Colorless Green Recurrent Networks Dream Hierarchically

Kristina Gulordava

Piotr Bojanowski　Edouard Grave

Tal Linzen　Marco Baroni

Do RNN language models learn syntax?

Do RNN language models learn syntax?

Linzen et al. 2016 (TACL) studied performance on difficult agreement constructions:

the **dogs** playing in my <u>neighbor</u>'s <u>garden</u> ...

barks ? **bark**

✗ ✔

To evaluate language models on binary prediction task:

P(barks | … neighbor's garden) <> P(bark | … neighbor's garden)

Do RNN language models learn syntax?

the **dogs** playing in my neighbor's garden ...

barks ? **bark**

Problem: the model can rely on frequency/semantic cues

dogs+bark **garden+barks**

✔ ✗

What does it mean to "learn syntax"?

Chomsky (1957):

grammar is independent from semantics and language use (frequencies)

"colorless green ideas sleep furiously"

We can say what is grammatical and what is not, even if nonsensical:

✔ colorless green ideas sleep furiously

✗ colorless green ideas sleep**s** furiously

Do RNN language models learn syntax?

the **dogs** playing in my <u>neighbor</u>'s <u>garden</u> ...

↓

the **ideas** rowing in my <u>lamp</u>'s <u>economy</u> ...

sleeps ? **sleep**

✗ ✔

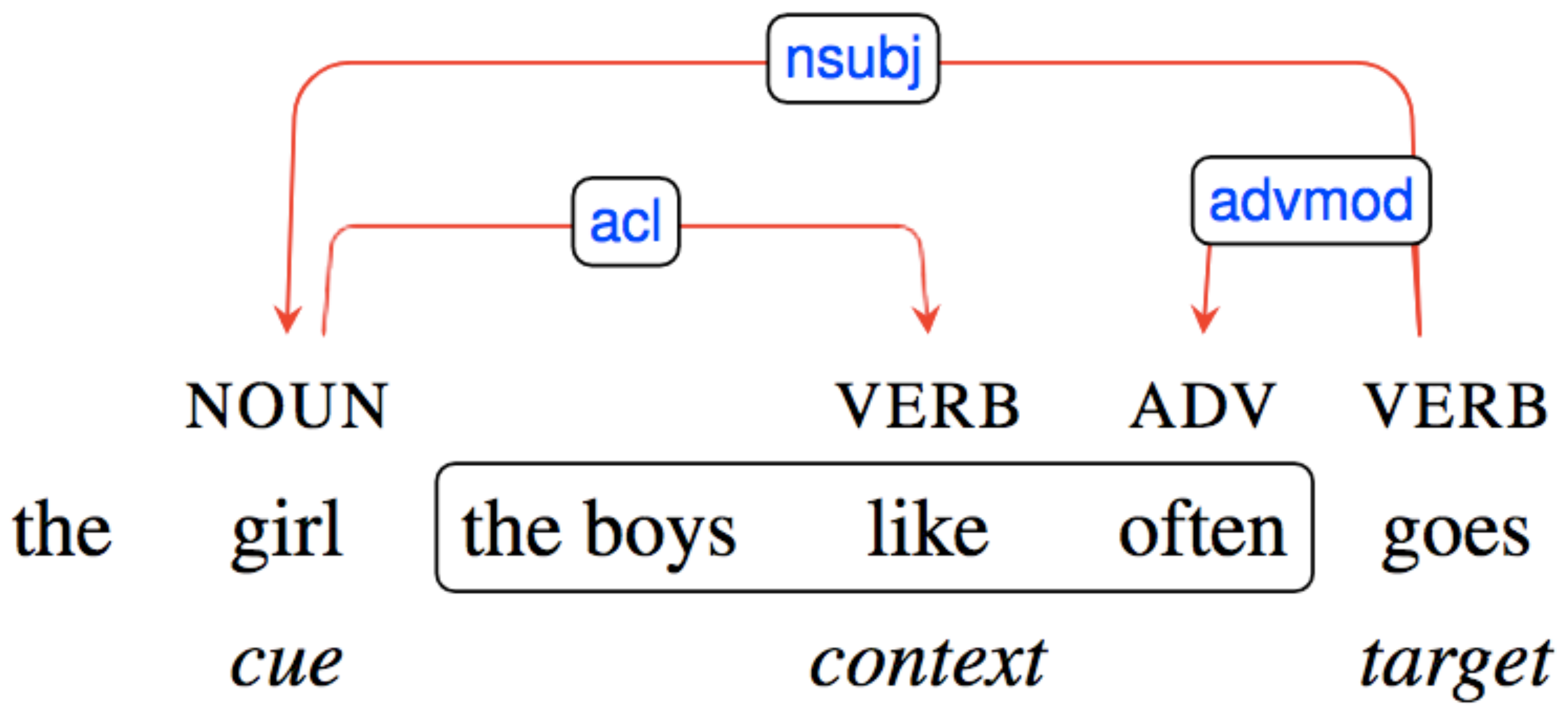
This work

Evaluates RNN language models on:

