Semantic-based NLP techniques discriminate schizophrenia and Wernicke's aphasia based on spontaneous speech

Frank Tsiwah, Anas Mayya, Andreas van Cranenburgh
Center for Language and Cognition Groningen, University of Groningen, The Netherlands
{f.tsiwah, a.w.van.cranenburgh}@rug.nl

Abstract

People with schizophrenia spectrum disorder (SSD)—a psychiatric disorder, and people with Wernicke's aphasia—an acquired neurological disorder, are both known to display semantic deficits in their spontaneous speech outputs. Very few studies directly compared the two groups on their spontaneous speech (Gerson et al., 1977; Faber et al., 1983), and no consistent results were found. Our study uses word (based on the word2vec model with moving windows across words) and sentence (transformer based-model) embeddings as features for a machine learning classification model to differentiate between the spontaneous speech of both groups. Additionally, this study uses these measures to differentiate between people with Wernicke's aphasia and healthy controls. The model is able to classify patients with Wernicke's aphasia and patients with SSD with a cross-validated accuracy of 81%. Additionally, it is also able to classify patients with Wernicke's aphasia versus healthy controls and SSD versus healthy controls with cross-validated accuracy of 93.72% and 84.36%, respectively. For the SSD individuals, sentence and/or discourse level features are deemed more informative by the model, whereas for the Wernicke group, only intra-sentential features are more informative. Overall, we show that NLP-based semantic measures are sensitive to identifying Wernicke's aphasic and schizophrenic speech.

Keywords: word embeddings, Schizophrenia, Wernicke's aphasia, word connectedness, coherence.

1. Introduction

The language of individuals with schizophrenia spectrum disorder (SSD) and Wernicke's aphasia are both characterized by semantic impairments, although they have distinct etiologies (Faber and Reichstein, 1981). While the former is a long-term psychiatric disorder that requires medication and sometimes hospitalization (American Psychiatric Association, 2013), the latter is an acquired neurological language disorder resulting most commonly from a cerebrovascular accident (Acharya and Wroten, 2023). Despite the differences in etiology and overall symptomatology, both disorders are known to affect the ability of individuals to comprehend and to produce semantically coherent speech. For example, speech by people with SSD may include incoherence, derailment, tangentiality and neologisms, and these features are routinely used by clinicians as one of the strongest diagnostic markers of schizophrenia in their mental health examinations (Kuperberg, 2010). Similarly, speech by people with Wernicke's aphasia is characterized by incoherence, use of neologisms and jargon. Interestingly, in the literature on both schizophrenia and Wernicke's aphasia, "word salad" (meaningless speech) has been used to describe panients' speech (Butler and Zeman, 2005).

This evident resemblance between the two patient groups poses a challenge in distinguishing them, potentially leading to misidentification of Wernicke's aphasia as a manifestation of a psychiatric thought disorder, particularly in the absence of neuroimaging examination (Butler and Zeman, 2005). The advent of natural language processing (NLP) and other machine learning (ML) techniques, and their sensitivity to detect subtle patterns in language data, enables us to quantify and observe semantic patterns in speech and language in general (e.g., Tang et al., 2021; Corcoran et al., 2020; Sarzynska-Wawer et al., 2021; Fraser et al., 2013; Themistocleous et al., 2021). Therefore, the goal of the current study is to use NLP-derived semantic measures to assess the degree of (dis)similarity between speech characterized by schizophrenia and Wernicke's aphasia.

A typical approach in examining language disruptions in individuals with schizophrenia involves assessing a deficit in "connectedness" of language, as a measure of coherence (Covington et al., 2005). Given that words that occur together within the same sentence tend to share the same meaning, connectedness can be measured both at the intra- and inter-sentential level. Recent advances in NLP have provided a means to quantify connectedness between words, but also across sentences, using word and sentence embeddings, respectively. This methodology has demonstrated comparable or even superior efficacy to current clinical scales in the diagnosis of schizophrenia (Voppel et al., 2021; Tang et al., 2021). Therefore, the current study aims to address the question of whether NLP-derived measures can be used to distinguish people with Wer-

Proceedings of the RaPID-5 @LREC-COLING 2024, pages 1–8

nicke's aphasia, schizophrenia and healthy controls, based on spontaneous speech transcripts.

