Russian Sentence Corpus

Anna Laurinavichyute

Higher School of Economics, University of Potsdam

Contents

Contents

Russian Sentence Corpus, academic interest

- Cyrillic, the 5th in the world use (250 millions actively using) alphabetic script
- Russian language:
  - SVO, but relatively "free" word order
  - rich inflecting morphology, lots of agreement
  - widespread homonymy between case & number endings in written forms
  - phonetic stress that is not represented in the script
  - letters that do not have a corresponding sound

Russian Sentence Corpus, practical use

- a benchmark for reading proficiency tests (test reading in children, dyslexics, aphasic patients, etc.)
- open dataset for everyone to explore
- a testing tool for computational models of eye-movements in reading

Potsdam Sentence Corpus twin brother

- 144 sentences
- target words with orthogonally manipulated length and frequency: nouns, verbs, and adjectives
- sentence length: from 5 to 13 words, mean length – 9, median – 10
- word length – from 1 to 16 letters, mean – 5.5, median – 5

Russian Sentence Corpus, target words

		Frequency		POS
		Low	High	
Length	Short	24	24	(16N, 4V, 4ADJ)
	Medium	24	24	(12N, 6V, 6ADJ)
	Long	24	24	(12N, 6V, 6ADJ)

Table : short – 3-4 letters, medium – 5-7 letters, long – 8-10 letters

Russian Sentence Corpus, target words selection

- random words with given POS-tag, length and frequency range selected from stimul.cognitivestudies.ru/

Russian Sentence Corpus, sentence selection

- 3 sentences handpicked from Ruscorpora.ru with each target word in any form (a letter longer than the base form at most), aiming at having a range from syntactically basic to "interesting" constructions
- lexical complexity score assigned to each sentence
- private vote for the best candidate sentence within the project team
- sentence editing (shortening, removing complex lexemes)
- acceptability norming (acceptability scores $>$ 3.5 out of 5 are allowed)

Russian Sentence Corpus, norming

I have no idea why these sentences got low scores:

- Наше правительство сделало крен на дополнительные инвестиции в развитие науки.
- Твоё тело расслабляется, и исчезает напряжение в области мышц.
- На ведущей вниз лестнице сосед просил прекратить разговоры.
- В бассейне микробы живут недолго, они привыкли к организму.
- В резервациях миссионеры взялись их обращать в новую веру.

Contents

What is predictability and why is it important?

- we have a measure that shows what people think the next word will be, given the preamble
- predictability of a given word in a given context shows in what percent of cases people choose this word as the most likely continuation
- predictability reliably affects reading times and skipping rates
- predictability is an important component of at least two language processing theories (*surprisal theory*: Hale, 2001; Levy, 2008)

Predictability data collection

- asking participants to guess the next word in a sentence
- constantly updating the correct cloze, so the task might get easier towards the end of sentence
- let's try our test
- about 40 sentences for each participant
- no reaction times measured
- you can do it in your free time!

Technical issues

- problems with automatic accuracy scoring:
  В сюжете этого фильма какие-то осы устраивают себе гнездо (the "right" answer is гнезда)
- misspellings (at least I expect there are plenty)
- too many ways to specify your native language as Russian (26!)

Technical issues

```
("Ru", "RUS", "ru", "rus", "russian", "russkij", "Russkij",
 "рус", "Русс5", "Руский", "русскии", "русский", "русский ",
 "руссКий", "РУССКИЙ", " русский", "Русский язык",
 "русссский", "Русский", "Русский ", "РУССКИЙ", " русский",
 "русски й", "русский язык", "русскиф", "Русском"), ]
```

Preliminary results

- mean predictability across all words in all sentences – 15% (with values from 0 to 100%)
- mean predictability across target words – 11%

> 80% predictability ratings

- Ваня раскрыл было рот, но понял, что
- И не надо ставить это целью всей своей жизни
- Город, раскинувшийся вдоль реки, состоял из
- Зоопарк – это кусочек другого мира
- У мамы есть подруга, которая
- Существует легенда, что Ноев ковчег
- Ей никак не

Predictability of target words

For target words,

- the closer to the end of the sentence, the higher the predictability scores
- the higher the frequency, the higher the predictability
- adjectives recieve lower predictability scores than verbs
- nouns recieve higher predictability scores than verbs

Future directions

- part-of-speech predictability (POS-tagging problem. Any potential collaborators here?)
- close synonym set predictability
- word form predictability (гнездо-гнёзда, пришёл-ушёл)
- dealing with misspellings

