Reformulating NLP tasks to Capture Longitudinal Manifestation of Language Disorders in People with Dementia.

Anonymous ACL submission

Abstract

Dementia is associated with language disorders which impede communication. Here, we automatically learn linguistic disorder patterns by making use of a moderately-sized pre-trained language model and forcing it to focus on reformulated natural language processing (NLP) tasks and associated linguistic patterns. Our experiments show that NLP tasks that encapsulate contextual information and enhance the gradient signal with linguistic patterns benefit performance. We then use the probability estimates from the best model to construct digital linguistic markers measuring the overall quality in communication and the intensity of a variety of language disorders. We investigate how the digital markers characterize dementia speech from a longitudinal perspective. We find that our proposed communication marker is able to robustly and reliably characterize the language of people with dementia, outperforming existing linguistic approaches; and shows external validity via significant correlation with clinical markers of behaviour. Finally, our proposed linguistic disorder markers provide useful insights into gradual language impairment associated with disease progression.

1 Introduction

Dementia is a neuro-degenerative disease affecting millions worldwide and is associated with cognitive decline, including language impairment (Forbes-McKay and Venneri, 2005). Language dysfunction may be difficult to detect in the early stages of dementia (Nestor et al., 2004); however, as the disease progresses, a gradual decline of semantic knowledge ensues, and eventually, all linguistic functions can be lost (Tang-Wai and Graham, 2008; Klimova et al., 2015). Recognizing language disorders as prodromal symptoms in people with dementia may help with earlier diagnosis and improve disease management.

Dementia can cause a variety of language deficits, such as: word-finding problems, a.k.a. *anomia* (Kempler and Goral, 2008); eloquent articulation lacking the expression of meaningful information, a.k.a. *empty speech* (Nicholas et al., 1985); dropping speech, when the last few words in an utterance become barely audible a.k.a. *trailing off speech*; or *circumlocution* of words/concepts within an utterance (Silagi et al., 2015); interruptions in the smooth flow of speech, a.k.a. *disfluency* (Ferreira and Bailey, 2004), characterized by repeated words, self-interruptions, and corrections of one's own speech, a.k.a. *self-repair* (Levelt, 1983); *agrammatism*, a syntactic disturbance, characterised by telegraphic speech, misuse of pronouns, or poor grammar (Garre-Olmo, 2018).

Table 1 provides the most common language disorders and associated manifestation (linguistic patterns) observed in the speech of subjects describing the Cookie Theft Picture (CTP, Appx. A) in the DementiaBank (Becker et al., 1994) and ADReSS (Luz et al., 2020) datasets. Here we use state-of-the-art Natural Language Processing (NLP) to learn linguistic patterns indicative of language disorders in transcribed speech from people with dementia and healthy controls. We subsequently use the resulting language models to characterise the language of individuals with dementia.

Early work in NLP for dementia relied on manual engineered features based on specific lexical, acoustic and syntactic features stemming from description tasks (such as CTP), to detect linguistic signs of cognitive decline (Fraser et al., 2016; Beltrami et al., 2018; Yeung et al., 2021). Recent work uses naive neural approaches to classify and analyse linguistic and acoustic characteristics so as to either predict cognitive scores or achieve binary classification of participants (Alzheimer's Disease (AD) vs non-AD) (Karlekar et al., 2018; Balagopalan et al., 2020; Nasreen et al., 2021b; Rohanian et al., 2021). However, such approaches tend to learn language discrimination across cohorts ignoring explicit information entailed in linguistic

Disorder	Example Utterances	Symptoms/Manifestation in Language
Anomia	a) He's trying to get this and he's gonna fall off of there b) If that little girl don't xxx. c) The boy hasn't gotten down to his fall yet.	a) Empty speech, b) trailing off speech, c) circumlocution in speech
Disflueny	a) The wife is wiping a dish plate. b) His his sister's asking for one. c) Here's a sp water spigot here .	a) Word/phrase revision, b) word/phrase repetition, c) phonological fragment
Agrammatism	a) Water running down from the sink. b) Her doing the dishes. c) Three pieces of to eat on.	a) Telegraphic speech, b) Misuse of pronouns c) poor grammar

Table 1: Language disorders associated with dementia and corresponding manifestation observed in the speech of subjects in the DementiaBank and ADReSS datasets. Words in blue denote linguistic disorder patterns.

patterns within the language itself. This is because the optimization objective is to learn a unified label space and thus important linguistic patterns never have any gradient signal (Tam et al., 2021). Moreover previous work ignores the longitudinal aspect of language disorders. Here, we address these limitations and make the following contributions:

- We learn a variety of linguistic patterns characteristic of language disorders from transcribed utterances by people with dementia and healthy controls. To achieve this we force a moderately-size pre-trained LM, namely RoBERTa (Liu et al., 2019), to focus on reformulated NLP tasks (Sec. 3.3). To the best of our knowledge, ours is the first attempt to apply the recent successful NLP paradigm shift of reformulating classification as text-to-text generation (Tam et al., 2021; Wang et al., 2021; Liu et al., 2023) in the context of dementia and mental health more broadly. We show that tasks encapsulating context and forcing the model to extract signal from the language itself benefit performance (Sec. 4.1).
- We introduce human interpretable digital linguistic markers to measure the quality of communication as well as the extent of a variety of language disorders in people with dementia. To construct the digital markers we leverage the model's probability estimates (Sec. 3.1).
- We conduct a comprehensive longitudinal analysis to investigate how the linguistic communication marker characterizes individuals' speech. This shows significant discrimination across healthy controls, people with mild cognitive impairment (MCI), and people with AD (Sec. 4.2).
- We compare our proposed communication marker against existing approaches based on semantic similarity and word-level disfluency; ours shows better diagnostic performance (Sec. 4.2).
- We evaluate the reliability of the communication marker against two clinical markers of behaviour widely used for assessing dementia and show significant correlation (Sec. 4.3).
- We show that the proposed linguistic disorder markers provide useful insights into the gradual language impairment associated with disease progression (Sec. 4.4).

2 Related Work

2.1 NLP for Dementia

Early NLP work for dementia detection analysed manually aspects of language such as lexical, grammatical, and semantic features (Ahmed et al., 2013; Orimaye et al., 2017; Kavé and Dassa, 2018), paralinguistic features (Gayraud et al., 2011; López-de Ipiña et al., 2013; Pistono et al., 2019), and interactional patterns in conversations (Elsey et al., 2015).

Recent work has made use of manually engineered features (Luz et al., 2020, 2021; Nasreen et al., 2021a), disfluency features (Nasreen et al., 2021b; Rohanian et al., 2021), or acoustic embeddings (Yuan et al., 2020; Shor et al., 2020; Pan et al., 2021; Zhu et al., 2021). All such previous work has focused on differentiating across cohorts, without considering language changes over time or the importance of emergent linguistic patterns. Some very recent work does now examine longitudinal changes, but relies on speech from public figures (Petti et al., 2023).