There have been few studies that have directly examined potential differences and similarities between schizophrenic and fluent aphasic speech. Gerson et al. (1977) compared people with conduction, transcortical sensory, and Wernicke's aphasia with people with schizophrenia, and showed that the former (three fluent aphasic) group had more paraphasic errors while the latter had more bizarre themes. Faber et al. (1983) compared the verbal abilities of 14 people with schizophrenia, diagnosed with formal thought disorder, with 13 (11 of which were fluent) of those with aphasia. The spontaneous speech transcripts of the patients were presented for blind classification to a language and speech therapist, two psychiatrists and two neurologists. Their findings showed that only three raters performed better than chance level in correctly identifying fluent aphasics, and with poor inter-rater reliability. Most errors were associated with misclassification of aphasia as schizophrenia than the other way round (23 errors out of 65 ratings vs 9 errors out of 70). No aphasic patient was unanimously classified correctly, while 8 schizophrenic patients were. In terms of speech differences, out of 14 language abnormalities rated by the blind assessors, five differentiated both groups: word approximations/private use of words, derailment/tangentiality were seen more in schizophrenia, while the other (aphasic) group demonstrated poverty of speech content, reduced auditory comprehension, and word finding difficulty. Contrary to the findings of Gerson et al. (1977), there is no indication that the schizophrenia group displayed a distinct thinking disorder: Both groups had equal number of paraphasias and neologisms, and only a third of the schizophrenic group demonstrated illogical thinking.

This raises the question of whether clinicians can reliably differentiate between the two disorders solely based on examining their speech and language (Gerson et al., 1977; Faber et al., 1983). However, to the best of our knowledge, no study has used an NLP-based or other ML approaches to investigate this research problem. Since the current determination of the etiology of individuals presented with this type of language impairment (either Wernicke or schizophrenia) requires language assessment, neurological examination and thorough psychiatric evaluation, using an NLP method for automatic classification can provide physicians and neuropsychologists with objective and cost-effective measures to assess and diagnose patients, and to track their responses to treatments.

2. Data and Participants

We obtained secondary data from two sources for this study. The first source was the Aphasiabank (MacWhinney et al., 2011), from which we obtained data of 26 patients with Wernicke's aphasia (WA) and 37 healthy controls (HC: randomly selected). The second source was the data published and shared by Tang et al. (2021), from which we included 27 patients with schizophrenia spectrum disorder (SSD). All participants were native speakers of English. The data included spontaneous speech transcripts based on participants' responses to semi-structured interview where questions such as "Tell me about an important event in your life" were asked (see Appendix A for an example of interviewer-participant dialogue for both group). Although the data from these two sources included picture descriptions which were different depending on data source, we decided to focus only on the open-ended personal questions since participants' responses to these questions would always be different regardless of whether (1) the data originates from the same source or not, (2) the testing conditions remained consistent or not. Data were pre-processed, and fillers or any symbols inserted by annotators in the transcripts were all removed.

3. Semantic Feature Extraction

The NLP-derived semantic scores in this study are cosine similarity scores, based on two pre-trained word and sentence embedding models: word2vec (Mikolov et al., 2013) and Sentence-Bidirectional Encoder Representation from Transformers (sBERT: Reimers and Gurevych, 2019), respectively. Semantic space models like word2vec aim to capture the interconnectedness within language by exploiting 'similarities' among words. A cosine similarity of 1 means the two vectors are identical, while a cosine similarity of 0 means the two vectors are orthogonal. In this study we use cosine similarity computed from the word2vec and the sBERT models as a measure of how similar words and sentences are to other words and sentences, respectively. We assume that a lower average cosine similarity in the speech output of a speaker implies lower coherence. We used two approaches for calculating similarities: (1) word and sentence similarities within only participants' utterances, (2) word and sentence similarities within participants' utterances in relation to the interview question or prompt. This was done with both the word2vec and the sBERT models, which are described below.

