Man Meets Machine

Dagmar Divjak

Antti Arppe

Ewa Dąbrowska

Coding options

- near synonyms (e.g., *walk, march, stride, strut…*)
- past tense (simple past, past progressive, past perfect, past perfect progressive…)
- causatives (*cause, make, have, get;* lexical causatives)

Distinguishing near-synonyms

- Languages abhor (complete) synonymy: in the vast majority of cases, different options preferred in different contexts
- The criteria governing the selection of the appropriate form are far from obvious
- The most promising approach:
  - extract a large number of examples from a corpus,
  - code for as many possibly relevant features as possible
  - use statistical techniques to develop a model
  - test the model on new set of corpus examples

Problem:
How do we evaluate such models?

- How good is e.g. 65%
- It depends:
  - how many options there are to choose from
  - the degree to which the phenomenon is predictable

- Language is never ever random, but it's also rarely, if ever, fully predictable.

Solution: compare the model to native speakers of the language

Possible outcomes:

- model performs less well than humans
- model performs as well as humans
- model performs better than humans

Russian verbs of trying

- *probovat′, silit′sja, pytat′sja, norovit′, starat′sja, poryvat'sja*)
- can all be translated with the English verb *try,*
- similar but not identical in meaning

Divjak 2010: 1351 examples, 14 categories, 85 tags

- **Try verb**: aspect, mode, tense
- **Subject**: animate (human being vs animal) vs inanimate (abstract vs concrete, man-made vs non-man made etc.)
- **Infinitive**: aspect, degree of control (low, medium, high), type of action (physical action, perception, communication, intellectual activity, emotions etc.
- **Optional elements**: adverbs, particles and connectors, negation
- **Clause type**: main v s subclause, declarative v imperative v interrogative v exclamative

Divjak and Arppe 2013

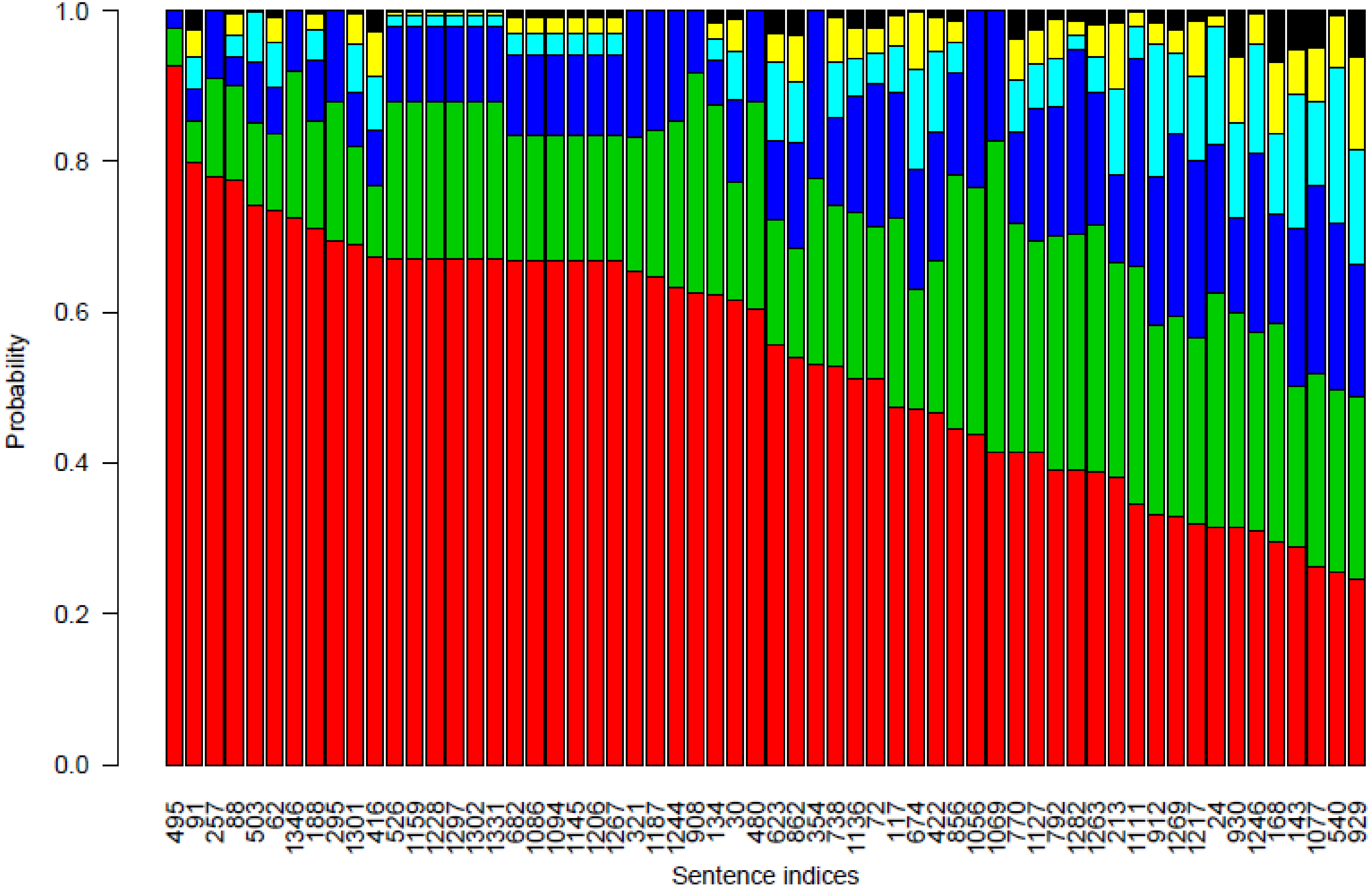
Polytomous logistic regression model which predicts the choice of verb → 51% prediction

