# LLMs Out-of-the Box Do Not Generate Context-Appropriate Word Order in Russian

Alina Shabaeva, John Frederick Bailyn, Owen Rambow

Stony Brook University
Department of Linguistics
Institute for Advanced Computational Science

# Word Order Variation in Russian

| | | |
|---|---|---|
| 1. | Maria čitaet knigu.<br>Maria.NOM read.PRES.3SG book.ACC<br>*'Maria reads a book'* | *SVO* |
| 2. | Maria knigu čitaet. | *SOV* |
| 3. | Knigu Maria čitaet. | *OSV* |
| 4. | Knigu čitaet Maria. | *OVS* |
| 5. | Čitaet Maria knigu. | *VSO* |
| 6. | Čitaet knigu Maria. | *VOS* |

# Examples

- Нет, я работаю в кино.
- Я и режиссер, и актер, и сценарист и продюсер.

**- Вы меня дурачите!** ***SOV***

- No, I work in film.
- I am a director, an actor, a screenwriter, and a producer.

- You're fooling me!

- Ты рассказала его пятидесяти миллионам зрителей, а мне не хочешь?
- Обещаю, я не буду смеяться.

**- Этого-то я и опасаюсь.** ***OSV***

- You told it to fifty million viewers, but you won't tell me?
- I promise I won't laugh.

- That is exactly what I fear.

# Research Questions

1. What **features** influence surface word order in Russian?

2. Can **LLMs** generate various word orders in Russian, or do they default to the canonical SVO order?

# Outline

1. Corpus

2. Machine Learning Experiments

3. LLM Experiments

# Corpus & Data Preprocessing

- Based on the Movie Scripts dataset (Lisovsky, 2019);
- 20 film scripts translated into Russian;
- Genres: crime, comedy, drama, thriller (1992-2012);
- 10 scripts manually edited so that it can be used as a dialogue corpus;
  - We removed screenplay elements such as camera instructions, scene descriptions and speaker names.

# Building the Corpus

1. **Word order** information based on dependency relations (Stanza)

*Вопросы тут задаю я.* **OVS** *I'm the one asking the questions here.*

2. **Preceding context**

- Это не твой текст. Скажи, что любишь меня.

- It's not your text. Tell me you love me.

*- Я тебя люблю.* **SOV** *- I love you.*

# Word Order Distribution

| Word order | Count | Percentage | Binary |
|---|---|---|---|
| **SVO** | 1123 | 72.8% | 1123 (72.8%) |
| **SOV** | 275 | 17.8% | |
| **OSV** | 71 | 4.6% | |
| **OVS** | 50 | 3.2% | 420 (27.2%) |
| **VSO** | 14 | 0.9% | |
| **VOS** | 10 | 0.7% | |

# Outline

1. Corpus

**2. Machine Learning Experiments**

3. LLM Experiments

# Decision Tree: binary classification

| Class | Precision | Recall | F1 |
|---|---|---|---|
| **SVO** (74%) | 0.88 | 0.83 | 0.86 |
| **non-SVO** (26%) | 0.59 | 0.68 | 0.63 |

| Metric | Accuracy on Test set | Majority Baseline (SVO) | Random Baseline |
|---|---|---|---|
| Accuracy | 0.79 | 0.43 | 0.50 |

| Feature | Imp. |
|---|---|
| **object_pos** | 0.61 |
| **predicate_type** | 0.10 |
| **coref_score_object** | 0.06 |
| has_negation | 0.05 |
| verb_constituent_size | 0.04 |
| object_syllable_weight | 0.03 |
| object_ner | 0.02 |
| subject_role | 0.02 |
| focus_subject | 0.02 |
| verb_type | 0.01 |

# Outline

1. Corpus

2. Machine Learning Experiments

**3. LLM Experiments**

# Prompt Template

You are working with a movie script in Russian. Below is a context from the script, where:
  SPEECH indicates character dialogue
  COMMENT indicates scene description or comment

**Context**:
{context}

Continue this movie scene by generating a new sentence that naturally follows this context.
The generated sentence **MUST** contain:
- the verb “{verb_lemma}” (with the specified aspect, tense, and negation, if applicable),
- “{subject}” as the subject constituent (nominative case), and
- “{object}” as the object constituent (accusative case).

**Requirements:**
1. You must use the exact given words:
   - subject: “{subject}”
   - verb: “{verb_lemma}”
   - object: “{object}”
2. These words must be used within the same clause as direct arguments of the verb.
3. Generate exactly one sentence in Russian. Do not add translation, comment, or explanation.
4. The sentence must:
   - be in the indicative mood,
   - be natural continuation of the context,
   - be grammatically correct,
   - not be a question, command, or use the subjunctive mood,
   - no be a copy of the context.

# Prompt continued

**NOTE ON WORD ORDER:**
While Russian allows flexible word order, subject-verb-object (SVO) is the most common and natural patterns. Prefer SVO unless the context (pragmatic or stylistic features or morphosyntactic properties) require a different order for emphasis, information structure, style, or emotion.