2.2 Language Models

Language models, the prevalent technology within NLP, are usually trained with the Cloze objective where part of the context in a text is removed, and the model is tasked with predicting the missing text (Taylor, 1953). Masked language modeling (MLM) is a Cloze-based denoising objective that has been widely used in pre-training language model (Yang et al., 2022). Several works have reformulated learning tasks as cloze questions to re-purpose pre-trained language models (Schick

and Schütze, 2020a; Liu et al., 2023). Other work has exploited task descriptions (prompts) and annotated examples with demonstrations to enable few-shot learning for downstream tasks (Gao et al., 2020; Wang et al., 2021). Such approaches have become an important research field as they overcome the challenge of expensive data annotation (Li et al., 2022). However, finding ways to reformulate tasks as cloze questions that make the best use of knowledge stored in language models can be difficult (Schick and Schütze, 2020b). Here we follow the task reformulation paradigm to force a model to learn linguistic patterns of language disorders.

3 Problem Setup

3.1 Task Definition

Our task is that of learning linguistic patterns of language disorders framed as a multi-class classification problem. This involves fine-tuning a pre-trained language model $\mathcal{L}$ on a collection of $\mathcal{N}$ transcribed speech utterances $\{u_i\}_{i=1}^{N}$ from people with dementia and healthy controls elicited by the CTP description task. Each utterance is mapped to a single label $y_i \in \mathcal{Y} = \{anomia, disfluency, agrammatism, fluent\}$ [1] and the goal is to predict the corresponding label. During fine-tuning emphasis is placed on strategies for reformulating the classification task into different NLP tasks.

For evaluation purposes we construct digital markers using the probability estimates of the model, to capture the overall quality in communication and the intensity of each of the language disorders. For the communication marker, we first extract the model's output probability estimate of an utterance to be fluent, i.e., $p(y_i^{\mathcal{L}} \mid y_i = fluent)$, and then obtain averaged probabilities over the entire session (description of the CTP). To investigate the discriminating ability of the communication marker across cohorts, we calculate average and longitudinal changes in the marker. To assess its reliability, we investigate the association between changes in this marker compared to two widely used clinical behavioural markers over time (Sec. 3.4) . We similarly construct anomia, disfluency, and agrammatism markers (see Appx. D), and compare their changes across cohorts as above.

[1] The label *fluent* indicates an utterance does not exhibit any of the linguistic disorder patterns. Only $165/4037$ samples in the DementiaBank and ADReSS corpora have two labels, so we frame it as a single-label multi-class task.

3.2 Data

We conduct experiments and train models on transcribed speech from two datasets, namely ADReSS (Luz et al., 2020) and DementiaBank (Becker et al., 1994). They both contain transcribed speech of people with dementia and healthy controls describing the Cookie Theft Picture (Appx. A). ADReSS includes a single speech sample per participant while DementiaBank contains longitudinal speech, up to five times per person (see Appx. B for a detailed description of the datasets). For training models, we use data from ADReSS and also transcripts from subjects who contributed up to two descriptions in DementiaBank. Table 2 provides an overview of the datasets. Utterance annotations are based on the paralinguistic information available in transcribed scripts using the CHAT protocol (MacWhinney, 2017). For details about the coding scheme please refer to Appx. C. During pre-prossessing, we remove the paralinguistic information and discard the carers' utterances as well as patients' non-descriptive utterances. We split the data into training (80%), validation (10%) and testing (10%) keeping same class proportions across the splits.

Cohort	# Sub.	# Ses.	# Flt.	# Ano.	# Dis.	# Agr.
Healthy	107	136	908	9	246	195
Dementia	224	277	1337	203	734	405

Table 2: Statistical overview of ADReSS and DementiaBank used for training. Abbreviations: Sub.=Subjects, Ses.=Sessions, Flt.=Fluent, Ano.=Anomia, Dis.=Disflunecy, Agr.=Agrammatism.

To conduct a longitudinal evaluation we use a subset from DementiaBank of healthy controls and people with dementia who have 3, 4 and 5 sessions. The corresponding numbers for controls are 28/10/8 and for people with dementia 12/8/3.

3.3 Fine-Tuning Strategies and NLP tasks

We take a moderately sized pre-trained language model (PLM) $\mathcal{L}$ = RoBERTa (Liu et al., 2019) and fine-tune it according to different strategies.

Standard Fine-tuning ($\mathcal{L}_{standard-finetune}$): Given the PLM $\mathcal{L}$, we first convert an utterance u into a sequence of tokens $\overline{u}$ = $[CLS]\ t_1\ t_2 \ldots t_n [SEP]$ where $t_1 \ldots t_n$ are the tokens in utterance u [2]. The model takes $\overline{u}$ and maps the original utterance to a sequence of logits

[2] $\overline{u}$ is defined in the same way for all the tasks.

$\mathcal{L}(\overline{u}) \in \mathbb{R}^{|\mathcal{Y}|}$. At prediction time, softmax is applied for multiclass classification. We fine-tune the model with cross-entropy loss as follows:

$$Loss = CE(p(y^{\mathcal{L}}|\overline{u}), y) \quad (1)$$

where $p(y^{\mathcal{L}}|\overline{u})$ is softmax over y calculated as:

$$p(y^{\mathcal{L}}|\overline{u}) = \frac{exp([\![\mathcal{L}(\overline{u})]\!])_y}{\sum_{y' \in \mathcal{Y}} exp([\![\mathcal{L}(\overline{u})]\!])'_y} \quad (2)$$

Multitask Fine-tuning with MLM: We fine-tune the PLM $\mathcal{L}$ with two objectives. The first one is the masked language model (MLM) objective to understand particular linguistic patterns in the domain. We first convert an utterance u to a sequence of tokens $\overline{u}$ as above and then dynamically [3] mask 15% of tokens within the utterance (Devlin et al., 2018). For a given utterance u (e.g., A mother is wiping a dish), the model receives a MLM input as

$$[CLS]\ A\ mother\ \textbf{[MASK]}\ wiping\ a\ dish\ [SEP]$$

and maps $[MASK]$ to a sequence of logits $\mathcal{L}(\overline{u}) \in \mathbb{R}^{|\mathcal{V}|}$, where $\mathcal{V}$ is the vocabulary of $\mathcal{L}$. The training process thus becomes a high-dimensional multiclass classification problem of predicting the original token corresponding to $[MASK]$ with cross-entropy loss (Eq. 1). The second objective is to predict the class label $y_i \in \mathcal{Y}$ corresponding to an utterance u. (See 3.3). We experiment with two variants: a) separate multitask learning, where each task is learned independently ($\mathcal{L}_{multitask-MLM-separately}$). We first fine-tuning the model on the MLM objective and then resuming fine-tuning for the second objective; b) jointly learning both objectives ($\mathcal{L}_{multitask-MLM-joint}$). The combined loss is a linear weighted sum of loss functions of the two objectives. The assignment of weights is an open research question. Here, we set the weights empirically, based on the minimum loss function values when fine-tuning the model on the two objectives separately (See Appx. D).

