Careful Whisper - leveraging advances in automatic speech recognition for robust and interpretable aphasia subtype classification

Laurin Wagner[1], Mario Zusag[1], Theresa Bloder[1]

[1]myReha GmbH, Austria

laurin@myreha.ai, mario@myreha.ai, tbloder@myreha.ai

Abstract

This paper presents a fully automated approach for identifying speech anomalies from voice recordings to aid in the assessment of speech impairments. By combining Connectionist Temporal Classification (CTC) and encoder-decoder-based automatic speech recognition models, we generate rich acoustic and clean transcripts. We then apply several natural language processing methods to extract features from these transcripts to produce prototypes of healthy speech. Basic distance measures from these prototypes serve as input features for standard machine learning classifiers, yielding human-level accuracy for the distinction between recordings of people with aphasia and a healthy control group. Furthermore, the most frequently occurring aphasia types can be distinguished with 90% accuracy. The pipeline is directly applicable to other diseases and languages, showing promise for robustly extracting diagnostic speech biomarkers.

Index Terms: speech recognition, aphasia classification, interpretable speech biomarkers

1. Introduction

Aphasia is defined as an acquired language disorder following brain injury [1], for instance, after stroke or traumatic head injury. Following a stroke, the most frequently occurring aphasia types are global, anomic, Wernicke's, and Broca's aphasia [2]. Kirshner and Wilson [3] summarized the characteristics according to these syndromes. In Broca's aphasia, the speech pattern is nonfluent, often referred to as "agrammatic" or "telegraphic". Patients with Broca's aphasia often make phonemic errors that are inconsistent from utterance to utterance. In contrast, the speech pattern of patients with Wernicke's aphasia is effortless, sometimes even overly fluent, containing verbal paraphasias, neologisms, and jargon productions. Auditory comprehension is impaired, which often renders patients unconscious to their nonsense speech productions. Global aphasia might be considered a summation of the symptoms observed in Broca's and Wernicke's aphasia, characterized by nonfluent speech paired with poor comprehension skills. Finally, in anomic aphasia, the main deficit concerns word access or retrieval from the mental lexicon which is characterized by pauses and circumlocutions. The distinction between these four aphasia types is, however, not always that clear cut and, several mixed aphasia types exist. In clinical practice, therapeutic strategies to approach any language-related deficits have to be aligned with the particular symptoms to maximize rehabilitative potential [4]. The accurate classification of aphasia is therefore highly relevant. However, the process of correctly evaluating these subtleties of speech and language can be quite time-consuming. Employing automated approaches on speech recordings with interpretable insights would enable clinicians to initiate speech therapy more quickly, thus expediting rehabilitation.

Several attempts have been made to classify different aphasia types and other speech pathologies from manual transcripts [5, 6]. Fraser et al. [5] conducted a study with the objective of discerning between individuals diagnosed with semantic dementia, progressive nonfluent aphasia and healthy controls, utilizing features extracted from manual transcripts for syntactic complexity and measures of fluency. Fromm et al. [6] employed the CHAT format manual transcriptions [7] to extract features on subtasks of AphasiaBank [8] data and study k-means clusters of these features. Other works have developed approaches to detect aphasia automatically from spontaneous speech [9] [10]. Notably, Chatzoudis et al. [11] use the XLSR-53 model from Conneau et al. [12], which was pre-trained on multiple languages using the contrastive self-supervised task and wav2vec2.0 architecture from Baevski et al. [13] to extract language-invariant linguistic features for classifying aphasia. These models learn high-quality audio representations using Transformer encoder blocks [14] and are successively fine-tuned for downstream tasks. To fine-tune these models, Conneau et al. [12] have added a classifier representing the output vocabulary of the respective downstream task on top of the XLSR-53 model by training on labeled datasets using a Connectionist Temporal Classification (CTC) loss [15].

Even though the underlying wav2vec2.0 architecture contains multiple Transformer layers, which can encode contextual information, fine-tuned XLSR-53 models without an additional language model tend to produce more literal, acoustic outputs than the encoder-decoder Whisper model from Radford et al. [16], therefore leaving filler words and recurring utterances as artifacts. Whisper was trained in a supervised fashion on weakly-supervised large-scale raw text labels without significant text standardization, such that the decoder part of Whisper works as an audio-conditioned language model and eliminates many filler words, recurring utterances and other artifacts. We propose to combine the output of these two models and apply carefully selected feature extraction methods on aligned and annotated transcripts to build prototypes of healthy speech. These prototypes are used to normalize features for inference in order to create interpretable insights for anomalous speech recordings.

