Cross-Methodological Comparison and Validation of Common Methods in NLP-based Psychosis Detection

Galina Ryazanskaya

1st advisor: Dr. Sherzod Hakimov
2nd advisor: Prof. Dr. Manfred Stede

Motivation

- Incoherent language is an important diagnostic feature of psychotic disorders
    - e. g. schizophrenia, schizoaffective disorder, etc

- Many NLP-based methods of detecting such incoherent language were proposed

- NLP-based psychosis detection tasks include
    - distinguishing patients from controls or healthy family members
    - predicting conversion in clinical high-risk populations
    - assessing or predicting symptom severity
        - positive vs negative symptom severity

Motivation

The commonly used NLP-based methods often

- lack replicability
    - different diagnosis and symptoms in the samples
    - different elicitation tasks
    - different metric implementation and text preprocessing
- have limited cross-linguistic validity
    - rarely evaluated on cross-linguistic samples
- are associated with confounding factors both internal and external
    - text or sentence length (patients are commonly reported to speak less)
    - sex, age, education, race, etc.
- are rarely directly evaluated against each other or against a simple baseline

Benchmarking study on two languages of the most common metrics

Literature Review

- 62 papers, theses, and conference proceedings
    - keywords & references to expand the search
    - excluded purely manual linguistic metrics & purely auditory / spoken metrics
- metrics were grouped by the units they operated on & assessed in performance
    - lexical, syntactic, semantic (graph and LM)
- languages, experiment designs, and preprocessing were explored
- degree of reported dependence on sentence length

- frequency of use
- reported performance
- cross-linguistic / cross-study applicability

Metrics

- Lexical: word count, LTR, MALTR
- Syntactic: POS rates, sentence length (mean, max, min, std) and count
- Graph-based:
    - co-occurrence graph (lemmas as nodes, co-occurrence as directed edges)
    - NN, NE, LCC, LSC, L1, L2, L3, PE, node degree (mean, std)
- LM:
    - *models*: w2v, GloVe, BERT
    - *sentence embeddings*:
        - average word vectors, TF-IDF weighted average
        - BERT last layer hidden-state CLS token embedding
    - local (first order) coherence, second order coherence, global (centroid) coherence, cumulative global coherence
    - pseudo-perplexity, next sentence probability

Data

- German sample:
  - 59 NAP and 20 controls
  - PANSS, SANS, SAPS, verbal IQ - more negative symptoms
  - Narrative Emotion Task
- Russian sample:
  - 31 NAP, 18 Dep, 30 controls
  - PANSS, TD (psychosis) and depression severity
  - 2 picture-elicited tasks (‘*sportsman*’, ‘*adventure*’)
  - 1 instruction task (‘*chair*’)
  - 1 personal story task (‘*present*’)

Preprocessing: sentence segmentation, lemmatization, hesitation pauses and interviewer speech removal

Association with social confounding factors detected neither in symptoms nor in metrics

Research Questions

- which metric groups work best?
- which metrics outperform the simplest baselines?
    - word count, sentence count, mean sentence length
- which metrics are associated with sentence length?
- which metrics work for which scales?
- which metrics work across tasks?
    - elicitation task effects
- which metrics work on both languages?
    - cross-linguistic differences

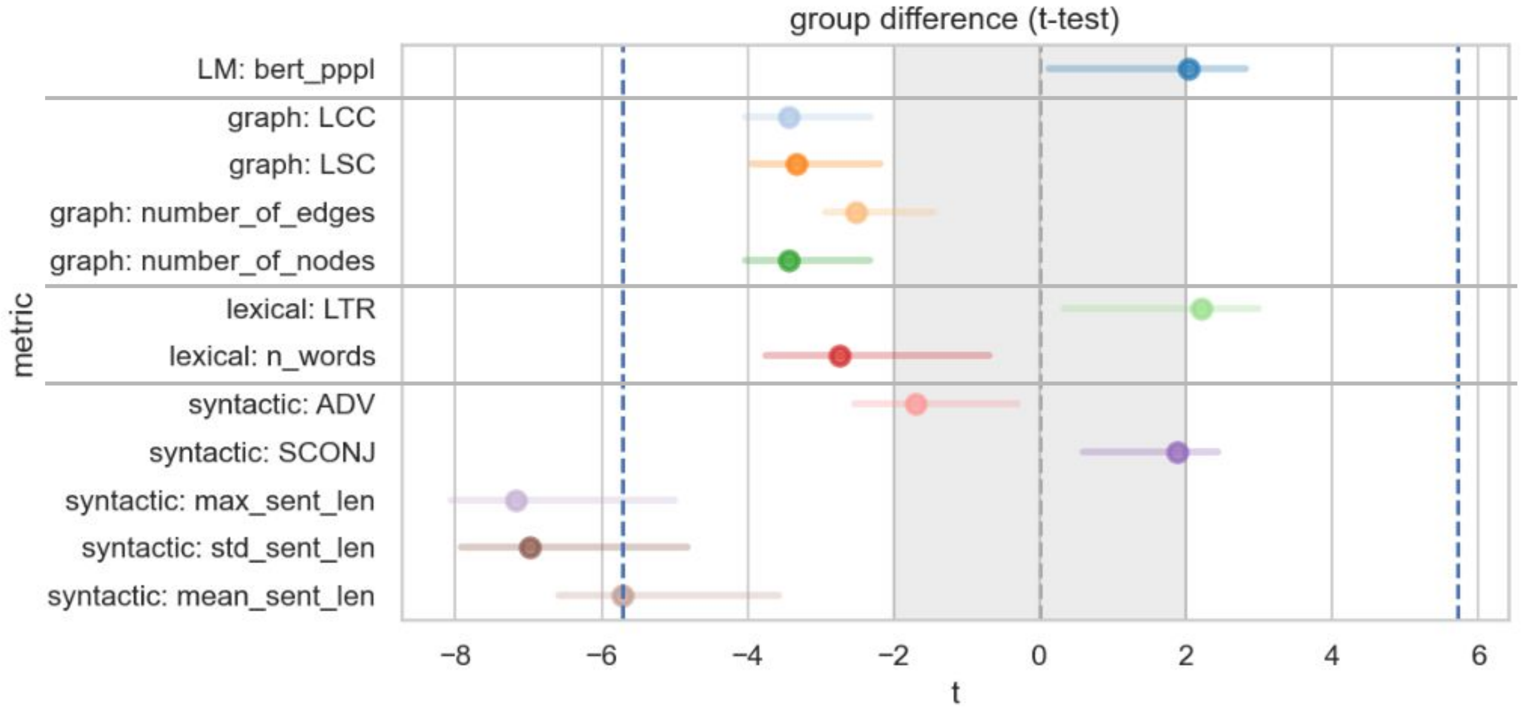
Analysis

Bootstrap to assess metric stability wrt influential points (1000 samples)

