

**EPSRC**

Engineering and Physical Sciences  
Research Council



# Conceptual Pacts for Reference Resolution using Small, Dynamically Constructed Language Models: A Study in Puzzle Building Dialogues

**Julian Hough<sup>1</sup>**

**Sina Zarrieß<sup>2</sup>, Casey Kennington<sup>3</sup> David Schlangen<sup>4,5</sup> and Massimo Poesio<sup>6</sup>**

1 School of Mathematics and Computer Science, **Swansea University**

2 Faculty of Linguistics and Literature, **Bielefeld University**

3 Department of Computer Science, **Boise State University**

4 Department of Linguistics, **University of Potsdam**

5 German Research Center for Artificial Intelligence (**DFKI**), Berlin

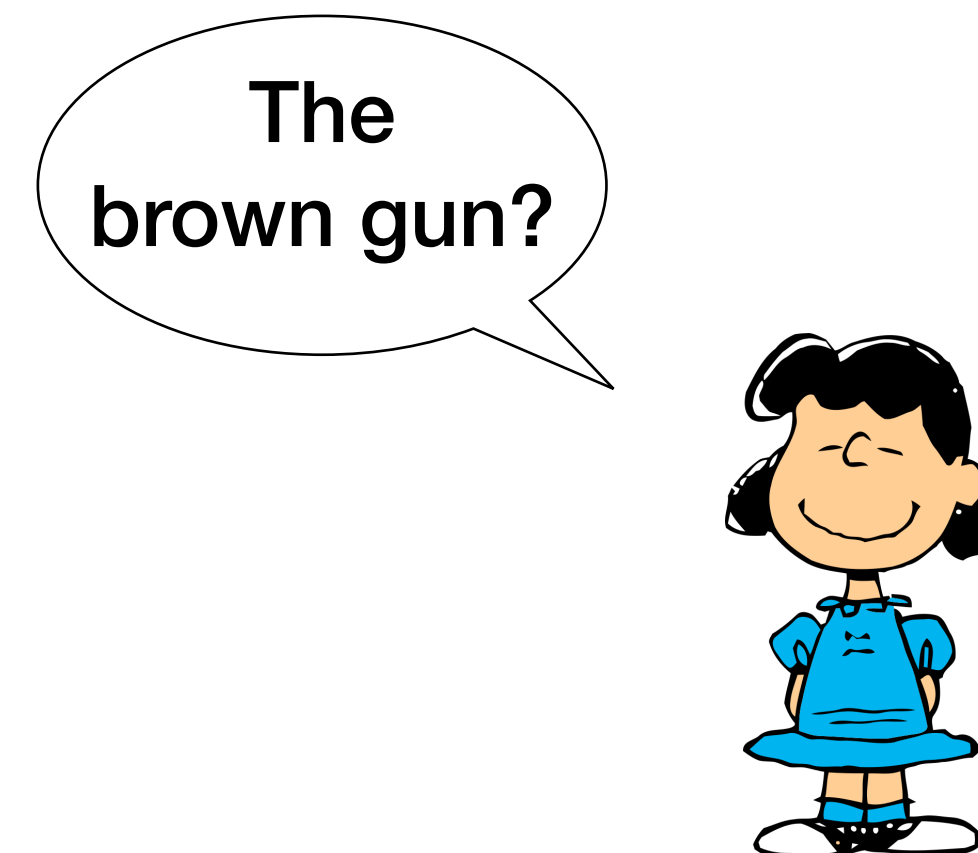
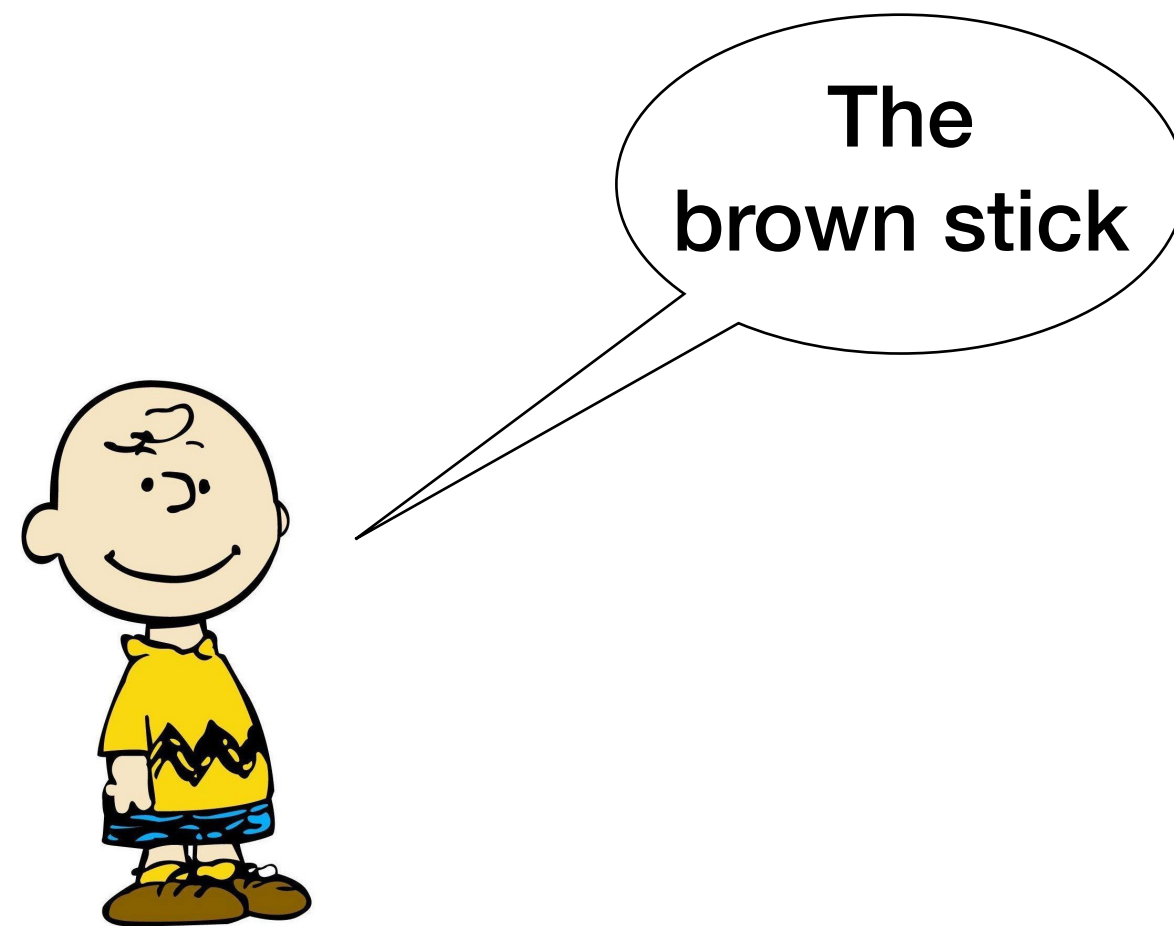
6 School of Electronic Engineering and Computer Science, **Queen Mary University of London**

