UNIVERSITY OF POTSDAM

MASTER THESIS

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

---

Author:
Galina RYAZANSKAYA

1st Supervisor:
Sherzod HAKIMOV

2st Supervisor:
Prof. Dr. Manfred STEDE

A thesis submitted in fulfillment of the requirements for the degree of Master of Science in Cognitive Systems

June 22, 2024

Declaration of Authorship

I, Galina RYAZANSKAYA, declare that this thesis titled, "Cross-Methodological Comparison and Validation of Common Methods in NLP-based Psychosis Detection" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UNIVERSITY OF POTSDAM

Abstract

Faculty of Human Sciences

Department of Linguistics

Master of Science in Cognitive Systems

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

by Galina RYAZANSKAYA

The present study is dedicated to a cross-methodological and cross-linguistic comparison of the NLP metrics commonly used for the detection of psychosis and symptom severity assessment. The results partly agree with previous cross-methodological studies, suggesting that graph-based methods are the most reliable and reproducible, followed by lexical and syntactic methods. LM-based methods are demonstrated to be least reliable, as could be expected from the mixed results of the studies reporting on them previously. Most tested metrics are also shown to be dependent to some extent on the verbosity. These results suggest that simple metrics, such as word count, unique lemma count, sentence length and count, provide a strong baseline that few more complex metrics can beat.

Acknowledgements

I would like to thank Sherzod Hakiomov and Prof Dr Manfred Stede for agreeing to advise me on the present work.

I am very grateful to my psychiatric collaborators, Dr Sandra Just and Tatyana Shishkovskaya, for allowing me to work with the data they had collected, as well as to my former advisor, Mariya Khudyakova, for continuing to work on the psychiatric linguistics project we have started together even after my graduation.

Finally, I would like to thank my husband, Daniil Bobrovskiy, whose constant companionship and patience helped me immensely throughout this project. This work would have been impossible without all the long discussions we have had about it.

Contents

List of Figures

List of Tables

List of Abbreviations

CHR	**C**linical **H**igh **R**isk
E	Number of **E**dges
FEP	**F**irst **E**pisode **P**sychisis
FTD	**F**ormal **T**hought **D**isorder
HC	**H**ealthy **C**ontrol
LCC	**L**argest **C**onnected **C**omponent
LM	**L**anguage **M**odel
LSA	**L**atent **S**emantic **A**nalysis
LSC	**L**argest **S**trongly connected **C**omponent
LTR	**L**emma-**T**oken **R**atio
MALTR	**M**oving **A**verage **L**emma-**T**oken **R**atio
N	Number of **N**odes
NAP	**N**on-**A**ffective **P**sychisis
NLP	**N**atural **L**anguage **P**rocessing
PANSS	**P**ositive **a**nd **N**egative **S**yndrome **S**cale
PANSS_pos	**P**ositive **a**nd **N**egative **S**yndrome **S**cale, Positive Subscale
PANSS_neg	**P**ositive **a**nd **N**egative **S**yndrome **S**cale, Negative Subscale
PANSS_o	**P**ositive **a**nd **N**egative **S**yndrome **S**cale, General Subscale
POS	**P**art-**O**f-**S**peech
SANS	**S**cale for the **A**ssessment of **N**egative **S**ymptoms
SAPS	**S**cale for the **A**ssessment of **P**ositive **S**ymptoms
SIF	**S**mooth **I**nverse **F**requency
SD	**S**chizotypal **D**isorder
SDD	**S**chizophrenia **S**pectrum **D**isorder
SZ	**S**chi**z**ophrenia
SZA	**S**chi**z**o**a**ffective Disorder
TALD	**T**hought **a**nd **L**anguage **D**iosorder Scale
TD	**T**hought **D**isorder
TF-IDF	**T**erm **F**requency - **I**nverse **D**ocument **F**requency
TLC	**T**hought, **L**anguage, and **C**ommunication Scale
TTR	**T**ype-**T**oken **R**atio
w2v	**w**ord**2v**ec

Chapter 1

Introduction

The aim of the present work is to benchmark the commonly used natural language processing methods of detecting psychosis and its symptoms. The introduction is structured as follows. First, I discuss psychotic disorders, the role that incoherent language plays in diagnosing and detecting them, the ways of conceptualizing the incoherence, and the possible psychiatric explanations for the origin of the incoherent language (1.1). Then, I turn to the various NLP approaches that have been suggested for psychosis detection over the years, indicating for each approach, which symptoms or conceptualizations of psychotic incoherence the approach tries to approximate (1.2). Finally, I review some issues with both internal and external validity of the suggested approaches (1.3) and conclude with a short motivation for the present work (1.4).

1.1 Incoherent Language in Psychosis

Psychosis is a common functionally disruptive symptom of many psychiatric conditions. It is characterised by delusions, hallucinations, and formal thought disorder. Psychosis is the defining feature of schizophrenia spectrum disorders (SSD), such as schizophrenia, schizoaffective disorder, and schizotypal disorder, and a common but variable feature of mood disorders,

such as bipolar disorder (Arciniegas, 2015)[1].

Kraepelin et al. (1919), defining what would later be referred to as schizophrenia, stated that derailment, loose associations, and incoherence of thought manifest themselves through disordered speech, and Bleuler stressed the importance of aberrant language as an important feature of schizophrenia (Bleuler, 1911). These features of psychotic speech and thought are covered by the term formal thought disorder (FTD) introduced by Andreasen (1986a). Importantly, incoherent or disordered speech serves as a key diagnostic criterion for schizophrenia spectrum disorders in the two major diagnostic manuals, the International Classification of Diseases (ICD-10[2]) and the Diagnostic and Statistical Manual of Mental Disorders (DSM-5[3]). A general overview of language anomalies characteristic of schizophrenia can be found in Kuperberg (2010a), Kuperberg (2010b), and de Boer et al. (2020).

1.1.1 Psychosis and Formal Thought Disorder

Formal thought disorder conceptualization of speech and thought anomalies observed in psychosis is often used as the guiding principle for developing automated psychosis detection tools. Formal thought disorder is defined as a set of specific disturbances in thought, speech, and communication that are typical of psychosis (Hart et al., 2017). The symptoms of thought disorder can be divided into positive and negative.

In general, positive symptoms are the symptoms that are not experienced normally but are typically present during a psychotic episode. These include delusions, hallucinations, and disorganized thoughts and speech. Positive thought disorder is the respective subset of FTD features that includes positive symptoms manifesting in speech and thought such as *incoherence*, *tangentiality*, *circumstantiality*, and *peculiar word use*.

[1] As most papers test the methods on schizophrenia spectrum disorders, I will use the terms SSD and psychotic disorder interchangeably and specify separately if non-SSD disorders are included in a study.

[2] Organization, 1992

[3] Association, 2013

Negative symptoms are the symptoms that manifest deficits in emotional responses or thought processes. Negative thought disorder includes such negative symptoms of FTD as *poverty of speech and speech content*, *blocking*, and *perseveration*.

The positive and negative symptoms that are typical of psychotic disorders are usually assessed with the positive and negative syndrome scale (PANSS, Kay et al., 1987) or using the two scales for the assessment of positive and negative symptoms, abbreviated as SAPS (Andreasen, 1986b) and SANS (Andreasen, 1989). While PANSS only identifies the language-specific disturbances very broadly (using *conceptual disorganization* for all positive speech disturbances and *lack of spontaneity and flow of conversation* and *stereotyped thinking* for the negative ones), SAPS and SANS include detailed subscales to assess the most common positive and negative anomalies of speech and thought typical of positive and negative FTD.

1.1.2 Conceptualizations of Incoherent Language

There is no unified linguistic theory to encompass all the features that contribute to coherence. Defining incoherence as 'speech that is essentially incomprehensible at times', Andreasen (1986a) states that incoherence may arise through 'several different mechanisms, which may sometimes all occur simultaneously'. This is because coherence is normally maintained simultaneously on many levels, including lexical connectors, syntactic structure, intonation, reference, and logical structure of a text. As a result, different researchers rely on different conceptualizations when they develop automated psychosis detection tools.

Some researchers rely on broad conceptualizations, such as *global coherence* defined as the relationship of every clause to the overall topic of the text (Glosser et al., 1991), *local coherence* defined as similarity in content and logical connectedness of adjacent sentences, or *cohesion* which encompasses grammatical and lexical linking within a text that holds it together.

Others use the scales aimed at measuring formal thought disorder as a source of conceptualization. There are a number of scales that aim at identifying different features of formal thought disorder: including TLC[4], SANS & SAPS[5], TLI[6], TALD[7], and K-FTDS[8]. There are several linguistic features that are underlined in all of these scales and are commonly used in automated psychosis detection research, which I discuss below. I cover these features here so that in discussing the existing metrics, I can refer to them, indicating which manifestations of positive and negative thought disorder could potentially be detected by higher or lower metric values.

Commonly used positive FTD features in these scales include many features indicating some form of inability to stay on topic. It may come in the form of being distracted by nearby stimuli interrupting the flow of speech (*distractible speech*); in the form of ideas slipping off track onto ideas obliquely related or unrelated (*derailment*, as well as *loosening of associations* and *flight of ideas*); or in the form replying in an oblique or irrelevant manner (*tangentiality*). Additionally, the inability to stay on topic can take the form of *circumstatiality*, where speech is very indirect and delayed in reaching its goal idea. The disconnectedness between speech fragments may take the form of *illogicality*, where the conclusions are reached that do not follow logically from the premises, or general *incoherence*, indicating speech that is essentially incomprehensible at times and may manifest in its severest form as schizophasia. Incoherence may stem from syntactic or grammatical errors, semantically unmotivated word choices, and lack of proper cohesion devices, such as coordinating and subordinating conjunctions or coreferences. Additionally, the rate and amount of speech might be increased, which is referred to as *pressured speech*. Some of the common positive FTD features also emphasize peculiar word choice, manifesting as incorrect or unconventional word use and phonetic associations (*paraphasias*), as well as the invention of new words (*neologisms*).

[4]Thought Language and Communication scale Andreasen (1986a)
[5]Andreasen, 1989; Andreasen, 1986b
[6]Thought and Language Index, Liddle et al. (2002)
[7]Thought and Language Disorder Scale, Kircher et al. (2014)
[8]Kiddie Formal Thought Disorder Rating Scale, Caplan et al. (1989)

Commonly used negative FTD features in these scales include general *poverty of speech*, which indicates decreased amount and rate of speech, brevity, low verbosity; *poverty of speech content* referring to vague, overly concrete or overly generalized speech, conveying little information[9]; and inability to reach the goal, which may be a result of sudden interruption (*blocking*) or slow drift (*loss of goal*). The patients may also have a tendency to repeat ideas or words over and over, which is referred to as *perseveration* or *verbigeration*.

Finally, some researchers test a broad variety of linguistic features to identify the ones that might help detect psychosis without referring to any pre-existing theories regarding these features.

It is not entirely clear whether negative or positive FTD aspects of language are more easily detected by NLP methods. In the present work, I set out to analyze the most commonly utilized features regardless of their conceptualization. For the features that clearly correspond to a facet of FTD, it is reported whether they are expected to be higher or lower in patients and whether this, in general, holds true across the features corresponding to an FTD symptom.

1.1.3 Theories of Incoherence in Psychosis

In a dedicated review, Ditman et al. (2010) outline two main types of theoretical frameworks explaining the origins of discourse incoherence observed in schizophrenia: executive dysfunction theories (also known as impaired cognition theories); and loose association theories. The former theories state that the lack of control over the process of thinking, typical of negative thought disorder, is the driving reason for the observed incoherence of speech in FTD. The latter, on the other hand, explain the incoherence in terms of tangentiality and loose associations that are characteristic of positive thought disorder. However, there is no definitive evidence for either theory being the main cause of the speech incoherence seen in schizophrenia spectrum disorders. Moreover, aspects of both positive and negative formal thought disorder may be contributing to incoherence in any given patient. The lack of clear

[9] Both poverty of speech and poverty of content may be sometimes referred to as *alogia*.

evidence for either theory is partially due to the inter-dependency between processes involved in speech production and partially due to the multiplicity of processes reflected in speech. It is further complicated by the fact that coherence heavily depends on both textual and social context (Cohen et al., 2017). Interestingly, there is modelling evidence for different NLP methods being more sensitive to different aspects of FTD, which could be regarded as indirect evidence for both mechanisms contributing to the incoherence that is typical of psychosis (Fradkin et al., 2023).

1.2 Natural Language Processing for Psychosis Detection

Natural language processing offers a variety of automated methods and tools for working with various aspects of language. Utilizing these tools for psychosis detection has been pioneered by Elvevåg et al. (2007) and many different metrics have been developed since, though reproducibility remained limited (Hitczenko et al., 2020; Parola et al., 2023; Fradkin et al., 2023). In the present work, psychosis detection is used as a general term for several somewhat different tasks: firstly, distinguishing between healthy controls and patients with a psychotic disorder, secondly, predicting conversion in populations at high risk for psychosis, and, finally, predicting the differences in symptom prevalence or severity between different patients. Natural language processing approaches to psychosis detection supposedly offer a fast, sensitive, and objective solution to psychosis detection and symptom severity assessment. However, such approaches must be applied with great caution, as prognostic assessment may significantly affect or even stigmatize the patients and the results may be sensitive to many external factors (Just et al., 2023), a topic further discussed in section 1.3.

Below, I introduce some of the natural language processing methods that have been applied to the task of psychosis detection. The methods are grouped, according to the type of linguistic material they operate on, into lexical, syntactic, and semantic. The metrics are discussed in greater detail in the next

chapter, 2. Note, that this work is mostly concerned with the linguistic features that occur both in written and spoken texts, leaving the acoustic[10] and temporal[11] features outside of the scope of the present work.

1.2.1 Lexical Methods

Under lexical methods I include all the methods that operate over words and focus on word use. These include *verbosity* or word count; *lexical diversity* metrics, such as type-token ratio (TTR) or average word frequency; *sentiment analysis* metrics, such as use of negative or positive emotion words; and *topic analysis*, usually assessed with a specified tool based on dictionaries. The tools used for lexical analysis often incorporate a variety of word-based metrics, LIWC having 72 metrics (Tausczik et al., 2010), TAACO - 34 (Crossley et al., 2018), and Coh-Metrix L3 - 108 (McNamara et al., 2014).

The lexical methods may aim at capturing both positive and negative FTD features. Verbosity in penitents may either be increased (pressure of speech) or decreased (poverty of speech). Lexical diversity and density may be increased (paraphasias) or decreased (poverty of speech and content). Sentiment analysis may reveal the blunted affect which is a typical non-linguistic negative symptom of schizophrenia. Finally, topic analysis may reveal common topics shared between different patients which may be related to both negative and positive symptoms.

1.2.2 Syntactic Methods

Under syntactic methods, I include methods relating to syntactic types and structures. These include *part-of-speech* based metrics; *referential* metrics, such as the use of ambiguous pronouns and cataphora[12]; measures of *syntactic*

[10] Acoustic features, such as spectral, frequency, and amplitude parameters were explored in Voppel et al. (2023), Tang et al. (2023b), Boer et al. (2023), and Wouts et al. (2021).

[11] Such temporal features as speech rate and pausation patterns were used by Aich et al. (2022), Liebenthal et al. (2023), Boer et al. (2023), and Tang et al. (2023b) to distinguish between patients and controls as well as predict symptom severity assessed by psychiatric scales.

[12] Cataphora is the use of the referential phrase that refers to or stands for a later word or phrase as in "If you want *them*, there are *cookies* in the kitchen".

complexity calculated based on use of various syntactic structures such as coordinated or subordinated clauses as well as the use of negation or passive and active voice; and, finally, the *lengths and counts* of various syntactic units, such as clauses and sentences.

The differences in the relative frequency of parts of speech use commonly indicate reduced syntactic complexity, which may be connected with negative FTD (poverty of speech), while reduced use of descriptive words (i.e. adjectives and adverbs) as well as reduced use of referential devices, which can be regarded as indicative of poverty of content. Similarly, reduced syntactic complexity may be indicative of poverty of speech or speech content. On the other hand, ambiguous pronouns may be one of the driving forces behind perceived incoherence, caused by the loosening of associations, typical of positive FTD.

1.2.3 Semantic Methods

Under semantic methods, I include methods operating on the level of the entire text and trying to assess some of its properties based on its meaning. The semantic methods can be further divided, methodologically, into the graph-based methods and the ones that use language models (LMs) to represent the text.

Graph-Based Methods

Graph-based methods include all the methods that represent a text with a graph and subsequently calculate properties of the graph, such as size or some measure of connectivity. The graph representations can be based on the *co-occurrence* of words in the text, with words used as the nodes of the graph and neighbouring words being connected by a directed edge. Alternatively, the graphs may be based on a *semantic* representation of the text, which can be obtained with semantic role labelling tools (SRL, Gildea et al., 2002).

Reduced graph connectivity is typically regarded as indicative of positive FTD symptoms such as incoherence or derailment and flight of ideas, as,

when very little is said on any given topic, the resulting graph is largely disconnected. On the other hand, very high connectivity may be indicative of negative FTD symptoms of poverty of speech content or perseveration.

Language Model-Based Methods

Language model-based methods use mathematical models to represent the texts numerically and consequently calculate some metric over the representation. The models are trained on large corpora to approximate probability distributions over sequences of linguistic units. They can learn and encode some of the semantic and syntactic features of the texts based on the co-occurrence of the words and sentences in the corpus. These models can be used to represent words or larger text fragments with vectors which are called *embeddings* and which encode the properties learnt by the model. The embeddings can be divided into *non-contextualized* which operate like dictionaries of words to vectors and *contextualized* that let the context in which a word appears to influence the embedding vector, allowing for disambiguation and better handling of unseen words. Non-contextual embeddings include representations obtained from n-grams, latent semantic analysis (LSA, Landauer et al. (1998)), word2vec (w2v, Mikolov et al. (2013)), GloVe (Pennington et al., 2014) for word representation and sent2vec (Moghadasi et al., 2020) for sentence representation. Contextualized embeddings can be obtained from BiLSTM and ELMo (Peters et al., 2017), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and other attention-based models (Vaswani et al., 2017; OpenAI et al., 2024).

There are several ways in which language models can be applied to psychosis detection. First, there are *embedding-based metrics* that represent words or sentences with vectors using an LM and then calculate some metric over the vectors, usually based on the cosine similarity of these vectors. Such metrics include the similarity of consecutive units, the similarity of units at a fixed distance, the similarity of a given unit to some standard unit, and the slope of difference of units from the question or initial prompt. Secondly, LMs can provide some measure of *text predictability* such as perplexity or probability from BERT next sentence prediction. Finally, LMs can be *fine-tuned* to be

used as *classifiers* for psychosis detection or symptom severity assessment. The embeddings can also be used as the input to a simple classifier, which I will refer to as *non-fine-tuned classifiers*.

The similarity-based metrics typically try to capture positive FTD symptoms such as incoherence or derailment as well as tangentiality. Lowered text predictability may also be seen as indicative of incoherence or derailment while higher values may indicate negative symptoms such as perseveration and poverty of speech content.

1.3 Theoretical Validity of NLP Approaches to Psychosis Detection

In this section, I outline some of the questions about the validity of the proposed NLP approaches. In psychology, *internal* validity examines whether a study answers the research questions without bias and *external* validity examines whether the study findings can be generalized to other contexts (Andrade, 2018). Applying these definitions to the context at hand, a metric would be internally valid if it successfully approximates the concept, such as coherence or poverty of speech (*construct* validity); and is not biased by any known confounding factors. The external validity of a metric would refer to the ability of the metric to be applied in a variety of contexts, including such external factors as sensitivity to the choice of pre-processing, embedding model, and elicitation task (1.3.2), as well as cross-linguistic applicability (1.3.3).

1.3.1 Internal Validity

In order to apply the NLP-derived metrics to clinical settings, it is important to assess the internal validity of the metrics that rely on the psychiatric or linguistic conceptualizations mentioned above. This could be achieved by correlating the metrics with human ratings, as suggested in the foundational study Elvevåg et al. (2007), or with the respective TLI, TLC, TALD, or SANS & SAPS subscales (see Bilgrami et al. (2022) for a study of construct validity

in psychosis detection). Additionally, it is important to assess if the metric is associated with positive or negative FTD as would be expected based on the conceptualization. This could be achieved by measuring positive and negative symptoms in general using PANSS, as done in most studies in the field, or using the aforementioned thought disorder scales, such as SANS & SAPS. Though many studies use some of the thought disorder scales , few correlate the metrics with the TD subscales but rather use the overall score. While some studies report correlation with the expected subscales (e.g. Vail et al. (2018), Just et al. (2020), and Jeong et al. (2023)), many fail to find such relations (e.g. Bedi et al. (2015), Corcoran et al. (2018), Iter et al. (2018), and Hitczenko et al. (2020)). In fact, several reviews point out that the relations between psychiatric constructs of language disturbances and the metrics might not be straightforward (Cohen et al., 2017; Holmlund et al., 2023).

As for confounding factors, introducing possible biases, many of the proposed metrics are known to be sensitive to such factors as sentence or text *length*[13], *repetitions*[14], and *preprocessing*, including sentence segmentation and removal of stop words and filler words[15]. Such dependencies stem naturally from the definitions of some of the LM-based and graph-based metrics. For example, cosine similarity-based metrics yield higher results on more repetitive texts and longer sentences. Surprisal or perplexity is also known to be higher on longer sentences. The number of nodes in graph-based metrics is closely related to the total number of words. As for preprocessing, any metric that relies on sentence boundaries is reliant on the human annotator's decisions if the original text is spoken rather than written. All these factors are quite problematic in the context of psychosis detection. Patients with SSD are commonly reported to be less verbose, producing significantly fewer words in most studies that report verbosity and often produce significantly fewer or shorter sentences. The repetitiveness which may be indicative of schizophrenia (perseveration) could artificially drive the cosine-similarity metrics up.

[13] Elvevåg et al. (2007), Mota et al. (2014), Iter et al. (2018), Corcoran et al. (2018), Hitczenko et al. (2020), Parola et al. (2023), and Jeong et al. (2023)

[14] Elvevåg et al. (2007), Iter et al. (2018), and Just et al. (2019)

[15] Parola et al. (2023) and Holmlund et al. (2023)

The verbal filler production may also be altered in some SSD patients introducing a problematic preprocessing dependence.

Some metrics might also not be sensitive to features that they should presumably be sensitive to, such as word or sentence shuffling or word replacement. There are studies that explore artificial perturbations and language modelling as a way of testing construct validity of LM-based metrics assessed using correlation with manual scores (Fradkin et al., 2023) or test for differences introduced by perturbations of control texts (Bedi et al., 2015; Hitczenko et al., 2020).

A notable complication for construct validity assessment for some of the metrics is the fact that the interpretation cannot be linear. While high scores along psychiatric dimensions of tangentiality or incoherence always indicate abnormality, the metrics might have 'optimal' values, while both low and high values outside this range could be considered abnormal (Fradkin et al., 2023). For example, very low perplexity would indicate repetitive or stereotyped speech and very high perplexity would indicate incoherence or word salad. Conversely, very low cosine similarity metric scores would indicate incoherence and very high scores would indicate high repetitiveness. Thus, unlike psychiatric scales, both high and low metric scores may indicate abnormality.

1.3.2 External Validity

The assessment of external validity for most metrics is limited by the diagnostic and design dissimilarities in the studies that report contradictory results for a given metric.

First of all, the dissimilarities may concern the population the study was conducted on. The studies may investigate in- or out-patients, patients with or without manifest thought disorder, and with prevailing positive or negative symptoms. The diagnostic categories used in each study also vary from one diagnosis (e.g. schizophrenia) to broader diagnostic categories such as

schizophrenia spectrum disorder or non-affective psychosis[16] and psychotic disorder[17]. Additionally, some studies explore patients at high risk for psychotic disorders[18] or healthy relatives (Elvevåg et al., 2010). The proportion of female patients, as well as average age, racial identity, education, and income level, may also vary greatly from study to study, and some of the metrics are reported to be sensitive to such factors (Hitczenko et al., 2020; Palaniyappan, 2021; Minor et al., 2023).

The study design may vary with respect to elicitation tasks used to obtain speech samples, with some studies using spoken and some written discourse, some utilizing monologue and some dialogue, and some obtaining very short and some very long speech samples. The speech may also be structured or unstructured, connected to some topic for all patients or entirely free from any prompt. The most typical tasks include picture description tasks and semi-structured interviews. The differences stemming from the elicitation tasks can be observed in most studies using several tasks[19].

The psychosis detection tasks used in a study may include tests identifying group differences between patients and controls, correlation with symptom severity, or predictive models for either of these tasks. The studies with clinical high-risk populations also often feature prediction of conversion to psychosis, and longitudinal studies try to predict disease course or longitudinal symptom severity.

Finally, the studies also differ with respect to the suite of metrics applied in the study, including the choice of embedding models, as well as the choice of preprocessing, such as transcription protocols, punctuation, filler and stop word removal. Some of the differences observed in the studies may also stem

[16] Includes schizophrenia (SZ), schizoaffective disorder (SZA), schizophreniform disorder (SFD), and schizotypal disorder (SD).

[17] Additionally includes the affective disorders with psychotic symptoms, such as some manifestations of bipolar disorder

[18] Bedi et al., 2015; Rosenstein et al., 2015; Corcoran et al., 2018; Gupta et al., 2018; Rezaii et al., 2019; Haas et al., 2020; Hitczenko et al., 2020; Bilgrami et al., 2022

[19] Such task dependence is observed in many studies including Elvevåg et al., 2007; Mota et al., 2016; Mota et al., 2017; Just et al., 2019; Ryazanskaya, 2020.

from cross-linguistic differences or differences in the availability and quality of NLP tools for different languages.

It is worth noting, that even the models that do not aim at approximating any psychiatric construct and simply try to identify group differences or predict symptom severity are subject to external validity limitations. As the datasets in the field are typically small[20] and the study designs are very diverse, the trained classifiers are likely to be inapplicable to any other study or setting. Such models lack transferability, limiting their practical value, and could also lack interpretability, requiring additional analysis for any theoretical value.

1.3.3 Cross-Linguistic Applicability

Another important consideration when it comes to external validity is the cross-linguistic applicability of the proposed methods.

First of all, there are ways in which some methods are inherently inapplicable to some languages. For example, while the use of determiners has been used as an indicator in many papers analyzing English (e.g. Bedi et al., 2015; Tang et al., 2021; Bilgrami et al., 2022), Polish does not have an equally broad category of determiners and thus this metric can not be used on a Polish-speaking population (Sarzynska-Wawer et al., 2021). Similarly, the frequency of use of some part of speech present in both languages or some construction, such as cataphora, may differ between the languages, making a metric that is useful in one language practically inapplicable in another. Incoherence may manifest differently in different languages, and what would be incoherent in one language may be more or less normal in another.

Secondly, there are some limitations imposed by the tools used for natural language processing. Many methods that may be applicable across languages, would require that the tools that calculate them be translated for every new language which is time-consuming and in many cases practically infeasible. Some tools are indeed translated, such as LIWC which is translated into 15 languages. However, it is not a free open-source tool, which

[20] Mode - 59 patients and 44 controls, mean - 20 patients and 21 controls.

is also a limiting factor. There are also methods that are applicable across languages but might be influenced by the training resources available for a given language. This is especially important for LM-based metrics, as LM performance is closely linked to the size of the training dataset (Kaplan et al., 2020). It is crucial to take into account these limitations when applying the NLP methods of psychosis detection cross-linguistically.

Recently, many studies applied the NLP-based methods of psychosis detection to a variety of languages, including Chinese[21], Danish[22], Dutch[23], German[24], Hebrew[25], Polish[26], Portuguese[27], Russian[28], and Spanish[29]. Despite that, the vast majority of metrics were developed and tested for English, and the results for other languages are often mixed, negative, or even contradictory to the findings for English (Kořánová, 2017; Bar et al., 2019; Panicheva et al., 2019; Doré, 2019; Parola et al., 2023).

I believe it is important to aim for cross-linguistically applicable NLP methods of psychosis detection, as well as to thoroughly test the cross-linguistic applicability of existent methods, keeping in mind the robustness with respect to resource availability and training corpora size for various languages.

I believe it is also important to aim for metrics that would be interpretable, internally valid, and transferable to settings other than the original. That is, the metrics should approximate a meaningful concept, should correlate with human judgements of the approximated concept, and should be independent of or corrected for known linguistic confounding factors. To achieve this goal, the proposed metrics must be validated with respect to psychiatric scales, theoretical assumptions, known confounding factors, and transferability potential.

[21] Parola et al., 2023
[22] Parola et al., 2023
[23] Doré, 2019; Voppel et al., 2021; Wouts et al., 2021; Corona-Hernández et al., 2023; Voppel et al., 2023; Boer et al., 2023
[24] Kořánová, 2017; Just et al., 2019; Just et al., 2020; Parola et al., 2023; Schneider et al., 2023
[25] Bar et al., 2019; Ziv et al., 2022; Shriki et al., 2022
[26] Sarzynska-Wawer et al., 2021
[27] Mota et al., 2014; Mota et al., 2017; Mota et al., 2023; Argolo et al., 2023
[28] Panicheva et al., 2019; Panicheva et al., 2020; Ryazanskaya et al., 2020
[29] Palominos et al., 2023

1.4 Motivation

The goal of the present work is to compare some of the most frequently used methods across the groups described above on two psychotic speech samples in German and Russian languages, prioritizing the metrics that can potentially be applied cross-linguistically and cross-methodologically. Moreover, this work aims at theoretical validation of some of these methods, assessing how the length of a sentence or utterance affects the metrics, as some of the metrics were reported to be strongly dependent on this confounding factor.

The present study aims to accomplish the following research goals. First, to establish the relative performance of the most commonly used metrics in NLP-based psychosis detection, as well as the relative performance of metric groups, delineating patterns of correlation with negative and positive symptoms. Second, to check whether the best metrics perform consistently across languages, elicitation tasks, and other methodological choices, such as the choice of embedding model. Third, to investigate which of the commonly used metrics are associated with mean sentence length, checking also whether the previously suggested patterns of length dependence in LM-based metrics hold on the present samples. Finally, to check whether the commonly used NLP-based metrics of psychosis detection can outperform the simplest baseline metrics, such as word count, sentence count, and mean sentence length.

Taken together, this may shed some light on the relative robustness and methodological validity of the tested metrics.

Chapter 2

Literature Review

In this section, I provide an overview of a number of papers dedicated to psychosis detection. The section is structured around the linguistic types of metrics outlined in the introduction, and papers may appear several times if they use a variety of metrics of different types. The papers were searched for with keywords relating to psychosis[1] and NLP[2]. The citations of the retrieved articles were used to expand the search. The articles concerning purely manual linguistic metrics or non-psychotic conditions were excluded from analysis, yielding a total of 61 papers and conference materials. The information about the patient populations of the papers reviewed can be found in appendix A.

Herein, I pay most attention to the metrics used in multiple papers. The sections cover each of the groups outlined in the introduction: lexical (2.1), syntactic (2.2), and semantic methods, including graph-based (2.3) and LM-based ones (2.4). I also provide a cross-group comparison of the methods (2.5) and a short summary of the trends observed (2.6). The results are summarized in table format in appendix B.

2.1 Lexical Methods

Lexical metrics focus on word use, including verbosity, lexical diversity, and choice of words. Most studies that report the difference in word count report

[1] "Psychosis", "schizophrenia", and "thought disorder".
[2] "Automated" + "language", "embedding", "model", "NLP".

lower *verbosity* in the patient group as compared to control[3] or association with negative symptom severity (Morgan et al., 2021; Minor et al., 2023), though some studies find no difference in verbosity[4].

Of the various measures of *lexical diversity*, the most frequently used one is the type-token ratio (*TTR*), defined as the number of unique words divided by the total number of words. While many studies use this metric or its variations, such as moving average TTR (MATTR)[5], a large portion of them only uses it as a feature for a classifier or latent analysis. In cases, where TTR is analyzed on its own, it is often reported to be lower in the patient group (Willits et al., 2018; Aich et al., 2022; Minor et al., 2023), though some report no significant differences (Hitczenko et al., 2020; Jeong et al., 2023; Schneider et al., 2023), or higher *type* and *lemma-token ratio* (Ziv et al., 2022). The *number of unique words* is also reported to be lower in the patient population by some (Willits et al., 2018) but not others (Schneider et al., 2023). Additional lexical diversity measures such as *Honoré statistics* are reported to be correlated with symptom severity (Jeong et al., 2023).

Another major group of lexical metrics is *sentiment analysis* metrics. Many researchers rely on pre-defined emotion word dictionaries provided by such tools as LIWC[6], Senti-WordNet, or NRC Word-Emotion Lexicon (Aich et al., 2022). Mitchell et al. (2015) report higher negative emotion words and lower positive emotion words in the schizophrenia group and Aich et al. (2022) report lower trust words use. Vail et al. (2018) and Girard et al. (2022) report negative emotion words correlating with the negative symptom severity, and Minor et al. (2023) report lower positive emotion word use and higher

[3]Mota et al., 2014; Iter et al., 2018; Willits et al., 2018; Doré, 2019; Just et al., 2019; Just et al., 2020; Panicheva et al., 2020; Morgan et al., 2021; Spencer et al., 2021; Voppel et al., 2021; Liang et al., 2022; Parola et al., 2023

[4]Mota et al., 2012; Gupta et al., 2018; Tang et al., 2021; Alonso-Sánchez et al., 2023; Alonso-Sánchez et al., 2022; Schneider et al., 2023; Nettekoven et al., 2023

[5]Rosenstein et al., 2015; Willits et al., 2018; Kramov, 2020; Hitczenko et al., 2020; Liang et al., 2022; Aich et al., 2022; Ziv et al., 2022; Jeong et al., 2023; Minor et al., 2023; Schneider et al., 2023; Tang et al., 2023b.

[6]Mitchell et al., 2015; Vail et al., 2018; Girard et al., 2022; Mota et al., 2023; Minor et al., 2023

negative emotion word use being associated with symptom severity. Alternatively, pre-trained sentiment analysis models can be utilized (Tang et al., 2023a; Tang et al., 2023b) to produce classifier features.

Additionally, different ways of quantifying *atypical word use* are reported to differentiate between controls and patients, including the use of neologisms (Just et al., 2019; Just et al., 2020) and out-of-vocabulary words (Just et al., 2020), and the use of foreign words is reported to be correlated with symptom severity (Jeong et al., 2023).

Finally, many researchers rely on such tools as LWIC, TAACO, or Coh-Metrix to provide them with a wide range of lexical and syntactic features, including various repetition and overlap patterns, cohesion or connective word categories, POS tags, and topic analysis[7]. Some researchers also utilize latent content analysis to explore prevalent topics, which is especially useful on heterogeneous texts (Mitchell et al., 2015; Rezaii et al., 2019). While these tools are popular and often prove useful for psychosis prediction, different researchers focus on different output features and topics, some also only using the features as input to classifiers or latent models, rendering direct comparisons uninformative.

The results for lexical methods are summarized in table B.1 in Appendix. In the present work, I assess verbosity and lexical diversity.

2.2 Syntactic Methods

Syntactic metrics analyze the use of syntactic types and structures. Analyzing the variation in the distribution of *part-of-speech* (POS) is the most popular syntactic method, as it is quite easy to perform and quantify. The most consistently reported difference is the lower use of determiners and wh-words in the patient group (Bedi et al., 2015; Corcoran et al., 2018; Sarzynska-Wawer et al., 2021; Tang et al., 2021), though with some reporting no group difference

[7] Mitchell et al., 2015; Gupta et al., 2018; Vail et al., 2018; Willits et al., 2018; Just et al., 2020; Mota et al., 2023; Girard et al., 2022; Liang et al., 2022; Minor et al., 2023

on a CHR population (Bilgrami et al., 2022) or no correlation with symptoms (Corcoran et al., 2018; Bilgrami et al., 2022). Mitchell et al. (2015) report higher article use rather than lower. Some studies report lower adjective and adverb use (Corcoran et al., 2018; Tang et al., 2021; Ziv et al., 2022) and reduced typicality in the use of adjectives and adverbs (Bar et al., 2019), while others report higher adjective use in at-risk populations (Argolo et al., 2023). The patients were also reported to show lower verb and past tense use (Ziv et al., 2022) though some report the opposite results (Mitchell et al., 2015) or no group differences in verb use (Tang et al., 2021; Argolo et al., 2023). Additionally, higher subordinating conjunction use (Silva et al., 2023), lower possessive pronoun use (Corcoran et al., 2018) and higher 1st person word use (Ziv et al., 2022) or negative correlation of pronoun use with symptom severity (Jeong et al., 2023) were reported for patient populations. Finally, several studies use POS tags as features in predictive models or latent analysis (Bedi et al., 2015; Sarzynska-Wawer et al., 2021; Tang et al., 2023a; Tang et al., 2023b). The only study to analyze POS tags and report no differences focused on a clinical high-risk population, rather than patients exhibiting symptoms (Haas et al., 2020). It is important to remark that two different major tagging schemes were used in the studies, namely, universal POS tagging[8] and Penn Treebank POS tagging[9] schemes, and some languages lack some of the POS categories commonly used for English, rendering some of the results incomparable.

Secondly, several studies assess the use of ambiguous pronouns and cataphora (Iter et al., 2018; Morgan et al., 2021; Nettekoven et al., 2023), which is either detected automatically, using a coreference tool, such as e2e_coref (Lee et al., 2017), or manually (Just et al., 2020). While some report higher referential failures in the patient population (Iter et al., 2018; Just et al., 2020), others report no difference in the referential failure rates (Morgan et al., 2021).

Several works propose various measures of syntactic complexity. Some measures are based on the use of different syntactic roles, reporting lower use of passive nominal subjects, clausal negation, and prepositional complements,

[8]https://universaldependencies.org/u/pos/
[9]https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

but higher use of simple nominal subjects in the patient population (Silva et al., 2023). Others measure syntactic complexity based on clause types, reporting lower use of coordinated clauses, lower index of general syntactic complexity, and higher rates of simple sentences (Schneider et al., 2023) as well as correlating lower syntactic complexity with symptom severity (Jeong et al., 2023). Some studies also rely on freely available tools, such as TAASSC[10], to assess syntactic complexity (Liang et al., 2022; Silva et al., 2023).

Finally, some rely on sentence, clause, or phrase length and count as a metric. Generally, decreased sentence length is reported in the patient population[11], though some report no differences between patients with FEP (Liang et al., 2022) or CHR (Gupta et al., 2018; Haas et al., 2020) and controls. Additionally, some report lower sentence length correlating with positive (Liebenthal et al., 2023) and negative (Bilgrami et al., 2022) symptoms, as well as social functioning (Silva et al., 2023). Maximal sentence length was reported to correlate with poverty of speech as assessed by TALD (Xu et al., 2020). Some also report lower clause length in patient population (Silva et al., 2023) or in the patient sub-population with more severely affected brain (Liang et al., 2022) with no difference between the patient group and control. Several studies successfully use utterance length (Tang et al., 2023b) or maximal phrase length (Bedi et al., 2015) as a feature in latent analysis or classifiers. Interestingly, the number of sentences is rarely used, and the results are contradictory, with some reporting lower count (Iter et al., 2018) or correlation with symptoms (Jeong et al., 2023) while others find no group differences (Gupta et al., 2018; Tang et al., 2021; Schneider et al., 2023) or higher sentence counts (Morgan et al., 2021; Nettekoven et al., 2023). This might indicate that lower verbosity in these studies is a result of shorter simpler sentences, rather than lower sentence counts.

The results for syntactic methods are summarized in table B.2 in Appendix.