Entailment-based Fine-tuning ($\mathcal{L}_{entailment}$): The goal here is to map the relationship between an utterance u and the corresponding language disorder label to a relationship space by reformulating multi-class classification as an entailment-task (Wang et al., 2021), a.k.a. natural language inference (NLI). Here, a language disorder definition is assumed to entail utterance u if the definition can be logically derived from utterance u, (e.g., for the utterance *"His his sister's asking for one"* entails *"Word repetition or revision"*).

Given an instance (u, y), we construct a set of tuples $\{(u, p_j)\}_{j=1}^{|\mathcal{Y}|}$ for each class $y \in \mathcal{Y}$ where $\{p_j\}$ is a set of label definitions, including [4] {*Talking around words/empty speech/incomplete speech, Word repetition or revision, Agrammatism or paragrammatism in speech, Fluent speech*}. For each utterance, the model $\mathcal{L}$ receives a set of $|\mathcal{Y}|$ tuples [5] in the form:

$$[CLS]\ u\ [SEP]\ p_j\ [SEP],$$

and outputs a sequence of logits $\mathcal{L}(u, p_j) \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{E}|}$, where $\mathcal{E} = \{entails,\ does\ not\ entail\}$. At inference time, we extract the probability of $p(entails|(u, p_j))$ for each class in $\mathcal{Y}$ and apply $argmax$ across the extracted probabilities. We fine-tune the model with cross-entropy loss.

Prompt-based Learning: Here the PLM $\mathcal{L}$ is tasked with "auto-completing" natural language prompts (Liu et al., 2023). In particular, for each utterance u let $\mathcal{T}(u)$ be a MLM input with one $[MASK]$ token. Let $\mathcal{M} : \mathcal{Y} \rightarrow \mathcal{V}^{|\mathcal{Y}|}$ be a one-to-one mapping from the task label space $\mathcal{Y}$ to individual words in the vocabulary $\mathcal{V}$ of $\mathcal{L}$. The model $\mathcal{L}$ receives a template $\mathcal{T}(u)$ and maps the $[MASK]$ token to a sequence of logits $\mathcal{L}(\mathcal{T}(u)) \in \mathbb{R}^{|\mathcal{V}|}$. We cast the problem of predicting the probability of $y \in \mathcal{Y}$ as a MLM task:

$$p(y \mid u) = p([MASK] = \mathcal{M}(y) \mid \mathcal{T}(u)). \quad (3)$$

For a set of instances $\{u, y\}$, $\mathcal{L}$ is fine-tuned to minimize the cross-entropy loss.

We experiment with the following variants:

- **Standard Prompt-based** ($\mathcal{L}_{standart-prompt}$): Here the MLM consists of an utterance u and a task-specific prompt as follows:

$$\mathcal{T}(u) = [CLS]\ u\ .\ \underline{It\ is\ \textbf{[MASK]}\ .}\ [SEP] \quad (4)$$

  , where the underlined text is the task specific template and $[MASK] \in \mathcal{M}(y)$.
- **Prompt-based with Demonstration Examples** ($\mathcal{L}_{prompt-demonstrations}$): We adopt the idea of incorporating demonstrations as additional context (Gao et al., 2020). For each utterance u,

[3] Different tokens are randomly masked in each epoch.

[4] The label definitions were created on the basesis of the CHAT protocol guidelines and manual analysis of the data

[5] This approach requires $|\mathcal{Y}|$ forward passes during inference time.

we randomly sample one example $(u, \mathcal{M}(y_i))_{i=1}^{|\mathcal{Y}|}$ from each class $y \in \mathcal{Y}$ and combine the original utterance and examples to create templates according to Eq. 4. For the random samples, we replace the $[MASK]$ token with $\mathcal{M}(y_i)$. The model $\mathcal{L}$ receives as input a combination of the templates:

$$\mathcal{T}(u) \oplus \mathcal{T}(u, \mathcal{M}(y_1)) \oplus ... \oplus \mathcal{T}(u, \mathcal{M}(y_i)) \quad (5)$$

where $\oplus$ denotes concatenation. Given a contextual utterance in the form of Eq 5, the task involves predicting the $[MASK]$ token in the original utterance. At test time we sample demonstration examples from the training subset.

- **Prompt-based with Inverse Learning Objective ($\mathcal{L}_{prompt-inverse}$):** The standard prompt-based objective encapsulates the question *"Given the input what is the right label"*. Here, we inverse the question, *"Given the answer label, what is the correct content"*. The model $\mathcal{L}$ is trained on the objective of predicting the input given the label. Formally, an utterance u is reformulated through $\mathcal{T}$ according to Eq. 4. Then, we replace the $[MASK]$ token in Eq. 4 with the original class token $\mathcal{M}(y)$ and apply a 50% random masking across the utterance's tokens. Thus, we force the model to predict the tokens in the context of the original label $\mathcal{M}(y)$. The model outputs for each of the $[MASK]$ tokens a sequence of logits $\mathcal{L}(u) \in \mathbb{R}^{|\mathcal{V}|}$, where $\mathcal{V}$ is the vocabulary of $\mathcal{L}$. Similarly to the MLM objective, we apply cross-entropy loss to predict the masked tokens. At test time, we give the model the correct and incorrect labels $\mathcal{M}(y)$ and reform the utterance u through $\mathcal{T}$. Out of $|\mathcal{Y}|$ combinations, we choose the one with minimum loss.

Random Rate: Finally we include weighted guessing as a baseline classifier where accuracy is guessed at the weighted percentages of classes.

For the experimental settings when training RoBERTa across different NLP tasks, we refer readers to D.

3.4 Evaluation Metrics

To evaluate the success of different NLP task reformulation strategies in capturing the different language disorders, we report per class accuracy and $F1$. We also calculate the macro-averaged accuracy and $F1$ score. We chose macro-averaged scores since we are interested in minority classes, such as anomia, important in charecterizing the communication ability of people with dementia.

We evaluate the digital linguistic markers defined in Sec. 3.1 against two widely used clinical behavioural markers, namely, the Mini-Mental State Examination (MMSE), and the Clinical Dementia Rating (CDR) scale (Morris, 1997). The higher the MMSE score, the higher the cognitive function. In contrast, the higher the CDR, the lower the cognitive function. For a detailed description of the behavioural markers see Appx. E.

