Combining automatic speech recognition with semantic natural language processing in schizophrenia

Ciampelli, S.; Voppel, A. E.; de Boer, J. N.; Koops, S.; Sommer, I. E.C.

Published in:
Psychiatry Research

DOI:
10.1016/j.psychres.2023.115252

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Combining automatic speech recognition with semantic natural language processing in schizophrenia

S. Ciampelli[a,*], A.E. Voppel[a], J.N. de Boer[a,b], S. Koops[a], I.E.C. Sommer[a]

[a] Department of Psychiatry, University Medical Center Groningen, Groningen, the Netherlands
[b] Department of Psychiatry, Department of Intensive Care Medicine, UMC Brain Center, University Medical Center Utrecht, Utrecht, the Netherlands

ARTICLE INFO

Keywords:
Speech analysis
Psychosis
Word error rate
Semantic similarity

ABSTRACT

Natural language processing (NLP) tools are increasingly used to quantify semantic anomalies in schizophrenia. Automatic speech recognition (ASR) technology, if robust enough, could significantly speed up the NLP research process. In this study, we assessed the performance of a state-of-the-art ASR tool and its impact on diagnostic classification accuracy based on a NLP model. We compared ASR to human transcripts quantitatively (Word Error Rate (WER)) and qualitatively by analyzing error type and position. Subsequently, we evaluated the impact of ASR on classification accuracy using semantic similarity measures. Two random forest classifiers were trained with similarity measures derived from automatic and manual transcriptions, and their performance was compared. The ASR tool had a mean WER of 30.4%. Pronouns and words in sentence-final position had the highest WERs. The classification accuracy was 76.7% (sensitivity 70%; specificity 86%) using automated transcriptions and 79.8% (sensitivity 75%; specificity 86%) for manual transcriptions. The difference in performance between the models was not significant. These findings demonstrate that using ASR for semantic analysis is associated with only a small decrease in accuracy in classifying schizophrenia, compared to manual transcripts. Thus, combining ASR technology with semantic NLP models qualifies as a robust and efficient method for diagnosing schizophrenia.

1. Introduction

Recently, natural language processing (NLP) models have been increasingly used to quantify speech incoherence in schizophrenia-spectrum disorders (SSD) (Corcoran et al., 2018; Elvevåg et al., 2007; Tang et al., 2021; Voppel et al., 2021). Although these methods are highly accurate and thus hold extensive potential for future clinical applications (Corcoran and Cecchi, 2020; de Boer et al., 2018), to date, participant interviews have been manually transcribed, which is a dedicated, extremely time-consuming, and hence high-cost process.

Automatic speech recognition (ASR) technology allows for the quick conversion of speech signal into text with human-like performance level in optimal settings (Kodish-Wachs et al., 2018). ASR could therefore greatly speed up the NLP research process and pave the way to developing actual clinical tools. Additionally, ASR technology would facilitate the remote assessment of patient symptoms through language and speech data, potentially increasing the frequency of monitoring and enabling more timely interventions (Insel, 2017).

Although ASR can reach human-like performance levels in optimal settings such as human-machine interface tasks (e.g., automatic call processing) or prepared speech (e.g., political speeches), the automatic transcription of conversational speech has generated less satisfactory results (Chiu et al., 2018; Kodish-Wachs et al., 2018). These challenges are due to various factors such as background noise, type and method of recording device, speakers' accents and disfluencies (e.g., false starts, repetitions, filled pauses) (Radha and Vimala, 2012). In addition, acoustic differences between male and female voices are likely to affect speech intelligibility and, in turn, ASR performance. For example, women were shown to have a greater fundamental frequency range than men, which was correlated with higher overall speech intelligibility (Bradlow, Torretta & Pisoni, 1996). Similarly, sex-specific ASR yielded better recognition results for women as compared to men (Adda-Decker and Lamel, 2005). The accuracy of ASR may be further reduced when applied to conversational speech from individuals with SSD because they may speak at a slower rate, have reduced intonation, and produce longer pauses (Parola et al., 2020).

Previous research showed that ASR can be used for the automated transcription of oral retellings of stories from individuals with SSD,

* Corresponding author at: Department of Psychiatry, University Medical Center Groningen, Hanzeplein 1, 9713 GZ Groningen, the Netherlands.
E-mail address: s.ciampelli@umcg.nl (S. Ciampelli).

https://doi.org/10.1016/j.psychres.2023.115252
Received 16 January 2023; Received in revised form 21 April 2023; Accepted 13 May 2023
Available online 16 May 2023

substance abuse disorders, and affective disorders (Chandler et al., 2019; Holmlund et al., 2020). Although ASR transcripts in these studies had an overall Word Error Rate (WER) ranging between 10% and 23%, these errors did not affect the performance of the final rating and classification models, which remained highly correlated with results obtained from manually transcribed recordings. Similarly, despite an average WER of 36%, most ASR-derived coherence measures from oral descriptions of participants' auditory hallucinations retained a moderate to high correlation with manual-transcript derived measures of coherence (Xu et al., 2022). A study by Corcoran & Cecchi (2020) suggested that up to a 25% WER in automated transcriptions can be tolerated for language analyses in psychiatry applications, because in contrast to other fields such as finance, they rely on the statistical properties of whole samples, hence higher error rates have a lower impact on the final outcome.

However, despite these results indicating that ASR errors do not affect the final performance of language models, a more in-depth examination of the type of ASR errors is essential to enhance the explainability of these findings and elucidate what factors influence specific classification tasks. When we take into consideration models of semantic similarity, we might indeed expect that not all ASR errors are relevant in the same manner. For example, in the sentence 'She was happy to win the prize' the function words 'she','was','to','the' play grammatical roles, while the content words 'happy','win','prize' convey the most important semantic information, i.e., the meaning; which is the focus of semantic NLP research.

This knowledge will aid to detailed understanding of the current limits of ASR in speech analysis and pave the way for implementing ASR applications in clinical settings.

In this work, we employ both quantitative and qualitative methods to investigate the performance of a state-of-the-art ASR tool and its influence on classification accuracy when applied to the semi-spontaneous speech of patients with SSD and healthy controls. First, we assess the performance of the ASR tool compared to that of "gold standard" human transcriptions quantitatively (using the WER) and qualitatively by analyzing the most frequently mistaken words, their grammatical parts of speech (POS), and their linguistic position. Subsequently, we evaluate the impact of ASR on diagnostic classification accuracy in SSD. In previous work by our group, we used semantic similarity measures on manual transcripts to classify patients with SSD and healthy controls (Voppel et al., 2021). Here, we replicate these analyses on a larger sample, and compare the classification accuracy of manual transcripts to that of ASR-derived transcripts. Because previous research suggested sex to have a major impact on ASR accuracy, we compare the performance in men and women separately.

