How You Say It Matters: Measuring the Impact of Verbal Disfluency Tags on Automated Dementia Detection

Shahla Farzana, Ashwin Deshpande, and **Natalie Parde**
Natural Language Processing Laboratory
Department of Computer Science
University of Illinois at Chicago
{sfarza3, adeshp27, parde}@uic.edu

Abstract

Automatic speech recognition (ASR) systems usually incorporate postprocessing mechanisms to remove disfluencies, facilitating the generation of clear, fluent transcripts that are conducive to many downstream NLP tasks. However, verbal disfluencies have proved to be predictive of dementia status, although little is known about how various types of verbal disfluencies, nor automatically detected disfluencies, affect predictive performance. We experiment with an off-the-shelf disfluency annotator to tag disfluencies in speech transcripts for a well-known cognitive health assessment task. We evaluate the performance of this model on detecting repetitions and corrections or retracing, and measure the influence of gold-annotated versus automatically detected verbal disfluencies on dementia detection through a series of experiments. We find that removing both gold and automatically-detected disfluencies negatively impacts dementia detection performance, degrading classification accuracy by 5.6% and 3% respectively.

1 Introduction

As populations grow older worldwide, the number of people with Alzheimer's disease (AD) and related dementia is also on the rise (Alzheimer's Association, 2018). Significant changes to speech and language use caused by dementia occur early in disease progression (Bucks et al., 2000). Interesting case studies have demonstrated how diachronic analysis of patients' language use may reveal signs of dementia, using writing samples from British novelists Iris Murdoch, who ultimately perished with Alzheimer's, and Agatha Christie, who was suspected of it (Le et al., 2011). Numerous studies have also sought to automatically detect early signs of the disease and model its progression using speech and writing samples (Becker et al., 1994; Herd et al., 2014; Yancheva et al., 2015; Masrani, 2018; Di Palo and Parde, 2019; Zhu et al., 2019; Fraser et al., 2019; Eyre et al., 2020; Farzana and Parde, 2020; Sarawgi et al., 2020).

Although some studies have pointed to disfluency patterns as an important predictor of AD status (Lopez-de Ipina et al., 2017; Mueller et al., 2018), research in this area has been limited by several factors. Disfluency detection is a challenging and resource-intensive task in itself (Wang et al., 2017; Jamshid Lou and Johnson, 2017; Zayats and Ostendorf, 2019), and may lie out of scope for many interdisciplinary researchers already straddling boundaries between NLP and clinical practice (Valizadeh and Parde, 2022; Kaelin et al., 2021). Rich manual disfluency annotations are present in some datasets common in automated dementia detection (Becker et al., 1994), but off-the-shelf ASR systems do not typically transcribe disfluencies. Moreover, inconsistencies between automatically generated and gold standard transcripts may pose significant challenges for modeling dementia in real-world applications (Balagopalan et al., 2020b), for which ASR will be a necessary component of any speech-based pipeline.

We address these limitations, by investigating the impacts of automatically derived disfluencies on modeling cognitive decline. Our key contributions are as follows:

1. We experiment with an off-the-shelf disfluency detection model to automatically assign word- and phrase-level disfluency tags to samples from the most popular dementia detection dataset, focusing on repetitions and retraces.

2. We measure the influence of these disfluency types on the downstream task of dementia detection by systematically ablating gold-labelled and automatically tagged disfluencies from manual transcripts.

3. We compare AD classification performance on manually and automatically generated transcripts, and compare the removal of gold and

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 37–48

automatically detected disfluencies from manual transcripts, to investigate the influence verbal disfluencies have on dementia detection.

This analysis[1] not only paves the way for the discovery of approaches to automated dementia detection that are more suitable for realistic scenarios, but also enhances our understanding of the individual contributions of different disfluency types to this task. We report on related studies and provide relevant background for automatic disfluency detection in §2. We describe our datasets and task setup in §3, and detail our methods in §4. We report the results of our experiments in §5, and further analyze our findings in §6 before concluding in §7.

2 Related Work

2.1 Studies of Disfluency in the Context of Cognitive Decline

Disfluency, defined as any interruption in the normal flow of speech, is prevalent in spoken language. Verbal disfluency comprises several major subcategories: *false starts*, *repetitions*, *filled pauses* (e.g., "uh," "um," etc.), and *sentence corrections* (Shriberg, 1994). Although verbal and nonverbal (*unfilled pauses*) disfluencies are common in spontaneous speech, there is a fine line between normal and abnormal disfluencies. This boundary can be exploited to facilitate modeling cognitive decline.

Studies have found that verbal fluency is an effective indicator of cognitive decline, as fluency declines rapidly for subjects suffering from early stage Mild Cognitive Impairment (MCI) relative to healthy controls (Mueller et al., 2018). Researchers have previously leveraged both acoustic and transcript-based fluency features to automatically detect MCI (Lopez-de Ipina et al., 2017; Mueller et al., 2018). Another study revealed that anomic aphasic subjects tend to produce more disfluent speech than non-aphasic subjects during word retrieval tasks, when examining disfluencies or "stutterings" including part-word repetitions, vocal segregate repetitions, and prolongations (Brown and Cullinan, 1981).