3.1. Word2vec

3.1.1. Participants' utterances

For every interview question, we calculate the average and variance of cosine similarities between the words in the participants' utterances. To capture a wide range of similarity within and between sentences, we use a moving window ranging from 1 to 19 (we adapted this method from Voppel et al., 2021). To illustrate, if the moving window is one, we would calculate the cosine similarity in the sentence "I enjoy doing the laundry" as shown in Table 1.

For each given window, cosine similarity between individual words uttered by the participants are calculated, and then averaged to produce a single average similarity value, reflecting the degree of word connectedness within that window. Additionally, the variance in similarity scores is computed over all similarities across the utterances of the participant. For every participant, we ended up with 19 average scores and 19 variance scores.

3.1.2. Participants' utterance in relation to interview questions

In addition to the word embeddings derived from only participants' utterances as described above, we compute cosine similarities across the words within the interviewer's questions or prompts, and then average them. We then measure the cosine similarity between the interviewer's question against the participant's utterance, which we split into three segments using the moving windows. The first, second and third segments corresponded with 1–7, 7–13, and 13–19 moving windows, respectively. The rationale behind this is to be able to capture potential derailment in answers given by participants in relation to the question by the interviewer, from the start of their utterance to the end. For instance, if the individuals with schizophrenia derailed more, then they would have lower cosine similarity scores on the second or third segments in relation to the averaged cosine similarity score based on the interviewer's question. That is, their first response to the interviewer's question would be semantically closer to the question than the second or third segment of their utterance, indicating derailment.

3.2. sBERT

3.2.1. Participants' utterances

Contrary to word2vec, we used sBERT to create sentence embeddings from the participants' utterances. We used moving windows from one to three, where each moving window represents a sentence rather than a word. Sentences were segmented based on ".,!,?" separators. The moving window paradigm was used to create 1–3 windows of sentence embedding, using both averages and variance of cosine similarity between the sentences of each participant.

3.2.2. Participants' utterance in relation to interview questions

For every utterance by the participant and interviewer, we averaged the vectors of all the sentences, and measured their variance as well. Additionally, similar to word2vec, we calculated the cosine similarity between the average of the interviewer's questions and each of the 1–3 moving windows of sentences based on participants' utterance, in order to capture derailment.

4. Method

We run a Random Forest (RF) model with all 51 predictors (features extracted using both word2vec and sBERT) included, with diagnosis as the target containing Wernicke, SSD and Healthy controls (Healthy_C). We compared the performance of the RF model with Naive Bayes and Support Vector Machine, but the RF was the best performing model. Therefore, we only report the experiment with the RF model. We run three classifications in total comparing the Wernicke group vs the SSD group; the Wernicke group vs the Healthy_C group; and the SSD group vs the Healthy_C group. Prior to running the RF model, we run a majority class baseline classifier for each comparison.

We use k-fold stratified cross-validation with $k = 5$ to train the model. This involves dividing the training set into k parts, referred to as folds, and subsequently training a model using each fold as a validation set. For each fold, the remainder of the data serves as its training set, with the goal of mitigating overfitting to noise in the dataset. We split the data into 80–20 for training and test sets, respectively, due to the small sample size of our dataset (Wernicke = 26, SSD = 27, Healthy_C = 37, total features = 51). The experiment is performed using the Scikit-learn module (Pedregosa et al., 2011) for the Python programming language. The code used is publicly available on GitHub: https://github.com/FrankTsi/NLP-Schizophrenia-Wernicke-s-aphasia.

5. Results and Discussion

Table 2 shows the classification scores for each group comparison. Using a random forest binary classification algorithm based on mean, variance in connectedness, and sBERT scores, a k-

Sentence: I enjoy doing the laundry			
Moving Window 1			
I-enjoy	enjoy-doing	doing-the	the-laundry
Moving Window 2			
I-enjoy-doing	enjoy-doing-the		doing-the-laundry

Table 1: Moving window example

	SSD vs Wernicke	Healthy_C vs Wernicke	Healthy_C vs SSD
accuracy	81.27	93.72	84.36
precision	81.74	94.67	87.23
recall	81.00	93.32	83.58
f1-score	81.06	93.13	83.40

Table 2: The RF classification scores for the three group classifications based on the k-fold (k=5) cross validation. Scores represent the means of all folds.