2. Methods

2.1. Subjects

Speech recordings were obtained from a total of 163 participants, namely 93 patients with SSD and 70 healthy controls between 2015 and 2021 at the University Medical Center Utrecht. To be eligible for the study, participants had to be at least 18 years old, speak Dutch as their native language, and have no uncorrected hearing impairment or speech disorder, such as stuttering. Bilingual speakers were included if Dutch was (one of) their dominant language(s). Patients were included if they met criteria for a DSM-IV diagnosis of: 295.x (schizophrenia, schizophreniform disorder, schizoaffective disorder) or 298.9 (psychotic disorder not otherwise specified). The diagnosis was established in all patients by their treating psychiatrist and was confirmed by a trained researcher using the Comprehensive Assessment of Symptoms and History interview (CASH) (Andreasen et al., 1992) or the Mini-International Interview (MINI) (Overbeek and Schruers, 2019). The severity of psychopathology was measured in patients with the Positive and Negative Syndrome Scale (PANSS) (Kay et al., 1987). Participants in the control group were included provided that they had no mental health complaints and no family history of psychotic symptoms. Past episodes of depression or anxiety disorders in full remission were not an exclusion criterion. Prior to enrollment, all individuals provided written informed consent. Participants received a small monetary award for participation (10 euros). The study was approved by the ethical review board of the University Medical Center Utrecht.

2.2. Language data acquisition

Spontaneous speech was elicited by a trained researcher using a semi-structured interview of approximately 15 min. To avoid biases in participants' responses, subjects were told that the study concerned the evaluation of their spoken language only at the end of the interview process. The interview consisted of open-ended questions about 'neutral', everyday topics (e.g., 'Which Dutch TV shows do you often watch? And which do you hate? Why?'); topics that could trigger an emotional response (e.g., 'health', 'quality of life') were avoided to control for possible variations in language caused by the topics covered. The questions were asked in a semi-randomized order; if a patient refused to answer a question, the interviewer would ask the following question. This interview procedure was applied to both control and patient participants. For a full list of questions, see Table S1 in the supplementary material.

2.3. Speech recording, pre-processing, and manual transcription

Two AKG-C544l head-worn cardioid microphones were used to record the interview, one for the participant and one for the interviewer. To minimize the alteration of vocal loudness, the distance between mouth and microphone was kept as stable as possible (2 cm). Speech was digitally recorded onto a Tascam DR40 solid-state recording device at a sampling rate of 44,100 kHz with 16-bit quantization, using two different channels. The PRAAT software (Boersma, 2001) was used to remove cross-talk (i.e., the interviewer's speech on the participants' audio channel). For this purpose, the function 'annotate to text grid silences' was applied to the interviewer's channel with the following settings: minimum pitch 100 Hz, time step 0.0, silence threshold −30.0 dB, minimum silence duration 1.0 s, minimum sounding duration 0.1 s. This function selected all speech segments where the interviewer was silent. These speech segments were subsequently selected on the participant's channel and concatenated into a new audio file that only included the participant's speech signal.

The digital recordings were manually transcribed by researchers who were blind to the participant condition using the CLAN—CHILDES transcription guidelines (Brundage and Bernstein Ratner, 2018; MacWhinney, 2000). Once transcribed, interviews were converted to plaintext (e.g., no punctuations, no capitalization). Hesitations (e.g., uhm) and interjections (e.g., yes/no) were discarded to facilitate subsequent comparison with ASR-derived transcripts. Contracted words were converted into their full form (e.g., m'n'-mijn).

2.4. Automatic speech recognition (ASR)

ASR was applied to the audio files using the Kaldi NL Speech Recognition Toolkit (version 0.4.1) (Povey et al., 2011). Kaldi NL is a speech recognition system trained on the Corpus Gesproken Nederlands (CGN; Corpus Spoken Dutch) (van Eerten, 2007) and it is composed of an acoustic model, a lexicon, and a language model. Since domain-specific ASR tools have been shown to perform better than generic ones (Dingliwal et al., 2021), we used the KALDI model based on daily conversations, which matched the specific domain of our audio data (i.e., neutral-topic interviews). In order to determine whether improving audio quality led to better ASR performance, we tested noise reduction, sound intensity normalization, and alternate splitting on a subsample of the audiofiles using PRAAT (Boersma, 2001). However, none of these pre-processing steps improved the word error rate and therefore we used the original unmodified files as input for the ASR analysis (see supplementary methods). The ASR transcripts were subsequently cleaned to plaintext (e.g., no punctuations, no capitalization). Hesitations and interjections were removed. Contracted words were converted into their full form (e.g., m'n'-mijn).

2.5. Quantitative error analysis

The Word Error Rate (WER) was used to quantitatively assess the performance of the ASR system by comparing automatic transcriptions with manual transcriptions, which we considered the "gold standard". The WER was calculated in Python (version 3.9) (Van Rossum and Drake, 2009) with the "Symbl.ai" utility (https://github.com/symblai/speech-recognition-evaluation), by computing a Levenshtein alignment between two-word sequences i.e., the minimum amount of edit operations that must be performed on a string to obtain a target string. To obtain the WER, we divided the number of insertions, substitutions, and deletions by the total number of words per interview.

2.6. Qualitative error analysis

We qualitatively characterized ASR errors based on grammatical part-of-speech classes (POS), most frequently misrecognized words, and their linguistic position within a sentence. Part-of-speech labels were obtained in Python (version 3.9) (Van Rossum and Drake, 2009) using the CLAM Client API version of Frog (version 0.24) (Van der Sloot et al., 2018). The part-of-speech tagger uses 12 POS categories, which were derived from the 316-tag set of the Corpus Gesproken Nederlands (Table S1) (van Eynde, 2004). Subsequently, we investigated the relative position in a sentence for each ASR error by dividing the index of each error in a sentence by the total sentence length.

