Universitat Pompeu Fabra Barcelona

Treball de fi de màster: *Recerca*

Examining Coreference Resolution in Schizophrenia with a Language Model: Detection Gap, Model Performance and Clinical Correlations

Nom i Cognoms: *Karla Fröhlich*

Màster: *Lingüística Teòrica i Aplicada*

Edició: 2022-2023

Director o directora: *Dr. Wolfram Hinzen, Dr. Thomas Brochhagen*

Any de Defensa*:* 2023

Col.lecció: Treballs de fi de màster

Departament de Traducció i Ciències del Llenguatge

Acknowledgements

First and foremost, I would like to express my heartfelt gratitude to my supervisors, Dr. Wolfram Hinzen and Dr. Thomas Brochhagen. I am incredibly thankful to them for providing me with the opportunity to delve into the realms of clinical and computational linguistics, two fields that were entirely new to me. Their guidance and support have not only shaped this study but also granted me a glimpse into these fascinating worlds, which I now shyly consider myself a part of.

I would also like to extend my sincere appreciation to my colleagues Andreea Nösner, Remo García Pellicer, Claudio Palominos Flores, and Rui He, who have played a crucial role in supporting me and this thesis throughout its research process and aiding in data analysis. I am truly grateful for their contributions, which have significantly enriched the findings and enhanced the overall value of this research endeavor.

I would like to express my deepest gratitude to the Department of Psychiatry of the Phillips University of Marburg, who generously provided the data for this study. I extend my heartfelt thanks to Prof. Tilo Kircher and Dr. Frederike Stein, for their invaluable contributions throughout the data collection process and the enriching correspondence.

Last, but not least, I am deeply grateful to my family and friends for their unwavering support, love, and belief in me.

Abstract

The occurrence of referential anomalies in spontaneous speech in psychosis, and more specifically in patients with formal thought disorder, has been documented since the late 1970s, and recent linguistic studies have shown in several typologically different languages that the quantitative distribution and quality of use of different types of noun phrases differs in people with psychosis, specifically in anaphoric noun phrases that pick up a previously mentioned entity. To identify changes in such spontaneous speech patterns, automated natural language processing (NLP) technologies are now frequently used with the intention that this line of research may eventually help with clinical goals of diagnosis and prognosis. Here we are taking a coreference resolution model applicable to German and test the prediction that it will fare worse when processing coreferential expressions from schizophrenic speech. Comparisons between patient and control groups, as well as correlations with clinical variables, revealed statistically non-significant results, suggesting that the referential coherence of schizophrenic patients' speech is comparable to that of healthy controls, from the viewpoint of this model. A post-hoc analysis of averaged semantic similarity between consecutive words yielded similarly non-significant group differences. Together, these results suggest semantic structure at both the referential and lexical-conceptual level to be intact in schizophrenia, questioning the generalizability of current evidence to the contrary.

Keywords: schizophrenia, formal thought disorder, coreference resolution, NLP

Table of Contents

Acknowledgements 1

Abstract 2

1. Introduction 5

The present study 8

2. Methodology 9

2.1. Participants and corpus 9

2.2. Procedure 11

2.3. Computational tools 13

2.4. Annotation scheme 14

2.5. Scoring and metrics 15

2.6 Post-hoc analysis 17

2.7 Statistical analysis 17

3. Results 18

4. Discussion 20

5. Conclusions 25

References 27

Appendix 33

Appendix A 33

Appendix B 35

1. Introduction

In the realm of human language, words serve as the means to identify and designate entities in the world, enabling subsequent co-references to these entities as the discourse unfolds. This process of (re-)referencing provides a sense of temporal stability, allowing for a continuity of narratives over time and a sense of coherence, as the same entities are tracked across a series of distinct events. The way such (re-)referencing occurs relies on the grammatical structures involved in the construction of noun phrases (NPs) and sentences (Hinzen & Sheehan, 2013). In clinical conditions such as schizophrenia, disturbances of coherence at the discourse level and referential anomalies may arise, leading to deviations from normative constraints on this pattern of (re-) referencing. This particularly concerns patients with high levels of the symptom of formal thought disorder (FTD), which is measured largely linguistically (Andreasen, 1986).

Numerous studies have revealed that there are differences not only in the quantity and distribution (e.g., overuse of pronouns) but also in the quality of use of referential NPs, resulting in referential anomalies. A lot of this work traces back to Rochester & Martin (1979), who documented problems of referential cohesion in FTD, specifically in the use of pronouns. In the instance of Turkish speakers with FTD, Çokal et al. (2022) discovered that they overproduce "bare" NPs, which lack grammatical function words like "the" or "a" in English. Schizophrenic speakers without elevated levels of FTD exhibited the same patterns but to a lesser extent. These findings imply that there are specific grammatical effects in the speech of persons with schizophrenia, which may be symptomatic of underlying cognitive and language problems associated with the condition. Similarly, Sevilla et al. (2018) in a study of Spanish assessed the use of referential NPs in a fairytale narrative and found that definite NPs (e.g., "the girl with the hat")

including pronouns were more often misused than indefinite NPs (e.g., “some girl”). Using graph theory, Palominos et al. (2023) demonstrated that there is a difference in Chilean Spanish schizophrenic patients’ cognitive control over the process of generating and maintaining references over narrative time: while patients referred to the same amount of entities, they co-referenced them more and the distances between subsequent references to the same entity decreased (called a “narrowing of the temporal window”).

Recent advances in natural language processing (NLP) techniques enable automated, economical, and quantitative measurements of speech incoherence, implying that NLP methods might support clinical observations and provide an additional layer of knowledge that may mitigate the impact of human subjectivity. Several prior studies have employed diverse NLP techniques in the domain of psychosis prediction, where referentiality through pronouns and determiners often appears: Bedi et al. (2015) predicted the emergence of psychosis in high-risk youths with 100% accuracy using a classifier that combined automated semantic coherence measurements with the variables of maximum phrase length and determiner frequency. Iter et al. (2018) were unable to reproduce Bedi et al.’s (2015) finding but achieved 93% accuracy using a random forests binary classifier to separate a small sample of patients with schizophrenia from healthy controls. The semantic coherence measure and the unresolved pronominal reference measure were both incorporated into this classifier. Corcoran et al. (2018) differeciated indicators among individuals who exhibited symptoms of psychosis and those who did not progress towards the development of psychosis in a dataset of 59 high-risk participants with 83% accuracy using a classifier based on fourteen factors (including pronouns and determiners), while Sarzyńska-Wawer et al. (2021) used a word embedding model to represent interviews with schizophrenia patients and healthy controls. Results were compared to Bedi et al. (2015)’s approach. The model obtained an accuracy

of 80%, while its manual (human-rated) counterpart, the canonical Thought, Language, and Communication (TLC) rating scale of Andreasen (1986) achieved 74%.