4 Experimental Results

4.1 Quantitative Results

Table 3 summarizes the experimental results for NLP task reformulation for identifying language disorder patterns in transcribed speech from the DementiaBank and ADReSS datasets. All fine-tuning and learning strategies yielded significantly better performance than random weighted guessing. However, class imbalance has caused bias towards the majority class (i.e., fluent speech), leading to underperformance for the minority class (i.e., anomia). We also noticed a trade-off in performance between the majority and minority classes. We suppose this is because speech with anomia is still fluent and prosodically correct but overall meaningless.

Both multitask with MLM and inverse prompt-based learning tasks were trained with the objective of forcing the model to obtain signal from linguistic patterns associated with a unified label space. Joint multitask learning with MLM is robust with respect to the minority class. In particular, it achieves the best accuracy and f_1 scores for the anomia class compared to all other settings. On the other hand, prompt-based with inverse learning objective underperforms all other approaches. We assume this is because the latter does not have a gradient signal from the labels during optimization. This setting may be more appropriate when masking is targeted rather than random. However, this would require word-level annotations which are not currently available in these datasets.

Tasks incorporating context in the form of additional information exhibit superior performance over tasks learning a unified space without context. In particular, entailment-based fine-tuning which includes label descriptions achieves an increased macro accuracy of 68.3% compared to 65.1% for standard fine-tuning. Similarly, prompt-based learning with demonstrations incorporat-

	Fluent		Anomia		Disfluency		Agrammatism		Macro	
	Acc.	**F_1**	**Acc.**	**F_1**	**Acc.**	**F_1**	**Acc.**	**F_1**	**Acc.**	**F_1**
Random Rate	30.8	-	0.2	-	5.7	-	2.1	-	1.2 (↓ 63.9)	-
$\mathcal{L}_{standard-finetune}$	**96.8**	94.2	20.8	31.2	86.5	78.0	56.5	67.2	65.1	67.5
$\mathcal{L}_{multitask-MLM-separately}$	94.8	91.7	29.2	37.8	85.6	78.8	50.7	61.9	65.1 (↔0.0)	67.6 (↑ 0.1)
$\mathcal{L}_{multitask-MLM-joint}$	93.7	92.0	**45.8**	**48.9**	74.8	71.6	55.1	62.3	67.3 (↑ 2.2)	68.7 (↑ 1.2)
$\mathcal{L}_{entailment}$	94.7	94.7	30.2	41.0	**88.9**	76.3	59.0	66.0	68.3 (↑ 3.2)	70.3 (↑ 2.8)
$\mathcal{L}_{standard-prompt}$	96.4	93.1	29.2	41.2	86.5	79.3	55.1	66.7	66.8 (↑ 1.7)	70.1 (↑ 2.6)
$\mathcal{L}_{prompt-demonstrations}$	96.6	**95.2**	27.0	37.4	87.5	**81.0**	**66.2**	**71.9**	**69.9** (↑ 4.8)	**72.2** (↑ 4.7)
$\mathcal{L}_{prompt-inverse}$	48.0	54.6	33.3	13.6	18.9	24.4	46.4	35.8	36.7 (↓ 28.4)	25.7 (↓ 41.8)

Table 3: Performance of models resulting from reformulated NLP tasks using RoBERTa for identifying language disorder patterns in transcribed speech from the DementiaBank and ADReSS datasets. Numbers in bold indicate best performance. Numbers in parentheses denote deviation from the performance of standard fine-tuning.

ing examples from each class yields an increased macro accuracy of 69.9% compared to 66.8% for standard prompt-based learning.

Overall, the experiments show that tasks which include context in the form of additional information and force the model to obtain signal from linguistic patterns yield better performance. In particular, prompt-based learning with demonstrations, which meets both of the above characteristics, achieves an increased macro accuracy of 69.9%, compared to 65.1% for standard fine-tuning trained with an objective that ignores patterns from the language itself during the optimization process.

4.2 Longitudinal Discrimination Ability

Using the probability estimates of RoBERTa trained on prompt-based learning with demonstration examples to recognise linguistic disorders (which yielded the highest macro-F1), we have created a digital communication marker and language disorder markers (See Sec 3.1 for more details). We analyze changes in the digital communication marker over time and across cohorts of people with AD, MCI and healthy controls. We calculate the average of the communication marker across the three cohorts (Table 4). The higher the score of the marker (1st column), the lower the impact of language disorders on communication. We observe that the marker decreases alongside disease severity. In particular, there is a significant difference in the marker's scores across the healthy, MCI, and AD cohorts.[6]

We subsequently calculate changes in the communication marker from the end to the start of the study and across cohorts (i.e., $\Delta_{(end-onset)}$ in Table 4). There is a significant decrease for the AD group compared to the healthy and MCI cohorts ($p < 0.05$) [6]. There was no significant change in linguistic ability for the MCI and healthy cohorts: for controls, there is presumably no cognitive decline; for the MCI group, changes in linguistic function are likely trivial (Nestor et al., 2004).

We also calculate changes in the communication marker between adjacent sessions over time and then aggregated them per individual. In Table 4, we report the average change across cohorts, i.e., $\Delta_{(long)}$. We obtain similar results as the ones from the end to the start of the study.

We compare the discrimination ability of our communication marker against two baseline markers based on semantic similarity and word-level disfluency. For a baseline developed on semantic similarity, we use the Incoherence Model (Iter et al., 2018), which scores adjacent pairs of utterances based on the cosine similarities of their sentence embeddings (Reimers and Gurevych, 2019). The higher the score, the better the thematic consistency within a session (CTP description). We note that the thematic consistency is higher for the MCI cohort compared to the healthy controls. However, there is no substantial difference across cohorts (see Table 4, Semantic similarity marker). We observe similar results when analysing the semantic marker's longitudinal discrimination ability. For word-level disfluency, we use a pre-trained transformer model for word-by-word disfluency detection in the form of reparandum-interregnum-repair (Rohanian and Hough, 2021). To construct the baseline marker, we use the normalized probability estimates of words within an utterance to be fluent and then average the scores obtained over a session (CTP description). The higher the

[6] We use the nonparametric Mann-Whitney test to measure if the distribution of a variable is different in two groups.

	Our communication marker			Semantic similarity marker			Word-level disfluency marker		
Cohort	**Marker**	$\Delta_{(end-start)}$	$\Delta_{(long)}$	**Marker**	$\Delta_{(end-start)}$	$\Delta_{(long)}$	**Marker**	$\Delta_{(end-start)}$	$\Delta_{(long)}$
Healthy	**0.759 (0.164)**	+0.011 (0.162)	+0.000 (0.106)	0.296 (0.077)	+0.013 (0.107)	+0.009 (0.054)	0.913 (0.064)	-0.005 (0.072)	-0.003 (0.030)
MCI	**0.630 (0.224)**	+0.010 (0.164)	+0.010 (0.068)	0.299 (0.080)	-0.051 (0.077)	-0.017 (0.031)	0.879 (0.081)	+0.019 (0.100)	+0.005 (0.030)
AD	**0.536 (0.201)**	**-0.229 (0.117)**	**-0.120 (0.094)**	0.270 (0.067)	+0.011 (0.890)	+0.001 (0.038)	0.892 (0.075)	-0.026 (0.081)	-0.008 (0.038)

Table 4: Comparison of our proposed digital linguistic communication marker versus baselines from semantic similarity and word-level Marker: Average of marker within a population. $\Delta_{(end-start)}$: Average change of the marker from the end to the beginning of the study. $\Delta_{(long)}$: Average change of the digital marker between adjacent individuals' sessions. Positive number implies improvement over time. Numbers in () refer to corresponding standard deviations. Numbers in bold denote significant difference across cohorts.

score, the less the occurrence of disfluent patterns in speech. We obtain results similar to the ones from the semantic similarity marker. In particular, the score is higher for people with AD compared to those with MCI. However, there is no significant difference across cohorts.