Transcript-based normalized verbal disfluency features (e.g, *filled pause count*, *retracing count*, and *repetition count*) have proved to be discriminative in predicting outcomes from cognitive screening tests such as the Mini Mental State Examination (MMSE) and AD classification, as have concatenations of automatically detected verbal disfluency segments (e.g., *repair onset*, *edit term*, and *fluent words*) with word vectors (Farzana and Parde, 2020; Rohanian et al., 2020, 2021). Automatically extracted non-verbal disfluency features from both transcripts and speech (e.g., *silent pauses*, *speed of articulation*, and *pronounciation*) have also shown performance boosts in AD classification (Yuan et al., 2020; Qiao et al., 2021).

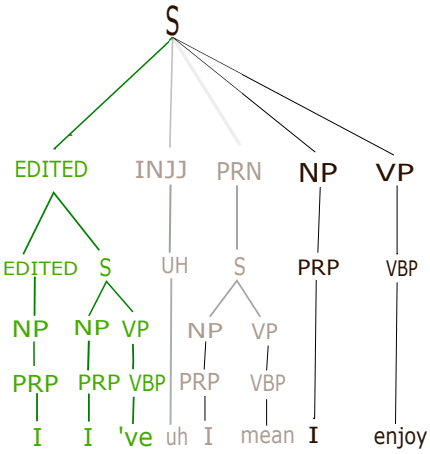

Figure 1: An example of gold labelled parse tree (Jamshid Lou et al., 2019).

2.2 Automatic Disfluency Detection

Disfluency detection is a key challenge in parsing transcribed speech. Disfluencies are defined structurally with three main components (Shriberg, 1994): the *reparandum*, the *interregnum* and the *repair*. The *reparandum* is replaced by the *repair* segment and the *interregnum* is an optional part of the structure consisting of filled pauses (e.g., "uh") and discourse connectives (e.g., "I mean"). We present an example disfluency with all three components present below:

$$\overbrace{I\ I\ 've}^{reparandum}\ \overbrace{uh I mean}^{interregnum}\ \overbrace{I enjoy}^{repair} \quad (1)$$

Disfluencies are further categorized into repetition, correction/retracing, and false start (Jamshid Lou and Johnson, 2020a), following established typology of speech repairs (Shriberg, 1994). Repetition contains identical *reparandum* and *repair* segments, whereas the *reparandum* and *repair* differ in correction/retracing. The latter is much harder to detect automatically.

Disfluency detection on pre-segmented utter-

[1] https://github.com/AshwinDeshpande96/Measuring_the_Impact_of_Verbal_Disfluency_Tags_on_Automated_Dementia_Detection

ances from the Switchboard treebank corpus (Godfrey and Holliman, 1993; Marcus et al., 1999) has been the focus of many prior works (Johnson and Charniak, 2004; Charniak and Johnson, 2001; Qian and Liu, 2013; Honnibal and Johnson, 2014). In the Switchboard corpus, reparanda, filled pauses, and discourse connectives are marked by EDITED, INTJ, and PRN labels respectively (illustrated in Figure 1). Conventional syntactic parsers often fail to capture the unconventional relation between reparandum and repair, where repair uses similar words to the reparandum in the same order, functioning as a "rough copy" rather than providing additional information (Johnson and Charniak, 2004; Charniak and Johnson, 2001). Because of the difficulty of addressing disfluency within the task of syntactic parsing, systems have instead been developed to detect and remove disfluency prior to parsing (Charniak and Johnson, 2001; Kahn et al., 2005; Lease and Johnson, 2006). Nonetheless, transition-based dependency parsers designed with special mechanisms to handle disfluencies have proven useful for detecting and removing disfluent words and their dependencies from sentences (Honnibal and Johnson, 2014; Rasooli and Tetreault, 2013; Yoshikawa et al., 2016; Tran et al., 2018). Moreover, encoder-decoder constituency parsing models using lexical and prosodic cues (Tran et al., 2018) have resulted in small performance gains both in parsing and disfluency detection. Augmenting parsing models with location-aware attention mechanisms has also been especially effective for disfluency detection (Tran et al., 2018).

Specialized disfluency detection models frame the problem as a sequence labelling task where each word in the input is labelled as disfluent or not. Neural models (CNNs and LSTMs) have been employed for this (Zayats et al., 2016; Jamshid Lou et al., 2018; Wang et al., 2016) but until recently have not performed very well. A recent state-of-the-art semi-supervised approach introduced a self-attentive model (Wang et al., 2018) that jointly performs syntactic parsing and disfluency detection.

The incremental approach for disfluency detection has been explored on both unsegmented and pre-segmented utterances from manual and automated transcripts using LSTM with different decoding schemes (Hough and Schlangen, 2015, 2017) leveraging joint and multitask settings. Another recent approach introduced the incremental processing of words to a Transformer model (BERT (Devlin et al., 2019)) to detect speech disfluency (Rohanian and Hough, 2021). However, these incremental approaches perform poorly on detecting *reparanda* of longer lengths.

3 Data and Task Setup

We used the ADReSS Challenge corpus for our experiments (Luz et al., 2020). The ADReSS Challenge corpus, developed as part of a shared task for INTERSPEECH 2020, is a benchmark dataset of spontaneous speech in the domain of AD classification and MMSE score prediction. It has been acoustically preprocessed, and is balanced in terms of age and gender. The data consists of audio recordings and manual transcriptions of spoken picture descriptions elicited from participants through the Cookie Theft task from the Boston Diagnostic Aphasia Exam (Roth, 2011). The corpus is a subset of the Pitt corpus,[2] which is itself a subset of the DementiaBank dataset (Becker et al., 1994).

