NATIONAL RESEARCH UNIVERSITY

Center for Language and Brain

Corpus-based probabilities can substitute for cloze in reading experiments

Anastasiya Lopukhina, Anna Laurinavichyute, Konstantin Lopukhin

Center for Language and Brain, National Research University Higher School of Economics, Moscow

Prediction in language

People can generate predictions about the upcoming input based on available context.

- **lexical** prediction = predict a particular word in a particular grammatical form

e.g. *The athlete pulled a …* [*muscle*]

- **morphosyntactic** prediction = predict some grammatical features of a word

e.g. *We bought a …* [noun, singular]

Predictability from human participants

Cloze task. Produce the most likely next word: *The cause of the accident was a mobile phone, which distracted the ...*

Predictability = the proportion of times the target word was produced over all productions.

Problems:

- no probability is available for words that did not come up in the cloze task
- lexical biases: participants prefer short, frequent, familiar word

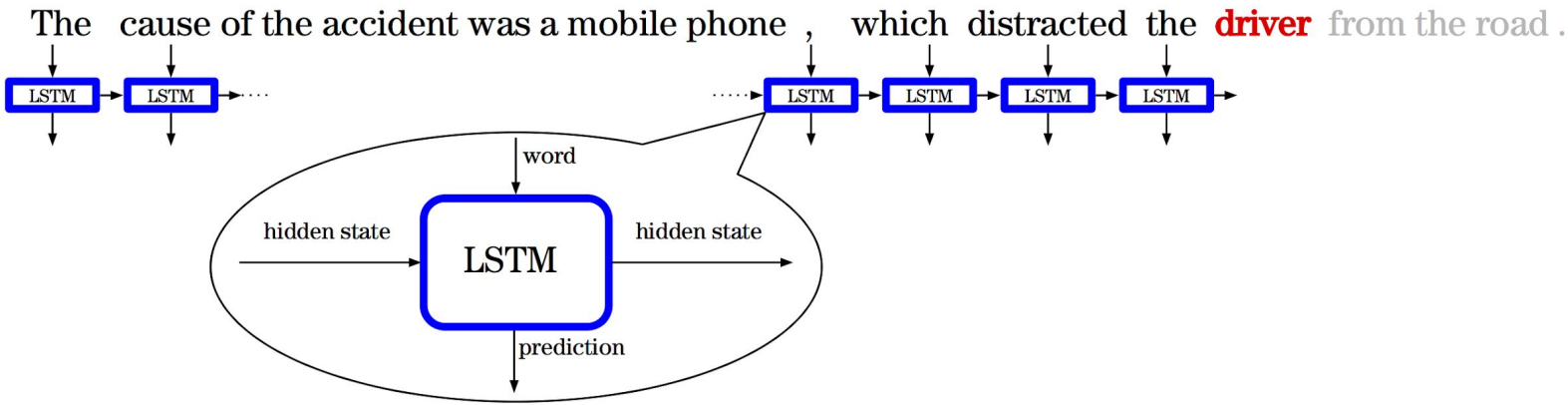
Predictability from a language model

5-gram language model:

The cause of the accident was a mobile phone , which distracted the **driver** from the road .

context | prediction

LSTM language model:



Our study

We compare cloze predictability and corpus-based predictability from a language model:

- by directly correlating them;
- by testing, which of them better predicts eye movements in natural reading.