The contributions of our work are as follows:

- With a fully automated approach, we achieve human-level accuracy on distinguishing speech recordings of aphasic patients from a healthy control group.
- To the best of our knowledge, it is the first approach exploiting the different decoding mechanisms of CTC and encoder-decoder based methods to create rich annotations for automatic feature extraction.

- With a fully automated approach, we achieve over 90% accuracy on classifying speech between a healthy control group, anomic, Wernicke's and Broca's aphasia.
- We provide several unseen scores for improving the accuracy of automatically assessing speech anomalies.

2. Proposed method

In this section, we describe the proposed pipeline for creating robust speech characteristics for assessing speech impairments. The pipeline consists of a prototype construction and classification phase. Both phases use feature extraction methods in three successive steps with the difference that the prototype phase uses extracted features from a population of healthy speech samples to estimate healthy speech distributions, while the classification phase uses distance measures from these distributions to generate classifiable features. We have used the 89 hours of free speech samples from the Europarl speech translation corpus [17] to construct the healthy prototypes.

The three successive steps consist of first constructing acoustic and clean transcripts leveraging the speech recognition models XLSR-53 and Whisper on speech chunks as described in Section 2.1. We have used the large version of XLSR-53 with 24 Transformer encoder blocks and the large version of Whisper. Second, we enrich the aligned transcripts with fully-automated annotations for pauses, filler words, and part-of-speech (POS) tags. Third, we are computing scores for coherence, fluency, syntax, lexical richness, and pronunciation as described in Section 2.2. For classification, speech recordings are fed into the alignment, annotation and score pipeline to compute the distance between each score and the score prototypes. These distances are then used for training simple machine learning classifiers and are used for the experiments in Section 3. The steps for computing prototypes are portrayed in Figure 1.

2.1. Alignment and annotation

The whole pipeline operates directly on speech waves, which are sampled at 16 kHz. The speech wave is fed into Whisper to produce syntactically correct transcripts containing punctuations, removing single utterances, repetitions and other artifacts, and into XLSR-53 to produce the more literal and standardized outputs. In the following, we will call the decoded XLSR-53 output *acoustic* and the Whisper output *clean* transcripts. Due to the configuration of wav2vec2.0, tokens are decoded with a 20 ms stride resulting in precise timing information on token level and by extension on word level [18]. Whisper provides timings on a sentence level only.

We combine the strengths of these two transcripts by aligning their standardized output via the minimization of the Levenshtein distance [19] as portrayed in Figure 1. This enables us to identify unmatched words or utterances and tokens as potential filler words or recurring utterances and allows us to augment the clean transcripts with accurate timings for the matched tokens from the acoustic transcripts in a fully-automated fashion.

The recordings were segmented into contiguous speech chunks, each containing a minimum of 200 words, based on clean transcriptions. To ensure accurate syntax scores, we confirmed that the final word in each chunk corresponded to the end of a sentence, with Whisper providing punctuation marks. The aligned acoustic, accurate timings, and clean text data were used in our annotation process, illustrated in Figure 1. In this step, we are augmenting the aligned transcripts with POS tags using RoBERTa from Liu et al. [20], detect pauses from silence tokens with more than 300 ms of duration and filler words from timings and unmatched utterances. These augmented transcripts are then fed into our score pipeline, as described in Section 2.2.

2.2. Scores for language abilities

The goal of the proposed scores is to capture multiple aspects of language abilities as shown in Table 1. We accumulate the distribution of each score and approximate it with a normal distribution, since they fit the empiric distributions well and serve as prototypes for healthy speech.

