Predicting cloze task results with language models

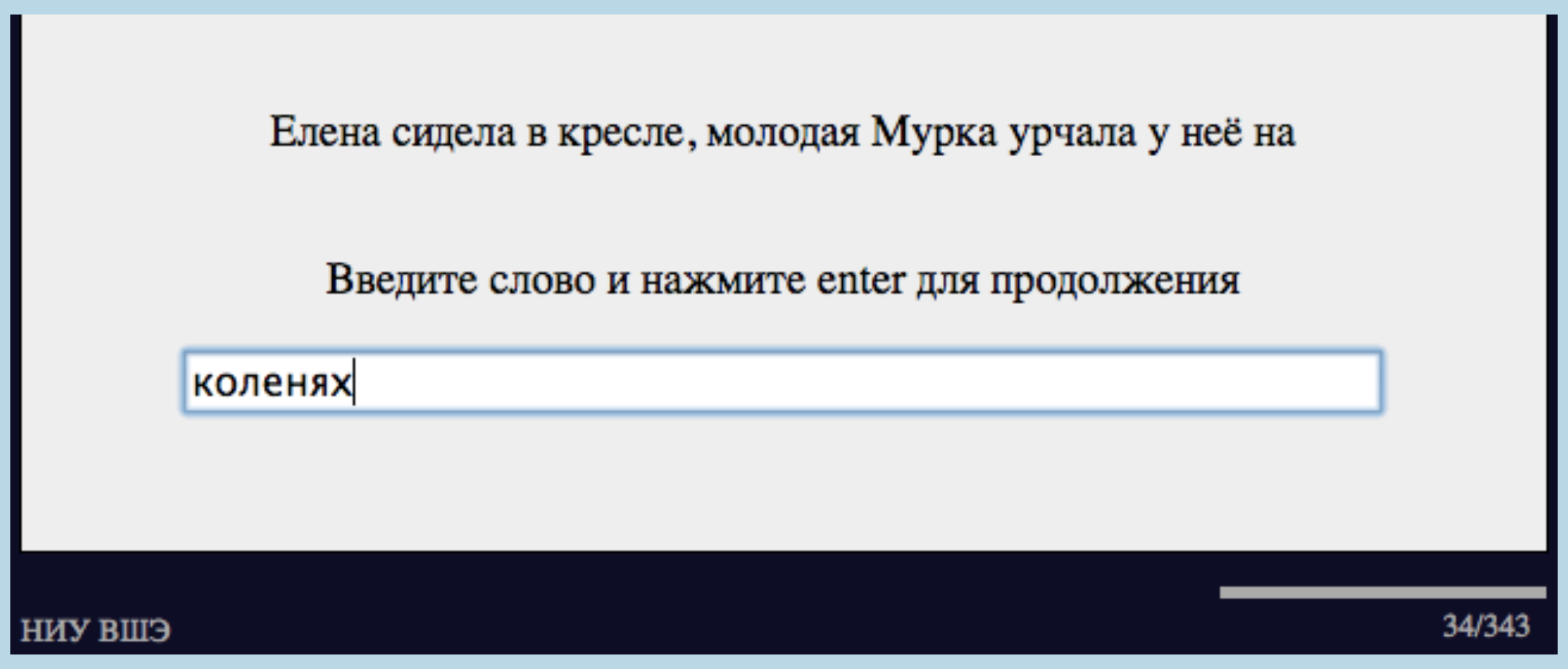
Anastasiya Lopukhina, Konstantin Lopukhin, Anna Laurinavichyute
National Research University Higher School of Economics (HSE)
alopukhina@hse.ru

Background

People can anticipate upcoming words based on the context. So each word has a degree of **predictability** — its probability of occurrence in the context.
Compare:
The boy wants to ...
The boy went outside to ...
Word predictability affects the way people read.

Cloze task

A dataset of 144 sentences from the Russian National Corpus was subjected to predictability norming.
Online experiment
750 native speakers of Russian
20 to 151 guessing attempts for each word
1218 words (excluding first in the sentence)
Mean probability is 18%



Елена сидела в кресле, молодая Мурка урчала у неё на

Введите слово и нажмите enter для продолжения

коленях

НИУ ВШЭ

34/343

Translation: *Elena was sitting in an armchair, a young cat was rumbling on her ...*

Problems [1]:
(1) Cloze task is a production task: participants produce short, familiar, frequent words. They are primed by the preceding context.
(2) It is impossible to count probabilities for words that no participant provided in the cloze experiment.