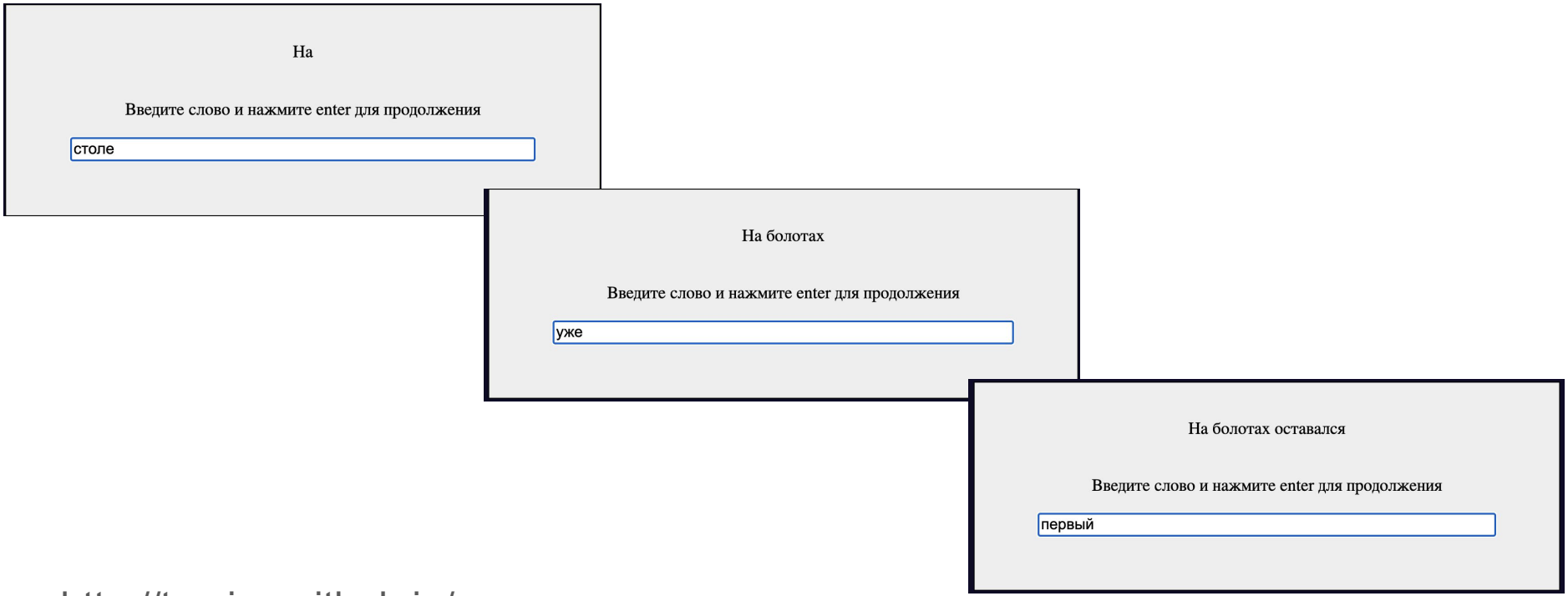
We do these comparisons for **lexical** predictability as well as for **morphosyntactic** predictability.

Method

Cloze predictability

- 605 Russian-speaking participants
- cumulative cloze task
- each word in 144 sentences: 1,218 words
- all words were tagged for word class and morphological features

Cloze task



http://tayrinn.github.io/

Morphosyntactic features

Word classes: nouns, verbs in finite forms, infinitives, adjectives, adverbs, numerals, personal pronouns, prepositions, conjunctions, and particles

Nouns: number, gender, and case

Verbs: tense, person, number, and gender

Method

Cloze predictability

- 605 Russian-speaking participants
- cumulative cloze task
- each word in 144 sentences
- all words were tagged for word class and morphological features

Corpus-based predictability

- long-short-term-memory (LSTM) recurrent neural network model
- trained on the Russian National Corpus (577 million tokens)
- each word in the same 144 sentences
- all words were tagged for word class and morphological features