Several studies in the field of computational linguistics have also focused on assessing semantic similarity, using large computational language models to represent the lexical meaning of content words (Just et al., 2019; Sarzyńska-Wawer et al., 2021). Computerized examinations of linguistic coherence and content have been interpreted to show that persons with schizophrenia and those who are at high clinical risk had lower levels of coherence and semantic richness (Tang et al., 2021; Bilgrami et al., 2022). A substantial proportion of these models approximate word meaning through the method of word embeddings, which consist of multi-dimensional vectors representing the co-occurrence patterns of words. Semantic similarity is then usually measured as the cosine similarity between these vectors, under the apriori hypothesis that decreased levels of coherence in speech in psychosis should map onto lower levels of overall semantic similarity between words. While this has been a replicated finding, Palominos et al. (2023) in the above-mentioned study applied this method as well, using the language model FastText, and found no significant differences between groups in terms of the averaged semantic similarity between word pairs. Their findings were therefore confined to the referential-semantic level, and these authors concluded that this referential and the conceptual-semantic level are complementary, each needing separate attention. Other studies have found a surprising *increase* in semantic similarity in a first-episode psychosis group (Alonso-Sánchez et al., 2022), thereby questioning the interpretation of semantic similarity as a measure of coherence. Parola et al. (2022) report a similar result for a Danish sample, and, in a meta-analysis of existing results, additionally show low levels of generalizability of these lexical-semantic metrics. In short, while several studies have been

promising for the project of classifying psychosis through computational semantic measures, both the interpretation and generalizability of these results are currently unclear.

In this context, the referential-semantic measures, such as those used in Palominos' (2023) study, are important through their complementarity to lexical-semantic measures. The very nature of linguistic meaning suggests this need, as it includes two fundamental components of meaning, which are integrated into any act of referential language use: a lexical-conceptual component, which identifies entities descriptively; and a grammatical-functional component, which provides a mechanism for linking general semantic concepts (e.g. *HOUSE*) to a specific entity (e.g. *this house*; *a house I own*). For a full story of coherence in psychosis, extant lexical-semantic computational metrics need to be complemented by referential-semantic ones, across more languages than have so far been studied.

The present study

The current study builds on the extensive body of literature that highlights difficulties in resolving referential anomalies in people with schizophrenia, especially those who show symptoms of FTD. The difficulty of resolving referential anomalies is influenced by a variety of elements within this broad spectrum. At its foundation, the key principle is, however, the inherent difficulty that the hearer faces in correctly recognizing and understanding the intended referent that the speaker meant. This effectively motivates the notion that NPs, within the context of schizophrenia samples, especially the ones experiencing FTD, should inherently pose a greater degree of difficulty when it comes to their resolution, in comparison to individuals without the disorder. Establishing anaphoric reference and allowing the identification of the antecedent of a given co-referential

expression is a process so fundamental to narrative coherence that, without it, there would not be any form of coherence at all. It is therefore an important target of investigations in this domain.

Based on this we here aimed to take a state-of-the-art coreference resolution model (Schröder et al., 2021) and to investigate its performance on German speech samples from schizophrenia patients and compare it to that on samples from healthy controls. The primary objective was to examine how the model fares when processing linguistic data derived from schizophrenic speech, specifically in relation to its performance in resolving reference. Moreover, we aimed to explore the potential influence of clinical variables, such as negative symptoms, positive symptoms, and the presence of FTD, on the model's performance. By analyzing these variables alongside the model's outputs, we seek to discern any correlations or associations that shed light on the impact of these clinical factors on the resolution of anaphoric NPs. We hypothesized that schizophrenia samples will exhibit lower coreference resolution performance from the language model than control samples. We also predicted that clinical assessments of conceptual disorganization and formal thought disorder will be correlated with how well the model performs in the coreference resolution task. Finally, because of the unexpected results of this thesis, which did not confirm these predictions, a post-hoc analysis of semantic similarity at the lexical-semantic level was conducted as well, to open a complementary window on the semantic organization in the speech of this sample and potentially illuminate the result obtained.

2. Methodology

2.1. Participants and corpus

This study was based on a spontaneous speech dataset collected using the Thematic Apperception Test (Liddle, et al., 2002) from 48 German-speaking, clinically stable in-patients with

schizophrenia spectrum disorders (Kircher, et al., 2019). This data was compiled and kindly provided to us by the research team of the Department of Psychiatry of the Phillips University of Marburg (PI: Prof. Tilo Kircher; lead investigator: Dr. Friederike Stein). Speech samples were composed of picture descriptions and stories about what is happening in them. Speech data were transcribed verbatim manually. The transcription conventions employed are provided in Appendix B for detailed documentation. Demographic and clinical characteristics of the sample are summarized in Table 1. For purposes of the present thesis, given its time constraints, annotations were restricted to 20 patients (Sz) and 20 healthy controls (HC). Two subjects were excluded from the analysis due to missing data in their diagnostics, resulting in a final sample size of 18 patients and 20 healthy controls. SCID diagnosis according to DSM-IV-TR and a semi-structured clinical interview codes are "Paranoid schizophrenia" (295.3/30, 9 patients), "Disorganized schizophrenia" (295.10, 3 patients), "Residual schizophrenia" (295.6, 4 patients) and "Undifferentiated schizophrenia" (295.90, 2 patients). The SANS (Scale for the Assessment of Negative Symptoms) was used to examine clinical factors related to negative symptoms, whereas the SAPS (Scale for the Assessment of Positive Symptoms) was used to assess clinical variables linked to positive symptoms. The SAPS information was collected during the semi-structured interview. The values are therefore not based on the spontaneous speech samples from the picture description task but on the clinical impression of the interviewer during the clinical data collection. These ratings confirmed FTD in all but six patients, who did not demonstrate any signs of this symptom and got a score of zero out of 40.