fold cross validation ($k = 5$) accuracy of 81.27% is attained in distinguishing individuals with Wernicke's aphasia—a neurological language disorder, and schizophrenia—a psychiatric thought disorder. This performance significantly surpasses the baseline model, which achieves only 51% accuracy. Notably, this level of accuracy is higher than previous attempts using clinical measures, which often results in challenges with differentiating schizophrenic speech from that of Wernicke's aphasia, usually accompanied by poor inter-rater reliability (Faber et al., 1983). Our results suggest that the underlying language impairments in schizophrenia and Wernicke's aphasia are distinct, despite both being associated with "word salad" (meaningless speech), implying a perceived similarity in their speech characteristics (Butler and Zeman, 2005). Thus it can be argued that based solely on spontaneous speech, psychiatric language disorders can largely be distinguished from neurological language disorders.

Turning now to the classification between the healthy controls and each of the patient groups, our model achieves a remarkably high accuracy of 93.7% in classifying Wernicke's aphasic individuals and healthy controls (see Table 2). To the best of our knowledge, this is the first study to report the use of an NLP approach to automatically detect Wernicke's aphasia. Furthermore, our random forest classifier demonstrated an accuracy of 87.6% in classifying the SSD group against the healthy control group. It is worth noting that these accuracy scores are based on a k-fold cross-validation ($k = 5$) report. This level of accuracy for distinguishing SSD from healthy controls is consistent with findings from other studies using NLP methods to detect schizophrenia (Voppel et al., 2021; Tang et al., 2021; Iter et al., 2018).

After demonstrating the sensitivity of our random forest classifier to discriminate between Wernicke's aphasic and the SSD speech transcripts, we now turn to the question: which word connectedness features are more important for our classifier to distinguish schizophrenic spontaneous speech from that of Wernicke's. We approached this by first comparing both the Wernicke's aphasic and the SSD speech against the healthy control speech, and then calculating the random forest's Gini importance of features to evaluate the importance of each feature used by the classifier. We report only the top ten Gini importance features and their scores, as demonstrated on Figure 1 (see Appendix B for the scores of all features). Our findings demonstrate that for Wernicke's aphasic speech and the healthy control, the features that were consistently deemed more important for our classifier were the word level embeddings capturing the average (ave_windows 1, 3, 5, 9, 10, 12, 14) cosine similarities, and variance (var_window 1). The feature 'INT_PAR_distance_score' (indicating the distance between the average cosine similarity score of the Interviewer's questions vs the participant's response) was the most informative to the model. The sBERT score from the first sentence (sBERT_ave_window_1) uttered by Wernicke's aphasic participants was also informative to the model, with a rank of three. Overall, for individuals with Wernicke's aphasia, intra-sentence word connectedness is deemed more informative in distinguishing them from healthy controls.

Conversely, the features that were most important for our classifier to distinguish individuals with SSD from healthy controls are the sentence-level characteristics extracted from sBERT sentence embeddings. Interestingly, all three sentence-level windows from sBERT ranked among the top 4 features deemed most significant by the random forest classifier. Specifically, for the SSD group, unlike the Wernicke group, discourse incoherence spanning across sentences emerged as the most