*LREC-COLING 2024, Turin, Italy*

# Conceptual Pacts for referring to objects

Brennan and Clark (1996):

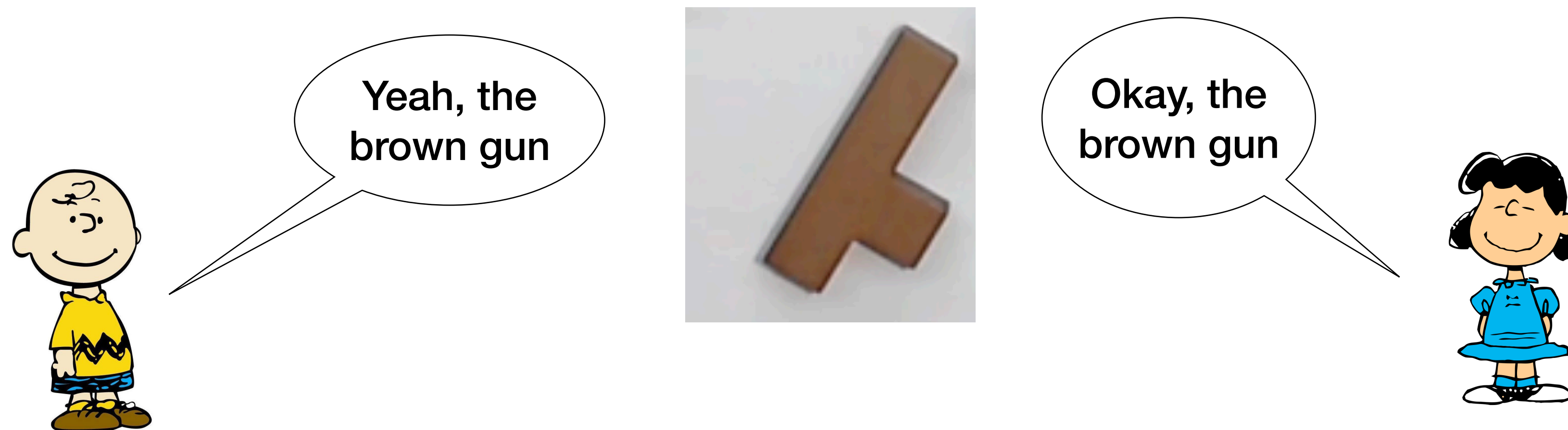
“When people in conversation refer repeatedly to the same object, they come to use the same terms.”



# Conceptual Pacts for referring to objects

Brennan and Clark (1996):

“When people in conversation refer repeatedly to the same object, they come to use the same terms.”