Table 1: Demographics of participants

	HC	Sz	Test	Statistic	*p*
Number	20	18	/	/	/
Age	43.25	41.5	IS t-test	.455	.651
Gender	30%	27.77%	χ^2 test	1.17e-31	1
Education	15.6	12.125	IS t-test	3.894	.0004***
Age of onset of first SCID diagnosis	/	19.12	/	/	/

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Mean values were used to represent age and age of onset. Gender was represented by the percentage of female subjects. For years of education, two samples had to be removed from the analysis due to missing values. In the case of the age of onset, one sample was removed due to missing data.

2.2. Procedure

Prior to the referential analysis, the interviews were subjected to a preprocessing phase. Given the large variability noted in interview transcript formats, a manual preprocessing approach was employed.

The first step involved standardizing the data into a unified format for further processing, by removing all elements that the files had in common that were not relevant to the analysis: line numbering and the transcription head, as well as information like participant ID, the date of the recording, the duration of the recording, the transcriber ID, the transcription date, and the number of prompts given during the interview for subjects to encourage them to continue talking. To isolate the participant's speech, a manual extraction process was conducted, removing any sections that contained interviewer speech. Furthermore, to make the participant's speech as seamless as

possible for the coreference resolution model to read and process, annotations of pauses and other non-verbal elements, such as prolonged silences or hesitations, were removed. This resulted in a mean number of 1161,42 words per participant (SD = 435.2).

Following these preprocessing steps, all files were manually annotated for coreference relations to form the human gold standard dataset prior to the automated analysis. Further details regarding the annotation process can be found in Section 2.4. The annotation was performed while blinded to the participant's diagnostic status, to reduce any unintentional influence and to facilitate an unbiased analysis. After the manual annotation procedure, the preprocessed files were run through the computational model (specified below).

To assess its performance, a scoring mechanism was employed that involved comparing the model's output with the gold standard annotations, using established evaluation metrics for measuring coreference resolution (Pradhan et al., 2012). As described in Section 2.5, four separate metrics were utilized to evaluate the model's performance, each providing a different viewpoint on coreference resolution. Specifically, the link-based MUC metric (Vilain et al., 1995), the mention-based B-CUBED metric (Bagga & Baldwin, 1998), and the entity-based CEAFe metric (Luo, 2005). CoNLL is the arithmetic mean of these metrics, offering an overall assessment of system performance[1].

[1] Initially, an automated scoring system was used to calculate these scores (Moosavi et al., 2019). However, during its evaluation, formatting issues were encountered that resulted in several biases and skewed data. Consequently, a manual scoring procedure was applied, calculating each of the metrics individually and averaging them to form the final CoNLL score.

2.3. Computational tools

A pre-trained model from Schröder et al. (2021) was utilized for German coreference resolution, enabling the identification of referential relationships within the dataset. Schröder et al. developed a neural end-to-end approach that trains contextual word embeddings jointly with mention and entity similarity scores. This approach has proven to be highly effective, achieving state-of-the-art performance across three established datasets for German: CoNLL F1 scores of 78.79 on the TüBa-D/Z dataset, a manually annotated collection of newspaper articles (Telljohann et al. 2012), 74.46 on the SemEval-2010 dataset, used to assess coreference resolution algorithms (Recasens et al., 2010) and 64.72 on the DROC dataset, containing 90 literary documents annotated for coreference (Krug et al., 2018). This approach involves two distinct model architectures: a mention linking-based approach and an incremental entity-based approach. The mention linking-based approach involves identifying mentions in the text and then linking them to entities based on their similarity. This approach is well-suited for short documents such as news articles. On the other hand, the incremental entity-based approach involves building up entities incrementally as more text is processed. This approach is better suited for longer documents such as literary works. In the present case, the mention linking-based approach was chosen based on several factors. Before all else, the mention-linking approach is computationally efficient, as it only considers the cross-products of mentions with the highest likelihood of being linked. Furthermore, by considering only the highest-scoring antecedent for each mention, the approach can make accurate local decisions (Schröder et al., 2021). This makes it well-suited for short documents, where the number of entities is relatively small, and thus, it fits well with the interview transcripts studied here. It should be noted that due to restricted access to the underlying model architecture, the model could only be

employed through the provided author's interface (https://ltdemos.informatik.uni-hamburg.de/coref-de/).

The second resource was used to aid the manual annotation: CorefAnnotator, as proposed by Reiter (2018). This tool was primarily developed to facilitate the annotation of entity references in textual data. It provides a user-friendly interface for annotating texts with coreference information. Each entity is represented by a color, and all references of the same entity in the text view are underlined with the same color. Additionally, each entity can be named and chosen entities in the entity list are highlighted in the text view, making it easier to rapidly find references to certain entities. Following the usage of the tool, these annotations are utilized as a gold standard against which the neural coreference resolution model is benchmarked.

2.4. Annotation scheme

The annotation scheme is inspired by a pre-existing manual (Naumann, 2006), adapted for present purposes so as to comprise the following steps:

a. Identification of anaphoric expressions, which may manifest as pronouns or noun phrases.
b. Marking of the identified anaphoric expressions.
c. Identification of their corresponding antecedent(s).
d. Marking of the identified antecedent(s) to establish their explicit relationship with the respective anaphoric expressions.

It is important to note that in the case of German, a significant number of pronouns serve as anaphoric expressions: personal pronouns ("er/sie/es"), possessive pronouns ("seine/seiner"), reflexive pronouns ("mir/dir"), demonstrative pronouns ("diese/dieser"), relative pronouns ("der/den"), interrogative pronouns ("wer"), and indefinite pronouns ("einer/eine"). Additionally,

proforms, which encompass adverbs such as “hier” (here) or “dort” (there), can also function as anaphoric expressions.

2.5. Scoring and metrics

Well-established scoring metrics were used to evaluate the efficacy of Schröder et al. 's (2021) model of coreference resolution on our dataset, namely MUC, B-CUBED, CEAFe, and CoNLL. The B-CUBED and CEAFe evaluation metrics involve calculating precision, recall, and F1 scores at both the mention level, where individual entity mentions are assessed, and the cluster level, where groups of related mentions are evaluated as a whole.

MUC (Vilain et al., 1995) focuses on links between pairs of mentions - a link is formed when two mentions relate to the same entity. The number of common links between entities in the gold standard and in the model output divided by the number of links in the gold standard represents recall, whereas precision is the number of common links between entities in the gold standard and model output divided by the number of links in the model output.