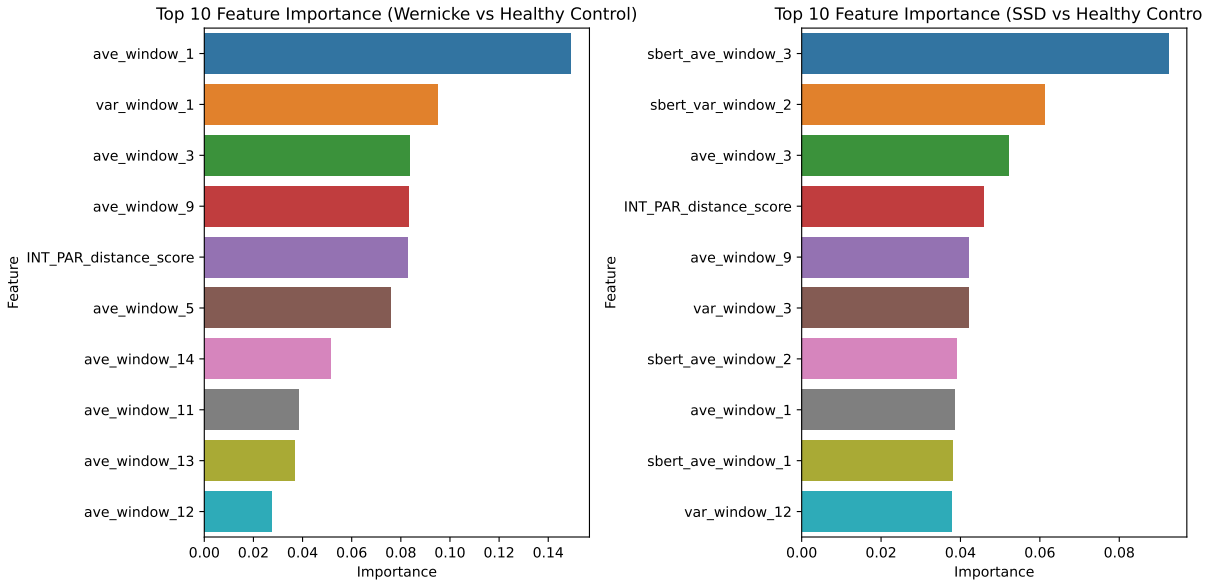

Figure 1: Feature importance scores.

critical feature in distinguishing them from healthy controls. This finding is in line with the spontaneous speech characteristics of individuals with schizophrenia, as reported in the literature (Covington et al., 2005; Voppel et al., 2021; Tang et al., 2021; Iter et al., 2018).

6. Limitations

We now consider the limitations of this study. First, the sample sizes of both patient groups were small for a classification model that splits data into training and testing data. We only had a testing sample of 6 or 7 for each of the Wernicke and SSD groups. This limits the generalizability of the current results. Second, we used interviewer-related measures with the assumption that all interviewers frame questions in the same way and make an equal number of turns in the conversation. This may not always be the case. Interviewer styles might differ across questions, interviews, and protocols. Such variation can affect the reliability of the measure. Additionally, our approach did not account for the occurrences of neologisms and misspellings, which could potentially affect the similarities scores from the word2vec model. We suggest that future efforts address these issues. Lastly, it is known that medication also influences the speech of patients with SSD (de Boer et al., 2020; Sinha et al., 2015). We recommend that future studies take into account the potential effect of medication on the performance of the patients with SSD, although such data was not available for the cohort involved in this project.

7. Conclusion

In summary, our results demonstrate that semantics-based, NLP-derived metrics alone can potentially serve as a diagnostic tool to differentiate not only individuals with Wernicke's aphasia and schizophrenia from healthy controls but also between these two patient cohorts. In spite of the limitations discussed in the previous section, the results of this study are particularly promising, as the current method of distinguishing Wernicke's aphasia and schizophrenia necessitates language assessment, neurological examination, and comprehensive psychiatric evaluation, which can be resource-intensive and time-consuming.

8. Acknowledgments

The authors are deeply grateful to Aphasiabank and Tang and colleagues who provided us with this clinical data, as well as to the anonymous reviewers for their helpful comments.

9. Bibliographical References

A.B. Acharya and M. Wroten. 2023. Wernicke aphasia. StatPearls.

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*, volume 5. American psychiatric association Washington, DC.

C Butler and AZJ Zeman. 2005. Neurological syndromes which can be mistaken for psychiatric conditions. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(suppl 1):i31–i38.

Christopher M Corcoran, Vijay A Mittal, Carrie E Bearden, Ruben Gur, Kasia Hitczenko, Zain Bilgrami, and et al. 2020. Language as a biomarker for psychosis: A natural language processing approach. *Schizophrenia Research*, 226:158–166.

Michael A Covington, Congzhou He, Cati Brown, Lorina Naçi, Jonathan T McClain, Bess Sirmon Fjordbak, James Semple, and John Brown. 2005. Schizophrenia and the structure of language: the linguist's view. *Schizophrenia research*, 77(1):85–98.

J. N. de Boer, A. E. Voppel, S. G. Brederoo, et al. 2020. Language disturbances in schizophrenia: the relation with antipsychotic medication. *npj Schizophrenia*, 6:24.