In the Cookie Theft task, an investigator and a participant (in this case, an older adult) carry on a conversation in which the investigator asks the participant to describe what is depicted in an eventful image containing, among other subjects, a boy stealing a cookie from a cookie jar.[3] There is no specific time limit for the conversation, allowing participants to talk as long as they want. In the Pitt corpus and by extension the ADReSS Challenge corpus, these conversations were recorded and manually transcribed using the CHAT transcription protocol (MacWhinney, 2000). Participants were labelled as HC (healthy control with no cognitive decline) or AD (declined cognitively) based on their prior diagnostic test results.

We report the transcript-level mean utterance count and standard deviation (SD) for data collected from AD and HC participants in Table 1, showing that the lengths of conversations across groups were fairly balanced (HC $= 13.79 \pm 5.21$ utterances; AD $= 13.93 \pm 9.54$ utterances). We also report the mean MMSE score and SD for each speaker category, showing a significant difference in cognitive health between groups (HC $= 29.11 \pm 0.98$ MMSE; AD $= 17.06 \pm 5.46$ MMSE). To assess significance, we applied the Mann–Whitney U test (as the normality assump-

[2] https://dementia.talkbank.org/access/English/Pitt.html

[3] We refer interested readers to Karlekar et al. (2018), Mueller et al. (2018), or some others cited in this paper for a copy of the original image.

	AD	HC	Test Statistics
Utterance Count	13.93 (SD=9.54)	13.79 (SD=5.21)	U=135.0 p=0.25
MMSE Score	17.06 (SD=5.46)	29.11 (SD=0.98)	U=47.5 p=0.00

Table 1: Mean utterance count and MMSE score for the AD and HC groups, with standard deviations in parentheses. Statistical significance (p) for differences between groups is reported along with the Mann-Whitney U test statistic.

Ref.: and UM THAT 'S UH that 's about all i can see

Aligned: *** ** **** ** not **** ** ***** all i can see

Figure 2: The reference (*Ref.*) and aligned ASR output for a sample utterance from the ADReSS Challenge corpus. The reference transcript is human-transcribed speech with gold disfluent words (red, capitalized) and fluent words (black). *Aligned* refers to the desired alignment of ASR output with the reference text for making meaningful FER and DER evaluations (Jamshid Lou and Johnson, 2020a).

tion was violated) across the two speaker groups, and we also report the test statistic (U) and significance value (p) for each group in Table 1.

3.1 ASR Setup

We used the phone call enhanced model ($16khz$) of the Google Cloud-based Speech Recognizer to automatically transcribe the audio files in the ADReSS Challenge corpus to facilitate our comparisons of manually and automatically generated transcriptions. Manually segmented utterances were fed to the speech recognizer for transcription. The overall word error rate (WER) for the automatically generated transcripts was 69.47%. To evaluate more fine-grained performance of the speech recognizer, we estimated the fluent and disfluent error rates (FER and DER). We provide the equations for computing both below, where d_f, s_f, i_f, and n_f refer to the number of deleted, substituted, inserted, and total fluent words, respectively, and d_d, s_d, i_d, and n_d refer to the number of deleted, substituted, inserted, and total disfluent words, respectively (Jamshid Lou and Johnson, 2020a):

$$\text{FER} = \frac{d_f + s_f + i_f}{n_f} \tag{2}$$

Group	FER	DER
AD	53.30%	77.60%
HC	47.30%	80.70%
Overall	50.20%	78.80%

Table 2: Rates of ASR error on the ADReSS Challenge dataset, both at the class level (AD and HC) and overall. For DER calcualtion, we consider all the disfluencies in Table 6 as well as the Filled pauses (e.g. *uh, um*)

	Repetition	Retracing
DER	76.50%	61.10%

Table 3: DER of broad disfluency categories (repetition and retracing, as defined in Table 6).

$$\text{DER} = \frac{d_d + s_d + i_d}{n_d} \tag{3}$$

To calculate DER,[4] we considered *word repetition*, *multiple repetition*, *phrase repetition*, *word retracing*, and *phrase retracing*, with additional details regarding each disfluency type provided in Table 4. We show an alignment between gold and automatically generated transcriptions for an example utterance from the ADReSS Challenge corpus in Figure 2. Computing FER for this example would set $d_f = 4$, $s_f = 0$, $i_f = 0$, and $n_f = 8$, resulting in FER=0.5. Computing DER for the same sample would set $d_d = 3$, $s_d = 1$, $i_f = 0$, and $n_f = 4$, resulting in DER=1.0. We report FER and DER across the ADReSS Challenge corpus for AD, HC, and all participants in Table 2 and the breakdown of DER for broad disfluency types (repetition, encompassing *word repetition*, *multiple repetition*, and *phrase repetition*, and retracing, encompassing *word retracing* and *phrase retracing*) in Table 3.

3.2 Disfluency Annotator Setup

We leverage the self-attentive neural parsing model (Jamshid Lou and Johnson, 2020b) to automatically detect disfluencies in the ASR-generated transcripts. The model is trained to jointly parse and detect disfluency using contextualized word embeddings (BERT (Devinney et al., 2020) or ELMO (Peters et al., 2018)) and currently produces state-of-the-art performance with a parsing accuracy of

[4] Although the original DER formulation counts the number of copies, we replace this with the number of deletions since we expect the ASR to transcribe disfluent as well as fluent words.