Overall, our proposed communication marker is robust and reliable in discriminating between people with dementia, MCI and healthy controls, identifying changes in linguistic ability over time and does so better than existing approaches.

4.3 Communication marker Reliability

We investigate the reliability of the digital communication marker by associating longitudinal changes in the marker with two widely used behavioural measures collected over the study. We consider individuals across different cohorts with at least three sessions each (for the description of the evaluation dataset, see Sec. 3.2).

We first investigate the association between longitudinal changes in the digital communication marker and the Mini-Mental State Examination (MMSE). We calculate the average of MMSE scores per individual [7] and the average difference in the communication marker between the same individual's adjacent sessions. Positive values of change indicate improvement in communication over time while negative values denote the opposite. Similarly, high MMSE scores are indicative of better cognitive function (refer to Appx. E for details on MMSE). Figure 1 illustrates the correlation between averaged longitudinal changes in the communication marker and average MMSE scores. We notice that people with a high MMSE score either improve or exhibit minor changes in communication over time. On the other hand, the communication marker decreases for those people with low MMSE scores. Overall, we found a Pearson correlation of 0.61 ($p = 4.48e^{-8}$) between changes in MMSE and the average difference in the communication marker over time.

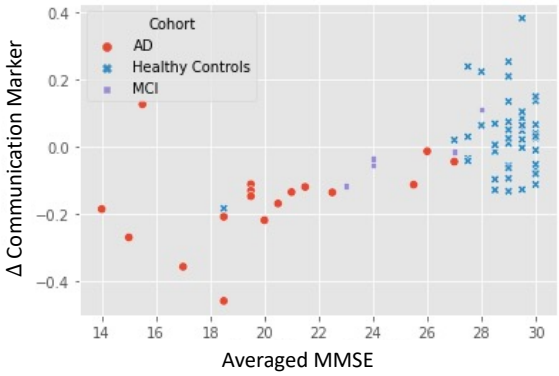

Figure 1: Association between average longitudinal change in communication marker and average Mini-Mental State Examination (MMSE) scores, across cohorts.

Similarly, we investigate the association between average longitudinal changes in the communication marker with the Clinical Dementia Rating (CDR). Here, the higher the CDR, the lower the cognitive function (see Appx. E for details on CDR). Figure 2 illustrates the association between average longitudinal changes in the communication marker and CDR. We note that people with low average values of CDR (i.e., CDR $\in [0, 1)$) improved their communication over time. This is presumably because subjects are able to remember and do better at the CTP description task when seeing it again (Goldberg et al., 2015). However, people with moderate to high levels of CDR (i.e., CDR $\in [1, 3]$) exhibit impairment in communication over time. Overall, we found a Pearson correlation of 0.56 ($p = 6.67e^{-7}$) between average CDR values and average values in changes for the communication over time.

We observe that people with AD with severe cognitive impairment, i.e. MMSE ranging from 14-18 and CDR from 2-2.5, did not exhibit a severe

[7] We don't calculate longitudinal changes in the behavioural measures due to missing values in the datasets.

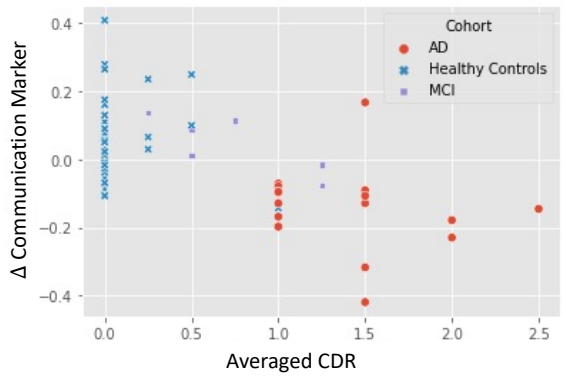

Figure 2: Association between average longitudinal changes in communication marker and average values of the Clinical Dementia Rating (CDR), across cohorts.

decrease in the communication marker over time. We attribute this to a ceiling effect. Indeed, a meta-analysis shows that the communication marker for people with the lowest behaviour scores was already much lower at the onset compared to those AD participants with higher behaviour scores.

4.4 Linguistic Disorder Markers

We investigate how different linguistic disorder markers capture the impact on individuals' speech. We compute the markers as the percentage of occurrence of each of the language disorders in Table 1 using the normalized probability estimates of the model (for details on how the markers are obtained, see Sec. 3.1). Table 5 provides the average percentage value of each linguistic disorder marker per cohort as well as corresponding percent changes from the end to the start of the study. The higher the percentage of a marker the more prevalent the language disorder.

	Anomia		Disfluency		Agrammatism	
Cohort	**Marker**	**Δ**	**Marker**	**Δ**	**Marker**	**Δ**
Healthy	1.11	+1.12	15.43	+1.91	11.41	-3.85
MCI	1.94	+1.35	21.66	-4.75	13.31	-0.44
AD	**5.58**	+2.18	**25.11**	+8.82	15.86	**+8.95**

Table 5: Percentage of language disorders as captured by the corresponding linguistic markers across cohorts. Marker: Average of marker within a cohort. $\Delta_{(end-start)}$: Average change of the marker from the end to the beginning of the study. Negative numbers imply improvement over time. Numbers in bold denote significant difference across cohorts.

We note that people across all cohorts exhibit disfluency. However, the disfluency marker was significantly higher for people with AD compared to healthy controls ($p < 0.05$).[6] The MCI cohort exhibits improvement in disfluency over the study (Δ=-4.75% in Table 5). Anomia is characteristic of people with AD (Botha and Josephs, 2019) and despite being less prevalent overall is significantly [6] higher for the AD cohort. Although agrammatism is more prominent in people with AD, there is no significant difference across cohorts. We attribute this to the same relative ratio of aggramatism in healthy controls and people with dementia in the training data (see Table 2 where Sub:Aggr≈.55 in both cases) rather than the sensitivity of the marker itself. Indeed the aggramatism marker captures that people with AD exhibit a significant change in syntactic disturbance over time (+8.95% in the value of the marker) whereas the rest of the cohorts improved over time.