# Conceptual Pacts for referring to objects

Modern Large Language Models (LLMs) don't have this dynamic approach, rather relying on **large static data**.

In this paper we propose a **conceptual pact model for referring to objects** for pairs of participants using language models.

We use using **classical language models**, though it is applicable to LLMs too.

We want to show how its utility for both:

1. Cognitive modelling of **human-human interaction**
2. **In-robot dialogue systems** which may have very little training data.

# Data: PentoRef PentoCV

PentoRef (Zarrieß et al., 2016)'s **PENTO-CV** data:

8 conversational pairs referring to Pentomino puzzles pieces on a white playing surface.



# Data: PentoRef PentoCV

Two roles of **Instruction Giver (IG)** who can see a video feed and **Instruction Follower (IF)** who builds the puzzles. Phases of a game:

**1. Selection:** IF picks 3 puzzle pieces from the 12. IG looks up a final configuration/shape including at least those 3 pieces from a database

**2. Building:** IG instructs the IF to construct the shape (this paper)

This repeats several times for each pair and they swap roles approximately half-way through.

# Data: PentoRef PentoCV

Total number of references to pieces: 1899

Median references each pair makes to a piece: 19 (min=2, max=41)

**Can we see conceptual pacts at work for each pair, for each object?**

# Joint establishment of a pact, then substitution and stabilisation



| Speaker | End time (s) | Referring expression       | English translation    |
|---------|--------------|----------------------------|------------------------|
| B       | 167.5        | das hellblaue L            | the light blue L       |
| B       | 407.2        | das zweite L               | the second L           |
| B       | 454.0        | das oben rechts liegende L | the L at the top right |
| B       | 519.6        | das große L                | the big L              |
| B       | 785.7        | das L                      | the L                  |
| B       | 1083.5       | das element                | the part               |
| B       | 1101.9       | dem blauen                 | the blue one           |
| B       | 1233.6       | das blaue andere L         | the other blue L       |
| B       | 1283.8       | das blaue element          | the blue part          |
| A       | 1545.5       | das blaue element          | the blue part          |
| A       | 1626.0       | das blaue L                | the blue L             |
| A       | 1635.4       | das blaue L                | the blue L             |
| A       | 1646.4       | das blaue L                | the blue L             |
| A       | 1661.9       | das blaue L                | the blue L             |
| A       | 1853.9       | das blaue L                | the blue L             |
| A       | 1922.8       | das blaue L                | the blue L             |
| A       | 1970.9       | der blaue                  | the blue one           |
| A       | 2114.2       | das blaue L                | the blue L             |
| A       | 2296.3       | das L                      | the L                  |
| A       | 2297.3       | das blaue                  | the blue one           |
| A       | 2298.1       | das blaue L                | the blue L             |
| A       | 2546.6       | das blaue L                | the blue L             |
| A       | 2645.0       | das blaue L                | the blue L             |
| A       | 2806.5       | das blaue L                | the blue L             |
| A       | 2834.6       | das blaue L                | the blue L             |

# Joint establishment of a pact, then substitution and stabilisation



| English translation    |
|------------------------|
| the light blue L       |
| the second L           |
| the L at the top right |
| the big L              |
| the L                  |
| the part               |
| the blue one           |
| the other blue L       |
| the blue part          |
| the blue part          |
| the blue L             |
| the blue L             |
| the blue L             |
| the blue L             |
| the blue L             |
| the blue one           |
| the blue L             |
| the L                  |
| the blue one           |
| the blue L             |
| the blue L             |
| the blue L             |
| the blue L             |
| the blue L             |

# Joint establishment of a pact, then substitution and stabilisation



| English translation    |
|------------------------|
| the light blue L       |
| the second L           |
| the L at the top right |
| the big L              |
| the L                  |
| the part               |
| the blue one           |
| the other blue L       |
| the blue part          |
| the blue part          |
| the blue L             |
| the blue L             |
| the blue L             |
| the blue L             |
| the blue L             |
| the blue L             |
| the blue one           |
| the blue L             |
| the L                  |
| the blue one           |
| the blue L             |
| the blue L             |
| the blue L             |
| the blue L             |
| the blue L             |
| the blue L             |

Stabilisation point

# Different pairs form particular and different pacts



## Pair 1

| English translation |
|---------------------|
| the plus            |
| the red plus        |
| the plus            |
| plus                |
| plus                |
| the plus            |
| plus                |
| the plus            |
| the plus            |
| the plus            |
| the plus            |
| the plus            |
| the plus            |
| plus                |
| the plus            |

## Pair 2

| English translation |
|---------------------|
| the red cross       |
| the red cross       |
| the cross           |
| the cross           |
| the cross           |
| the cross           |
| the cross           |
| the cross           |
| the red cross       |
| the cross           |
| the cross           |
| the X               |
| the cross           |
| the X               |
| the cross           |
| the X               |
| the cross           |
| the X               |
| the cross           |
| the cross           |
| the cross           |
| cross               |
| the cross           |
| the X               |
| the cross           |
| the cross           |

# Model: one language model per pact

For one conversational pair, we assume they build **one pact per object as a language model** i.e. the likelihood a given string will be used to refer to that object:



$$p_X^{pact}(w_0..w_n)$$



$$p_L^{pact}(w_0..w_n)$$

Etc.