2.7. Classification performance using semantic similarity analysis

Semantic similarity analysis was performed separately for automated transcripts and manual transcripts. Semantic similarity was computed using a 300-dimensional Word2vec model (Mikolov et al., 2013), which was trained on a large text corpus from the Corpus Gesproken Nederland; words were then converted into vectors of numbers (i.e., word embeddings) (for a more extensive description of how the model was trained, see Voppel et al., 2021). Once words were vectorized, similarity measures between vectors were calculated. We employed the similarity measures that were found most informative in earlier analyses, which included minimum similarity, variance of similarity, mean similarity, and maximum similarity (Corcoran et al., 2018; de Boer et al., 2018; Voppel et al., 2021).

Following these studies, a moving window size approach of 5 to 10 words was applied. This meant that a single similarity value per window was generated, by calculating and averaging the similarity between a word and all of its nearby words inside that window. The window was then slid to the next word, a new similarity value was calculated, and so on until the interview was complete.

2.8. Statistical analyses

Statistical analyses were performed with R (version 4.1.2) (R Core Team, 2020). Two-tailed independent t-tests were performed to test for group differences in age, years of education, and parental years of education. A Chi-Square goodness of fit test was carried out to test whether the proportion of males and females was equal in the patient and in the control group. A non-parametric Kruskal-Wallis one-way analysis of variance (ANOVA) was conducted to evaluate differences in WER between groups and sexes. Wilcoxon's approach (non-parametric) was used to test if the difference in effect sizes of WER between the groups was significant (Tomczak and Tomczak, 2014). Post-hoc Mann-Whitney U tests were used to compare word deletion rate, word substitution rate, and word insertion rate between groups and sexes. Paired samples Wilcoxon's signed-rank tests were performed to compare the distributions of semantic similarity metrics between ASR-derived transcripts vs. manual transcripts. Non-parametric two-tailed Spearman bivariate correlations were performed to investigate associations between WER, age, education, and the severity of patients' psychotic symptoms (PANSS). To control for multiple comparisons in correlation tests, False Discovery Rate (FDR) was employed. Alpha was set to 0.05 for all analyses.

Two Random Forest Classifiers (RFC) (Breiman, 2001) were built to discriminate controls from patients with SSD based on similarity measures derived from automatic and manual transcripts (Fig. 1). During model training 10-fold cross-validation was performed, which randomly divides the data set into 10 separate folds. 9 subsamples of the data per fold were used for training and 1 for testing. This process was iterated 10 times, so that the model was tested each time on a different data subsample. Downsampling was applied to ensure equal distribution between classes (patients vs. controls) per fold. We calculated Gini importance scores to compare the relative ranking of semantic similarity measures in the ASR and manual classifiers. McNemar's test was performed to verify whether the diagnostic accuracy of the two classifiers significantly differed based on the type of transcript used.

3. Results

3.1. Demographics

Clinical and demographic information is shown in Table 1. Patients and controls differed in their education level, although their parental educational level did not differ significantly. This difference was expected, as psychosis frequently develops throughout scholastic years and may therefore impact (higher) educational achievements. Patients had an average PANSS total score of 51.1 (with a standard deviation of 14.5), indicating that most of them were in remission.

3.2. Quantitative error analysis

The ASR tool had an overall WER of 30.4%. The Kruskal-Wallis test revealed that the WER significantly differed between groups (H(1)= 7.438, p=.006), with patients having a WER of 32.3% and controls of 27.8% (Fig. 2). The difference in effect size between the groups was estimated to be 0.22, indicating a significant difference in effect size since the confidence intervals calculated with the bootstrap method (1000 iterations) did not include the value 0 [0.08; 0.36]. Automated transcripts of patients with SSD had a higher rate of word deletion and substitution than controls', while the rate of word insertion did not differ between the two groups (Table S3).

3.2. Association with demographics and symptomatology

No significant association was found between WER and participants' age, education, and parental education (all p>.05). After FDR correction, correlational analyses between WER and psychotic symptoms in patients with SSD indicated that the WER was significantly associated with PANSS total, general, positive, and negative subscales (Table 2).

Analysis of sex differences in WER (all men compared to all women) indicated that sex significantly impacted WER (H(1) = 19.019, p=.009), with men reaching a WER of 32.7% and women a WER of 25.5% (Fig. 3). Post-hoc analyses showed that men had a greater word deletion rate (W = 3846, p<.001) and a higher word substitution rate (W = 4403, p=.002) than women, with no significant difference in the rate of word insertion (W = 3170.5, p=.366). Analyses of sex differences in WER within the patient and the control group revealed that sex had a significant effect on the WER in transcripts of both the patients (H(1) = 11.307, p<.001) and the controls (H(1) = 7.742, p=.005). Healthy men

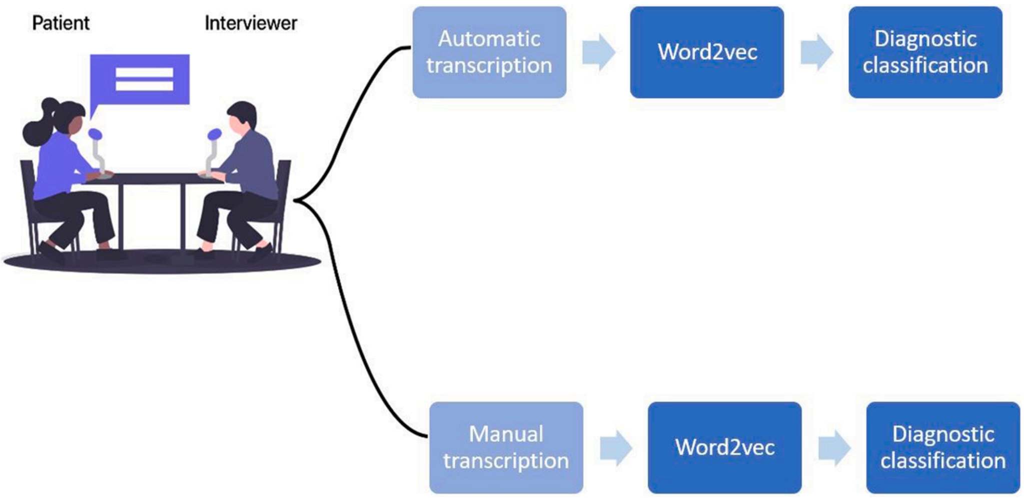

Fig. 1. Overview of the methodological pipeline.

Table 1
Demographic characteristics of patients with SSD and healthy controls.