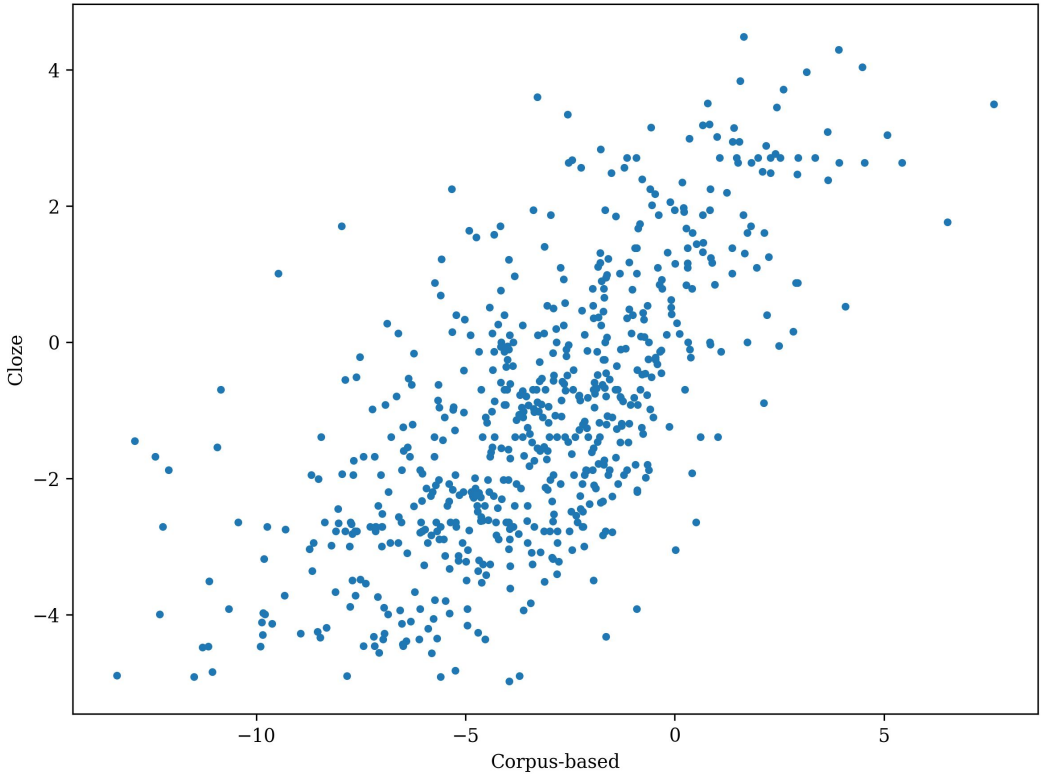
LSTM model

One layer LSTM-2048-512 from (Jozefowicz et al., 2016). The size of the hidden state is 2048; the size of the input and output token embeddings is 512.

Perplexity: 328

Accuracy: 0.173

Correlation: lexical predictability



Mean cloze = 0.184
Mean LSTM = 0.195
Pearson correlation is 0.68

Morphosyntactic predictability: word class

Mean word class probabilities and standard deviations

Word classes	# words	Mean word class cloze probabilities	Mean word class corpus-based probabilities	Pearson correlations
Content words				
nouns	439	0.76 (0.01)	0.81 (0.02)	0.71
verbs (finite forms)	190	0.66 (0.02)	0.70 (0.03)	0.63
verbs (infinitives)	52	0.65 (0.05)	0.71 (0.06)	0.72
adjectives	165	0.35 (0.02)	0.32 (0.04)	0.57
adverbs	44	0.30 (0.05)	0.16 (0.06)	0.72
numerals	6	0.45 (0.15)	0.50 (0.20)	0.00
All content words	896	0.63 (0.01)	0.66 (0.02)	0.70
Function words				
personal pronouns	69	0.47 (0.03)	0.36 (0.06)	0.67
prepositions	117	0.71 (0.03)	0.60 (0.05)	0.59
conjunctions	64	0.74 (0.03)	0.55 (0.06)	0.64
particles	32	0.52 (0.05)	0.50 (0.09)	0.71
All function words	282	0.63 (0.02)	0.52 (0.03)	0.65
All words	1178	0.63 (0.01)	0.62 (0.01)	0.68

Morphosyntactic predictability: nouns and verbs

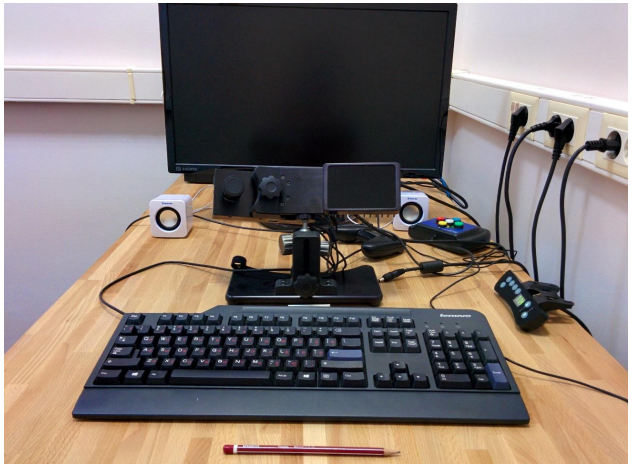
Mean morphological probabilities for nouns and standard deviations

	Mean morphological cloze probabilities	Mean morphological corpus-based probabilities	Pearson correlations
gender	0.62 (0.02)	0.51 (0.02)	0.69
number	0.86 (0.01)	0.83 (0.02)	0.64
case	0.86 (0.01)	0.81 (0.02)	0.58
All features	0.78 (0.01)	0.51 (0.01)	0.67