Discussion

The major advantage of language models is that they are free of the cloze task biases and allow us to capture relative differences for highly unprobable words.
We found that the LSTM model with lower perplexity and higher accuracy has higher correlation with cloze task results.
Corpus-based probabilities explain more variance in eye movements in reading than cloze.

We aim to model not only **lexical predictability**, but also morphosyntactic predictability (part of speech; inflectional information for nouns and verbs) – **graded prediction**. We assume it will explain more variance in eye movements in reading.

References

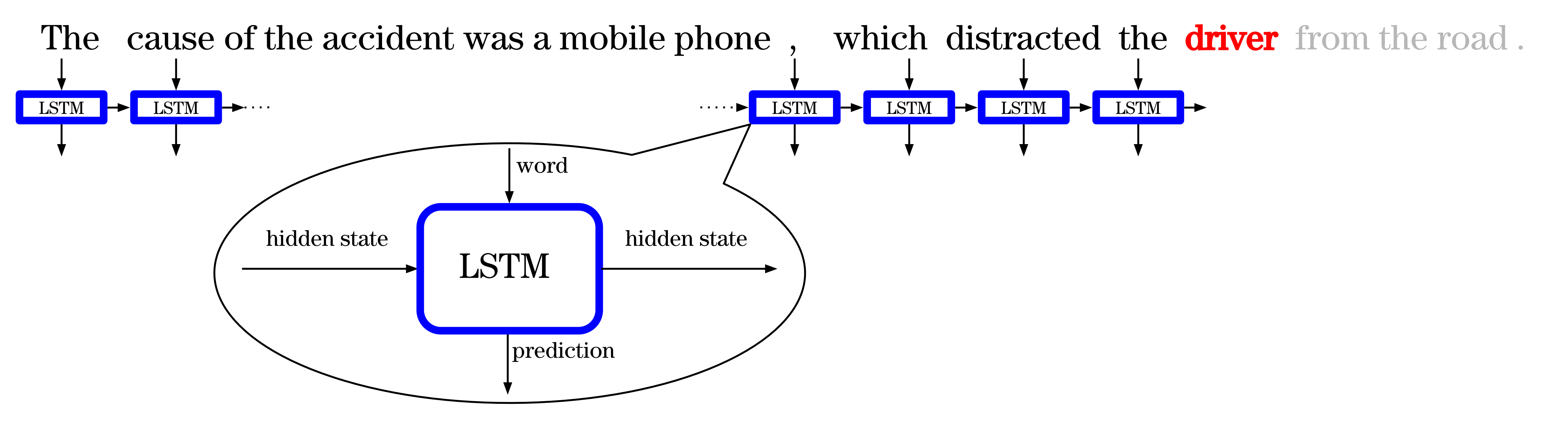
[1] N. J. Smith and R. Levy *Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing*, Proceedings of the 33rd Annual Conference of the Cognitive Science Society (2011)

[2] A. K. Laurinavichyute, I. A. Sekerina, S. Alexeeva, K. Bagdasaryan, and R. Kliegl *Russian Sentence Corpus. Benchmark measures of eye movements in reading in Cyrillic.* submitted

[3] M. Hofmann, Ch. Biemann, and S. Remus *Benchmarking n-grams, topic models and recurrent neural networks by cloze completions, EEGs and eye movements*, Cognitive Approach to Natural Language Processing (2017)

Language model

LSTM recurrent neural network language model (long short-term memory; one layer LSTM-2048-512) is able to use the whole sentence as context for prediction. It gives more accurate predictions (has lower perplexity and higher accuracy) in comparison with n-gram models.


The cause of the accident was a mobile phone , which distracted the driver from the road .
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
word
hidden state
LSTM
hidden state
prediction