	SSD patients (*n* = 93)	Healthy Controls (*n* = 70)	Test statistics
Age (years)	33.8 ± 12.77	35.7 ± 15.29	*t* = 0.848, *p* = .398
Female sex, n (%)	28 (30.11%)	25 (35.71%)	χ2 = 0.843, *p* = .359
Education (years)	12.8 ± 2.41	14.7 ± 2.25	*t* = 5.087, *p* = .009
Parental Education (years)	12.3 ± 2.86	12.2 ± 3.09	*t* = 0.028, *p* = .978
Illness duration (years)	8.5 ± 10.76		
Diagnosis			
Psychosis NOS	32(34%)		
Schizoaffective disorder	19(20%)		
Schizophrenia	40(43%)		
Schizophreniform Disorder	2(2%)		
PANSS total	51.1 ± 14.53		
Positive	11.8 ± 4.67		
Negative	12.8 ± 4.84		
General	26.5 ± 7.99		

Legend. Reported values are means ± SD or n (%). *N*=sample size, SD=standard deviation, PANSS=positive and negative syndrome scale, NOS=not otherwise specified.
Indicates significance at the level *p*<.01.

had a WER of 29.7% and healthy women a WER of 24.4%; the difference in mean WER between the sexes was larger in the patients, with male patients reaching a WER of 34.8% and female patients a WER of 26.5%. Further post-hoc tests revealed that transcripts of male patients with SSD had a significantly higher word deletion rate (*W* = 1258, *p*=.004) and a higher word substitution rate (*W* = 1391, *p*=.039) compared to transcripts of female patients with SSD, while there was no difference in word insertion rate (*W* = 862, *p*=.691). In the control group, transcripts of healthy men contained more word substitution errors (*W* = 829, *p*=<0.001) than healthy women, whereas no difference in word insertions (*W* = 643, *p*=.098) and deletions (*W* = 629, *p*=.142) was found. For a more in-depth examination of the effect of sex on WER, see supplemental methods & results.

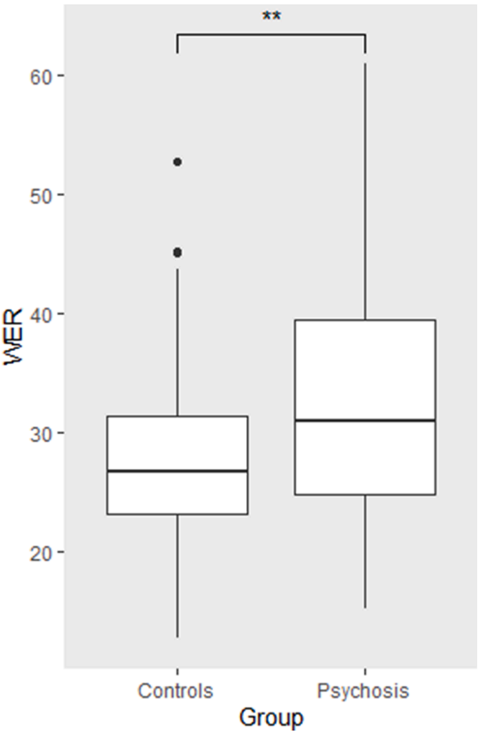

Fig. 2. Word Error Rate (WER) in transcripts of healthy controls and patients with SSD.
Legend. ** Indicates significance at the level *p*<.010.

3.3. Qualitative error analysis

Grammatical analysis of ASR errors revealed that, out of the total number of errors, pronouns were the most mistaken part-of-speech class (26.1%), followed by verbs (18.3%), adverbs (15.4%), and common nouns (12.7%). Adjectives, prepositions, conjunctions, and articles were recognized by the ASR with similar accuracy, with respectively 7.6%, 6.8%, 6.1, and 5.1% error rates. Errors in proper nouns accounted for 1.2% of the total error rate, while numerals were mistaken with a 0.7% error rate. Compared to controls, patients' ASR transcripts had a significantly higher error rate for verbs (*p*=.045). When examining the top 20 errors in the automated transcripts of individuals with SSD, 'I/Ik' was found to be the most frequently misrecognized word, followed by 'that/dat,die', 'the/het/de', and 'and/en' (Fig. 4). A similar pattern was observed in the ASR-derived transcripts of healthy participants (Figure S1). Analysis of ASR errors' linguistic position revealed that substitutions, deletions, and insertions were more likely to occur at the

Table 2
Correlations between quantitative error metrics and symptom severity in SSD.

	PANSS Negative	PANSS Positive	PANSS General	PANSS Total
Word error rate (WER)	0.295**	0.326*	0.320**	0.412**
Word insertion rate (WIR)	0.154	0.191	0.076	0.154
Word deletion rate (WDR)	0.265*	0.286*	0.328**	0.333**
Word substitution rate (WSR)	0.303*	0.310**	0.353**	0.395*

Correlation coefficients are Spearman's rho. Legend. PANSS=Positive and Negative Syndrome Scale.

* Indicates significance at the level of $p < 0.050$,.

** $p < 0.010$.

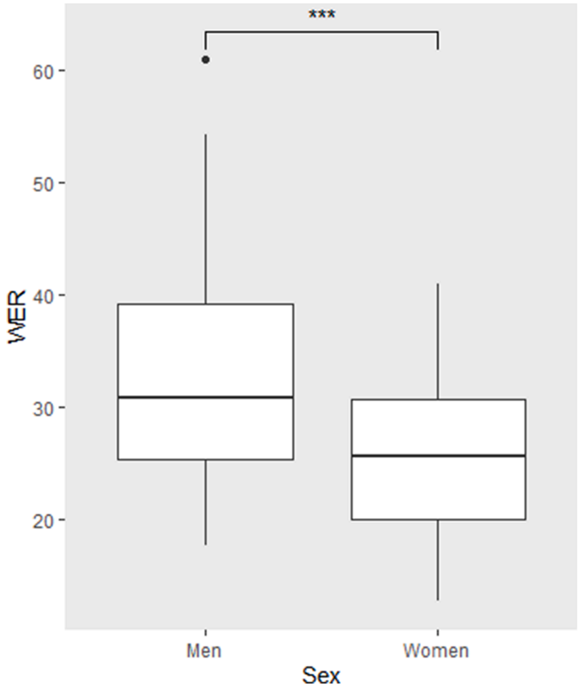

Fig. 3. Word Error Rate (WER) in transcripts of men and women. Legend. *** Indicates significance at the level $p < .001$.

end of a sentence. The probability of an error happening at the start of a sentence was the lowest (Figure S2).

