INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Exploring MMSE Score Prediction Using Verbal and Non-Verbal Cues

Shahla Farzana, Natalie Parde

Department of Computer Science
University of Illinois at Chicago
851 S. Morgan St., Chicago, IL 60607
{sfarza3, parde}@uic.edu

Abstract

The Mini Mental State Examination (MMSE) is a standardized cognitive health screening test. It is generally administered by trained clinicians, which may be time-consuming and costly. An intriguing and scalable alternative is to detect changes in cognitive function by automatically monitoring individuals' memory and language abilities from their conversational narratives. We work towards doing so by predicting clinical MMSE scores using verbal and non-verbal features extracted from the transcripts of 108 speech samples from the ADReSS Challenge dataset. We achieve a Root Mean Squared Error (RMSE) of 4.34, a percentage decrease of 29.3% over the existing performance benchmark. We also explore the performance impacts of acoustic versus linguistic, text-based features and find that linguistic features achieve lower RMSE scores, providing strong positive support for their inclusion in future MMSE score prediction models. Our best-performing model leverages a selection of verbal and non-verbal cues, demonstrating that MMSE score prediction is a rich problem that is best addressed using input from multiple perspectives.

Index Terms: spoken language processing, spoken language analysis, healthcare applications, dementia detection

1. Introduction

Scientific progress and improved healthcare standards in many areas of the world have resulted in older populations than ever before [1]. Although this is in many ways cause for celebration, it also introduces new challenges to administering effective clinical care. A growing elderly population creates an increased demand for a wide range of healthcare services, including cognitive assessment. Managing clinician burden and allowing medical professionals to allocate their time effectively is key to maximizing health outcomes and minimizing patient distress. One way to do this is by automating lower-risk tasks, such as routine cognitive assessment.

Cognitive assessment is often performed using straightforward, clinically validated tests such as the Mini Mental State Examination (MMSE) [2]. Clinicians administering the MMSE ask patients a series of questions in five different areas (orientation, registration, attention, memory, and language); their responses to these questions ultimately result in a score ranging from 0 (greatest cognitive decline) to 30 (no cognitive decline). Although simple to administer, the assessment can be burdensome, requiring the patient to travel to a clinical setting for in-person assessment. It may also be subject to biases from various demographic factors [3]. As an alternative to the structured, in-person MMSE, preliminary evidence suggests that automated methods can be used to predict MMSE scores from open-ended narrative descriptions [4]. The availability of easily-accessible, automated mechanisms could also enable assessment of individuals at more frequent, regular intervals, potentially facilitating quicker diagnosis of early-stage dementia [5].

We work toward this goal of simple, efficient dementia diagnosis by investigating a wide range of spoken language features for automated MMSE score prediction. It is well-known that dementia can influence spontaneous speech production, with declines in verbal fluency often manifesting with longer hesitations, lower speech rates, more frequent repetition, and other aphasic conditions [6, 7]. We design features that account for these discourse characteristics, in addition to incorporating promising linguistic features from prior work. Our findings suggest that a combination of verbal and non-verbal features results in strong predictive ability. Our contributions are as follows:

1. We propose a suite of features for MMSE score prediction, and run experiments to assess their utility for the task. We find that a blend of features drawn from multiple linguistic and discourse perspectives exhibits the strongest performance.
2. We extract features designed to encode properties of hesitation and verbal fluency, which are important biomarkers of Alzheimer's disease. Since identifying these subtle characteristics directly from audio files remains a challenging task [8, 9], we leverage the extensive set of annotations for non-verbal cues already present in the transcripts. To the best of our knowledge, the use of these features for MMSE score prediction is novel.
3. We compare the performance of acoustic and textual features for the task, finding that models trained only on text features outperform those trained only on acoustic features. This provides strong support for the inclusion of linguistic features in future models.
4. We analyze patterns in the features found to be most beneficial, finding that function words and discourse connectives offer high predictive value.

Our best-performing model outperforms the existing task benchmark by a wide margin (RMSE=4.34, a 29.3% decrease from the acoustic benchmark (RMSE=6.14) at the time of submission, and a 16.5% decrease from the linguistic benchmark (RMSE=5.20) added before the camera-ready deadline [4]).