B-CUBED (Bagga & Baldwin, 1998) assesses how well the model recognizes individual references that pertain to the same entity. Using pairwise mention similarity, for each mention in the model output, the metric calculates the similarity between the model output mentions and all gold standard mentions. Precision is determined for each mention in the model output by examining its similarity to the corresponding mentions in the gold standard, hence quantifying the fraction of correctly detected mentions in the model output. In terms of recall, it is determined for each mention in the gold standard by considering the resemblance to the corresponding mentions in the model output and assessing the fraction of properly detected mentions from the gold

standard. B-CUBED presupposes a one-to-one relationship between model output and gold standard mentions. It regards each mention as a separate unit.

Finally, CEAFe (Luo, 2005) assesses the alignment between clusters created by the model output and the gold standard. CEAFe's purpose is to match each entity in the model output with at least one gold standard entity in the reference. Alignment is the process of creating a correlation between a model output entity and a gold standard entity. The alignment method is centered on determining the optimal one-to-one mapping between entities, which means that each entity of the model output should be aligned with just one gold standard entity and vice versa. CEAFe aligns every model output entity with at least one gold standard entity by determining the optimal one-to-one mapping between the entities using an entity similarity measure. Recall is calculated by dividing total similarity by the number of mentions in the gold standard, and precision is calculated by dividing total similarity by the number of mentions in the model output. These three metrics contribute to the final CoNLL score, which is their average (Pradhan et al., 2012).

Taking into account the impact of psychosis on speech, which has been known to manifest itself through referential anomalies (Rochester & Martin, 1979), the CEAFe metric might be particularly suitable in the analysis of coreference resolution in Sz samples. While link- and mention-based metrics represent important pillars in coreference resolution evaluation, their emphasis on single links or individual mentions may obscure the overall coherence of entity-level coreference connections. Entity-based metrics, on the other hand, such as CEAFe, give a more comprehensive assessment of coreference resolution by considering the alignment and structure of complete entities. By considering the alignment of entire entities, CEAFe could allow for an examination of how well the system captures the overall coherence of coreference clusters. This

may be more relevant for understanding the influence of disorder severity on coreference resolution.

2.6 Post-hoc analysis

The post hoc analysis of semantic similarity at the lexical-semantic level consisted in representing word meaning with embeddings from German corpora through the language model FastText (Grave et al., 2018). The averaged semantic similarity was obtained for consecutive pairs of words produced. For each participant, their interview was split into individual picture descriptions. Then, stopwords, which are frequently used (e.g., grammatical function) words that have little relevance in the context of the lexical-semantic analysis, were eliminated from these descriptions as part of the preprocessing stage. Following the completion of the preprocessing, the picture descriptions were analyzed utilizing a Python script that calculated the mean, standard deviation, minimum, and maximum values for consecutive semantic similarities within the picture descriptions. These results were then averaged for every participant and added to the statistical analysis together with the coreference resolution scores.

2.7 Statistical analysis

Mann-Whitney U tests were performed to assess group differences in individual scores, as well as the overall CoNLL score. In the Sz group, Spearman's rank correlation test was employed to assess the relationship between the CEAFe scores and the following three clinical variables: the sum score of negative symptoms, the scale of FTD, and the sum score of positive symptoms. Considering the difference in educational years across groups, an additional correlational analysis in the form of Spearman's rank correlation coefficient was carried out, to look into the relationship between educational years and performance scores, specifically the CEAFe and CoNLL. The

objective of this was to establish the association between educational attainment and performance scores while taking into consideration the observed variation in educational years between the groups. For the post-hoc analysis, groups were compared in their consecutive semantic similarity scores using Mann-Whitney U tests. These tests were performed using the statistical softwares R (version 4.1.3) and JASP (version 0.17.2.1).

3. Results

As seen in Figure 1, at the chosen alpha-threshold of $\alpha = 0.05$, the results indicated no statistically significant between-group differences in the MUC score ($Z = -0.248$, $p = .803$), the B-CUBED score ($Z = -0.301$, $p = .762$), the CEAFe ($Z = -0.348$, $p = .727$), or the CoNLL ($Z = -0.409$, $p = .682$).

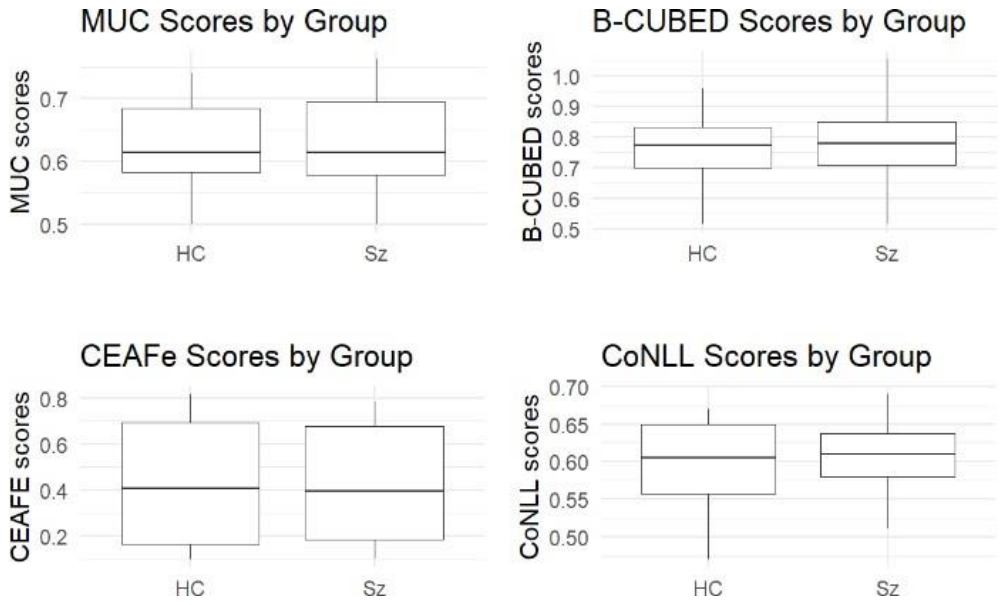


Figure 1: Box plots of the distribution of individual scores and CoNLL score between HC and Sz groups

The correlation analysis between CEAFe scores and the sum score of negative symptoms revealed a non-significant correlation at the selected alpha-threshold of $\alpha = 0.05$, for mention level (rho = 0.078, p = .756) and cluster level respectively (rho = 0.337, p = .171). Similarly, the correlation between CEAFe scores and the scale of positive formal thought disorder was not significant for mentions (rho = -0.244, p = .328) and clusters (rho = 0.225, p = .369). As for CEAFe and the sum score of positive symptoms, the results indicated no statistically significant association between these variables on mention level (rho = -0.343, p = .162) or cluster level (rho = -0.076, p = .761).