3.4. *Semantic similarity analysis and classification*

The RFC distinguished patients with SSD from controls with a mean accuracy of 76.7% (sensitivity 70%; specificity 86%) using automated transcriptions and 79.8% (sensitivity 75%; specificity 86%) based on manual transcriptions. The area under the curve-receiver operating characteristic (AUC-ROC) was 0.80 for the manual RFC, with a 0.73–0.87 95% confidence interval; and 0.79 for the automatic RFC, with a 0.72–0.86 95% confidence interval. McNemar's test revealed that the two classifiers did not statistically differ in their performance (p=.791), which indicates that they had a similar proportion and type of errors.

The distribution of semantic similarity measures in a sentence-length window of 5 words significantly differed between ASR transcripts and manual transcripts, including mean similarity (V = 89, p=<0.001), maximum similarity (V = 4500.5, p=.021), variance similarity (V = 3950, p=<0.001), and minimum similarity (V = 3287, p=<0.001). Similar results were found for the distribution of similarity measures in windows across the 6–10 word range (Supplementary Table S4). The manual and ASR classifiers shared 80% of the top 10 semantic similarity measures, indicating that differences in distribution only partially affected how useful similarity features were at predicting a schizophrenia diagnosis (Fig. 5).

4. Discussion

The aim of the current study was to quantitatively and qualitatively assess the performance of a state-of-the-art ASR tool and its impact on diagnostic classification in SSD using a semantic NLP model. Automated transcripts could be used to classify patients with SSD with an accuracy of 77%, compared to the 80% accuracy achieved with manual transcripts. When comparing the models, this drop in accuracy was not significant. This indicates that despite relatively high WERs, the final classification accuracy of the semantic NLP model remained high, which could be explained by the fact that most ASR errors occurred in function words (e.g., pronouns), whereas semantic NLP models rely more on content words (e.g., nouns) when calculating coherence measures.

4.1. *Quantitative error analysis*

In this study, we showed that with ASR there was an overall WER of 30.4%, with poorer performance in patients with SSD (32.3%) than in healthy controls (27.8%) (Holmlund et al., 2020; Moro-Velazquez et al., 2019). The higher WERs in patients' transcriptions were due to a higher word deletion and substitution rate, which may be associated with phonetic-acoustic differences in the speech of patients with SSD. Compared to controls, patients with SSD tend to speak less, at a slower rate, with longer pauses, and less pitch variability, factors that may cause the ASR to repeatedly miss or misrecognize words (de Boer et al., 2021; Martínez-Sánchez et al., 2015). Furthermore, our study supports previous research showing a lower ASR error rate in healthy women compared to healthy men (Adda-Decker and Lamel, 2005), and extends these results to speech from patients with SSD. In healthy individuals, the differences in ASR results may be explained by the fact that women have a more standard-like pronunciation compared to men, who tend to articulate less carefully and have a higher number of pauses and repetitions (Adda-Decker and Lamel, 2005). On the other hand, the larger sex differences in WER found among patients with SSD relative to controls may be related to sex differences in SSD symptoms. Men with SSD have on average more negative symptoms than women with SSD (Barajas et al., 2010), which have a disadvantageous impact on speech (Tahir et al., 2019), possibly resulting in higher WERs. Indeed, (negative) symptom severity of both male and female patients correlated significantly with the WER. Overall, these results confirm our initial hypothesis and results of prior research indicating that patients' and men's speech are less accurately recognized by the ASR tool than healthy controls' and women's speech, respectively.

4.2. *Qualitative error analysis*

Pronouns were the most difficult part of speech to recognize by the ASR tool. A closer analysis of the 20 most mistaken words indicated that, in both patients and controls' ASR transcriptions, these errors all belong to the class of function words, including pronouns (e.g., I/Ik, my/mijn), articles (e.g., the/het,de, a/een), and auxiliary verbs (e.g., is/is, was/was). We therefore extended previous findings showing that function words were more likely to be misrecognized by ASR tools than content words (Goryainova et al., 2014; Santiago et al., 2015) to automated transcripts from patients with SSD. One possible explanation for this result is that, with Dutch being a stress-timed language, content words are typically stressed in speech whereas function words are often unstressed and pronounced in a contracted form, making the latter more difficult for the ASR's acoustic model to recognize (Cutler and Foss,

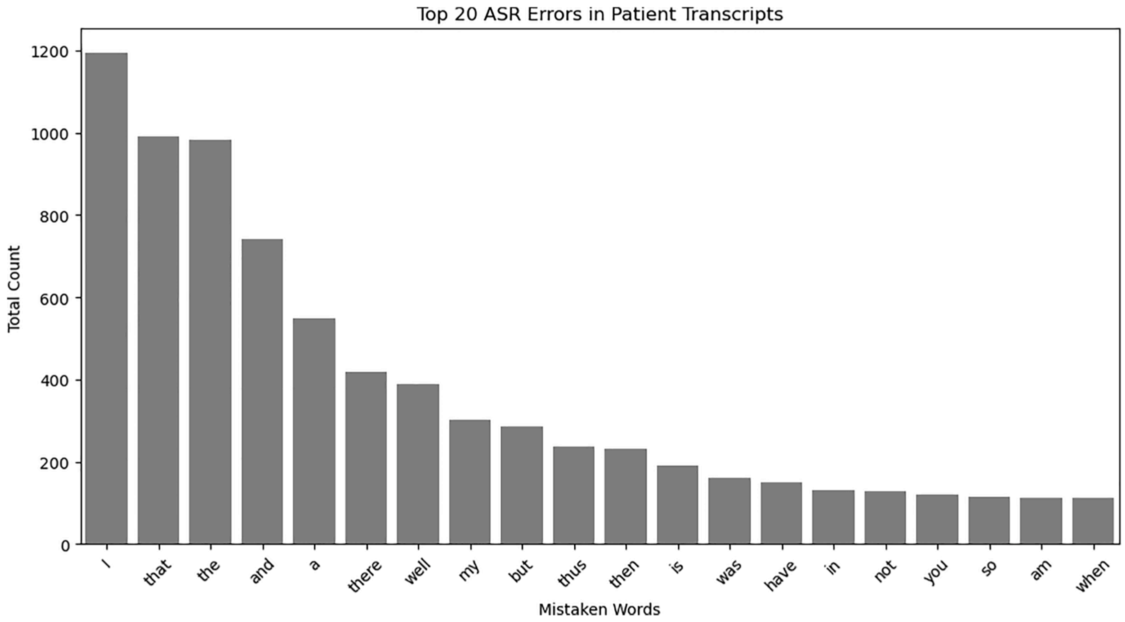

Fig. 4. Top 20 ASR errors in the transcripts of patients with SSD.
Dutch translation: ‘I-Ik’, ’that-die,dat’, ’the-het,de’, ’and-en’, ’a-een’, ’there-er,daar’, ’well-wel’, ’my-mijn’, ’but-maar’, ’thus-dus’, ’then-dan’, ’is-is’, ’was-was’, ’have-heb’, ’in-in’, ’not-niet’, ’you-je’, ’so-zo’, ’am-ben’, ’when-toen’.