Lexeme ~ CLAUSE.MAIN + FINITE.ASPECT_PERFECTIVE + FINITE.MOOD_GERUND + FINITE.MOOD_INDICATIVE + FINITE.TENSE_PAST + INFINITIVE.ASPECT_IMPERFECTIVE + INFINITIVE.CONTROL_HIGH + INFINITIVE.SEM_COMMUNICATION + INFINITIVE.SEM_EXCHANGE + INFINITIVE.SEM_METAPHORICAL_MOTION + INFINITIVE.SEM_METAPHORICAL_PHYSICAL_EXCHANGE + INFINITIVE.SEM_METAPHORICAL_PHYSICAL_OTHER + INFINITIVE.SEM_MOTION + INFINITIVE.SEM_MOTION_OTHER + INFINITIVE.SEM_PHYSICAL + INFINITIVE.SEM_PHYSICAL_OTHER + SENTENCE.DECLARATIVE + SUBJECT.SEM_ANIMATE_HUMAN

Method 1: Humans

- Extracted 60 sentences from the Divjak 2010 dataset representing the whole spectrum of probability distribution
- TRY verb replaced with a blank.
- Each item consisted of a sentence with the blank for the TRY verb, followed by the six options.
- 4 experimental lists, each with a different random order
- Participants: 159 adult native speakers of Russian
- Asked to choose the most appropriate verb for each sentence
- Completed the experiment online, took about 15 minutes

Probability distributions



Method 2: Machine

- Excluded the 60 test sentences from the Divjak 2010 dataset
- Trained the model on the remaining sentences
- Used the model to compute the probability of each of the six verbs in all the test sentences