The findings of the supplementary correlational analysis showed that there was a statistically insignificant correlation between years of education and CEAFe scores on mention level (rho = -0.173, p = .286), CEAFe scores on cluster level (rho = 0.010, p = .953); and CoNLL scores (rho = -0.120, p = .461).

The resulting output, as shown in Figure 2, indicated that in the post-hoc analysis on measures of consecutive semantic similarity such as mean (U = 165.000, p = .675), standard deviation (U = 133.000, p = .176), maximum (U = 196.500, p = .637), and minimum (U = 196.000, p = .654), there were no statistically significant differences between the schizophrenia and healthy control groups in terms of semantic similarity measures.

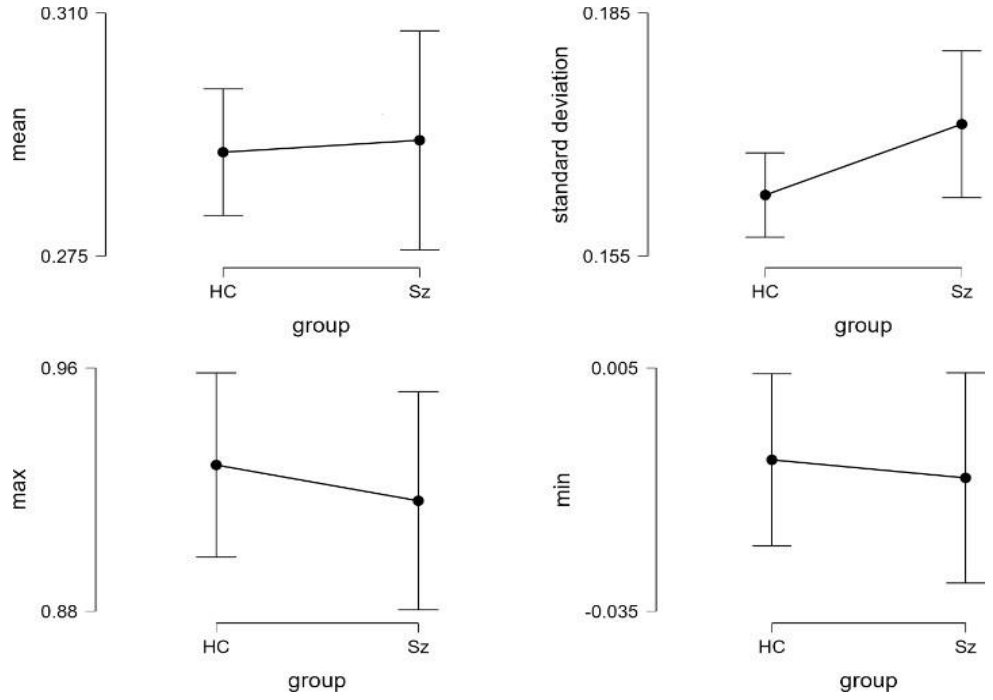


Figure 2: Descriptive plots of semantic similarity using FastText word embeddings for both groups

4. Discussion

Since there were no statistically significant differences in between-group comparisons, the results do not confirm our initial prediction that the Sz samples would perform worse in terms of coreference resolution than the HC samples. The clinical evaluations of conceptual disorganization and formal thought disorder in Sz patients that we compared to the language model's performance also bear this out: the correlations were non-significant. Furthermore, our additional test did not suggest that years of education were a confound in these results. This prompted us to look into the potential contributing elements to this result.

The similarity in coreference resolution performance scores between the Sz and HC groups raises intriguing questions about the capabilities of the language model employed in this study. It is essential to scrutinize whether the model's architecture, training data or underlying algorithms contribute to the observed differences, specifically looking at and taking into account the impact of the language model's training data, which consists of written material like news articles (Schröder et al., 2021), on its performance in the context of the spoken transcripts utilized in this work. In terms of linguistic organization, syntactic complexity, and the existence of conversational signals, spoken and written languages naturally differ from one another. Applying the language model to verbal transcripts, which are characterized by spontaneous speech patterns, disfluencies, and contextual dependencies connected to real-time communication, may cause particular problems as a result of these characteristics. Exploring these potential limitations and biases of the language model becomes crucial to understanding the underlying mechanisms contributing to the homogenous performance. On the other hand, it is important to acknowledge that these scores are nonetheless significant even though they may be lower than those found in English language data (Joshi et al., 2019). This raises the possibility that there may not be any compelling evidence to draw the conclusion that the model's performance is inherently subpar. Additionally, these fairly high coreference resolution results suggest that automated methods, like the language model utilized in this study, may be useful in a clinical setting, particularly when working with data of a similar nature. Despite the drawbacks previously mentioned in relation to the distinctions between spoken and written language, these results suggest that such automated tools can be of significant use, especially when it comes to quantifying data.

Turning to the unique qualities of the sample, particularly the individuals with schizophrenia may have a major impact on the observed similar results. The performance of the

language model was predicted to be correlated with clinical evaluations of conceptual disorganization and formal thought disorder. Thus, low levels of formal thought disorder, or the lack thereof, in the sample may point to substantially retained syntactic structure and coherence, which would increase coreference resolution abilities. Furthermore, the sample's age distribution, with a mean age of 41.5 in the schizophrenia group, may have an impact on the observed likeness in coreference resolution performance. People in this age group may have been taking medication and receiving therapy for a while, which could have resulted in a degree of symptom stability. Taking into account the fact that the mean age of onset for this group is 19.12 years, it is important to consider the implications of early intervention and treatment on the observed variations in coreference resolution performance. This extended period of intervention could have played a role in mitigating the severity of linguistic deficits and enhancing language-related outcomes. Early treatment in schizophrenia patients has been linked to better symptom control and functional results (Amminger et al., 2011), so it's possible that having access to the right interventions at the right time gave these patients the chance to build a foundation for better communication. More severe linguistic deficits, such as issues with coreference resolution, might represent some of the characteristics of early stages of, or acute, psychosis, which may explain that NLP markers are sensitive to them (Morgan et al., 2021; Mackinley et al., 2020). As a result, the observed sample may show weaker language deficiencies, perhaps diminishing the discernible differences in coreference resolution skills compared to people in the initial stages of psychosis.