93.9% and a disfluency detection F_1-score of 0.924 on the Switchboard development set (Jamshid Lou and Johnson, 2020b) in the joint task. We use the pretrained version of the disfluency detector and parser.[5] This version is self-trained on the Switchboard gold parse trees (Marcus et al., 1999) and Fisher Corpus Part 1 (Cieri et al., 2004) and Part 2 (Cieri et al., 2005) silver parse trees, using *BERT-base-uncased* word representations.

4 Methods

4.1 Verbal Disfluency Types

We consider several disfluency types in this investigation: *word repetition*, *phrase repetition*, *word retracing*, and *phrase retracing*. We limit our scope to these disfluency types for two primary reasons: (1) these verbal disfluency types are annotated in our corpus of interest, and (2) automatic detection of these types is challenging. We provide examples of each of these in Table 4.[6] Word and phrase repetition indicate repeated utterance of the same word or phrase in such a way that is disfluent with the natural flow of speech, whereas word and phrase retracing indicate verbal "backtracking" to correct a previously uttered word or phrase. In Table 5, we report the frequencies of these disfluency types across speaker groups.

Disfluency Type	Example
Word Repetition	**the [/]** the cabinet door has just swung open
Multiple Repetition	there's nothing going on outside there's just **bushes [x 3]**.
Phrase Repetition	⟨**what are**⟩ **[/]** what are the instructions ?
Word Retracing	and there are **dishes [/ /]** &uh &uh two cups and a saucer on the sink
Phrase Retracing	and outside the window there's a ⟨**walk with a**⟩ **[/ /]** &c curved walk with a garden .

Table 4: Example of different types of disfluencies from transcripts annotated using the CHAT protocol (MacWhinney, 2000). Disfluencies are bold-faced followed by disfluency markers. Angle brackets indicate phrase-level disfluencies, whereas [x n] indicates that the word before the marker is repeated n times.

4.2 Automatic Disfluency Annotation

We leveraged the self-attentive neural disfluency annotator described in §3.2, trained on the Penn Treebank-3 SWBD corpus (Marcus et al., 1999) and the Fisher I and II corpora (Cieri et al., 2004, 2005) using a semi-supervised approach (Jamshid Lou and Johnson, 2020b). This multi-task learning setup enables the model to predict both parse trees and disfluency tags for utterances. The disfluency annotator adds word-level annotations to disfluent words, or those acting as *EDITED*, *INTJ*, or *PRN* nodes (illustrated in Figure 1).

We preprocessed both the reference and ASR-generated transcripts by removing punctuation and (for the reference transcripts) existing disfluency tags. We then fed the disfluency annotator one utterance per line, in turn producing both a parse tree and a disfluency-tagged version of the utterance as output. Figure 3 shows an example utterance with: (1) the actual text and disfluency tags from the ADReSS Challenge corpus, considering the disfluency types referred in Table 4; (2) the gold disfluency tags formatted as the expected output from the automatic disfluency annotator; and (3) the predicted word-level disfluency tags from the automatic disfluency annotator. Phrase repetition accuracy for the utterance in Figure 3 would be 100% as both the words in the repeated phrase (highlighted in red) are predicted correctly, whereas phrase retracing accuracy would be 0%, as no words in the retraced phrase (highlighted in blue) are predicted as disfluent.

Table 6 illustrates the performance of the automatic disfluency annotator at predicting different disfluency types for the ADReSS Challenge training set, providing evidence that retracing/correction (especially at the phrase level) is harder to predict than repetition. The annotator often fails to detect cases of *multiple repetition* (accuracy=11.11%, making it lowest among all disfluency types in Table 6), likely because it was intermixed with word-level repetition in the training data.

4.3 Disfluency Removal

We implement two methods for removing disfluencies from transcribed speech, described further in

[5] https://github.com/pariajm/english-fisher-annotations

[6] Although *multiple repetition* is coded distinctly from single *word repetition* under the CHAT transcription protocol, we consider both as members of the *word repetition* category.

Disfluency	AD	HC
Word Repetition	96	29
Phrase Repetition	27	17
Word Retracing	48	35
Phrase Retracing	67	46
Total	238	127
Disfluency-Tagged	317	176

Table 5: Frequencies of disfluency types across AD and HC participants, where *Total* refers to the sum of all of our disfluency types of interest (rows 1–4), and *Disfluency-Tagged* refers to the sum of all disfluencies reported (including those not in the focus of this investigation).

Disfluency Type	Accuracy
Word Repetition	72.65%
Phrase Repetition	73.61%
Word Retracing	50.00%
Phrase Retracing	42.64%

Table 6: Percentages of disfluent words in the manually-transcribed ADReSS Challenge training set tagged with different disfluency labels (considering *multiple repetition* as a subset of word repetition) by the Fisher annotator.

§4.3.1 and §4.3.2.