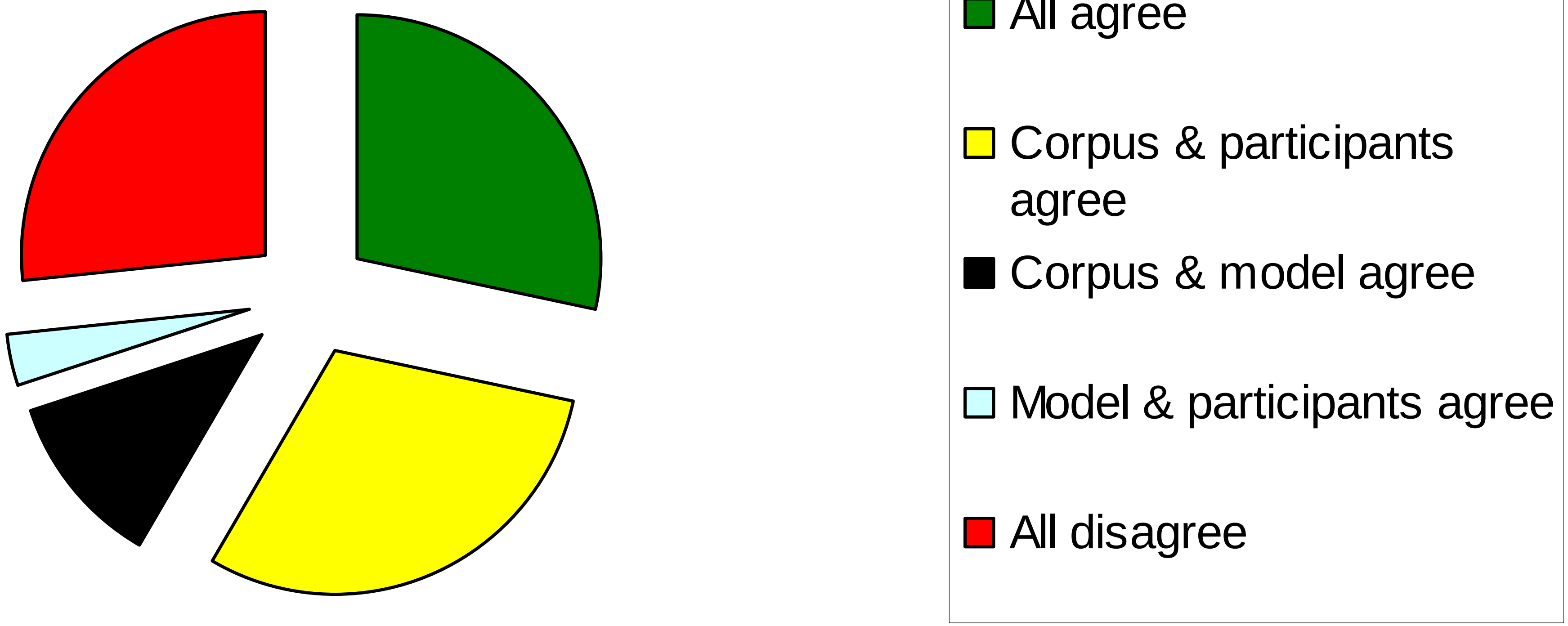
Results

- Chance: 10 (60 items)
- Model: 23
- Participants
  - mean 25.8 (SD 7.5)
  - median 27
  - range 0-38
  - “first past the post”: 35