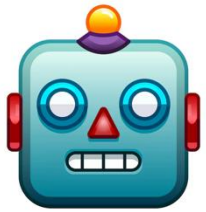
**OUTPUT:**
Provide only one generated Russian sentence.

# Examples

- Ты хотел собой покончить.
- А ты мне пытался мозги вышибить.

- You wanted to kill yourself.
- And you tried to blow my brains out.

***- Пистолет*** *ты не получишь.* **OSV** *- You won't get a gun.*



1. Ты не получил **пистолет**, чтобы всё это закончить.
2. Ты не смог получить **пистолет**.
3. Ты не получил **пистолет**, чтобы сдаться.

**SVO**

1. *You didn't get a gun to end it all.*
2. *You couldn't get a gun.*
3. *You didn't get a gun to give up.*

# Model Selection

**Goal**: **NOT** to evaluate whether the model generates the "correct" word but investigate **word order distribution.**

- Models evaluated: GPT-3.5 Turbo, GPT-4o-mini, and GPT-5-mini
- 3 runs per sentence

| | Gold | GPT-3.5-turbo | GPT-4o-mini | GPT-4o |
|---|---|---|---|---|
| **SVO** | 74.0% | 83.6% | 90.1% | 86.4% |
| **other** | 26.0% | 7.0% | 0.5 % | 5.9% |
| **UNK** | | 9.4% | 9.3% | 7.7% |

**Unknown (UNK)** outputs generated sentences that could not be evaluated due to:

- Incoherent output;
- Failure to follow the instructions in the prompt

# Include Decision Tree Prediction in Prompt

| WO | GOLD | GPT-4omini | Decision Tree predictions in Prompt |
|---|---|---|---|
| SVO | 74.0% | 90.1% | 85.6% |
| other | 26.0% | 0.5% | 11.6% |
| UNK | | 9.3% | 2.7% |
| **accuracy** | | 69.3% | 73.4% |

**Unknown (UNK)** outputs

Generated sentences that could not be evaluated due to:

- Incoherent output;
- Failure to follow the instructions in the prompt

## Few-shot prompting

Original prompt with four examples, one for each word order (with context):

SVO
SOV
OVS
OSV

## Fine-tuning

- Supervised fine-tuning on the training corpus;
- Training examples formatted as input-output pairs;
- Input: context + constituent + morphosyntactic features from the prompt;
- Output: target word order label

# Few-shot prompting & Fine-tuning

| WO | GOLD | GPT-4omini | Decision Tree predictions in Prompt | Few-shot | Fine-tuned |
|---|---|---|---|---|---|
| SVO | 74.0% | 90.1% | 85.6% | 95.5% | 71.7% |
| other | 26.0% | 0.5% | 11.6% | 0.8% | 18.5% |
| UNK | | 9.3% | 2.7% | 3.6% | 9.8% |
| **accuracy** | | 69.3% | 73.4% | 71.8% | 73.6% |

*Word order distributions across baseline annotations and multiple prompting strategies (**GPT-4o-mini**)*

# Summary: Contributions

- Releasing an annotated corpus;
- Evaluation of machine learning experiments evaluating our set of features predicting word order;
- LLMs struggle to generate various word orders;
- Possible solution: fine-tuning.