[10] Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC), available at https://www.linguisticanalysistools.org/taassc.html, (Kyle, 2016)

[11] Iter et al., 2018; Morgan et al., 2021; Spencer et al., 2021; Tang et al., 2021; Bilgrami et al., 2022; Silva et al., 2023; Nettekoven et al., 2023; Schneider et al., 2023

Among syntactic methods, I assess, firstly, parts of speech rates paying particular attention to determiners, adjectives, and pronouns. Secondly, I assess the length of sentences, as it is both one of the frequently used metrics and a metric that may influence other metrics, and, finally, I take into account the number of sentences.

2.3 Graph-Based Methods

Graph-based metrics represent the texts with a graph and then calculate the properties of the resulting graph to use as predictive features. There are two main approaches to representing text with graphs being used in the field of psychosis detection. The first approach is based on co-occurrence, where the words are used as graph nodes and they are connected with an edge if they follow one another in a sentence. Many studies rely on software designed specifically for the construction of such graphs, called SpeechGraphs (Mota et al., 2012; Mota et al., 2014). Alternatively, the graph may be formed on the basis of semantic connections, with words used as the nodes and semantic relationships between them as the edges (Nikzad et al., 2022; Nettekoven et al., 2023), with some proposing that semantic graphs function better than co-occurrence graphs for psychosis detection (Nikzad et al., 2022).

The most popular graph characteristics of the graph used for psychosis detection include the numbers of nodes (N) and edges (E), as well as the number of nodes in the largest connected component (LCC)[12] and largest strongly connected component (LSC)[13]. Some studies also measure the organization by comparing the size of LCC and LSC to that of random graphs of the same size and calculating z-scores (LCC and LSC z-score, respectively).

The number of nodes was reported to be lower in the patient population by Nikzad et al. (2022) and Nettekoven et al. (2023), yet for the latter the effect disappeared after controlling for overall verbosity. Mota et al. (2023) used the number of nodes as a feature in a classifier, while Tang et al. (2023b) excluded

[12] A subgraph in which all nodes are connected by some path.
[13] A subgraph in which all nodes are connected by an edge.

it on the initial step. Mota et al. (2012) and subsequently Mota et al. (2014) reported no difference in the number of nodes and no significant correlation with symptom scales, and so did Nettekoven et al. (2023). Palominos et al. (2023) reported no difference in the number of nodes that were calculated not from words but from NPs.

Several studies report lower number of edges in the patient population[14], but some find no such differences after controlling for length (Mota et al., 2012; Nettekoven et al., 2023). Some also found smaller largest strongly connected components and largest connected components (Mota et al., 2014; Mota et al., 2016; Mota et al., 2017; Spencer et al., 2021; Morgan et al., 2021; Nikzad et al., 2022). Nettekoven et al. (2023) report lower numbers and smaller median size of connected components in patient populations. As for organization, the LSC and LCC z-scores were reported to be lower, indicating more random-like graphs in patient populations (Mota et al., 2016; Mota et al., 2017; Spencer et al., 2021; Morgan et al., 2021; Nikzad et al., 2022) with LSC z-score being more reliable in differentiating between the groups.

Many report that lower values of these graph features are associated with negative PANSS (Mota et al., 2014; Mota et al., 2016; Nikzad et al., 2022) and negative linguistic symptoms measured by TLC[15] (Morgan et al., 2021; Spencer et al., 2021; Nikzad et al., 2022), though some report no relation with PANSS (Mota et al., 2012; Argolo et al., 2023; Nettekoven et al., 2023).

Some studies also report lower average total degree (Mota et al., 2014) or average weighted degree (Nikzad et al., 2022). The founding papers (Mota et al., 2012; Mota et al., 2014), also report differences in the number of repeated and parallel edges, as well as the number of loops of lengths one, two, and three, though theses effects disappeared after controlling for verbosity[16].

Some studies also use graph *density*[17], which was reported by Nikzad et al.

[14]Mota et al. (2014), Mota et al. (2016), Mota et al. (2017), and Nikzad et al. (2022)

[15]Or TLI, Thought Language Index (Liddle et al., 2002).

[16]Other metrics include NP recurrence and distance between NPs (Palominos et al., 2023), graph centrality measures (Argolo et al., 2023), number of nodes in the largest clique (Tang et al., 2023a; Tang et al., 2023b), and number of connected and strongly connected components (Argolo et al., 2023; Nettekoven et al., 2023).

[17]2E/N(N - 1)

(2022) to be lower in patient populations, while Mota et al. (2012), Mota et al. (2014), and (Argolo et al., 2023) found no such differences. Graph *diameter*[18] was used in multiple studies but was reported to differ between groups only in one study (Mota et al., 2014); and the same is true of *average shortest path*[19] (Mota et al., 2012; Nikzad et al., 2022; Tang et al., 2023a; Argolo et al., 2023).

Several papers also use various graph features as input for classifier models to predict negative symptoms (Mota et al., 2023; Tang et al., 2023b) and social cognition (Tang et al., 2023a).

As many of these metrics, such as the total number of nodes and the number of nodes in the largest connected component, depend on verbosity, controlling for the effects of verbosity is very important, and many studies adjust the metrics for the number of words or calculate graphs over windows of around 100 words.

The results for graph methods are summarized in table B.3 in Appendix.

In the present work, I use co-occurrence-based graphs, as there is more consensus in the use of this method, though it may possibly be less efficient. I assess the number of nodes and edges, as well as the sizes of the largest connected and strongly connected components, loops of size one, two, and three, and parallel edges. I employ the common procedure of calculating graphs over windows of 100 words to somewhat lessen the effects of verbosity.

2.4 Language Model-Based Methods

There is a great variety of language model-based methods of psychosis detection involving different language models as well as many ways of applying these models to the task at hand. This part of the literature review is organized according to the types of LM-based metrics and covers the possible options for calculating them. The section is structured as follows: subsection 2.4.1 covers the methods that use embeddings and subsection 2.4.2 various other ways of using LMs. Subsection 2.4.3 is dedicated to the ways scores for

[18] The length of the longest shortest path between the node pairs of a network.
[19] The average length of the shortest path connecting each pair of nodes in the graph.

separate units are aggregated into a single participant or text score. Subsection 2.4.4 compares different LMs used in the field. Finally, subsection 2.4.5 discusses the ways in which the choice of metric and the choice of the model may interact in the task of psychosis detection.

2.4.1 Embedding-Based Methods

Most LM-based methods of psychosis detection operate over embeddings of words or sentences.

Word-Based Methods

While many apply word embedding-based metrics to structured tasks, such as verbal fluency task or single-word associations to assess the similarity between pairs of individual words[20], this technique is less popular for texts with fully-formed sentences. Nevertheless, several studies report lower cosine similarity between consecutive words (*word-based coherence*) in the patient population in several languages (Hebrew, Bar et al., 2019; German and Chinese Parola et al., 2023) or observe a correlation of the scores with human coherence judgement (Xu et al., 2020), though some find no group differences (Liebenthal et al., 2023; Argolo et al., 2023) or even report the opposite direction of difference (Danish, Parola et al., 2023). Alternatively, one may calculate the average similarity of words at a fixed distance (*k-inter-word similarity*), as suggested by Corcoran et al. (2018), who found lower k-inter-word similarities in the CHR group as compared to controls at distances of 5 to 7, even after controlling for sentence length, though with no correlation with symptom severity. Parola et al. (2023) show similar results for Chinese and German, but higher k-inter-word similarity in the Danish-speaking patient population as compared to control, and Argolo et al. (2023) find no group differences on a CHR population. Some studies also assess similarity between

[20] Elvevåg et al., 2007; Holmlund et al., 2019; Pietrowicz et al., 2019; Ryazanskaya, 2020; Ryazanskaya et al., 2020; "Computational linguistic analysis applied to a semantic fluency task: A replication among first-episode psychosis patients with and without derailment and tangentiality" 2021

all pairs of words (*all-word similarity*, Alonso-Sánchez et al. (2023), Alonso-Sánchez et al. (2022), and Liebenthal et al. (2023)), but none find group differences in this metric and only one study finds a correlation with symptom severity (Alonso-Sánchez et al., 2022). Two studies also assess *vector magnitude*, finding no differences between groups (Rezaii et al., 2019; Liebenthal et al., 2023). Similarity of each word vector to the *centroid* of all word vectors, or *cumulative centroid* of all words preceding a given word also were tested and proved moderately efficient in approximating human coherence judgements (Xu et al., 2020; Xu et al., 2022). One paper also successfully used cosine similarity between word windows including or excluding different groups of connective words (Corona-Hernández et al., 2023).

Another popular approach is the *moving-window coherence* also known as *sliding-window coherence*, where the metric is calculated as the average similarity between each word pair in a window of several words, rather than between sentences, and the window is moved by one word along the text. This method is particularly well adapted for shorter texts where the number of sentences might be insufficient to produce reliable coherence assessments (Panicheva et al., 2019). The studies typically assess various window sizes between 2 and 10, and the results vary greatly. While some report lower mean values in patient populations for windows of size 1 to 5 (on content words in Hebrew, Bar et al., 2019) and 5 and 10 (for Chinese, Parola et al., 2023), others report higher mean moving-window coherence values for windows of size 5 and 10 (Alonso-Sánchez et al., 2023; Alonso-Sánchez et al., 2022) or no difference at all (for window of size 8 on Dutch-speaking patient populations, Doré, 2019 and windows of size 5 and 10 on German and Danish-speaking patient populations Parola et al., 2023). Some also report higher variance of moving-window coherence values in Dutch-speaking patient populations (Voppel et al., 2021; Voppel et al., 2023) with window sizes from 5 to 10, as well as lower maximum and higher minimum values for Russian-speaking patient populations (Panicheva et al., 2019) with window size of 4.

The results for word-based LM methods are summarized in table B.4 in Appendix.

Sentence or Phrase-Based Methods

More commonly than on individual word vectors, papers focus on embeddings of sentences or phrases.

By far the most popular metric is so-called *first-order coherence* or simply *coherence*[21], which is calculated as the similarity between adjacent sentences or phrases. This metric was used in half the studies that relied on LM-based metrics with mixed results. It is important to note that while most studies report mean cosine similarity of all adjacent sentence pairs, some report minimal or maximal values, complicating direct comparisons.

The first studies to adopt this metric used it as a feature in predictive models (Bedi et al., 2015; Rosenstein et al., 2015), and this is frequently used in this was to this day (Sarzynska-Wawer et al., 2021; Tang et al., 2023a; Tang et al., 2023b). While some report lower coherence scores in the patient population (Just et al., 2019; Morgan et al., 2021), some of these studies only observe this effect with few embedding methods among many tested settings (Iter et al., 2018; Just et al., 2019; Ryazanskaya, 2020) or report that the results are significantly affected by sentence length to the extent that the results are insignificant after controlling for it (Just et al., 2019). Some studies report finding no group difference at all both for SDD (Just et al., 2020) and CHR populations (Hitczenko et al., 2020; Bilgrami et al., 2022; Haas et al., 2020). The metric was also reported to fail as a predictor of longitudinal outcomes (Just et al., 2023).

As for correlation with symptoms, while some studies find a correlation with PANSS, especially negative subscales (Ryazanskaya, 2020; Just et al., 2023), others find no such relation but report a correlation with SANS (Parola et al., 2023). Just et al. (2020) report lower coherence in patients with positive FTD as opposed to patients without it, though no significant group difference with controls was found. On CHR populations, no correlation with symptom severity was observed (Hitczenko et al., 2020).

[21] Also known as *local coherence* as opposed to *global coherence*.

Several validation studies report that the coherence metric correlates manual assessments of coherence (Xu et al., 2020; Xu et al., 2022) and maximal coherence scores were reported to correlate with TLC subscales of tangentiality, derailment, circumstantiality, loss of goal, poverty of speech, pressure of speech (Bilgrami et al., 2022).

Altogether, despite the popularity of this metric, there is still no consensus on whether it can serve as a reliable psychosis predictor, as the results vary across models (Iter et al., 2018; Just et al., 2019; Ryazanskaya, 2020) and languages (Just et al., 2020; Parola et al., 2023). The metric also seems to fail for CHR populations (Hitczenko et al., 2020; Bilgrami et al., 2022; Haas et al., 2020).

Additionally, Bedi et al. (2015) proposed *second-order coherence*, which compared sentences one sentence apart rather than consecutive sentences. The authors used this metric in a classifier model with very high accuracy (Bedi et al., 2015), and a similar approach was since applied to Polish (Sarzynska-Wawer et al., 2021). Parola et al. (2023) tested second-order coherence in Chinese, Danish, and German, and found it to be significantly lower in the patient populations. Morgan et al. (2021) use a somewhat similar approach to assess *repetitiveness* by calculating maximal similarity among all sentence pairs but find no significant difference in this metric between the groups.

Another family of approaches tries to approximate how well each sentence is related to the main topic, which can be seen as assessing *global* rather than *local* coherence. One approach is to assess how close the vector of an entire response is to the vectors of responses produced by other participants (*group global coherence*) or to a gold standard response (*gold standard global coherence*). These two approaches are especially useful for picture-elicited texts or story retellings, as the contents could reasonably be expected to be highly similar for most responses. The *group global coherence* was introduced in the pioneering work by Elvevåg et al., 2007, where the similarity to other patient responses was successfully used to predict human ratings of organizational structure, tangentiality, and content, and it was replicated in Elvevåg et al.

(2010). The method was also applied to Russian material, correlating with local coherence judgements (Ryazanskaya et al., 2020), differentiating between the groups, and correlating with PANSS scores (Ryazanskaya, 2020). The similarity to a gold standard description correlated with judgements of local coherence and violations of completeness on Russian material (Ryazanskaya et al., 2020) and was reported to be able to differentiate between the groups on English-speaking patient populations (Morgan et al., 2021)[22]. An alternative approach to global coherence was proposed by Xu et al. (2020), where each sentence is compared to the centroid of vectors of all sentences in one text (*centroid global coherence*) or to the centroid of all sentences preceding a sentence (*cumulative centroid global coherence*). Both metrics were shown to correlate with human coherence judgements (Xu et al., 2020; Xu et al., 2022) as well as to predict TALD scores (Xu et al., 2022). The metrics were also successfully applied to Russian, being able to differentiate between the groups and correlating with PANSS scores (Ryazanskaya, 2020); as well to German, where the centroid global coherence was shown to correlate with negative, disorganized, and exited PANSS subscales (Just et al., 2023), though with no ability to predict longitudinal outcomes.

Finally, several metrics were developed to assess *tangentiality*, meaning the relevance to the topic under discussion or the question asked. The most popular approach to tangentiality assessment was introduced in the pioneering work by Elvevåg et al. (2007). The researchers used the sliding window and calculated the similarity of each window to the question posed by the interviewer. Then, they calculated the slope to assess how quickly each participant moved away from the topic. They found a significant correlation between the slope of the similarity values and the blind human ratings of tangentiality, as well as greater variance of the values at higher window sizes and in patients with high FTD. While some studies report higher values of tangentiality (steeper slopes) in the patient populations (Iter et al., 2018; Tang et al., 2021), others find no significant group differences both in English (Hitczenko et al., 2020; Morgan et al., 2021) and in other languages (German, Kořánová, 2017; Just et al., 2019 and Dutch, Doré, 2019). The same

[22] Nettekoven et al., 2023 utilize this metric but do not report on its efficiency.

studies also report no correlation with symptom severity both in SDD (Doré, 2019) and in CHR populations (Hitczenko et al., 2020; Morgan et al., 2021). Kořánová (2017) and Just et al. (2019) also assess mean similarity to the question rather than the slope, yet also find no group differences.

Table 2.1 provides formulae for the metrics described above. The results for sentence-based LM methods are summarized in table B.5 in Appendix.

Metric Name	**Definition**	**Introduced**
(First-Order) Coherence	$agg_{i=1}^{N}(cossim(S_i, S_{i+1}))$	Bedi et al., 2015
Second-Order Coherence	$agg_{i=1}^{N-1}(cossim(S_i, S_{i+2}))$	Bedi et al., 2015
Repetitiveness	$max_{i,j=1,i\neq j}^{N}(cossim(S_i, S_j))$	Morgan et al., 2021
Moving Window Coherence	$agg_{i=1}^{M-k}(mean(cossim(W_x, W_y)))$ and $0 < y - x \leq k; i \leq x, y < M$	Bar et al., 2019 Panicheva et al. (2019)
K-Inter Word Similarity	$agg_{i=1}^{M-k}(cossim(W_i, W_{i+k}))$	Corcoran et al., 2018
Group Global Coherence	$cossim(R_a, mean_{b=1}^{Q}(R_b))$	Elvevåg et al., 2007
Gold Standard Global Coherence	$cossim(R_a, R_{gold})$	Ryazanskaya et al. (2020)
Centroid Global Coherence	$agg_{i=1}^{N}(cossim(S_i, mean_{j=1}^{N}(S_j)))$	Xu et al., 2020
Cumulative Centroid Global Coherence	$agg_{i=1}^{N}(cossim(S_i, mean_{j=1}^{i}(S_j)))$	Xu et al., 2020
Slope Tangentiality	$slope_{i=1}^{N}(cossim(S_q, S_i))$	Elvevåg et al., 2007[23]
Q-similarity Tangentiality	$agg_{i=1}^{N}(s_q, s_i)$	Kořánová, 2017

TABLE 2.1: Definitions of the embedding-based methods.
S - sentence embedding; W - word embedding; R - response embedding; $cossim$ - cosine similarity; agg - aggregation scheme (such as mean, max, min, or median); $slope$ - slope of a linear regression function; S_q - question embedding; R_{gold} - gold standard description embedding; M - number of words; N - number of sentences; Q - number of people in the group. Score aggregation is discussed in more detail below (2.4.3).

Among embedding-based methods, I assess coherence, as the most frequently used metric, as well as second-order coherence and both variants of centroid global coherence, as they are some of the more consistently well-functioning metrics. Unfortunately, the data at my disposal does not allow for testing tangentiality, as it is only well adapted for semi-structured interviews with longer responses.

2.4.2 Feature-Based Methods

An alternative group of methods does not rely on the cosine similarity of embedded text but rather uses language models to generate predictive features

or fine-tune the models for psychosis prediction.

An early approach was to use an embedding of the entire text as a feature in a classifier model with no fine-tuning of the language model (Elvevåg et al., 2010; Rosenstein et al., 2015), yielding relatively high classification accuracy of 0,72 - 0,83. Recently, BERT embeddings were successfully used to distinguish between SDD patients and healthy controls (Srivastava et al., 2022b). Now, the studies that have enough data for such a procedure, *fine-tune transformer-based language models* (Wouts et al., 2021; Aich et al., 2022; Shriki et al., 2022). While some report relatively high accuracy with fine-tuned BERT (0,84 in Hebrew, Shriki et al., 2022) or fine-tuned RoBERTa (0,76 in Dutch, Wouts et al., 2021), others report that fine-tuning of various transformer models performs very poorly compared to feature-based methods, reporting 0,6 accuracy for MentalRoBERTa; 0,38 for MentalBERT, and 0,33 for BERT base as compared to 0,70-0,96 achieved by a random forest classifier over simpler features on an English-speaking population (Aich et al., 2022). One of the challenges to the fine-tuning approach is the very small sample size of most studies (with a mode sample size of 20 patients and 21 controls). Aich et al. (2022) have the largest sample size, with 247 patients and 110 controls, and yet they report low accuracy for fine-tuned models. Some combat the problem of sample size by using small chunks of each text as fine-tuning material (Wouts et al., 2021), yet that still requires a significant sample size for sufficient fine-tuning, which is unattainable in many clinical settings, and the models trained on mined mental health-related texts, such as MentalBERT, seem to fail to transfer to the speech domain (Aich et al., 2022).

Another type of metric that has been proposed by Mitchell et al. (2015) is *perplexity* or *surprisal*. This metric is a model assessment of how likely a given text fragment is. Perplexity is typically used to assess the quality of the model fit, as the model should give low perplexity or surprisal scores on real texts. In other words, the model should not be very surprised by normal real-world texts from the same domain as the one it was trained on. Some hypothesized that as psychotic speech is atypical, it could receive higher perplexity or surprisal scores, while others suggested that stereotyped speech would result

in lower perplexity values. While Mitchell et al. (2015) apply perplexity to psychosis detection, they find no group differences in perplexity. In contrast, Srivastava et al. (2022a) report increased perplexity in SZ as compared to CHR and controls, and both Vail et al. (2018) and Girard et al. (2022) find an association between increased perplexity and higher PANSS scores in psychotic disorder populations[24]. On the other hand, Jeong et al. (2023) find no association of BERT surprisal scores with PANSS, though they report that it is positively correlated with pressure of speech, circumstantiality, illogicality, tangentiality and negatively with poverty of speech as measured by TLC, SANS, and SAPS.

Finally, some researchers suggested that one of the BERT training tasks, namely, *next sentence prediction* could be utilized for psychosis detection. Like perplexity, next sentence prediction assesses the likelihood of the text, but rather than the likelihood of one sentence or the entire text, it measures how likely the two sentences are to follow one another. The model is trained to differentiate between sentences that did occur sequentially from sentence pairs that were sampled randomly. For highly unpredictable speech such as that sometimes seen in psychosis, one could expect lower values of next sentence probability obtained from a BERT model. The metric was first used for psychosis detection by Hitczenko et al. (2020) with no significant differences found between healthy controls and the CHR population. Subsequently, it was applied to SZ patients, also yielding no group difference (Tang et al., 2021) and no correlation with PANSS, though next sentence probability was negatively associated with derailment, illogicality, and circumstantiality as measured by TLC, SANS, and SAPS (Jeong et al., 2023).

The results for feature-based LM methods are summarized in table B.6 in Appendix.

Among feature-based metrics, I assess next-sentence prediction and pseudo-perplexity. Unlike the models used in the research above, many pre-trained

[24]Colla et al., 2022 focus on the method itself and run an in-depth analysis of perplexity metric variants applied to Alzheimer Disease detection.

models only provide embedding, and not probability distributions over tokens. Therefore, true perplexity is not defined for them. However, it has been suggested for BERT (Wang et al., 2019), that pseudo log-likelihood form per-word masking can be used to mimic the calculation performed on traditional language models. Using this approach, one can calculate pseudo-perplexity and use it for similar purposes as regular perplexity. Fine-tuning, however, is not possible on the datasets at hand, as it requires more data points than I have at my disposal.

2.4.3 Score Aggregation

Many of the embedding-based methods produce not a single score for a text, but a collection of scores, one for each sentence, for a pair of sentences, or for each window. Multiple ways of aggregating these scores into a single score for the entire text can be utilized. While three-quarters of the examined studies used mean of the scores, many also tested other approaches, such as minimum[25] or maximum[26] of the scores.

Minimum values in such metrics as first-order coherence are used frequently, as they were reported to perform well for group differentiation in early work in the field (Bedi et al., 2015). The results, however, are mixed with the minimum being utilized for a variety of metrics, and some reporting higher minimum values in patient populations (Panicheva et al., 2019; Corona-Hernández et al., 2023) and negative association with clinical tangentiality assessment (Bilgrami et al., 2022), with others reporting lower minimum values in patient populations (Bedi et al., 2015; Corcoran et al., 2018; Iter et al., 2018; Ryazanskaya et al., 2020), and still others finding no group differences on CHR populations (Haas et al., 2020; Bilgrami et al., 2022). The minimum

[25] Bedi et al., 2015; Iter et al., 2018; Corcoran et al., 2018; Panicheva et al., 2019; Haas et al., 2020; Ryazanskaya, 2020; Xu et al., 2020; Morgan et al., 2021; Sarzynska-Wawer et al., 2021; Voppel et al., 2021; Bilgrami et al., 2022; Corona-Hernández et al., 2023; Xu et al., 2022; Voppel et al., 2023

[26] Corcoran et al., 2018; Panicheva et al., 2019; Haas et al., 2020; Ryazanskaya, 2020; Morgan et al., 2021; Bilgrami et al., 2022; Corona-Hernández et al., 2023; Voppel et al., 2023

scores were also reported as highly useful classifier features in some studies (Bedi et al., 2015; Xu et al., 2020; Xu et al., 2022; Sarzynska-Wawer et al., 2021).

While some report no difference in *maximum* of scores (Haas et al., 2020; Morgan et al., 2021; Voppel et al., 2021; Bilgrami et al., 2022), others report lower maximum cosine similarity-derived scores in patient populations (Corcoran et al., 2018; Panicheva et al., 2019; Ryazanskaya, 2020; Corona-Hernández et al., 2023) and positive association with clinical tangentiality assessment (Bilgrami et al., 2022).

Some works also utilize *standard deviation* or *variance* of the scores as a predictive feature[27]. While some report no difference in this aggregate for CHR populations (Haas et al., 2020; Hitczenko et al., 2020), others find higher variance values (Voppel et al., 2021; Voppel et al., 2023) and still others find lower values of variance in patient populations (Corcoran et al., 2018).

Other aggregation methods include median[28], 10 and 90 percentile values (Corcoran et al., 2018; Panicheva et al., 2019), range (Corona-Hernández et al., 2023), interquartile range (Parola et al., 2023), and sum (Jeong et al., 2023).

The use of different averaging approaches is listed in table B.8 in Appendix.

Parola et al. (2023) report that among many tested options, 'median and interquartile range more robustly present independent measures of mode and variance of the distribution, respectively'. And a simulation study, conducted by Fradkin et al. (2023), found that 'whereas previous studies attempted to capture more complex dynamics of speech disorganization by accounting for how semantic distances vary within a narrative, our simulations showed that these alternative metrics are, in most cases, less sensitive than simple averaging'.

[27] Bedi et al., 2015; Corcoran et al., 2018; Panicheva et al., 2019; Haas et al., 2020; Hitczenko et al., 2020; Sarzynska-Wawer et al., 2021; Voppel et al., 2021; Corona-Hernández et al., 2023; Voppel et al., 2023

[28] Bedi et al., 2015; Sarzynska-Wawer et al., 2021; Corona-Hernández et al., 2023; Parola et al., 2023

Overall, there is most evidence for mean and minimum values serving as a successful aggregation technique for finding group differences. Yet, to limit the number of comparisons, only the mean averaging is used in the present work.

2.4.4 Language Models

It is important to note that different studies use different language models and different ways of acquiring sentence embeddings from them which might have a significant influence on the study results.

Non-Contextualized Embeddings

The pioneering works in the area of psychosis detection, such as Elvevåg et al. (2007) and Elvevåg et al. (2010), used latent semantic analysis models (*LSA*) which work by applying singular value decomposition over word co-occurrence counts in a large collection of documents (Landauer et al., 1998). The resulting vectors can be used to represent the distributional semantics of the words, as the words that occurred in similar types of context are represented with similar vectors, and the semantic similarity between these vectors is measured with cosine distance between them (i.e. cosine similarity). The downside of LSA is that singular value decomposition is a very computationally expensive operation, especially for large corpora. Additionally, the resulting vector representations are not context-aware and cannot distinguish between different meanings of the same word or between homonyms. Nevertheless, LSA has been used in many psychosis detection works, either as the only model[29] or in comparison with other models[30]. While the early works report that LSA was useful as a feature in a classifier[31], the model has failed to differentiate healthy controls from CHR population (Hitczenko et al., 2020; Haas et al., 2020). Additionally, while LSA has been reported to outperform other non-contextual embeddings (word2vec) when both are

[29] Elvevåg et al., 2007; Elvevåg et al., 2010; Rosenstein et al., 2015; Bedi et al., 2015; Haas et al., 2020

[30] Iter et al., 2018; Xu et al., 2020; Hitczenko et al., 2020; Tang et al., 2021; Tang et al., 2023b

[31] Elvevåg et al., 2007; Elvevåg et al., 2010; Rosenstein et al., 2015; Bedi et al., 2015

trained on a small corpus (Xu et al., 2020), yet, given sufficient training data, word2vec outperformed LSA (Iter et al., 2018; Xu et al., 2020).

Word2vec is by far the most popular embedding model in the field of psychosis detection. Introduced by Mikolov et al. (2013), word2vec relies on a neural network architecture to create a representation of the words that can predict if the words are likely to occur in the same contexts in a text. The resulting vector representations of the words can be used to assess the semantics of text, also using cosine similarity, though, like LSA, these representations are not context-dependent. With modern computational capacities and large corpora, word2vec is quite easy to train, and ready-to-use word2vec representations trained on internet common crawl, as well as Wikipedia, are available[32] for 157 languages (Bojanowski et al., 2017)[33]. This makes word2vec a very attractive model choice for many researchers, and it was used in half of the articles that relied on LM-metrics, and for more than half of the articles that used word2vec, it was the only model they used.

Quite a few papers compared word2vec to newer models, such as GloVe[34], sent2vec[35], ELMo[36], and BERT[37]. While some report that the choice of the model does not affect the outcomes (Hitczenko et al., 2020; Fradkin et al., 2023), other find significant differences between the models, interaction between models, metrics, and tasks. Iter et al. (2018) and Just et al. (2023) report word2vec outperforming GloVe, but Just et al. (2019) observe the opposite pattern. ELMo and BERT were reported to outperform word2vec in symptom severity assessment (Ryazanskaya, 2020) but proved equally unsuccessful as word2vec in distinguishing CHR from controls (Hitczenko et al., 2020). Additionally, BERT and its variants were shown to outperform word2vec

[32] https://fasttext.cc/docs/en/crawl-vectors.html

[33] Even though, technically, fasttext is different from word2vec, it is only different in the tokenization procedure, not the training architecture, and most works do not differentiate between the two, stating that they use word2vec when using a fasttext model. Because of the sub-word tokenization employed in it, fasttext is better suited for texts that can have neologisms, as they can still be represented, rather than being always OOV.

[34] Iter et al., 2018; Just et al., 2019; Tang et al., 2023a; Tang et al., 2023b; Just et al., 2023

[35] Iter et al., 2018; Just et al., 2019; Hitczenko et al., 2020

[36] Ryazanskaya, 2020; Hitczenko et al., 2020; Sarzynska-Wawer et al., 2021

[37] Ryazanskaya, 2020; Hitczenko et al., 2020; Xu et al., 2022

in predicting human coherence judgement (Xu et al., 2022). Finally, though Iter et al. (2018) report that word2vec outperforms *sent2vec*[38], both Just et al. (2019) and Hitczenko et al. (2020) show that the two are equally inefficient for psychosis detection.

GloVe is conceptually similar to word2vec and is another non-contextualized word embedding model (Pennington et al., 2014). It is incorporated into such tools as SpaCy[39] and CoVec[40]. The papers that use exclusively GloVe embeddings report quite limited success, finding no group differences (Just et al., 2020; Alonso-Sánchez et al., 2023; Alonso-Sánchez et al., 2022), tough Alonso-Sánchez et al. (2022) report correlation with PANSS and Just et al. (2020) report lower coherence values in high FTD patients.

For perplexity assessments, several papers used *trigram models* which directly assess the co-occurrence likelihood of three-word sequences, rather than modelling words with vectors (Mitchell et al., 2015; Vail et al., 2018; Girard et al., 2022). This method was reported to correlate with symptom severity in one study (Vail et al., 2018), while others found no significant effect (Girard et al., 2022) or group difference in perplexity (Mitchell et al., 2015).

In the present study, the choice of non-contextualized embeddings is limited to the most popular models, i.e. word2vec and GloVe.

Sentence Embedding Aggregation

Another significant factor in word embedding models is the way the vectors representing individual words are combined to obtain a single sentence or phrase representation. By far the most popular technique is *averaging* all the word vectors in a sentence or window, yet it has been reported to be noisy and confounded (Fradkin et al., 2023) as well as to produce undesirable connection between cosine similarity metrics and sentence length (Hitczenko et al., 2020; Parola et al., 2023; Fradkin et al., 2023). The reason for this is that

[38] *sent2vec* (Moghadasi et al., 2020) was only used in a few papers in the field and only in comparison with other methods and is less popular than transformer-based architectures.
[39] https://spacy.io/
[40] https://www.covingtoninnovations.com/software/CoVec-manual.pdf

averaging static word vectors with no weights results in a more generic and meaningless representation for longer sentences and it fails to reflect that function words, such as articles, might contribute little to the meaning of a sentence while driving all sentence vectors closer to a meaningless average. Therefore, many researchers utilize such weighing schemes as *TF-IDF*, where words are weighted inversely proportional to their corpus frequency and directly proportional to the word frequency in the sentence. This method is used in quite a few papers, with several reporting successful application over LSA (Iter et al., 2018) or word2vec (Just et al., 2019; Xu et al., 2022), others using it in classifier models (Ryazanskaya et al., 2020; Tang et al., 2023b), though some report no group difference in metrics obtained TF-IDF weighted word vectors (Just et al., 2020; Hitczenko et al., 2020). Xu et al. (2020) actually find no weighting more efficient than TF-IDF in predicting human coherence judgements, but unlike other studies that use a weighted average, Xu et al. (2020) use a weighted sum of the vectors. Another method used in several studies is the smooth inverse frequency (*SIF*), introduced by Arora et al. (2017). It combines frequency weighting with removing the common meaningless component, producing better-performing sentence representations. Metrics over this representation were shown to successfully differentiate between clinical and control groups (Iter et al., 2018; Ryazanskaya, 2020; Morgan et al., 2021; Nettekoven et al., 2023), but could not help to differentiate healthy controls from CHR (Hitczenko et al., 2020) and were not correlated with symptom severity (Iter et al., 2018; Ryazanskaya, 2020; Hitczenko et al., 2020; Morgan et al., 2021). Unlike Iter et al. (2018), who find both TF-IDF and SIF somewhat efficient, Just et al. (2019) only finds group differences with TF-IDF, but not SIF or mean averaging.

All in all, one could cautiously advise against the use of an unweighted average, as it is more prone to the problems of meaningless similarity and length dependence. The use of different word embedding sentence averaging methods is listed in table B.9 in Appendix.

In the present paper, TF-IDF is compared to simple averaging to assess the patterns in length-dependence and the relative efficacy of the two sentence averaging schemes.

Contextualized Embeddings

Contextualized embeddings such as *ELMo* and *BERT* are capable of representing the words in context and differentiating between different senses of words dependent on the surrounding context. Additionally, these embeddings are capable of representing the entire sentences as a whole rather than individual words. ELMo uses bi-directional LSTM units to obtain contextualized word and sentence embeddings (Peters et al., 2017). This model was used in several studies with some success. Ryazanskaya (2020) report comparable performance of ELMo and BERT in detecting group differences and correlation with PANSS scores. Sarzynska-Wawer et al. (2021) use ELMo in a classifier model with a reasonable accuracy. Srivastava et al. (2022a) use biLSTM-based perplexity as a feature for predicting symptom severity. On the other hand, Hitczenko et al. (2020) report ELMo to be equally unable as all other models to differentiate between CHR and healthy controls.

BERT, introduced by Devlin et al. (2019), along with other transformer architectures, such as RoBERTa, is commonly used in more recent works dedicated to psychosis detection. Many different approaches to using BERT have been utilized over the years, as BERT can produce different features, including the second-to-last layer of hidden state output and CLS token, that can be used for representing the entire sentence, or one might use word embeddings and encode the sentence as the sum of the token embeddings that form the sentence (Xu et al., 2022). The results of using BERT embedding are mixed, some reporting successfully differentiating between the groups or scores correlating with symptoms (Ryazanskaya et al., 2020; Xu et al., 2022; Srivastava et al., 2022b) and similarly good results observed for SentenceBERT model (Xu et al., 2022), introduced by Reimers et al. (2019). Others report lack of group difference on CHR populations (Hitczenko et al., 2020; Bilgrami et al., 2022) or no correlation of symptom severity with next sentence prediction and surprisal metrics (Jeong et al., 2023). As reported above, fine-tuning BERT models, including base BERT, RoBERTa (Liu et al., 2019), MentalBERT, and Mental-RoBERTa (Ji et al., 2021) also shows mixed results with some reporting high (Wouts et al., 2021; Shriki et al., 2022) and some rather low classification accuracy (Aich et al., 2022). The same mixed findings also hold

for other methods that rely on BERT models, namely, the next-sentence prediction and surprisal methods discussed above, some reporting successful application while others find no significant group differences or symptom correlations. It is important to note that such large models might be more sensitive to the computational restrictions than smaller models (Kaplan et al., 2020), and different BERT models were reported to show different performance levels in psychosis detection (Aich et al., 2022). The upside of complex models is that it is possible to incorporate other types of information, such as audio or temporal information into the model (Xu et al., 2022; Wouts et al., 2021).

The use of different language models is listed in table B.7 in Appendix. To limit the number of comparisons, among contextualized embeddings, only BERT sentence embeddings are used in the present work. BERT, word2vec, and GloVe are the most frequently used models. BERT can be taken to represent contextualized embeddings, while w2v and GloVe represent non-contextualized embeddings in the present benchmarking.

2.4.5 Model-Metric Interactions

Taking into account the wide choice of both models and metrics that use them, it is important to add that most studies that compare both several models and several metrics that utilize them, find significant model-metric interaction. Iter et al. (2018) report that the coherence metric was different between the groups only with word2vec embeddings and SIF averaging, while for tangentiality both word2vec with SIF and LSA with TF-IDF produced classifying scores, and report GloVe and sent2vec not working for any metric. Both Just et al. (2019) and Hitczenko et al. (2020), on the other hand, observe that only coherence calculated with GloVe model and TF-IDF weighting could distinguish the groups, and even this was insignificant after correcting for the effect of length, with all other LM-metrics and models showing no effect at all. Just et al. (2023) report that both first-order coherence and centroid global coherence calculated with GloVe or word2vec with mean averaging could predict negative symptom severity and only word2vec model

correlated with other subscales. Ryazanskaya (2020) reported that on one of the elicitation tasks metrics could work better with word2vec with SIF or ELMo while on another task BERT showed better performance in distinguishing the groups, and the results were highly inconsistent across tasks, metrics, models, and averaging. Both Xu et al. (2020) and Xu et al. (2022) also report some metrics performing better with one model than another without a clear indication of one model being universally better. While Xu et al. (2022) report SentenceBERT cumulative centroid global coherence as the best metric for both TALD and correlation with symptom severity, they also show that the best metrics in approximating human coherence judgement depend on model selection BERT CLS token and Sentence BERT working for cumulative centroid global coherence, and BERT word vector summation along with Sentence BERT for centroid global coherence. Nevertheless, it could cautiously be suggested that contextualized embeddings, such as ELMO and BERT, should be favoured, as they become more and more available for a wider range of languages, as they could be less prone to length dependence stemming from word averaging (Fradkin et al., 2023).

Just et al. (2023) suggest that 'the choice of NLP model should not be arbitrary', showing that even models trained on the same material can still yield different results depending on the model architecture alone. And stating that 'the effect of different embedding models both intra- and cross-linguistically requires further investigation'.

2.5 Cross-Group Comparison of Methods

Having discussed various methods in detail, let us compare the groups of methods by aggregating the results of studies that include several methods of different types. It is important to note, that the results between these studies might not be directly comparable, as different studies may be using different metrics from the same group. Therefore, even if the studies report opposite trends it does not necessarily imply contradictory results.

The most widely used groups are language model-based methods and syntactic methods. Table 2.2 summarizes the results reported by studies comparing some LM-based to some syntactic methods. Some studies find both groups efficient in differentiating between the groups (Bar et al., 2019) or correlating with clinical scales (Argolo et al., 2023), and others find that neither group of methods can efficiently distinguish CHR from controls (Haas et al., 2020) or assess CHR symptom severity (Bedi et al., 2015; Corcoran et al., 2018). Some also find mixed results with only some metrics from both groups functioning well in identifying group differences (Tang et al., 2021) or correlating with clinical scales (Rezaii et al., 2019; Bilgrami et al., 2022). The majority of studies that compare these two groups of methods find syntactic methods more successful than LM-based methods both in differentiating between the groups[41] and in predicting various clinical scales (Iter et al., 2018; Liebenthal et al., 2023; Jeong et al., 2023), with very few exceptions (Rezaii et al., 2019).