1. "colorless green" sentences
2. varied agreement constructions, harvested automatically
3. multiple languages: English, Italian, Hebrew, Russian

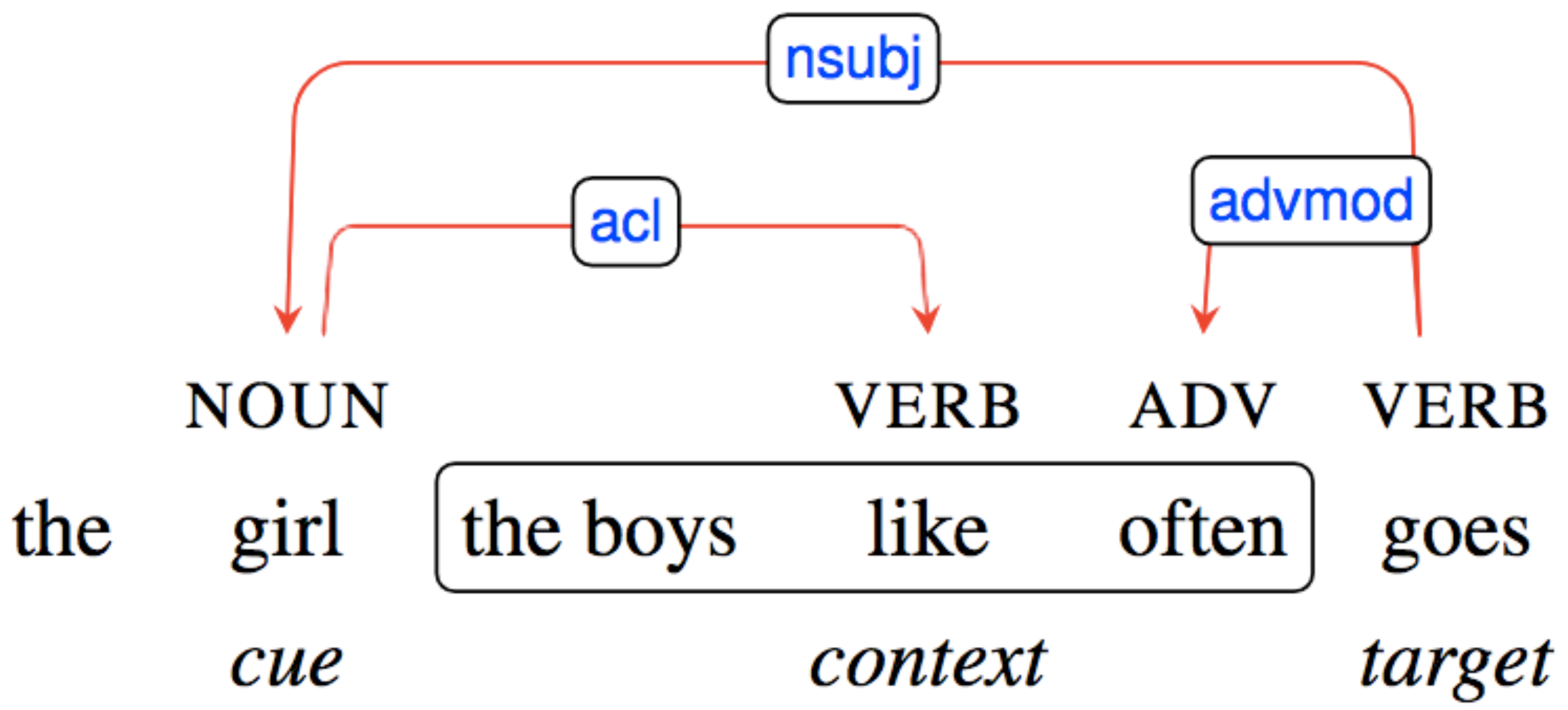
+ comparison with human performance

Extracting agreement constructions



- extract automatically from treebanks with dependency grammar annotation
- only **number agreement** (plural vs singular) but in many constructions
- **cue** and **target** in a dependency relation, sharing number feature