Table 1: *Scores for identifying speech characteristics.*

Language ability	Feature
Fluency	Words and phonemes per second
	Percentage time spoken
	Productive time ratio
	Pause length and distance
	Pause per word
	Mean phoneme length nouns
Lexical richness	TTR, MATTR
	gzip ratio
	HD-D, MTLD
	Word information
Syntax	Noun, verb and adjective ratio
	Grammar acceptance
	Mean sentence length
Pronunciation	Word error rate acoustic
	Character error rate acoustic
Coherence	CTRLEval
	Word vector coherence
	GPT2 perplexity

2.2.1. Fluency

The **words per second** score is computed on clean transcripts, while **percentage time spoken** uses the aggregated time of the acoustic transcripts divided by the total time of the provided audio chunk. For measuring the **productive time ratio**, we divide the strict spoken time (discounting pauses) of the acoustic transcript by the strict spoken time of the clean transcript. We count every successive set of silence tokens from acoustic transcripts that are longer than 300 ms as a pause to compute **pause length** scores. Sampling all pause lengths as a distribution gives insights into the frequency and duration of speech breaks. We use different quantiles (10, 25, 50, 75, 95) of this distribution as features. **Pause distance** by contrast computes the time between successive pauses and again, samples the pause distances as a distribution and uses quantiles (10, 25, 50, 75, 95) as features. The **pause per word** score is defined as the ratio of pauses longer than 300 ms and the total number of words from the clean transcript. We also included two phoneme scores, for which we use the automatic phonemizer from Bernard and Titeux [21] to transform the clean and acoustic text into phonemes. For **mean phoneme length nouns**, we aggregate and average the lengths of all words tagged as nouns from clean transcripts. Finally, the **phonemes per second** score is computed on the phonemized version of clean transcripts.
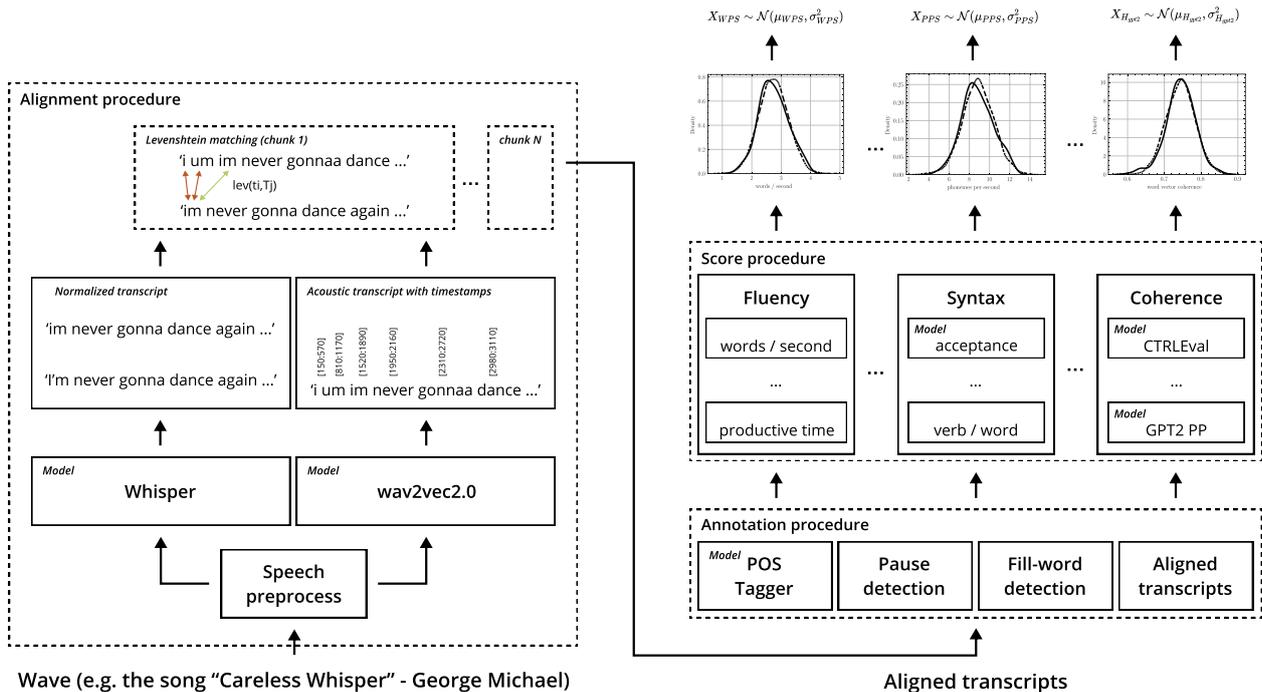
Figure 1: *Alignment, annotation and score procedure*