LM	Syntactic	Group Difference	Clinical Scales
+	+	Bar et al., 2019	*Argolo et al., 2023*
?	+	Iter et al., 2018 *Corcoran et al., 2018*; *Morgan et al., 2021*	Jeong et al., 2023
?	?	Tang et al., 2021	*Rezaii et al., 2019*; *Bilgrami et al., 2022*
?	!	*Rezaii et al., 2019*	
!	+	Mitchell et al., 2015; Just et al., 2020; *Bilgrami et al., 2022; Argolo et al., 2023*	Iter et al., 2018; Liebenthal et al., 2023
!	!	*Haas et al., 2020*	*Bedi et al., 2015*; *Corcoran et al., 2018*

TABLE 2.2: Comparison between language model-based (LM) and syntactic methods (Synt).
"+" indicates significant group difference or correlation for most metrics tested within the group. "?" indicates mixed results with some metrics showing significant results but not others. "!" indicates an absence of significant differences in the metrics tested. The studies on clinical high-risk populations are shown in italics.

As shown in table 2.3, quite a few studies compare some LM-based to some lexical methods with similar results. While some find both groups efficient

[41] Mitchell et al., 2015; Iter et al., 2018; Corcoran et al., 2018; Just et al., 2020; Morgan et al., 2021; Bilgrami et al., 2022; Argolo et al., 2023

in predicting clinical scales (Vail et al., 2018), many report neither of these approaches being efficient for CHR populations (Hitczenko et al., 2020; Argolo et al., 2023). Similarly, many find lexical methods more efficient than LM-based ones both in differentiating between the groups (Mitchell et al., 2015; Just et al., 2019; Just et al., 2020; Aich et al., 2022) and in assessing symptom severity (Girard et al., 2022; Hitczenko et al., 2020), again with some exceptions (Voppel et al., 2023).

LM	Lexical	Group Difference	Clinical Scales
+	+		Vail et al., 2018
+	!	Voppel et al., 2023	
?	+	Just et al., 2019	*Rezaii et al., 2019*; Jeong et al., 2023
!	+	Mitchell et al., 2015; Just et al., 2020; Aich et al., 2022	Girard et al., 2022
!	!	*Hitczenko et al., 2020; Argolo et al., 2023*	*Hitczenko et al., 2020*

TABLE 2.3: Comparison between language model-based (LM) and lexical methods. "+" indicates significant group difference or correlation for most metrics tested within the group. "?" indicates mixed results with some metrics showing significant results but not others. "!" indicates an absence of significant differences in the metrics tested. The studies on clinical high-risk populations are shown in italics.

Table 2.4 compares lexical and syntactic methods with some studies successfully using both approaches (Mitchell et al., 2015; Just et al., 2020; Jeong et al., 2023) and some reporting no effect with either (Liang et al., 2022). Overall, neither approach can be deemed better, as some report lexical methods being more effective than syntactic (Gupta et al., 2018; Rezaii et al., 2019), while others observe the opposite trend (Schneider et al., 2023; Argolo et al., 2023).

Finally, Table 2.5 compares graph methods to syntactic and LM-based ones. Most studies report graph-based methods being equally successful in differentiating between groups as syntactic methods (Spencer et al., 2021; Morgan et al., 2021; Nettekoven et al., 2023), though Argolo et al. (2023) find better success with syntactic methods than graph-based ones in predicting symptom severity. As for LM-based methods, Morgan et al. (2021) report limited success as compared to graph-based methods for finding group differences,

Lexical	Syntactic	Group Difference	Clinical Scales
+	+	Mitchell et al., 2015; Just et al., 2020	Jeong et al., 2023
+	?		*Rezaii et al., 2019*
?	!	*Gupta et al., 2018*	
!	+	Schneider et al., 2023; *Argolo et al., 2023*	
!	!	Liang et al., 2022	

TABLE 2.4: Comparison between lexical and syntactic methods.
"+" indicates significant group difference or correlation for most metrics tested within the group. "?" indicates mixed results with some metrics showing significant results but not others. "!" indicates an absence of significant differences in the metrics tested. The studies on clinical high-risk populations are shown in italics.

while Argolo et al. (2023) report the opposite pattern for predicting symptom severity.

Graph	Synt	LM	Group Difference	Clinical Scales
+	+		*Spencer et al., 2021; Nettekoven et al., 2023*	
+	+	?	*Morgan et al., 2021*	
?	+	+		*Argolo et al., 2023*

TABLE 2.5: Comparison between graph-based (Graph), syntactic (Synt), and language model-based methods (LM).
"+" indicates significant group difference or correlation for most metrics tested within the group. "?" indicates mixed results with some metrics showing significant results but not others. "!" indicates an absence of significant differences in the metrics tested. The studies on clinical high-risk populations are shown in italics.

2.6 Summary

All in all, despite a great variety of metrics proposed, no single metric seems to have strong evidence for being a robust indicator of psychotic alterations in speech especially with respect to cross-linguistic and cross-model reliability.

Coherence is by far the most frequently used NLP metric, yet the results are mixed. The same can be said of the moving window coherence, slope tangentiality, and word-based coherence metrics. TTR and emotion words are the most popular metrics among lexical methods. Part-of-speech frequencies, sentence length and count are the most frequently used syntactic metrics.

Graph-based approaches utilize a similar set of metrics, which typically include the size of LSC and LCC, as well as their randomness, and the number of nodes and edges. Additionally, many studies conducted on spoken discourse report disfluency features such as perseverations, hesitation pauses, and false-starts. Many studies report lower word counts as well as shorter sentences in affected populations.

Overall, the results reported both for syntactic and lexical methods are more consistent than the ones for LM-based methods, which might be partially caused by the differences stemming from embedding model and preprocessing differences. Too few studies compare graph-based methods to the other groups, yet the results of the graph-based studies are quite consistent within the group. Across methods, the results are mixed and contradictory in clinical high-risk populations both for the task of predicting psychosis conversion and for the task of differentiating CHR from controls, and CHR speech metrics were repeatedly reported as more similar to HC than FEP (Morgan et al., 2021; Srivastava et al., 2022b; Nettekoven et al., 2023). It seems that negative symptoms are also more prevalent in the analyzed samples and can generally be approximated more robustly than positive symptoms. This could be because negative symptoms may occur before positive ones in the course of a psychotic disorder (Just et al., 2023).

The present work presents a cross-group cross-linguistic comparison of some of the most frequently used and most effective metrics.

Chapter 3

Methods

3.1 Data

In this chapter, I discuss the data used in the present study (3.1), as well as the metrics chosen from each of the metric groups (3.2) and the ways in which they were computed and analyzed (3.3)[1].

3.1.1 German

The German clinical sample consisted of Narrative of Emotions Task (Buck et al., 2014) interview recordings from 59 NAP patients[2] and 20 controls, characterised in table 3.1. There was no difference in gender balance, age, or years of education between the groups[3].

A short version of the narrative emotions task, translated into German, was used. The interview included 3 questions on 4 emotions: sadness, fear, anger, and happiness. The questions were as follows: (1) what does this emotion mean to you? (2) describe a situation where you felt this emotion, and (3) why do you think you felt this emotion in this situation? All interviews were

[1] The code used for pre-processing the data, calculating the metrics, and post-processing the results is available on GitHub (https://github.com/flying-bear/MA_thesis).

[2] 46 diagnosed with schizophrenia and 13 with schizoaffective disorder.

[3] The NAP sample partially coincides with the one described in Just et al. (2023) for the second time-point. The outpatient data was collected at the Charité hospital in Berlin. The transcription was performed by native German speakers. The data was provided by Dr. Sandra Just.

recorded and manually transcribed according to transcription guidelines for establishing the sentence boundaries[4]. The interviewer's speech, as well as filled hesitation pauses, were removed from the transcripts. The participant responses were concatenated for each emotion, and the emotions were processed separately and then the values were averaged across emotions.

	N	**female**	**age**	**edu_years**
NAP	59	24	39.5 (11.1)	14.6 (3.0)
HC	20	9	43.85 (13.3)	15.5 (2.8)

TABLE 3.1: Social statistics of the German clinical dataset. Standard deviation is provided in parenthesis for each mean value. "edu_years" indicates years of education.

For each patient, the SANS, SAPS, and PANSS scores were collected, along with a Verbal IQ score (Schmidt et al., 1992). The scores for each scale, as well as for PANSS subscales, are characterised in table 3.2. There was no difference in age, education years, or clinical scale scores between the sexes in the clinical sample. There was also no correlation with years of education or verbal IQ in any of the symptom severity scales. Total SAPS score correlated with PANSS positive subscale, and total SANS score with PANSS negative and general subscales, as well as total PANSS score. As may be expected, PANSS subscales were also intercorrelated, except for PANSS positive, which only correlated with the total score.

Figure 3.1 provides a comparison between the scales for the samples, showing that negative symptoms prevail in the German clinical sample.

3.1.2 Russian

The Russian clinical sample consisted of monologue speech samples, elicited with 4 tasks, from 31 NAP patients, 18 depressive patients[5], and 30 controls

[4] As the LM metrics of the present study operate over sentences, the sentence segmentation decisions can significantly affect the results. To ensure uniformity, the sentence boundaries were determined based on syntax. The sentence was defined as a main clause with all its dependent clauses. Unfinished clauses were delineated as sentences as well. Conjuncted main clauses were separated.

[5] The exact diagnosis counts are provided in the appendix B in table C.1.

	N	age	edu_years	Verbal IQ	SANS	SAPS	PANSS	PANSS_n	PANSS_p	PANSS_o
range					0-120	0-170	30-210	7-49	7-49	16-112
all	59	39.5 (11.1)	14.6 (3.0)	105.2 (15.7)	27.7 (20.4)	16.8 (16.7)	57.3 (16.2)	16.9 (6.0)	12.7 (5.5)	27.8 (7.5)
male	35	37.3 (10.1)	14.5 (3.1)	107.0 (12.7)	28.6 (21.2)	18.6 (16.7)	59.2 (17.0)	17.1 (6.6)	13.4 (5.6)	28.6 (7.6)
female	24	42.7 (11.9)	14.7(2.8)	102.4 (19.3)	26.4 (19.4)	14.2 (16.7)	54.5 (14.8)	16.4 (5.2)	11.6 (5.2)	26.5 (7.3)

TABLE 3.2: Clinical statistics of the psychiatric sample in the German clinical dataset. Standard deviation is provided in parenthesis for each mean value. The range is provided for the possible values of the psychiatric scales. "edu_years" indicates years of education.

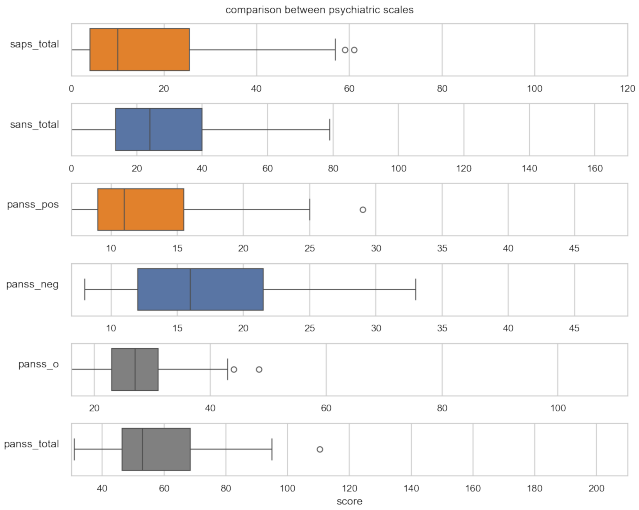

FIGURE 3.1: Clinical statistics of the psychiatric sample in the German clinical dataset. The range corresponds to possible values of the psychiatric scales. Negative symptom scales are shown in blue and positive symptom scales in orange, while grey indicates general symptom scales.

who also underwent a clinical interview process to exclude a possibility of undiagnosed disorders[6]. The sample is characterised in table 3.3. There were no differences in age between NAP patients and clinical controls but there was a significant difference in the years of education between the NAP and control groups (t=3.6, p=0.0006). There were no differences between the sexes in any of the samples. It is worth noting that the Russian sample is predominantly female, which is an understudied group when it comes to psychotic disorders, yet this imbalance is a limitation of the present study.

[6]The sample partially coincides with the one described in Ryazanskaya (2020). The data was collected jointly by Tatyana Shishkovskaya, Mariya Khudyakova, and me, and transcribed by me. The patient data was collected at the Mental Health Research Center, Moscow.

	N	**female**	**age**	**edu_years**
NAP	31	25	27.13 (7.14)	13.32 (2.41)
Dep	18	18	20.89 (3.71)	12.67 (1.94)
HC	30	26	25.0 (7.4)	15.43 (2.13)

TABLE 3.3: Social statistics of the Russian clinical dataset (only including the participants doing the selected tasks). Standard deviation is provided in parenthesis for each mean value. "edu_years" indicates years of education. "Dep" is the sample with predominant depressive symptoms. "HC" stands for the sample of the healthy participants.

The four tasks used to elicit monologue speech included two picture-elicited tasks ('sportsman' and 'adventure'), one instruction task ('chair') and one personal story task ('present') . The picture description tasks were elicited using two Bidstrup comics, provided in the appendix B, asking the participant to tell the story depicted in them. The instruction task was elicited using an IKEA chair brochure, asking the participant to instruct a third person, who cannot see the image. Finally, the personal story task was elicited by asking the participant to tell a story about the most memorable present that they had received. As not all participants were able to complete all the tasks, the number of samples available for each task is given in table 3.4. The speech recordings were manually transcribed and separated into sentences according to the same transcription guidelines as the ones used for the German sample. The interviewer's speech, as well as filled hesitation pauses, were removed from the transcripts.

	N	adventure	chair	present	sportsman
NAP	31	30	17	21	28
Dep	18	14	14	13	14
HC	30	25	16	19	26

TABLE 3.4: Task availability for each selected task in Russian clinical dataset. "Dep" is the sample with predominant depressive symptoms. "HC" stands for the sample of the healthy participants.

For each participant, PANSS scores, as well as clinical impressions of thought disorder and depression severity, ranging from 0 to 3, were collected. The scores for each subsample are characterised in table 3.5, and figure 3.2 shows

a comparison between the scales in the NAP subsample of the Russian clinical dataset, showing that negative symptoms prevail in this sample. After Bonferroni correction, two of the psychiatric correlated significantly with years of education: PANSS general (r=-0.41, p[7]=0.0008) and depression severity (r=-0.33, p=0.003). The psychiatric scores were also intercorrelated. Depression severity correlated with PANSS general (r=0.6, p<0.000001) and PANSS total score (r=0.35, p=0.005), as well as thought disorder severity (r=0.33, p=0.003). Thought disorder severity correlated significantly with all other psychiatric scales, most strongly with PANSS positive subscale (r=0.84), total PANSS score (r=0.81), general (r=0.71), and negative (r=0.69) subscales[8]. All PANSS subscales were significantly intercorrelated with r >0.65.

	sex	**N**	**age**	**edu_years**	**Dep**	**TD**	**P_N**	**PANSS_TD**	**PANSS**	**PANSS_n**	**PANSS_p**	**PANSS_o**
range								4-28	30-210	7-49	7-49	16-112
NAP	all	31	27.13 (7.14)	13.32 (2.41)	0.58 (0.85)	0.84 (0.73)	29	10.03 (3.74)	69.79 (16.13)	22.93 (8.59)	15.90 (4.92)	30.97 (8.42)
	f	25	27.80 (7.53)	13.56 (2.48)	0.72 (0.89)	0.8 (0.76)	23	9.43 (3.62)	69.13 (15.38)	22.52 (7.79)	15.3 (4.91)	31.3 (9.08)
	m	6	24.33 (4.72)	12.33 (1.97)	0.0 (0.0)	1.0 (0.63)	6	12.33 (3.56)	72.33 (20.16)	24.5 (11.93)	18.17 (4.67)	29.67 (5.65)
Dep	f	18	20.89 (3.71)	12.67 (1.94)	0.56 (0.62)	0.06 (0.24)	13	4.42 (0.9)	37.92 (5.89)	8.31 (1.97)	8.46 (1.94)	21.15 (3.58)
HC	all	30	25.0 (7.4)	15.43 (2.13)	0.0 (0.0)	0.0 (0.0)	22	4.36 (1.0)	30.77 (1.54)	7.23 (0.53)	7.23 (0.61)	16.32 (0.95)
	f	26	25.42 (7.8)	15.42 (1.7)	0.0 (0.0)	0.0 (0.0)	20	4.4 (1.05)	30.85 (1.6)	7.25 (0.55)	7.25 (0.64)	16.35 (0.99)
	m	4	22.25 (3.3)	15.5 (4.43)	0.0 (0.0)	0.0 (0.0)	2	4.0 (0.0)	30.0 (0.0)	7.0 (0.0)	7.0 (0.0)	16.0 (0.0)

TABLE 3.5: Clinical statistics of the psychiatric sample in the Russian clinical dataset (only including the participants doing the selected tasks). "HC" only refers to the subset of the healthy patients. Standard deviation is provided in parenthesis for each mean value. The range is provided for the possible values of the psychiatric scales. "f" stands for female; "m" for male. "edu_years" indicates years of education; "P_N" indicates the number of participants for whom PANSS scores are available, "PANSS_td" stands for the sum for PANSS questions related to formal thought disorder.

3.2 Data Processing: Metric Pool

Each text in the clinical sample was automatically separated into sentences based on punctuation, tokenised, and lemmatised. The selected metrics, described below, were then computed for each text separately.

[7] All p values in this passage are reported as p before correction.
[8] All p<0.000001.

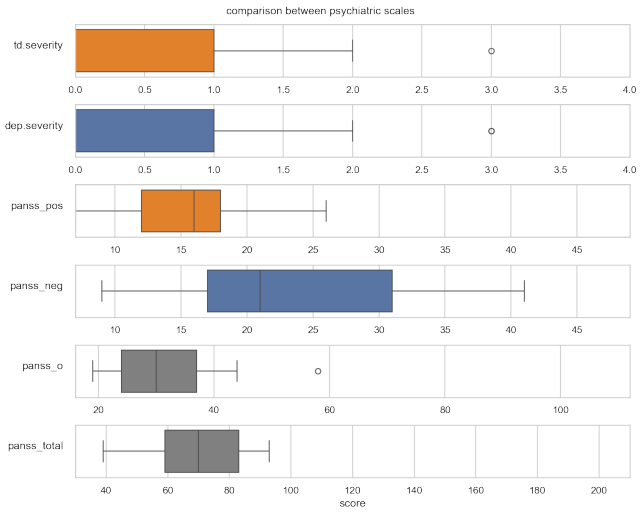

FIGURE 3.2: Clinical statistics of the psychiatric sample in the NAP subsample of the Russian clinical dataset. The range corresponds to possible values of the psychiatric scales. Negative symptom scales are shown in blue and positive symptom scales in orange, while grey indicates general symptom scales.

3.2.1 Lexical Methods

As there was little consistency in the use of semantic lexical metrics, only the Lemma-Token Ratio (LTR), Moving Average Lemma-Token Ratio (MALTR), and total word count (n words) were used. LTR was computed as the total number of unique lemmas over the total word count. MALTR was computed as LTR averaged across a moving window of size 10[9], with overlapping windows shifting one word at a time regardless of sentence boundaries. The lemmatization was performed using the SpaCy[10] `de_core_news_md` and `ru_core_news_md` models.

3.2.2 Syntactic Methods

The syntactic metrics included sentence parameters such as mean, maximal, and minimal sentence length, along with standard deviation in sentence length, and total sentence count (n sents). Additionally, part-of-speech (POS) rates were used for the POS tested previously: adjectives (ADJ), adverbs

[9]The window size was selected to be close to mean sentence length.
[10]Spacy version 3.7.2.

(ADV), auxiliary verbs (AUX), coordinating and subordinating conjunctions (CCONJ, SCONJ), determiners (DET), nouns and proper nouns (NOUN, PROPON), pronouns (PRON), particles (PART), and verbs (VERB). The POS tagging was performed using SpaCy models, and the rate was computed as the number of instances of a particular part-of-speech over the total word count.

3.2.3 Graph-Based Methods

The graphs were constructed based on word co-occurrence, as done in Mota et al. (2012). The graph construction was performed over lemmas and was performed over a moving window of 100 words[11]. A pair of lemmas was connected with a directed edge every time they appeared one after the other. The graph metrics included the number of nodes (N)[12], number of edges (E), largest connected component (LCC), largest strongly connected component (LSC), number of parallel edges (PE), number of loops of length one (repeated words), two, and three (L1, L2, L3), average node degree, and standard deviation in the node degree. These characteristics were computed for each window graph and then averaged across the overlapping windows. The characteristics of the graphs were computed using the `networkx` library.

3.2.4 Language Model-Based Methods

For LM-based methods, three models were compared. First, word2vec, which was loaded via SpaCy interface (obtained from fasttext[13] for `de` and `ru`)[14]. Second, built-in SpaCy GloVe models (`de_core_news_md` and `ru_core_news_md`). Finally, BERT sentence embeddings (`bert-base-german-cased` for German

[11] The moving average approach to graph methods is commonly accepted, and though the window size is not always reported, the 100 words is a common window size. Only few texts were below 100 words in total.

[12] The number of nodes corresponds to the unique lemma count calculated over a moving window of size of 100.

[13] https://fasttext.cc/docs/en/crawl-vectors.html

[14] Fasttext models at fasttext website are all pretrained on common crawl and Wikipedia dumps.

and `DeepPavlov/rubert-base-cased` for Russian). The two simplest methods of obtaining sentence embeddings from word embeddings were compared for word2vec: simple averaging word vectors (`w2v_avg` and `glove_avg`) and TF-IDF weighted averaging of word vectors (`w2v_tf` and `glove_tf`). The word frequencies for TF-IDF were obtained from the `wordfreq` library[15]. The TF-IDF weighting for word averaging implied giving each word vector a weight of one over the looked-up corpus word frequency, and the 'term frequency' was accounted for by the repetition of the word vectors where they were repeated in the sentence. The 'CLS' token embedding was used as the BERT sentence vector representation.

Each model sentence vectors were used to compute several coherence scores, namely, local (first order) coherence (`lcoh`), second order coherence (`scoh`), global coherence (`gcoh`) and cumulative global coherence (`cgcoh`). Local coherence was computed as the mean cosine similarity between the pairs of consecutive sentence vectors. Second-order coherence was computed as the mean cosine similarity between the pairs of sentence vectors with one intervening sentence between them. Global coherence was computed as the mean cosine similarity between each sentence vector and the mean vector of all sentences. Finally, cumulative global coherence was computed as the mean cosine similarity between each sentence vector and the mean vector of all preceding sentences.

A separate representation of the same BERT model was used to estimate pseudo-perplexity (`pppl`). It was obtained by masking one word at a time and obtaining the pseudo log-likelihood of the word that was masked for each word in a sentence. The pseudo log-likelihoods were then averaged across the sequence and exponentiated to obtain a pseudo-perplexity score[16]. The metric was averaged across all sentences.

Finally, a separate BERT representation was used to obtain next sentence

[15] wordfreq version 3.1.1

[16] The calculation was simplified by passing the true input token ids as labels to the model, and the loss was simply exponentiated, thus calculating the average masking pseudo log-likelihood for each token in the sentence in one line.

probability scores for each pair of consecutive sentences (`sprob`), and the metric was averaged across all sentence pairs.

3.3 Data Analysis

3.3.1 Control Variables

Both samples were analyzed using a two-sided independent t-test for between-group differences in the psycho-social characteristics, as well as for any effects of sex within each group. The psychiatric scales were analyzed for the degree of inter-correlation, as well as for age, education years, and IQ (using Pearson's r). To rule out the effect of age, education years, and IQ, the correlation of the tested metrics with them was computed. Additionally, the correlation of each psycho-social variable with mean sentence length was analyzed to account for results that could stem from the differences in mean sentence length.

3.3.2 Target Variables

Bootstrap was used to evaluate the uncertainty of each metric performance estimate and to account for the effect of the influential points. The data points were drawn with replacements from the sample until the actual sample size was reached; then, the performance was estimated on this sample. For each metric, median value and 25/75 percentiles were used to estimate the metric performance across 1000 iterations.

The performance of each metric was assessed using:

- two-sided independent t-test for between-group differences[17].
- Pearson's r correlation with each psychiatric scale[18];
- ordered categorical regression McFadden's pseudo r squared for TD and depression severity variables[19];

[17] The t-test was computed using `scipy.stats` (version 1.9.1).
[18] The Pearson's r was also computed using `scipy.stats`.
[19] The ordered model provided in `statsmodels` (version 0.13.5) was used.

- Pearson's r correlation with mean sentence length to control for effects explained merely by differences in sentence length.

The p-values were not analyzed for the bootstrap, as only the point of the present work is to assess the relative rather than the absolute performance of the metrics[20]. A metric was considered well-performing if it correlated above 0.3 with any psychiatric scale or had pseudo r squared above 0.09. For the t-test, the metric was well-performing if the 25-75 percentile interval did not cross zero. The metric was considered to be length-dependent if it correlated with mean sentence length above 0.3, yet metrics outperforming the mean sentence length baseline were still considered of interest and analyzed. All were compared for their performance on different psychiatric scales, and for the Russian clinical sample, the metrics were additionally compared for their performance on different tasks. Then the results were compared across languages.

[20]Due to multiple comparisons, very few if any metrics would remain significant in this benchmark study.

Chapter 4

Results

In this chapter, I first discuss the textual characteristics of both samples and verbosity patterns (4.1), then turn to the interaction between the social variables and the target variables, as well as the relation between the target variables and the text length (4.2). Afterwards, each group of metrics is discussed separately for both languages (4.4-4.7). Finally, I present a cross-group (4.8) and a cross-linguistic (4.8.3) comparison of the studied metrics.

4.1 Textual Characteristics of the Samples

As NAP patients are believed to be less verbose, it is important to take the differences in text length into account. Table 4.1 summarizes the length characteristics of both German and Russian samples. Unlike what has been reported previously, the overall text length, i.e. word count, seems to be explained more by the number of sentences than by their length, though sentence length also plays some role, especially in the German sample. It also seems to play some role in the overall verbosity on two of the tasks in the Russian sample, adventure and chair, - the shortest and the longest one, respectively.

Figures 4.1 and 4.2 show the relation between the word count and both sentence count and mean sentence length, as well as the correlation coefficient, underlining the differences in the components contributing to the text length

across the tasks and languages. These differences imply that the mean sentence length would serve as a strong baseline for performance only on some tasks, while on others only the number of sentences, and, possibly, the word count, would function as such. For the tasks, where the differences in mean sentence length are prominent, it is especially important to account for the intrinsic relation between some of the metrics and mean sentence length. However, no such correction is required for the number of sentences, as there seems no intrinsic relation between the number of sentences and the metric values for any of the metrics used, except the verbosity[1].

task	**German**	**Russian**			
		adventure	**chair**	**present**	**sportsman**
n words	184.2 (117.4)	129.8 (82.1)	168.4 (120.9)	140.8 (107.7)	134.4 (86.0)
n sents	17.9 (9.3)	18.7 (11.0)	20.4 (15.0)	15.1 (12.7)	17.7 (10.1)
mean sent len	10.0 (2.8)	6.9 (1.5)	8.1 (1.6)	9.9 (4.7)	7.8 (2.2)
r n sents	0.88	0.93	0.95	0.95	0.93
r n mean sent len	0.59	0.36	0.38	0.08	0.11

TABLE 4.1: Statistics of the mean length of each task between the languages. Standard deviation is provided in parenthesis for each mean value.
'n words' stands for the word count, 'n sents' for sentence count, and 'mean sent len' for mean sentence length. 'r n sents' indicates the correlation coefficient of the word count with the sentence count and 'r mean sent length' that of the word count with mean sentence length.

[1]It is possible that the moving window procedure does not entirely remove the effects of verbosity from the graph-based metrics. These effects may come via the number of sentences, rather than sentence length, though with the latter the graph-based metrics are indeed not intrinsically related. Yet, we do not explore this question further in the present work.

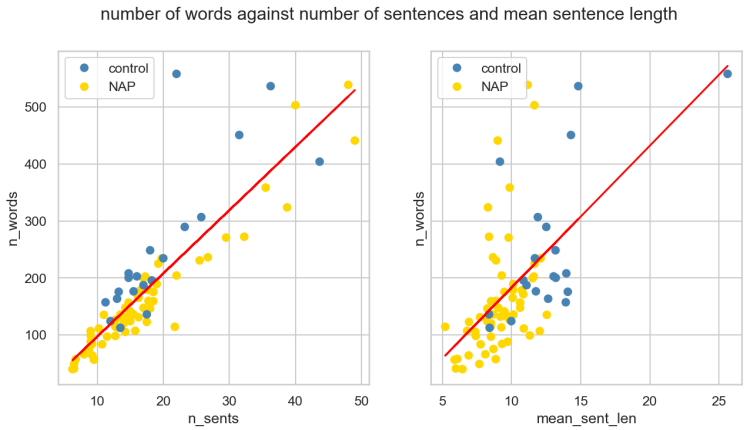

FIGURE 4.1: The correlation between word count and both sentence count and mean sentence length on the German sample, with colors indicating controls and NAP groups.

4.2 Control Variables

4.2.1 German

After correcting for multiple testing, there remained a significant negative correlation of symptom severity with mean sentence length for negative symptom scales: SANS and PANSS negative (r <-0.4, p before correction <0.001). There was no relation between sentence length and sex, age, years of education, or verbal IQ.

As for the relation between the control variables and the tested metrics, there was no difference in metric between the sexes, and no significant correlation with age, years of education, or verbal IQ, after Bonferroni correction.

4.2.2 Russian

After correcting for multiple testing, there was no significant correlation with mean sentence length for any of the psychiatric scales or social variables, though this may be both because of a high number of comparisons and a lower importance of mean sentence length for overall word count in the Russian sample.

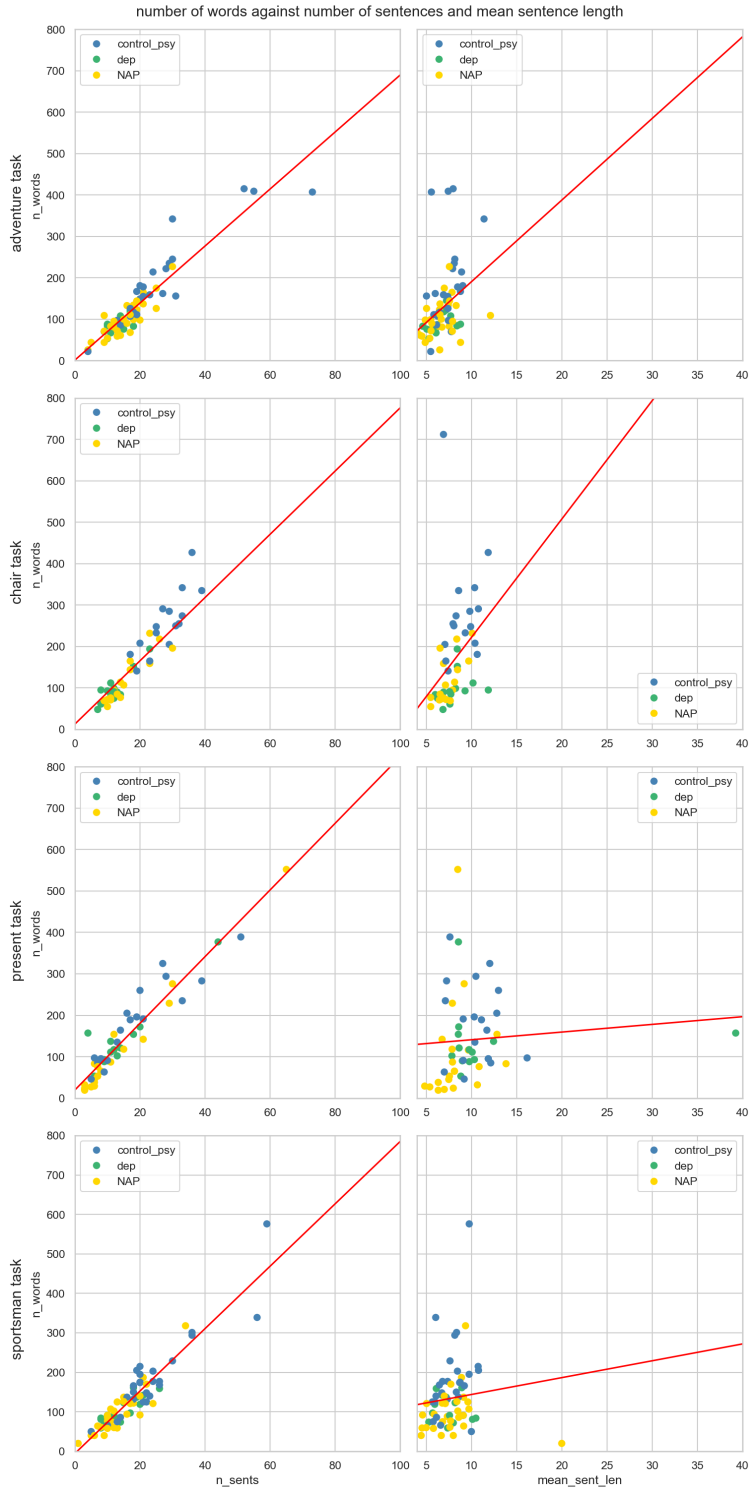

FIGURE 4.2: The correlation between word count and both sentence count and mean sentence length on the Russian sample for each of the tasks, with colors indicating controls, depression and NAP groups.

There also was no correlation between the metrics and the control variables, after controlling for multiple comparisons, with no significant correlation with age or ears of education, and no effect of sex, though here as well it may be due to a high number of comparisons.

4.3 Scale and Task Effects

Table 4.2 summarizes the performance of the tested metrics across tasks and scales, showing the number of metrics correlated with each psychiatric scale, as well as the number of metrics among them which were also uncorrelated with mean sentence length.

	German	**Russian**				
	total (De)	**adventure**	**chair**	**present**	**sportsman**	**total (Ru)**
PANSS total	16 (3)	9 (2)	12 (5)	18 (13)	8 (7)	30 (17)
PANSS_pos	1 (0)	1 (1)	12 (5)	13 (11)	10 (8)	24 (15)
PANSS_neg	12 (3)	8 (1)	5 (3)	17 (12)	8 (7)	28 (15)
PANSS_o	14 (2)	7 (2)	11 (4)	18 (13)	7 (6)	28 (16)
SANS	10 (3)	-	-	-	-	-
SAPS	5 (1)	-	-	-	-	-
dep severity	-	0 (0)	9 (2)	0 (0)	0 (0)	9 (2)
td severity	-	0 (0)	3 (2)	13 (11)	4 (3)	19 (15)
total	21 (4)	10 (2)	17 (5)	23 (18)	10 (8)	30 (20)

TABLE 4.2: The number of metrics correlating above 0.3 with the target scale (or having pseudo r squared above 0.9) and either do not correlate with mean sentence length above the threshold of 0.3 or outperform this baseline metric. In parenthesis are the numbers of metrics not correlating with sentence length above 0.3.

4.3.1 German

The overall differences between the scales are summarized in table 4.2. On the German sample, negative symptoms predominate, and, therefore, across the tested metrics, the correlation with the negative symptoms is stronger than with the positive. Additionally, SAPS, which is dedicated solely to the positive symptoms, was more detailed than the corresponding PANSS subscale. Thus, the correlation with SAPS across the metrics was stronger than with the positive PANSS subscale. Yet, SANS and PANSS negative subscale

were close to each other. The total PANSS score, as it encompasses both positive and negative symptoms, was correlated with most metrics. On the German sample, the correlation with mean sentence length was frequent for the well-performing metrics, yet many metrics outperformed this baseline. As for the group differences in the metrics on the German sample, they are discussed below for each metric type.

4.3.2 Russian

The same table (4.2) shows the overall effects of both tasks and scales for the Russian sample. On it, negative symptoms also predominate, and across all PANSS subscales, the symptoms are more severe than in the German sample. Thus, both positive and negative, as well as general symptoms correlated with some of the metrics. Like on the German sample, PANSS total score correlated with most metrics, however, all PANSS subscales followed closely, the positive symptoms being hardest to predict. TD severity was slightly more less predictable, being easiest to predict on one of the tasks (present), and depression severity could only be predicted on one task (chair). No metric was sensitive enough to reliably differentiate between the groups on the Russian sample, as all bootstrap quantile-based error bars crossed zero for all tasks and metrics[2].

The symptom severity was most strongly correlated with the metrics calculated on present task, followed by chair task, sportsman task, and adventure task. This means that the two picture-elicited tasks, sportsman and adventure, were the hardest to predict the symptom severity on.

As verbosity was more explained by mean sentence length for chair and adventure tasks, these two showed the most difference between the metrics that did not correlate with sentence length and those that did but outperformed it. On the other two tasks, the sentence length did not serve as a strong baseline and the aforementioned difference was therefore not as pronounced.

[2]Due to this absence of any meaningful difference between the groups, the t-test results for the Russian sample are not discussed in any more detail in the present work.

4.4 Lexical Methods

This section covers the performance of the selected lexical metrics, namely, lemma-token ratio (LTR), moving average lemma-token ratio (MALTR), and word count (n words) across target variables in both languages.

4.4.1 German

Figure 4.3 shows the performance of each lexical metric on SANS, SAPS, and PANSS subscales. On the German sample, LTR positively correlated with the negative symptom scales (PANSS negative and SANS), total PANSS, and PANSS general subscale. MALTR, on the other hand, weakly negatively correlated with both SANS and SAPS. Finally, word count negatively correlated with the negative symptom scales.

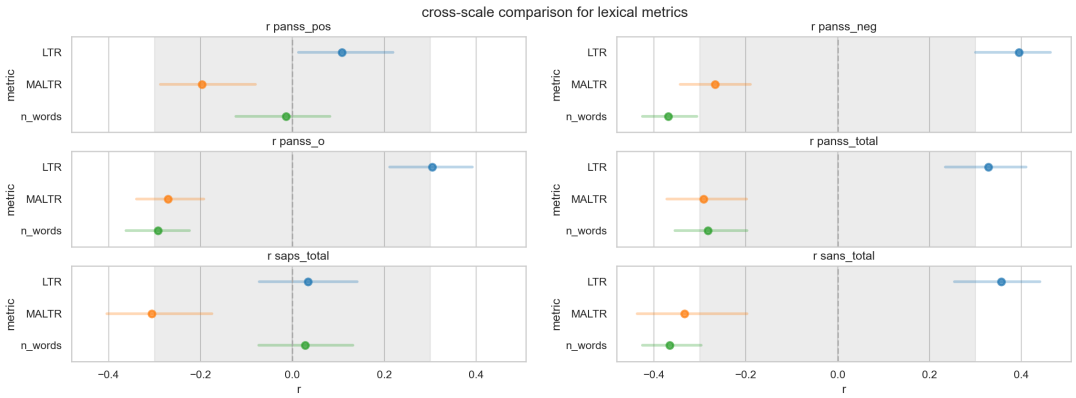

FIGURE 4.3: Pearson's r correlation coefficient with each scale for the lexical metrics on the German dataset. Grey indicates the values below the 0.3 threshold in absolute value.