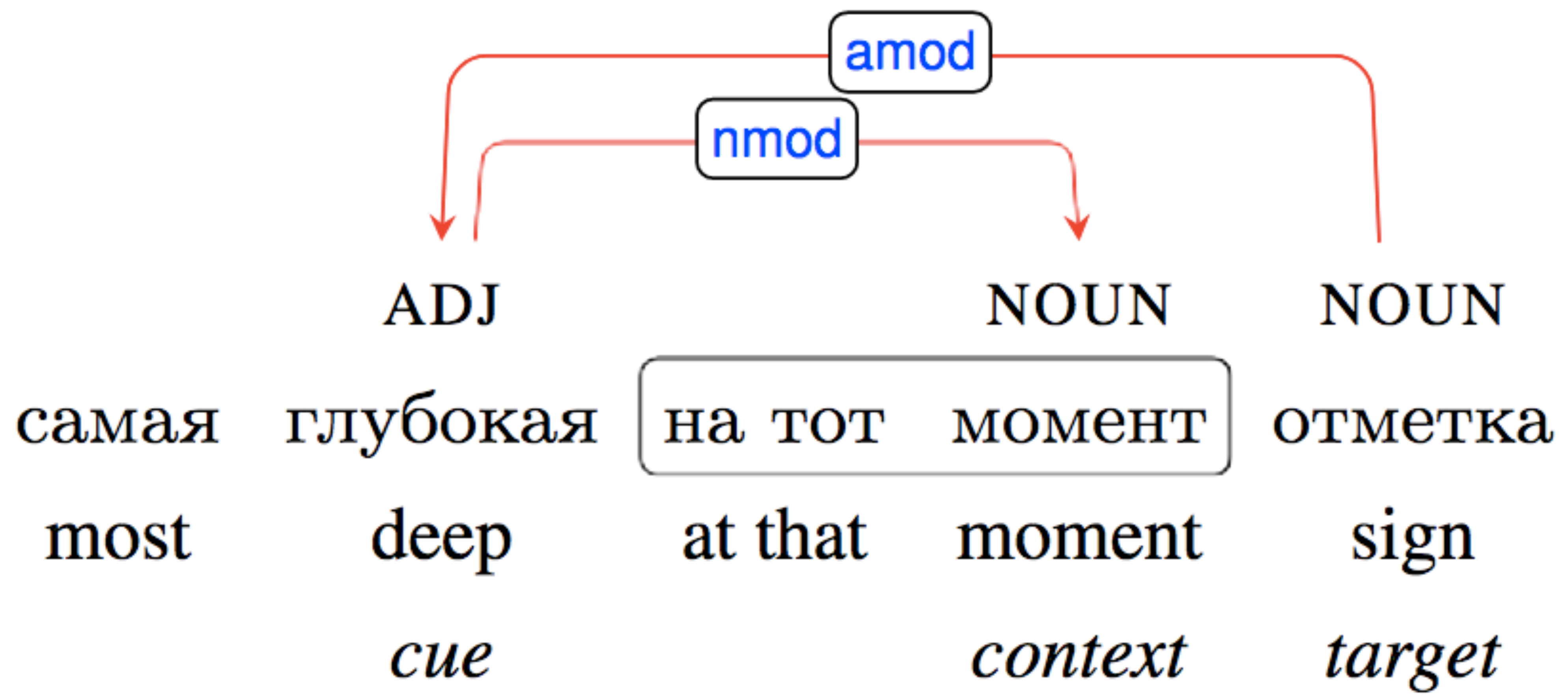
Extracting agreement constructions



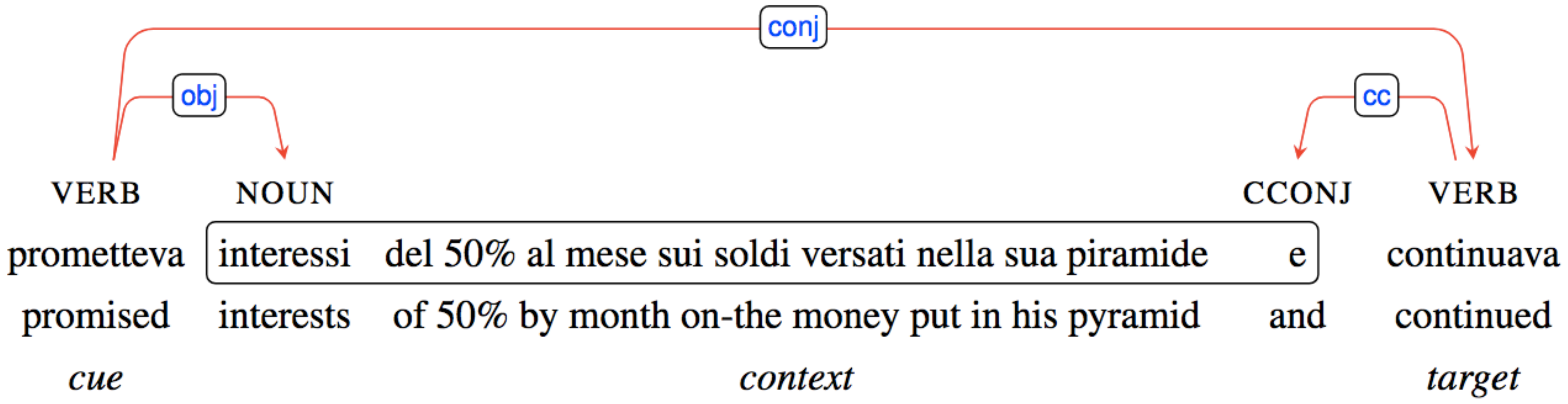
- agreement construction = (POS1 : POS2 : context), (NOUN : VERB : VERB ADV)
- # words in context ≥ 3 to ensure long-distance relations