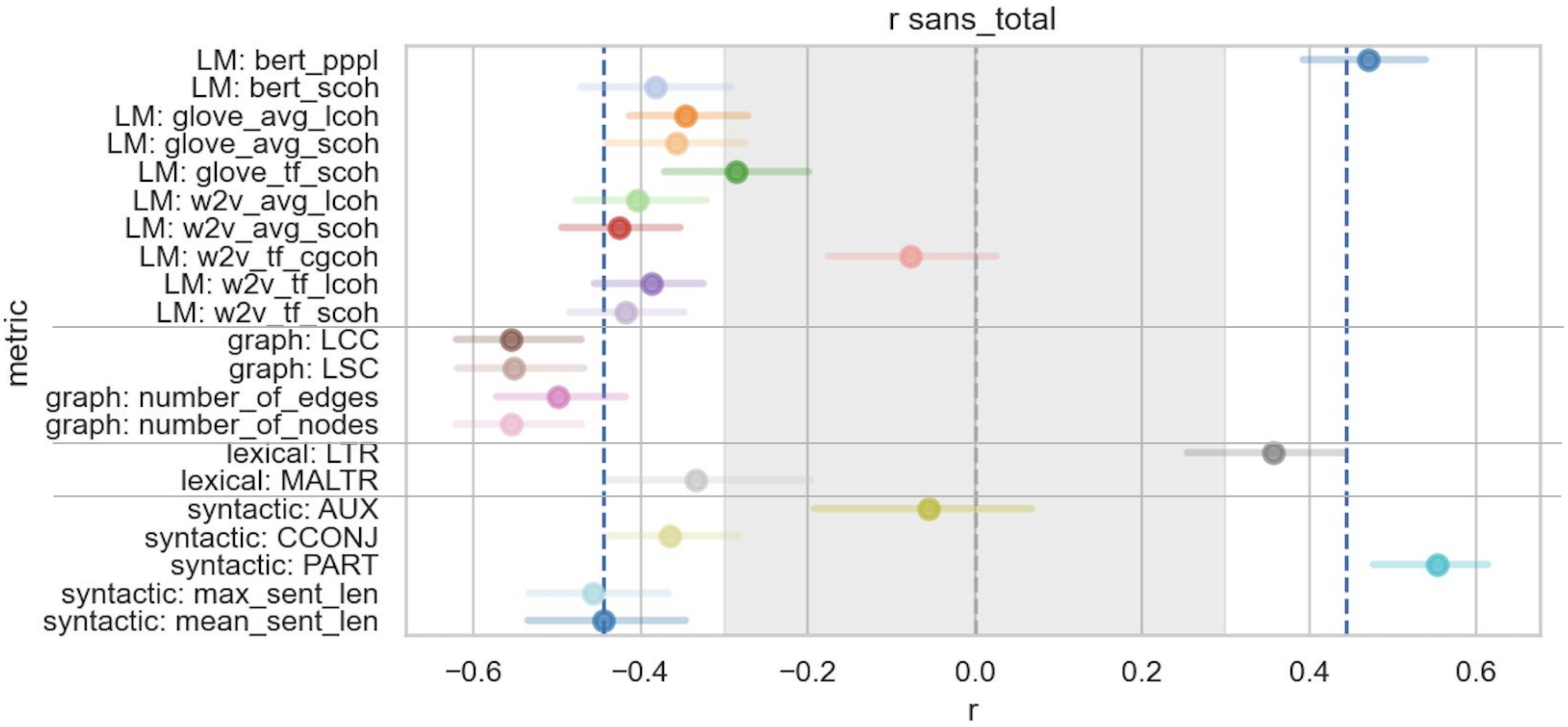
Analysis:

- two-sided t-test
  - NAP vs control
- r squared:
  - control variables
  - symptoms
  - mean sentence length
- pseudo-r squared for ordinal response variables:
  - TD and depression severity

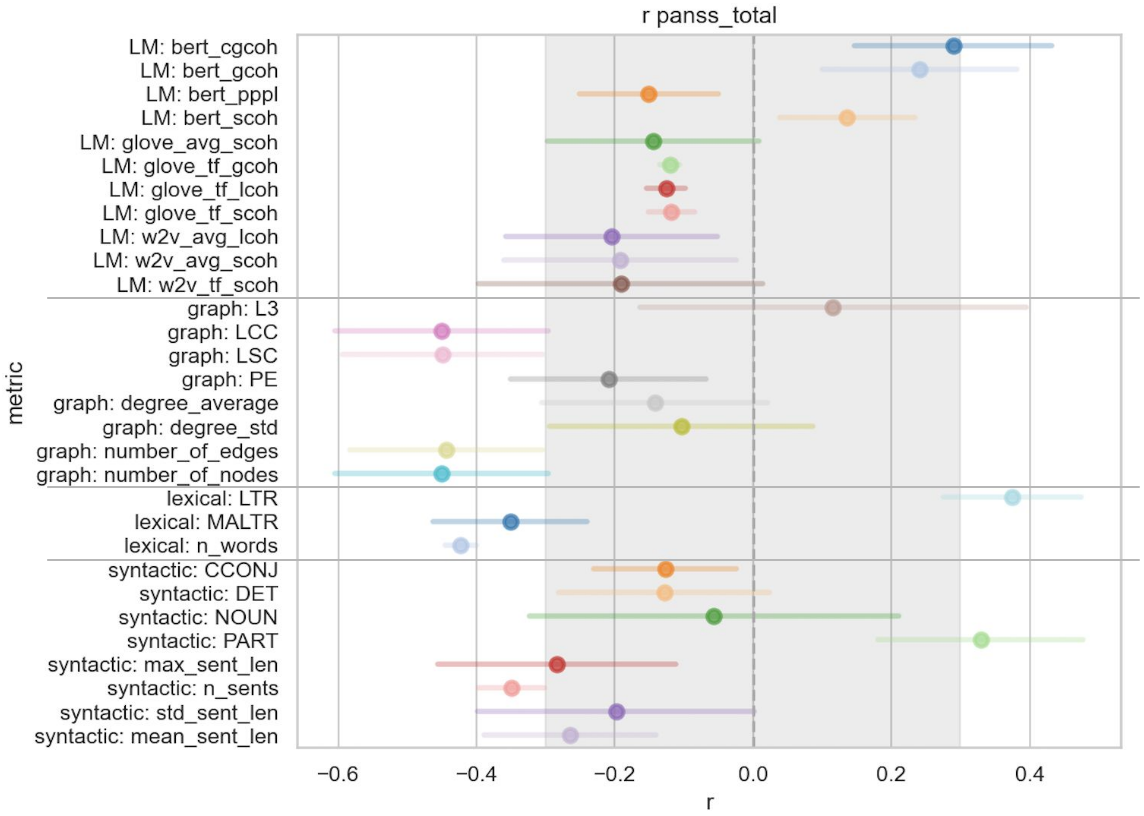
Results: German, cross-metric comparison (t)