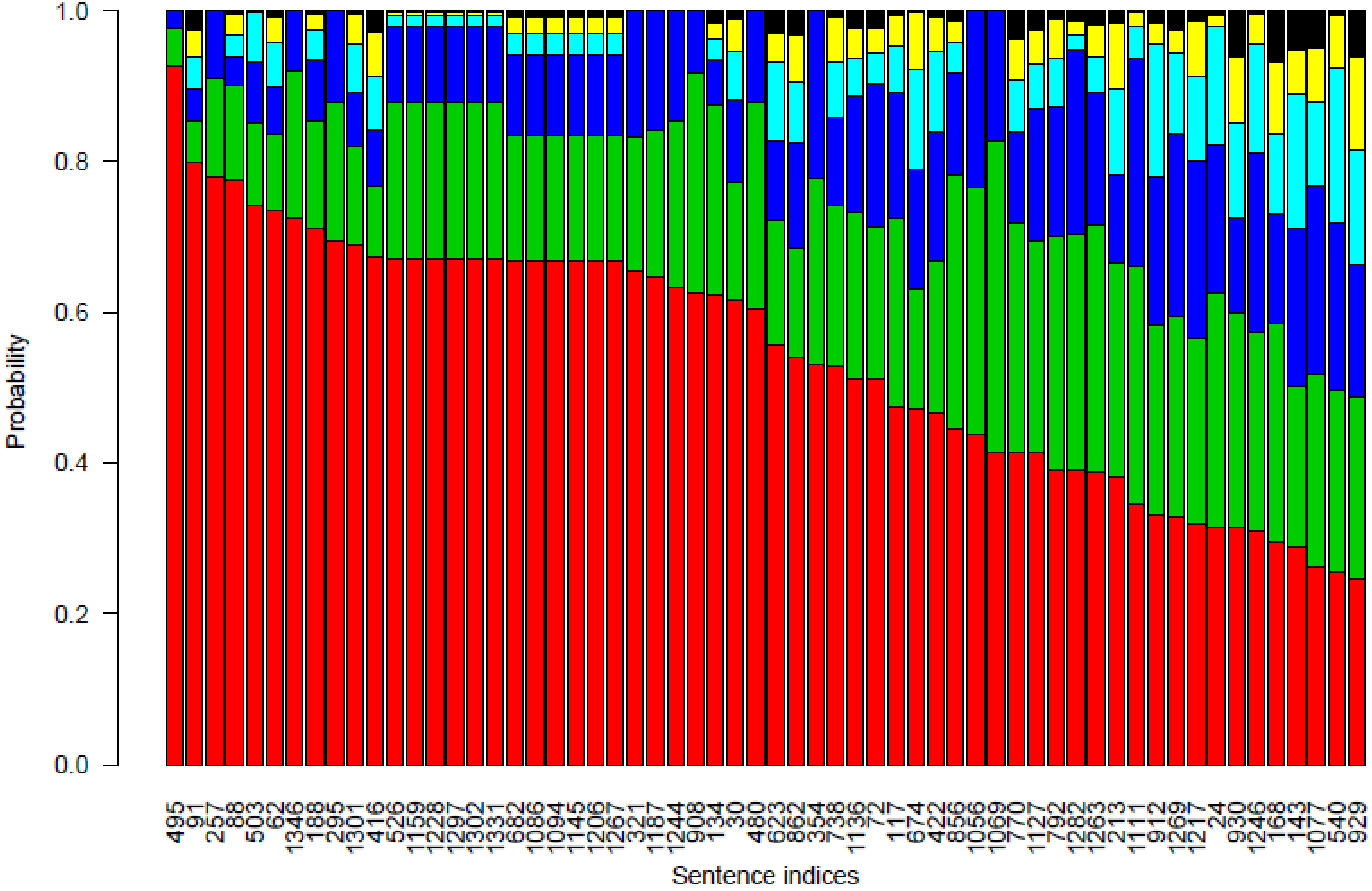
Why is the “plurality vote” more accurate?

- 159 speakers have more experience
- Different speakers pick up on different regularities?