These pact models **begin empty**, with no vocabulary, then build up with each reference to the piece over the whole interaction. The likelihood should go up, or, **uncertainty should go down** for each new reference given the experience is growing with each mention if pacts are being established.

In this paper, we use **simple n-gram language models** with Laplace smoothing.

# Model: incorporating prior experience

For **reference resolution**, we want to model the effect of **prior experience** of references to the object before the current pact began, and mix that with the pact model. In a Bayesian model:

$$\arg \max_{r \in refs} p_r^{ex}(w_0..w_n) + \lambda p_r^{pact}(w_0..w_n) \cdot p(r)$$

Experience model of object r      Pact model of object r (initially empty)

Weight of the pact model      Prior of referring to object

# Model: adding dynamic increasing weight of the pact model

$$\arg \max_{r \in refs} p_r^{ex}(w_0..w_n) + \lambda p_r^{pact}(w_0..w_n) \cdot p(r)$$

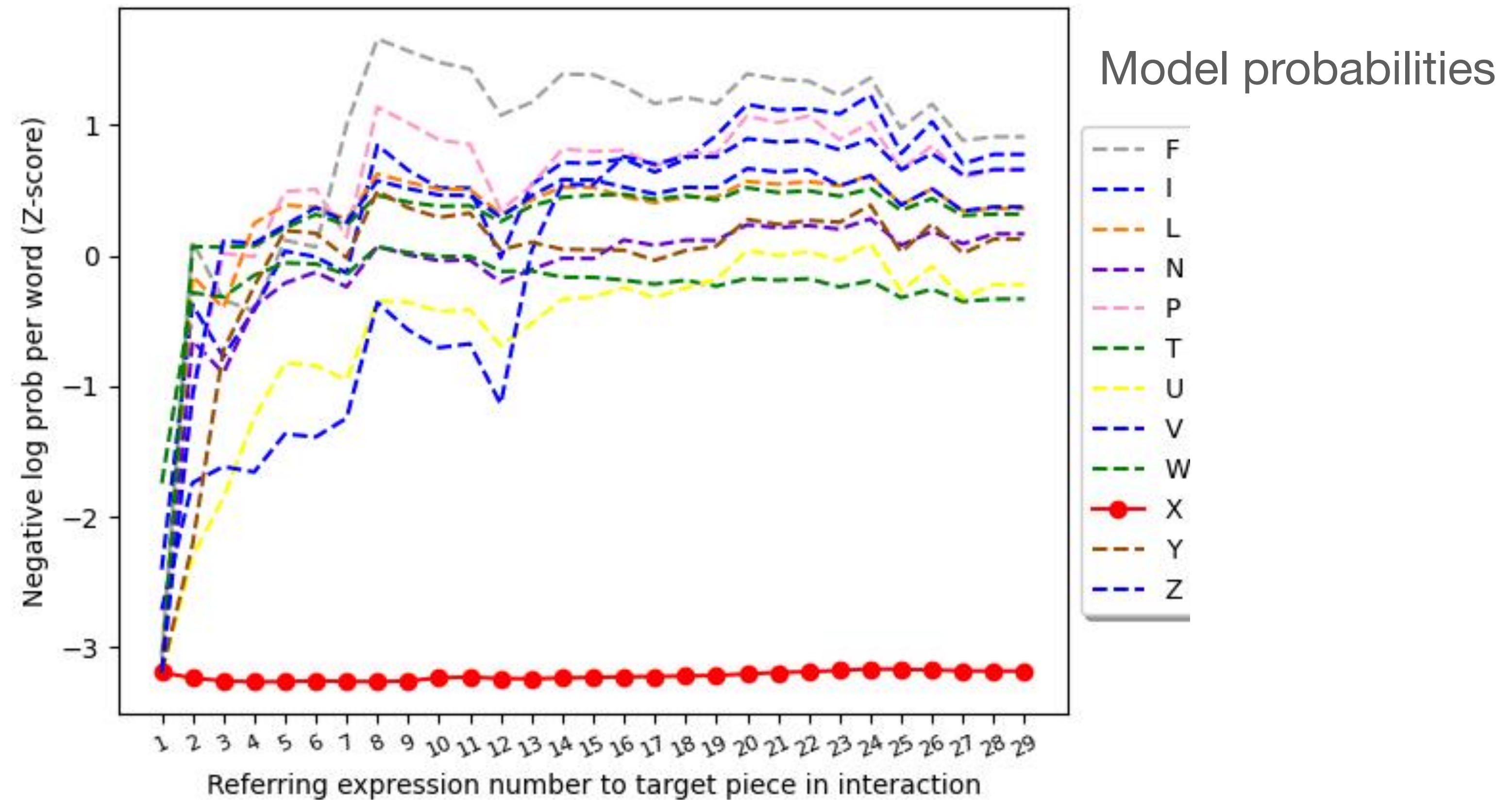
Weight of the pact model

Make this weight **dynamic**: initially 0 when the pact model is empty, then increasing it up to a **final weight** gradually up to a number of references **stable**