Results: German, cross-metric comparison (R^2)



Results: Russian, cross-metric comparison (R^2)



Results: Relative Performance

German:

- t-test: sentence length metrics by far outperform all others
- correlation with symptoms:
    - graph, PART > syntactic, lexical, LM
    - best graph-based metrics (NN, NE, LCC, LSC), PART rate, and PPPL outperform the mean sentence length baseline

Russian:

- correlation with symptoms:
    - graph, lexical, PART > syntactic > LM
    - mean sentence length weak baseline; sentence count stronger baseline
    - best graph-based metrics (NN, NE, LCC, LSC) an lexical counts outperform sentence count

Results: Mean Sentence Length

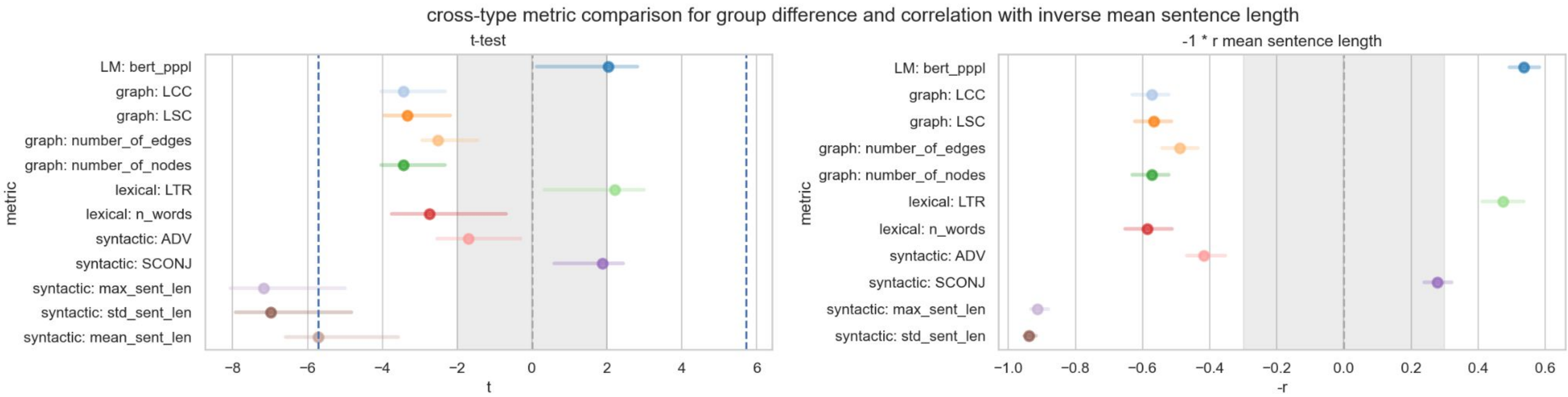
Many metrics inherently correlate with verbosity either mean sentence length or sentence count

- Graph: number of nodes is unique lemma count over a moving window
- Synt (POS): proportional to sentence length
- LM: vector averaging hypothesis
- Lexical: word count incorporates sentence length and count

Verbosity itself can be regarded as a reasonable baseline

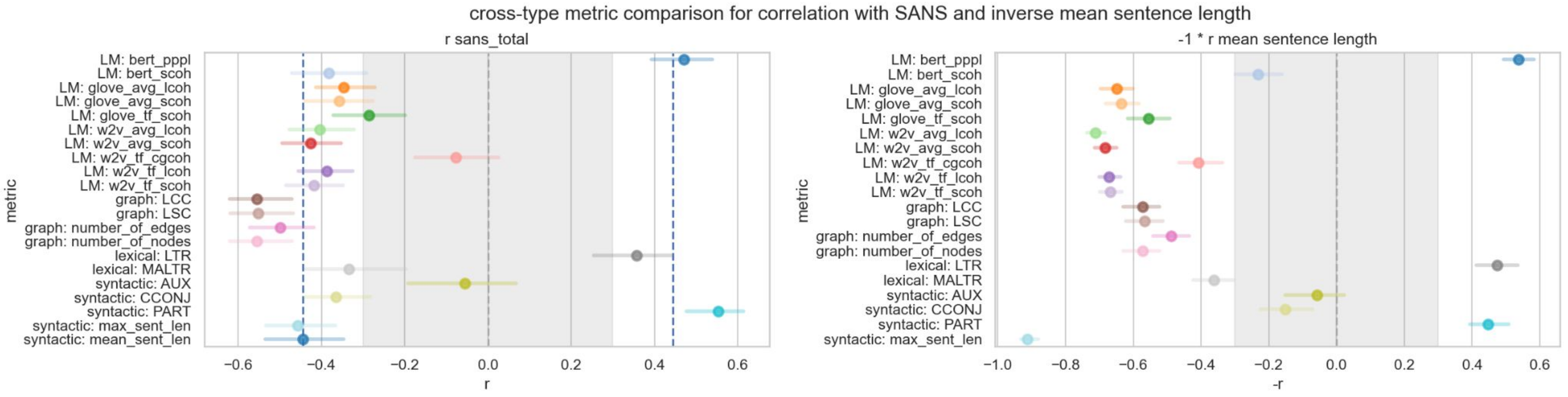
Results: German, Mean Sentence Length vs t-test