2.2.2. Lexical richness

All lexical richness scores are computed on clean transcripts. We have computed **type token ratio** (TTR) and **MATTR** [22] with window sizes of 10, 25 and 50 words. Using the intuition that redundant text is easier to compress, we have used the gzip algorithm, creating **gzip ratio** as a measurement for redundancy, dividing the length of compressed by uncompressed byte string representations. Further we use **HD-D** and **MTLD** as described in [23], which were proven to capture lexical diversity well, even under varying text length. **Word information** is defined as the entropy of the normalized token count distribution of the lemmatized tokens [24].

2.2.3. Syntax

All syntax scores are computed on clean transcripts. **Noun, verb and adjective ratio** are simply the ratio of the specified POS tags to the number of words in the clean transcript. The **grammar acceptance** feature is defined as the softmax probability of a RoBERTa-large model [20] fine-tuned for grammar acceptability on the CoLA Corpus [25]. Finally, we use the **mean sentence length** as an indicator of noticeable short or sprawling phrases.

2.2.4. Pronunciation

The **word/character error rate acoustic** scores are the word/character error rates between the acoustic and clean transcripts, intuitively capturing speech intelligibility proportional to the diverging outputs of Whisper and XLSR-53.

2.2.5. Coherence

All coherence scores are evaluated on Whisper transcripts. **CTRLEval** is a reference-free metric to assess the coherence for controlled text generation tasks as detailed by Ke et al. [26].

For **word vector coherence**, we use the cosine similarity to assess the local coherence between averaged word vectors of neighboring sentences. Word vectors are taken from spaCy's [27] en_core_web_sm corpus. At last, with the **GPT2 perplexity** score, we evaluate the perplexity from GPT2 by Radford et al. [28] with a sliding window size of 512 tokens using the standard GPT2 tokenizer.

3. Experiments and results

3.1. Experimental setup

To validate our approach, we are using speech samples extracted from the AphasiaBank database [8] for English. AphasiaBank comprises a collection of interviews between clinicians and subjects afflicted with aphasia, as well as healthy control subjects. These interviews are transcribed in the CHAT transcription format [7] and include timestamps for interviewer and patient speech segments. Most interviews contain a label indicating which form of aphasia the subject is suffering from. There are 272 individuals from the control group, 136 with anomic, 80 with Broca's, 60 conduction, 29 with Wernicke's, 9 with transmotor, 3 individuals with global aphasia and 35 individuals who do not exhibit symptoms of aphasia as determined by the Western Aphasia Battery (WAB-R) [29]. For 348 subjects the (WAB-R) Aphasia Quotient (AQ) is included. We remove all speech uttered by the interviewer as given by the timestamps from the CHAT protocols and concatenate the resulting audio clips, which are input to our pipeline. For each speech chunk, we compute the scores from Section 2.2 as outlined in Figure 1 and average multiple score vectors from the same interview. The distance for each averaged score s_i to the healthy prototype of this score, estimated by $X \sim \mathcal{N}(\mu_s, \sigma_s^2)$, are then used as features for classification and computed by $s_i = \frac{\sigma_s}{|\mu_s - s_i|}$, if $|s_i - \mu_s| > \sigma_s$ else it is set to 1.

To ensure comparability with the results from Chatzoudis et al. [11], we adopt a Leave-One-Subject-Out (LOSO) cross-validation strategy for all classification experiments. We use a standard linear Support Vector Classifier SVC [30] with regularization parameter $C = 0.1$ and max_iter $= 50000$ for classification. We do not use task-specific feature selection and train on all of our features. All experiments were run on a virtual machine using an NVIDIA-A100 GPU.

3.2. Results

3.2.1. Aphasia versus Control

To facilitate a direct comparison with Chatzoudis et al. [11], we have included the classification results of our approach for aphasic (including all subtypes) versus healthy control patients in Table 2. The outcome of this analysis reveals that our fully automatic method exhibits superior performance relative to the one of [11], even when comparing ourselves to features extracted directly from the manually annotated CHAT transcriptions. Furthermore, since we classify on averaged chunks of similar length, instead of the full transcript, we are not benefiting from the discriminatory information provided by total word output. These findings provide compelling evidence of the robustness and efficacy of our fully-automatic approach for real-world applications.