# THANK YOU!

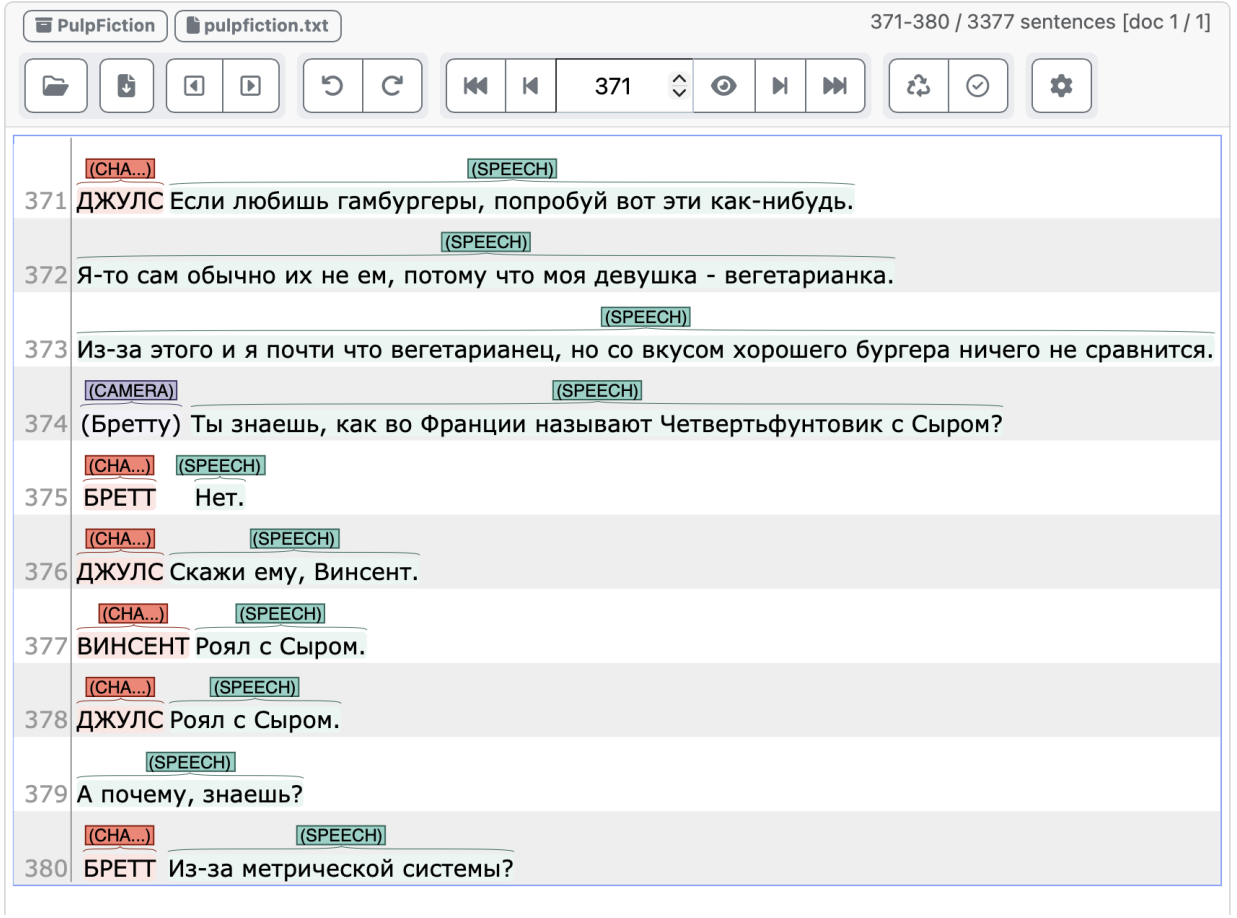
# BACKUP SLIDES

# Data Annotation

Manually annotated 10 scripts (for script structure) using INCEpTION:

- COMMENT
- SPEECH
- CHARACTER NAME
- CAMERA

***Movie Scripts (Lisovsky, 2019)***



# Building the Corpus: Adding Features

## Morphosyntactic

- Part of speech (POS): noun, pronoun, verb, etc.;
- Constituent size (number of tokens) for subject, object, verb;
- Verb aspect (perfective / imperfective);
- Presence of negation for the verb;
- Clause type (matrix / embedded clause)

## Semantic

- Subject and object named entity types (PER, ORG, LOC, MISC)
- Animacy for subject and object;
- Verb type (perception, communication, action, etc.)
- Semantic roles of subject and object (Experiencer, Agent, Theme, etc.)

## Phonological

- Syllable weight

## Discourse

- Discourse-related elements;
- The coreference score;
- Is target sentence preceded by a question?

# Example

**Ona ne polozhila tvoj dvojnoj burger!**
She.NOM not put in. PAST.FEM. your double burger.ACC.
*'She didn't put your double burger in.'*

Ona | ne | polozhila | tvoj dvojnoj burger

- **POS**: PRONOUN
- **constituent size**: 1
- **syllable weight**: 2
- **animacy**: not_applicable
- **NER**: PER
- **type**: agent
- **coreference score**: 1.2

- **syllable weight**: 4
- **aspect:** perfective;
- **verb type**: action;
- **predicate type**: 2-place

- **POS**: NOUN
- **constituent size**: 3
- **syllable weight**: 5
- **animacy**: inanimate
- **NER**: not_applicable
- **type**: theme
- **coreference score**: 1.2

- **preceded by question:** FALSE;
- **presence of negation:** TRUE;
- **clause type:** matrix;
- **WORD ORDER:** SVO
- **focus subject:** FALSE
- **focus object:** FALSE

# Decision Tree: multiclass classification

| class | precision | recall | F-1 score |
|---|---|---|---|
| SVO (74%) | 0.88 | 0.85 | 0.87 |
| SOV (18%) | 0.53 | 0.78 | 0.63 |
| OSV (5%) | 0.75 | 0.21 | 0.33 |
| OVS (3%) | 0.33 | 0.10 | 0.15 |

| Metric | Accuracy on Test set | Majority Baseline (SVO) | Random Baseline |
|---|---|---|---|
| accuracy | 0.78 | 0.21 | 0.28 |

# Top-10 most important features

| Feature | Imp. |
|---|---|
| **object_pos** | 0.47 |
| **coref_score_object** | 0.09 |
| obj_animacy | 0.09 |
| **predicate_type** | 0.08 |
| object_syllable_weight | 0.05 |
| verb_constituent_size | 0.05 |
| subject_role | 0.04 |
| coref_score_subject | 0.03 |
| object_ner | 0.03 |
| subject_syllable_weight | 0.02 |

Multiclass classification

| Feature | Imp. |
|---|---|
| **object_pos** | 0.61 |
| **predicate_type** | 0.10 |
| **coref_score_object** | 0.06 |
| has_negation | 0.05 |
| verb_constituent_size | 0.04 |
| object_syllable_weight | 0.03 |
| object_ner | 0.02 |
| subject_role | 0.02 |
| focus_subject | 0.02 |
| verb_type | 0.01 |

Binary classification