The hypothesis itself and its implications for the observed data are other components that should be taken into account. Although the predictions indicated that people with schizophrenia would score less well on coreference resolution tasks, as well as having clinical evaluations of conceptual disorganization and formal thought disorder correlate with these lower scores, it is

crucial to assess whether the experimental setup and methods adequately captured the nuances of this language phenomenon. The sample size and period of data analysis were limited due to time constraints and limited resources in the frame of this thesis. As a result, the statistical power of the investigation may have been compromised, limiting the capacity to identify subtle changes in coreference resolution performance between the groups.

Furthermore, it is important to acknowledge that the findings of this study may be influenced by the specific language under investigation, which in this case is German. English has been the primary language of choice for much of the existing research on coreference resolution and language processing in psychiatric populations. This English-centric viewpoint prevents findings from being generalized and may ignore linguistic nuances that exist in other languages, which play a significant role in shaping linguistic organization and cognitive processes, and variations across languages can impact the manifestation of coreference resolution difficulties. Languages differ in their coreference resolution patterns and tactics, according to research in cross-linguistic studies in coreference resolution tasks (Villodre et al., 2013). These particularities could be explained by differences in grammatical constructions, word order, and referential expressions among various linguistic systems.

Because of these surprising findings in coreference resolution, an analysis of semantic similarity capturing the average semantic distances crossed from word to word as the discourse proceeds was undertaken. Again, surprisingly, and in conflict with previous studies (Bedi et al., 2015; Elvevåg et al., 2017; Corcoran et al., 2018; Iter et al., 2018) using semantic similarity metrics, this analysis did not produce significant group differences either. This provides supporting evidence that these previous findings may have limited generalizability across schizophrenia samples and languages – as noted above, Palominos et al. (2023) reported a similar null result for

a Chilean Spanish sample. It is possible that individuals with schizophrenia display patterns of semantic organization that are comparable to those of their healthy counterparts, both in terms of generating and retaining reference over narrative time and retrieving appropriate lexical concepts for this purpose. However, another possibility is that the proxy of lexical meaning that language models such as FastText offer, are not suitable to capture what is deviant in semantic processing in schizophrenia; and that the referential-semantic level as analyzed graph-theoretically in Palominos et al. (2023), may be more promising – including more promising than the coreference methodology employed here.

Another perspective for future research is to broaden the scope of the study to include samples of other clinical populations, such as major depressive disorder or bipolar disorder, in addition to schizophrenia. This comparison method could shed light on whether coreference resolution deficiencies detected by language models are specific to schizophrenia or if they are shared cognitive impairments among diverse psychiatric populations. Future research can provide useful insights into the specificity of coreference resolution deficiencies by integrating a broader range of mental disorders. If these deficiencies are regularly detected across illnesses, it may indicate the presence of common underlying cognitive mechanisms or shared language. If, on the other hand, coreference resolution deficiencies are shown to be more pronounced or distinct in schizophrenia compared to other disorders, this could indicate to schizophrenia's specific cognitive profile. This could also expand beyond reference: a recent study's (Schneider et al., 2023) findings revealed that people with schizophrenia spectrum disorders have lower syntactic complexity than people with major depressive disorder and healthy controls. These results could suggest that syntactic deficits are linked to specific language and cognitive profiles within the schizophrenia spectrum. These syntactic findings highlight the value of conducting studies that

simultaneously look at syntactic and semantic organization. Language and sentence-level meaning are indissolubly linked, according to the interconnectedness of language. As a result, it is unusual to have a favorable outcome on the syntactic side and receive null results on the semantic side. Future research can benefit from including analyses of both syntactic and semantic aspects to better understand the intricate interactions between linguistic structure and meaning in a variety of mental diseases. Such investigations can reveal whether coreference resolution deficits are consistently seen across diseases, pointing to the existence of shared cognitive mechanisms or linguistic impairments. In contrast, it would suggest the existence of a different cognitive profile associated with schizophrenia if coreference resolution deficits are demonstrated to be more pronounced or distinct in schizophrenia compared to other mental illnesses. This thorough investigation can go beyond reference resolution and include a larger understanding of syntactic and semantic deficiencies within the schizophrenia spectrum, helping to characterize language and cognitive profiles in psychiatric populations in a more thorough manner.

5. Conclusions

The study's consistent results of non-significance highlight the complexities of measuring coreference resolution difficulties in people with schizophrenia automatically. While earlier research has revealed that this demographic has language processing issues that are picked up by an advanced computer system, the lack of statistically significant differences calls into question the assumption of a simple failure in coreference resolution on the side of language models. These results call for careful study of the underlying variables that contribute to these findings, as well as additional research to improve our understanding of language processing abilities in automated

language models. In addition to the non-significant differences observed in coreference resolution, the study's post-hoc analysis of semantic similarity also yielded non-significant results. These highlight the complexity of measuring semantic processing changes in schizophrenia patients using automated techniques. In this group, there appears to be a complex interplay between coreference resolution complexity and lexical-semantic coherence that calls for careful future research. Furthermore, these findings bring up crucial questions for the creation and improvement of automated language models. They emphasize how important it is to carry out more crosslinguistic studies to improve our comprehension of language processing capabilities in these models, especially in the setting of intricate linguistic phenomena like coreference resolution and semantic coherence. We can work to enhance automated language models' functionality and applicability in the clinical setting by better comprehending their constraints and nuances.

Future research in this area will be critical in unraveling the complexity of how language models tackle coreference resolution deficiencies in schizophrenia. Given the current study's limitations, which include sample characteristics, data nature, and experimental limits, it is important to address these limitations and refine the procedures used. Larger sample sizes, as well as comparative analyses involving other mental dissorders, can provide useful insights into the specificity and generalizability of the automatization of coreference resolution deficiencies.