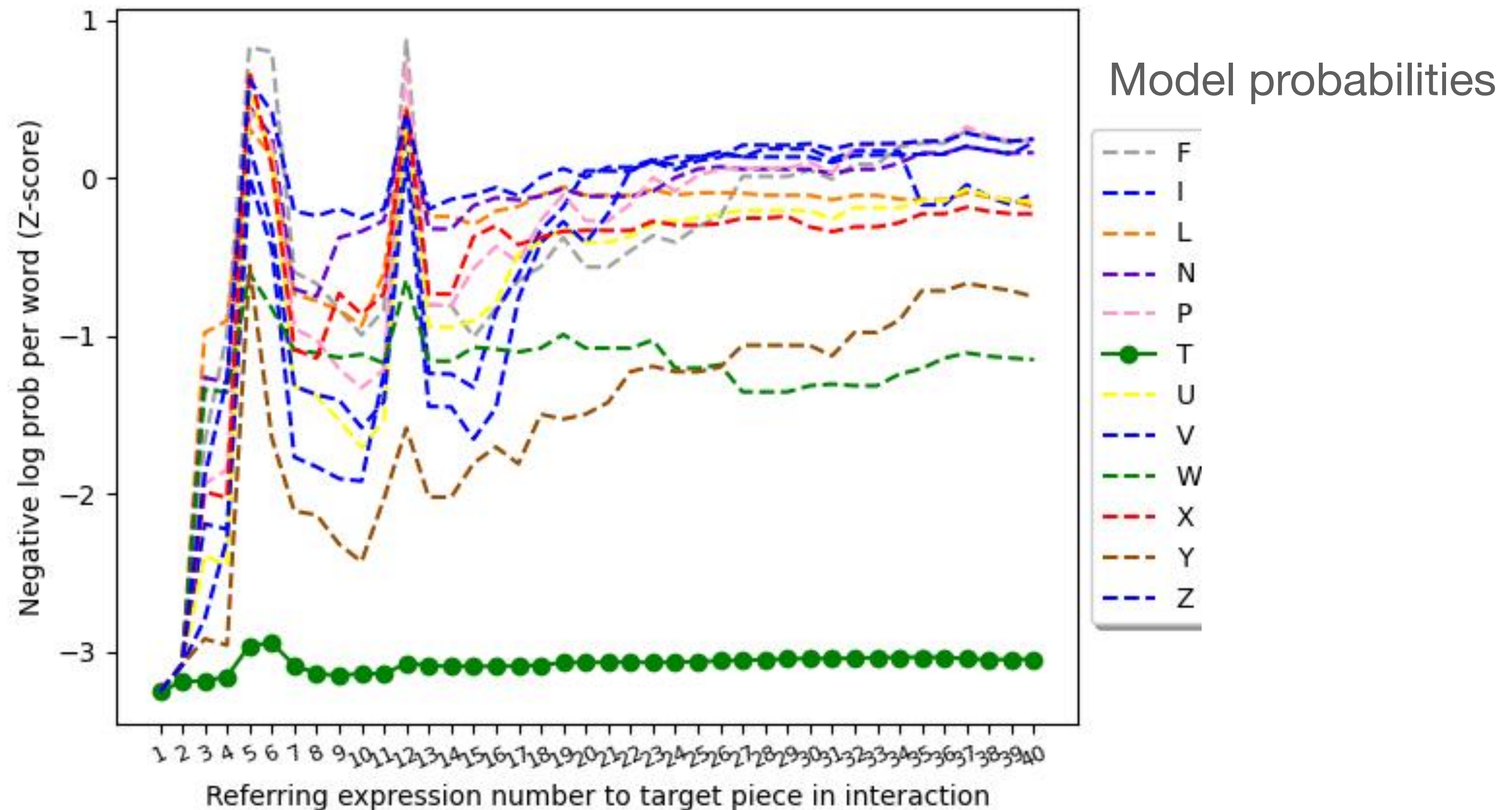
$$\lambda_r(o_r) = \lambda - \left( \frac{1}{stable} \cdot \max(stable - o_r, 0) \cdot \lambda \right)$$

$$(1 - \lambda_r(o_r)) p_r^{ex}(w_0..w_n) + \lambda_r(o_r) p_r^{pact}(w_0..w_n)$$