As shown in figure 4.4, the correlation with mean sentence length closely matched the performance of each metric on the t-test. The LTR was higher in the patient group, while the word count was lower, with no difference in MALTR. All three metrics correlated with mean sentence length, LTR negatively, and MALTR and word count positively.

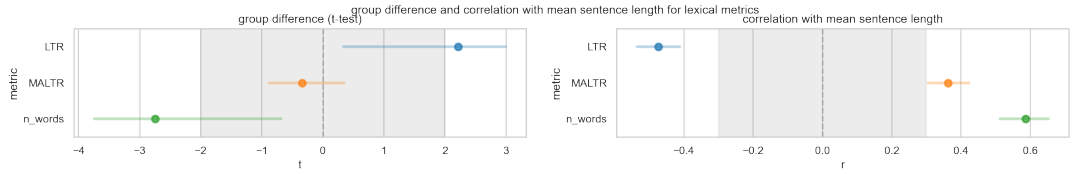

FIGURE 4.4: T-test and Pearson's r correlation coefficient with mean sentence length for the lexical metrics on the German dataset. Grey indicates the values below 2 for the t score and below the 0.3 threshold in absolute value for the correlation coefficient.

4.4.2 Russian

On adventure task, shown in figure 4.5, LTR did not perform on any of the scales. MALTR correlated negatively with PANSS negative and general scales, as well as the total PANSS score. Word count correlated negatively with all PANSS scales. None of the metrics could detect TD or depression severity.

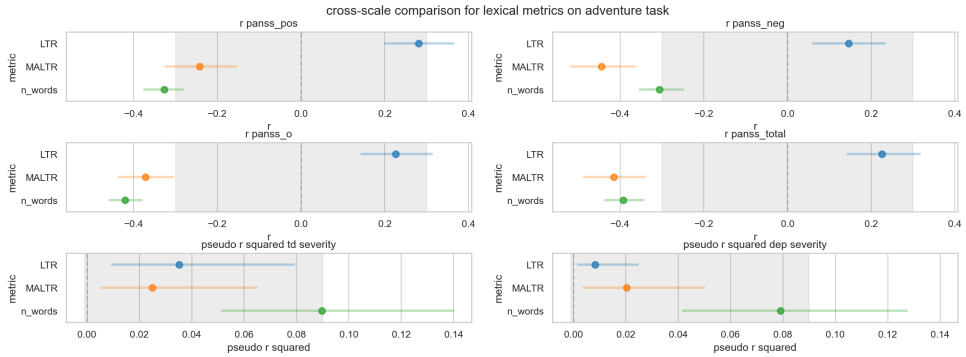

FIGURE 4.5: Pearson's r correlation coefficient and pseudo r squared for each scale for the lexical metrics on the Russian dataset, adventure task. Grey indicates the values below the 0.3 threshold in absolute value or pseudo r squared below 0.09.

On chair task, shown in figure 4.6, LTR correlated positively with all PANSS subscales, and was also predictive of depression severity. Conversely, MALTR and word count both correlated negatively with all PANSS subscales, and word count was also predictive of depression severity on chair task.

On present task, shown in figure 4.7, LTR correlated positively with all PANSS subscales and was predictive of TD severity. MALTR correlated with PANSS

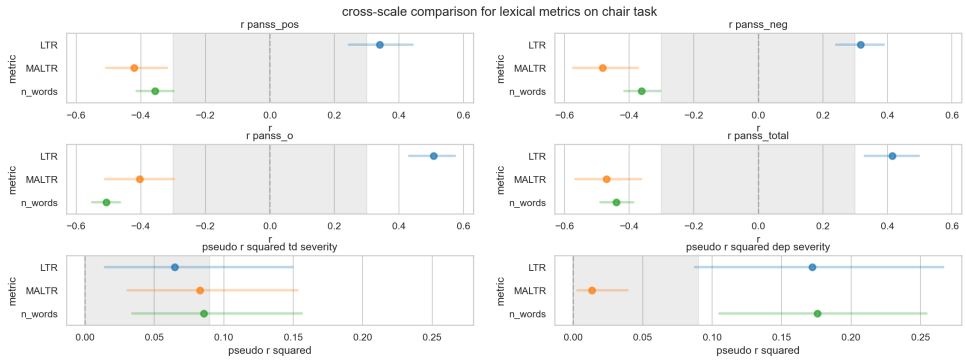

FIGURE 4.6: Pearson's r correlation coefficient and pseudo r squared for each scale for the lexical metrics on the Russian dataset, chair task. Grey indicates the values below the 0.3 threshold in absolute value or pseudo r squared below 0.09.

negative only, and word count correlated negatively with all PANSS subscales and was also predictive of TD severity.

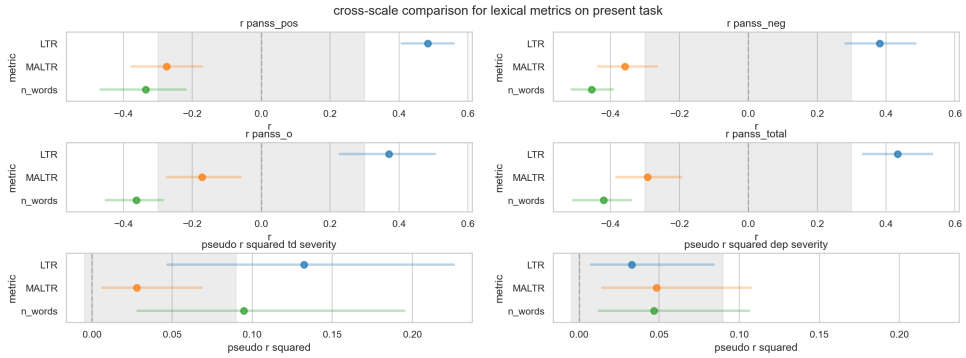

FIGURE 4.7: Pearson's r correlation coefficient and pseudo r squared for each scale for the lexical metrics on the Russian dataset, present task. Grey indicates the values below the 0.3 threshold in absolute value or pseudo r squared below 0.09.

Finally, on sportsman task, shown in figure 4.8, LTR also correlated positively with all PANSS subscales and was predictive of TD severity, and word count correlated negatively with all PANSS subscales and was also predictive of TD severity. MALTR did not perform on any of the scales on this task.

Figure 4.9 shows the strength of correlation with mean sentence length across tasks, indicating that MALTR positively correlated with mean sentence length on all tasks, while word count only does on chair task, and LTR only on sportsman task.

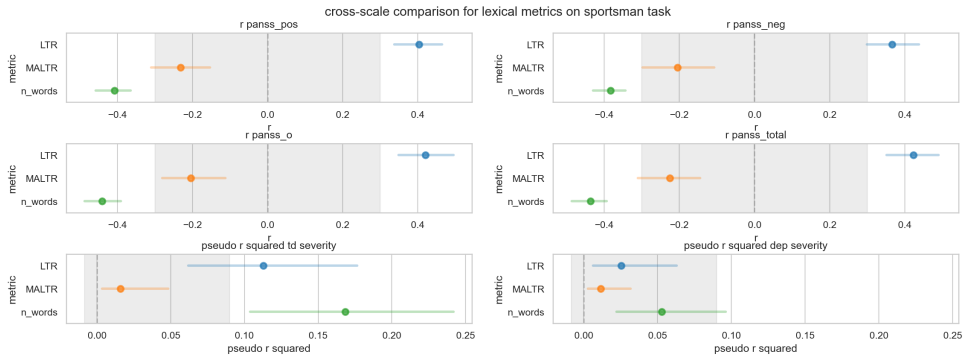

FIGURE 4.8: Pearson's r correlation coefficient and pseudo r squared for each scale for the lexical metrics on the Russian dataset, sportsman task. Grey indicates the values below the 0.3 threshold in absolute value or pseudo r squared below 0.09.

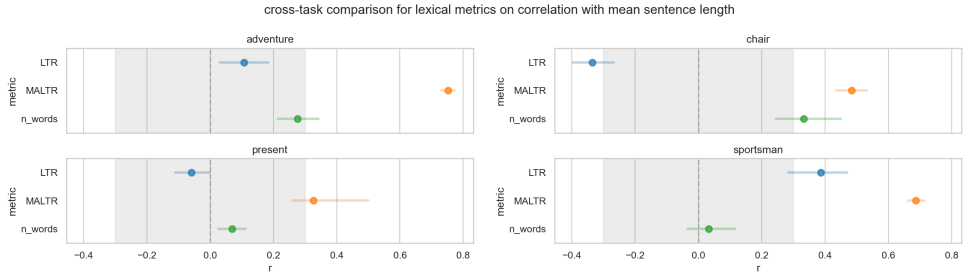

FIGURE 4.9: Pearson's r correlation coefficient with mean sentence length for the lexical metrics on the Russian dataset across tasks. Grey indicates the values below 0.3.

Overall, on the Russian sample, LTR correlated positively across PANSS subscales but for one task, and word count correlated negatively across PANSS subscales but for one task, while MALTR only performed on two tasks. LTR and word count were also predictive of TD severity on two tasks and of depression severity on one task, while not being overly strongly correlated with mean sentence length.

4.4.3 Cross-Linguistic Comparison

The direction of the correlation for the lexical metrics, was similar between the languages, while the strength of correlation as well as the correlation with mean sentence length differed across scales, tasks, and languages.

4.5 Syntactic Methods

This section covers the performance of syntactic metrics, including mean, maximal, and minimal sentence length, along with standard deviation in sentence length (mean_, min_, max_, and std_ sent_len, respectively), and total sentence count (n sents). Part-of-speech (POS) rates are also tested for the parts-of-speech that were shown to perform in some of the literature, namely: adjectives (ADJ), adverbs (ADV), auxiliary verbs (AUX), coordinating and subordinating conjunctions (CCONJ, SCONJ), determiners (DET), nouns and proper nouns (NOUN, PRPON), pronouns (PRON), particles (PART), and verbs (VERB).

4.5.1 German

The performance of the syntactic metrics on the German sample is shown in figure 4.10.

Among the POS rate metrics, only PART, AUX, and CCONJ correlated consistently with any of the scales. PART use correlated positively with all scales, but PANSS positive. AUX correlated negatively with all scales, but the positive ones, while CCONJ correlated negatively only with SAPS total score. As for the other syntactic metrics, maximum, mean, and minimal sentence length showed the best performance, as they correlated negatively with all negative scales as well as with PANSS total score, except for minimal sentence length. None of the other metrics correlated with any of the scales.

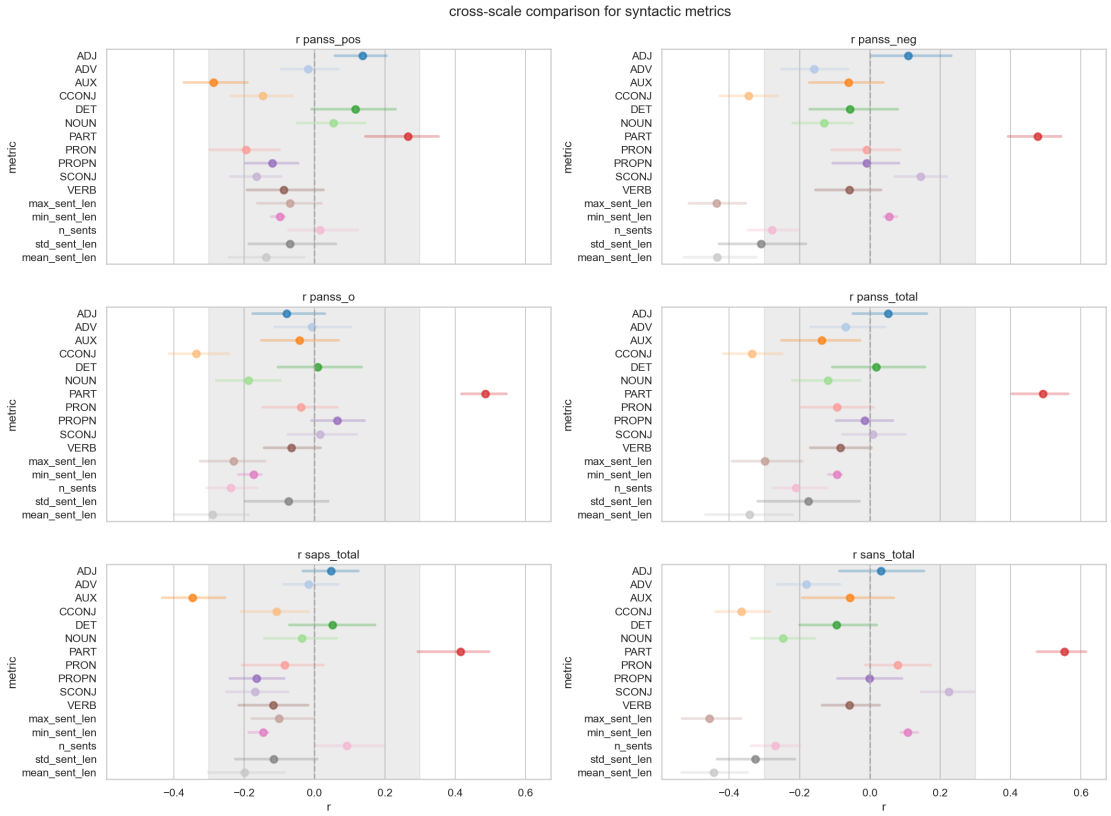

FIGURE 4.10: Pearson's r correlation coefficient with each scale for the syntactic metrics on the German dataset. Grey indicates the values below the 0.3 threshold in absolute value.

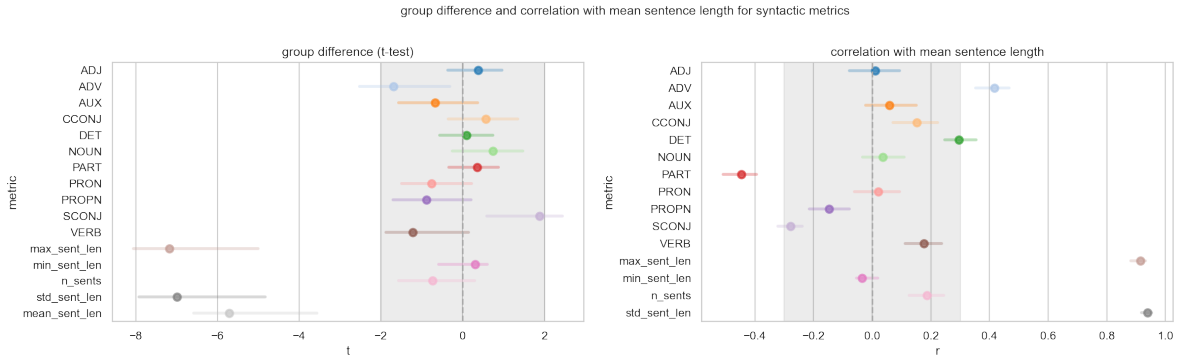

FIGURE 4.11: T-test and Pearson's r correlation coefficient with mean sentence length for the syntactic metrics on the German dataset. Grey indicates the values below 2 for t score and below the 0.3 threshold in absolute value for correlation coefficient.

Figure 4.11 compares the strength of the t-test to the corresponding correlation with mean sentence length. Mean sentence length is a reasonable baseline for the German sample, and both maximal and minimal sentence length, as could be expected, correlated strongly with it, somewhat outperforming it on the t-test. PART correlated negatively with mean sentence length, while ADV correlated positively with it.

4.5.2 Russian

Figure 4.12 shows the performance of the syntactic metrics on the adventure task, and the only metric that correlated with any of the symptom scales is the total number of sentences, which correlated negatively with general symptoms and total PANSS score.

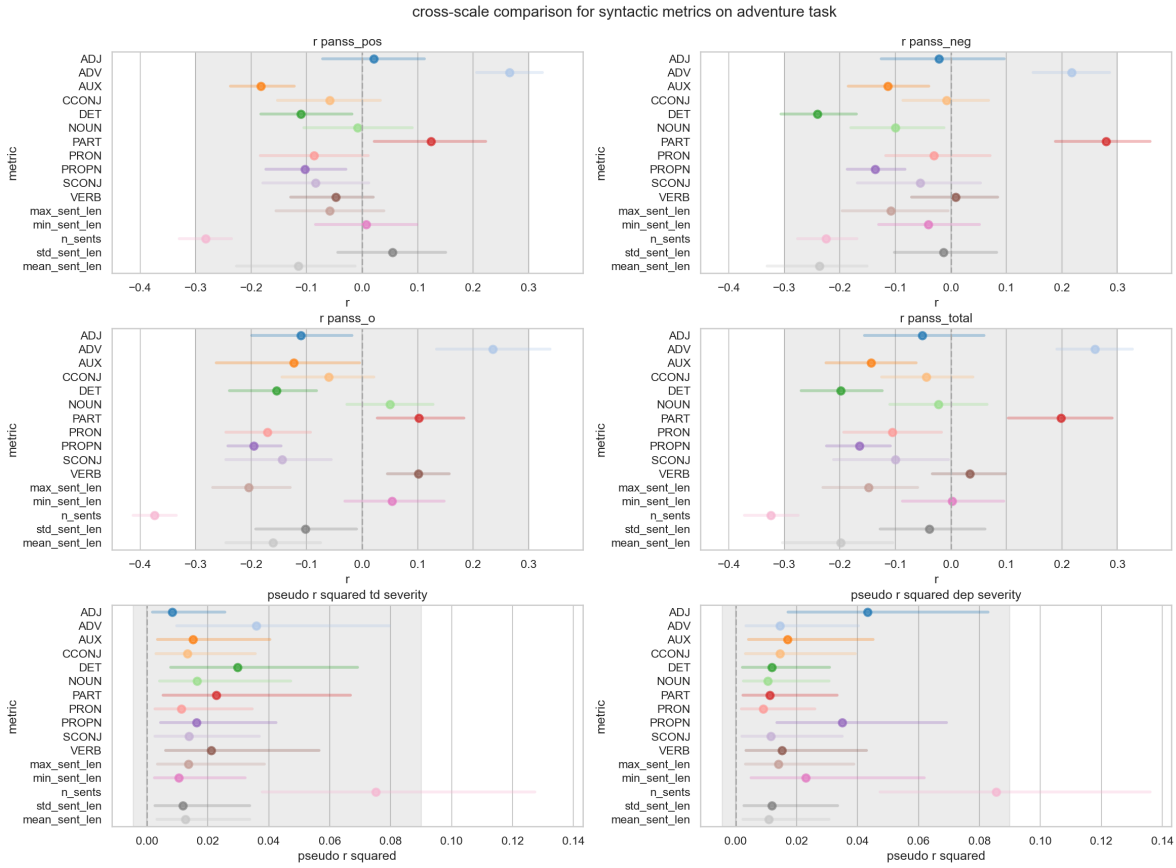

FIGURE 4.12: Pearson's r correlation coefficient and pseudo r squared for each scale for the syntactic metrics on the Russian dataset, adventure task. Grey indicates the values below the 0.3 threshold in absolute value or pseudo r squared below 0.09.

Figure 4.13 shows the performance of the syntactic metrics on the chair task. Among the parts of speech, PART rate correlated with all PANSS subscales, and NOUN rate with all but the general symptoms subscale. These two metrics were also the only ones performing well in TD severity detection. The mean sentence length correlated with all PANSS subscales as well as predicted depression severity along with maximal sentence length. The total number of sentences correlated with all PANSS subscales but PANSS negative.

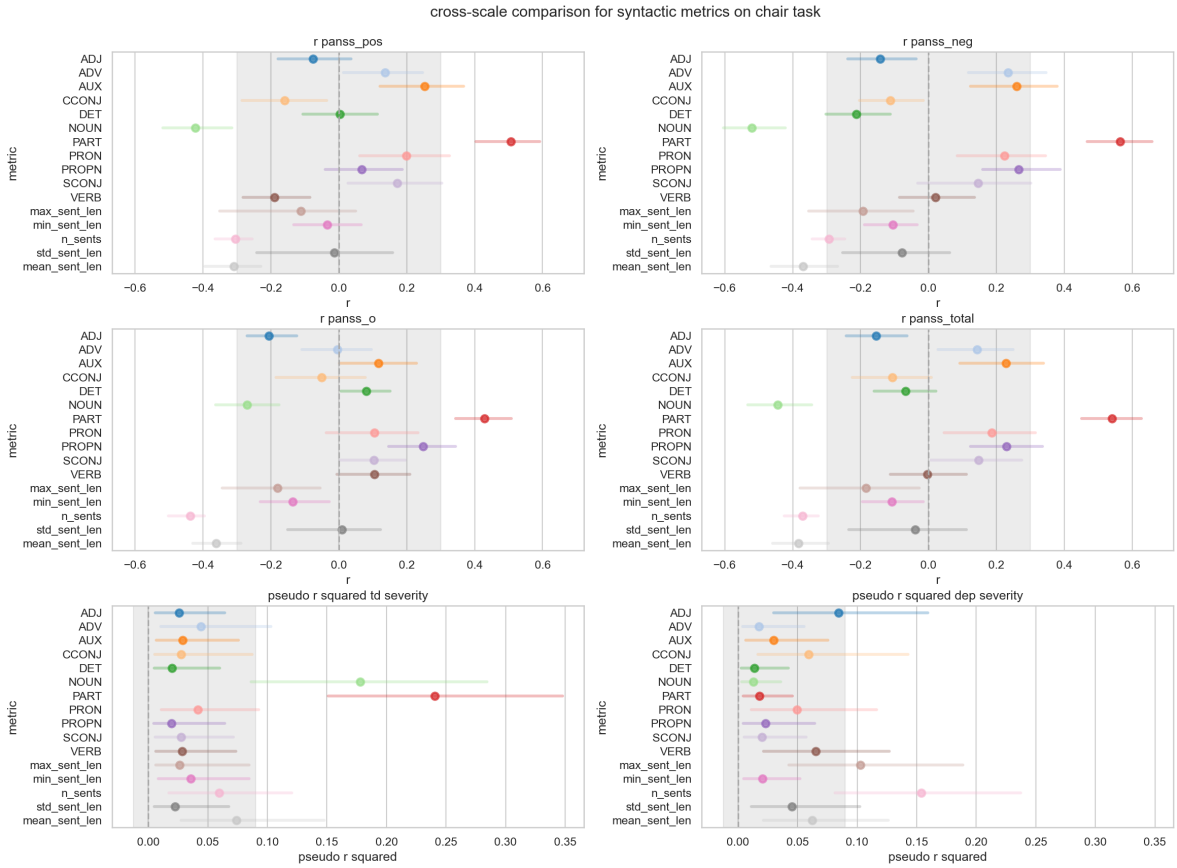

FIGURE 4.13: Pearson's r correlation coefficient and pseudo r squared for each scale for the syntactic metrics on the Russian dataset, chair task. Grey indicates the values below the 0.3 threshold in absolute value or pseudo r squared below 0.09.

Figure 4.14 shows the performance of the syntactic metrics on the present task. PART rate correlated positively with PANSS positive and total scores, DET rate correlated negatively with PANSS negative and total scores, and CCONJ rate correlated negatively with PANSS general only. The total number of sentences correlated negatively with PANSS negative only, while mean sentence length did so with all PANSS scales but the positive. Maximal sentence length and the standard deviation correlated negatively with all PANSS scales and also performed well in TD severity detection, largely outperforming mean sentence length.

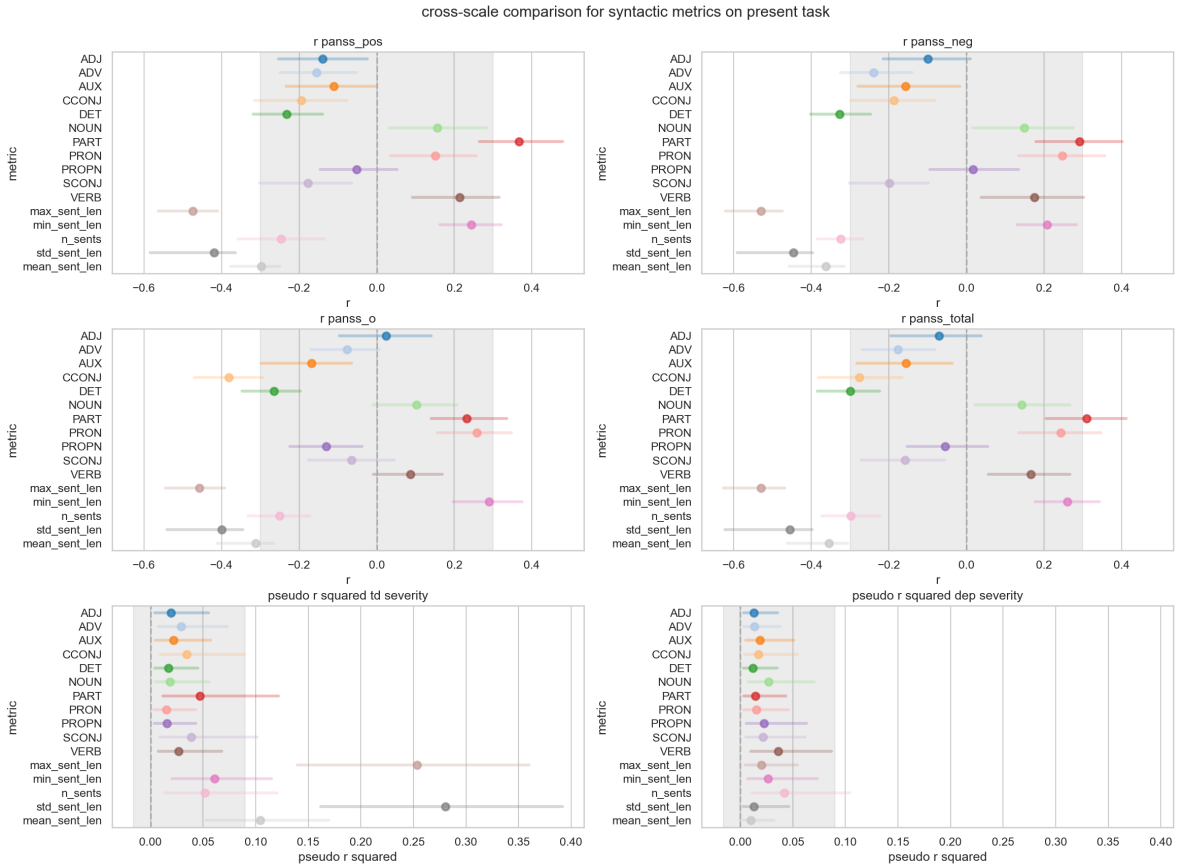

FIGURE 4.14: Pearson's r correlation coefficient and pseudo r squared for each scale for the syntactic metrics on the Russian dataset, present task. Grey indicates the values below the 0.3 threshold in absolute value or pseudo r squared below 0.09.

Figure 4.15 shows the performance of the syntactic metrics on the sportsman task. On this task, PART correlated positively only with PANSS positive, while maximal sentence length correlated with it negatively. The number of sentences on the other hand correlated negatively with all PANSS scales and was also the only metric sensitive to TD severity.

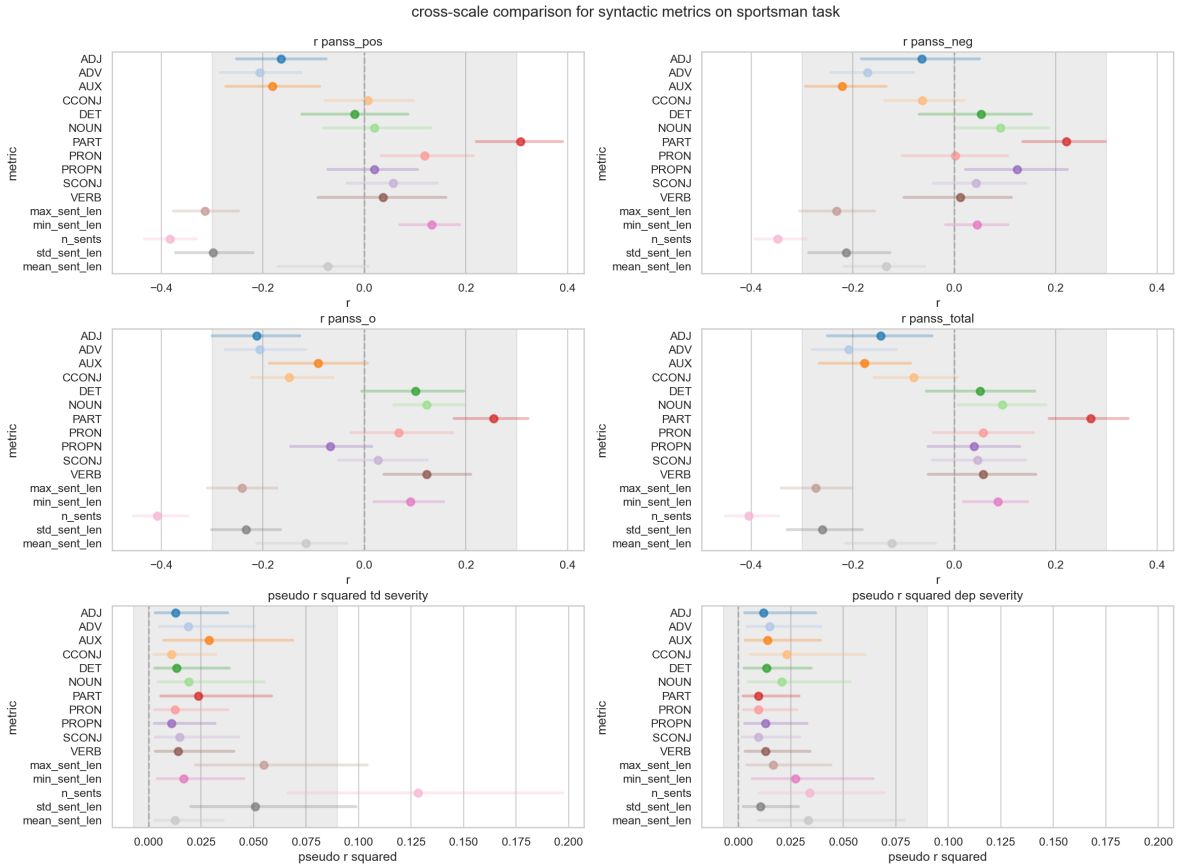

FIGURE 4.15: Pearson's r correlation coefficient and pseudo r squared for each scale for the syntactic metrics on the Russian dataset, sportsman task. Grey indicates the values below the 0.3 threshold in absolute value or pseudo r squared below 0.09.

Figure 4.16 shows the strength of correlation with mean sentence length across tasks. As could be expected, on all tasks, maximal, standard deviation, and minimal sentence length correlated positively with the mean. VERB rate correlated negatively with mean sentence length on sportsman and adventure tasks, and ADV did so on chair task, while ADJ correlated positively with mean sentence length on all tasks present, and NOUN did so on chair task. Importantly, it is on the chair task that NOUN performed well.

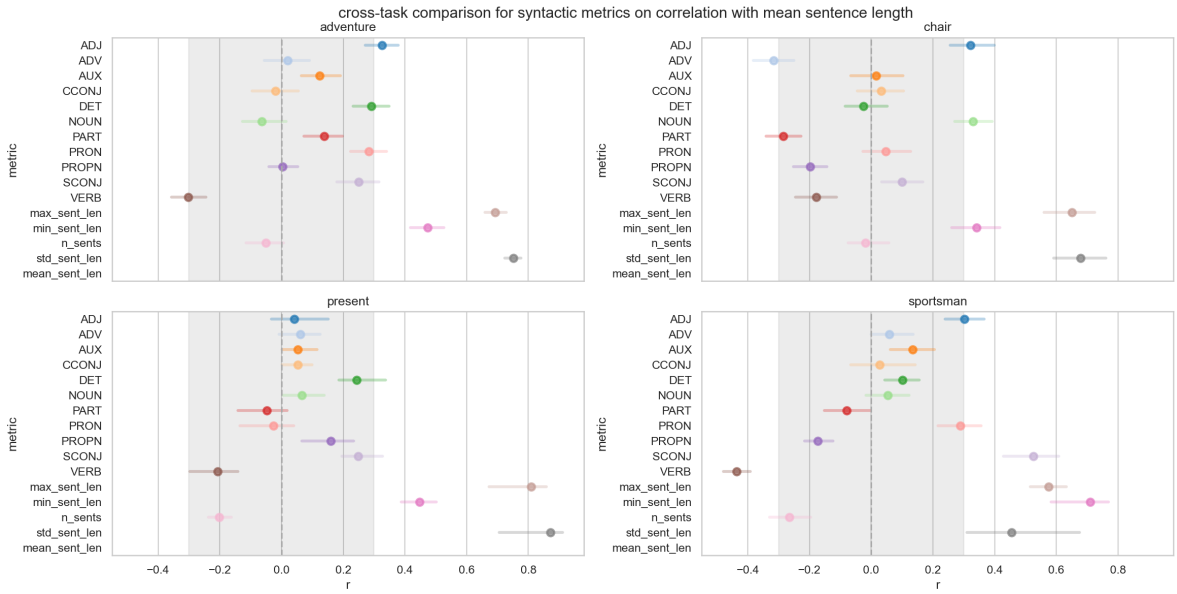

FIGURE 4.16: Pearson's r correlation coefficient with mean sentence length for the syntactic metrics on the Russian dataset across tasks. Grey indicates the values below 0.3.

Overall, on the Russian sample, the number of sentences was the strongest metric, being the only one, that correlated with symptom scales on all tasks. The rate of PART was positively correlated with symptom severity on three of the four tasks. As could be expected, mean sentence length served as a good baseline on only two of the tasks (chair and present), and maximal sentence length performed on the same tasks, though it correlated with fewer subscales that mean sentence length. As for, CCONJ, DET, and NOUN rates, each only performed on one of the tasks, and the same is true of standard deviation in mean sentence length.

4.5.3 Cross-Linguistic Comparison

In both languages, PART rate correlated positively with PANSS scales, and similarly in both languages mean, maximal, and standard deviation sentence length correlated negatively with PANSS scales, with mean sentence length performing better on the German sample, as could be expected. CCONJ rate correlated negatively both on the German and the Russian samples, though for the latter only on one scale for one task. AUX rate correlated negatively only on the German sample, and NOUN and DET did so only on the Russian sample, both on only one task. Similarly, the number of sentences correlated negatively with PANSS scales only on the Russian sample.

4.6 Graph-Based Methods

This section covers the performance of co-occurrence graph-based metrics, number of nodes (N), number of edges (E), largest connected component (LCC), largest strongly connected component (LSC), number of parallel edges (PE), number of loops of length one (repeated lemmas), two, and three (L1, L2, L3), average node degree (degree average), and standard deviation in the node degree (degree std).

4.6.1 German

Among the graph metrics, shown in figure 4.17, four correlated negatively with negative symptom scales, as well as PANSS general and total scores, namely, the largest connected and strongly connected component size as well as the number of nodes and edges.

Figure 4.18 shows the strength of correlation with mean sentence length and the power of bidirectional t-test, with the patterns corresponding very closely between the two graphs. The metric values were calculated for moving window of size 100, to avoid direct correlation with verbosity, yet there was still a significant correlation with mean sentence length for all the metrics that performed well on t-test and the psychiatric scales, i.e. LCC, LSC, N, and E.

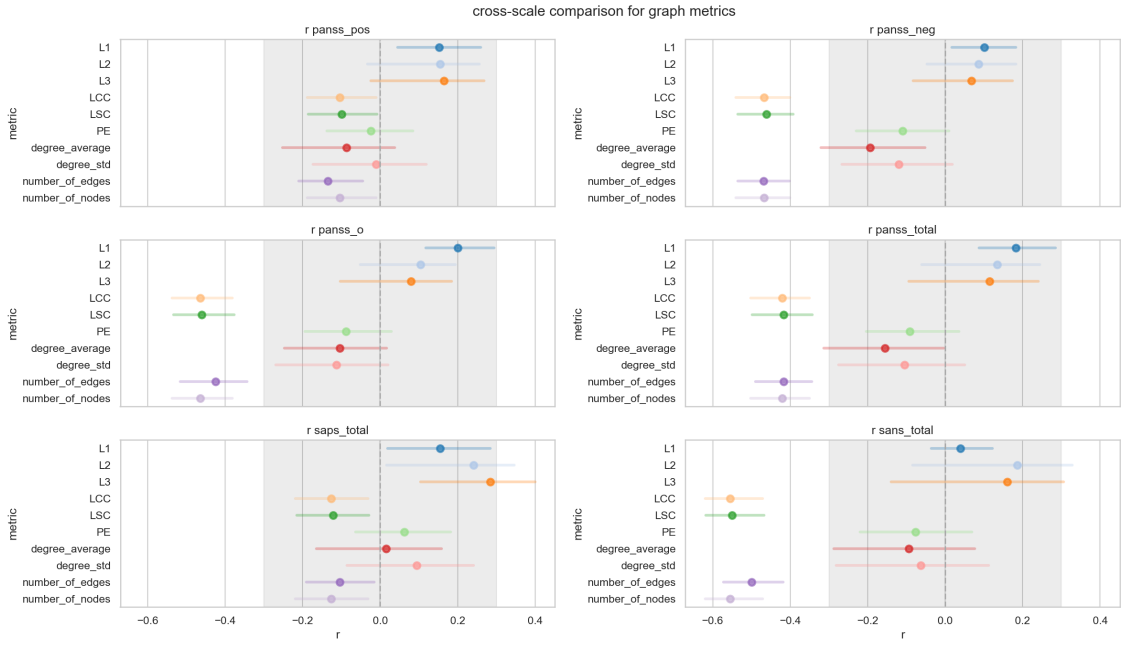

FIGURE 4.17: Pearson's r correlation coefficient with each scale for the graph-based metrics on the German dataset. Grey indicates the values below the 0.3 threshold in absolute value.

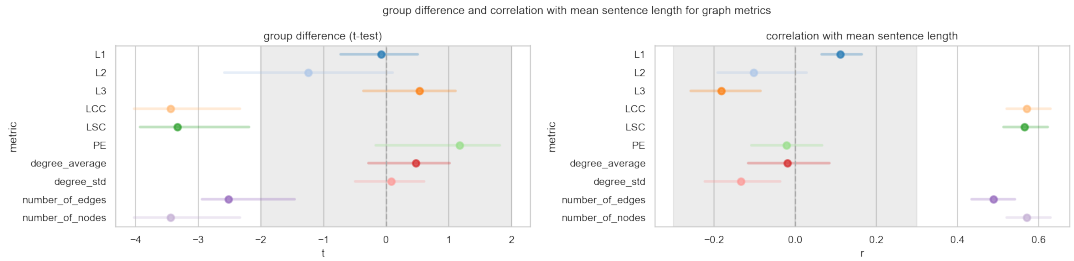

FIGURE 4.18: T-test and Pearson's r correlation coefficient with mean sentence length for the graph-based metrics on the German dataset. Grey indicates the values below 2 for t score and below the 0.3 threshold in absolute value for correlation coefficient.

4.6.2 Russian

Figure 4.19 shows the performance of graph-based metrics on adventure task. The largest connected and strongly connected component size as well as the number of nodes and edges correlated negatively with PANSS general and total scores, and all of these but the number of edges also correlated with PANSS negative.