Extracting agreement constructions

- varied constructions



- long dependencies



Extracting agreement constructions

	English	Italian	Hebrew	Russian
# distinct constructions (POS1 : POS2 : context)	2	8	18	21
# unique treebank sentences	41	119	373	442

Generating colorless green sentences

It **presents** the case for marriage equality and **states** ...
cue *target*

It **kills** the shuttle for honesty insurance and **finds** ...
cue *target*

- randomly subtitute content words, preserving POS and **morphology**
- for each extracted sentence in the treebank (*original*), generate 9 sentences (*nonce*)

Evaluation

- Compute LM probabilities *P(target singular | prefix), P(target plural | prefix)*

P(finds | … kills the shuttle for honesty insurance and) *0.0001*

P(find | … kills the shuttle for honesty insurance and) *0.0002*

- Following Linzen et al. 2016, we report **accuracy** over all sentences assuming that a model is correct if *P(correct target | prefix) > P (wrong target | prefix)*

Experiments

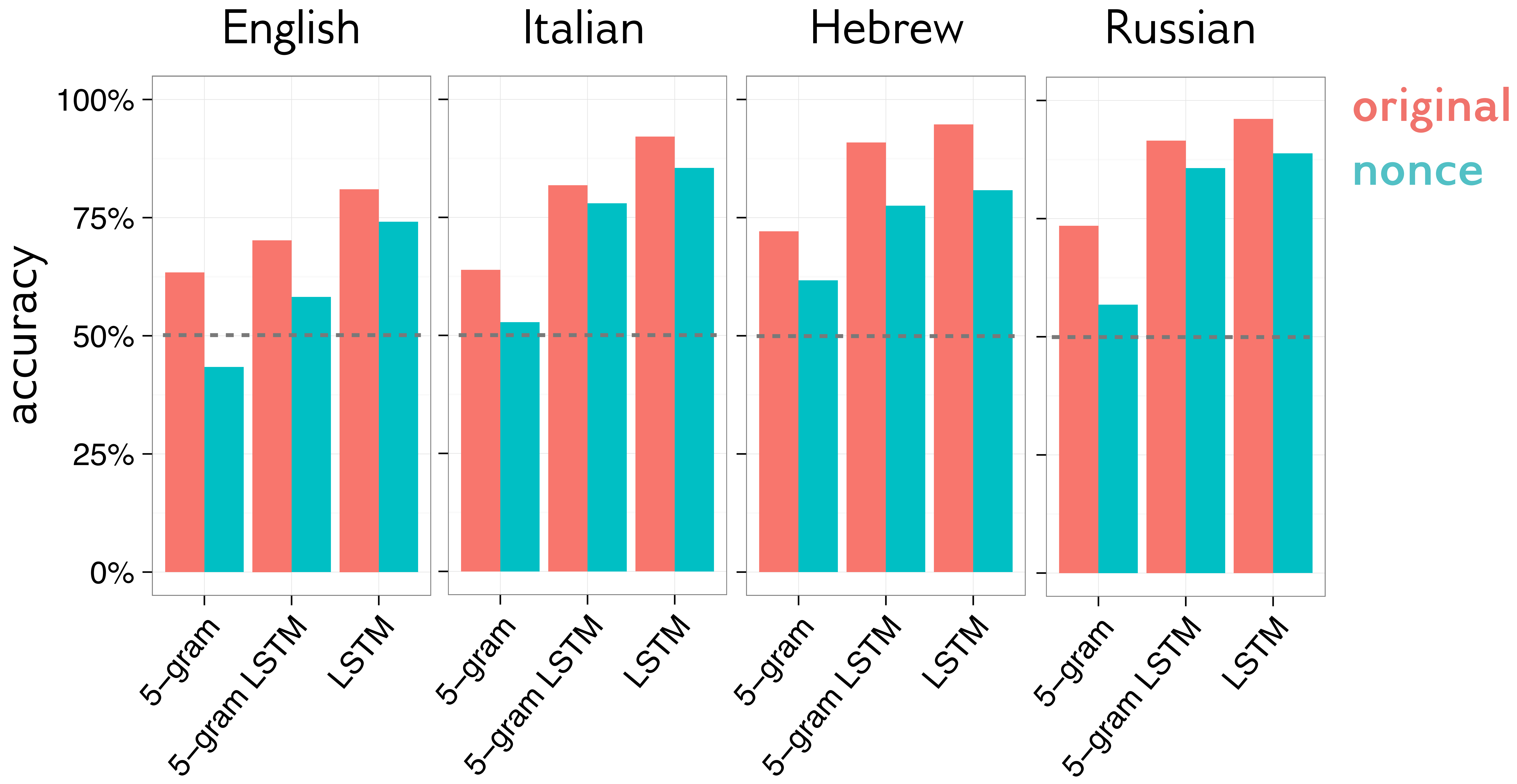
- LM training corpus: 80M words Wikipedia, 50K vocabulary
- LSTMs trained with (word-level) LM objective
  - 2 layers, 650 hidden, embedding units
  - we chose best models in a hyperparameter search based **on LM validation perplexity**