# Model: how do the pact models estimate the probability of different referring expressions over a conversation?

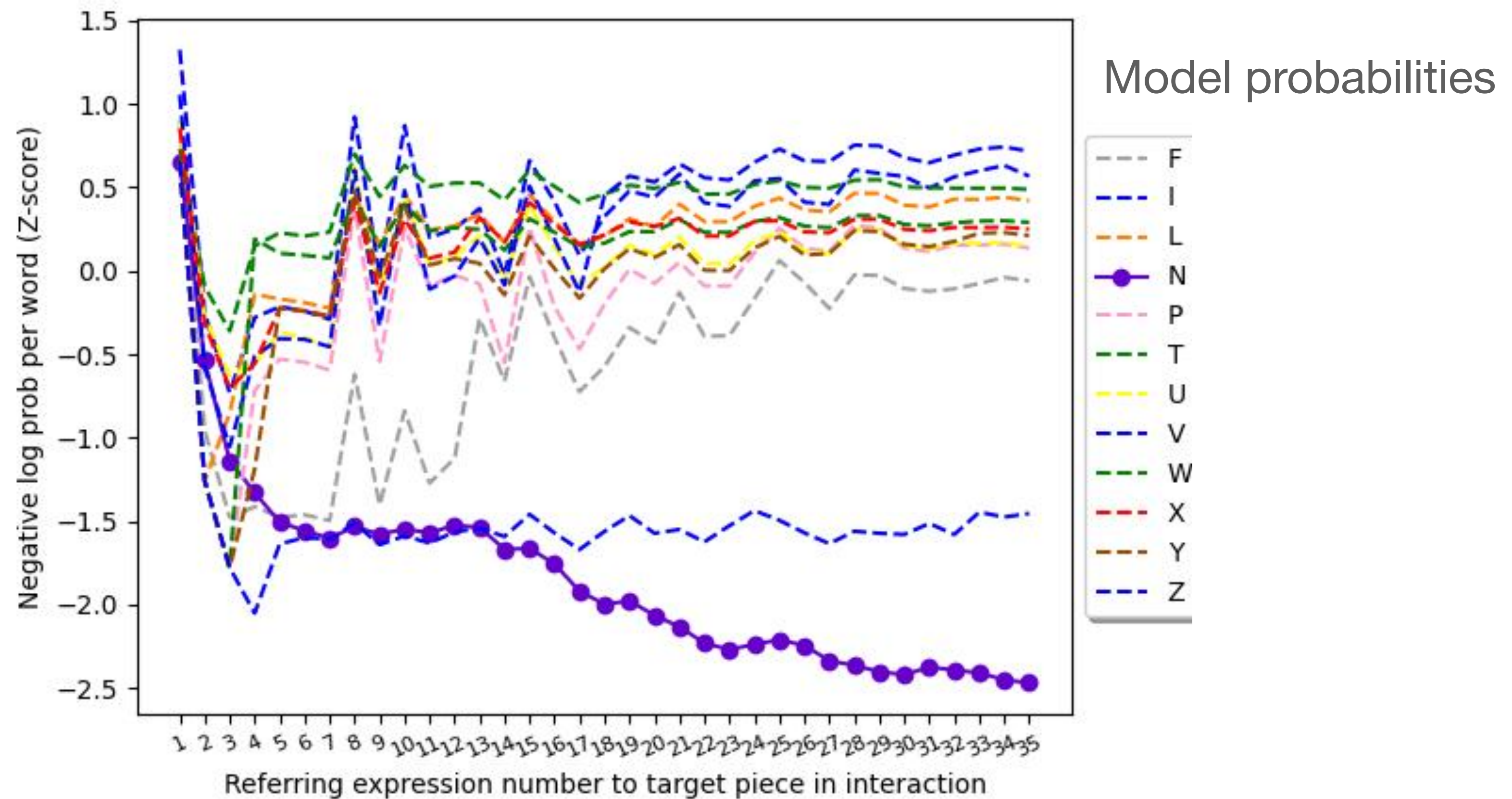


# Model: how do the pact models estimate the probability of different referring expressions over a conversation?





# Model: how do the pact models estimate the probability of different referring expressions over a conversation?



# Model: how do the pact models estimate the probability of different referring expressions over a conversation?

For some objects, the pact model distinguishes the object **very clearly from the others early on** - the target object has high probability in its corresponding model, but low probability in the other models.

For objects which are more difficult to describe uniquely from the outset, it **takes longer for a clearer** separation of these probabilities.

In either case, the pact models eventually distinguish the different pieces over time - **a pact stabilises with consistent referring expressions** emerging for each model resulting in relatively low entropies eventually.

# Reference Resolution experiments

We add the probabilities from all the past models for each reference to features of a simple **dynamic reference resolution model**.

Model is a Support Vector Machine (SVM) using the words of the current reference in addition to these probabilities.