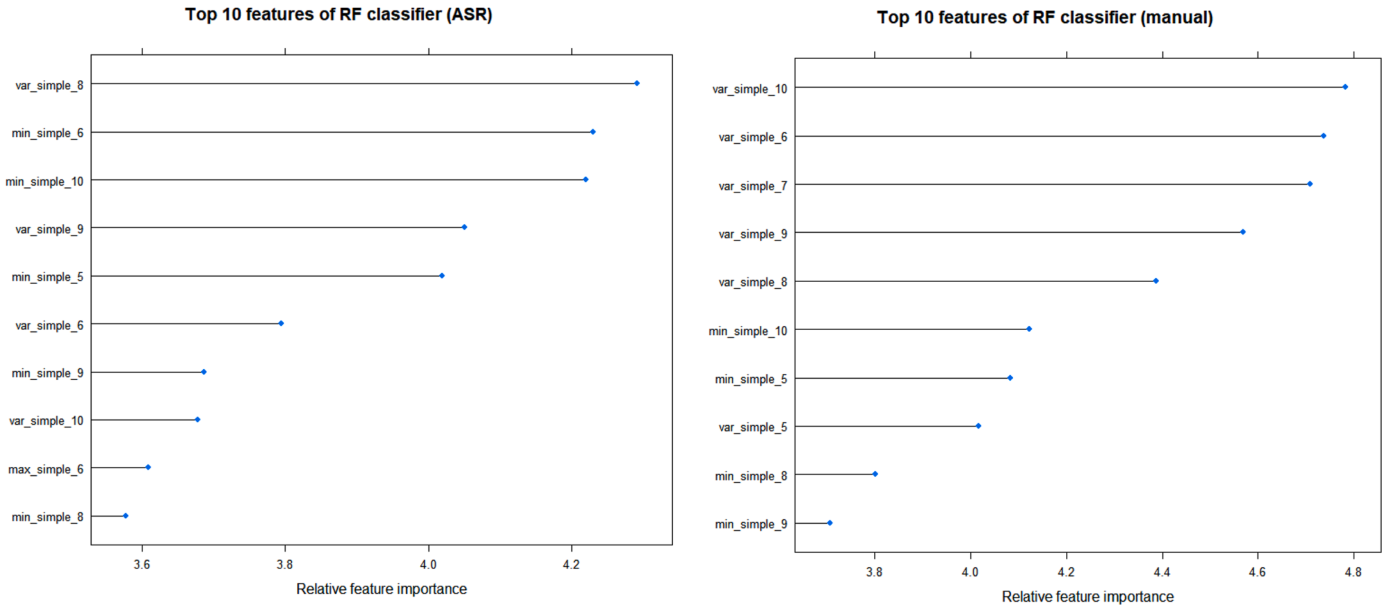

Fig. 5. Top 10 semantic similarity features of Random Forest Classifiers trained on ASR-derived transcripts and manual transcripts.

1977). Moreover, function words are usually short and subject to co-articulation, further reducing their acoustic-phonetic predictability (Harper and Maxwell, 2008).

From a semantic viewpoint, function words contribute to a lesser extent to the meaning of a sentence, because they are not associated with real-world identifiable concepts and primarily serve grammatical purposes. Although the mean WER was higher than 25%, errors did not significantly reduce the diagnostic value of semantic similarity measures, which can be explained by the fact that NLP tools rely more on content words, such as nouns, adjectives, verbs, and adverbs when calculating word embedding representations (Lenci, 2018).

The distribution of semantic similarity measures differed based on the type of transcript (ASR-derived vs. manual). This is likely due to the high number of word deletions and substitutions occurring in the automatic transcriptions, which caused differences in the calculation of word embeddings and, in turn, cosine similarity per window. However, these changes in the distribution of semantic similarity features between automated vs. manual transcripts had minimal impact on their relative importance in the RFCs, as evidenced by an 80% overlap between the top 10 semantic similarity features in the two classifiers. Thus, this study suggests that semantic NLP models can tolerate WERs up to 30% on specific classification tasks. For future clinical applications, the 3% absolute drop in performance may still necessitate manual correction of ASR errors or the development of better ASR tools. High diagnostic accuracy is among the most important premises for the future application of automated language assessments in clinical practice.

In addition to semantic similarity between words, emerging research has shown that connectives-related similarities (Corona-Hernández et al., 2022) and referential meaning as measured by linguistic devices such as pronouns and determiners (Palominos et al., 2023) are particularly informative for assessing coherence in the speech of patients with SSD. Based on our current findings, we believe that the ASR tool is not robust enough yet to support such grammatical analyses, as evidenced by the highest pronoun error rate and the most common mistaken words, which included various determiners (that/dat, the/het,de), and connectives (and/en, thus/dus).

The linguistic position of ASR errors revealed that deletion, substitution, and insertion errors occurred most frequently at the end of a sentence and least frequently at the beginning, with a roughly uniform distribution across a sentence. These results are in contrast with those of Goldwater et al. (2010) and Markl and Lai (2021) who found that words at the beginning of a turn have the highest error rate. This discrepancy could be related to a different ASR system (e.g., Google Cloud Speech-to-Text versus Kaldi) or, alternatively, to a different type of speaking task (e.g., self-recorded audio) or language employed in these studies. The peak of errors we observed in sentence-final position could be linked to falling intonation and lower speech intensity, common in most languages at the end of a sentence (Carranza et al., 2014), or the fact that speakers often lower their voices (i.e., loudness) at the end of a turn (Gravano and Hirschberg, 2011), both of which could affect ASR performance. The fact that the most mistaken word 'I/Ik' rarely occurs at the end of a sentence was assumed to be unrelated to the highest rate of error in sentence-final position but could rather be explained by the fact that it is one of the most widely used words in the Dutch language.