Mean morphological probabilities for finite forms of verbs and standard deviations

	Mean morphological cloze probabilities	Mean morphological corpus-based probabilities	Pearson correlations
tense	0.56 (0.02)	0.38 (0.04)	0.72
number	0.79 (0.02)	0.73 (0.03)	0.56
person	0.43 (0.04)	0.23 (0.06)	0.73
gender	0.66 (0.03)	0.53 (0.05)	0.71
All features	0.65 (0.01)	0.52 (0.02)	0.71

Eye tracking during reading



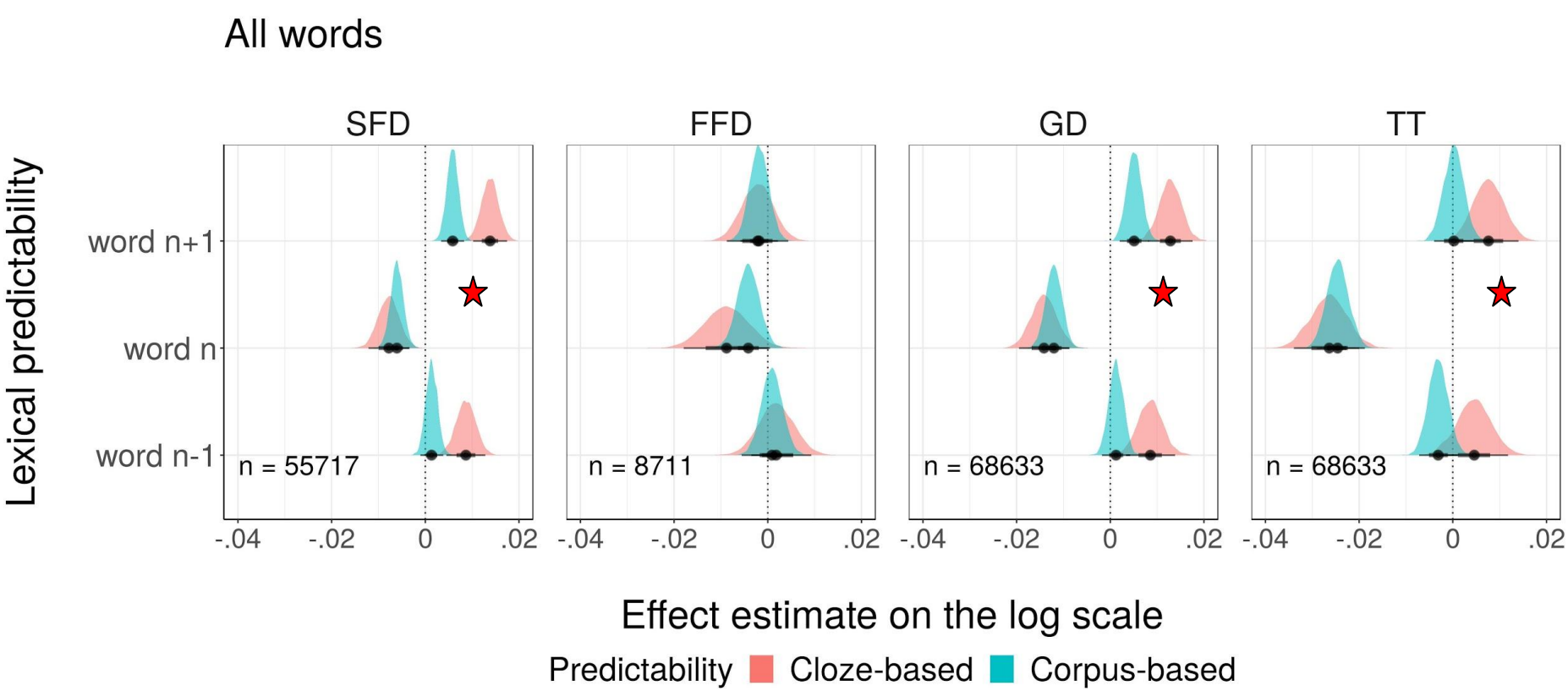
Materials: same 144 sentences

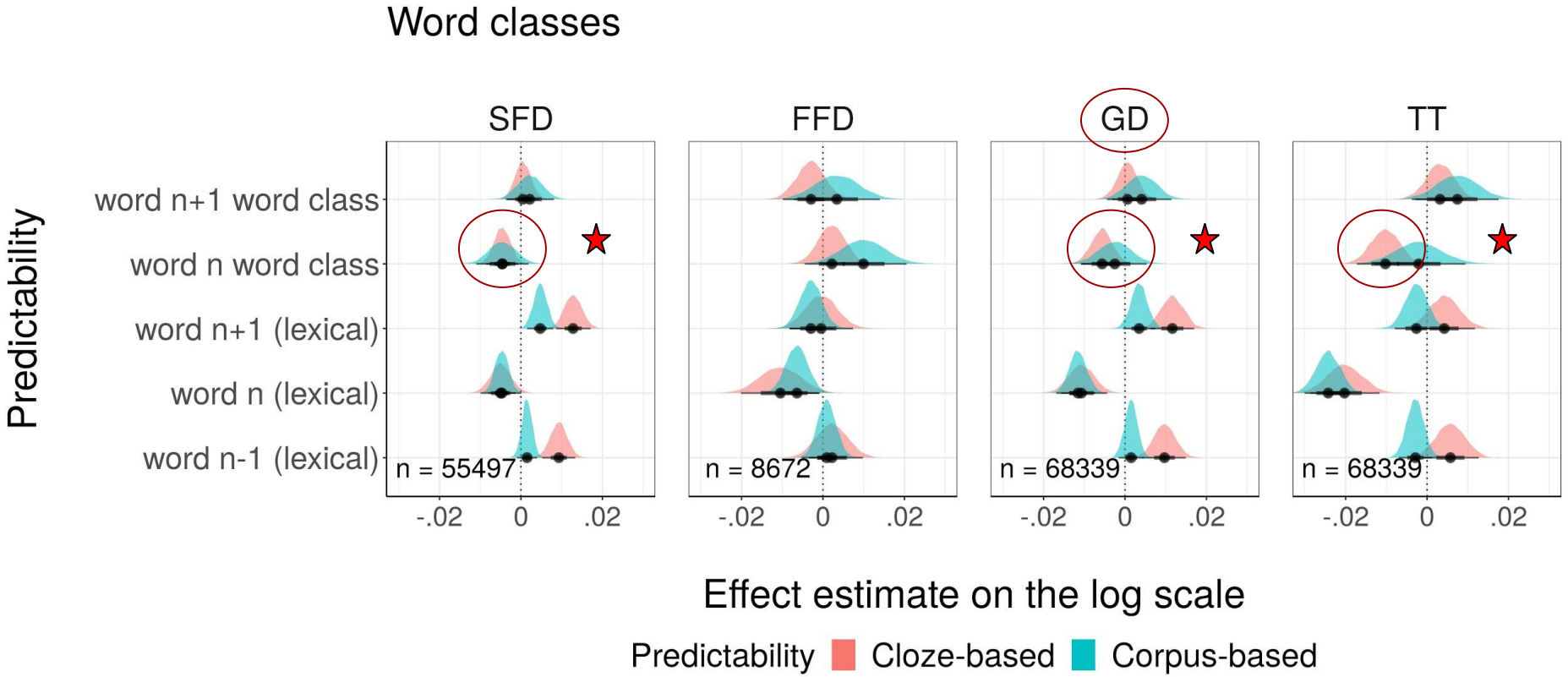
Participants: 96 Russian monolinguals

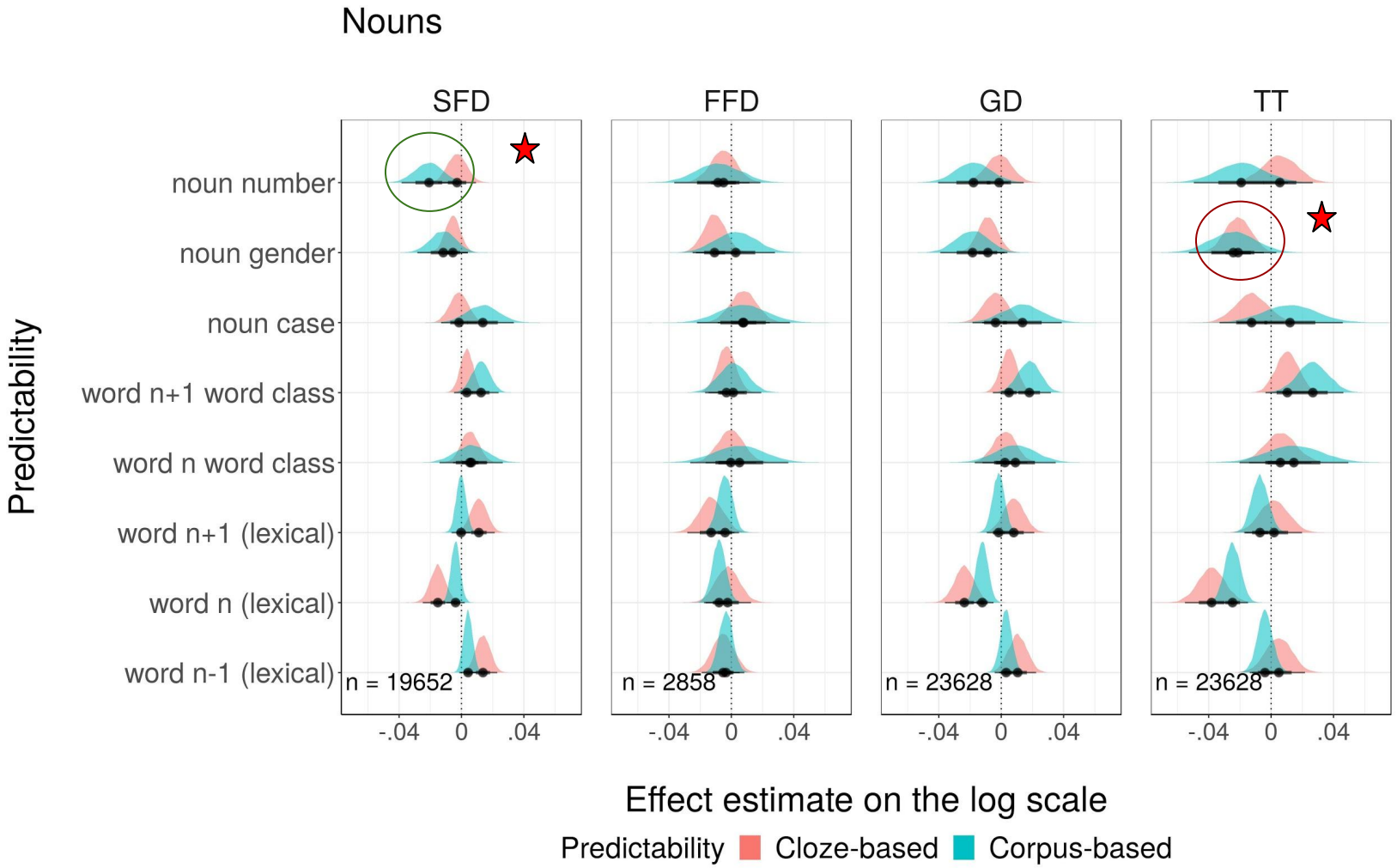
Measures:

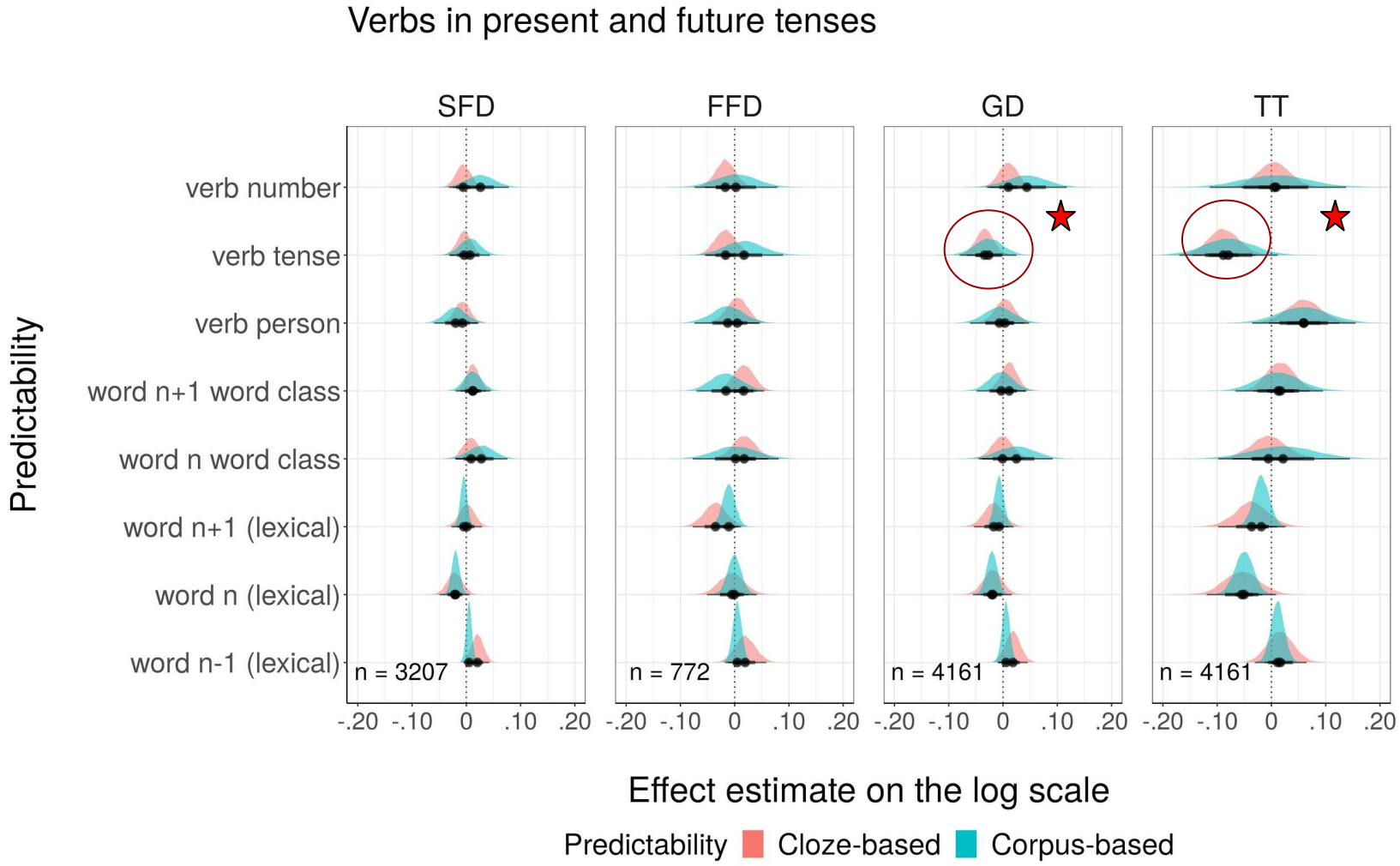
- single fixation duration;
- first fixation duration;
- gaze duration;
- total reading time

Models with cloze and with corpus-based predictiability measures were compared using the k-fold cross-validation (k = 10)


All words
SFD
FFD
GD
TT
Lexical predictability
word n+1
word n
word n-1
n = 55717
n = 8711
n = 68633
n = 68633
-.04
-.02
0
.02
Effect estimate on the log scale
Predictability
Cloze-based
Corpus-based


Word classes
SFD
FFD
GD
TT
Predictability
word n+1 word class
word n word class
word n+1 (lexical)
word n (lexical)
word n-1 (lexical)
n = 55497
n = 8672
n = 68339
n = 68339
-.02
0
.02
Effect estimate on the log scale
Predictability
Cloze-based
Corpus-based


Nouns
SFD
FFD
GD
TT
noun number
noun gender
noun case
word n+1 word class
word n word class
word n+1 (lexical)
word n (lexical)
word n-1 (lexical)
Predictability
n = 19652
n = 2858
n = 23628
n = 23628
-.04
0
.04
Effect estimate on the log scale
Predictability
Cloze-based
Corpus-based


Verbs in present and future tenses
SFD
FFD
GD
TT
verb number
verb tense
verb person
word n+1 word class
word n word class
word n+1 (lexical)
word n (lexical)
word n-1 (lexical)
Predictability
n = 3207
n = 772
n = 4161
n = 4161
-.20 -.10 0 .10 .20
Effect estimate on the log scale
Predictability Cloze-based Corpus-based

Main conclusions

- Cloze and corpus-based predictability measures strongly correlate
- Cloze and corpus-based predictability measures explain the same amount of variance in reading

→ corpus can substitute for cloze in estimating predictability in reading experiments

- Word class can be highly predictable from context
- Higher word class predictability facilitates reading over and above lexical predictability

→ in languages with rich inflectional morphology, such as Russian, pre-activation of word class features is much more common than prediction of words' full identity

http://lm.ll-cl.org/