We compare the model to two competitors:

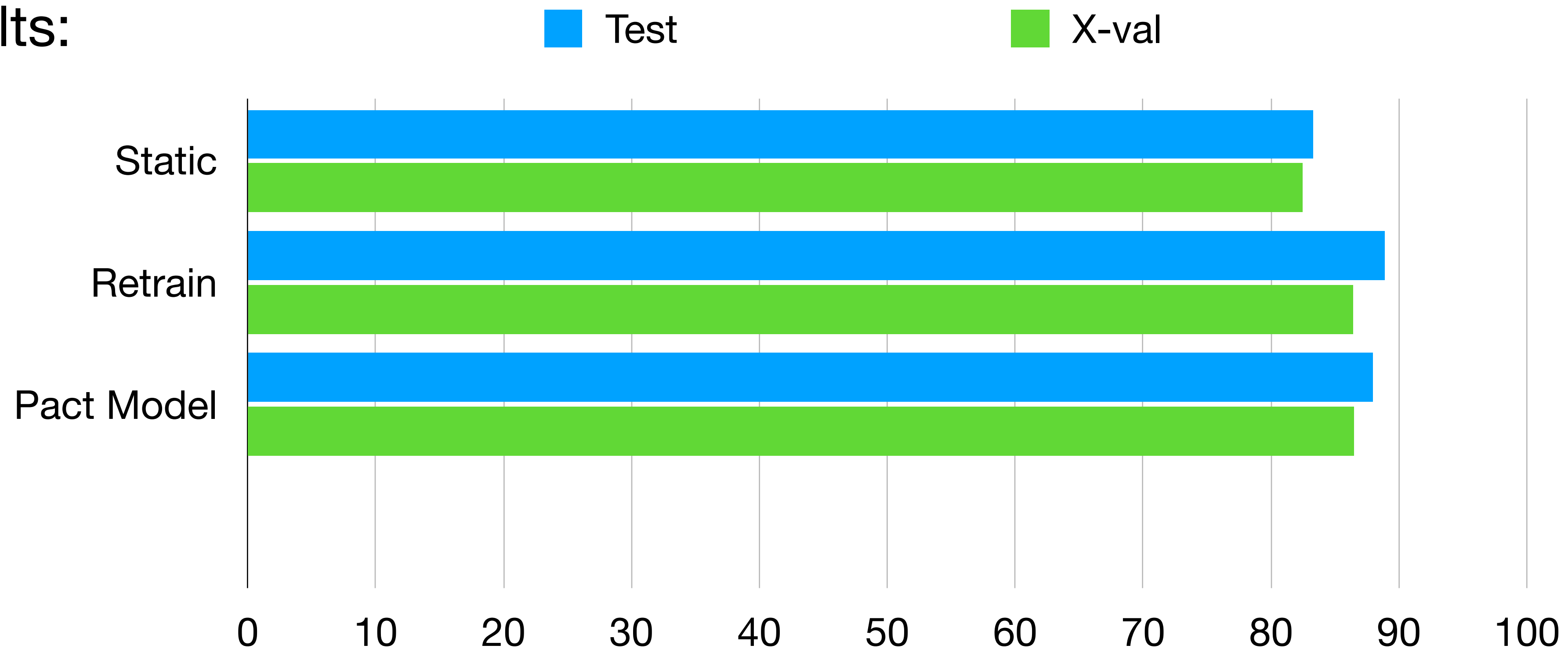
1. **Static baseline:** SVM model only using the lexical features, trained on the experience model's data
2. **Retrain:** SVM using lexical features only, exhaustively retraining after each mention, adding the current example to its training data

We test in two set-ups:

- 1) **X-val:** 7-fold cross-validation (leave-one-pair-out)
- 2) **Test:** training on those 7 pairs and testing on the 8th pair

# Reference Resolution experiment 1

Results:



In both X-val and Test set-ups, **the Pact Model outperforms the Static baseline**

It performs the same as the Retrain model, but is **3 times faster**.