Overall, the linguistic disorder markers were effective in screening and monitoring AD where gradual language impairment ensues.

5 Conclusion

We are the first to introduce reformulated NLP tasks for learning language disorder patterns from transcribed speech in dementia datasets by forcing a pre-trained language model to obtain signal from the language itself. Our experiments show that NLP tasks encapsulating contextual information and enhancing the gradient signal with linguistic patterns benefit performance. We use the probability estimates of the model with highest macro-F1 to construct digital markers measuring communication ability and the occurrence of various language disorders in the speech of people with dementia and healthy controls. Longitudinal analysis shows that the digital communication marker is able to assess the quality of communication and distinguish between people with MCI, Alzheimer's Disease (AD) and healthy controls. A comparison against existing linguistic approaches for capturing language impairment shows the superiority of our proposed communication marker. Moreover, the latter correlates significantly with two widely used clinical behaviour markers. Finally, our proposed linguistic disorder markers prove effective for screening and monitoring AD and provide useful insights into longitudinal change in linguistic ability. In the future we will explore large pre-trained generative transformers and automatic generation of templates to improve performance on capturing linguistic disorder patterns.

Limitations

Monitoring dementia using computational linguistics approaches is an important topic. Previous work has primarily focused on learning language discrimination across healthy controls and people with AD, ignoring longitudinal language disorders. In this work, we use DementiaBank to capture longitudinal linguistic disorder patterns that characterize people living with dementia. Currently, DementiaBank is the largest available longitudinal dementia dataset. A limitation of DementiaBank is that the longitudinal aspect is limited, spanning up to 5 sessions/descriptions maximum per individual, with most participants contributing up to two narratives. Moreover, the number of participants is relatively small, especially for the mild-cognitive impairment (MCI) cohort. Finally, descriptions are elicited through the Cookie Theft Picture (CTP), ignoring interactive aspects of everyday conversational interaction. The Carolinas Conversation Collection dataset (Pope and Davis, 2011) contains more natural conversations between patients and clinical practitioners. However, it only contains speech data from people with AD and no equivalent data for healthy controls. In the future, we aim to address these limitations by investigating the generalisability of our proposed digital language disorder markers on a novel fine-grained longitudinal multi-modal dataset from people with dementia over several months in a natural setting (currently under review).

In this study, we used manually transcribed data from DementiaBank and its paralinguistic information to annotate transcribed turns. In a real-world scenario, participants mostly provide speech via a speech elicitation task. This implies that the introduced method requires an automatic speech recognition (ASR) system robust to various sources of noise to be operationalized. ASR for mental health is currently underexplored, with most transcription work being done by humans.

It may be that the proposed digital linguistic markers become a less accurate means for monitoring dementia when people experience other comorbidities, neurodegenerative and mental illnesses, that significantly affect speech and language. Indeed, cognitive-linguistic function is a strong biomarker for neuropsychological health (Voleti et al., 2019).

Finally, there is a great deal of variability to be expected in speech and language data affecting the sensitivity of the proposed digital linguistic markers. Both speech and language are impacted by speaker identity, context, background noise, spoken language etc. Moreover, people may vary in their use of language due to various social contexts and conditions, a.k.a., style-shifting (Coupland, 2007). Both inter and intra-speaker variability in language could affect the sensitivity of the proposed digital markers. While it is possible to tackle intra-speaker language variability, e.g., by integrating speaker-dependent information to the language, the inter-speaker variability remains an open-challenging research question.

Ethics Statement

Our work does not involve ethical considerations around the analysis of the DementiaBank and ADReSS corpora as they are widely used. For DementiaBank, ethics was obtained by the original research team by James Backer and participating individuals consented to share their data following a larger protocol administered by the Alzheimer and Related Dementias Study at the University of Pittsburgh School of Medicine (Becker et al., 1994). Access to the data is password protected and restricted to those signing an agreement. For ADReSS, ethics was obtained by the original research team by Brian MacWhinney that collected the data for ADReSS challenge. Access to the data requires membership of DementiaBank and a non-disclosure agreement between the stakeholders and the research team.

This work uses transcribed dementia data to identify changes in cognitive status considering individuals' language disorders. Research Potential risks from the application of our work in being able to identify cognitive decline in individuals are akin to those who misuse personal information for their own profit without considering the impact and the social consequences in the broader community. Potential mitigation strategies include running the software on authorised servers, with encrypted data during transfer, and anonymization of data prior to analysis. Another possibility would be to perform on-device processing (e.g. on individuals' computers or other devices) for identifying changes in cognition and the results of the analysis would only be shared with authorised individuals. Individuals would be consented before any of our software would be run on their data.

References

Samrah Ahmed, Anne-Marie F Haigh, Celeste A de Jager, and Peter Garrard. 2013. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain*, 136(12):3727–3737.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020. To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer's disease detection. *arXiv preprint arXiv:2008.01551*.

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.

Daniela Beltrami, Gloria Gagliardi, Rema Rossini Favretti, Enrico Ghidoni, Fabio Tamburini, and Laura Calzà. 2018. Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline? *Frontiers in aging neuroscience*, 10:369.

Hugo Botha and Keith A Josephs. 2019. Primary progressive aphasias and apraxia of speech. *Continuum: Lifelong Learning in Neurology*, 25(1):101.

Nikolas Coupland. 2007. *Style: Language variation and identity*. Cambridge University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Christopher Elsey, Paul Drew, Danielle Jones, Daniel Blackburn, Sarah Wakefield, Kirsty Harkness, Annalena Venneri, and Markus Reuber. 2015. Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics. *Patient Education and Counseling*, 98(9):1071–1077.

Fernanda Ferreira and Karl GD Bailey. 2004. Disfluencies and human language comprehension. *Trends in cognitive sciences*, 8(5):231–237.

Katrina E Forbes-McKay and Annalena Venneri. 2005. Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. *Neurological sciences*, 26:243–254.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2016. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Josep Garre-Olmo. 2018. Epidemiology of Alzheimer's disease and other dementias. *Revista de neurologia*, 66(11):377–386.

Frederique Gayraud, Hye-Ran Lee, and Melissa Barkat-Defradas. 2011. Syntactic and lexical context of pauses and hesitations in the discourse of Alzheimer patients and healthy elderly subjects. *Clinical linguistics & phonetics*, 25(3):198–209.

Terry E Goldberg, Philip D Harvey, Keith A Wesnes, Peter J Snyder, and Lon S Schneider. 2015. Practice effects due to serial cognitive assessment: implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1):103–111.

H. Goodglass, E. Kaplan, S. Weintraub, and B. Barresi. 2001. The Boston diagnostic aphasia examination. *Philadelphia, PA: Lippincott, Williams & Wilkins*.