Agreement between corpus, model and participants



Probability distributions

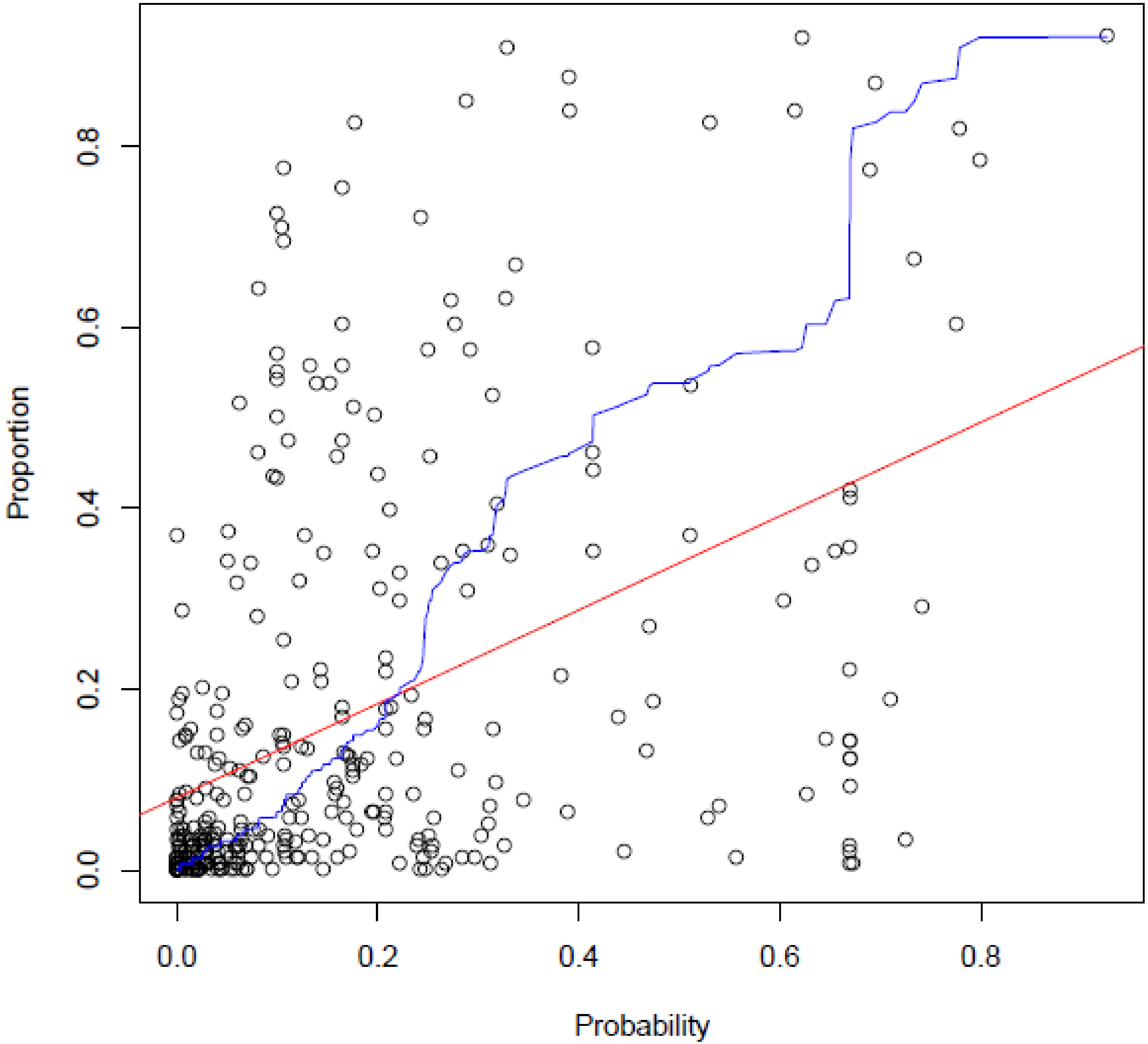


Near-categorical choice (ex. 495)

Verb	Model	Humans	Corpus
norovit’	0.00	0.00	0.00
poryvat’sja	0.00	0.01	0.00
probobat’	0.93	0.92	0.70
pytat’sja	0.04	0.03	0.15
silit’sja	0.00	0.01	0.00
starat’sja	0.02	0.03	0.15

Equiprobable choice (ex. 929)

Verb	Model	Humans	Corpus
norovit’	0.26	0.00	0.080
poryvat’sja	0.13	0.08	0.195
probobat’	0.06	0.11	0.195
pytat’sja	0.15	0.54	0.204
silit’sja	0.16	0.12	0.177
starat’sja	0.23	0.16	0.150


Proportion
0.0
0.2
0.4
0.6
0.8
Probability

Performance by verb

Verb	Model	Humans
norovit’	7/12	10/12
poryvat’sja	2/4	4/4
probovat’	6/8	4/8
pytat’sja	1/10	8/10
silitsja	3/6	0/6
starat’sja	4/20	11/20

The verbs differ in frequency

Verb	Frequency
pytat’sja (pf+imp)	32550
starat’sja (pf+imp)	20011
probovat’ (pf+imp)	4023
norovit’ (imp)	1266
silit’sja (imp)	492
poryvat’sja (imp)	241

Model v. participants

- But model and participants don’t necessarily get the same sentences right.
- Participants tended to use the most frequent verb (*pytat’sja*) as default
- Model that combines information about context with overall frequency = 27/60

Conclusions

- The model performed at about the same level as the average participant, but not as well as participants as a group
- Once predictions were adjusted for frequency, it performed better than the average participant, but not as well as participants as a group

Conclusions

- 38% correct can be pretty good – not everything can be predicted
- Possible that different speakers pick up on different predictors – need further research