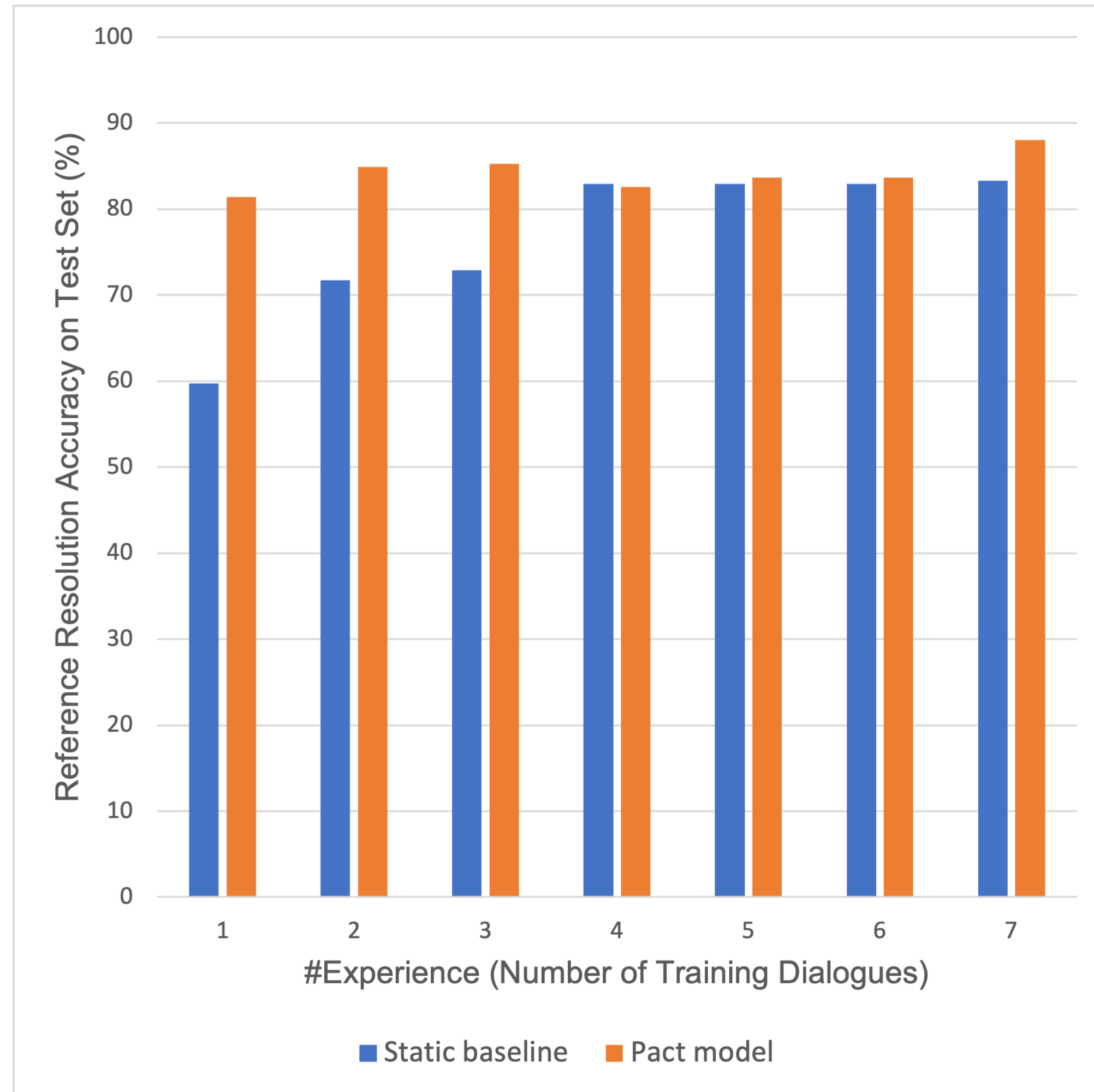
# Reference Resolution experiment 2

## Investigating the effect of limiting prior experience

We investigate how the effect of limiting the number of training dialogues in the experience model affects the performance of the model, including limiting experience to just **one previous interaction**.

Performance will go down with less data, but how well can **relying on the local pact** do overall in the course of the conversation with very little prior training data?

# Reference Resolution experiment 2



Using **just one prior conversation** in the experience model, reference resolution reaches **81.4%** accuracy with the pact model, **outperforming the 59.7% for the static model**. Faster convergence.

The exhaustive retraining method outperforms our model overall, though the models are at parity for X-val for using 3 or more prior dialogues.

# Discussion

In a simple puzzle building domain, probability values from a simple conceptual pact model with updating small language models (one per object), improve the performance of a comparable static reference resolution model without them.

Our model is **at parity with a far more resource-intensive** exhaustive retraining model in a number of settings, and is also **completely transparent**.

This shows potential for **generalisation to more reference domains**.

# Discussion

These results adds to work in **embodied reference resolution** that is cognitively motivated (Steels and Vogt, 1997).

The work supports the idea that a full **dynamic interactive model** is needed for live interactions and a challenge is given to LLM models to adapt this kind of model.

If models only rely on static knowledge, even if using very large amounts of it, they will ultimately be limited.

**EPSRC**

Engineering and Physical Sciences  
Research Council



**Thank you!**

# Conceptual Pacts for referring to objects

Brennan and Clark (1996) claim these **conceptual pacts** have the following properties:

1. In conversation, conceptual pacts are established by **speakers and addressees jointly**.
2. When speakers propose a conceptualization, they often **mark it for how confident** they are that it will be understood and adopted by their addressees.
3. People do not establish conceptual pacts all at once, but often **little by little**. The more firmly two partners establish a pact, the more likely they are to appeal to it later (and to appeal to it with confidence).
4. Conceptual pacts are **accessible** to both **speakers and addressees**.
5. Speakers form conceptual pacts with **particular** addressees.

Our model  
(This paper)



(Future work)