2. Related Work

There is growing interest in automated dementia detection, although most work to date has focused on the binary task of dementia classification (wherein an individual is predicted to either have or not have dementia) [10, 11, 12, 13, 4] rather than the more nuanced problem of assigning continuous MMSE scores [14, 4]. Unlike most recent natural language processing tasks, which have migrated almost exclusively to using neural models with implicitly learned features, small dataset sizes and a

http://dx.doi.org/10.21437/Interspeech.2020-3085

strong interest in maintaining model interpretability have kept the problem space of automated dementia detection refreshingly diverse. Recent high-performing models have relied on a wide range of engineered features [10, 14, 11, 12, 15, 4], at the same time that others have explored neural solutions [16, 13].

Although we examine one neural solution for comparative purposes, our focus in this work is on identifying high-performing interpretable feature sets. Previously, others have explored both acoustic [11, 15, 4] and linguistic [10, 11, 12, 13] engineered features, primarily for dementia classification [14, 15, 4] rather than regression [4]. Acoustic features that have proved successful for the task include fundamental frequency [4], measures of vocal quality [4], Mel Frequency Cepstral Coefficients [11, 4], and pause- and duration-based features [15], among others. High-performing linguistic features have included verbal markers (e.g., indicators of repetition or backtracking) [10], syntax patterns [10, 11], lexical characteristics [10, 12], part-of-speech tags [11, 13], syntactic complexity [11], psycholinguistic traits [11, 13], vocabulary richness [11, 12], information content [11], repetitiveness [11], n-grams [12], and sentiment [13]. We draw inspiration from many of these prior approaches in selecting and designing features for our MMSE prediction models. Specifically, we make use of an expanded n-gram set, non-verbal speech and discourse markers via CHAT transcript [17] annotations, and measures of word familiarity, imageability, concreteness, sentiment, and typical age of word acquisition, as well as MFCC acoustic features.

3. Methods

We employ a set of automatically-extracted lexicosyntactic, psycholinguistic, discourse-based, and acoustic features for estimating continuous MMSE scores on a scale of 0 to 30. Although MMSE scores are often present in dementia detection datasets, the task is generally approached as a binary classification problem; its framing as a regression task is under-explored. We experiment with several machine learning techniques for representing relationships between our observed features and the underlying clinical scores. We explored this task in the context of the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) Challenge.

3.1. Data

The ADReSS Challenge dataset is a subset of DementiaBank's Pitt Corpus [18]. The Pitt Corpus consists of anonymized recordings and transcripts of spoken picture descriptions elicited from participants who were shown the Cookie Theft picture from the Boston Diagnostic Aphasia Exam [19]. In the recordings and transcripts, the interviewer asks the participant to describe what is in the picture, with no time constraints and relatively little structure (on occasion, the interviewer prods the participant for clarification or additional details). The audio from these conversations was manually transcribed, with discourse markers added for false starts, pauses, word repetition, phrase tracing, incomplete sentences, and other nonverbal cues, using the CHAT coding system [17]. For the ADReSS Challenge, the original speech recordings were also segmented into volume-normalized clips of at most ten seconds in length.

The dataset was divided by the task organizers into training and test sets. The training set contained 108 transcripts with an average conversation length (in terms of number of words uttered by the participant) of 98.5 (SD=55.37), and the test set contained 48 transcripts with an average conversation length of 93.38 (SD=56.20). The dataset (unlike the Pitt Corpus as a whole) was gender- and age-balanced across participants with and without dementia. Individual participant demographic information, cognitive status (Dementia or Control), and MMSE score were provided for all training samples; cognitive status and MMSE score were not provided for test samples. We preprocessed the transcripts to remove interviewer utterances, as well as numbers, punctuation, and unwanted symbols.

Table 1: *Token-level psycholinguistic and sentiment features.*

Feature	Description
Age of Acquisition	The age at which a particular word is usually learned.
Concreteness	A measure of a word's tangibility.
Familiarity	A measure of how often one might expect to encounter a word.
Imageability	A measure of how easily a word can be visualized.
Sentiment	A measure of a word's valence.

3.2. Features

We automatically extracted a variety of features from each transcript, described in more detail below.