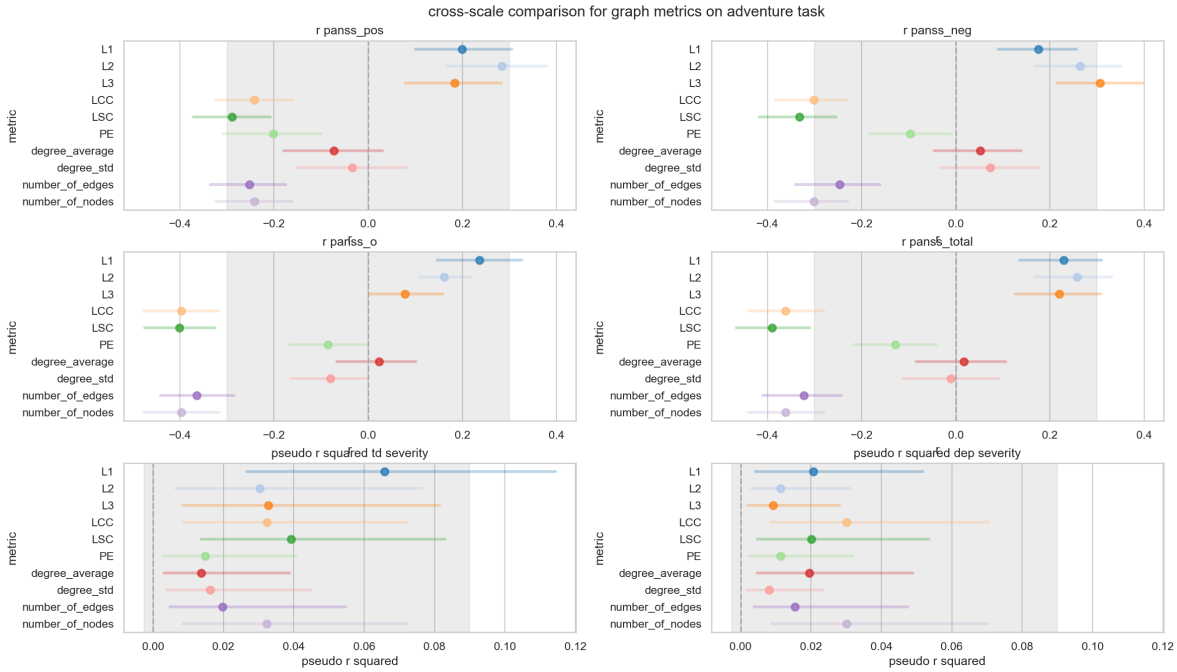

FIGURE 4.19: Pearson's r correlation coefficient and pseudo r squared for each scale for the graph-based metrics on the Russian dataset, adventure task. Grey indicates the values below the 0.3 threshold in absolute value or pseudo r squared below 0.09.

Figure 4.20 shows the performance of graph-based metrics on chair task. The largest connected and strongly connected component size as well as the number of nodes and edges correlated negatively with all PANSS scales. L3 correlated positively with all PANSS scales and was predictive of TD severity, while PE negatively correlated with PANSS positive and total scores, and was, alongside the number of edges, predictive of depression severity.

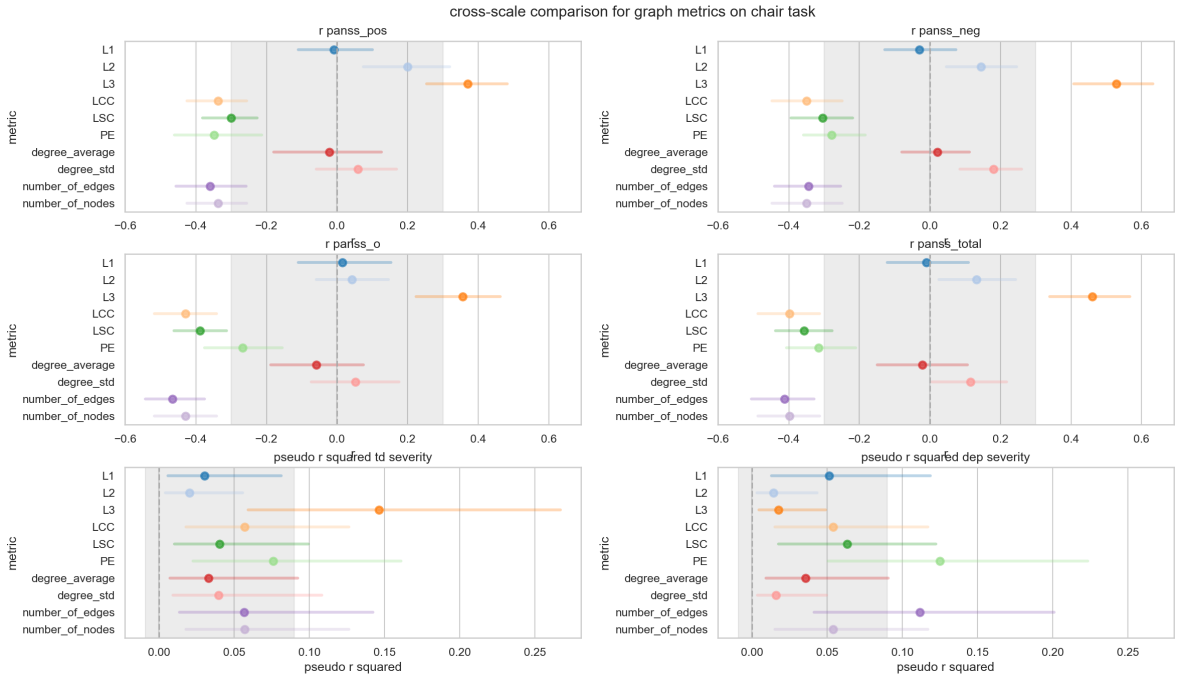

FIGURE 4.20: Pearson's r correlation coefficient and pseudo r squared for each scale for the graph-based metrics on the Russian dataset, chair task. Grey indicates the values below the 0.3 threshold in absolute value or pseudo r squared below 0.09.

Figure 4.21 shows the performance of graph-based metrics on present task. The largest connected and strongly connected component size as well as the number of nodes and edges correlated negatively with all PANSS scales and were also predictive of thought disorder severity. Additionally, on present task, average node degree and standard deviation in node degree correlated negatively with PANSS positive and total scores, and average degree was somewhat predictive of TD severity as well. L1 correlated slightly with PANSS general score.

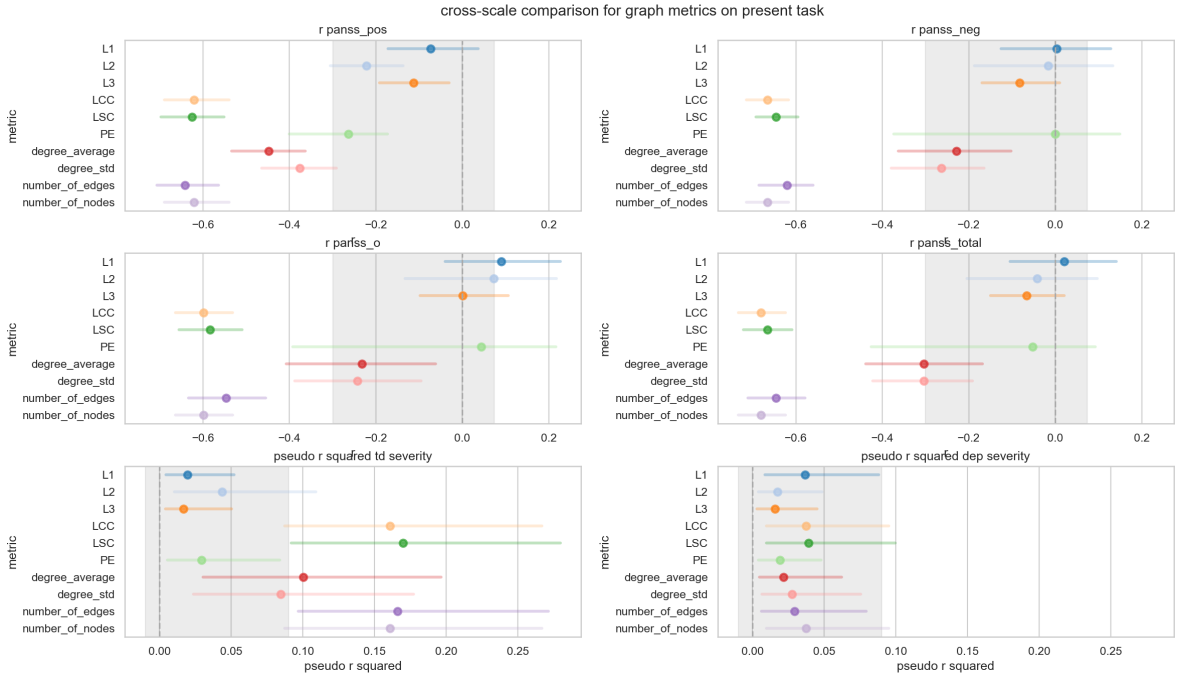

FIGURE 4.21: Pearson's r correlation coefficient and pseudo r squared for each scale for the graph-based metrics on the Russian dataset, present task. Grey indicates the values below the 0.3 threshold in absolute value or pseudo r squared below 0.09.

Finally, figure 4.22 shows the performance of graph-based metrics on sportsman task. As on other tasks, the largest connected and strongly connected component size as well as the number of nodes and edges correlated negatively with all PANSS scales. The number of parallel edges correlated negatively with all PANSS subscales but general and was also predictive of TD severity.

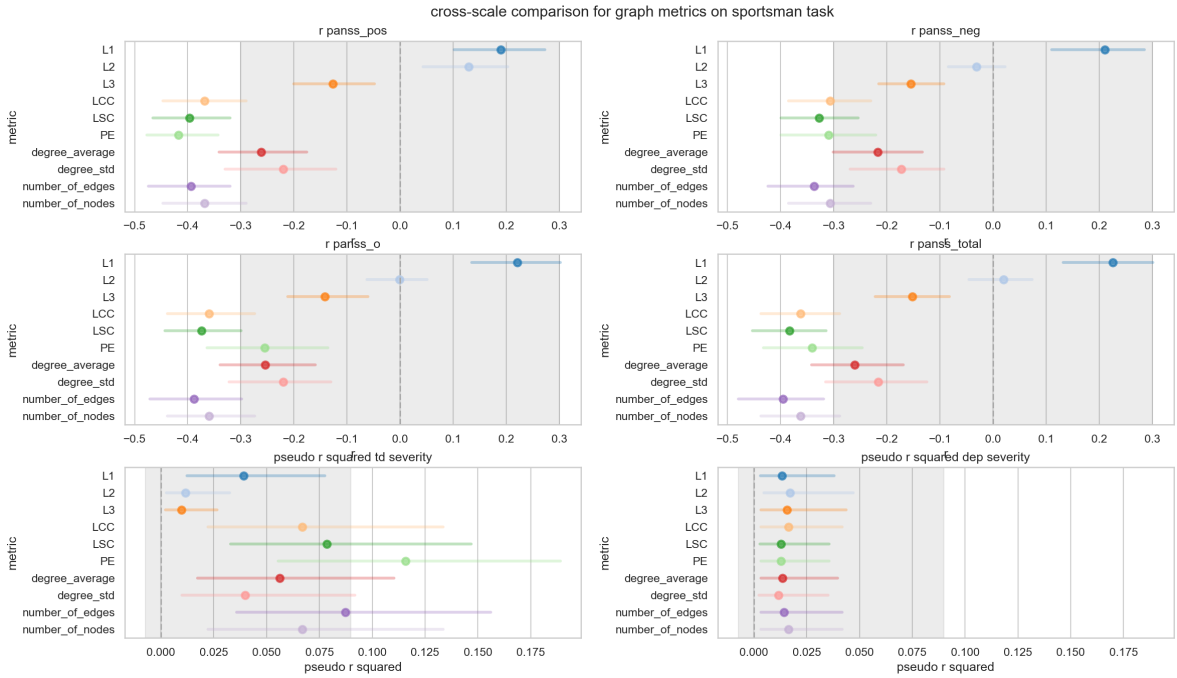

FIGURE 4.22: Pearson's r correlation coefficient and pseudo r squared for each scale for the graph-based metrics on the Russian dataset, sportsman task. Grey indicates the values below the 0.3 threshold in absolute value or pseudo r squared below 0.09.

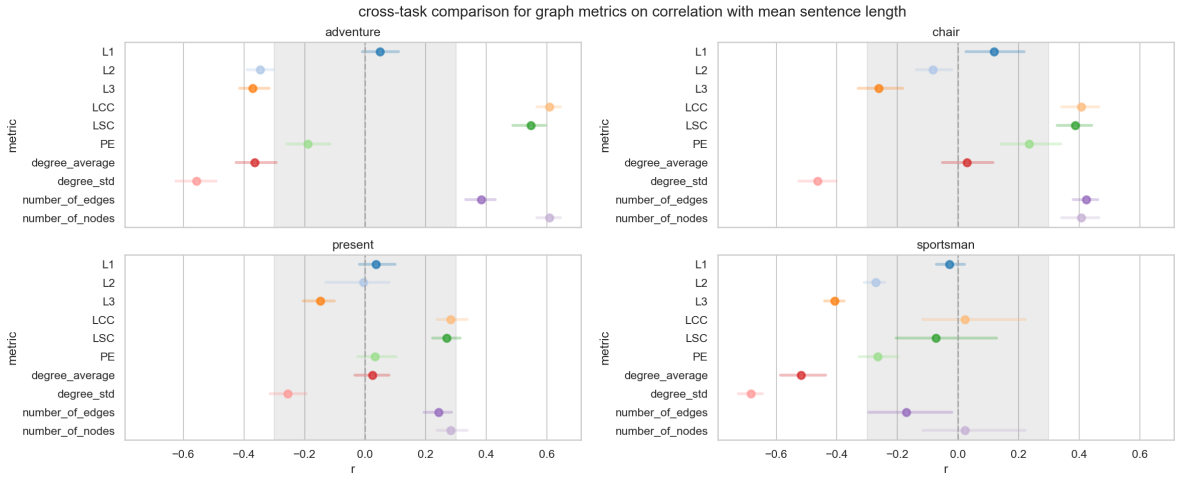

FIGURE 4.23: Pearson's r correlation coefficient with mean sentence length for the graph-based metrics on the Russian dataset across tasks. Grey indicates the values below 0.3.

Figure 4.23 shows the strength of correlation with mean sentence length across tasks. LCC, LSC, N, and E correlated positively with mean sentence length

on adventure and chair tasks. L2 correlated negatively with mean sentence length on adventure task, and L3, average node degree, and standard deviation of node degree did so on adventure and sportsman tasks. The standard deviation of node degree also correlated negatively with mean sentence length on chair task.

Overall, among the graph-based metrics, LCC, LSC, N, and E showed the best performance on the Russian sample, as they correlated with some psychiatric scales across all of the tasks, though they also were correlated with mean sentence length on the two tasks where it could be expected, as they probably depended somewhat on the verbosity. The number of parallel edges performed on two of the four tasks, chair and sportsman, and the average and standard deviation in node degree, as well as loops of size one (lemma repetitions) and three, only performed on one task, namely, present.

4.6.3 Cross-Linguistic Comparison

Both on the German and Russian samples, LCC, LSC, N, and E were negatively correlated with negative and general symptoms scales and were also positively correlated with mean sentence length, though on the Russian sample, this correlation was present only for some tasks. As positive symptoms were more pronounced in the Russian sample, the same metrics were also correlated negatively with PANSS positive scale and on some tasks predictive of TD and depression severity. L1, L3, PE, average node degree, and standard deviation in node degree, were only correlated with symptom scales on the Russian sample.

4.7 Language Model-Based Methods

This section covers the results of the language model-based metrics for both languages. The metrics included cumulative global coherence (`cgcoh`), global coherence (`gcoh`), local coherence (`lcoh`), second order coherence (`scoh`), next sentence probability (`sprob`), and pseudo-perplexity (`pppl`). The models included BERT and two weighting schemes for word2vec. `BERT` is denoted as such, `w2v` stands for word2vec (fasttext), `glove` stands for GloVe, and `tf` on figures stands for TF-IDF weighted sentence scheme, while `avg` for simple sentence averaging.

4.7.1 German

Figure 4.24 shows the performance of Language Model-based metrics on the German sample.

Overall, there was a hierarchy of metrics, with pseudo-perplexity performing the best, positively correlating with negative and general symptom severity. The second-best performing metric was second-order coherence, which correlated negatively across models with negative symptoms and total PANSS, and on all models but averaged GloVe with general PANSS. Local coherence and next sentence probability showed similar levels of performance, with local coherence correlating stronger on word2vec and averaged GloVe, but not BERT, TF-IDF weighted word2vec even correlating negatively with positive symptom severity. Next sentence probability only reached the threshold in negative correlation with PANSS negative and total. Finally, global and cumulative global coherence performed the worst and did not correlate strongly with any of the scales.

As for the models, on the cosine similarity-based metrics, word2vec showed the best performance with TF-IDF averaging followed by simple averaging, and GloVe, where the pattern was reversed, simple averaging outperforming TF-IDF. For both models, the difference between the averaging schemes was slight. Finally, BERT performed worst for the cosine similarity-based metrics, but the two feature-based metrics performed relatively well.

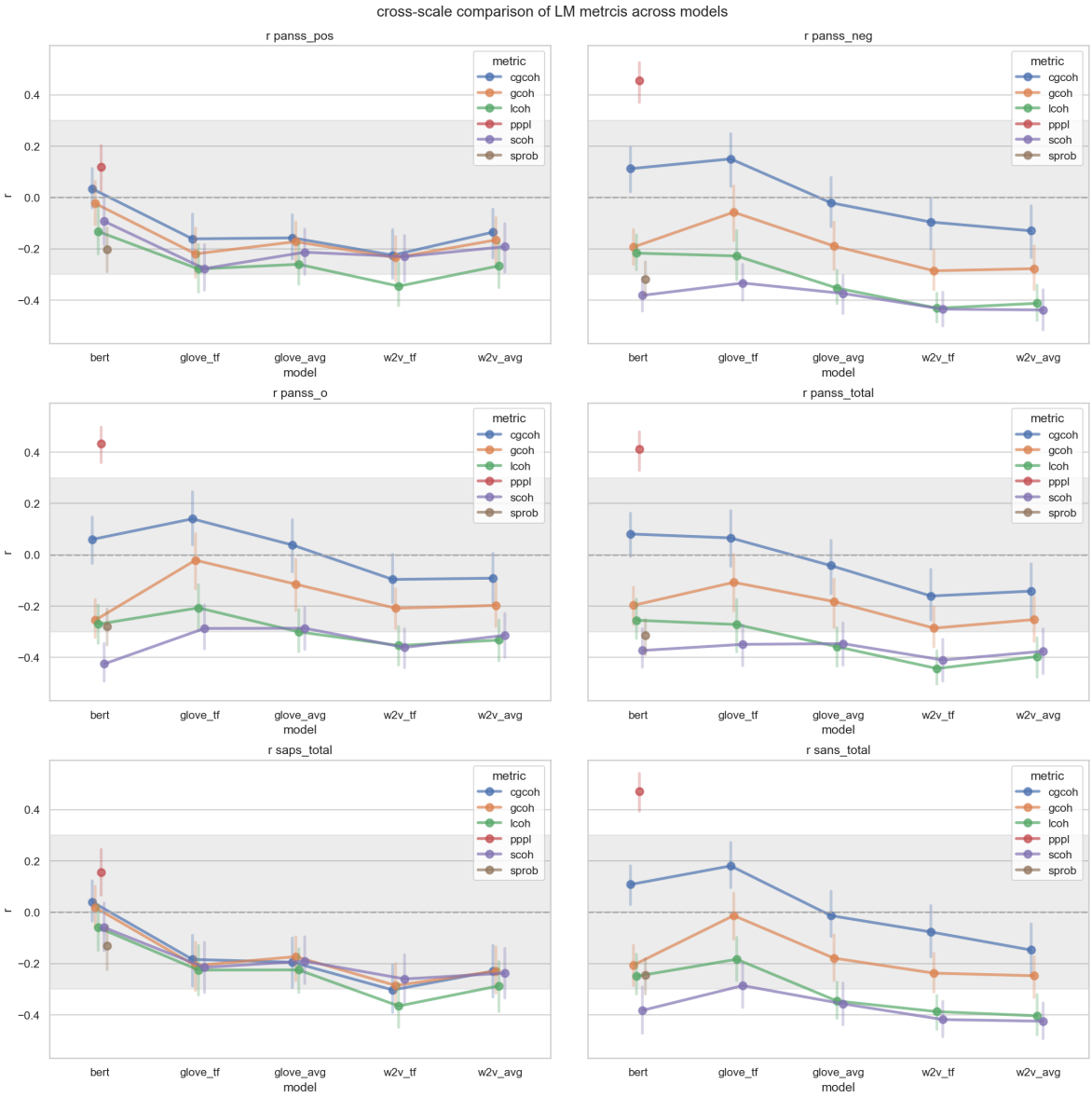

FIGURE 4.24: Pearson's r correlation coefficient with each scale for the LM-based metrics on the German dataset. Grey indicates the values below the 0.3 threshold in absolute value.

Figure 4.25 shows the results of a bidirectional t-test against the strength of correlation with mean sentence length. Except for pseudo-perplexity, which was somewhat higher in patients than in controls, no metric could differentiate between the groups.

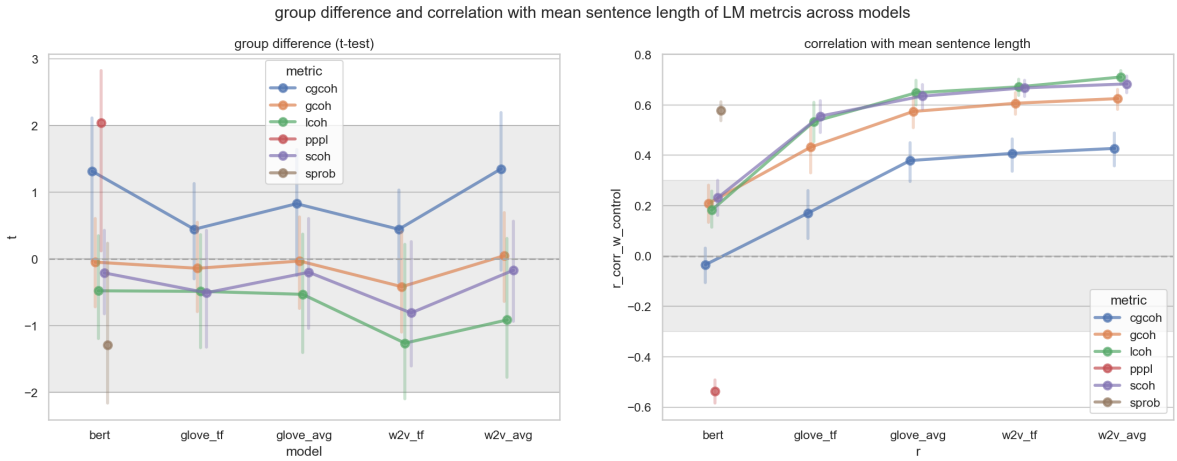

FIGURE 4.25: T-test and Pearson's r correlation coefficient with mean sentence length for the LM-based metrics on the German dataset. Grey indicates the values below 2 for t score and below the 0.3 threshold in absolute value for correlation coefficient.

The cumulative global coherence was least correlated with mean sentence length, followed by global coherence, and then all other metrics. Interestingly, both next sentence probability and pseudo-perplexity were strongly correlated with mean sentence length, though the former was correlated positively and the latter negatively. For the cosine similarity-based metrics, BERT was uncorrelated with mean sentence length. GloVe-based metrics were less correlated than w2v, and the TF-IDF weighted metrics were somewhat less correlated than simple averaged ones. There was an inverse relation trend between overall metric performance and its correlation with mean sentence length, being more pronounced for the correlation with symptom scales, than for the T-test results.

There was an interesting pattern, that cumulative global coherence tended to correlate less negatively (and, in some cases, more positively) than other metrics, and it was followed by global coherence, while local and second-order coherence tended to correlate more negatively with symptom severity.

4.7.2 Russian

Out of the four tasks, on one no LM metric correlated with any of the psychiatric scales, while on the other three, there was some predictive power.

Figure 4.26 shows the comparative LM-based metric performance across scales, metrics, and models on adventure task.

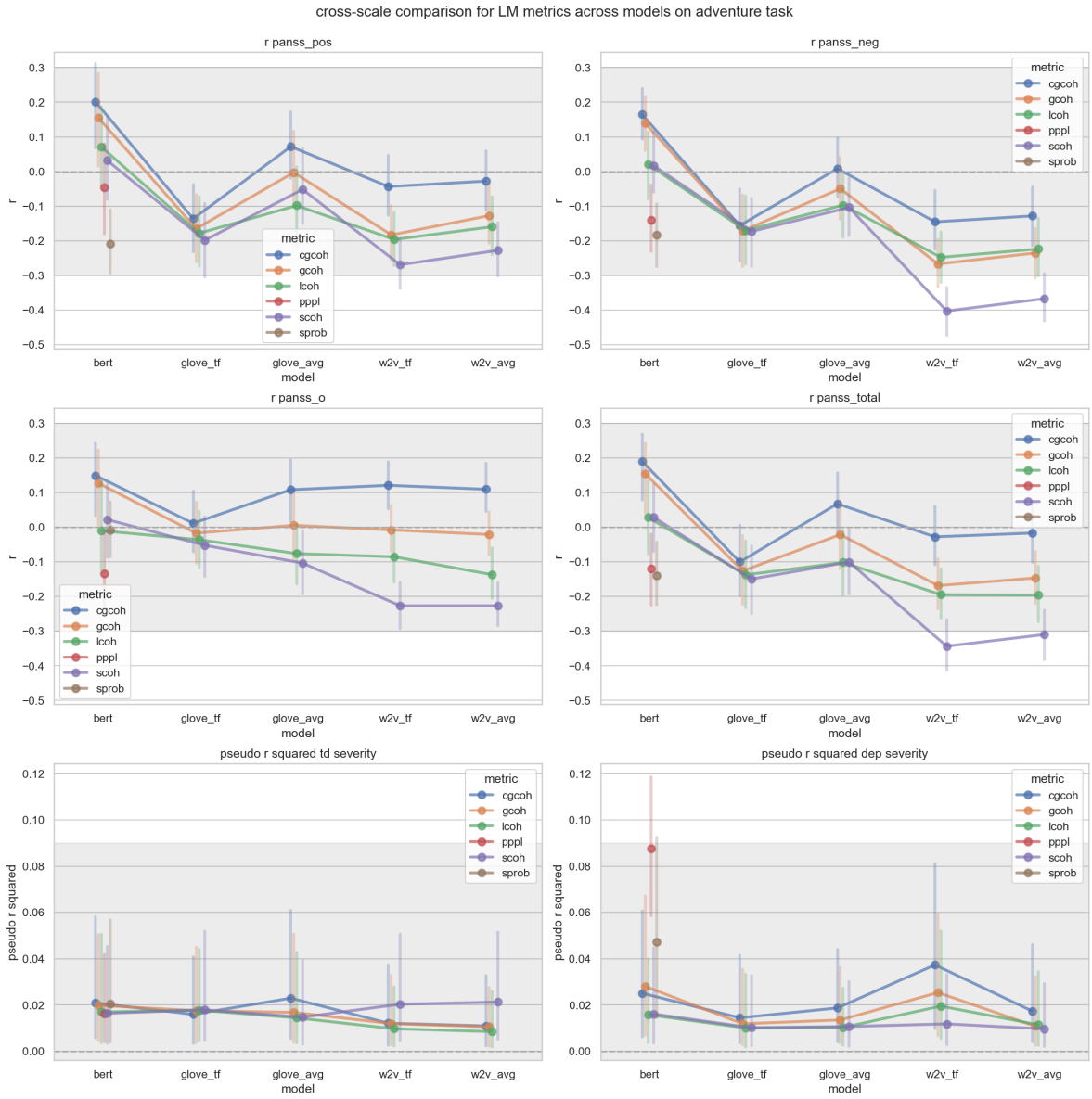

FIGURE 4.26: Pearson's r correlation coefficient and pseudo r squared for each scale for the language model-based metrics on the Russian dataset, adventure task. Grey indicates the values below the 0.3 threshold in absolute value or pseudo r squared below 0.09.

There was only one metric, that interacted with any scales, namely, second-order coherence, which correlated negatively with PANSS negative and total

when calculated with either word2vec weighting scheme. For other metrics, even tough they did not reach the threshold, there was a hierarchy in performance, with second-order coherence being followed by next sentence probability and pseudo-perplexity, and then by local coherence and the two global coherence metrics.

word2vec, on average, seemed to outperform the other models, with TF-IDF averaging improving the performance compared to simple averaging for both word2vec and GloVe. BERT under-performed on this task, except for the feature-based metrics. Cosine similarity-based metrics calculated using BERT embeddings, unlike those calculated with either non-contextualized embedding model, tended to correlate positively, rather than negatively, with symptom severity.

Figure 4.27 shows the comparative LM-based metric performance across scales, metrics, and models on chair task.

Only on this task, depression severity could be predicted, and GloVe TF-IDF could predict depression severity with all metrics but cumulative global coherence. Additionally, averaged word2vec cumulative global coherence correlated positively with general symptoms.

Generally, on this task cumulative global coherence correlated most strongly with symptom severity, followed by pseudo-perplexity, while local coherence and second-order coherence performed worse, and the next sentence prediction showed the worst performance.

Among the models, there was no clear patter on this task, though GloVe outperformed word2vec, and TF-IDF outperformed simple averaging, with BERT showing middle correlation strength.

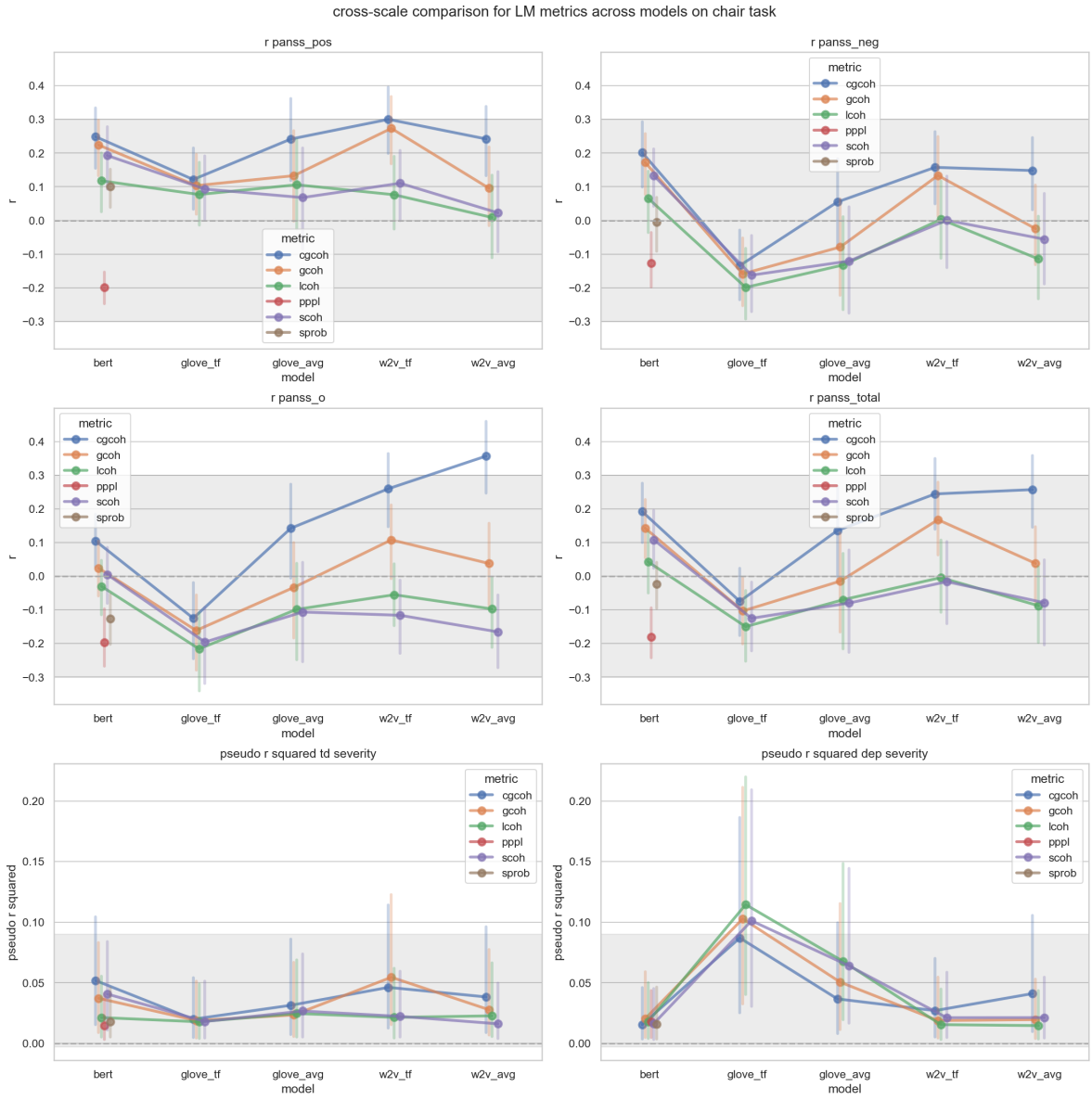

FIGURE 4.27: Pearson's r correlation coefficient and pseudo r squared for each scale for the language model-based metrics on the Russian dataset, chair task. Grey indicates the values below the 0.3 threshold in absolute value or pseudo r squared below 0.09.

Figure 4.28 shows the comparative LM-based metric performance across scales, metrics, and models on present task.

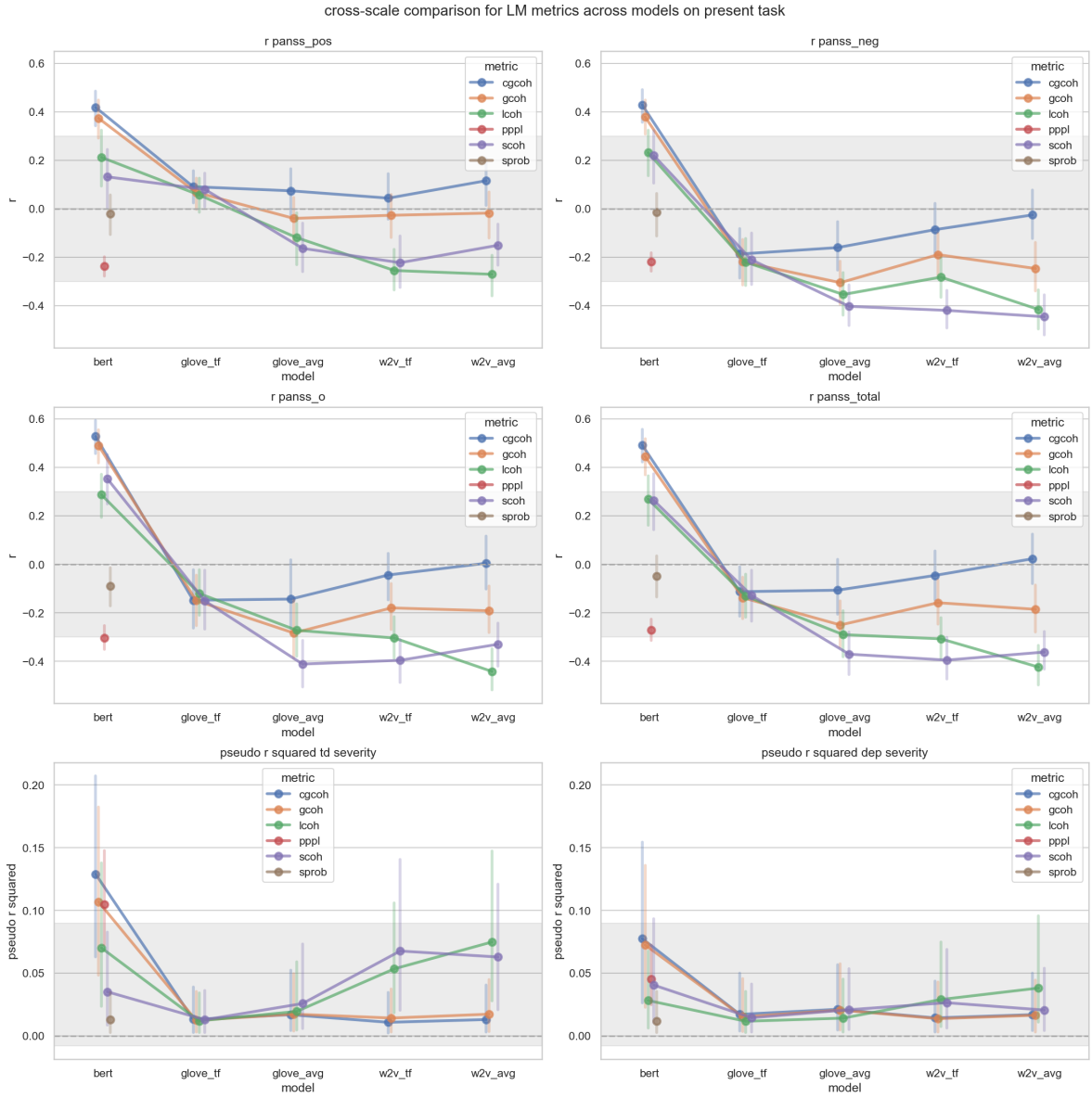

FIGURE 4.28: Pearson's r correlation coefficient and pseudo r squared for each scale for the language model-based metrics on the Russian dataset, present task. Grey indicates the values below the 0.3 threshold in absolute value or pseudo r squared below 0.09.

On present task, BERT global and cumulative global coherence correlated positively with all scales, while pseudo-perplexity correlated negatively with

general symptoms, and they were all predictive of TD severity. Second-order coherence calculated using word2vec with either averaging or GloVe with simple averaging correlated negatively with all PANSS subscales but positive. Local coherence calculated using word2vec correlated negatively with PANSS general and total scores, and when calculated using GloVe with simple averaging it correlated negatively with PANSS negative score.

Pseudo-perplexity showed comparatively good performance on this task. When assessed across all models, second-order and local coherence performed best, while global and cumulative global coherence showed worse performance, and next sentence probability barely differed from zero.

On this task, as BERT global coherence scores correlated positively with all scales, it showed the best performance of all models. As for the non-contextualized embeddings, word2vec outperformed GloVe, and simple averaging outperformed TF-IDF in absolute correlation strength.

Figure 4.29 shows the strength of correlation with the mean sentence length across tasks, models, and metrics.