- general pattern of performance on t-test is proportional to correlation with mean sentence length



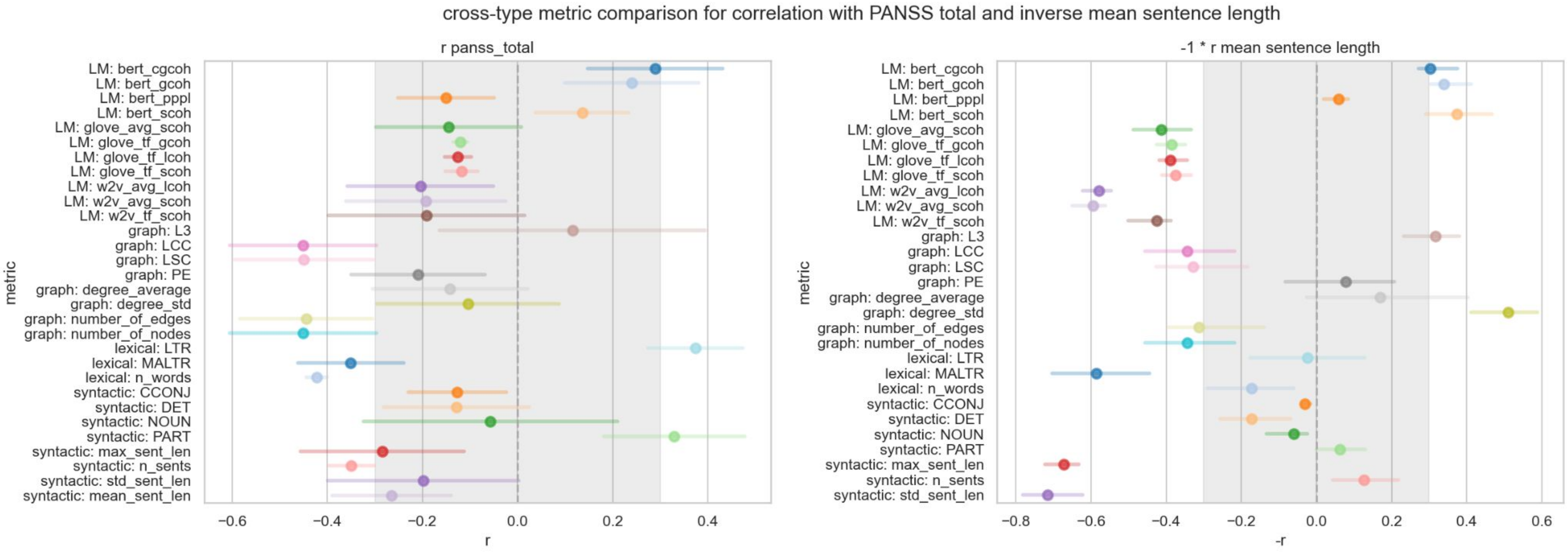
Results: German, Mean Sentence Length vs SANS r^2

- general pattern of performance is proportional to correlation with mean sentence length
- important exception: graph-based metrics
  - correlate less than LM, perform better

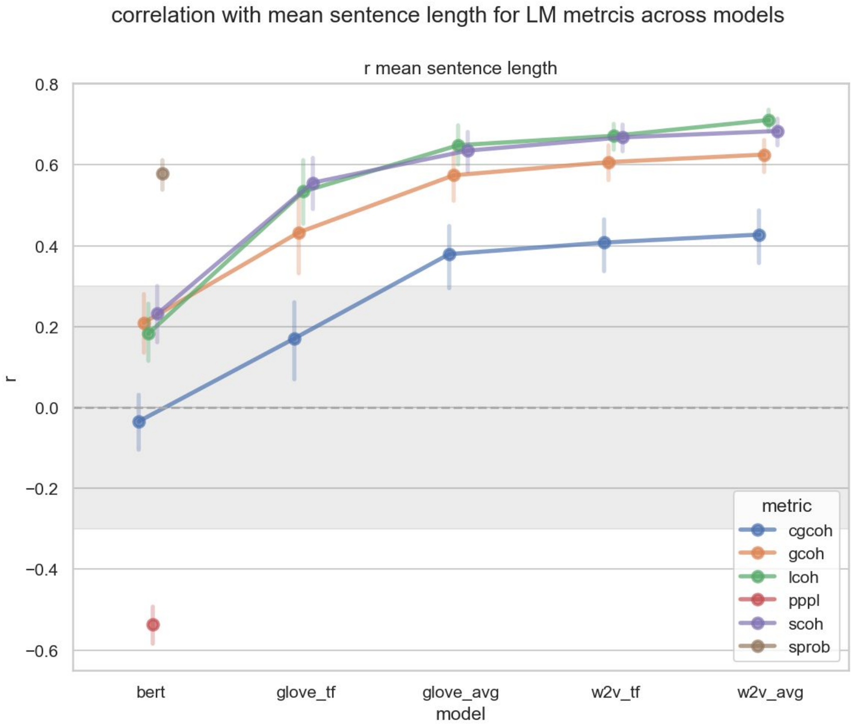


Results: Russian, Mean Sentence Length vs SANS r^2

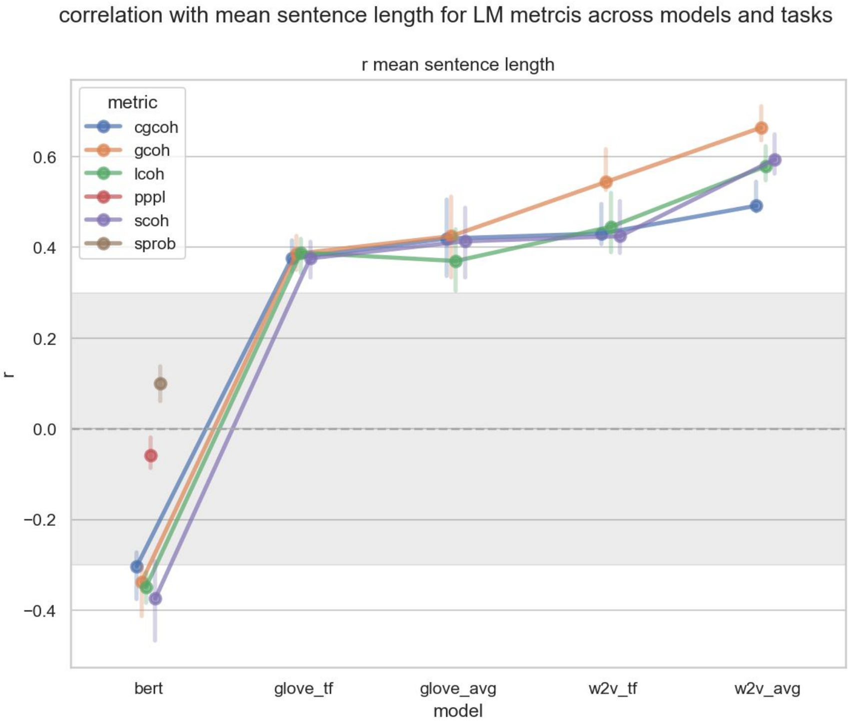
- similar but much weaker general pattern
- graph-based methods are still the exception