Table 2: *Classification metrics for aphasia versus healthy control group using linguistic features on AphasiaBank.*

Method	Manual annotation	Accuracy	F_1
SVC [11]	✗	93.9	-
SVC [11]	✓	97.4	-
SVC [31]	✓	-	85.9
Ours (SVC)	**✗**	**98.6**	**98.5**

3.2.2. Aphasia subtype classification

To validate the effectiveness and robustness of the feature extraction pipeline, we use a linear SVC to differentiate three types of aphasia and the healthy control group as a multi-class classification problem. We have used three of the four main aphasia subtypes, i.e. anomic, Wernicke's, and Broca's aphasia. We exclude global aphasia since there are only 3 people with this diagnosis in AphasiaBank. Further details on the classification performance are summarized in Table 3.

Table 3: *Classification metrics for a 4-class classification problem differentiating between aphasia subtypes and the control group of the AphasiaBank dataset.*

Class	Precision	Recall	F_1
Control	97.5	98.5	98.0
Anomic	83.1	86.8	84.9
Broca	80.0	75.0	77.5
Wernicke	91.7	68.8	78.6
Weighted average	90.6	90.6	90.6

3.2.3. Classification results per score category

Each score category should contribute meaningful information to the multi-class subtype classification problem from section 3.2.2. To this end, we trained a linear SVC that only utilizes the subset of scores from the respective score category to differentiate the three aphasia subtypes and the healthy control group. The results for the weighted average precision, recall and F_1 scores for the aphasia subtypes and control group per score category are condensed in Table 4.

Table 4: *Weighted average precision, recall and F_1 scores for a 4-class classification problem differentiating between aphasia subtypes and the control group of the AphasiaBank dataset using only features from a single score category.*

Score Category	Precision	Recall	F_1
Fluency	81.8	82.0	81.6
Coherence	68.9	71.0	69.1
Lexical Richness	76.5	77.9	76.5
Syntax	76.9	78.3	77.4
Pronunciation	61.7	67.3	63.9

3.2.4. WAB-R AQ Regression

The automatic estimation of the aphasia quotient (AQ) from spontaneous speech has numerous potential benefits, including progress monitoring without the need for frequent repetition of the WAB-R assessment procedures. We follow the evaluation procedure from Le et al. [32] and frame the WAB-R AQ prediction as a regression problem, with our proposed score distances as features and the AQ scores as the target labels for the 348 individuals, for which the AQ was included. The results are portrayed in Table 5.

Table 5: *WAB-R AQ Prediction. MAE=Mean Absolute Error, PC = Pearsons's Correlation, SVR = Support Vector Regression*

Method	Manual annotation	PC	MAE
SVR [32]	✗	0.776	9.90
SVR [32]	✓	0.787	9.54
Ours (SVR)	**✗**	**0.815**	**8.19**

4. Conclusions

The proposed approach showed high level of robustness for identifying aphasia and aphasic subtypes from free speech recordings and presents a pipeline that can be adapted to other diseases and languages with minimal effort. We used the same pipeline to evaluate the proposed set of features for a dementia classification challenge, beating the baseline by a large margin. In the near future, we want to investigate the distance measures from healthy prototypes to build accurate and interpretable speech biomarkers for multiple diseases. Also, the optimal number of words per chunk for practical applications, the transferability of features across multiple languages and the correlation between multiple features need further investigation.

5. References

[1] M. V. Ivanova and N. F. Dronkers, "Aphasia: How our language system can "break"," *Frontiers for young minds*, vol. 10, 2022.

[2] P. M. Pedersen, K. Vinter, and T. S. Olsen, "Aphasia after stroke: type, severity and prognosis," *Cerebrovascular diseases*, vol. 17, no. 1, pp. 35–43, 2004.

[3] H. S. Kirshner and S. M. Wilson, "Aphasia and aphasic syndromes," *Bradley's Neurology in Clinical Practice E-Book*, vol. 133, 2021.

[4] J.-F. Landrigan, F. Zhang, and D. Mirman, "A data-driven approach to post-stroke aphasia classification and lesion-based prediction," *Brain*, vol. 144, no. 5, pp. 1372–1383, 2021.