References

Alonso-Sánchez, M. F., Ford, S., Mackinley, M., Silva, A. B., Limongi, R., & Palaniyappan, L. (2022). Progressive changes in descriptive discourse in First Episode Schizophrenia: a longitudinal computational semantics study. *Schizophrenia, 8*(1). doi:https://doi.org/10.1038/s41537-022-00246-8

Amminger, G., Henry, L., Harrigan, S., Harris, M., Alvarez-Jimenez, M., Herrman, H., . . . McGorry, P. (2011). Outcome in early-onset schizophrenia revisited: Findings from the Early Psychosis Prevention and Intervention Centre long-term follow-up study. *Schizophrenia Research*. doi:https://doi.org/10.1016/j.schres.2011.06.009.

Andreasen, N. C. (1986). Scale for the Assessment of Thought, Language, and Communication (TLC). *Schizophrenia Bulletin, 12*(3), 473–482. doi:https://doi.org/10.1093/schbul/12.3.473

Bagga, A., & Baldwin, B. (1998). Entity-based cross-document coreferencing using the Vector Space Model. doi:https://doi.org/10.3115/980845.980859

Bedi, G., Carillo, F., Cecchi, G. A., Slezak, D. F., Signam, M., Mota, N. B., . . . Corocan, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youth. *npj Schizophrenia 1*. doi:https://doi.org/10.1038/npjschz.2015.30

Bilgrami, Z. R., Sarac, C., Srivastava, A., Herrera, S. N., Azis, M., Haas, S. S., . . . Corocan, C. M. (2022). Construct validity for computational linguistic metrics in individuals at. *Schizophrenia Research*, 90-96. doi:https://doi.org/10.1016/j.schres.2022.01.019

Çokal, D., Palominos-Flores, C., Yalınçetin, B., Türe-Abacı, Ö., Bora, E., & Hinzen, W. (2022). Referential noun phrases distribute differently in Turkish speakers with schizophrenia. *Schizophrenia Research*. doi:https://doi.org/10.1016/j.schres.2022.06.024

Corcoran, C., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C. C., Javitt, D. C., . . . Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, *17*(1), 67–75. doi:https://doi.org/10.1002/wps.20491

Elvevåg, B., Foltz, P. W., Rosenstein, M., Ferrer-I-Cancho, R., De Deyne, S., Mizraji, E., & Cohen, A. S. (2017). Thoughts About Disordered Thinking: Measuring and Quantifying the Laws of Order and Disorder. *Schizophrenia Bulletin*, *43*(3), 509–513. doi:https://doi.org/10.1093/schbul/sbx040

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L18-1550

Halliday, M. A., & Hasan, R. (1976). *Cohesion in English.* London: Longman.

Hinzen, W., & Sheehan, M. (2013). *The Philosophy of Universal Grammar.* Oxford University Press.

Hitczenko, K. M. (2020). Understanding Language Abnormalities and Associated Clinical Markers in Psychosis: The Promise of Computational Methods. *Schizophrenia Bulletin*, 344–362. doi:https://doi.org/10.1093/schbul/sbaa141

Hitczenko, K., Cowan, H., Mittal, V., & Goldrick, M. (2021). Automated coherence measures fail to index thought disorder in individuals at righ risk for psychosis. *Proceedings of the*

Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access, 129-150. doi:10.18653/v1/2021.clpsych-1.16

Iter, D., Yoon, J. H., & Jurafsky, D. (2018). Automatic Detection of Incoherent Speech for Diagnosing Schizophrenia. doi: https://doi.org/10.18653/v1/w18-0615

Joshi, M. S., Levy, O., Zettlemoyer, L., & Weld, D. S. (2019). BERT for Coreference Resolution: Baselines and Analysis. doi: https://doi.org/10.18653/v1/d19-1588

Just, S., Haegert, E., Kořánová, N., Bröcker, A., Nenchev, I., Funcke, J., . . . Stede, M. (2019). Coherence models in schizophrenia. doi:https://doi.org/10.18653/v1/w19-3015

Kircher, T., Wöhr, M., Nenadic, I., Schwarting, R. K., Schratt, G., Alferink, J., . . . Krug, A. (2019). Neurobiology of the major psychoses: a translational perspective on brain structure and function—the FOR2107 consortium. *European Archives of Psychiatry and Clinical Neuroscience*, *269*(8), 949-962. doi:https://doi.org/10.1007/s00406-018-0943-x

Krug, M., Weimer, L., Reger, I., Macharowsky, L., Feldhaus, S., Puppe, F., & Jannidis, F. (2018). Description of a corpus of character references in German novels-DROC [Deutsches ROman Corpus]. *DARIAH-DE Working Papers, 27*, 1-16.

Liddle, P. F., Ngan, E. T., Caissie, S. L., Anderson, C., Bates, A. T., Quested, D., . . . Weg, R. (2002). Thought and Language Index: an instrument for assessing thought and language in schizophrenia. *British Journal of Psychiatry, 181*(4), 326–330. doi:https://doi.org/10.1192/bjp.181.4.326

Luo, X. (2005). On coreference resolution performance metrics. doi:https://doi.org/10.3115/1220575.1220579

Mackinley, M., Chan, J., Ke, H., Dempster, K., & Palaniyappan, L. (2020). Linguistic determinants of formal thought disorder in first episode psychosis. *Early Intervention in Psychiatry, 15*(2), 344–351. doi:https://doi.org/10.1111/eip.12948

Morgan, S. E., Diederen, K., Vértes, P., Ip, S. H., Wang, B., Thompson, B., . . . McGuire, P. (2021). Natural Language Processing markers in first episode psychosisand people at clinical high-risk. *Transl Psychiatry*. doi:10.1038/s41398-021-01722-y

Murray, H. A. (1943). *Thematic apperception test.* Harvard University Press.

Naumann, K. (2006). Manual for the Annotation of in-document Referential Relations. https://www.lingexp.uni-tuebingen.de/sfb441/a1/Publikationen/tuebadz_relations_man.pdf.