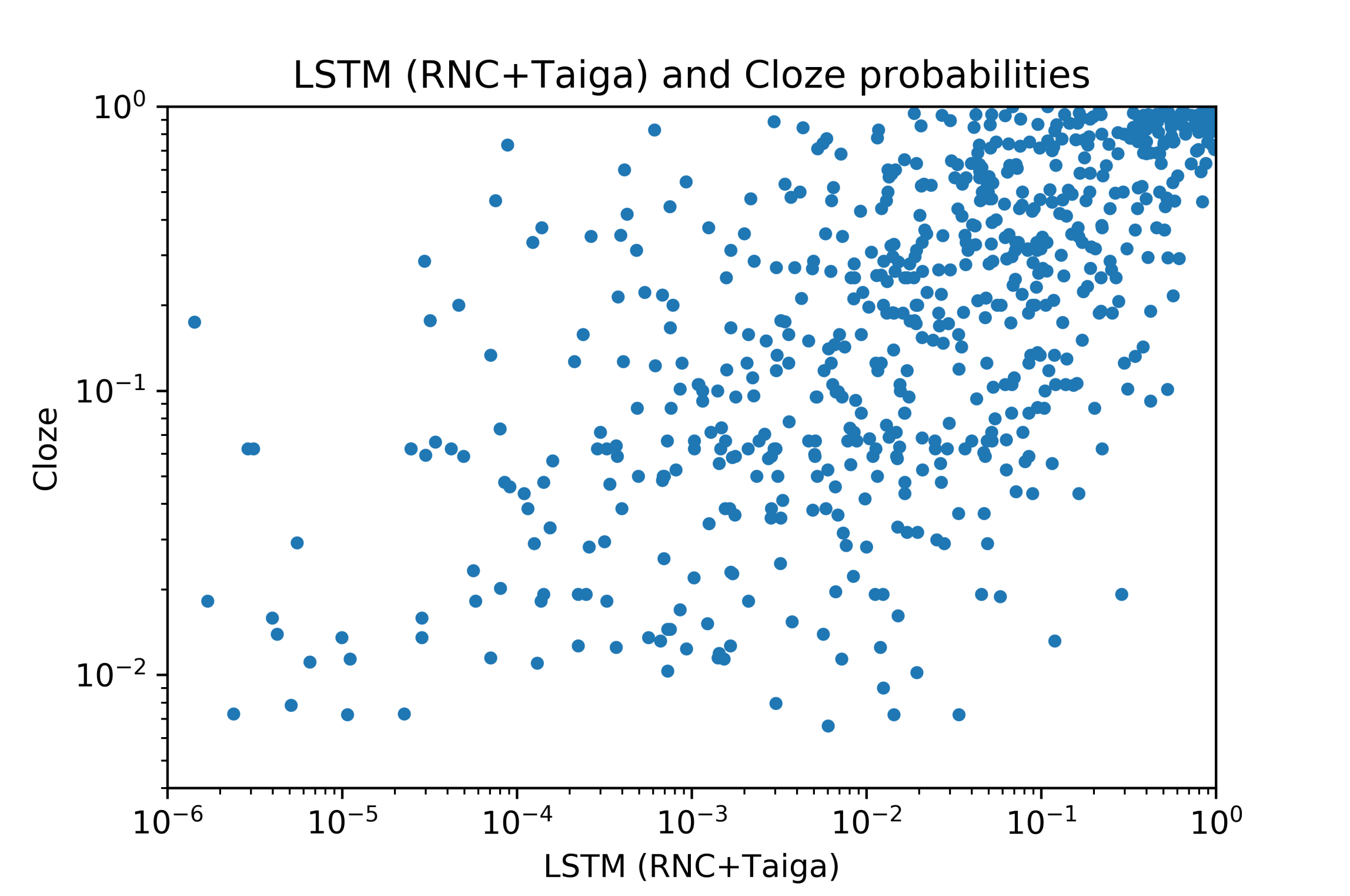
LSTM was trained on three corpora:
– Russian National Corpus (RNC, `ruscorpora.ru`, 576 million tokens);
– web-based corpus Taiga (`tatianashavrina.github.io/taiga_site`, 656 million tokens);
– Russian National Corpus + Taiga (1232 million tokens).