4.3. Limitations and final remarks

Some limitations to our study should be considered when interpreting the results. First, we emphasize the need to replicate these results in a larger cohort, across languages, and settings. While this study used high-quality speech recordings collected in a research setting, further research should test whether these findings are generalizable to speech of diverse quality obtained in more naturalistic environments (e.g., the participant's home). Second, we evaluated the use of ASR to support semantic similarity analysis only, which does not guarantee its validity for other types of NLP techniques, e.g., analysis of connectives. Third, although the final classification model in our sample was not significantly impacted by WERs over 25%, a 3% reduction in diagnostic accuracy is an important loss for clinical utility; lower error rates should be aimed for, especially in light of future clinical applications. ASR technology is rapidly developing, indeed providing opportunities to improve transcript accuracy; tools such as Whisper (https://openai.com/blog/whisper/), a novel open-source multilingual ASR tool, show improved robustness to noise and speakers' accents (Radford et al., 2022) and could play a role in enhancing the validity and clinical feasibility of NLP methods for automatically predicting and diagnosing SSD.

In conclusion, our analysis helps understand the transcription accuracy reachable with a state-of-the-art ASR tool in patients with SSD and demonstrates that using automated transcripts instead of manual transcripts, despite high WERs, does not significantly reduce diagnostic accuracy based on a semantic NLP model. Our study therefore suggests that the combination of ASR technology and semantic NLP models is a viable, robust, and efficient method for diagnosing schizophrenia-spectrum disorders.

CRediT authorship contribution statement

S. Ciampelli: Conceptualization, Methodology, Formal analysis, Writing – original draft. **A.E. Voppel:** Conceptualization, Methodology, Writing – review & editing. **J.N. de Boer:** Conceptualization, Methodology, Writing – review & editing. **S. Koops:** Writing – review & editing. **I.E.C. Sommer:** Conceptualization, Writing – review & editing.

Declaration of Competing Interest

All authors declare that they have no conflicts of interest.

Acknowledgments

We are grateful to all participants and research interns for their help with data collection and preparation.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.psychres.2023.115252.

References

Adda-Decker, M., Lamel, L., 2005. Do speech recognizers prefer female speakers? Interspeech, 2005.

Andreasen, N.C., Flaum, M., Arndt, S., 1992. The comprehensive assessment of symptoms and history (CASH). An instrument for assessing diagnosis and psychopathology. Arch. Gen. Psychiatry 49, 615–623. https://doi.org/10.1001/archpsyc.1992.01820080023004.

Barajas, A., Baños, I., Ochoa, S., Usall, J., Huerta, E., Dolz, M., Sánchez, B., Villalta, V., Foix, A., Obiols, J., Haro, J.M., 2010. Gender differences in incipient psychosis. Eur J Psychiatry 24, 176–194.

Boersma, P., 2001. Praat: doing Phonetics by Computer. Glot Int. 5 (9), 341–345.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Brundage, S., Bernstein Ratner, N., 2018. A Clinician's Complete Guide to CLAN and PRAAT 1–43.

Carranza, M., Cucchiarini, C., Llisterri, J., Machuca, M., Rios, A., 2014. A corpus-based study of Spanish L2 mispronunciations by Japanese speakers. In: 6th International Conference on Education and New Learning Technologies. Presented at the Edulearn14.

Chandler, C., Foltz, P.W., Cheng, J., Bernstein, J.C., Rosenfeld, E.P., Cohen, A.S., Holmlund, T.B., Elvevåg, B., 2019. Overcoming the bottleneck in traditional assessments of verbal memory: modeling human ratings and classifying clinical group membership. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology. Presented at the CLPsych 2019. Minneapolis, Minnesota. Association for Computational Linguistics, pp. 137–147. https://doi.org/10.18653/v1/W19-3016.

Chiu, C.-.C., Tripathi, A., Chou, K., Co, C., Jaitly, N., Jaunzeikare, D., Kannan, A., Nguyen, P., Sak, H., Sankar, A., Tansuwan, J.J., Wan, N., Wu, Y., Zhang, F., 2018. Speech recognition for medical conversations.

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

Corcoran, C.M., Cecchi, G.A., 2020. Using language processing and speech analysis for the identification of psychosis and other disorders. Biol. Psychiatry Cogn. Neurosci. Neuroimag. 5, 770–779. https://doi.org/10.1016/j.bpsc.2020.06.004.

Corona-Hernández, H., de Boer, J.N., Brederoo, S.G., Voppel, A.E., Sommer, I.E.C., 2022. Assessing coherence through linguistic connectives: analysis of speech in patients with schizophrenia-spectrum disorders. Schizophr. Res. https://doi.org/10.1016/j.schres.2022.06.013.

Cutler, A., Foss, D.J., 1977. On the role of sentence stress in sentence processing. Lang. Speech. 20, 1–10. https://doi.org/10.1177/002383097702000101.

de Boer, J.N., Voppel, A.E., Begemann, M.J.H., Schnack, H.G., Wijnen, F., Sommer, I.E.C., 2018. Clinical use of semantic space models in psychiatry and neurology: a systematic review and meta-analysis. Neurosci. Biobehav. Rev. 93, 85–92. https://doi.org/10.1016/j.neubiorev.2018.06.008.

de Boer, J.N., Voppel, A.E., Brederoo, S.G., Schnack, H.G., Truong, K.P., Wijnen, F.N.K., Sommer, I.E.C., 2021. Acoustic speech markers for schizophrenia-spectrum disorders: a diagnostic and symptom-recognition tool. Psychol. Med. 1–11. https://doi.org/10.1017/S0033291721002804.

Dingliwal, S., Shenoy, A., Bodapati, S.B., Gandhe, A., Gadde, R., Kirchhoff, K., 2021. Domain Prompts: towards memory and compute efficient domain adaptation of ASR systems.

Elvevåg, B., Foltz, P.W., Weinberger, D.R., Goldberg, T.E., 2007. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. Schizophr. Res. 93, 304–316. https://doi.org/10.1016/j.schres.2007.03.001.

Goldwater, S., Jurafsky, D., Manning, C.D., 2010. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. Speech. Commun. 52 (181) https://doi.org/10.1016/j.specom.2009.10.001.

Goryainova, M., Grouin, C., Rosset, S., Vasilescu, I., 2014. Morpho-syntactic study of errors from speech recognition system. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Presented at the

LREC 2014, European Language Resources Association (ELRA). Reykjavik, Iceland, pp. 3045–3049.