Palominos, C., Figueroa-Barra, A., & Hinzen, W. (2023). Coreference Delays in Psychotic Discourse: Widening the Temporal Window. *Schizophrenia Bulletin*, 153-162. doi:https://doi.org/10.1093/schbul/sbac102

Parola, A., Lin, J., Simonsen, A., Bliksted, V., Zhou, Y., Wang, H., . . . Fusaroli, R. (2022). Speech disturbances in schizophrenia: assessing cross-linguistic generalizability of. *Schizophrenia Research*. doi:https://doi.org/10.1016/j.schres.2022.07.002

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., & Zhang, Y. (2012). CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. *Empirical Methods in Natural Language Processing*, (pp. 1–40). doi:https://aclanthology.org/W12-4501/

Recasens, M. V., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., . . . Versley, Y. (2010). SemEval-2010 Task 1: Coreference Resolution in Multiple Languages. *Meeting of the Association for Computational Linguistics*, (pp. 1–8). Retrieved from https://www.aclweb.org/anthology/S10-1001.pdf

Reiter, N. (2018). CorefAnnotator - A New Annotation Tool for Entity References. *Abstracts of EADH: Data in the Digital Humanities.*

Rochester, S., & Martin, J. (1979). *Crazy Talk: A Study of the Discourse of Schizophrenic Speakers.* Springer New York, NY.

Sarzyńska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., & Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Res.* doi:https://doi.org/10.1016/j.psychres.2021.114135

Schneider, K., Leinweber, K., & Jamalabadi, H. e. (2023). Syntactic complexity and diversity of spontaneous speech production in schizophrenia spectrum and major depressive disorders. *Schizophr, 9*(35). doi:https://doi.org/10.1038/s41537-023-00359-8

Schröder, F., Hatzel, H. O., & Biemann, C. (2021). Neural End-to-end Coreference Resolution for German in Different Domains. *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, (pp. 170–181). doi:https://aclanthology.org/2021.konvens-1.15/

Sevilla, G., Rosselló, J., Salvador, R., Sarró, S., López-Araquistain, L., Pomarol-Clotet, E., & Hinzen, W. (2018). Deficits in nominal reference identify thought disordered speech in a narrative production task. *PLOS ONE*. doi:https://doi.org/10.1371/journal.pone.0201545

Tan, E. J., Sommer I., E. C., & Palaniyappan, L. (2023). Language and Psychosis: Tightening the Association. *Schizophrenia Bulletin*. doi:10.1093/schbul/sbac211

Tang, S. X., Kriz, R., Cho, S., Park, S. J., Harowitz, J., Gur, R. E., . . . Liberman, M. (2021). Natural language processing methods are sensitive to subclinical linguistic differences in schizophrenia spectrum. *Npj Schizophrenia, 7*(1). doi:https://doi.org/10.1038/s41537-021-00154-3

Telljohann, H., Hinrichs, E., Kübler, S., Zinsmeister, H., & Beck, K. (2012). Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z).

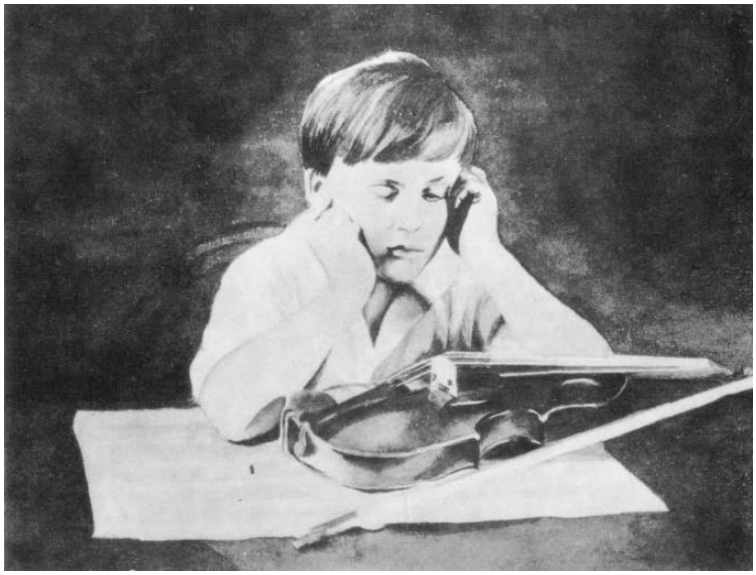
Vilain, M., Burger, J. D., Aberdeen, J. S., Connolly, D., & Hirschman, L. (1995). A model-theoretic coreference scoring scheme. doi:https://doi.org/10.3115/1072399.1072405

Villodre, L., Recasens, M., & Sapena, E. (2013). Coreference resolution: an empirical study based on SemEval-2010 shared Task 1. *Language Resources and Evaluation*. doi:https://doi.org/10.1007/s10579-012-9194-z.

Appendix

Appendix A

A series of images, taken from the Thematic Apperception Test (Murray, 1943), were used to prompt the speech samples. Participants are asked to create stories based on what they see in each image, including details about the character's thoughts, feelings, and motives.



Picture 1

Picture 2

Picture 4BF

Picture 6

Appendix B

Transcription conventions used to manually obtain the speech samples.

Transcription structure:

- transcript header
    - participant ID
    - recording date
    - duration of recording
    - transcriber ID
    - transcription date
    - number of reinforcements given by the interviewer
- general conditions:
    - one transcription file of 4 pictures per respondent. Designation: 0891_2_raw
    - segmentation of utterances by perceptible pauses and intonation, according to conventional descriptions of sentence boundaries (e.g., content)
- possible utterances: grammatical sentences, interjections, fillers, fragments. If an utterance is continued after a pause, it is counted as part of the preceding utterance "(.)", or "(?)" for questions. Hesitations are "lexical" (okay, yes, no, stop) and "non-lexical" filler words (mostly, uh, um, huh). Sentence-final conjunctions/non-lexical filler words at the end of an utterance are counted towards the next utterance. Lexical filler words (ok), which are often found at the beginning of an utterance, are included in the following utterance. Revisions of utterances are also transcribed (included in the original utterance) and analyzed, except when completely broken off and introduced in a new utterance.
- Conventions

- Pauses:
  - (.) - micro pause
  - (-), (--), (---) - short, medium, longer pauses of approx. 0.25 - 0.75 seconds, up to approx. 1 second. Pauses longer than approx. 1 second (e.g., 2.85) are indicated by corresponding decimals after the dot.
  - segments: uh, öh, etc. are transcribed as :, ::, ::: signifying the elongation of the word, depending on the duration
  - laughter:
    - so(h)o – laughing particles when speaking
    - haha hehe hihi – syllabic laughter
    - [laughs] – laughter
  - other:
    - [coughs] – para- and extra-linguistic actions
    - () – unintelligible utterance
    - (XXXXX) – presumed utterance
    - (…) – omission in the transcript