Dan Iter, Jong Yoon, and Dan Jurafsky. 2018. Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 136–146.

Sweta Karlekar, Tong Niu, and Mohit Bansal. 2018. Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. *arXiv preprint arXiv:1804.06440*.

Gitit Kavé and Ayelet Dassa. 2018. Severity of Alzheimer's disease and language features in picture descriptions. *Aphasiology*, 32(1):27–40.

Daniel Kempler and Mira Goral. 2008. Language and dementia: Neuropsychological aspects. *Annual review of applied linguistics*, 28:73–90.

Blanka Klimova, Petra Maresova, Martin Valis, Jakub Hort, and Kamil Kuca. 2015. Alzheimer's disease and language impairments: social intervention and medical treatment. *Clinical interventions in aging*, pages 1401–1408.

W. J. Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.

Zicheng Li, Shoushan Li, and Guodong Zhou. 2022. Pre-trained token-replaced detection model as few-shot learner. *arXiv preprint arXiv:2203.03235*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Karmele López-de Ipiña, Jesus-Bernardino Alonso, Carlos Manuel Travieso, Jordi Solé-Casals, Harkaitz Egiraun, Marcos Faundez-Zanuy, Aitzol Ezeiza, Nora Barroso, Miriam Ecay-Torres, Pablo Martinez-Lage,

et al. 2013. On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis. *Sensors*, 13(5):6730–6745.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. *arXiv preprint arXiv:2004.06833*.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021. Detecting cognitive decline using speech only: The adresso challenge. *arXiv preprint arXiv:2104.09356*.

Brian MacWhinney. 2017. Tools for analyzing talk part 1: The CHAT transcription format. *Carnegie.[Google Scholar]*, 16.

John C Morris. 1997. Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the Alzheimer type. *International psychogeriatrics*, 9(S1):173–176.

Shamila Nasreen, Julian Hough, Matthew Purver, et al. 2021a. Detecting Alzheimer's Disease using Interactional and Acoustic features from Spontaneous Speech. Interspeech.

Shamila Nasreen, Morteza Rohanian, Julian Hough, and Matthew Purver. 2021b. Alzheimer's dementia recognition from spontaneous speech using disfluency and interactional features. *Frontiers in Computer Science*, page 49.

Peter J Nestor, Philip Scheltens, and John R Hodges. 2004. Advances in the early detection of Alzheimer's disease. *Nature medicine*, 10(Suppl 7):S34–S41.

Marjorie Nicholas, Loraine K Obler, Martin L Albert, and Nancy Helm-Estabrooks. 1985. Empty speech in Alzheimer's disease and fluent aphasia. *Journal of Speech, Language, and Hearing Research*, 28(3):405–410.

Sylvester O Orimaye, Jojo SM Wong, Karen J Golden, Chee P Wong, and Ireneous N Soyiri. 2017. Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC bioinformatics*, 18(1):1–13.

Yilin Pan, Bahman Mirheidari, Jennifer M Harris, Jennifer C Thompson, Matthew Jones, Julie S Snowden, Daniel Blackburn, and Heidi Christensen. 2021. Using the outputs of different automatic speech recognition paradigms for acoustic-and bert-based Alzheimer's dementia detection through spontaneous speech. In *Interspeech*, pages 3810–3814.

Ullaa Petti, Simona Baker, Anna Korhonen, and Jessica Robin. 2023. The generalizability of longitudinal changes in speech before Alzheimer's disease diagnosis. *Journal of Alzheimer's Disease*, 92(2):547–564.

Aurélie Pistono, Jeremie Pariente, C Bézy, B Lemesle, J Le Men, and Mélanie Jucla. 2019. What happens when nothing happens? an investigation of pauses as a compensatory mechanism in early Alzheimer's disease. *Neuropsychologia*, 124:133–143.

Charlene Pope and Boyd H Davis. 2011. Finding a balance: The carolinas conversation collection.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 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.

Morteza Rohanian, Julian Hough, and Matthew Purver. 2021. Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs. *arXiv preprint arXiv:2106.15684*.

Timo Schick and Hinrich Schütze. 2020a. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Timo Schick and Hinrich Schütze. 2020b. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*.

Joel Shor, Aren Jansen, Ronnie Maor, Oran Lang, Omry Tuval, Felix de Chaumont Quitry, Marco Tagliasacchi, Ira Shavitt, Dotan Emanuel, and Yinnon Haviv. 2020. Towards learning a universal non-semantic representation of speech. *arXiv preprint arXiv:2002.12764*.

Marcela Lima Silagi, Paulo Henrique Ferreira Bertolucci, and Karin Zazo Ortiz. 2015. Naming ability in patients with mild to moderate Alzheimer's disease: what changes occur with the evolution of the disease? *Clinics*, 70:423–428.

Derek Tam, Rakesh R Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. 2021. Improving and simplifying pattern exploiting training. *arXiv preprint arXiv:2103.11955*.

David F Tang-Wai and Naida L Graham. 2008. Assessment of language function in dementia-Alzheimer. *Geriatrics*, 11(2):103–110.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Rohit Voleti, Julie M Liss, and Visar Berisha. 2019. A review of automated speech and language features for assessment of cognitive and thought disorders. *IEEE journal of selected topics in signal processing*, 14(2):282–298.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*.

Dongjie Yang, Zhuosheng Zhang, and Hai Zhao. 2022. Learning better masking for better language model pre-training. *arXiv preprint arXiv:2208.10806*.

Anthony Yeung, Andrea Iaboni, Elizabeth Rochon, Monica Lavoie, Calvin Santiago, Maria Yancheva, Jekaterina Novikova, Mengdan Xu, Jessica Robin, Liam D Kaufman, et al. 2021. Correlating natural language processing and automated speech analysis with clinician assessment to quantify speech-language changes in mild cognitive impairment and Alzheimer's dementia. *Alzheimer's research & therapy*, 13(1):109.

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 *INTERSPEECH*, volume 2020, pages 2162–6.

Youxiang Zhu, Abdelrahman Obyat, Xiaohui Liang, John A Batsis, and Robert M Roth. 2021. Wavbert: Exploiting semantic and non-semantic speech using wav2vec and bert for dementia detection. In *Interspeech*, pages 3790–3794.

A The Cookie Theft Picture

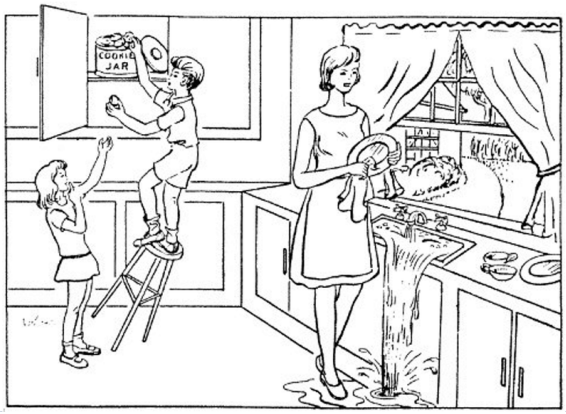

Figure 3: The Cookie Theft Picture from the Boston Diagnostic Aphasia Examination (Goodglass et al., 2001).