3.2.1. Lexicosyntactic Features

We extracted n-grams for $n \in \{1, 2, 3\}$ from all training set samples, retaining only n-grams that appeared at least five times and at most 50 times across the training data and including coded non-verbal cues (e.g., *laugh*, *cough*, *breath intake*, or *sigh*). This resulted in a sparse feature vector for each utterance containing one dimension for each n-gram. Feature values were filled using TFIDF counts for a given transcript, computed as follows where TF is the term frequency within the transcript and DF is the number of documents containing the term:

$$TFIDF = TF \times \frac{1}{DF} \tag{1}$$

Each vector was L2-normalized with unit modulus. The final vocabulary size across all n-grams was 613.

3.2.2. Psycholinguistic Features

Psycholinguistic characteristics play a key role in verbal processing [20], and thus we suspected that they may have high utility for predicting MMSE scores. We considered four classic psycholinguistic properties (age of acquisition, concreteness, familiarity, and imageability), as well as sentiment scores. These features (five total, described further in Table 1) were all extracted from third-party lexical resources as token-level scores, which we then averaged across all tokens in a given transcript. Sentiment scores were obtained using NLTK's SentimentAnalyzer library,[1] and psycholinguistic scores were obtained from an open source repository[2] containing scores from multiple aspects of the MRC Psycholinguistic Database [21].

[1] `https://www.nltk.org/api/nltk.sentiment.html`

[2] `https://github.com/vmasrani/dementia\_classifier`

3.2.3. Discourse-Based Features

To model global discourse patterns across the entire transcript, we extracted an array of count-based features for discourse tags. These features include CHAT transcript [17] markers for different pause types (including filled pauses containing, e.g., *uh* or *umm*), word repetition, retracing (restarting the same phrase or segment), and incomplete utterances. Our full list of discourse-based features included: short_pause_count, long_pause_count, very_long_pause_count, word_repetition_count, retracing_count, filled_pause_count, and incomplete_utterance_count. We normalized these counts by the number of words uttered in the conversation. We also examined both word count and utterance count as features, ultimately dropping utterance count due to its high correlation ($r > 0.5$) with the former, but retaining word count, resulting in a total set of eight discourse features.

3.2.4. Acoustic Features

Finally, we extracted acoustic features due to their success in prior work on dementia detection [11, 22, 15, 4] and MMSE score prediction [4]. Specifically, we computed Mel Frequency Cepstral Coefficients (MFCCs) and extracted the first 14 MFCCs for each speech segment. We identified mean values, variance, skewness, and kurtosis for these features, and then computed the same for velocity and acceleration. This resulted in a total of 171 audio features for each segment.

3.3. Model

We designed separate models for our textual (lexicosyntactic, psycholinguistic, and discourse-based) and acoustic features due to underlying differences in how the data was handled. Since we extracted our acoustic features from local audio segments (maximum duration 10 seconds), we employed a segment-based model similar to that seen in the existing performance benchmark [4]. The model predicted individual MMSE scores for each discrete segment, and these scores were then averaged across an entire transcript to produce a transcript-level MMSE score. We employed a transcript-level model for our textual features since they were extracted from the transcript as a whole. We experimented with two high-performing statistical regression algorithms: Support Vector Regression (SVR) with a polynomial kernel, regularization parameter $C = 100$, and kernel coefficient γ="auto"; and Gaussian Process Regression (GP) with a squared exponential kernel, $\alpha = 0.1$, and optimizer restarts set to 10. All other parameters for the respective algorithms were kept at their default values.

To empirically validate the utility of our engineered features relative to neural alternatives, we also experimented with a fine-tuned DistilBERT sequence classification model [23] for the task. We illustrate the architecture for this model in Figure 1. The pre-trained DistilBERT tokenizer processes unseen tokens (e.g., discourse tags in our transcripts) as subword units, allowing it to make use of vocabulary not present in its original corpus. Input is thus tokenized and then encoded, and the resulting hidden representation is subsequently passed to a final fully-connected network, which applies linear transformations to the data to ultimately predict a single output neuron representing the predicted MMSE value for the specific patient.