Raymond Faber, Richard Abrams, Michael A Taylor, Arlene Kasprison, Charles Morris, and Reuben Weisz. 1983. Comparison of schizophrenic patients with formal thought disorder and neurologically impaired patients with aphasia. *The American journal of psychiatry*, 140(10):1348–1351.

Raymond Faber and Michele Bierenbaum Reichstein. 1981. Language dysfunction in schizophrenia. *The British Journal of Psychiatry*, 139(6):519–522.

Kathleen C Fraser, Frank Rudzicz, Naida Graham, and Elizabeth Rochon. 2013. Automatic speech recognition in the diagnosis of primary progressive aphasia. In *Proceedings of the fourth workshop on speech and language processing for assistive technologies*, pages 47–54.

S. N. Gerson, F. Benson, and S. H. Frazier. 1977. Diagnosis: schizophrenia versus posterior aphasia. *The American Journal of Psychiatry*, 134(9):966–969.

D. Iter, J. Yoon, and D. Jurafsky. 2018. Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology*.

Gina R Kuperberg. 2010. Language in schizophrenia part 1: an introduction. *Language and linguistics compass*, 4(8):576–589.

Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. 2011. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Fabian Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Justyna Sarzynska-Wawer, Aleksandra Wawer, Adam Pawlak, Joanna Szymanowska, Izabela Stefaniak, Marek Jarkiewicz, and et al. 2021. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135.

P. Sinha, V. P. Vandana, N. V. Lewis, M. Jayaram, and P. Enderby. 2015. Evaluating the effect of risperidone on speech: a cross-sectional study. *Asian Journal of Psychiatry*, 15:51–55.

Sunny X Tang, Reno Kriz, Sunghye Cho, Suh Jung Park, Jenna Harowitz, Raquel E Gur, Mahendra T Bhati, Daniel H Wolf, João Sedoc, and Mark Y Liberman. 2021. Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. *npj Schizophrenia*, 7(1):25.

Charalambos Themistocleous, Bronte Ficek, Kimberly Webster, Dirk-Bart den Ouden, Argye E Hillis, and Kyrana Tsapkini. 2021. Automatic subtyping of individuals with primary progressive aphasia. *Journal of Alzheimer's Disease*, 79(3):1185–1194.

AE Voppel, JN de Boer, SG Brederoo, HG Schnack, and IEC Sommer. 2021. Quantified language connectedness in schizophrenia-spectrum disorders. *Psychiatry Research*, 304:114130.

A. Example of spontaneous speech sample from both groups

Interviewer - SSD dialogue

Interviewer: Now I'm just going to ask you two open ended questions, so just try to respond to my prompts with as much detail as you can. Okay?
Interviewer: Tell me about yourself.
Participant: So
I'm the devil.
And
I can't talk like the devil so I have to change my face a lot.
But that's one of my faces on the inside and out.
So I guess
I have to be kind to that one and let him talk at all.
You know, the things he would have said if he was a naughty person.
But not be like him, and save the world.
Interviewer: Anything else?
Participant: breath I have a wife.
with three
hundred gazillion and twenty-six kids.
I have a mother that's name is [Patricia].
I love my dad the most,
because he never hits me.
Mom used to whip me.
But she's the devil's
daughter.
And that's just a role she had to play, not because she wanted to play.

Wernicke's participant - Interviewer dialogue

Interviewer: well thinking back um can you tell me about something important that happened in your life?
Participant: being born i guess.
best.
i when i was about three i was three years.
yeah.
he's he drawing you know.
oh yeah.
oh yeah.
i have three girls brothers who were babies you know.
and i got a i got we can see my brothers if you wanna.
over there i got here over there.
okay.
yeah for for a minute.
mhm.
well firstname J and firstname W they they fought all the time you know for high school.
and they at time that they're they were about seven eight high school you know.
they fought a little bit.
me and firstnamew got two fights.
wayne no firstnamej what one fight me and me and him.
yeah.
i wish they one time we had a girl and her and just three boys.
i i wish i was not the baby and a girl and they had four no kids you know.

B. All Features with scores for both the SSD and Wernicke groups