Gravano, A., Hirschberg, J., 2011. Turn-taking cues in task-oriented dialogue. Comput. Speech Lang. 25, 601–634. https://doi.org/10.1016/j.csl.2010.10.003.

Harper, M.P., Maxwell, M., 2008. Spoken Language Characterization. In: Benesty, J., Sondhi, M.M., Huang, Y.A. (Eds.), Springer Handbook of Speech Processing. Springer Handbooks. Springer, Berlin, Heidelberg, pp. 797–810. https://doi.org/10.1007/978-3-540-49127-9_40.

Holmlund, T.B., Chandler, C., Foltz, P.W., Cohen, A.S., Cheng, J., Bernstein, J.C., Rosenfeld, E.P., Elvevåg, B., 2020. Applying speech technologies to assess verbal memory in patients with serious mental illness. npj Digit. Med. 3, 1–8. https://doi.org/10.1038/s41746-020-0241-7.

Kay, S.R., Fiszbein, A., Opler, L.A., 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr. Bull. 13, 261–276. https://doi.org/10.1093/schbul/13.2.261.

Kodish-Wachs, J., Agassi, E., Kenny, P., Overhage, J.M., 2018. A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech. AMIA Annu. Symp. Proc. 2018 683–689.

Lenci, A., 2018. Distributional models of word meaning. Ann. Rev. Linguist. 4, 151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254.

MacWhinney, B., 2000. The CHILDES project: tools for analyzing talk: transcription format and programs. In: The CHILDES project: Tools for Analyzing talk: Transcription format and Programs, 3rd ed, 1. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US. Vol. 1, 3rd ed.

Markl, N., Lai, C., 2021. Context-sensitive evaluation of automatic speech recognition: considering user experience & language variation. In: Proceedings of the First Workshop on Bridging Human-Computer Interaction and Natural Language Processing. Presented at the HCINLP 2021. Association for Computational Linguistics, pp. 34–40. Online.

Martínez-Sánchez, F., Muela-Martínez, J.A., Cortés-Soto, P., García Meilán, J.J., Vera Ferrándiz, J.A., Egea Caparrós, A., Pujante Valverde, I.M., 2015. Can the acoustic analysis of expressive prosody discriminate schizophrenia? Span. J. Psychol. 18, E86. https://doi.org/10.1017/sjp.2015.85.

Mikolov, T., Chen, K., Corrado, G.S., Dean, J., 2013. Efficient estimation of word representations in vector space. In: Proceedings of Workshop at ICLR 2013.

Moro-Velazquez, L., Cho, J., Watanabe, S., Hasegawa-Johnson, M.A., Scharenborg, O., Kim, H., Dehak, N., 2019. Study of the performance of automatic speech recognition systems in speakers with Parkinson's disease. In: Interspeech 2019. Presented at the Interspeech 2019, ISCA, pp. 3875–3879. https://doi.org/10.21437/Interspeech.2019-2993.

Overbeek, T., Schruers, K., 2019. MINI-S voor DSM-5 Nederlandse versie 2019, Overbeek & Schruers /English - Version 2 © Hergueta & Weiller.

Palominos, C., Figueroa-Barra, A., Hinzen, W., 2023. Coreference delays in psychotic discourse: widening the temporal window. Schizophr. Bull. 49, S153–S162. https://doi.org/10.1093/schbul/sbac102.

Parola, A., Simonsen, A., Bliksted, V., Fusaroli, R., 2020. Voice patterns in schizophrenia: a systematic review and Bayesian meta-analysis. Schizophr. Res. 216, 24–40. https://doi.org/10.1016/j.schres.2019.11.031.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., et al. (Eds.), 2011. The Kaldi Speech Recognition Toolkit. Presented at the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society.

R Core Team, 2020. R: a language and environment for statistical computing. R Foundation for Statistical Computing., Vienna, Austria.

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2022. Robust Speech Recognition via Large-Scale Weak Supervision. https://doi.org/10.48550/arXiv.2212.04356.

Radha, V., Vimala, C., 2012. A review on speech recognition challenges and approaches. World Comput. Sci. Inform. Technol. J. (WCSIT) 1, 1–7.

Santiago, F., Dutrey, C., Adda-Decker, M., 2015. Towards a typology of ASR errors via syntax-prosody mapping. In: Errors by Humans and Machines in Multimedia, Multimodal and Multilingual Data Processing, pp. 175–192.

Tahir, Y., Yang, Z., Chakraborty, D., Thalmann, N., Thalmann, D., Maniam, Y., Rashid, N. A.Binte A., Tan, B.-.L., Keong, J.L.C., Dauwels, J., 2019. Non-verbal speech cues as objective measures for negative symptoms in patients with schizophrenia. PLoS One 14, e0214314. https://doi.org/10.1371/journal.pone.0214314.

Tang, S.X., Kriz, R., Cho, S., Park, S.J., Harowitz, J., Gur, R.E., Bhati, M.T., Wolf, D.H., Sedoc, J., Liberman, M.Y., 2021. Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. npj Schizophr. 7, 1–8. https://doi.org/10.1038/s41537-021-00154-3.

Tomczak, M., Tomczak, E., 2014. The need to report effect size estimates revisited. Overv Some Recommend. Measur. Effect Size 21, 19–25.

Van der Sloot, K., Hendrickx, I., van Gompel, M., van den Bosch, A., Daelemans, W., 2018. Frog, A natural language processing suite for Dutch. In: Language and Speech Technology Technical Report Series 18-02.

van Eerten, L., 2007. Corpus Gesproken Nederlands. Nederlandse taalkunde (Groningen) 12, 194–215.

van Eynde, F., 2004. Part of speech tagging en lemmatisering van het corpus gesproken nederlands. Technical report, Centrum voor Computerlinguıstiek.

Van Rossum, G., Drake, F.L., 2009. Python 3 Reference Manual. CreateSpace, Scotts Valley, CA.

Voppel, A.E., de Boer, J.N., Brederoo, S.G., Schnack, H.G., Sommer, I., 2021. Quantified language connectedness in schizophrenia-spectrum disorders. Psychiatry Res. 304, 114130 https://doi.org/10.1016/j.psychres.2021.114130.

Xu, W., Wang, W., Portanova, J., Chander, A., Campbell, A., Pakhomov, S., Ben-Zeev, D., Cohen, T., 2022. Fully automated detection of formal thought disorder with Time-series Augmented Representations for Detection of Incoherent Speech (TARDIS). J. Biomed. Inform. 126, 103998 https://doi.org/10.1016/j.jbi.2022.103998.