4. Evaluation

We selected a diverse set of five models for entry to the ADReSS Challenge:

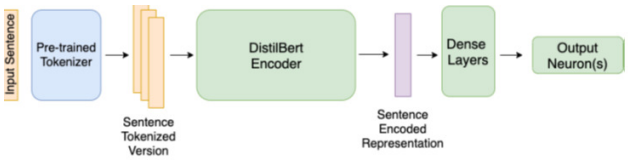

Figure 1: *Model Architecture for DistilBERT.*

- **ALL:** All textual features described in Section 3.2.
- **N-GRAM:** All lexicosyntactic features.
- **SELECTED-FEATURE:** A selection of the 90 highest-performing features from the training corpus. To obtain this feature subset, we employed a Random Forest regression model with 100 trees and selected features based on their mean decrease impurity (MDI), where impurity was measured as variance. We retained only features having MDI values exceeding a predefined threshold (10^{-3}). We show the top ten most important features measured using this process in Table 4.
- **DISTILBERT:** The DistilBERT model described in Section 3.3.
- **ACOUSTIC-ALL:** All acoustic features.

Although not entered into the ADReSS Challenge, we also experimented with a selection of the highest-performing acoustic features (ACOUSTIC-SELECTED), using the same feature selection technique as applied to SELECTED-FEATURE. We additionally ran some experiments using a late fusion neural network model to map acoustic and textual features to the same hidden space,[3] but the model performance was significantly lower than alternatives in the leave-one-out (LOO) experiment (RMSE>10). We report both our LOO cross-validation results on the training corpus, and our ADReSS Challenge results on the test data. We report both root mean squared error (RMSE) and R-squared values for the LOO setting, and RMSE for the results on the test data.

4.1. Results

We present the results from our LOO cross-validation experiment in Table 2, and our ADReSS Challenge results on the test set in Table 3. Our LOO experiment included both SVR and GP versions of each model; since SVR outperformed GP in more cases and we were limited to a batch of five results submissions, we submitted only SVR models (along with our DistilBERT alternative) to the ADReSS Challenge. Our best-performing model in the LOO experiment was ALL using an SVR classifier, achieving an RMSE of 4.97. Interestingly, ALL and ACOUSTIC-ALL exceeded the performance of SELECTED-FEATURE and ACOUSTIC-SELECTED, respectively, in the LOO experiments. Although ACOUSTIC-SELECTED was not entered in the ADReSS Challenge, this advantage did not persist for ALL vs. SELECTED-FEATURE on the test data. The R-squared values in Table 2 provide insight into the variance from the regression line; $R^2 = 0.52$ is considered moderate [25]. Our highest-performing model on the test set (Table

[3]Specifically, we encoded words using 300-dimensional English GloVe embeddings [24] and passed them to a bidirectional LSTM (Bi-LSTM) layer. We fed the acoustic features for each segment to a separate LSTM layer, and then we concatenated the resulting hidden representations of the Bi-LSTM and LSTM layers. We merged this concatenated vector with the vector of discourse-based features, and fed the merged vector into a feedforward layer with an output linear activation.

Table 2: *LOO results, formatted as RMSE (R^2).*

Features	SVR	GP
ALL	**4.97 (0.52)**	6.43 (-0.001)
N-GRAM	5.00 (0.514)	5.60 (0.225)
SELECTED-FEATURE	5.49 (0.415)	5.31 (0.451)
ACOUSTIC-ALL	6.59 (-0.093)	6.71 (-0.135)
ACOUSTIC-SELECTED	7.67 (-0.481)	7.31 (-0.271)

Table 3: *Test set results.*

Features	RMSE
ALL	4.87
NGRAM	4.61
SELECTED-FEATURE	**4.34**
DISTILBERT	4.63
ACOUSTIC-ALL	6.42

3) employed the SELECTED-FEATURE subset with SVR. This model (RMSE=4.34) outperformed the best-performing baseline model on the test set (RMSE=5.20 [4]) by 16.5%.

4.2. Analysis

We analyzed trends in RMSE scores across binned MMSE score groups to identify weaknesses in our best model and areas for potential improvement, and present our findings in Figure 2. We found that in general our model's predictive power was best for high MMSE scores, which is likely an artifact of the training set distribution—although samples in the ADReSS Challenge dataset are balanced across age and gender, they are not evenly distributed across the MMSE score continuum.

We also sorted the features in SELECTED-FEATURE in descending order based on their MDI importance score to analyze the strongest identified patterns, and present the top ten features in Table 4. Interestingly, we found many non-content function words and discourse connectives in this list, along with some discourse-based count features. In general, individuals with higher MMSE scores created longer descriptions of the picture and used more content words and complex phrases (e.g. *fall, cookie jar and*), whereas those with lower MMSE scores used shorter descriptions and more pauses and filler words. This provides evidence that verbal disfluency markers are important indicators of cognitive status, and also supports our hypothesis that a wide range of features can be productively leveraged in tandem for this task.