Results: Mean Sentence Length for LMs



German



Russian

Results: Metrics vs Scales

- German
  - best metrics work for both negative (SANS, PANSS_neg) and general scales (PANSS_o, PANSS total score)
  - lack of positive symptoms does not allow for thorough positive scale analysis (SAPS, PANSS_pos)

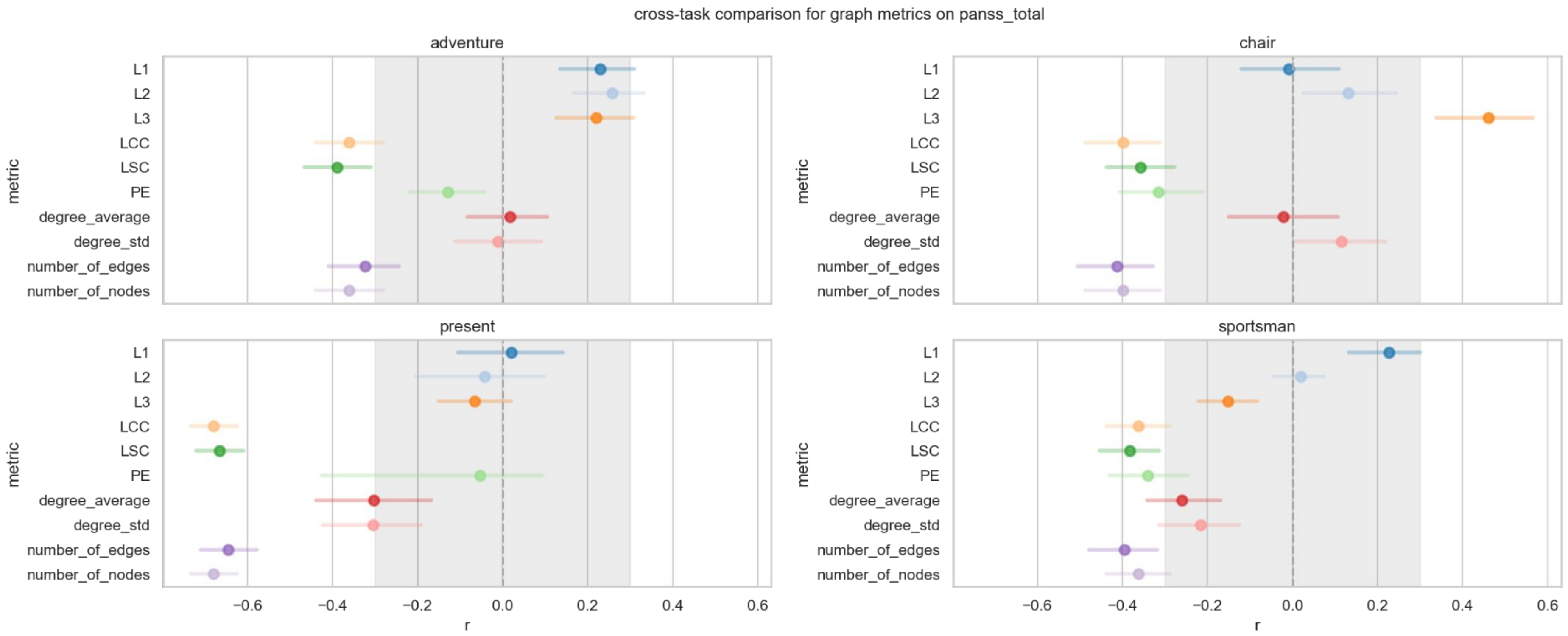
- Russian
  - best metrics typically work across all scales