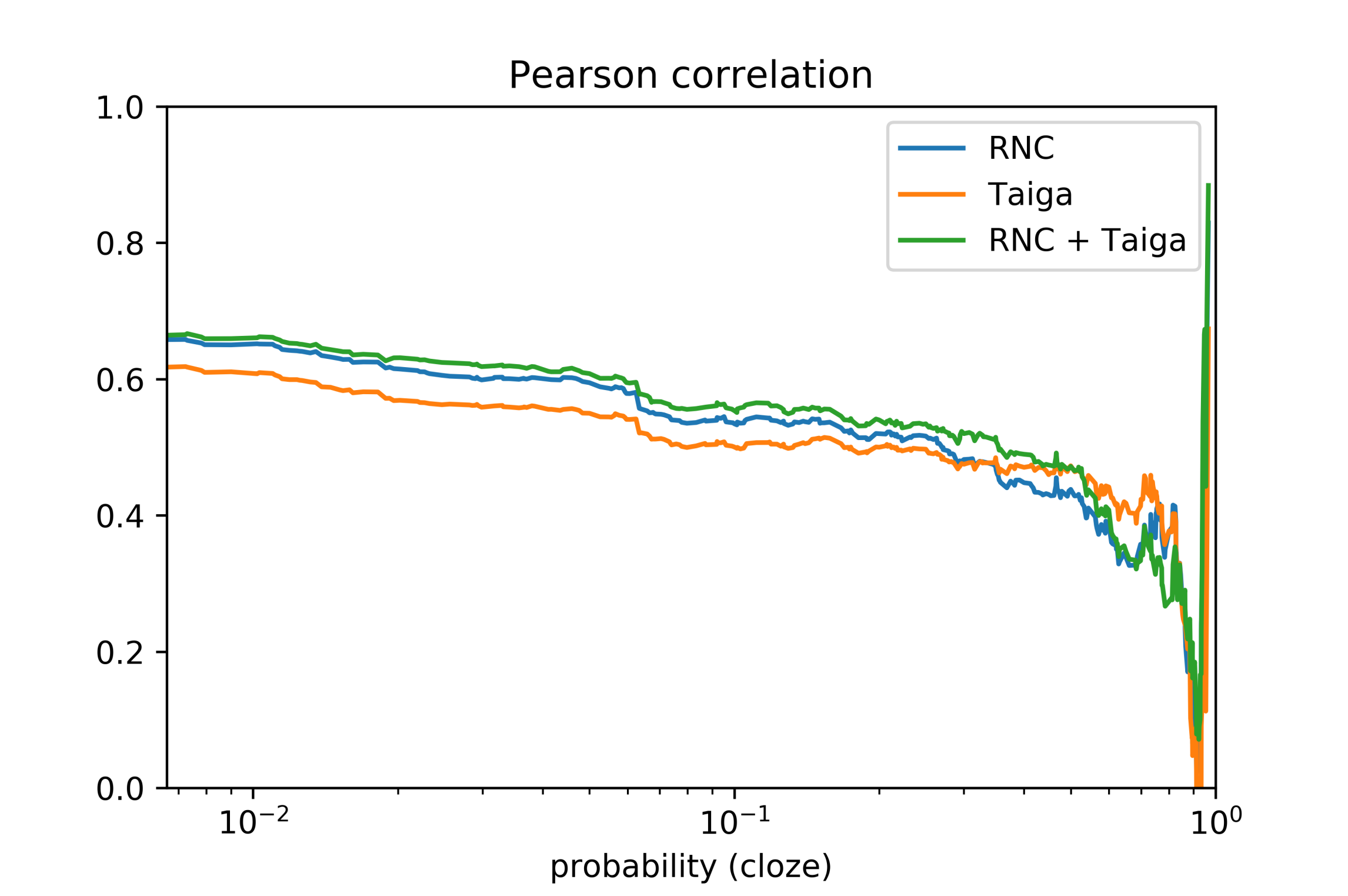
Corpus-based VS cloze probabilities

Pearson correlation is calculated on logit-transformed values and does not take into account zero probabilities (where no human guessed the correct answer). **Spearman** correlation is sensitive only to rank and includes zero probabilities.

	Corpus	Corpus size, M tokens	Perplexity	Accuracy	Pearson correlation	Spearman correlation
Cloze	-	-	-	0.181	-	-
LSTM	RNC	576	**348**	**0.153**	0.658	0.707
LSTM	Taiga	656	419	0.137	0.618	0.681
LSTM	RNC+Taiga	1232	364	**0.153**	**0.664**	**0.716**

The line chart shows that Pearson correlation is **higher** for less predictable words than for more predictable words.





Eye movements: corpus-based VS cloze probabilities

Data: Russian Sentence Corpus (144 sentences; 96 participants; Eyelink 1000+ eye-tracker) [2]
Measures of reading time: **single fixation duration** (the length of the single fixation on a word); **gaze duration** (the sum of all fixations on a word before leaving it for the first time); **total viewing time** (the sum of all fixations, including rereading).

A linear model that explains variance at the item-level, averaged across participants, was used (following the study by Markus Hofmann et al. [3]). Explained variance score was measured depending on features used (r^2). Baseline features: log-transformed word frequency; relative position of a word in a sentence. The LSTM model was trained on RNC + Taiga.

Predictability features	**Single fixation duration**	**Gaze duration**	**Total viewing time**
none	0.2481	0.3960	0.4173
Cloze	0.2770	0.4391	0.4755
LSTM	0.2808	0.4656	0.5058
Cloze + LSTM	0.2873	0.4703	0.5132