acts model, arXiv preprint arXiv:1510.06807 (2015).

[Kao et al.(2014)Kao, Bergen, and Goodman] J. Kao, L. Bergen, N. Goodman, Formalizing the pragmatics of metaphor understanding, in: Proceedings of the annual meeting of the Cognitive Science Society, volume 36.

[Bingham(2018)] Bingham, Pyro: Deep Universal Probabilistic Programming, arXiv preprint arXiv:1810.09538 (2018).

[Loper and Bird(2002)] E. Loper, S. Bird, Nltk: the natural language toolkit, arXiv preprint cs/0205028 (2002).

[Sutton et al.(1992)Sutton, Barto, and Williams] R. S. Sutton, A. G. Barto, R. J. Williams, Reinforcement learning is direct adaptive optimal control, IEEE Control Systems Magazine 12 (1992) 19–22.

[Scontras et al.(????)] G. Scontras, et al., Probabilistic Language understanding an introduction to the rational speech act framework, `http://www.problang.org/`, ???? Accessed: 2019-10-31.

[Martínez-Gómez et al.(2016)Martínez-Gómez, Mineshima, Miyao, and Bekki] P. Martínez-Gómez, K. Mineshima, Y. Miyao, D. Bekki, ccg2lambda: A compositional semantics system, in: Proceedings of ACL 2016 System Demonstrations, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 85–90.

[Hockenmaier and Steedman(????)] J. Hockenmaier, M. Steedman, CCGbank LDC2005T13 web download, Web Download. Philadelphia: Linguistic Data Consortium, 2005. `https://catalog.ldc.upenn.edu/LDC2005T13`, ???? Accessed: 2019-10-31.

[Eric and Manning(2017)] M. Eric, C. D. Manning, Key-value retrieval networks for task-oriented dialogue, arXiv preprint arXiv:1705.05414 (2017).