data and code: https://github.com/facebookresearch/colorlessgreenRNNs

Baselines

- 5-gram count-based LM (with Kneser-Ney smoothing)
  - how much do n-gram statistics capture, especially for nonce sentences?
- 5-gram LSTM-based LM
  - is longer context needed for our dataset?

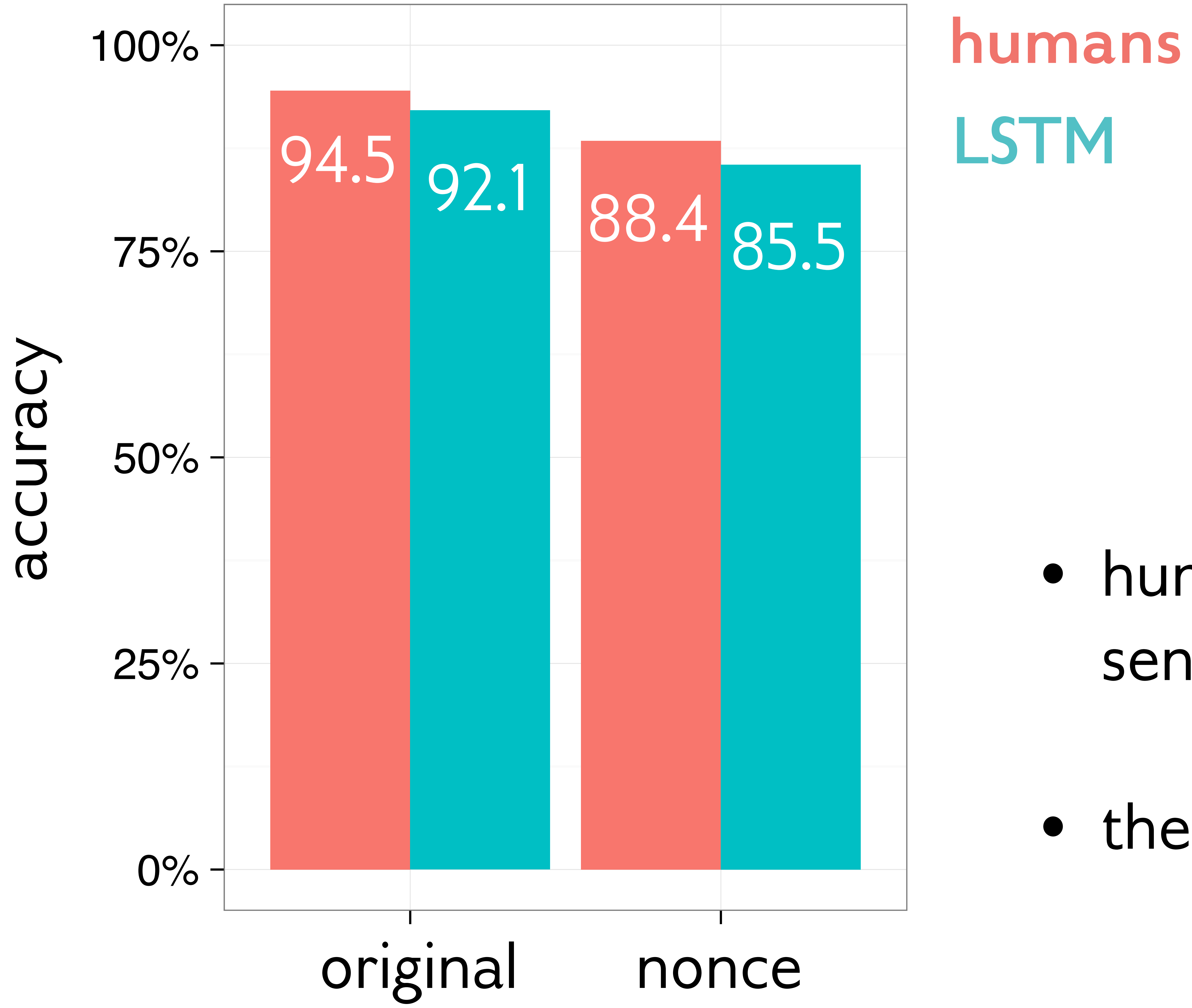
LSTM LMs vs Baselines



Evaluating human performance