5. Discussion and Conclusion

Overall, we found text-based features to be more informative than acoustic features for the MMSE score prediction task. We speculate that this may be an important distinction between this and the dementia classification task, for which acoustic features have achieved considerable success [11, 15]. Our source code

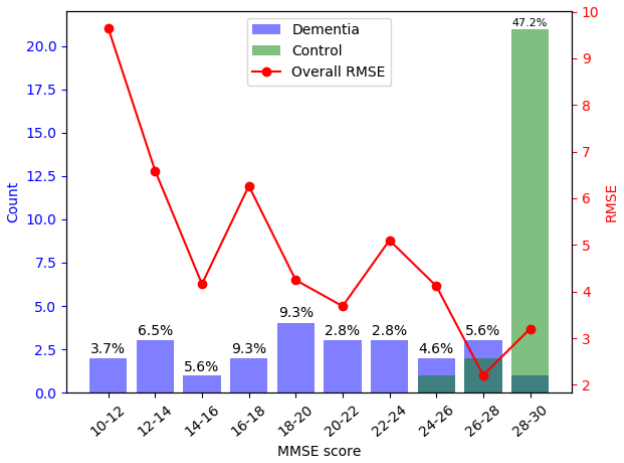

Figure 2: *Binned MMSE scores and frequency counts, with corresponding average RMSE per bin. Frequency counts (left y-axis and associated histogram bars) and RMSE (right y-axis and associated line graph) are for test instances, whereas percentages above histogram bars indicate the corresponding training set frequency for the same MMSE bins.*

Table 4: *Top 10 features based on MDI importance.*

Features	Importance
this	0.284
here	0.050
word_count	0.044
fall	0.037
well	0.034
laughs (non-verbal)	0.034
short_pause_count	0.021
in the	0.015
cookie jar and	0.014
it uh	0.013

is publicly available.[4] Further investigation into more informative features (e.g., acoustic disfluency markers) from the normalized speech signal could potentially transfer insights from our text-based features to high-performing acoustic analogues. Likewise, we are interested in leveraging the segment-based model with the text transcripts (casting utterances as segments). Finally, while automated MMSE score prediction may make testing more accessible, reliable, and resource-effective, future work could additionally explore more precise measures such as the Montreal Cognitive Assessment (MoCA) or the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) [26, 27], which have higher sensitivity than the MMSE to subtle changes in cognitive decline.

6. Acknowledgements

We thank Flavio Di Palo for contributing the DISTILBERT model, and the anonymous reviewers for their helpful feedback.

[4] https://github.com/treena908/MMSE-Prediction

7. References

[1] V. Fuster, "Changing demographics," *J. Am. Coll. Cardiol*, vol. 69, no. 24, pp. 3002–3005, 2017. [Online]. Available: http://www.onlinejacc.org/content/69/24/3002

[2] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""mini-mental state": A practical method for grading the cognitive state of patients for the clinician," *Journal of Psychiatric Research*, vol. 12, no. 3, pp. 189 – 198, 1975. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0022395675900266

[3] R. N. Jones and J. J. Gallo, "Education and Sex Differences in the Mini-Mental State Examination: Effects of Differential Item Functioning," *The Journals of Gerontology: Series B*, vol. 57, no. 6, pp. P548–P558, 11 2002. [Online]. Available: https://doi.org/10.1093/geronb/57.6.P548

[4] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge," in *Proceedings of INTERSPEECH 2020*, Shanghai, China, 2020. [Online]. Available: https://arxiv.org/abs/2004.06833

[5] S. Farzana, M. Valizadeh, and N. Parde, "Modeling dialogue in conversational cognitive health screening interviews," in *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*. Marseilles, France: European Language Resources Association, May 11-16, 2020 2020.