Cosine similarity-based metrics calculated using BERT, correlated negatively with mean sentence length, though the correlation was not equally strong across tasks and metrics, being above the threshold for all metrics on chair and sportsman tasks; only for local and global coherence on the adventure task; and for no metrics on present task. Next sentence probability and pseudo-perplexity did not correlate with length on any of the tasks. As for the metrics calculated using GloVe and word2vec, they correlated positively, rather than negatively, with mean sentence length. This correlation was stronger on averaged than on TF-IDF weighed embeddings for all tasks when word2vec was used, and for GloVe it was so on two tasks, sportsman and present, while on chair task the opposite was observed, and almost no difference was seen on the adventure task. All metrics calculated using word2vec correlated with mean sentence length on three tasks and on the fourth, present, second order coherence calculated using TF-IDF averaging did not exceed the threshold. The correlation was above the threshold for all

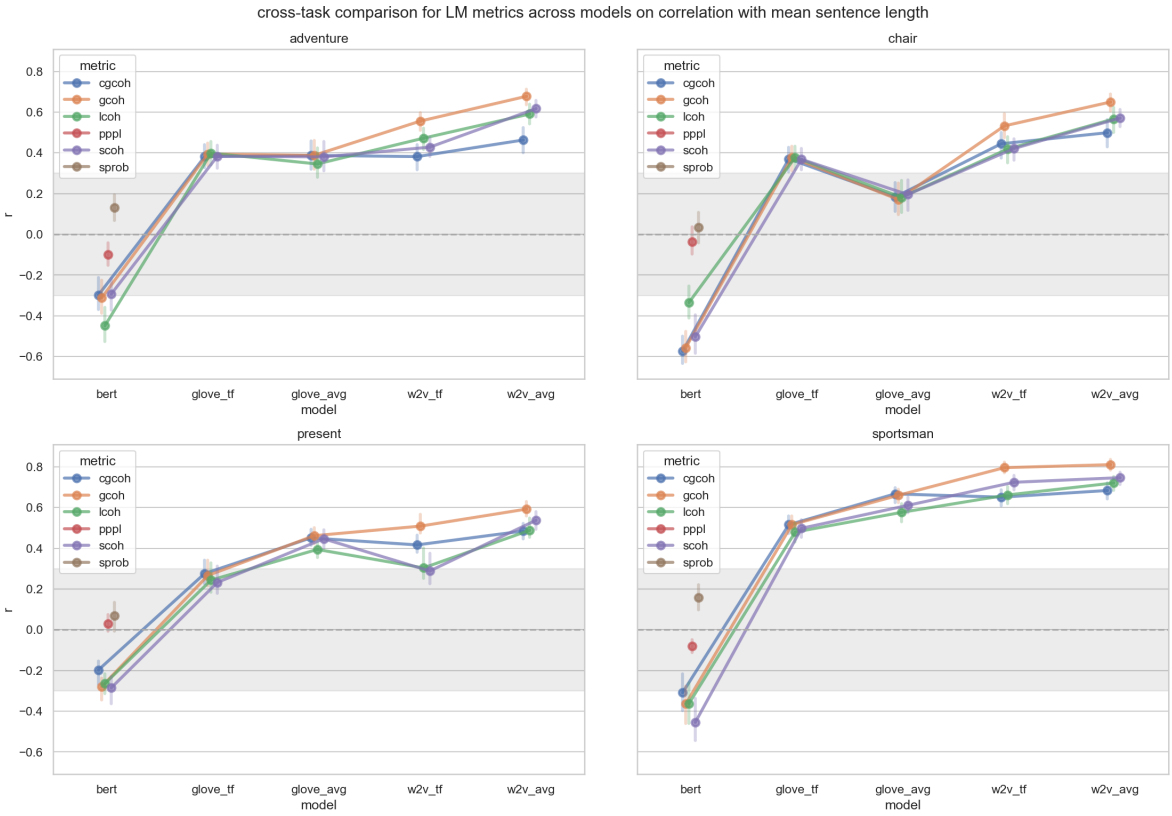

FIGURE 4.29: Pearson's r correlation coefficient with mean sentence length for the language model-based metrics on the Russian dataset across tasks. Grey indicates the values below 0.3.

GloVe metrics on sportsman and adventure tasks and was also so for TF-IDF GloVe on chair and for averaged GloVe on present tasks.

Interestingly, across all tasks, cosine similarity-based metrics calculated using BERT showed a positive correlation with the symptom scales and a negative correlation with mean sentence length, while the two non-contextualized embedding models tended to show a negative correlation with the symptoms scales and a positive one with the sentence length.

There was no clear hierarchy between models or metrics across all tasks, as there was significant interaction between metrics and models, which was also different for different tasks. On average, perplexity, cumulative global

coherence, and second-order coherence performed better in absolute correlation strength, while local and global coherence followed closely, and the next sentence probability tended to underperform. BERT and word2vec tended to outperform GloVe, and TF-IDF weighting tended to outperform simple averaging. The performance of word2vec and GloVe seems to be at least partially explained by correlation with mean sentence length.

Across the tasks, there was the same pattern, that cumulative global coherence tended to correlate more positively (or less negatively) than other metrics, and it was followed by global coherence, while local and second-order coherence tended to correlate more negatively. There was, however, a curious difference between the models, as cosine similarity-based metrics calculated using BERT tended also to correlate more positively (or less negatively) than either word2vec or GloVe, which both tended towards a negative correlation. These patterns, when combined, caused a significant model-metric interaction.

4.7.3 Cross-Linguistic Comparison

There was, for LM-based metrics, little similarity between the languages or tasks. Interestingly, even the direction of correlation with symptoms differed between the samples, as BERT metrics tended to correlate positively with symptoms scales on the Russian sample, but negatively on the German one, where there was no difference in correlation direction between BERT and non-contextualized models. Similarly, pseudo-perplexity correlated positively with symptom severity on the German sample, but negatively on the Russian one, and the reverse was true for next sentence probability. Both these metrics correlated with mean sentence length on the German sample, not on the Russian one, where they were the least strongly correlated of all the metrics.

There was a similarity in the tendency for more positive (or less negative) correlation for cumulative global and global coherence across the models and for more negative correlation with symptom severity for local and especially second-order coherence, which was present for both languages.

On the German sample, there was a much clearer hierarchy of models and metrics, which was absent from the Russian sample across the tasks, as there were large differences in performance between them. On both samples, BERT correlated least with mean sentence length, though here, also, the direction of correlation differed between the samples, as on the Russian sample BERT-based metrics tended to correlate negatively with mean sentence length. On both samples TF-IDF clearly helped mitigate the correlation with length, yet, on the Russian sample, this pattern was somewhat weaker and task-dependent. Cumulative global coherence was clearly the least length-dependent for the German sample, but this pattern, if at all present, was also much weaker on the Russian sample. The relative performance of GloVe and word2vec, the latter outperforming the former, seems to be inversely related to the strength of their respective correlation with mean sentence length.

4.8 Cross-Group Metric Comparison

This section compares the performance of the metrics across metric types taking into account only the metrics that showed good performance, comprehending t-test error bars not intersecting zero, above threshold correlation with any of the scales, or above threshold predictive power. Additionally, metrics that were correlated with mean sentence length and performed worse than this baseline on all scales were excluded.

4.8.1 German

Figure 4.30 compares the performance of the metrics across metric groups for all the psychiatric scales on the German sample. Among the metrics that performed well independently of sentence length, was BERT second-order coherence, which correlated negatively with the negative symptoms as well as general and total PANSS scores, and the same correlation pattern was also apparent for the rate of CCONJ. The only other length-independent metric was the AUX rate which correlated negatively with total SAPS score.

Mean sentence length could only serve as a moderate baseline on the negative scales and PANSS total score. The rate of PART was correlated with length but consistently outperformed this baseline, and correlated positively with all the scales but PANSS positive. Four graph metrics, LCC, LSC, N, and E correlated with length but consistently outperformed it, negatively correlating with all but the two positive symptom scales. LTR showed weak positive correlation patterns, as it only outperformed mean sentence length on general PANSS, and for negative MALTR correlation, the only such scale was total SAPS score.

BERT pseudo-perplexity score correlated more strongly than the mean sentence length baseline with negative and general symptom scales. BERT second-order coherence was uncorrelated with mean sentence length, and performed below the baseline on PANSS negative, but above it on SANS and PANS general and total scores. Averaged GloVe local coherence, which was correlated with length, barely correlated with general PANSS scale and was

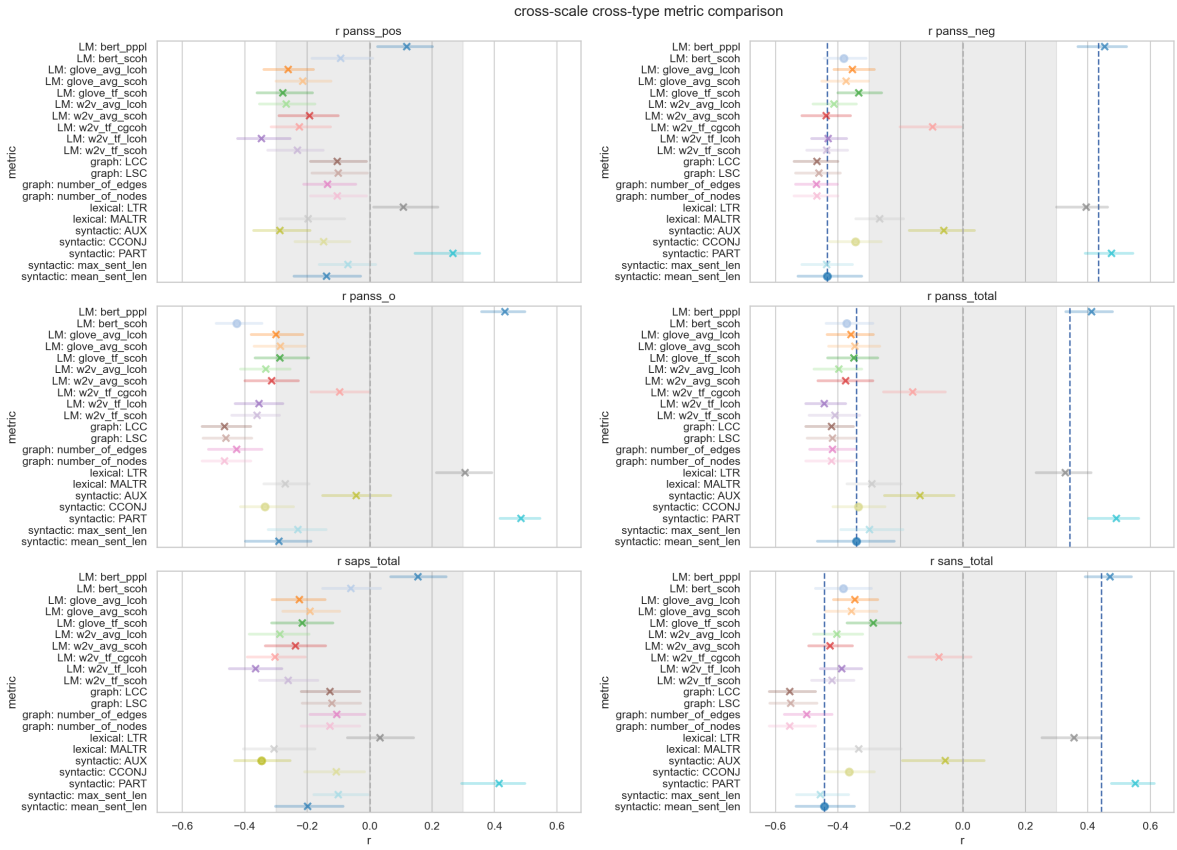

FIGURE 4.30: Pearson's r correlation coefficient with each scale across well-performing metrics on the German dataset. Grey indicates the values below the 0.3 threshold in absolute value. The crossed dots indicate the metrics indicate either above-threshold correlation with mean sentence length or below-threshold correlation on a given scale. The mean sentence length baseline is shown by blue dashed line on the scales where this metric serves as a baseline.

barely above mean sentence length in the negative correlation with PANSS total score. Averaged GloVe second-order coherence was similarly barely above mean sentence length on the negative correlation with PANSS total score. As for word2vec, all the metrics calculated using it correlated with mean sentence length, but frequently outperformed it, with TF-IDF weighted local coherence correlating with positive symptoms scales and outperforming length on the total PANSS score. Second-order coherence on TF-IDF weighted word2vec was weaker, outperforming the baseline only for PANSS general. Averaged local and second-order coherence calculated with word2vec

were barely above the threshold and baseline for PANSS general and total. On the negative scales, where mean sentence length served as a reasonable baseline, all GloVe metrics were below it, while word2vec TF-IDF weighted and simply averaged second-order coherence barely outperformed the baseline for PANSS negative. Maximum sentence length barely outperformed mean sentence length on SANS total but on no other scale.

As negative symptoms predominated in the German sample, it was unsurprising that the performance was overall better on the negative and general symptom scales, with the only exception of the AUX rate. As could be expected, more pronounced patterns could be seen on SANS and SAPS than on the corresponding PANSS subscales. Overall, on the German sample, syntactic metrics showed the most promise, taking into account the mean sentence length baseline. Graph-based and lexical metrics seemed to be to some significant extent explaining the same effects as mean sentence length, yet outperformed this baseline in some cases. Finally, LM, except for pseudo-perplexity, was the least reliable group, with weak correlations, barely ever outperforming mean sentence length.

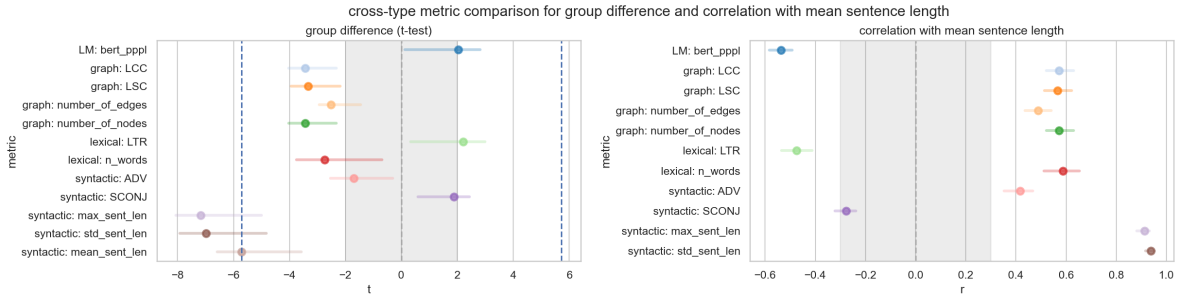

FIGURE 4.31: T-test and Pearson's r correlation coefficient with mean sentence length for the LM-based metrics on the German dataset. Grey indicates the values below 2 for t score and below the 0.3 threshold in absolute value for correlation coefficient. Only the metrics, for which error bars on T-test do not intersect zero, are shown.

Figure 4.31 compares the t-test effect size to the correlation with mean sentence length for the metrics, the bootstrap 25/75 percentile error bars of which

did not intersect zero on the t-test. For all the metrics, the t-test effect size corresponds quite closely with the strength and to the direction of the correlation with mean sentence length. The mean sentence length itself also served as a very strong baseline for the t-test, outperformed, but not significantly so, only by the maximum and standard deviation in mean sentence length.

4.8.2 Russian

Figure 4.32 compares the performance of the metrics averaged across tasks for all the psychiatric scales on the Russian sample. The metrics that did not, on average, strongly correlated with length, and performed somewhat well were as follows. The number of words correlated negatively with all PANSS scales and was also predictive of TD severity. LTR, being inversely related to the word count, performed similarly but in the opposite direction, correlating positively with all PANSS scales, yet less strongly than the simple word count. The number of sentences on average correlated negatively with all PANSS scales but positive. Interestingly, the number of parallel edges was on average slightly above the baseline for the positive PANSS subscale.

On the Russian sample, the mean sentence length did not, on average, serve as a reasonable baseline, because the differences were more pronounced in the number of sentences and words, rather than the length of sentences, as discussed above (4.1). LCC, LSC, N, and E consistently outperformed this weak baseline and correlated negatively with all PANSS scales. The rate of PART correlated above baseline for all PANSS scales but the general. The standard deviation in mean sentence length was, on average, barely above the baseline in predicting TD severity. No LM metric was above baseline for any of the scales, when averaged across the four tasks. As depression severity could only be predicted on one task, no metric averaged across tasks, was predictive above baseline, with the number of words and sentences being the strongest predictors. No metric, even on the individual tasks, could differentiate between the groups reliably, as all error bars intersected zero.

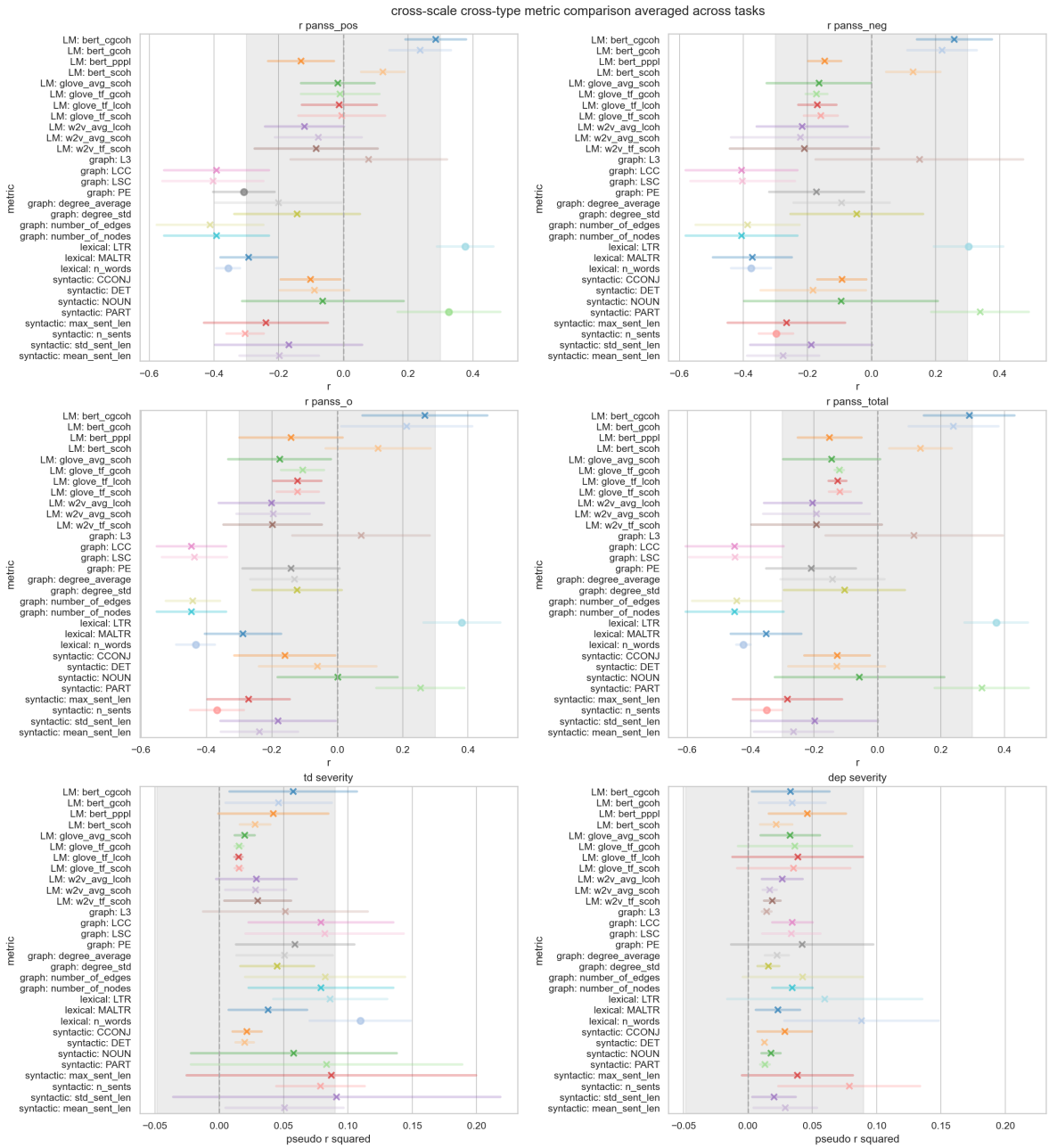

FIGURE 4.32: Pearson's r correlation coefficient with each scale across well-performing metrics on the Russian dataset, averaged across tasks. Grey indicates the values below the 0.3 threshold in absolute value. The crossed dots indicate the metrics indicate either above-threshold correlation with mean sentence length or below-threshold correlation on a given scale. Unlike other plots in this section, the error bars represent standard deviation, and the dot itself is the average value across the four tasks.

Though there were some interactions between the scales and the metrics, the best-performing metrics showed similar results across scales, and the correlation direction was also the same.

On the Russian sample, verbosity was the most robust though not the strongest metric, but otherwise, the tested graph-based metrics outperformed the lexical and syntactic ones, and LM-based ones were by far the weakest.

4.8.3 Cross-Linguistic Comparison

The difference between the languages was about as large as the difference between the tasks on the Russian sample, which, as the tasks were different between the languages, could be the main yet not the only cause for this difference.

On both samples, LCC, LSC, N, and E were correlated with length, yet outperformed it in the strength of negative correlation with the psychiatric scales. PART rate, on the other hand, was also correlated with length and yet outperformed it in the strength of correlation, correlating positively with the psychiatric scales for both samples. LTR and MALTR correlated with length and were much weaker, sometimes performing worse than the mean sentence length baseline, especially on the German sample, where it was stronger. On the German sample mean sentence length was a stronger baseline than word count or sentence count, while the opposite was true on the Russian sample.

CCONJ rate performed well for the German sample, but weak on average, for the Russian sample. The mean sentence length, due to the differences covered in section 4.1, provided a moderate baseline for the German, but not the Russian sample. Instead, on the Russian sample the numbers of words and sentences were much stronger.

Among LM metrics, there was a surprising pattern, as the direction of correlation for BERT metrics was positive on the Russian sample but negative on the German one, and the direction of correlation was inverted for feature-based metrics as well as for cosine similarity-based ones. The perplexity performed well for the German, but not the Russian sample. Across both

samples, cumulative global and global coherence correlated more positively with symptom severity than local and second-order coherence, which tended to correlate more negatively. Finally, there was a difference in the strength of correlation with mean sentence length between the languages. However, there was an overall pattern, more evident on the German sample, of BERT being least correlated with sentence length, followed by the TF-IDF and then by simply averaged non-contextualized models, and GloVe being less correlated than word2vec.

In both languages, relatively simple metrics, such as the word count, sentence count, or the number of nodes in the co-occurrence graph, i.e. moving window unique lemma count, showed the best performance, as compared to more complex metrics, such as LM-based ones. Co-occurrence graphs showed a promising result overall, and so did the syntactic metrics, though, the POS rates less so that simple unit length and count.

Chapter 5

Discussion

This chapter is dedicated to the interpretation of the results described above. The chapter covers each group of methods and compares the results with the ones reported in other studies (5.1-5.4). After that, there is a comparison of cross-methodological results (5.5), a discussion of study limitations (5.6), and a short conclusion (5.7).

5.1 Lexical Methods

As in other studies in the field, lower overall verbosity was observed on the German sample, though, on the Russian sample, the t-test error bar crossed zero even for this metric. Overall verbosity was also negatively correlated with all PANSS subscales, and, on some tasks, with TD and depression severity, meaning that any severe symptom was associated with a decreased word count for the Russian sample. However, on the German sample, where positive symptoms were less pronounced, lower verbosity was only associated with negative symptom severity.

With regards to LTR, the results on the German sample are in line with the findings that reported lower TTR in the patient group (Willits et al., 2018; Aich et al., 2022; Minor et al., 2023), but, on the Russian sample, there were no meaningful group differences, in line with the results reported by Hitczenko et al. (2020), Jeong et al. (2023), and Schneider et al. (2023). The results of the present study contradict the findings reported in Ziv et al. (2022), as LTR was

lower, rather than higher, in both samples, despite the gap in the size of this difference. LTR also correlated positively with negative and general scales on both samples and, on the Russian sample, also with positive symptoms, as well as being somewhat predictive of TD and depression severity. The positive correlation could be explained by the inverse proportion to overall verbosity which correlated negatively with symptom severity. This hypothesis is partially confirmed by the fact that the opposite direction of correlation was observed for moving average LTR, for which the effect of overall verbosity is mostly negated. In both samples, simple word count performed better than LTR. The fact that MALTR performed much worse in absolute value than LTR, coupled with the fact that the word count outperforms LTR, also suggests that the most predictive strength of LTR comes from the word count component of this metric.

5.2 Syntactic Methods

The part-of-speech rates obtained in the present study align only partially with the results reported previously.

Even though DET rates were often reported to be lower in the NAP or SDD patients for several languages[1], the present study found no group differences in DET rates in either sample, similarly to the results reported for CHR populations (Bilgrami et al., 2022; Haas et al., 2020). Nevertheless, there were some negative correlations with negative symptom severity on one of the tasks in the Russian sample, contradicting the absence of such association reported by Corcoran et al. (2018) and Bilgrami et al. (2022) and higher article use reported by Mitchell et al. (2015).

The experiments conducted here could find no effects for pronoun use, contradicting partially the results reported by Corcoran et al. (2018) and Jeong et al. (2023). This also does not support the idea that reduced pronoun use

[1] Bedi et al., 2015; Corcoran et al., 2018; Sarzynska-Wawer et al., 2021; Tang et al., 2021

could be used to detect poverty of speech typical of negative FTD, as negative symptoms were prevalent in both samples and yet pronoun rates remained unaltered. Unlike what has been reported (Silva et al., 2023), no effect could be observed for subordinating conjunction (SCONJ) use, but instead, there was a relatively consistent negative correlation with general symptoms for coordinating conjunction use (CCONJ). This could be indicative of lower syntactic complexity and poverty of speech, yet this is not explored in sufficient detail in the present work to make any certain claim. No differences or correlation with symptom severity was observed for ADJ and ADV rates, contradicting the papers claiming both higher (Corcoran et al., 2018; Tang et al., 2021; Ziv et al., 2022) and lower (Argolo et al., 2023) rates in patient populations. This result does not support the idea that the reduction in these POS types could consistently indicate the poverty of speech content often present in negative FTD. Similarly, no effects were found for verb rates, in line with the results reported by Tang et al. (2021), Argolo et al. (2023), and Haas et al. (2020), but in contradiction with lower verb rates reported for a Hebrew-speaking population (Ziv et al., 2022).

Interestingly, AUX rates were associated negatively with positive symptoms on the German sample but not on the Russian one. This is the more surprising as the positive symptoms were less pronounced on the former than on the latter, and this result may have to do with the difference between the languages, rather than the tasks or the populations. The reduction in AUX use could be a marker of lower syntactic complexity, but it is not entirely clear whether it is indicative of it, and if so, why it is associated with positive, rather than negative symptoms. There was a slight negative correlation of NOUN rates with both positive and negative symptoms on one of the tasks for the Russian sample, yet this result could be explained merely by the positive correlation with mean sentence length. Most surprisingly, the best result for POS rates was observed for PART use, which was not reported as a good metric in any of the reviewed studies, yet it correlated positively with all symptom scales to some extent across tasks and languages, as well as being somewhat predictive of TD severity. Once again, however, this could be partially explained by the negative correlation with mean sentence length,

which was present for both samples.

Overall, there was little agreement with the results previously reported for POS rates, and only PART use could be suggested as a metric with a reasonable potential.

The results regarding unit counts and length are mixed. As is commonly reported, decreased sentence length was found in the German patient population[2], though not in the Russian one, for which the result was more similar to the ones reported by Liang et al. (2022), Gupta et al. (2018), and Haas et al. (2020), showing no group differences. Additionally, on the German sample, the lower mean sentence length was associated with negative symptoms, similar to the results reported by Bilgrami et al. (2022), but not with positive symptoms, which were less pronounced, contradicting Liebenthal et al. (2023). On the Russian sample, it was associated both with positive and negative symptoms, but only on two of the four tasks. The reduced mean sentence length could be indirectly indicative of reduced syntactic complexity or poverty of speech typical for negative FTD, yet it was associated with both positive and negative symptoms on the Russian sample.

The maximal sentence length, which was only used as a feature in other studies (Bedi et al., 2015; Tang et al., 2023b), could differentiate between the groups on the German, but not the Russian sample, yet correlated negatively with symptom severity on both, though with negative symptoms only on the German sample. On both samples, maximal sentence length generally performed worse than the mean, though with a few exceptions. The results for standard deviation in sentence length were similar to that for mean and maximum, but still weaker.

As discussed above (4.1), the sentence count contributed more than the mean sentence length to the overall verbosity for two tasks of the Russian sample, but not the other two, and not on the German sample. It is unsurprising that the sentence count only correlated with symptom severity on the Russian sample, not the German one. On the Russian sample, the number of

[2]Iter et al., 2018; Morgan et al., 2021; Spencer et al., 2021; Tang et al., 2021; Bilgrami et al., 2022; Silva et al., 2023; Nettekoven et al., 2023; Schneider et al., 2023; Silva et al., 2023

sentences was the strongest metric, being the only one that correlated with symptom scales on all tasks. It was also rather a strong predictor of TD and depression severity. This negative association with symptom severity is similar to what was reported by Jeong et al. (2023). Following several studies (Gupta et al., 2018; Tang et al., 2021; Schneider et al., 2023), but unlike both Iter et al. (2018), who reported lower sentence count, and Morgan et al. (2021) and Nettekoven et al. (2023), who reported a higher one, no group differences were present in either of the samples in the present study.

5.3 Graph-Based Methods

In line with previously reported results, a lower number of nodes and edges, as well as lower sizes of connected and strongly connected components were observed in the patient population for the German sample, but not the Russian one, where there was no difference between the groups in any of the metrics.

The lower number of nodes is in line with the results reported by Nikzad et al. (2022), but contradicts Mota et al. (2012) and Mota et al. (2014), and like what was reported by Nettekoven et al. (2023), the number of nodes was linked to overall verbosity. The number of nodes was also negatively associated with symptom severity on both samples, more so with general than negative or positive scales, contradicting the reported absence of such a relation (Mota et al., 2012; Mota et al., 2014; Nettekoven et al., 2023). By the mode of graph construction, the number of nodes corresponds to the number of unique lemmas calculated over a moving window, and the simple count of unique words has been reported to be lower in the patient population by some (Willits et al., 2018) but not others (Schneider et al., 2023), and the difference in the present study was also found on the German but not the Russian sample.

The lower number of edges, similarly, was lower only in the German patient sample, in line with several studies (Mota et al., 2014; Mota et al., 2016; Mota et al., 2017; Nikzad et al., 2022), while on the Russian sample the result was

in line with the papers finding no difference in the number of edges (Mota et al., 2012; Nettekoven et al., 2023). In the present study, the lower number of edges consistently correlated with the general symptom scales, and less so with negative and positive ones. This partly is in line with Mota et al. (2014), who report significant anti-correlation with negative and cognitive symptom severity, yet contradicts other studies, where the correlation was absent (Mota et al., 2012; Nettekoven et al., 2023). The number of edges could also serve as a predictor of depression severity on the Russian sample, as well as of TD severity.

Like in most studies reporting the results for largest connected and strongly connected components[3], there was a lower value for both these metrics in the German patient population, though not in the Russian one, similarly to Mota et al. (2012). There was also, on both samples, a significant negative correlation with general and negative symptom severity, as well as with positive symptom severity, and predictive power for TD and depression severity on the Russian sample, similar to the correlations reported in several studies exploring the graph-based metrics (Morgan et al., 2021; Spencer et al., 2021; Nikzad et al., 2022) and in contradiction with the ones finding no such effects (Mota et al., 2012; Argolo et al., 2023; Nettekoven et al., 2023).

In the present study, there was no difference number of parallel edges or loops, unlike what has been reported in the founding papers (Mota et al., 2012; Mota et al., 2014), yet even in these studies, the effects disappeared after controlling for length. Also in contradiction with these studies, there was some correlation with symptom severity on the Russian sample for parallel edges, as well as for loops of size one[4] and three, with L3 positively correlating with most scales, while L1 only with general symptoms both only on one task, and PE being additionally predictive of depression severity.

There was no difference in average node degree or standard deviation therein

[3]Mota et al., 2014; Mota et al., 2016; Mota et al., 2017; Spencer et al., 2021; Morgan et al., 2021; Nikzad et al., 2022

[4]The number of loops of size one corresponds to the number of lemma repetitions calculated over a moving window, so this could be seen as somewhat similar to the studies exploring repetitions or perseveration.

in either sample, which is in line with the founding paper (Mota et al., 2012) but contradicts some follow-up studies (Mota et al., 2014; Nikzad et al., 2022). Unlike the reported absence of correlation with symptom severity (Mota et al., 2012; Mota et al., 2014; Nikzad et al., 2022), there was an anti-correlation with positive and negative symptom severity on one of the tasks on the Russian sample, and average node degree was also predictive of TD severity on this task.

Overall, these results do not offer much support for graph connectivity being indicative of either positive or negative FTD but rather suggest that the well-performing graph-based metrics are strong predictors, sensitive to any symptoms, regardless of symptom type. The results found in this work are stronger than what has been previously reported with respect to correlation with symptom severity, though, in some cases, weaker than what has been reported for the efficacy of graph-based metrics in identifying group differences.

5.4 LM-Based Methods

In this section, I first discuss the relative efficacy of the tested LM-based metrics (5.4.1), followed by the effects of the model selection and sentence averaging procedure (5.4.2), as well as the relation of both models and metrics with mean sentence length, and concluding with metric-model effects and general remarks (5.4.3).

5.4.1 Metrics

None of the LM-based metrics tested in the present work differentiated meaningfully between the groups.

For the most popular LM-based metric, i.e. first order (local) coherence, some correlations with symptom scales were present, though weak, on both samples. The absence of group difference was also reported by several papers in

the field[5], though a few papers reported a difference for some or all tested models[6]. Confirming what had been suggested previously (Ryazanskaya, 2020; Just et al., 2023; Parola et al., 2023), there was correlation with negative and general symptoms measured by PANSS or SANS on both samples, though it depended greatly on task and model, with averaged GloVe or either word2vec variant performing on the German sample and one task of the Russian one, while on another task there was no correlation with PANSS, and yet TF-IDF weighted GloVe local coherence could predict depression severity.

There was also no difference between the groups in second-order coherence, unlike what has been reported by Sarzynska-Wawer et al. (2021) for Polish and by Parola et al. (2023) for Chinese, Danish, and German. However, on the German sample, it did correlate negatively with SANS as well as negative, general, and total PANSS scores when calculated using BERT, either word2vec variant, or averaged GloVe; and TF-IDF weighted GloVe only correlated with SANS and PANSS general. Similar correlations with negative, general, and total PANSS scores were also obtained for either word2vec variant or averaged GloVe second-order coherence on one task in the Russian sample, but not on other tasks or models. BERT second-order coherence only correlated with the general symptom scale for the Russian sample.

As for centroid-based global coherence as well as cumulative global coherence, there are no papers showing its efficacy in differentiating between the groups, but the papers that use show both metrics to correlate with PANSS (for Russian by Ryazanskaya (2020); and for German by Just et al. (2023)) and with TALD scores and human judgement of coherence (Xu et al., 2020; Xu et al., 2022). In the present study, there was indeed no difference between the groups, and the correlation with PANSS, across the scales, was only present for one model on one task in the Russian sample and not at all for the German one. On one task, BERT centroid and cumulative centroid global coherence correlated positively with all PANSS subscales, as well as being predictive

[5]Iter et al., 2018; Just et al., 2020; Hitczenko et al., 2020; Bilgrami et al., 2022; Haas et al., 2020

[6]Iter et al., 2018; Just et al., 2019; Morgan et al., 2021; Ryazanskaya, 2020

of TD severity; while on another task no correlation with PANSS was found, but averaged w2v global coherence could predict depression severity.

Like previously reported (Hitczenko et al., 2020; Tang et al., 2021), there was no difference in next sentence probability between the groups, though, unlike what was previously reported (Tang et al., 2021; Jeong et al., 2023), there was, for the German sample, a slight negative correlation with PANSS negative and total scores, and no correlation with SANS or SAPS, with no effects at all on the Russian sample.

In line with the results reported by Vail et al. (2018), on both samples there was a correlation of pseudo-perplexity with symptom severity, though, surprisingly, the direction of correlation differed between the samples, as it was positive for SANS and general, negative, and total PANSS scores for the German sample, but negative for general PANSS on one for the tasks the Russian sample, where pseudo-perplexity was also predictive of thought disorder severity on one of the tasks. Overly negative results on the Russian sample are partially similar to Girard et al. (2022), who reported no correlation with symptom severity. The absence of group difference in perplexity has been reported previously (Mitchell et al., 2015) and corresponds to what was found for the Russian sample, though for the German sample, the pseudo-perplexity was somewhat higher in the patient population. The higher pseudo-perplexity is unexpected given the predominately negative symptoms in the German sample, which would be expected to render the speech somewhat more stereotyped and less complex. However, since pseudo-perplexity was negatively correlated with mean sentence length, which was lower in the patient sample, this negative length-dependence may underlie slightly higher perplexity in the patient population.

Overall, on the German sample, there was some metric performance hierarchy, with pseudo-perplexity showing the best performance, alongside second-order coherence, being followed by local coherence and next sentence probability, while both metrics of global coherence performed worst. On the Russian sample, however, the metric performance differed very much between tasks and models, with no clear hierarchy. It is very probable, that several

tasks of different types for the German sample would also have differed greatly in metric performance.

On both languages, there was a curious pattern, where cumulative global coherence and global coherence tended to correlate more positively with symptom severity relative to the other two cosine similarity-based metrics, while local and second-order coherence tended to correlate more negatively, independently of the absolute correlation strength or direction.

Finally, the correlation with mean sentence length was overall weakest for cumulative global coherence, but this pattern was weak on the Russian sample. There was also one perplexing difference between the two samples that cannot be explained by the difference between the tasks used, namely, the fact that both pseudo-perplexity and next sentence probability only correlated with mean sentence length on the German sample, not on the Russian one, where they were least strongly correlated with length of all the metrics, and they also differed in the direction of correlation with symptoms between the two samples.

5.4.2 Language Models

There was no influence of the model on the difference between the groups, as it was equally absent for all models, and the direction of differences often differed between the metrics in some cases. On the German sample, metrics computed using BERT performed slightly better than the ones computed with w2v or GloVe, though the opposite pattern could be seen for the positive symptom scales. On the Russian sample, BERT only outperformed other models on one of the tasks, followed by averaged w2v, yet there was no or even the opposite pattern for the other tasks. This tentative pattern is in line with the results reported in Ryazanskaya (2020). There was little difference between TF-IDF and average w2v in terms of performance. Neither schema was very efficient, which is in line with Just et al. (2020) and Hitczenko et al. (2020) but contradicts Just et al. (2019) and Xu et al. (2022). GloVe tended to perform worse than word2vec similar to what had been reported by Iter et al. (2018) and Just et al. (2023), but contradicting Just et al. (2019). GloVe

did not perform well overall, as reported by several studies before, finding no group differences (Just et al., 2020; Alonso-Sánchez et al., 2023; Alonso-Sánchez et al., 2022), though there were some correlations with PANSS, in line with Alonso-Sánchez et al. (2022).

In terms of the correlation with sentence length, the pattern was much more pronounced on the German sample, where BERT cosine similarity-based metrics were clearly least correlated with mean sentence length, followed by TF-IDF weighted GloVe, and averaged GloVe, and then TF-IDF weighted and averaged word2vec, which, as has been reported before, was most strongly correlated with sentence length. A similar pattern could be seen on some tasks in the Russian sample, but it is not very consistent. The positive association with mean sentence length in w2v, as has been suggested, seems to stem from the way in which longer sequences are closer to a meaningless average driving the cosine similarity higher (Hitczenko et al., 2020; Parola et al., 2023; Fradkin et al., 2023). The results of the present experiments suggest that TF-IDF helps mitigate but cannot entirely remove this effect. As mentioned above, the correlation with sentence length for next sentence prediction and pseudo-perplexity was very high for the German sample, but not the Russian one. Rather surprisingly, the direction of correlation with symptoms also differed between the languages for BERT, despite BERT being efficacious with some metrics for both languages. The direction of correlation with symptoms for the metrics calculated with word2vec or GloVe was consistently negative in both samples, while for BERT it was only negative for the German and positive for the Russian sample.