- Human evaluation on **Italian** data
- MTurkers did the same binary choice task as LMs
- For each sentence (original and nonce), we collected minimum 5 judgements, 9.5 on average

Comparison with human performance in Italian



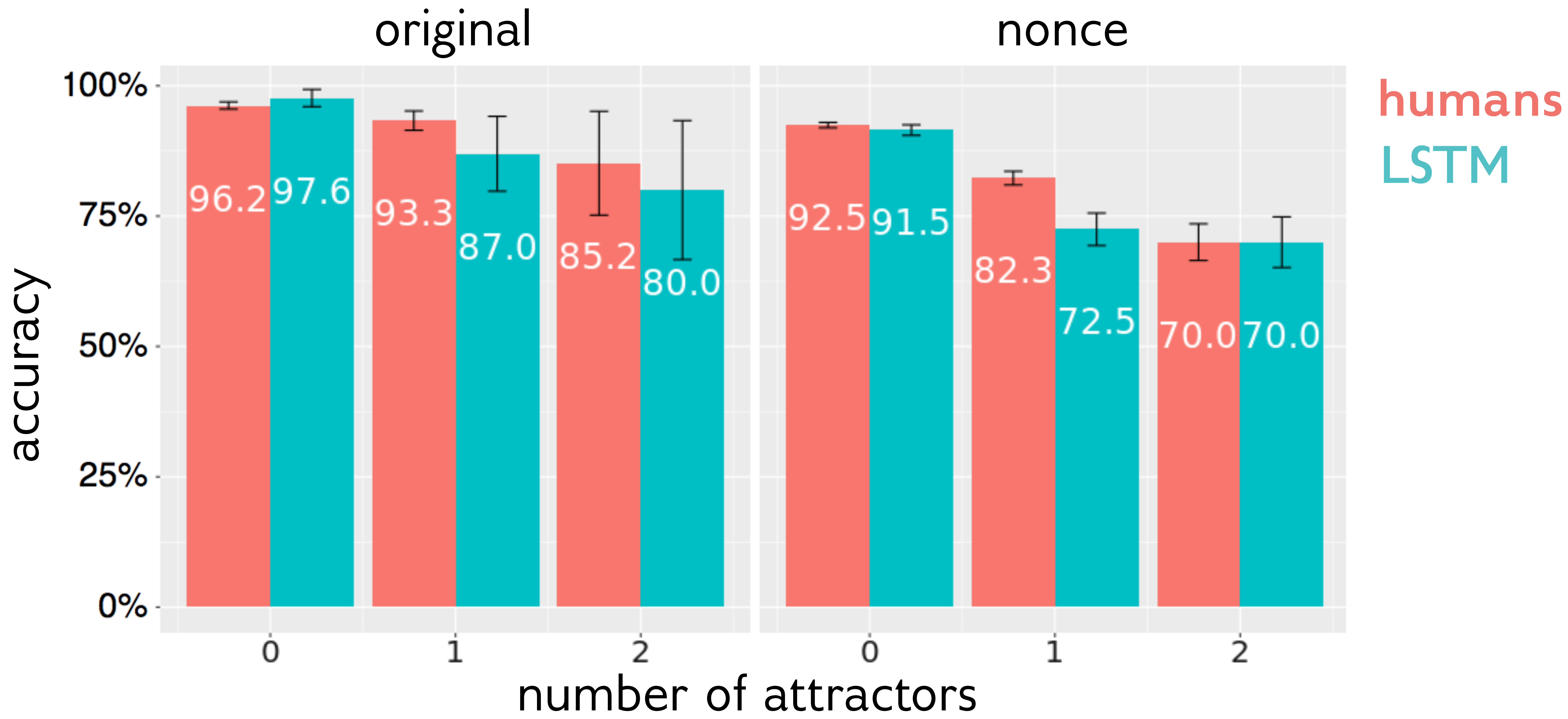
- humans make more mistakes in nonce sentences too!
- the gap in the two conditions is comparable