[6] I. Hoffmann, D. Nemeth, C. D. Dye, M. Pákáski, T. Irinyi, and J. Kálmán, "Temporal parameters of spontaneous speech in alzheimer's disease," *International Journal of Speech-Language Pathology*, vol. 12, no. 1, pp. 29–34, 2010, pMID: 20380247. [Online]. Available: https://doi.org/10.3109/17549500903137256

[7] A. Satt, R. Hoory, A. König, P. Aalten, and P. H. Robert, "Speech-based automatic and robust detection of very early dementia," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[8] L. Tóth, G. Gosztolya, V. Vincze, I. Hoffmann, G. Szatlóczki, E. Biró, F. Zsura, M. Pákáski, and J. Kálmán, "Automatic detection of mild cognitive impairment from spontaneous speech using asr," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] K. Fraser, F. Rudzicz, N. Graham, and E. Rochon, "Automatic speech recognition in the diagnosis of primary progressive aphasia," in *Proc. of the 4th Workshop on Speech and Language Processing for Assistive Technologies*. Grenoble, France: Assoc. for Computational Linguistics, 2013, pp. 47–54. [Online]. Available: https://www.aclweb.org/anthology/W13-3909

[10] S. O. Orimaye, J. S.-M. Wong, and K. J. Golden, "Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Baltimore, Maryland, USA: Association for Computational Linguistics, Jun. 2014, pp. 78–87. [Online]. Available: https://www.aclweb.org/anthology/W14-3210

[11] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[12] D. Weissenbacher, T. A. Johnson, L. Wojtulewicz, A. Dueck, D. Locke, R. Caselli, and G. Gonzalez, "Automatic prediction of linguistic decline in writings of subjects with degenerative dementia," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 1198–1207. [Online]. Available: https://www.aclweb.org/anthology/N16-1143

[13] F. Di Palo and N. Parde, "Enriching neural models with targeted features for dementia detection," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 302–308. [Online]. Available: https://www.aclweb.org/anthology/P19-2042

[14] M. Yancheva, K. Fraser, and F. Rudzicz, "Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias," in *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*. Dresden, Germany: Association for Computational Linguistics, Sep. 2015, pp. 134–139. [Online]. Available: https://www.aclweb.org/anthology/W15-5123

[15] J. Weiner, M. Angrick, S. Umesh, and T. Schultz, "Investigating the effect of audio duration on dementia detection using acoustic features," *Proceedings of Interspeech 2018*, pp. 2324–2328, 2018.

[16] S. Karlekar, T. Niu, and M. Bansal, "Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 701–707. [Online]. Available: https://www.aclweb.org/anthology/N18-2110

[17] B. MacWhinney, *The CHILDES Project: Tools for analyzing talk. transcription format and programs*. Psychology Press, 2000, vol. 1.

[18] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 06 1994. [Online]. Available: https://doi.org/10.1001/archneur.1994.00540180063015

[19] C. Roth, *Boston Diagnostic Aphasia Examination*. New York, NY: Springer New York, 2011, pp. 428–430. [Online]. Available: https://doi.org/10.1007/978-0-387-79948-3_868

[20] T. Salsbury, S. A. Crossley, and D. S. McNamara, "Psycholinguistic word information in second language oral discourse," *Second Language Research*, vol. 27, no. 3, pp. 343–360, 2011. [Online]. Available: https://doi.org/10.1177/0267658310395851

[21] M. Coltheart, "The mrc psycholinguistic database," *The Quarterly Journal of Experimental Psychology Section A*, vol. 33, no. 4, pp. 497–505, 1981.

[22] S. Al-Hameed, M. Benaissa, and H. Christensen, "Detecting and predicting alzheimer's disease severity in longitudinal acoustic data," in *Proceedings of the International Conference on Bioinformatics Research and Applications 2017*, ser. ICBRA 2017. New York, NY, USA: Association for Computing Machinery, 2017, p. 57–61. [Online]. Available: https://doi.org/10.1145/3175587.3175589

[23] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[24] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://www.aclweb.org/anthology/D14-1162

[25] H. Jörg, R. C. M., and S. R. R., *The use of partial least squares path modeling in international marketing*, ser. Advances in International Marketing. Emerald Group Publishing Limited, Jan 2009, vol. 20, pp. 277–319. [Online]. Available: https://doi.org/10.1108/S1474-7979(2009)0000020014

[26] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, "The montreal cognitive assessment, moca: A brief screening tool for mild cognitive impairment," *Jour. of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1532-5415.2005.53221.x

[27] A. R. Loughan, S. E. Braun, and A. Lanoye, "Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): preliminary utility in adult neuro-oncology," *Neuro-Oncology Practice*, vol. 6, no. 4, pp. 289–296, 12 2018. [Online]. Available: https://doi.org/10.1093/nop/npy050