Results: Russian, Metrics across Tasks

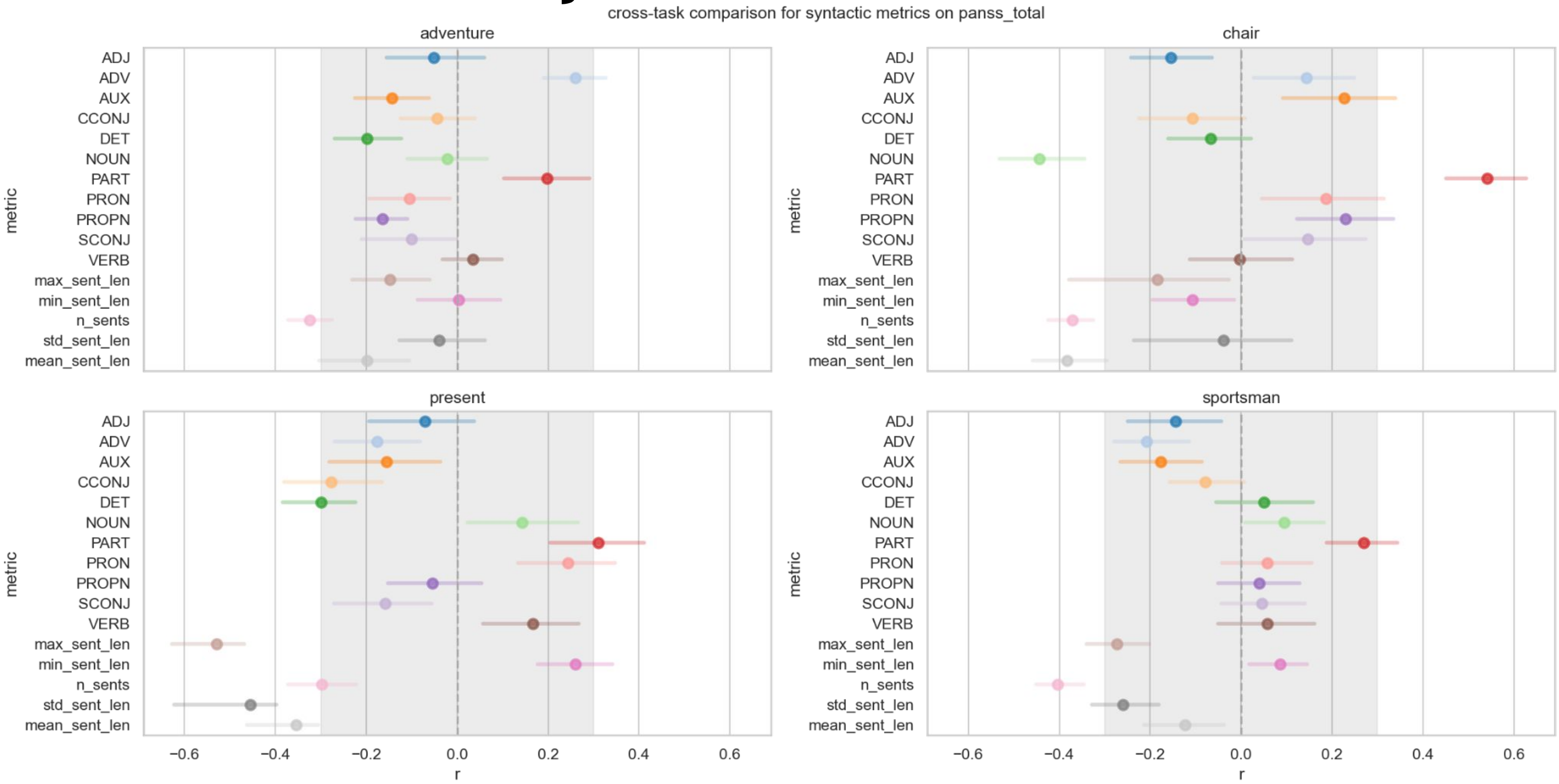
- Best performing metrics show more consistency across tasks
  - Graph: NN, NE, LCC, LSC
  - Syntactic: n sentences, PART, to some degree mean sentence length
  - Lexical: word count, LTR

- In other metrics high variability in performance across tasks

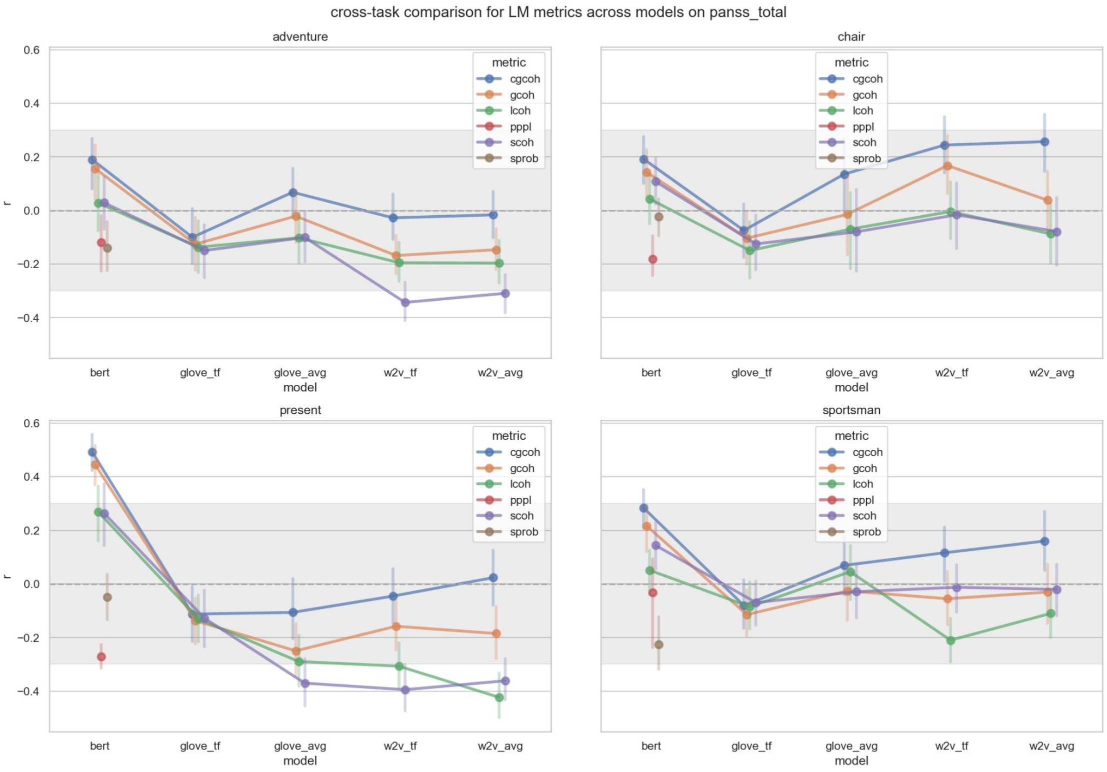
Results: Russian, Graph-Based Metrics across tasks