5.4.3 Model-Metric Interaction

In both samples, the degree of model-metric interaction was high, more so in the Russian sample, where the model choice determined the direction of the correlation, while the metrics had their own tentative correlation direction. In both samples, neither metric nor model selection had a decisive effect on the

performance. The size of difference between the cosine similarity-based metrics for one model was comparable between the models on the German sample, and highest for average word2vec on the Russian sample, being lowest for TF-IDF weighted GloVe. The pattern of only some combination of models and metrics working while others fail is very common for studies that test multiple models and metrics[7]. This trend suggests that LM-based metrics are not a particularly robust or consistently performing group of metrics.

5.5 Cross-Methodological Comparison

The hierarchy of the groups of metrics differed slightly between the two samples, as the syntactic metrics performed better on the German sample, being the best group, followed by graph-based and then lexical ones. On the Russian sample, verbosity was the strongest symptom predictor, followed by graph-based and syntactic metrics, and then by other lexical metrics. LM-based methods clearly showed the worst performance for both languages, though pseudo-perplexity could be seen as a promising metric on the German data.

The findings of the present study support the patterns reported in other cross-methodological studies, namely, that syntactic metrics outperform LM-based ones[8], and, to some extent, the graph-based ones (Argolo et al., 2023), though in the results of the present work syntactic and graph-based metrics are quite similar in their performance, as most comparing them report (Mitchell et al., 2015; Just et al., 2020; Jeong et al., 2023). The relative performance of lexical and syntactic methods was not very definitive, being, as reported, more or less comparable (Mitchell et al., 2015; Just et al., 2020; Jeong et al., 2023), with opposite slight trends in the two samples, and the same could be seen in the literature[9]. The results reported here also support

[7] Iter et al., 2018; Just et al., 2019; Ryazanskaya, 2020; Xu et al., 2020; Hitczenko et al., 2020; Xu et al., 2022; Just et al., 2023

[8] Mitchell et al., 2015; Iter et al., 2018; Corcoran et al., 2018; Just et al., 2020; Morgan et al., 2021; Bilgrami et al., 2022; Liebenthal et al., 2023; Argolo et al., 2023

[9] Gupta et al., 2018; Rezaii et al., 2019 report better performance of the lexical metrics, while Schneider et al. (2023) and Argolo et al. (2023) observed the opposite trend.

the claim that lexical metrics outperform LM-based ones[10]. The only partial contradiction with the cross-methodological literature (Argolo et al., 2023) is in the fact that for both samples graph-based metrics clearly outperformed LM-based ones. The results of this study were more similar to what has been reported for the detection of group difference by Morgan et al. (2021).

All in all, the results of the present study are largely in line with cross-methodological literature but suggest that some graph-based metrics may be stronger than is commonly reported in detecting symptom severity.

5.6 Limitations

The present study has several important limitations.

First of all, like in many studies in the field, the sample size for both languages is rather small, with 59 and 31 NAP patients in German and Russian samples, respectively. It is still larger than the mode sample size of 20 patients, yet it is not large enough to reliably detect small effects.

Secondly, there are quite a few differences between the two samples. The tasks used to elicit the texts are different between the Russian and German samples and some substantial part of the difference between the results for the two languages may be attributed to this cause, as the size of the difference is often comparable to that of the difference between the tasks within the Russian sample. Further research is required to separate and compare the size of the difference between tasks in one language and between one task cross-linguistically. Then, the target scales used to assess symptom severity only overlap partially, which renders impossible a cross-linguistic validation of some of the results. Additionally, there is a difference in symptom severity between the samples, the positive symptoms being significantly less prominent in the German sample. Because of this low level of positive symptoms, little can be said about the cross-linguistic reliability of positive symptom detection, as they could, for the most part, only be identified on the Russian

[10]Mitchell et al., 2015; Just et al., 2019; Just et al., 2020; Hitczenko et al., 2020; Aich et al., 2022; Girard et al., 2022

sample. More research on dedicated groups showing negative, positive, and mixed symptoms could shed more light on this issue. There was also a difference in the gender balance, with the Russian sample skewing towards females, while the German sample was more balanced. This, however, should not significantly influence the results presented here, as there was no difference between the sexes in any of the metrics, or target and control characteristics. The German sample was significantly older than the Russian one, but the metrics in the two samples did not seem to be much correlated with age, and between the groups in both samples the age was comparable.

Thirdly, there are limitations to the metrics selection and analysis used in the study. On the one hand, the selection of metrics was quite limited, especially for the lexical metrics, making the conclusions regarding this group less reliable. On the other hand, due to the very high number of comparisons, statistical significance would not have been a useful indicator of performance. Nevertheless, the bootstrapping procedure used here is not a perfect solution, and the reported results rely on a somewhat arbitrary correlation cutoff.

Overall, however, these limitations are not likely to render the results presented in the study generally unreliable.

5.7 Conclusion

The present study is a cross-linguistic benchmarking of many common NLP methods used for psychosis detection.

In this work, several groups of methods were compared, and the general patterns in their relative performance were mostly in accordance with the cross-methodological research in the field. The results obtained here indicate that co-occurrence graph-based metrics are the most reliably correlated with symptom severity and are also most in line with the previous results, performing in some cases even better than what was reported before. The lexical and syntactic methods tested in this work performed somewhat worse and were not as well reproducible neither between the two samples nor with respect to the previous literature, the results of which have also been mixed.

Finally, the LM-based metrics barely performed at all, and they were also clearly the least reliable and least reproducible group of metrics in the field.

Several graph-based metrics, namely, the number of nodes and edges, as well as the sizes of the largest connected and strongly connected component, despite being correlated with mean sentence length, outperformed this baseline, as well as word count, for both languages and showed consistent results across the tasks. Similarly, PART rate correlated positively with the psychiatric scales for both samples, outperforming sentence length for both languages, as well as outperforming word count on the German but not the Russian sample. The lemma-token ratio and moving average lemma-token ratio were much weaker, sometimes performing worse than the baseline, especially on the German sample, where it was stronger. On the German sample mean sentence length was a stronger baseline than word count or sentence count, while the opposite was true on the Russian sample.

The present work demonstrated a large difference in relative metric performance on different elicitation tasks, even of the same task type, as well as between the languages. Generally, the size of this difference was comparable for different tasks within one language and between different tasks for different languages. Although for some metrics, such as POS rates, the cross-linguistic differences in performance could be anticipated, in the present study, these differences were still comparable to cross-task ones. On the contrary, for LM-based metrics, there were clear differences in the performance that could only be attributed to the differences either between the languages or between the tools available for them, and not to the difference between the tasks. For instance, the direction of correlation for the BERT-based metrics differed between the languages, but not between the tasks within one language, with BERT pseudo-perplexity proving a very strong metric for the German, but not the Russian sample. Moreover, the patterns of length-dependence of the LM-based metrics also differed between the two languages. All in all, both cross-model and cross-linguistic differences may contribute to the lack of cross-linguistic replicability for LM-based metrics, and the results of the present study suggest that cross-model differences play a more important role.

The exploration of the length-dependence of the metrics suggests that many metrics do depend significantly on sentence length and at least a part of their explanatory power comes from this dependence. Consequently, further research in the field should test sentence length and the number of sentences as baseline metrics, control other metrics for them, and account for them in explanatory models.

The present study found no definitive tendencies in any metric to be better suited for the identification of positive rather than negative symptoms or vice versa. Instead, well-performing metrics tended to perform across all symptom scales. Further, more targeted research on separate positive and negative symptom populations is required to explore any such patterns.

To conclude, the benchmarking conducted in the present study indicates that the simplest metrics, such as mean sentence length, word count, and unique lemma count, provide a strong baseline that other, more complex metrics can rarely beat.

Bibliography

Aich, Ankit et al. (Dec. 2022). "Towards Intelligent Clinically-Informed Language Analyses of People with Bipolar Disorder and Schizophrenia". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 2871–2887. DOI: 10.18653/v1/2022.findings-emnlp.208. URL: https://aclanthology.org/2022.findings-emnlp.208.

Alonso-Sánchez, Maria Francisca et al. (Apr. 2022). "Progressive changes in descriptive discourse in First Episode Schizophrenia: a longitudinal computational semantics study". In: *Schizophrenia* 8.1. ISSN: 2754-6993. DOI: 10.1038/s41537-022-00246-8.

Alonso-Sánchez, María Francisca et al. (Sept. 2023). "Language network self-inhibition and semantic similarity in first-episode schizophrenia: A computational-linguistic and effective connectivity approach". In: *Schizophrenia Research* 259, 97–103. ISSN: 0920-9964. DOI: 10.1016/j.schres.2022.04.007.

Andrade, Chittaranjan (Sept. 2018). "Internal, External, and Ecological Validity in Research Design, Conduct, and Evaluation". In: *Indian Journal of Psychological Medicine* 40.5, 498–499. ISSN: 0975-1564. DOI: 10.4103/ijpsym.ijpsym_334_18.

Andreasen, N. C. (Jan. 1986a). "Scale for the Assessment of Thought, Language, and Communication (TLC)". In: *Schizophrenia Bulletin* 12.3, 473–482. ISSN: 1745-1701. DOI: 10.1093/schbul/12.3.473.

Andreasen, Nancy C. (1986b). *Scale for the Assessment of Positive Symptoms*. DOI: 10.1037/t48377-000.

Andreasen, Nancy C. (1989). "The Scale for the Assessment of Negative Symptoms (SANS): Conceptual and Theoretical Foundations". In: *British Journal of Psychiatry* 155.S7, 49–52. DOI: 10.1192/S0007125000291496.

Arciniegas, David B. (June 2015). "Psychosis". In: *CONTINUUM: Lifelong Learning in Neurology* 21, 715–736. ISSN: 1080-2371. DOI: 10.1212/01.con.0000466662.89908.e7.

Argolo, Felipe C et al. (Mar. 2023). *Burnishing the blueprint of speech assessment with natural language processing: methods to characterize subtle impairments on individuals in at-risk mental states from a large urban population.* DOI: 10.31234/osf.io/epgfy. URL: osf.io/preprints/psyarxiv/epgfy.

Arora, Sanjeev, Yingyu Liang, and Tengyu Ma (2017). "A Simple but Tough-to-Beat Baseline for Sentence Embeddings". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. URL: https://oar.princeton.edu/handle/88435/pr1rk2k.

Association, American Psychiatric (May 2013). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association. ISBN: 0890425574. DOI: 10.1176/appi.books.9780890425596.

Bar, Kfir et al. (2019). *Semantic Characteristics of Schizophrenic Speech*. DOI: 10.48550/arXiv.1904.07953. arXiv: 1904.07953 [cs.CL].

Bedi, Gillinder et al. (Aug. 2015). "Automated analysis of free speech predicts psychosis onset in high-risk youths". In: *npj Schizophrenia* 1.1. ISSN: 2334-265X. DOI: 10.1038/npjschz.2015.30.

Bilgrami, Zarina R. et al. (2022). "Construct validity for computational linguistic metrics in individuals at clinical risk for psychosis: Associations with clinical ratings". In: *Schizophrenia Research* 245. Computational Approaches to Understanding Psychosis, pp. 90–96. ISSN: 0920-9964. DOI: 10.1016/j.schres.2022.01.019. URL: https://www.sciencedirect.com/science/article/pii/S0920996422000299.

Bleuler, Eugen (1911). *Dementia praecox, oder Gruppe der Schizophrenien*. German. Vol. 4. Handbuch der Psychiatrie, Spezieller; T., 4. Abt., 1. Hälfte. Leipzig: Deuticke.

Boer, J. N. de et al. (Aug. 2023). "Acoustic speech markers for schizophrenia-spectrum disorders: a diagnostic and symptom-recognition tool". In: *Psychological Medicine* 53.4, 1302–1312. ISSN: 1469-8978. DOI: 10.1017/s0033291721002804.

Bojanowski, Piotr et al. (2017). "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5. Ed. by Lillian Lee, Mark Johnson, and Kristina Toutanova, pp. 135–146. DOI: 10.1162/tacl_a_00051. URL: https://aclanthology.org/Q17-1010.

Buck, Benjamin et al. (July 2014). "The use of narrative sampling in the assessment of social cognition: The Narrative of Emotions Task (NET)". In: *Psychiatry Research* 217.3, 233–239. ISSN: 0165-1781. DOI: 10.1016/j.psychres.2014.03.014.

Caplan, Rochelle et al. (May 1989). "The Kiddie Formal Thought Disorder Rating Scale: Clinical Assessment, Reliability, and Validity". In: *Journal of the American Academy of Child and Adolescent Psychiatry* 28.3, 408–416. ISSN: 0890-8567. DOI: 10.1097/00004583-198905000-00018.

Cohen, Alex S. et al. (Apr. 2017). "Can RDoC Help Find Order in Thought Disorder?" In: *Schizophrenia Bulletin* 43.3, 503–508. ISSN: 1745-1701. DOI: 10.1093/schbul/sbx030.

Colla, Davide et al. (Dec. 2022). "Semantic coherence markers: The contribution of perplexity metrics". In: *Artificial Intelligence in Medicine* 134, p. 102393. ISSN: 0933-3657. DOI: 10.1016/j.artmed.2022.102393.

"Computational linguistic analysis applied to a semantic fluency task: A replication among first-episode psychosis patients with and without derailment and tangentiality" (2021). In: *Psychiatry Research* 304, p. 114105. ISSN: 0165-1781. DOI: 10.1016/j.psychres.2021.114105. URL: https://www.sciencedirect.com/science/article/pii/S0165178121004029.

Corcoran, Cheryl M. et al. (2018). "Prediction of psychosis across protocols and risk cohorts using automated language analysis". In: *World Psychiatry* 17.1, pp. 67–75. DOI: 10.1002/wps.20491.

Corona-Hernández, H. et al. (2023). "Assessing coherence through linguistic connectives: Analysis of speech in patients with schizophrenia-spectrum disorders". In: *Schizophrenia Research* 259. Language and Speech Analysis in Schizophrenia and Related Psychoses, pp. 48–58. ISSN: 0920-9964. DOI:

10.1016/j.schres.2022.06.013. URL: https://www.sciencedirect.com/science/article/pii/S0920996422002481.

Crossley, Scott A., Kristopher Kyle, and Mihai Dascalu (Oct. 2018). "The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap". In: *Behavior Research Methods* 51.1, 14–27. ISSN: 1554-3528. DOI: 10.3758/s13428-018-1142-4.

de Boer, Janna N. et al. (May 2020). "Anomalies in language as a biomarker for schizophrenia". In: *Current Opinion in Psychiatry* 33.3, 212–218. ISSN: 1473-6578. DOI: 10.1097/yco.0000000000000595.

Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: DOI: 10.48550/arXiv.1810.04805. arXiv: 1810.04805 [cs.CL].

Ditman, Tali and Gina R. Kuperberg (May 2010). "Building coherence: A framework for exploring the breakdown of links across clause boundaries in schizophrenia". In: *Journal of Neurolinguistics* 23.3, 254–269. ISSN: 0911-6044. DOI: 10.1016/j.jneuroling.2009.03.003.

Doré, MN (2019). "Quantification of Coherence in Spoken Language asan Indicator for the Schizophrenia Spectrum Disorder". Bachelor's thesis. Utrecht University.

Elvevåg, Brita et al. (2007). "Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia". In: *Schizophrenia Research* 93.1, pp. 304–316. ISSN: 0920-9964. DOI: 10.1016/j.schres.2007.03.001. URL: https://www.sciencedirect.com/science/article/pii/S092099640700117X.

Elvevåg, Brita et al. (2010). "An automated method to analyze language use in patients with schizophrenia and their first-degree relatives". In: *Journal of Neurolinguistics* 23.3. Language, Communication and Schizophrenia, pp. 270–284. ISSN: 0911-6044. DOI: 10.1016/j.jneuroling.2009.05.002. URL: https://www.sciencedirect.com/science/article/pii/S0911604409000402.

Fradkin, Isaac, Matthew M. Nour, and Raymond J. Dolan (2023). "Theory-Driven Analysis of Natural Language Processing Measures of Thought Disorder Using Generative Language Modeling". In: *Biological Psychiatry:*

Cognitive Neuroscience and Neuroimaging 8.10. Natural Language Processing in Psychiatry and Clinical Neuroscience Research, pp. 1013–1023. ISSN: 2451-9022. DOI: 10.1016/j.bpsc.2023.05.005. URL: https://www.sciencedirect.com/science/article/pii/S2451902223001258.

Gildea, Daniel and Daniel Jurafsky (2002). "Automatic Labeling of Semantic Roles". In: *Computational Linguistics* 28.3, pp. 245–288. DOI: 10.1162/089120102760275983. URL: https://aclanthology.org/J02-3001.

Girard, Jeffrey M. et al. (2022). "Computational analysis of spoken language in acute psychosis and mania". In: *Schizophrenia Research* 245. Computational Approaches to Understanding Psychosis, pp. 97–115. ISSN: 0920-9964. DOI: 10.1016/j.schres.2021.06.040. URL: https://www.sciencedirect.com/science/article/pii/S0920996421002528.

Glosser, Guila and Toni Deser (1991). "Patterns of discourse production among neurological patients with fluent language disorders". In: *Brain and Language* 40.1, pp. 67–88. ISSN: 0093-934X. DOI: 10.1016/0093-934X(91)90117-J. URL: https://www.sciencedirect.com/science/article/pii/0093934X9190117J.

Gupta, Tina et al. (2018). "Automated analysis of written narratives reveals abnormalities in referential cohesion in youth at ultra high risk for psychosis". In: *Schizophrenia Research* 192, pp. 82–88. ISSN: 0920-9964. DOI: 10.1016/j.schres.2017.04.025. URL: https://www.sciencedirect.com/science/article/pii/S0920996417302128.

Haas, S. S. et al. (2020). "Linking language features to clinical symptoms and multimodal imaging in individuals at clinical high risk for psychosis". In: *European Psychiatry* 63.1, e72. DOI: 10.1192/j.eurpsy.2020.73.

Hart, Mara and Richard R. J. Lewine (Jan. 2017). "Rethinking Thought Disorder". In: *Schizophrenia Bulletin* 43.3, pp. 514–522. ISSN: 0586-7614. DOI: 10.1093/schbul/sbx003.

Hitczenko, Kasia, Vijay A Mittal, and Matthew Goldrick (Nov. 2020). "Understanding Language Abnormalities and Associated Clinical Markers in Psychosis: The Promise of Computational Methods". In: *Schizophrenia Bulletin* 47.2, pp. 344–362. ISSN: 0586-7614. DOI: 10.1093/schbul/sbaa141.

Holmlund, Terje B. et al. (2019). "Updating verbal fluency analysis for the 21st century: Applications for psychiatry". In: *Psychiatry Research* 273, pp. 767–769. ISSN: 0165-1781. DOI: 10.1016/j.psychres.2019.02.014. URL: https://www.sciencedirect.com/science/article/pii/S0165178118324181.

Holmlund, Terje B. et al. (2023). "Towards a temporospatial framework for measurements of disorganization in speech using semantic vectors". In: *Schizophrenia Research* 259. Language and Speech Analysis in Schizophrenia and Related Psychoses, pp. 71–79. ISSN: 0920-9964. DOI: 10.1016/j.schres.2022.09.020. URL: https://www.sciencedirect.com/science/article/pii/S0920996422003620.

Iter, Dan, Jong Yoon, and Dan Jurafsky (June 2018). "Automatic Detection of Incoherent Speech for Diagnosing Schizophrenia". In: *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. Ed. by Kate Loveys et al. New Orleans, LA: Association for Computational Linguistics, pp. 136–146. DOI: 10.18653/v1/W18-0615. URL: https://aclanthology.org/W18-0615.

Jeong, Lydia et al. (2023). "Exploring the Use of Natural Language Processing for Objective Assessment of Disorganized Speech in Schizophrenia". In: *Psychiatric Research and Clinical Practice* 5.3, pp. 84–92. DOI: 10.1176/appi.prcp.20230003.

Ji, Shaoxiong et al. (2021). *MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare*. DOI: 10.48550/arXiv.2110.15621. arXiv: 2110.15621 [cs.CL].

Just, Sandra et al. (June 2019). "Coherence models in schizophrenia". In: *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. Ed. by Kate Niederhoffer et al. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 126–136. DOI: 10.18653/v1/W19-3015. URL: https://aclanthology.org/W19-3015.

Just, Sandra A et al. (2020). "Modeling incoherent discourse in non-affective psychosis". In: *Frontiers in Psychiatry* 11, p. 846. ISSN: 1664-0640. DOI: 10.3389/fpsyt.2020.00846. URL: https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2020.00846.

Just, Sandra Anna et al. (2023). "Validation of natural language processing methods capturing semantic incoherence in the speech of patients with non-affective psychosis". In: *Frontiers in Psychiatry* 14, p. 1208856. ISSN: 1664-0640. DOI: 10.3389/fpsyt.2023.1208856. URL: https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2023.1208856.

Kaplan, Jared et al. (2020). *Scaling Laws for Neural Language Models*. arXiv: 2001.08361 [cs.LG].

Kay, Stanley R., Abraham Fiszbein, and Lewis A. Opler (Jan. 1987). "The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia". In: *Schizophrenia Bulletin* 13.2, pp. 261–276. ISSN: 0586-7614. DOI: 10.1093/schbul/13.2.261.

Kircher, Tilo et al. (2014). "A rating scale for the assessment of objective and subjective formal Thought and Language Disorder (TALD)". In: *Schizophrenia Research* 160.1, pp. 216–221. ISSN: 0920-9964. DOI: 10.1016/j.schres.2014.10.024. URL: https://www.sciencedirect.com/science/article/pii/S0920996414005933.

Kořánová, Nora (2017). "Analyzing coherence in spontaneous speech of schizophrenic patients". Master's Thesis. University of Potsdam.

Kraepelin, Emil, George M. Robertson, and R. Mary Barclay (1919). *Dementia praecox and paraphrenia*. English. Chicago: Chicago Medical Book Co.

Kramov, Artem (2020). "Evaluating text coherence based on the graph of the consistency of phrases to identify symptoms of schizophrenia". In: *CoRR* abs/2005.03008. DOI: 10.48550/arXiv.2005.03008.

Kuperberg, Gina R. (Aug. 2010a). "Language in Schizophrenia Part 1: An Introduction". In: *Language and Linguistics Compass* 4.8, 576–589. ISSN: 1749-818X. DOI: 10.1111/j.1749-818x.2010.00216.x.

– (Aug. 2010b). "Language in Schizophrenia Part 2: What Can Psycholinguistics Bring to the Study of Schizophrenia...and Vice Versa?" In: *Language and Linguistics Compass* 4.8, 590–604. ISSN: 1749-818X. DOI: 10.1111/j.1749-818x.2010.00217.x.

Kyle, Kristopher (2016). "Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of

syntactic sophistication". Dissertation. Georgia State University. DOI: 10.57709/8501051.

Landauer, Thomas K, Peter W. Foltz, and Darrell Laham (1998). "An introduction to latent semantic analysis". In: *Discourse Processes* 25.2-3, pp. 259–284. DOI: 10.1080/01638539809545028.

Lee, Kenton et al. (Sept. 2017). "End-to-end Neural Coreference Resolution". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, pp. 188–197. DOI: 10.18653/v1/D17-1018. URL: https://aclanthology.org/D17-1018.

Liang, Liangbing et al. (2022). "Widespread cortical thinning, excessive glutamate and impaired linguistic functioning in schizophrenia: A cluster analytic approach". In: *Frontiers in Human Neuroscience* 16, p. 954898. ISSN: 1662-5161. DOI: 10.3389/fnhum.2022.954898. URL: https://www.frontiersin.org/articles/10.3389/fnhum.2022.954898.

Liddle, Peter F et al. (2002). "Thought and Language Index: an instrument for assessing thought and language in schizophrenia". In: *The British Journal of Psychiatry* 181.4, pp. 326–330. DOI: 10.1192/bjp.181.4.326.

Liebenthal, Einat et al. (2023). "Linguistic and non-linguistic markers of disorganization in psychotic illness". In: *Schizophrenia Research* 259. Language and Speech Analysis in Schizophrenia and Related Psychoses, pp. 111–120. ISSN: 0920-9964. DOI: 10.1016/j.schres.2022.12.003. URL: https://www.sciencedirect.com/science/article/pii/S0920996422004509.

Liu, Yinhan et al. (2019). "Roberta: A robustly optimized bert pretraining approach". In: *CoRR* abs/1907.11692. DOI: 10.48550/arXiv.1907.11692.

McNamara, Danielle S. et al. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

Mikolov, Tomas et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges et al. Vol. 26. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce
Paper.pdf.

Minor, Kyle S et al. (2023). "Automated measures of speech content and speech organization in schizophrenia: Test-retest reliability and generalizability across demographic variables". In: *Psychiatry Research* 320, p. 115048. ISSN: 0165-1781. DOI: 10.1016/j.psychres.2023.115048. URL: https://www.sciencedirect.com/science/article/pii/S016517812300001X.

Mitchell, Margaret, Kristy Hollingshead, and Glen Coppersmith (May 2015). "Quantifying the Language of Schizophrenia in Social Media". In: *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Denver, Colorado: Association for Computational Linguistics, pp. 11–20. DOI: 10.3115/v1/W15-1202. URL: https://aclanthology.org/W15-1202.

Moghadasi, Mahdi Naser and Yu Zhuang (2020). "Sent2Vec: A New Sentence Embedding Representation With Sentimental Semantic". In: *2020 IEEE International Conference on Big Data (Big Data)*, pp. 4672–4680. DOI: 10.1109/BigData50022.2020.9378337.

Morgan, Sarah E et al. (Dec. 2021). "Natural Language Processing markers in first episode psychosis and people at clinical high-risk". In: *Translational Psychiatry* 11.1. ISSN: 2158-3188. DOI: 10.1038/s41398-021-01722-y.

Mota, Natália B, Mauro Copelli, and Sidarta Ribeiro (2016). *Quantifying word salad: The structural randomness of verbal reports predicts negative symptoms and Schizophrenia diagnosis 6 months later*. DOI: 10.48550/arXiv.1610.08566. arXiv: 1610.08566 [q-bio.NC].

– (Apr. 2017). "Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance". In: *npj Schizophrenia* 3.1. ISSN: 2334-265X. DOI: 10.1038/s41537-017-0019-3.

Mota, Natalia B et al. (Apr. 2012). "Speech graphs provide a quantitative measure of thought disorder in psychosis". In: *PloS one* 7.4, pp. 1–9. DOI: 10.1371/journal.pone.0034928.

Mota, Natália B et al. (Jan. 2014). "Graph analysis of dream reports is especially informative about psychosis". In: *Scientific Reports* 4.1. ISSN: 2045-2322. DOI: 10.1038/srep03691.

Mota, Natália Bezerra et al. (2023). "Happy thoughts: What computational assessment of connectedness and emotional words can inform about early stages of psychosis". In: *Schizophrenia Research* 259. Language and Speech Analysis in Schizophrenia and Related Psychoses, pp. 38–47. ISSN: 0920-9964. DOI: 10.1016/j.schres.2022.06.025. URL: https://www.sciencedirect.com/science/article/pii/S0920996422002602.

Nettekoven, Caroline R et al. (Mar. 2023). "Semantic speech networks linked to formal thought disorder in early psychosis". In: *Schizophrenia Bulletin* 49.Supplement$_2$, S142–S152. ISSN: 1745-1701. DOI: 10.1093/schbul/sbac056.

Nikzad, Amir H et al. (July 2022). "Who does what to whom? graph representations of action-predication in speech relate to psychopathological dimensions of psychosis". In: *Schizophrenia* 8.1. ISSN: 2754-6993. DOI: 10.1038/s41537-022-00263-7.

OpenAI et al. (2024). *GPT-4 Technical Report*. DOI: 10.48550/arXiv.2303.08774. arXiv: 2303.08774 [cs.CL].

Organization, World Health (1992). *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*. Vol. 1. World Health Organization.

Palaniyappan, Lena (Aug. 2021). "More than a biomarker: could language be a biosocial marker of psychosis?" In: *npj Schizophrenia* 7.1. ISSN: 2334-265X. DOI: 10.1038/s41537-021-00172-1.

Palominos, Claudio, Alicia Figueroa-Barra, and Wolfram Hinzen (Mar. 2023). "Coreference Delays in Psychotic Discourse: Widening the Temporal Window". In: *Schizophrenia Bulletin* 49.Supplement$_2$, S153–S162. ISSN: 0586-7614. DOI: 10.1093/schbul/sbac102.

Panicheva, Polina and Tatiana Litvinova (2019). "Semantic coherence in schizophrenia in Russian written texts". In: *2019 25th Conference of Open Innovations Association (FRUCT)*. IEEE, pp. 241–249. URL: https://fruct.org/publications/volume-25/fruct25/files/Pan.pdf.

– (2020). "A corpus study of semantic coherence in schizophrenia in Russian written texts". In: *The Night Whites Language Workshop: The Fifth Saint Petersburg Winter Workshop on Experimental Studies of Speech and Language*

(Night Whites 2019), pp. 81–81. URL: https://nightwhites2019.wordpress.com/wp-content/uploads/2019/12/nw_abstracts_alph.order_-1.pdf.

Parola, Alberto et al. (2023). "Speech disturbances in schizophrenia: Assessing cross-linguistic generalizability of NLP automated measures of coherence". In: *Schizophrenia Research* 259. Language and Speech Analysis in Schizophrenia and Related Psychoses, pp. 59–70. ISSN: 0920-9964. DOI: 10.1016/j.schres.2022.07.002. URL: https://www.sciencedirect.com/science/article/pii/S0920996422002742.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (Oct. 2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162.

Peters, Matthew E. et al. (2017). *Semi-supervised sequence tagging with bidirectional language models*. DOI: 10.48550/arXiv.1705.00108. arXiv: 1705.00108 [cs.CL].

Pietrowicz, Mary et al. (2019). "A New Approach for Automating Analysis of Responses on Verbal Fluency Tests from Subjects At-Risk for Schizophrenia." English (US). In: vol. 2019-September. Publisher Copyright: Copyright © 2019 ISCA; 20th Annual Conference of the International Speech Communication Association: Crossroads of Speech and Language, INTERSPEECH 2019 ; Conference date: 15-09-2019 Through 19-09-2019, pp. 3028–3032. DOI: 10.21437/Interspeech.2019-2987. URL: https://www.isca-archive.org/interspeech_2019/pietrowicz19_interspeech.pdf.

Reimers, Nils and Iryna Gurevych (Nov. 2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, pp. 3982–3992. DOI: 10.18653/v1/D19-1410. URL: https://aclanthology.org/D19-1410.

Rezaii, Neguine, Elaine Walker, and Phillip Wolff (June 2019). "A machine learning approach to predicting psychosis using semantic density and latent content analysis". In: *npj Schizophrenia* 5.1. ISSN: 2334-265X. DOI: 10.1038/s41537-019-0077-9.

Rosenstein, Mark et al. (2015). "Language as a biomarker in those at high-risk for psychosis." In: 165.2, pp. 249–250. ISSN: 0920-9964. DOI: 10.1016/j.schres.2015.04.023. URL: https://www.sciencedirect.com/science/article/pii/S0920996415002182.

Ryazanskaya, Galina (2020). "Automated Assessment of Discourse Coherence in Schizophrenia and Schizoaffective Disorder". Bachelor's Thesis. Higher School of Economics.

Ryazanskaya, Galina and Mariya Khudyakova (2020). "Automated Analysis of Discourse Coherence in Schizophrenia: Approximation of Manual Measures". In: *LREC 2020 Language Resources and Evaluation Conference*, pp. 98–101. URL: https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/RaPID3book.pdf.

Sarzynska-Wawer, Justyna et al. (2021). "Detecting formal thought disorder by deep contextualized word representations". In: *Psychiatry Research* 304, p. 114135. ISSN: 0165-1781. DOI: 10.1016/j.psychres.2021.114135. URL: https://www.sciencedirect.com/science/article/pii/S0165178121004315.

Schmidt, Karl-Heinz and Peter Metzler (1992). *Wortschatztest : WST*. German. Weinheim: Beltz.

Schneider, Katharina et al. (May 2023). "Syntactic complexity and diversity of spontaneous speech production in schizophrenia spectrum and major depressive disorders". In: *Schizophrenia* 9.1. ISSN: 2754-6993. DOI: 10.1038/s41537-023-00359-8.

Shriki, Yaara et al. (July 2022). "Masking Morphosyntactic Categories to Evaluate Salience for Schizophrenia Diagnosis". In: *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*. Ed. by Ayah Zirikly et al. Seattle, USA: Association for Computational Linguistics, pp. 148–157. DOI: 10.18653/v1/2022.clpsych-1.13. URL: https://aclanthology.org/2022.clpsych-1.13.

Silva, Angelica M et al. (2023). "Syntactic complexity of spoken language in the diagnosis of schizophrenia: A probabilistic Bayes network model". In: *Schizophrenia Research* 259. Language and Speech Analysis in Schizophrenia and Related Psychoses, pp. 88–96. ISSN: 0920-9964. DOI: 10.1016/j.schres.2022.06.011. URL: https://www.sciencedirect.com/science/article/pii/S0920996422002456.

Spencer, Tom John et al. (2021). "Lower speech connectedness linked to incidence of psychosis in people at clinical high risk". In: *Schizophrenia Research* 228, pp. 493–501. ISSN: 0920-9964. DOI: 10.1016/j.schres.2020.09.002. URL: https://www.sciencedirect.com/science/article/pii/S0920996420304588.

Srivastava, Agrima et al. (May 2022a). "Increased Metaphor Production in Open-Ended Speech Samples of Patients With Prodromal and Developed Schizophrenia Detected with NLP". In: *Biological Psychiatry* 91.9, S50. ISSN: 0006-3223. DOI: 10.1016/j.biopsych.2022.02.145.

Srivastava, Agrima et al. (May 2022b). "P473. Estimating Self-Disturbance in Psychosis and Its Risk States Using Natural Language Processing Analysis of Open-Ended Interviews". In: *Biological Psychiatry* 91.9, S280. ISSN: 0006-3223. DOI: 10.1016/j.biopsych.2022.02.709.

Tang, Sunny X. et al. (May 2021). "Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders". In: *npj Schizophrenia* 7.1. ISSN: 2334-265X. DOI: 10.1038/s41537-021-00154-3.

Tang, Sunny X et al. (2023a). "Clinical and computational speech measures are associated with social cognition in schizophrenia spectrum disorders". In: *Schizophrenia Research* 259. Language and Speech Analysis in Schizophrenia and Related Psychoses, pp. 28–37. ISSN: 0920-9964. DOI: 10.1016/j.schres.2022.06.012. URL: https://www.sciencedirect.com/science/article/pii/S092099642200247X.

Tang, Sunny X et al. (Mar. 2023b). "Latent factors of language disturbance and relationships to quantitative speech features". In: *Schizophrenia Bulletin* 49.Supplement$_2$, S93–S103. ISSN: 0586-7614. DOI: 10.1093/schbul/sbac145.

Tausczik, Yla R and James W Pennebaker (2010). "The psychological meaning of words: LIWC and computerized text analysis methods". In: *Journal of Language and Social Psychology* 29.1, pp. 24–54. DOI: 10.1177/0261927X09351676.

Vail, Alexandria K. et al. (2018). "Toward Objective, Multifaceted Characterization of Psychotic Disorders: Lexical, Structural, and Disfluency Markers of Spoken Language". In: ICMI '18. New York, NY, USA: Association for Computing Machinery, 170–178. ISBN: 9781450356923. DOI: 10.1145/3242969.3243020.

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Voppel, Alban E et al. (2021). "Quantified language connectedness in schizophrenia-spectrum disorders". In: *Psychiatry Research* 304, p. 114130. ISSN: 0165-1781. DOI: 10.1016/j.psychres.2021.114130. URL: https://www.sciencedirect.com/science/article/pii/S0165178121004261.

Voppel, Alban E et al. (2023). "Semantic and Acoustic Markers in Schizophrenia-Spectrum Disorders: A Combinatory Machine Learning Approach". In: *Schizophrenia Bulletin* 49.Supplement_2, S163–S171.

Wang, Alex and Kyunghyun Cho (2019). *BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model*. DOI: 10.48550/arXiv.1902.04094. arXiv: 1902.04094 [cs.CL].

Willits, Jon A. et al. (2018). "Evidence of disturbances of deep levels of semantic cohesion within personal narratives in schizophrenia". In: *Schizophrenia Research* 197, pp. 365–369. ISSN: 0920-9964. URL: https://www.sciencedirect.com/science/article/pii/S0920996417307107.

Wouts, Joppe et al. (2021). *belabBERT: a Dutch RoBERTa-based language model applied to psychiatric classification*. DOI: 10.48550/arXiv.2106.01091. arXiv: 2106.01091 [cs.CL].

Xu, Weizhe et al. (2020). "The centroid cannot hold: comparing sequential and global estimates of coherence as indicators of formal thought disorder". In: *AMIA Annual Symposium Proceedings*. Vol. 2020. American Medical Informatics Association, p. 1315. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8075468/.

Xu, Weizhe et al. (2022). "Fully automated detection of formal thought disorder with Time-series Augmented Representations for Detection of Incoherent Speech (TARDIS)". In: *Journal of Biomedical Informatics* 126, p. 103998. ISSN: 1532-0464. DOI: 10.1016/j.jbi.2022.103998. URL: https://www.sciencedirect.com/science/article/pii/S1532046422000144.

Ziv, Ido et al. (2022). "Morphological characteristics of spoken language in schizophrenia patients – an exploratory study". In: *Scandinavian Journal of Psychology* 63.2, pp. 91–99. DOI: 10.1111/sjop.12790.

Appendix A

Sample Characteristics

Paper	Diagnosis	CHR	HC	FEP/SZ	Language	TLC/TLI	Symptoms	Other Scales	Task Type
Nettekoven et al., 2023	CHR	24	13	16	English	TLI	PANSS		picture description task, story retelling
Bilgrami et al., 2022	CHR	60	27		English	TLC	SIPS/SOPS		open-ended interview
Srivastava et al., 2022a	CHR	172	126	65	English		SIPS		open-ended interview
Srivastava et al., 2022b	CHR	172	126	65	English		SIPS		open-ended interview
Hitczenko et al., 2020	CHR	36	41		English		SIPS	WRAT	semi-structured interview
Morgan et al., 2021	CHR	25	13	15	English	TLI	PANSS	WRAT IQ, WAIS Digit Span Test	picture description task, story retelling, open-ended interview
Spencer et al., 2021	CHR	24	13	16	English	TLI	PANSS	WRAT IQ, WAIS Digit Span Test	picture description task, story retelling, open-ended interview
Haas et al., 2020	CHR	46	22		English		SIPS/SOPS	GFS	open-ended interview
Rezaii et al., 2019	CHR	40		23	English		SIPS		semi-structured interview
Corcoran et al., 2018	CHR	59	21		English	K-FTDS	SIPS/SOPS		semi-structured interview
Gupta et al., 2018	CHR	41	43		English			WRAT IQ, HVLT-R, LNS, CPT-IP, GFS-R	written picture description
Bedi et al., 2015	CHR	34	34		English		SIPS/SOPS		open-ended interview
Rosenstein et al., 2015	CHR	30	25		English				picture description task
Argolo et al., 2023	ARMS	60	73		Portuguese		SIPS		semi-structured interview, open prompt, picture description task
Palominos et al., 2023	CHR	20	20	20	Spanish		SIPS/SOPS PANSS	GAF	semi-structured interview

TABLE A.1: sample characteristics for studies on clinical high risk populations. CHR - clinical high risk; ARMS - At Risk Mental States; FEP - first episode psychosis; HC - Healthy Control; SZ - schizophrenia. TLI - Thought and Language Index; TLC - Thought, Language, and Communication Scale; K-FTDS - Kiddie Formal Thought Disorder Rating Scale. PANSS - Positive and Negative Syndrome Scale; SIPS - Structured Interview for Prodromal Symptoms; SOPS - Scale of Prodromal Symptoms. WRAT - Wide Range Achievement Test; GFS - Global Functioning Scale; LNS - Letter Number Sequencing; HVLT-R - Hopkins Verbal Learning Test – Revised; CPT-IP - Continuous Performance Test, Identical Pairs version; GFS-R - Global Functioning Scale Role; GAF - Global Assessment of Functioning.