Performance in the presence of attractors

attractor = the same POS as cue, but different number

the **ideas** rowing in my <u>lamp</u>'s <u>economy</u> **sleep** 2 attractors

Performance in the presence of attractors

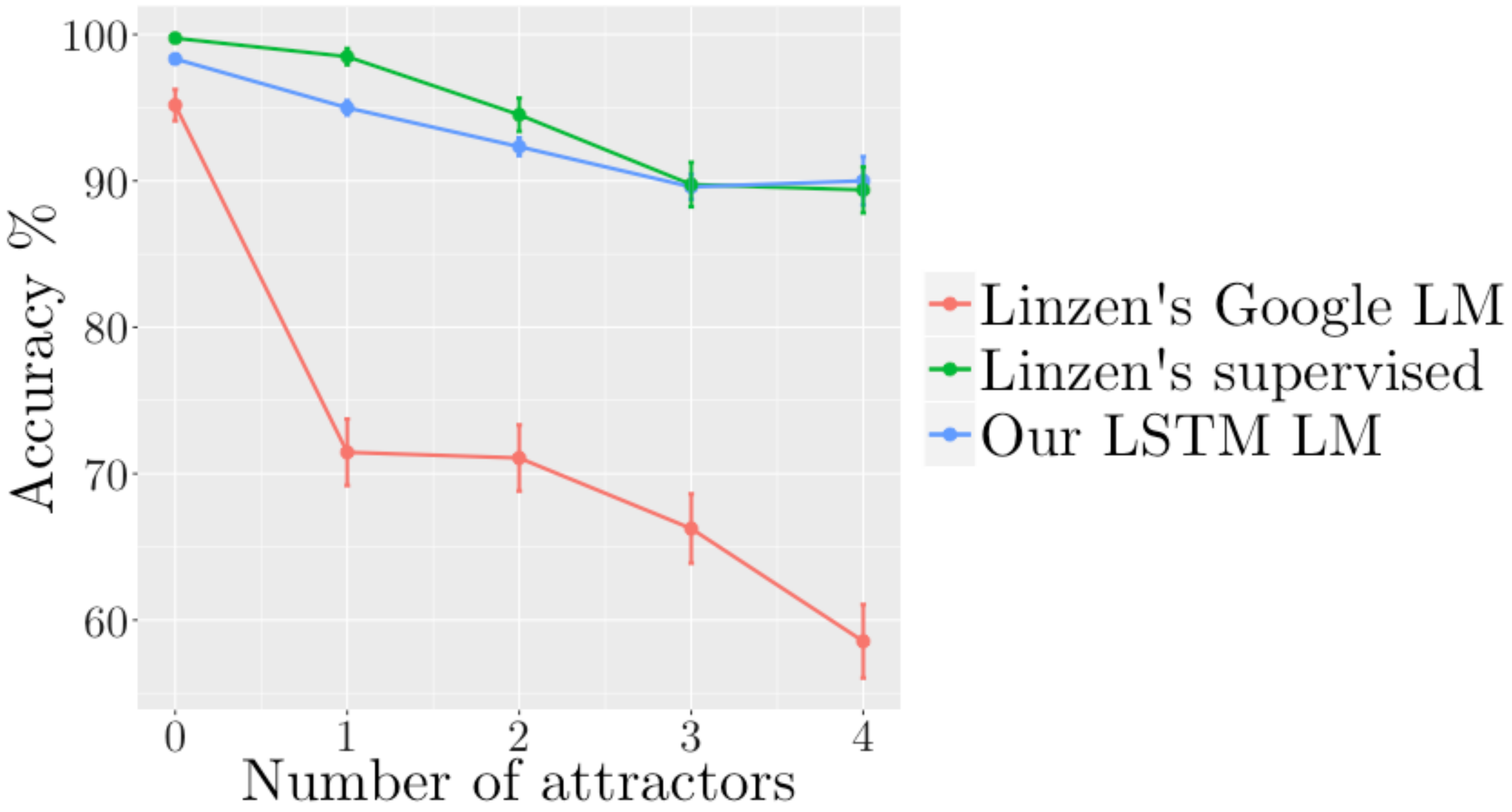


the **ideas** rowing in my <u>lamp</u>'s <u>economy</u> **sleep** 2 attractors

Conclusions

- LSTMs trained as language models capture abstract syntactic relations
  - not just (long-distance) collocation patterns or semantic associations
- LSTM architecture is better than other RNNs (e.g., simple RNN)
  - how LSTMs encode hierarchical information?
  - our data can be used to further analyse and compare RNNs