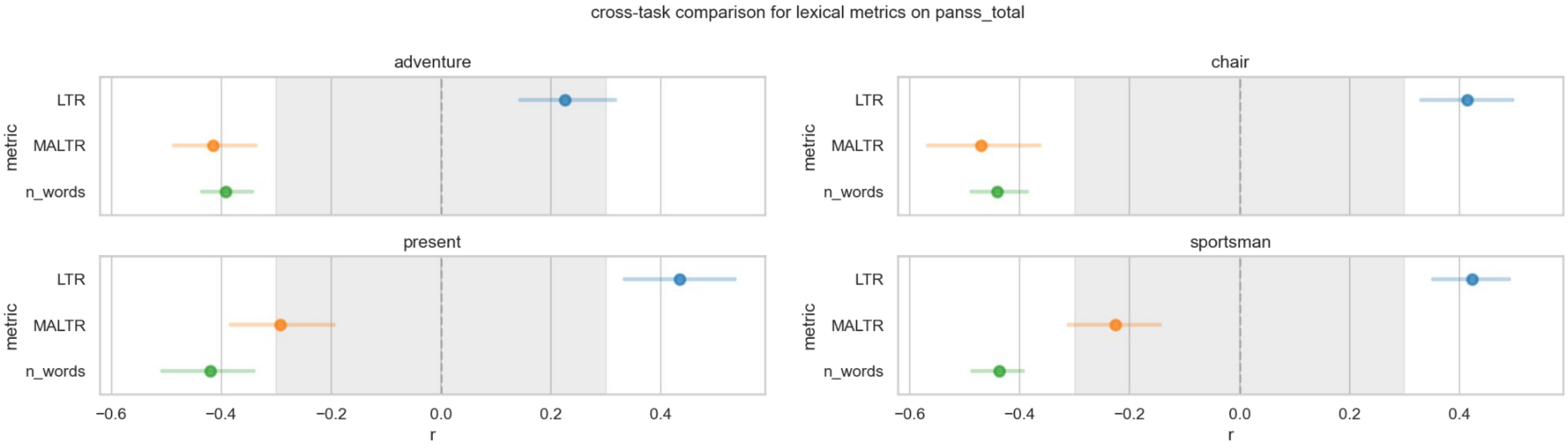
Results: Russian, Syntactic Metrics across Tasks



Results: Metrics across tasks



Results: Metrics across tasks



Results: Metrics across Languages

- Different tasks between languages
- Best performing metrics show more consistency across languages
  - Graph: NN, NE, LCC, LSC
  - Syntactic: PART
  - Lexical: word count, LTR
- Mean sentence length serves as a better baseline for German, and sentence count for Russian (both work for some tasks)
- Surprising BERT differences between languages (but not tasks)
  - direction of correlation with symptom severity and with mean sentence length
  - PPPL & next sentence prediction differ in correlation with mean sentence length

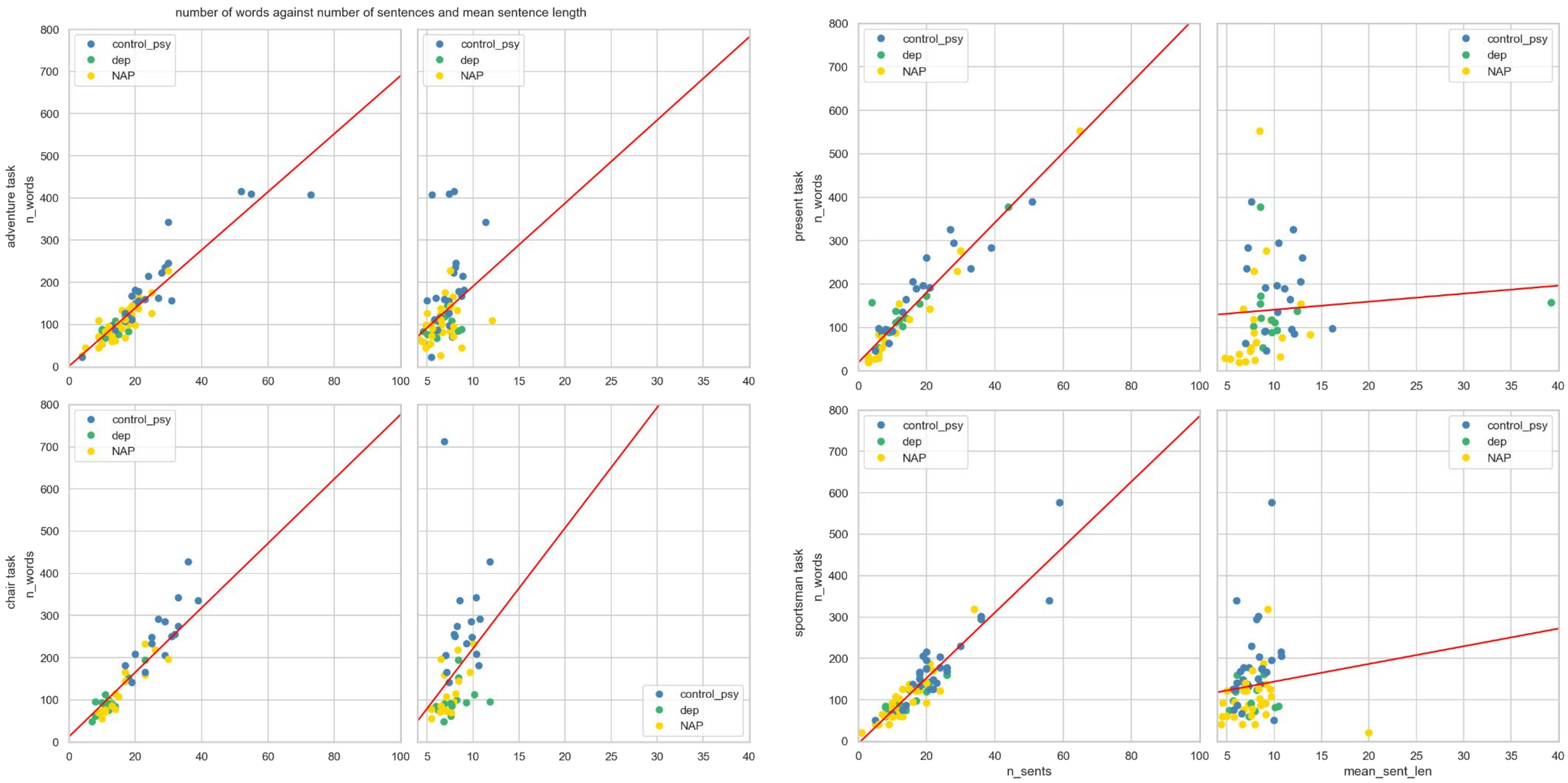
Results: Summary

- **which metrics work best:** graph > lexical, syntactic > LM
- **which metrics outperform the baselines:** within each group simple is generally better than complex
- **which metrics are associated with mean sentence length:** apart from graph-based metrics, correlation with mean sentence length is strongly associated with performance
- **which metrics work for which scales:** best metrics work for all scales
- **which metrics work across tasks:**
  - best metrics work for all tasks
  - large differences between tasks
- **which metrics work on both languages:**
  - graph (NE, NN, LCC, LSC), lexical (word count, LTR), PART
  - unexplained cross-linguistic differences in BERT performance