4.3.1 Gold Disfluency Removal

We removed gold labelled disfluencies from the manually created reference transcripts. We did this by removing different CHAT transcription tags corresponding to repetition and retracing behaviors. Thus, the text in Figure 3 was converted to:

- **Repetition Removal:** *his sister has her hand up finger up to her mouth like she's saying.*
- **Retracing Removal:** *his sister has her has her finger up to her mouth like she's saying.*

4.3.2 Fisher Disfluency Removal

We removed disfluencies predicted by the Fisher tagger (described in §4.2) from the automatically transcribed speech. To remove words of a particular disfluency type, we matched the relevant segment of text with the predicted tag (see Figure 3) and removed the words tagged as *E* (representing *errors*, or disfluencies). For instance, to remove retracing, the blue segments of actual text and predicted tags in Figure 3 are matched, and since none of the

Actual text: *his sister <has her> [/] has her <hand up> [//] finger up to her mouth like she's saying.*

Gold tag: *his _ sister _ has E her E has _ her _ hand E up E finger _ up _ to _ her _ mouth _ like _ she _ 's _ saying _*

Predicted tag: *his _ sister _ has E her E has _ her _ hand _ up _ finger _ up _ to _ her _ mouth _ like _ she _ 's _ saying _*

Figure 3: Example utterance annotated by automatic disfluency annotator. **Actual text** represents the gold label annotated utterance from the ADReSS Challenge training set. **Gold tag** represents the expected word level annotation given the gold labels, whereas **Predicted tag** shows the predicted disfluency annotations (fluent words are followed by _ tags and disfluent words are followed by E tags) by the disfluency tagger. Repetition is highlighted in red and retracing in blue.

words are predicted as *E*, none are removed. Thus, after the removal of disfluencies according to the Fisher tagger, the text in Figure 3 was converted to:

- **Repetition Removal:** *his sister has her hand up finger up to her mouth like she's saying.*
- **Retracing Removal:** *his sister has her has her hand up finger up to her mouth like she's saying.*

4.4 Classification Setup

4.4.1 Input and Output

The ADReSS Challenge training corpus included data from N=108 participants. The input for a given data point was a sequence of words from the processed transcript, and the output was the class of the speaker: 0 for HC, or 1 for AD. Transcripts were preprocessed to remove disfluency markers, punctuation, and digits. When *multiple repetition* markers followed a word in any utterance, the word was added the specified number of times, and the marker was then removed.

4.4.2 Model

We used Bert-for-Sequence-Classification[7] to implement our model, experimenting with *bert-base-*

[7] https://github.com/huggingface/transformers

uncased as our base model and using the following hyperparameters: learning rate = 2e-5, batch size = 4, epochs = 8, max input length of 256 (a length sufficient to cover most cases). The standard default tokenizer was used. Two special tokens, [CLS] and [SEP], were added to the beginning and the end of each transcript utterance. We chose these model and parameter settings since they attained promising performance in previously published work (Yuan et al., 2020) with leave-one-out cross-validation on the ADReSS Challenge dataset.

5 Experiments

5.1 Experimental Setup

To evaluate the impact of disfluency presence and type on classifying AD status, we performed experiments considering the following conditions:

- **ALLTEXT:** The baseline condition using the original manually-created transcripts, complete with gold disfluencies, preprocessed as defined in §4.4.
- **ASR:** Transcripts are generated using ASR (explained in §3.1), and the ASR-generated transcripts are fed to the model.
- **-REP.:** Repetitions (both word- and phrase-level) are removed from ALLTEXT transcripts using either the gold or Fisher disfluency removal method.
- **-RET.:** Incidents of retracing (both word- and phrase-level) are removed from ALLTEXT using either the gold or Fisher disfluency removal method.
- **-DISF.:** Transcripts are processed so that all cases of word- or phrase-level repetition or retracing are removed. When using the Fisher disfluency removal method, this includes all disfluency-tagged words.

We report accuracy, precision, recall, and F_1 for each condition. When performing development experiments, we observed large performance differences across folds. Such brittleness has also been reported previously (Yuan et al., 2020), and may be attributed to the use of a large model (BERT) for classification on a small dataset. To address this, we perform three runs, each using different random seeds, of five-fold cross-validation and report averages and standard deviations across runs.

5.2 Results

We report our evaluation results in Table 7. As expected, we observe the highest performance in the baseline condition (ALLTEXT), which is comparable to the results in previous literature (Balagopalan et al., 2020a). The ASR condition exhibits the worst performance, with accuracy, F_1 for AD, and F_1 for HC decreasing 17.7%, 14%, and 33% respectively relative to the baseline. This underscores one of our primary motivations in conducting this work—namely, that ASR has a high error rate in real-world settings and particularly in this task environment, and moreover that its mistagging (or in some cases, purposeful removal) of disfluency has a deleterious impact on dementia detection performance. We observe from Tables 2 and 3 that DER is much higher than FER for ASR output. ASR tends to delete or replace repetitive words, increasing overall word error rate and leading to poor performance in the AD detection task. Prior work has clearly suggested that disfluencies are important indicators of cognitive health status (Lopez-de Ipina et al., 2017; Mueller et al., 2018).

Furthermore, performance clearly degrades relative to the baseline when gold disfluencies are removed (-REP.$_G$, -RET.$_G$, and -DISF.$_G$). Although retracing removal caused a slightly higher decrease in accuracy than repetition removal, there is no significant difference in performance between the -REP.$_G$ and -RET.$_G$ conditions across metrics. Accuracy and F_1 decrease 5.6% and 6% (for both AD and HC) compared to the baseline when all gold disfluencies are removed from the transcripts.