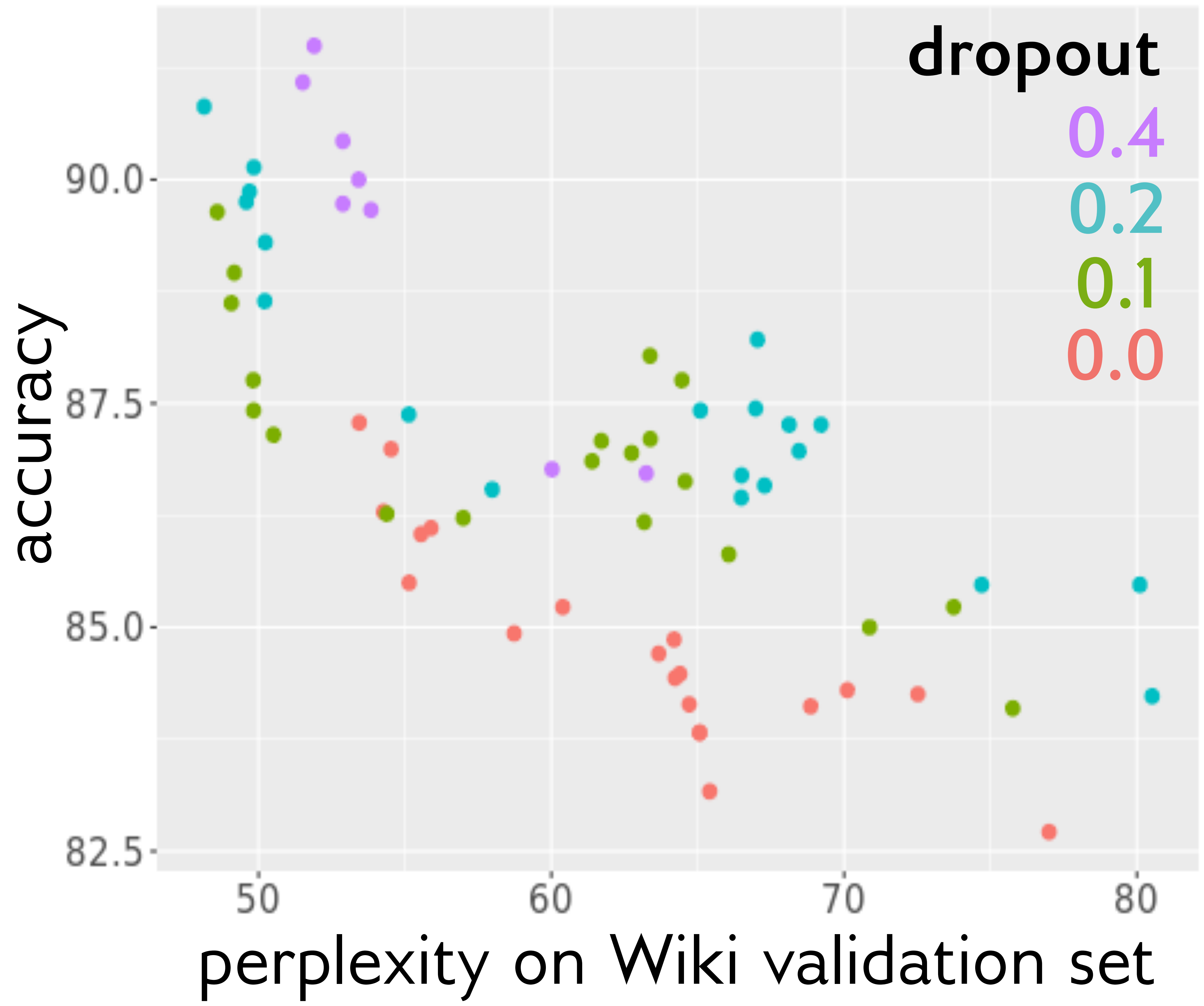
Comparison with Linzen et al.



Ambiguous sentences

- English: if you **have** any questions or **need**/needs
- Italian: **orto** di regolamenti davvero **pedonale**/i
  - orchard of rules truly pedestrian
  - “truly pedestrian orchard of rules”

Accuracy vs Perplexity



good correlation ppl ~ accuracy
but still a lot of variation

regularization helps learning
abstract features

Common constructions

		N V V	V NP conj V
Italian	Original	$93.3_{\pm 4.1}$	$83.3_{\pm 10.4}$
	Nonce	$92.5_{\pm 2.1}$	$78.5_{\pm 1.7}$
English	Original	$89.6_{\pm 3.6}$	$67.5_{\pm 5.2}$
	Nonce	$68.7_{\pm 0.9}$	$82.5_{\pm 4.8}$
Hebrew	Original	$86.7_{\pm 9.3}$	$83.3_{\pm 5.9}$
	Nonce	$65.7_{\pm 4.1}$	$83.1_{\pm 2.8}$
Russian	Original	-	$95.2_{\pm 1.9}$
	Nonce	-	$86.7_{\pm 1.6}$