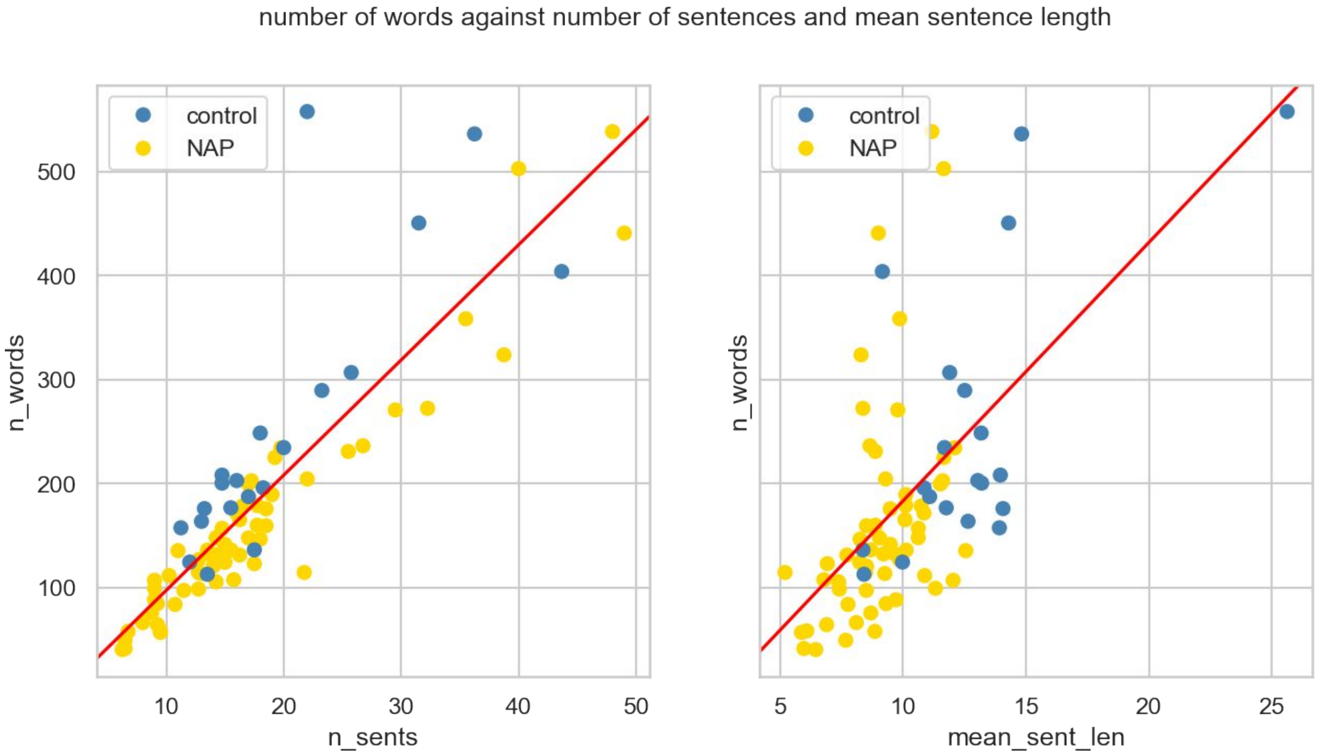
Bootstrap

- Various non-outlier influence points that render correlation unstable
- Bootstrap - subjects sampled with replacement for 1000 iterations
- 25-75% quantile interval to assess dispersion and approximate robustness of the metrics wrt influential points in question

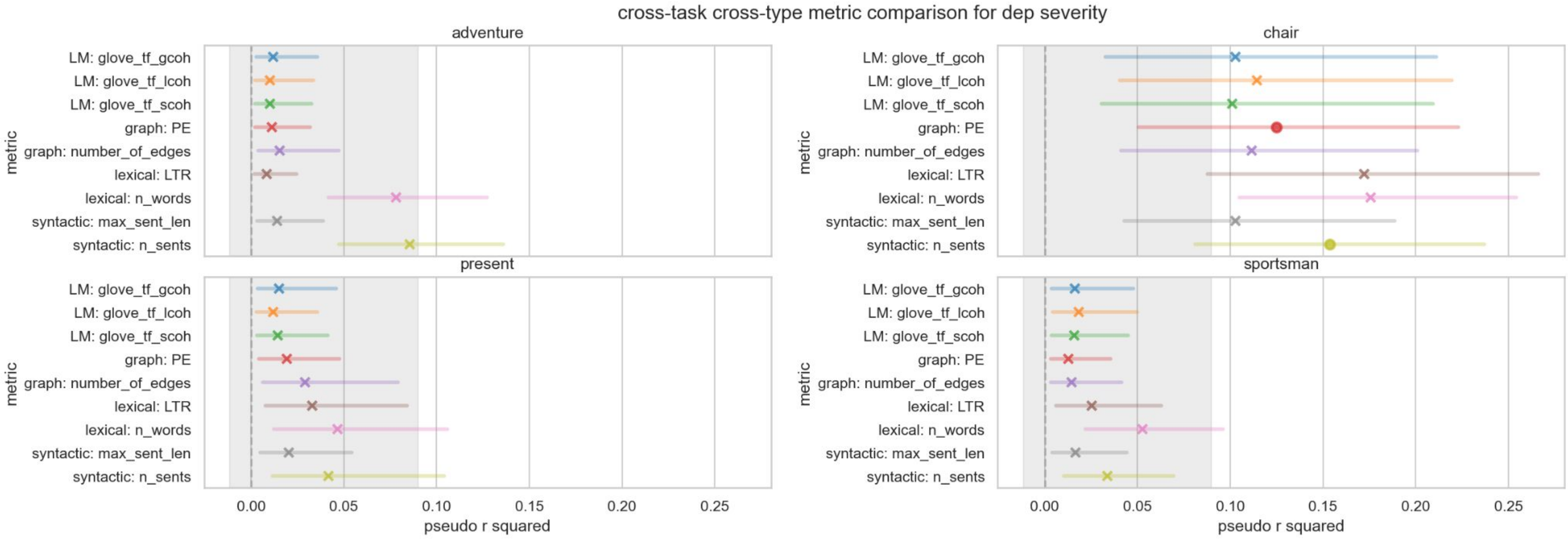
Sentence Length VS Sentence Count, Russian



Sentence Length VS Sentence Count, German



Russian: Depression Severity



Russian: TD Severity