Removal of Fisher disfluencies also leads to performance degradation across all metrics. Since the Fisher disfluency annotations are more limited than the gold disfluency labels, performance in this condition (-REP.$_F$, -RET.$_F$, and -DISF.$_F$) degrades less than is observed with gold disfluency removal. Accuracy, F_1 for AD, and F_1 for HC decrease 3%, 4%, and 2% respectively compared to the baseline when all Fisher-predicted disfluencies are removed.

5.3 Distinctive Effects of Disfluency Removal

To further investigate why disfluency removal influences classification performance, we experiment with measures of syntactic complexity, context-free grammar rules, and measures of vocabulary richness[8] to identify linguistic features having mod-

[8] https://github.com/vmasrani/dementia_classifier

	Accuracy	Precision		Recall		F1	
		AD	HC	AD	HC	AD	HC
ALLTEXT	0.843±.015	0.88±.017	0.82±.020	0.80±.028	0.89±.019	0.84±.016	0.85±.013
ASR	0.670±.037	0.69±.062	0.54±.032	0.72±.060	0.52±.121	0.70±.023	0.52±.065
-REP.$_G$	0.797±.034	0.81±.044	0.79±.034	0.78±.049	0.80±.053	0.80±.021	0.79±.036
-RET.$_G$	0.787±.024	0.77±.043	0.80±.012	0.81±.015	0.76±.060	0.80±.017	0.78±.035
-DISF.$_G$	0.787±.020	0.78±.025	0.77±.028	0.76±.040	0.81±.030	0.78±.026	0.79±.020
-REP.$_F$	0.827±.015	0.86±.021	0.80±.010	0.78±.011	0.88±.021	0.82±.014	0.84±.014
-RET.$_F$	0.820±.010	0.86±.013	0.79±.021	0.78±.032	0.87±.019	0.82±.013	0.83±.006
-DISF.$_F$	0.813±.006	0.85±.018	0.78±.004	0.76±.000	0.87±.018	0.80±.008	0.83±.010

Table 7: Five-fold cross-validation results, averaged across three runs with different random seeds on the ADReSS Challenge training set. The subscript *G* refers to gold disfluency removal and *F* refers to Fisher disfluency removal.

erate to high correlation with disfluency (as measured by normalised disfluency count, repetition count, and retracing count). We find that disfluency count (considering all disfluencies in Table 4) has significant, high negative Spearman correlation ($r = -0.55$, $p < 0.001$) with type token ratio (TTR). This indicates that verbal disfluencies are highly negatively correlated with vocabulary richness, which is in turn an important feature of AD detection (Masrani, 2018). Some context-free grammar rules (INTJ, INTJ_to_UH, VP_to_VBG, VP_to_AUX) and syntactic complexity features (constituency parse tree height), also key features for AD detection (Masrani, 2018), exhibit moderate correlation with disfluency frequency. Such results show that vocabulary richness and the syntactic structure of language are vulnerable to the deletion of disfluencies, which may in turn lead to classification performance degradation.

6 Discussion

From our corpus analyses, we find that members of the AD group exhibit more verbal disfluency (Table 2), with increased rates of repetition and correction relative to the HC group. This is in line with our expectations, since disfluencies and speech errors are correlated with cognitive functions such as cognitive load, arousal, and working memory (Arciuli et al., 2010; Daneman, 1991); with increased impairment of these functions, hesitations and disfluencies increase. Previous studies have also reported that verbal disfluency frequency can be an important predictor of fine-grained cognitive status of older adults (Farzana et al., 2020). Our evaluation provides evidence that removing both gold-labelled and Fisher-annotated verbal disfluencies leads to changes in AD detection performance, opening intriguing questions for follow-up work that may further tease apart the nature of these contributions.

We speculate that some of these findings may transfer to other conditions as well. For example, studies have also reported that filled pauses are less frequently uttered by children with autism spectrum disorder than typically developed children (Gorman et al., 2016; Irvine et al., 2016). It is possible that incorporating richer disfluency information in speech-based systems for autism detection and monitoring may improve performance similarly to that seen with AD detection.

7 Conclusion

Verbal disfluencies are an important indicator of AD, and current ASR systems fail to capture and label word- and phrase-level disfluencies adequately. Doing so is necessary to generate useful transcripts with minimal human intervention, such that they can be leveraged for successful AD detection. Our future work will focus on training an end-to-end ASR system on disfluent speech so that it can generate richer disfluency annotated transcripts, which will pave the way for building end-to-end speech-based dementia detection systems.

8 Acknowledgements

This work was supported in part by a startup grant from the University of Illinois at Chicago. We thank the anonymous reviewers for their helpful comments.

References

Alzheimer's Association. 2018. 2018 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 14(3):367–429.

Joanne Arciuli, David Mallard, and Gina Villar. 2010. "Um, i can tell you're lying": Linguistic markers of deception versus truth-telling in speech. *Applied Psycholinguistics*, 31(3):397–411.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020a. To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection. In *INTERSPEECH*.

Aparna Balagopalan, Ksenia Shkaruta, and Jekaterina Novikova. 2020b. Impact of ASR on Alzheimer's disease detection: All errors are equal, but deletions are more equal than others. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 159–164, Online. Association for Computational Linguistics.