[5] K. C. Fraser, J. A. Meltzer, N. L. Graham, C. Leonard, G. Hirst, S. E. Black, and E. Rochon, "Automated classification of primary progressive aphasia subtypes from narrative speech transcripts," *cortex*, vol. 55, pp. 43–60, 2014.

[6] D. Fromm, J. Greenhouse, M. Pudil, Y. Shi, and B. MacWhinney, "Enhancing the classification of aphasia: a statistical analysis using connected speech," *Aphasiology*, vol. 36, no. 12, pp. 1492–1519, 2022.

[7] B. MacWhinney, "The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database," 2000.

[8] B. MacWhinney, D. Fromm, M. Forbes, and A. Holland, "Aphasiabank: Methods for studying discourse," *Aphasiology*, vol. 25, no. 11, pp. 1286–1307, 2011.

[9] D. Le, K. Licata, and E. M. Provost, "Automatic quantitative analysis of spontaneous aphasic speech," *Speech Communication*, vol. 100, pp. 1–12, 2018.

[10] Y. Qin, T. Lee, and A. P. H. Kong, "Automatic assessment of speech impairment in cantonese-speaking people with aphasia," *IEEE journal of selected topics in signal processing*, vol. 14, no. 2, pp. 331–345, 2019.

[11] G. Chatzoudis, M. Plitsis, S. Stamouli, A.-L. Dimou, A. Katsamanis, and V. Katsouros, "Zero-shot cross-lingual aphasia detection using automatic speech recognition," 2022. [Online]. Available: https://arxiv.org/abs/2204.00448

[12] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[13] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[15] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[17] J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, J. Jorge, N. Roselló, A. Giménez, A. Sanchis, J. Civera, and A. Juan, "Europarl-st: A multilingual corpus for speech translation of parliamentary debates," 2019. [Online]. Available: https://arxiv.org/abs/1911.03167

[18] L. Kürzinger, D. Winkelbauer, L. Li, T. Watzel, and G. Rigoll, "CTC-segmentation of large corpora for german end-to-end speech recognition," in *Speech and Computer*. Springer International Publishing, 2020, pp. 267–278. [Online]. Available: https://doi.org/10.1007%2F978-3-030-60276-5_27

[19] V. I. Levenshtein *et al.*, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8. Soviet Union, 1966, pp. 707–710.

[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[21] M. Bernard and H. Titeux, "Phonemizer: Text to phones transcription for multiple languages in python," *Journal of Open Source Software*, vol. 6, no. 68, p. 3958, 2021. [Online]. Available: https://doi.org/10.21105/joss.03958

[22] M. A. Covington and J. D. McFall, "Cutting the gordian knot: The moving-average type–token ratio (mattr)," *Journal of quantitative linguistics*, vol. 17, no. 2, pp. 94–100, 2010.

[23] P. M. McCarthy and S. Jarvis, "Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment," *Behavior research methods*, vol. 42, no. 2, pp. 381–392, 2010.

[24] K. T. Cunningham and K. L. Haley, "Measuring lexical diversity for discourse analysis in aphasia: Moving-average type–token ratio and word information measure," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 3, pp. 710–721, 2020.

[25] A. Warstadt, A. Singh, and S. R. Bowman, "Neural network acceptability judgments," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 625–641, 2019.

[26] P. Ke, H. Zhou, Y. Lin, P. Li, J. Zhou, X. Zhu, and M. Huang, "CTRLEval: An unsupervised reference-free metric for evaluating controlled text generation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2306–2319. [Online]. Available: https://aclanthology.org/2022.acl-long.164

[27] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python," 2020.

[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[29] A. Kertesz, "Western aphasia battery–revised," 2007.

[30] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[31] A. Balagopalan, J. Novikova, M. B. A. McDermott, B. Nestor, T. Naumann, and M. Ghassemi, "Cross-language aphasia detection using optimal transport domain adaptation," 2019. [Online]. Available: https://arxiv.org/abs/1912.04370

[32] D. Le, K. Licata, and E. M. Provost, "Automatic quantitative analysis of spontaneous aphasic speech," *Speech Communication*, vol. 100, pp. 1–12, 2018.