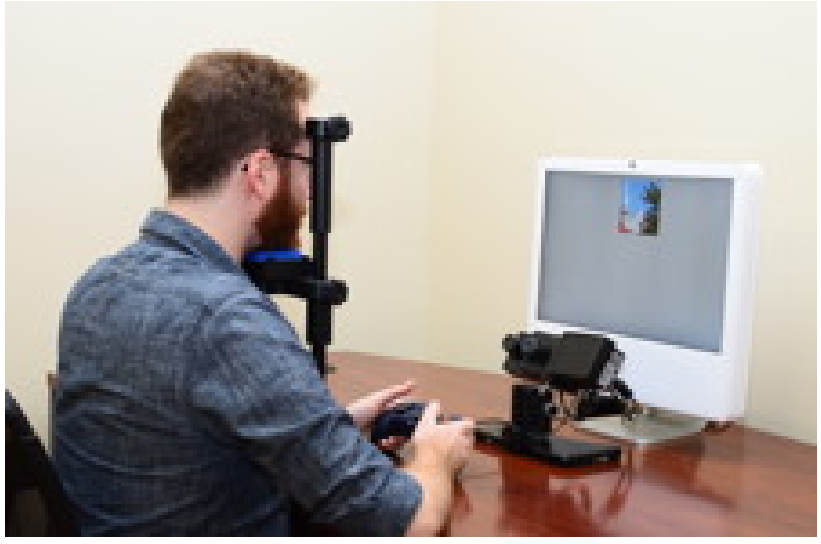
Contents

The eye-tracking measures

- So-called early measures:
  - first fixation duration
  - gaze duration (sum of all fixations on the word during first-pass reading)
  - probability of skipping the word
- So-called late measures:
  - regression path duration (time spent reading the target word and the earlier parts of the sentence, if there were any regressions from the target word)
  - total reading time

The experiment

- Eyelink 1000+ Desktop mount eye-tracker with chin rest
- tracking the right eye
- 9 points calibration, recalibration after every 15 sentences
- multiple choice comprehension questions for approx. 1/3 of all sentences
- 144 sentences



Word length effect

The longer the word,

- the longer the first fixation duration
- the longer the gaze duration
- the lower the probability of skipping
- the longer the total reading time
- the longer the regression path duration
- the closer to the beginning of the word lands the saccade

Frequency effect

The higher the frequency,

- the shorter the first fixation duration
- the shorter the gaze duration
- the higher the probability of skipping
- the shorter the total reading time
- the closer to the end of the word lands the saccade
- the shorter the regression path duration

Predictability effect

The higher the predictability,

- the shorter the first fixation duration
- the shorter the gaze duration
- the higher the probability of skipping
- the shorter the total reading time
- the shorter the regression path duration

Position within the sentence

The closer the word is to the end of the sentence,

- the closer is the saccade landing position to the end of the word
- the longer the first fixation on the following word

Sentence length

The longer the sentence,

- the shorter the first fixation on the following word

Presence of a comma

If the word contains a comma at its end,

- the total reading time is faster than for a word of the same length with all letters and no commas
- the saccade landing position is closer to the beginning of the word
- the probability of skipping this word is greater (that's interesting)

Part of speech

Nouns have

- higher probability of skipping than verbs and adjectives
- faster total reading times than verbs and adjectives

Fixations on the previous word

If the previous word was skipped,

- the total reading times of the current word is longer
- the saccade landing position at the current word is closer to the beginning of the word
- the regression path duration for the current word is longer

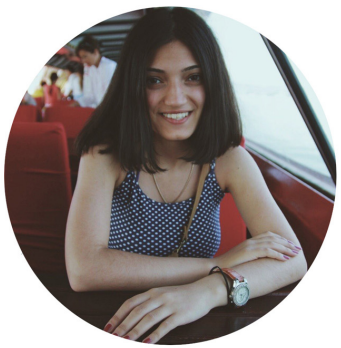
Future directions

- build syntactic trees and explore the influence of syntax (Potsdam Sentence Corpus – Boston et al., 2008; Hindi corpus – Husain et al., 2015)
- examine the part-of-speech factor more closely (are there any factors that affect difference in reading times between verbs and nouns?)
- create a duplicate corpus with the same target nouns in a complimentary case (direct vs. indirect) and explore the role of case
- examine the end of clause processing (any wrap-up effects?)
- are homonymous word forms read slower?
- how are words with ь and ъ read?
- how are consonant clusters with "silent" consonants read?
- the role of phonetic stress (stress alternation in word forms)

Acknowledgements

Irina Sekerina

Svetlana Alexeeva



Kristina Bagdasaryan