James T. Becker, François Boiler, Oscar L. Lopez, Judith Saxton, and Karen L. McGonigle. 1994. The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis. *Archives of Neurology*, 51(6):585–594.

Catherine S. Brown and Walter L. Cullinan. 1981. Word-retrieval difficulty and disfluent speech in adult anomic speakers. *Journal of Speech, Language, and Hearing Research*, 24(3):358–365.

R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock. 2000. Analysis of spontaneous, conversational speech in dementia of alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91.

Eugene Charniak and Mark Johnson. 2001. Edit detection and parsing for transcribed speech. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Christopher Cieri, David Graff, Owen Kimball, Dave Miller, and Kevin Walker. 2004. Fisher english training speech part 1 transcripts ldc2004s13.

Christopher Cieri, David Graff, Owen Kimball, Dave Miller, and Kevin Walker. 2005. Fisher english training speech part 2 transcripts ldc2005s13.

M. Daneman. 1991. Working memory as a predictor of verbal fluency. *Journal of Psycholinguistic Research*, 20:445–464.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2020. Semi-supervised topic modeling for gender bias discovery in English and Swedish. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 79–92, Barcelona, Spain (Online). Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Flavio Di Palo and Natalie Parde. 2019. 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*, pages 302–308, Florence, Italy. Association for Computational Linguistics.

Ben Eyre, Aparna Balagopalan, and Jekaterina Novikova. 2020. Fantastic features and where to find them: Detecting cognitive impairment with a subsequence classification guided approach. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 193–199, Online. Association for Computational Linguistics.

Shahla Farzana and Natalie Parde. 2020. Exploring MMSE Score Prediction Using Verbal and Non-Verbal Cues. In *Proceedings of Interspeech 2020*, pages 2207–2211.

Shahla Farzana, Mina Valizadeh, and Natalie Parde. 2020. Modeling dialogue in conversational cognitive health screening interviews. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1167–1177, Marseille, France. European Language Resources Association.

Kathleen C. Fraser, Nicklas Linz, Bai Li, Kristina Lundholm Fors, Frank Rudzicz, Alexandra König, Jan Alexandersson, Philippe Robert, and Dimitrios Kokkinakis. 2019. Multilingual prediction of Alzheimer's disease through domain adaptation and concept-based language modelling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3659–3670, Minneapolis, Minnesota. Association for Computational Linguistics.

John Godfrey and Edward Holliman. 1993. Switchboard-1 release 2 ldc97s62.

Kyle Gorman, L. Olson, A. Hill, R. Lunsford, P. Heeman, and J. van Santen. 2016. Uh and um in children with autism spectrum disorders or language impairment. *Autism Research*, 9.

Pamela Herd, Deborah Carr, and Carol Roan. 2014. Cohort profile: Wisconsin longitudinal study (wls). *International journal of epidemiology*, 43(1):34–41.

Matthew Honnibal and Mark Johnson. 2014. Joint incremental disfluency detection and dependency parsing. *Transactions of the Association for Computational Linguistics*, 2:131–142.

Julian Hough and David Schlangen. 2015. Recurrent neural networks for incremental disfluency detection. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 849–853. ISCA.

Julian Hough and David Schlangen. 2017. Joint, incremental disfluency detection and utterance segmentation from speech. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 326–336, Valencia, Spain. Association for Computational Linguistics.

Christina Irvine, Inge-Marie Eigsti, and Deborah Fein. 2016. Uh, um, and autism: Filler disfluencies as pragmatic markers in adolescents with optimal outcomes from autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 46.

Paria Jamshid Lou, Peter Anderson, and Mark Johnson. 2018. Disfluency detection using auto-correlational neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4610–4619, Brussels, Belgium. Association for Computational Linguistics.

Paria Jamshid Lou and Mark Johnson. 2017. Disfluency detection using a noisy channel model and a deep neural language model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 547–553, Vancouver, Canada. Association for Computational Linguistics.

Paria Jamshid Lou and Mark Johnson. 2020a. End-to-end speech recognition and disfluency removal. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2051–2061, Online. Association for Computational Linguistics.

Paria Jamshid Lou and Mark Johnson. 2020b. Improving disfluency detection by self-training a self-attentive model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3754–3763, Online. Association for Computational Linguistics.

Paria Jamshid Lou, Yufei Wang, and Mark Johnson. 2019. Neural constituency parsing of speech transcripts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2756–2765, Minneapolis, Minnesota. Association for Computational Linguistics.

Mark Johnson and Eugene Charniak. 2004. A TAG-based noisy-channel model of speech repairs. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 33–39, Barcelona, Spain.

Vera C Kaelin, Mina Valizadeh, Zurisadai Salgado, Natalie Parde, and Mary A Khetani. 2021. Artificial intelligence in rehabilitation targeting the participation of children and youth with disabilities: Scoping review. *J Med Internet Res*, 23(11):e25745.

Jeremy G. Kahn, Matthew Lease, Eugene Charniak, Mark Johnson, and Mari Ostendorf. 2005. Effective use of prosody in parsing conversational speech. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 233–240, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Sweta Karlekar, Tong Niu, and Mohit Bansal. 2018. 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)*, pages 701–707, New Orleans, Louisiana. Association for Computational Linguistics.

Xuan Le, Ian Lancashire, Graeme Hirst, and Regina Jokel. 2011. Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. *Literary and Linguistic Computing*, 26(4):435–461.