Paper	Diagnosis	PD	HC	Language	TLC/TALD	Symptoms	Other Scales	Task Type
Nikzad et al., 2022	PD (SSD, BD, MDD)	81	124	English	TLC	SANS	BPRS	open-ended interview, picture description task
Liebenthal et al., 2023	PD (SSD, BD, MDD)	59		English		PANSS		semi-structured interview
Girard et al., 2022	PD (SSD, BD, MDD)	38		English		PANSS	BPRS MADRS, YMRS	semi-structured interview
Silva et al., 2023	FEP (SZ, BD, MDD, NOS, CHR)	26	12	English	TLI	PANSS	SOFAS	picture description task
Vail et al., 2018	PD (SSD, BD, MDD)	28		English		PANSS		semi-structured interview
Xu et al., 2022	AVH	134		English	TALD		HPSVQ	story
Xu et al., 2020	AVH	142		English	TALD			story
Mota et al., 2023	PD FEP	24	33	Portuguese		PANSS		picture description task

TABLE A.2: sample characteristics for studies on psychotic disorder populations.
PD - psychotic disorder; SSD - Schizophrenia Spectrum Disorder; BD - Bipolar Disorder; MDD - Major Depressive Disorder; AVH - Auditory Verbal Hallucinations; PD FEP - Psychotic Disorder First Episode Psychosis; HC - Healthy Control. TLC - Thought, Language, and Communication Scale; TALD - Thought and Language Disorder Scale. PANSS - Positive and Negative Syndrome Scale; SANS - Scale for the Assessment of Negative Symptoms; SAPS - Scale for the Assessment of Positive Symptoms. BPRS - Brief Psychiatric Rating Scale; MADRS - Montgomery-Asberg Depression Rating Scale; YMRS - Young Mania Rating Scale; SOFAS - Social and Occupational Functioning Assessment Scale; HPSVQ - Hamilton Program for Schizophrenia Voices Questionnaire.

Paper	SSD	HC	Other	Language	TLC	Symptoms	Cognitive	Clinical	Other Scales	Task Type
Boer et al., 2023	142	142		Dutch		PANSS	BACS			semi-structured interview
Voppel et al., 2023	94	73		Dutch		PANSS				semi-structured interview
Corona-Hernández et al., 2023	50	50		Dutch		PANSS				semi-structured interview
Wouts et al., 2021	170	147	22 MDD	Dutch						semi-structured interview
Doré, 2019	50	50		Dutch		PANSS				semi-structured interview
Tang et al., 2023b	90	76		English	TLC	SANS				open-ended interview
Tang et al., 2023a	63			English	TLC	SANS	WRAT-3	BPRS	ER40, AIHQ	open-ended interview
Iter et al., 2018	9	5		English		SANS, SAPS				semi-structured interview
Willits et al., 2018	200	55	HIV+ as HC	English						semi-structured interview
Elvevåg et al., 2007	26	25	10 high TD 11 low TD	English	TLC		WAIS-R IQ	BPRS		semi-structured interview
Just et al., 2023	71		51 longitudinal	German		PANSS	Verbal IQ WST	MINI-ICF		semi-structured interview
Schneider et al., 2023	34	40	38 MDD	German		SANS, SAPS	Verbal IQ WST, VLMT		GAF, HAM-D, HAM-A	picture description task
Just et al., 2020	20	20		German		SANS, SAPS	Verbal IQ WST	CGI		semi-structured interview
Ryazanskaya, 2020	20	21		Russian		PANSS				story, instruction
Ryazanskaya et al., 2020	9	10		Russian						video-description task

TABLE A.3: sample characteristics for studies on schizophrenia spectrum disorder populations. SSD - Schizophrenia Spectrum Disorder; HC - Healthy Control; MDD - Major Depressive Disorder; HIV+ - Tested Positive for Human Immunodeficiency Virus; TD - Thought Disorder. TLC - Thought, Language, and Communication Scale. PANSS - Positive and Negative Syndrome Scale; SANS - Scale for the Assessment of Negative Symptoms; SAPS - Scale for the Assessment of Positive Symptoms. BACS - Brief Assessment of Cognition in Schizophrenia; WRAT - Wide Range Achievement Test; WAIS - Wechsler Adult Intelligence Scale; WST - Wortschatztest; VLMT - Verbale Lern- und Merkfähigkeitstest . BPRS - Brief Psychiatric Rating Scale; MINI-ICF - Shortened version of International Classification of Functioning, Disability and Health Social Functioning Scale; CGI - Clinical Global Impression. ER40 - Penn Emotion Recognition Tests; AIHQ - Ambiguous Intentions Hostility Questionnaire; GAF - Global Assessment of Functioning; HAM-D - Hamilton Depression Scale; HAM-A - Hamilton Anxiety Scale.

paper	SZ	HC	Other	Language	TLC/TLI	Symptoms	Cognitive	Social	Clinical	Task Type
Parola et al., 2023	51	42		Chinese		PANSS, SAPS, SANS	Verbal IQ			short story
Parola et al., 2023	111	129		Danish		PANSS, SAPS, SANS	Verbal IQ			short story
Voppel et al., 2021	50	50		Dutch		PANSS				semi-structured interview
Jeong et al., 2023	7			English	TLC	SANS, SAPS				open-ended interview
Minor et al., 2023	101			English		PANSS				semi-structured interview
Alonso-Sánchez et al., 2023	60	30		English	TLI	PANSS		SOFAS		picture description task
Alonso-Sánchez et al., 2022	46	36	20, 13 longitudinal	English	TLI	PANSS	DSST, Stroop	SOFAS		picture description task
Liang et al., 2022	66	36		English	TLI	PANSS		SOFAS		picture description task
Aich et al., 2022	247	110	286 BD	English				SSPA		dialogue
Tang et al., 2021	20	11		English	TLC					open-ended interview
Mitchell et al., 2015				English						-
Elvevåg et al., 2010	53	19	11 first degree relatives	English						heterogeneous unstructured prompts
Parola et al., 2023	25	29		German		PANSS, SAPS, SANS	Verbal IQ WST			short story
Just et al., 2019	20	10		German		SANS, SAPS	Verbal IQ WST		CGI	semi-structured interview
Kořánová, 2017	28			German		PANSS, SAPS, SANS				semi-structured interview, dialogue
Shriki et al., 2022	23	28		Hebrew		PANSS				written picture description, written open question
Ziv et al., 2022	24	25		Hebrew						picture description task
Bar et al., 2019	24	27		Hebrew		PANSS				written picture description, written open question
Sarzynska-Wawer et al., 2021	35	35		Polish	TLC					semi-structured interview
Mota et al., 2017	21	21	20 BD	Portuguese		PANSS				story, picture description task
Mota et al., 2016	21	21		Portuguese		PANSS				short story
Mota et al., 2014	20	20	19 BD	Portuguese		PANSS			BPRS	short story
Mota et al., 2012	8	8	8 BD	Portuguese		PANSS			BPRS	short story
Panicheva et al., 2019	12	12		Russian						written story

TABLE A.4: sample characteristics for studies on schizophrenia populations. SZ - Schizophrenia; HC - Healthy Control; BD - Bipolar Disorder. TLI - Thought and Language Index; TLC - Thought, Language, and Communication Scale. PANSS - Positive and Negative Syndrome Scale; SANS - Scale for the Assessment of Negative Symptoms; SAPS - Scale for the Assessment of Positive Symptoms. DSST - Digit Symbol Substitution Test. SOFAS - Social and Occupational Functioning Assessment Scal; SSPA - Social Skills Performance Assessment.BPRS - Brief Psychiatric Rating Scale; CGI - Clinical Global Impression.

Appendix B

Methods

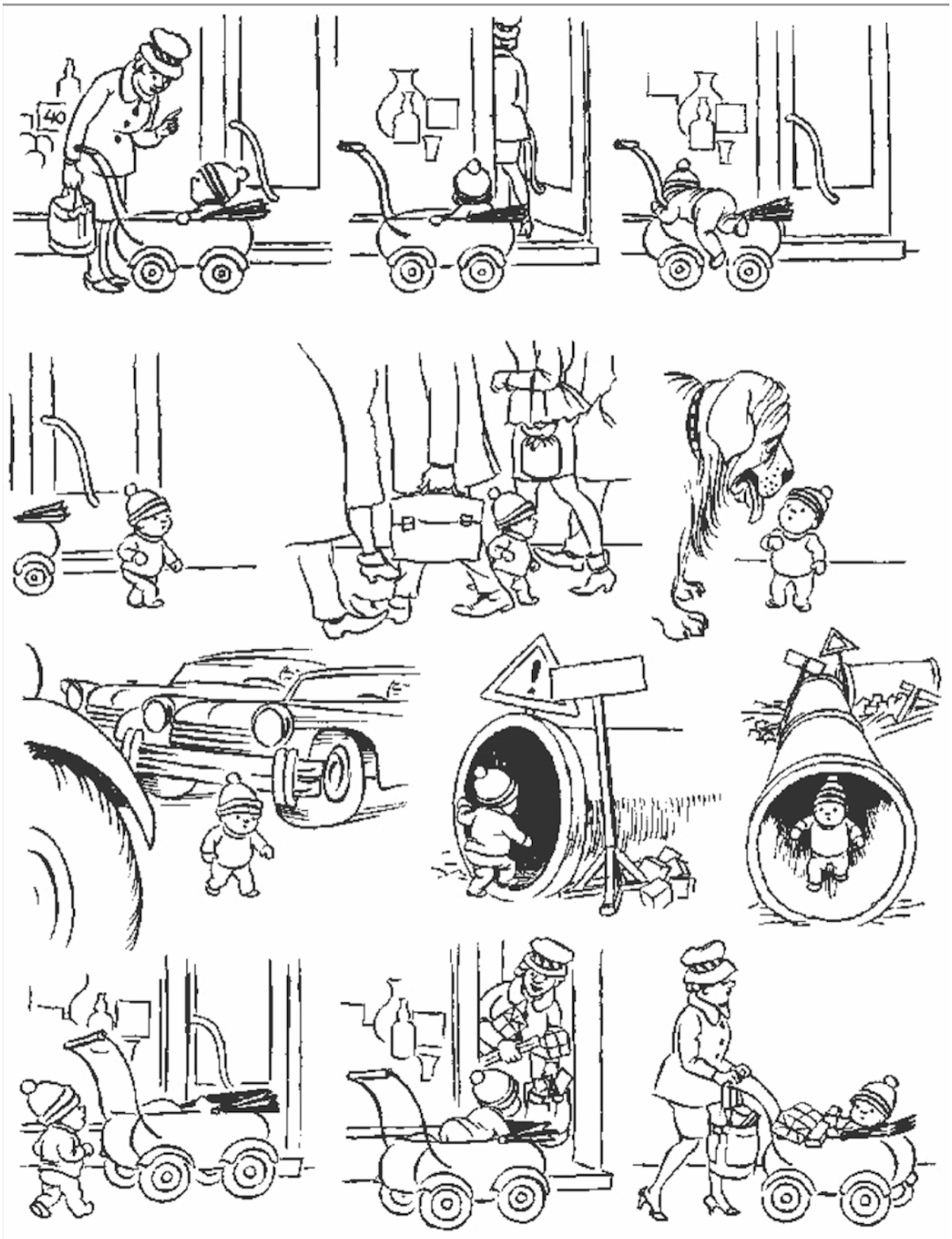

FIGURE B.1: The image used to elicit adventure task.

Lexical Metric	Count	+	-	!	NR	Corr. Symptoms	Classifier Feature
Word Count	29		Mota et al., 2014; Iter et al., 2018; Willits et al., 2018; Doré, 2019; Just et al., 2019; Just et al., 2020; Panicheva et al., 2019; *Morgan et al., 2021*; *Spencer et al., 2021*; Voppel et al., 2021; Liang et al., 2022; Parola et al., 2023	*Mota et al., 2012*; *Gupta et al., 2018*; *Ryazanskaya et al., 2020*; *Morgan et al., 2021*; *Tang et al., 2021*; *Alonso-Sánchez et al., 2023*; *Alonso-Sánchez et al., 2022*; *Schneider et al., 2023*; *Nettekoven et al., 2023*	Kořánová, 2017; *Argolo et al., 2023*	*Morgan et al., 2021*; *Minor et al., 2023*	Liebenthal et al., 2023; Tang et al., 2023a; Tang et al., 2023b; Voppel et al., 2023
LD: TTR+	8	Ziv et al., 2022	Willits et al., 2018; Aich et al., 2022; Minor et al., 2023	*Hitczenko et al., 2020*; Jeong et al., 2023*; Schneider et al., 2023			*Rosenstein et al., 2015*; Kramov, 2020; Liang et al., 2022; Tang et al., 2023b
LD: Unique Words	2		Willits et al., 2018	Schneider et al., 2023			
LD: Honoré Statistics	1					Jeong et al., 2023	
SA: Emotion Words	6	Mitchell et al., 2015	Mitchell et al., 2015; Aich et al., 2022	Mota et al., 2023		Vail et al., 2018; *Girard et al., 2022*; Minor et al., 2023	
SA: models	5			*Argolo et al., 2023*			Sarzynska-Wawer et al., 2021; Tang et al., 2023a; Tang et al., 2023b; Aich et al., 2022
Neologisms, OOV	2	Just et al., 2019; Just et al., 2020					
Foreign Word Use	1					Jeong et al., 2023	

TABLE B.1: Summary of the results reported for the lexical methods used in the reviewed papers. "+" indicates significant group difference with *higher* values in the patient population. "-" indicates significant group difference with *lower* values in the patient population. "!" indicates absence of significant differences in the metric tested. "NR" indicates that the metric is used but results are not reported. "Corr. Symptoms" indicates significant correlation with symptom severity in any direction. "Classifier Feature" indicates that the metric is used as a feature in a classifier or latent analysis. The studies on clinical high risk populations are shown in italics. * - no correlation with symptoms.

Syntactic Metric	Count	+	-	!	NR	Corr. Ling. Scale	Corr. Symptoms	No Corr. Symptoms	Classifier Feature
POS: det, wh-words	10	Mitchell et al., 2015	*Corcoran et al., 2018*; Sarzynska-Wawer et al., 2021; Tang et al., 2021	*Rezaii et al., 2019*; *Haas et al., 2020*; *Bilgrami et al., 2022*; *Argolo et al., 2023*	*Bedi et al., 2015*	*Corcoran et al., 2018*; *Bilgrami et al., 2022*		*Corcoran et al., 2018*; *Bilgrami et al., 2022*	*Bedi et al., 2015*; Tang et al., 2021
POS: adj	12	*Argolo et al., 2023*	*Corcoran et al., 2018*; Tang et al., 2021; Ziv et al., 2022	*Haas et al., 2020*; *Bilgrami et al., 2022*	*Bedi et al., 2015*; Mitchell et al., 2015; *Rezaii et al., 2019*	*Corcoran et al., 2018*	*Argolo et al., 2023*	*Corcoran et al., 2018*	*Bedi et al., 2015*; Tang et al., 2023b; Sarzynska-Wawer et al., 2021
POS: verb	9	Mitchell et al., 2015	Ziv et al., 2022	*Haas et al., 2020*; Tang et al., 2021; *Argolo et al., 2023*	*Bedi et al., 2015*; *Rezaii et al., 2019*				Sarzynska-Wawer et al., 2021; Tang et al., 2023a
POS: pron	9	Mitchell et al., 2015; Tang et al., 2021	*Corcoran et al., 2018*	*Haas et al., 2020*; *Argolo et al., 2023*	*Bedi et al., 2015*; *Rezaii et al., 2019*	*Corcoran et al., 2018*	Jeong et al., 2023	*Corcoran et al., 2018*	Sarzynska-Wawer et al., 2021
POS: noun	7			*Haas et al., 2020*; Tang et al., 2021; Ziv et al., 2022; *Argolo et al., 2023*	*Bedi et al., 2015*; Mitchell et al., 2015				Sarzynska-Wawer et al., 2021
POS: adv	6		Tang et al., 2021; Ziv et al., 2022	*Haas et al., 2020*; *Argolo et al., 2023*	Mitchell et al., 2015				Sarzynska-Wawer et al., 2021
POS: subconj	5	Silva et al., 2023		*Haas et al., 2020*; Tang et al., 2021; *Argolo et al., 2023*					Tang et al., 2023b
POS: coorconj	4			*Haas et al., 2020*; Tang et al., 2021; *Argolo et al., 2023*					Tang et al., 2023b
Ambiguous Pronouns, Referential Failures	4	Iter et al., 2018; Just et al., 2020	Morgan et al., 2021		*Nettekoven et al., 2023*	Iter et al., 2018	Iter et al., 2018		
Sent. / Unit Length	16		Iter et al., 2018; Palominos et al., 2023; *Spencer et al., 2021*; Tang et al., 2021; *Bilgrami et al., 2022*; Silva et al., 2023; *Nettekoven et al., 2023*; Schneider et al., 2023	Liang et al., 2022; *Gupta et al., 2018*; *Haas et al., 2020*; *Morgan et al., 2021*		Xu et al., 2020; *Bilgrami et al., 2022*	Liebenthal et al., 2023; *Bilgrami et al., 2022*; Silva et al., 2023; Jeong et al., 2023		*Bedi et al., 2015*; Tang et al., 2023b
Sent. Count	7	Morgan et al., 2021; *Nettekoven et al., 2023*	Iter et al., 2018	*Gupta et al., 2018*; Ryazanskaya et al., 2020; *Morgan et al., 2021*; Schneider et al., 2023	Tang et al., 2021		Jeong et al., 2023		

TABLE B.2: Summary of the results reported for the syntactic methods used in the reviewed papers. "+" indicates significant group difference with *higher* values in the patient population. "-" indicates significant group difference with *lower* values in the patient population. "!" indicates absence of significant differences in the metric tested. "NR" indicates that the metric is used but results are not reported. "Corr. Symptoms" indicates significant correlation with symptom severity in any direction. "No Corr. Symptoms" indicates absence of significant correlation with symptom severity in any direction. "Corr. Ling. Scale" indicates significant correlation with a linguistic symptom scale in any direction. "Classifier Feature" indicates that the metric is used as a feature in a classifier or latent analysis. The studies on clinical high risk populations are shown in italics. "POS" is part-of-speech prefix. "det" stands for determiners and determiner pronouns. "adj" stands for adjectives. "pron" stands for pronoun. "adv" stands for adverb. "Sent." stands for sentence.

Graph-Based Method	Count	-	!	Corr. Ling. Scale	Corr. Symptoms	No Corr. Symptoms	Classifier Feature
N	4	Nikzad et al., 2022	Mota et al., 2012; *Nettekoven et al., 2023**			Mota et al., 2012	
E	6	Mota et al., 2014; Mota et al., 2016; Mota et al., 2017; Nikzad et al., 2022	Mota et al., 2012; *Nettekoven et al., 2023**		Mota et al., 2014; Nikzad et al., 2022	Mota et al., 2012	
LCC / median CC	11	Mota et al., 2014; Mota et al., 2016; Mota et al., 2017; Spencer et al., 2021; Morgan et al., 2021; Nikzad et al., 2022; Nettekoven et al., 2023	*Spencer et al., 2021*; *Morgan et al., 2021*; *Nettekoven et al., 2023**; *Argolo et al., 2023*	*Spencer et al., 2021*; *Morgan et al., 2021*; *Nettekoven et al., 2023***	Mota et al., 2014; Mota et al., 2016	Mota et al., 2012; *Nettekoven et al., 2023*	Mota et al., 2017; Mota et al., 2023
LSC	11	Mota et al., 2014; Mota et al., 2016; Mota et al., 2017; Spencer et al., 2021; Morgan et al., 2021	*Spencer et al., 2021*; *Morgan et al., 2021*; Nikzad et al., 2022; *Argolo et al., 2023**	*Spencer et al., 2021*; *Morgan et al., 2021*; Nikzad et al., 2022	Mota et al., 2014; Mota et al., 2016; Nikzad et al., 2022	Mota et al., 2012; *Argolo et al., 2023**	Mota et al., 2017; Mota et al., 2023; Tang et al., 2023a; *Argolo et al., 2023*
LCCz/LCCr	5	Spencer et al., 2021; Morgan et al., 2021	Mota et al., 2016; Mota et al., 2017; *Spencer et al., 2021*; *Morgan et al., 2021*	*Spencer et al., 2021*; *Morgan et al., 2021*	Mota et al., 2016		
LSCz/LSCr	5	Mota et al., 2016; Mota et al., 2017; Spencer et al., 2021; Morgan et al., 2021; Nikzad et al., 2022	*Spencer et al., 2021*; *Morgan et al., 2021*	*Spencer et al., 2021*; *Morgan et al., 2021*; Nikzad et al., 2022	Mota et al., 2016; Nikzad et al., 2022		Mota et al., 2017; Mota et al., 2023
ASP	6	Mota et al., 2014	Mota et al., 2012; Tang et al., 2023a; Nikzad et al., 2022; *Argolo et al., 2023*		Nikzad et al., 2022	Mota et al., 2012	Tang et al., 2023a; *Argolo et al., 2023*
density	5	Nikzad et al., 2022	Mota et al., 2012; Mota et al., 2014; *Argolo et al., 2023**	Nikzad et al., 2022	Nikzad et al., 2022	Mota et al., 2012	Tang et al., 2023a; *Argolo et al., 2023*
diameter	5	Mota et al., 2014	Mota et al., 2012; Nikzad et al., 2022	Nikzad et al., 2022		Mota et al., 2012	Tang et al., 2023a
ATD	4	Mota et al., 2014	Mota et al., 2012*			Mota et al., 2012	Tang et al., 2023a
AWD	1	Nikzad et al., 2022		Nikzad et al., 2022			
Largest Clique	2						Tang et al., 2023a; Tang et al., 2023b
Clustering Coefficient	1						Tang et al., 2023b

TABLE B.3: Summary of the results reported for the graph-based methods used in the reviewed papers. "-" indicates significant group difference with *lower* values in the patient population. "!" indicates absence of significant differences in the metric tested. "NR" indicates that the metric is used but results are not reported. "Corr. Symptoms" indicates significant correlation with symptom severity in any direction. "No Corr. Symptoms" indicates absence of significant correlation with symptom severity in any direction. "Corr. Ling. Scale" indicates significant correlation with a linguistic symptom scale in any direction. "Classifier Feature" indicates that the metric is used as a feature in a classifier or latent analysis. The studies on clinical high risk populations are shown in italics. Some studies report both for SZ and CHR and are reported twice. "*" means the metric was insignificant after corrections for verbosity. "**" means the metric was insignificant after correcting for multiple comparisons. "N" stands for number of nodes. "E" stands for number of edges. "LCC" stands for largest connected component. "CC" stands for connected component. "LSC" stands for largest strongly connected component. "LCCz/LCCr" stands for z-score or randomness of the largest connected component. "LSCz/LSCr" stands for z-score or randomness of the largest strongly connected component. "ASP" stands for average shortest path. "ATD" stands for average total degree. "AWD" stands for average weighted degree.

LM-Based Method	Count	+	-	!	Corr. Ling. Scale	Corr. Symptoms	No Corr. Symptoms
Moving Window Coherence	8	Panicheva et al., 2020 (min); Alonso-Sánchez et al., 2023; Alonso-Sánchez et al., 2022; Voppel et al., 2021 (var); Voppel et al., 2023 (var)	Bar et al., 2019; Panicheva et al., 2020 (max); Parola et al., 2023 (Ch)	Doré, 2019; Parola et al., 2023 (D, G)			
Word-Based Coherence	5	Parola et al., 2023 (D)	Bar et al., 2019; Parola et al., 2023 (Ch, G)	Liebenthal et al., 2023; *Argolo et al., 2023*	Xu et al., 2020; Xu et al., 2022		
K-Inter Word Similarity	3		*Corcoran et al., 2018*	Parola et al., 2023 (?); *Argolo et al., 2023*			*Corcoran et al., 2018*
All word Similarity	3					Alonso-Sánchez et al., 2022	
Vector Magnitude	2			*Rezaii et al., 2019*; *Liebenthal et al., 2023*			
Word-Based Centroid Global Coherence & Word-Based Cumulative Centroid Global Coherence	2				Xu et al., 2020; Xu et al., 2022		

TABLE B.4: Summary of the results reported for the word embedding-based metrics used in the reviewed papers.

"-" indicates significant group difference with *lower* values in the patient population. "!" indicates absence of significant differences in the metric tested. "Corr. Symptoms" indicates significant correlation with symptom severity in any direction. "No Corr. Symptoms" indicates absence of significant correlation with symptom severity in any direction. "Corr. Ling. Scale" indicates significant correlation with a linguistic symptom scale in any direction. The studies on clinical high risk populations are shown in italics. Some studies report both for several languages - "Ch" stands for Chinese, "D" - for Danish, and "G" - for German. "min" indicates that minimum values are reported, "max" stands for maximum values, and "var" - for variance.

LM-Based Method	Count	+	-	!	NR	Corr. Ling. Scale	Corr. Symptoms	No Corr. Symptoms	Classifier Feature
(First-Order) Coherence	22		Iter et al., 2018; Just et al., 2019; Ryazanskaya, 2020; *Morgan et al., 2021*	*Haas et al., 2020*; Just et al., 2020; *Hitczenko et al., 2020*; Parola et al., 2023; *Bilgrami et al., 2022*	*Nettekoven et al., 2023*	Xu et al., 2020; Xu et al., 2022; *Bilgrami et al., 2022*	Ryazanskaya, 2020; Just et al., 2023	*Bedi et al., 2015*; Iter et al., 2018; *Haas et al., 2020*; *Hitczenko et al., 2020*; Parola et al., 2023 (?); *Bilgrami et al., 2022*	*Bedi et al., 2015*; *Rosenstein et al., 2015*; Iter et al., 2018; Just et al., 2020; Ryazanskaya et al., 2020; Sarzynska-Wawer et al., 2021; Tang et al., 2023a; Tang et al., 2023b
Second order Coherence	3		Parola et al., 2023						*Bedi et al., 2015*; Sarzynska-Wawer et al., 2021
Repetitiveness	1			*Morgan et al., 2021*					
Group Global Coherence	4		Ryazanskaya, 2020			Elvevåg et al., 2007; Elvevåg et al., 2010; Ryazanskaya et al., 2020	Ryazanskaya, 2020		
Gold Standard Global Coherence	3		*Morgan et al., 2021*		*Nettekoven et al., 2023*	Ryazanskaya et al., 2020			
Centroid Global Coherence	4		Ryazanskaya, 2020			Xu et al., 2020; Xu et al., 2022	Ryazanskaya, 2020; Just et al., 2023		
Cumulative Centroid Global Coherence	3		Ryazanskaya, 2020			Xu et al., 2020; Xu et al., 2022	Ryazanskaya, 2020		
Slope Tangentiality	9	Iter et al., 2018; Tang et al., 2021		Kořánová, 2017; Just et al., 2019; Doré, 2019; *Hitczenko et al., 2020*; Morgan et al., 2021		Elvevåg et al., 2007		Doré, 2019; *Hitczenko et al., 2020*; Morgan et al., 2021	
Q-similarity Tangentiality	4		*Morgan et al., 2021*	Kořánová, 2017; Just et al., 2019	*Nettekoven et al., 2023*				

TABLE B.5: Summary of the results reported for the phrase embedding-based metrics used in the reviewed papers.
"-" indicates significant group difference with *lower* values in the patient population. "!" indicates absence of significant differences in the metric tested. "Corr. Symptoms" indicates significant correlation with symptom severity in any direction. "No Corr. Symptoms" indicates absence of significant correlation with symptom severity in any direction. "Corr. Ling. Scale" indicates significant correlation with a linguistic symptom scale in any direction. The studies on clinical high risk populations are shown in italics. "?" indicates that the information in the article only allows for a presumable placement of the article.

LM-Based Method	Count	+	!	Corr. Ling. Scale	Corr. Symptoms	No Corr. Symptoms	Classifier Feature
Perplexity / Surprisal	5	Srivastava et al., 2022a	Mitchell et al., 2015; *Srivastava et al., 2022a*; Jeong et al., 2023	Jeong et al., 2023	Vail et al., 2018; Girard et al., 2022; Jeong et al., 2023		
Next-Sentence Prediction	3		*Hitczenko et al., 2020*; *Tang et al., 2021*	Jeong et al., 2023	Jeong et al., 2023	Tang et al., 2021	
Non-Fine-Tuned Classifier	3						Elvevåg et al., 2010; *Rosenstein et al., 2015*; *Srivastava et al., 2022b*
Fine-Tuned Classifier	3						Wouts et al., 2021; Aich et al., 2022; Shriki et al., 2022

TABLE B.6: Summary of the results reported for the LM-based feature metrics used in the reviewed papers.
"+" indicates significant group difference with *higher* values in the patient population. "!" indicates absence of significant differences in the metric tested. "Corr. Symptoms" indicates significant correlation with symptom severity in any direction. "No Corr. Symptoms" indicates absence of significant correlation with symptom severity in any direction. "Corr. Ling. Scale" indicates significant correlation with a linguistic symptom scale in any direction. The studies on clinical high risk populations are shown in italics.

LM	Count	Papers
LSA	10	Elvevåg et al., 2007; Elvevåg et al., 2010; Bedi et al., 2015; Rosenstein et al., 2015; Iter et al., 2018; Xu et al., 2020; Haas et al., 2020; Hitczenko et al., 2020; Tang et al., 2023a; Tang et al., 2023b
Word2Vec	24	Kořánová, 2017; Iter et al., 2018; Rezaii et al., 2019; Just et al., 2019; Bar et al., 2019; Panicheva et al., 2019; Doré, 2019; Ryazanskaya, 2020; Ryazanskaya et al., 2020; Xu et al., 2020; Hitczenko et al., 2020; Morgan et al., 2021; Sarzynska-Wawer et al., 2021; Voppel et al., 2021; Corona-Hernández et al., 2023; Liebenthal et al., 2023; Parola et al., 2023; Tang et al., 2021; Xu et al., 2022; Argolo et al., 2023; Just et al., 2023; Nettekoven et al., 2023; Tang et al., 2023b; Voppel et al., 2023
GloVe	8	Iter et al., 2018; Just et al., 2019; Hitczenko et al., 2020; Alonso-Sánchez et al., 2023; Alonso-Sánchez et al., 2022; Tang et al., 2023a; Just et al., 2020; Tang et al., 2023b
BERT	10	Ryazanskaya, 2020; Hitczenko et al., 2020; Tang et al., 2021; Wouts et al., 2021; Aich et al., 2022; Bilgrami et al., 2022; Srivastava et al., 2022b; Shriki et al., 2022; Xu et al., 2022; Jeong et al., 2023
ELMo	4	Ryazanskaya, 2020; Hitczenko et al., 2020; Sarzynska-Wawer et al., 2021; Srivastava et al., 2022a
sent2vec	3	Iter et al., 2018; Just et al., 2019; Hitczenko et al., 2020
Trigram backoff	3	Mitchell et al., 2015; Vail et al., 2018; Girard et al., 2022

TABLE B.7: Models used for LM-based metrics.

Averaging	Count	Papers
Mean	28	Bedi et al., 2015; Mitchell et al., 2015; Corcoran et al., 2018; Bar et al., 2019; Doré, 2019; Just et al., 2019; Panicheva et al., 2019; Haas et al., 2020; Just et al., 2020; Ryazanskaya, 2020; Xu et al., 2020; Hitczenko et al., 2020; Voppel et al., 2021; Morgan et al., 2021; Sarzynska-Wawer et al., 2021; Alonso-Sánchez et al., 2023; Alonso-Sánchez et al., 2022; Corona-Hernández et al., 2023; Girard et al., 2022; Liebenthal et al., 2023; Srivastava et al., 2022a; Srivastava et al., 2022b; Tang et al., 2023a; Argolo et al., 2023; Jeong et al., 2023; Just et al., 2023; Nettekoven et al., 2023; Tang et al., 2023b; Voppel et al., 2023
Min	14	Bedi et al., 2015; Corcoran et al., 2018; Iter et al., 2018; Panicheva et al., 2019; Haas et al., 2020; Ryazanskaya, 2020; Xu et al., 2020; Morgan et al., 2021; Sarzynska-Wawer et al., 2021; Voppel et al., 2021; Bilgrami et al., 2022; Corona-Hernández et al., 2023; Xu et al., 2022; Voppel et al., 2023
Max	11	Bedi et al., 2015; Corcoran et al., 2018; Panicheva et al., 2019; Haas et al., 2020; Ryazanskaya, 2020; Morgan et al., 2021; Bilgrami et al., 2022; Corona-Hernández et al., 2023; Argolo et al., 2023; Just et al., 2023; Voppel et al., 2023
SD / Variance	9	Bedi et al., 2015; Corcoran et al., 2018; Panicheva et al., 2019; Haas et al., 2020; Hitczenko et al., 2020; Sarzynska-Wawer et al., 2021; Voppel et al., 2021; Corona-Hernández et al., 2023; Voppel et al., 2023
Median	4	Bedi et al., 2015; Sarzynska-Wawer et al., 2021; Corona-Hernández et al., 2023; Parola et al., 2023
Percentiles	2	Corcoran et al., 2018; Panicheva et al., 2019
IQR	1	Parola et al., 2023

TABLE B.8: Averaging methods used for LM-based metrics.

Sent. Averaging	Count	Papers
Mean	23	Elvevåg et al., 2007; Elvevåg et al., 2010; Bedi et al., 2015; Rosenstein et al., 2015; Corcoran et al., 2018; Bar et al., 2019; Doré, 2019; Panicheva et al., 2019; Rezaii et al., 2019; Haas et al., 2020; Hitczenko et al., 2020; Tang et al., 2021; Sarzynska-Wawer et al., 2021; Voppel et al., 2021; Alonso-Sánchez et al., 2023; Alonso-Sánchez et al., 2022; Corona-Hernández et al., 2023; Liebenthal et al., 2023; Parola et al., 2023; Tang et al., 2023a; Just et al., 2023; Tang et al., 2023b; Voppel et al., 2023
IDF	8	Iter et al., 2018; Just et al., 2019; Just et al., 2020; Ryazanskaya et al., 2020; Xu et al., 2020; Hitczenko et al., 2020; Xu et al., 2022; Tang et al., 2023b
SIF	6	Iter et al., 2018; Just et al., 2019; Ryazanskaya, 2020; Hitczenko et al., 2020; Morgan et al., 2021; Nettekoven et al., 2023
Sum	1	Xu et al., 2020

TABLE B.9: Sentence averaging methods used for word embedding-based metrics.

FIGURE B.2: The image used to elicit sportsman task.

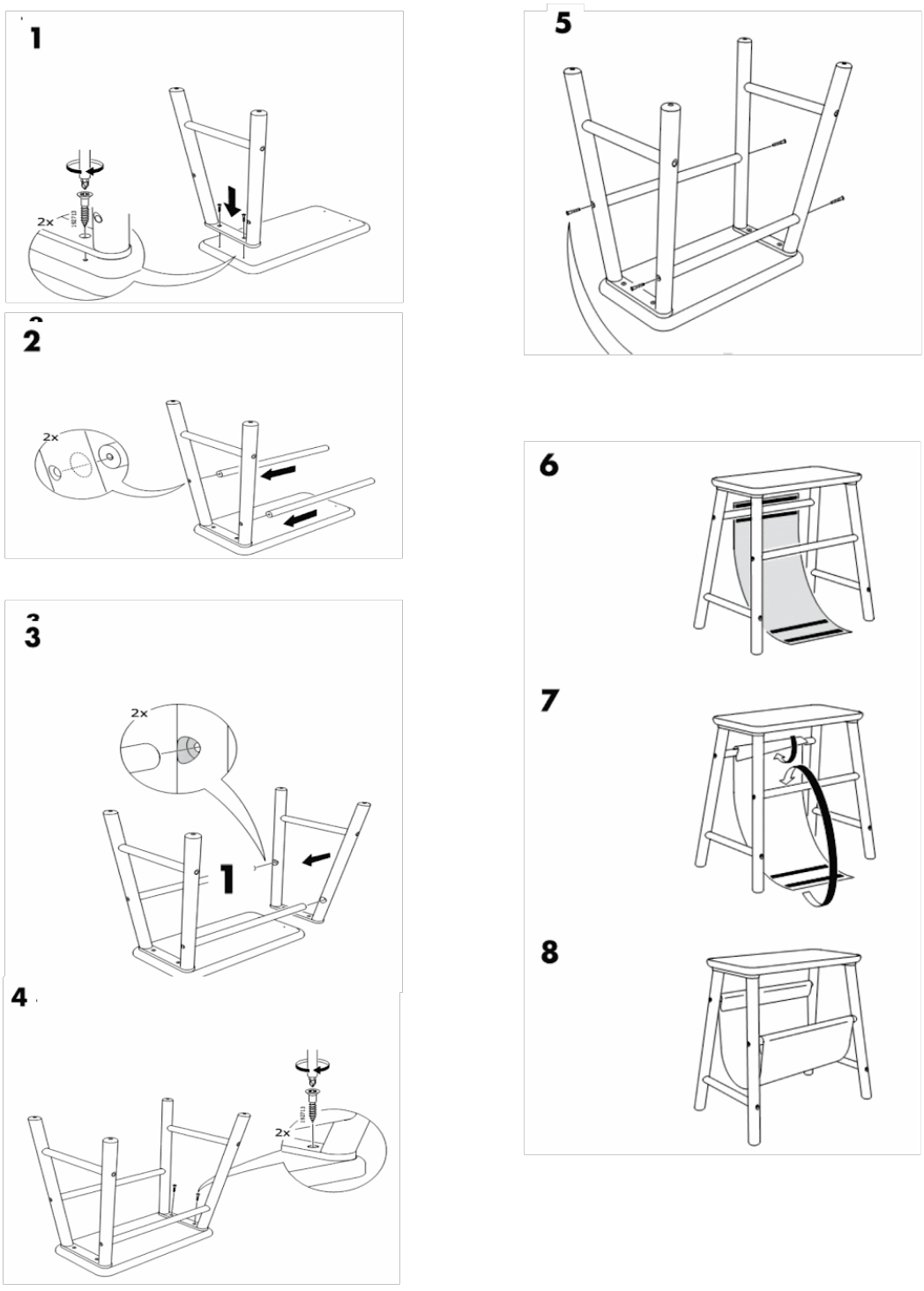

FIGURE B.3: The image used to elicit chair task.

Appendix C

Clinical Data

	code	diagnosis	N
NAP	F20	schizophrenia	20
	F25	schizoaffective.disorder	8
	F21	schizotypal.disorder	2
	F21.3	schizotypal.disorder.pseudoneurotic.schizophrenia	1
Dep	F31	bipolar.affective.disorder	6
	F60.31	borderline.personality.disorder	3
	F31.4	bipolar.affective.disorder.severe	2
	F31.5	bipolar.affective.disorder.severe.psychotic	2
	F33	recurrent.depressive.disorder	2
	F32.1	depressive.episode.moderate	1
	F33.3	recurrent.depressive.disorder.severe.psychotic	1
	F60	personality.disorder	1

TABLE C.1: Diagnosis frequencies in the clinical samples.