Matthew Lease and Mark Johnson. 2006. Early deletion of fillers in processing conversational speech. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 73–76, New York City, USA. Association for Computational Linguistics.

K. Lopez-de Ipina, U. Martinez-de Lizarduy, P. M. Calvo, B. Beitia, J. Garcia-Melero, M. Ecay-Torres, A. Estanga, and M. Faundez-Zanuy. 2017. Analysis of disfluencies for automatic detection of mild cognitive impartment: a deep learning approach. In *2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI)*, pages 1–4.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk. transcription format and programs*, volume 1. Psychology Press.

Mitchell P Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3 ldc99t42.

Vaden Masrani. 2018. *Detecting dementia from written and spoken language*. Ph.D. thesis, University of British Columbia.

Kimberly D. Mueller, Rebecca L. Koscik, Bruce P. Hermann, Sterling C. Johnson, and Lyn S. Turkstra.

2018. Declines in connected language are associated with very early mild cognitive impairment: Results from the wisconsin registry for alzheimer's prevention. *Frontiers in Aging Neuroscience*, 9.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Xian Qian and Yang Liu. 2013. Disfluency detection using multi-step stacked learning. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 820–825, Atlanta, Georgia. Association for Computational Linguistics.

Yu Qiao, Xuefeng Yin, Daniel Wiechmann, and Elma Kerz. 2021. Alzheimer's Disease Detection from Spontaneous Speech Through Combining Linguistic Complexity and (Dis)Fluency Features with Pretrained Language Models. In *Proc. Interspeech 2021*, pages 3805–3809.

Mohammad Sadegh Rasooli and Joel Tetreault. 2013. Joint parsing and disfluency detection in linear time. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 124–129, Seattle, Washington, USA. Association for Computational Linguistics.

Morteza Rohanian and Julian Hough. 2021. Best of both worlds: Making high accuracy non-incremental transformer-based disfluency detection incremental. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3693–3703, Online. Association for Computational Linguistics.

Morteza Rohanian, Julian Hough, and Matthew Purver. 2020. Multi-modal fusion with gating using audio, lexical and disfluency features for alzheimer's dementia recognition from spontaneous speech. *Interspeech 2020*.

Morteza Rohanian, Julian Hough, and Matthew Purver. 2021. Alzheimer's Dementia Recognition Using Acoustic, Lexical, Disfluency and Speech Pause Features Robust to Noisy Inputs. In *Proc. Interspeech 2021*, pages 3820–3824.

Carole Roth. 2011. Boston diagnostic aphasia examination. In Jeffrey S. Kreutzer, John DeLuca, and Bruce Caplan, editors, *Encyclopedia of Clinical Neuropsychology*, pages 428–430. Springer New York, New York, NY.

Utkarsh Sarawgi, Wazeer Zulfikar, Nouran Soliman, and Pattie Maes. 2020. Multimodal inductive transfer learning for detection of alzheimer's dementia and its severity. In *Proceedings of Interspeech 2020*, pages 2212–2216.

Elizabeth Ellen Shriberg. 1994. Preliminaries to a theory of speech disfluencies. Technical report, University of California, Berkeley.

Trang Tran, Shubham Toshniwal, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Mari Ostendorf. 2018. Parsing speech: a neural approach to integrating lexical and acoustic-prosodic information. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 69–81, New Orleans, Louisiana. Association for Computational Linguistics.

Mina Valizadeh and Natalie Parde. 2022. The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 2022 Conference of the Association for Computational Linguistics*, Dublin, Ireland. Association for Computational Linguistics.

Chunqi Wang, Ji Zhang, and Haiqing Chen. 2018. Semi-autoregressive neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 479–488, Brussels, Belgium. Association for Computational Linguistics.

Shaolei Wang, Wanxiang Che, and Ting Liu. 2016. A neural attention model for disfluency detection. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 278–287, Osaka, Japan. The COLING 2016 Organizing Committee.

Shaolei Wang, Wanxiang Che, Yue Zhang, Meishan Zhang, and Ting Liu. 2017. Transition-based disfluency detection using LSTMs. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2785–2794, Copenhagen, Denmark. Association for Computational Linguistics.

Maria Yancheva, Kathleen Fraser, and Frank Rudzicz. 2015. 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*, pages 134–139, Dresden, Germany. Association for Computational Linguistics.

Masashi Yoshikawa, Hiroyuki Shindo, and Yuji Matsumoto. 2016. Joint transition-based dependency parsing and disfluency detection for automatic speech recognition texts. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1036–1041, Austin, Texas. Association for Computational Linguistics.

Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church. 2020. Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's Disease. In *Proc. Interspeech 2020*, pages 2162–2166.

Vicky Zayats and Mari Ostendorf. 2019. Giving attention to the unexpected: Using prosody innovations in disfluency detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 86–95, Minneapolis, Minnesota. Association for Computational Linguistics.

Vicky Zayats, Mari Ostendorf, and Hannaneh Hajishirzi. 2016. Disfluency detection using a bidirectional LSTM. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association*, pages 2523–2527. ISCA.

Zining Zhu, Jekaterina Novikova, and Frank Rudzicz. 2019. Detecting cognitive impairments by agreeing on interpretations of linguistic features. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1431–1441, Minneapolis, Minnesota. Association for Computational Linguistics.