For the PD task, the examiner asks subjects to describe the picture (see Fig. 3) by saying, "Tell me everything you see going on in this picture". Then subjects might say, "there is a mother who is drying dishes next to the sink in the kitchen. She is not paying attention and has left the tap on. As a result, water overflows from the sink. Meanwhile, two children attempt to make cookies from a jar when their mother is not looking. One of the children, a boy, has climbed onto a stool to get up to the cupboard where the cookie jar is stored. The stool is rocking precariously. The other child, a girl, is standing next to the stool and has her hand outstretched ready to be given cookies.

B Dementia datasets

B.1 DentiaBank

The dataset was gathered longitudinally between 1983 and 1988 as part of the Alzheimer Research Program at the University of Pittsburgh. The study initially enrolled 319 participants according to the following eligibility criteria: all the participants were required to be above 44 years old, have at least seven years of education, have no history of major nervous system disorders, and have an initial Mini-Mental State Examination score above 10. Finally, the cohort consisted of 282 subjects. In particular, the cohort included 101 healthy control subjects (HC) and 181 Alzheimer's disease subjects (AD). An extensive neuropsychological assessment was conducted on the participants, including verbal tasks and the Mini-Mental State Examination (MMSE).

B.2 ADReSS

ADReSS is a benchmark dataset of spontaneous speech, which is acoustically pre-processed and balanced in terms of age and gender. The dataset entails transcribed speech of 78 non-AD subjects and 78 AD subjects of 35 males and 43 females for each of the cohorts. The dataset was made available for the ADReSS challenge consisted of two tasks: a) an AD classification task, where the task required one to produce a model to predict the label (AD or non-AD) for a speech session and b) an MMSE score regression task, where the task required one to create a model to infer the subject's Mini-Mental Status Examination (MMSE) score based on speech and/or language data.

C Coding Scheme for the Annotation of Transcribed Utterances.

Table 6 lists the codes we used to annotate transcribed speech utterances in accordance with the CHAT protocol (MacWhinney, 2017). Moreover, we used the code $[+exc]$ to filter out non-descriptive utterances from the Cookie Thief Pic-

ture (CTP) description task (e.g., "Yeah that's it."). As shown in Table 6, the manifestation granularity varies across different language disorders. For example, anomia is exhibited through various symptoms in language.

Disorder	Code	Manifestation in Language
Agrammatism	$[+gram]$	Agrammatic and paragrammatic speech.
Disfluency	$[/]$	Word or phrase repetition.
	$[//]$	Word or phrase revision.
	$\&+$	Phonological fragment.
Anomia	$+es$	Empty speech.
	$+...$	Termination of an incomplete utterance.
	$[+cir]$	Talking around words/concepts.
	$[+jar]$	Fluent and prosodically correct but largely meaningless speech.
Disruptive	$[+exc]$	Non-descriptive speech.

Table 6: Coding scheme used for the annotation of transcribed speech utterances following the CHAT protocol (MacWhinney, 2017).

D Experimental Settings

We used a grid search optimization technique to optimize the parameters. For consistency, we used the same experimental settings for all models. We first fine-tuned all models by performing a twenty-times grid search over their parameter pool. We empirically experimented with learning rate (lr): $lr \in \{0.00001, 0.00002, 0.00005, 0.0001, 0.0002\}$, batch size ($bs$): $bs \in \{16, 32, 64, 128\}$ and optimization (O): $O \in \{AdamW, Adam\}$. After the fine-tuning process, we trained again all the models for 50 epochs with 4 epochs early stopping, three times. We reported the average performance on the test set for all experiments. Model checkpoints were selected based on the minimum validation loss. Experiments were conducted on two GPUs, Nvidia V-100.

For fine-tuning RoBERTa with MLM jointly, we suggest the weights (1/0.5139) for the classification objective and (1/2.4149) for the MLM objective.

To investigate how various language disorders involve with the progression of dementia, we construct anomia, disfluency, and agrammatism markers, by first extracting the corresponding model's probability estimates for each utterance, i.e., $p(y_i^{\mathcal{L}} \mid y_i = y_i^*)$, where $y_i^* \in \{anomia, disfluency, agrammatism\}$. We then obtain averaged probabilities over the entire session (description of the CTP).

E Clinical Behavioural Markers

E.1 Mini-Mental State Examination (MMSE)

The Mini-Mental State Examination (MMSE) has been the most common method for diagnosing AD and other neurodegenerative diseases affecting the brain. It was devised in 1975 by Folstein et al. as a simple standardized test for evaluating the cognitive performance of subjects, and where appropriate to qualify and quantify their deficit. It is now the standard bearer for the neuropsychological evaluation of dementia, mild cognitive impairment, and AD.

The MMSE was designed to give a practical clinical assessment of change in cognitive status in geriatric patients. It covers the person's orientation to time and place, recall ability, short-term memory, and arithmetic ability. It may be used as a screening test for cognitive loss or as a brief bedside cognitive assessment. By definition, it cannot be used to diagnose dementia, yet this has turned into its main purpose.

The MMSE includes 11 items, divided into 2 sections. The first requires verbal responses to orientation, memory, and attention questions. The second section requires reading and writing and covers ability to name, follow verbal and written commands, write a sentence, and copy a polygon. All questions are asked in a specific order and can be scored immediately by summing the points assigned to each successfully completed task; the maximum score is 30. A score of 25 or higher is classed as normal. If the score is below 24, the result is usually considered to be abnormal, indicating possible cognitive impairment. The MMSE has been found to be sensitive to the severity of dementia in patients with Alzheimer's disease (AD). The total score is useful in documenting cognitive change over time.

E.2 Clinical Dementia Rating (CDR)

The Clinical Dementia Rating (CDR) is a global rating device that was first introduced in a prospective study of patients with mild "senile dementia of AD type" (SDAT) in 1982 (Hughes et al., 1982). New and revised CDR scoring rules were later introduced (Berg, 1988; Morris, 1993; Morris et al., 1997). CDR is estimated on the basis of a semistructured interview of the subject and the caregiver (informant) and on the clinical judgment of the clinician. CDR is calculated on the basis of testing six different cognitive and behavioral domains

such as memory, orientation, judgment and problem solving, community affairs, home and hobbies performance, and personal care. The CDR is based on a scale of 0–3: no dementia (CDR = 0), questionable dementia (CDR = 0.5), MCI (CDR = 1), moderate cognitive impairment (CDR = 2), and severe cognitive impairment (CDR = 3). Two sets of questions are asked, one for the informant and another for the subject. The set for the informant includes questions about the subject's memory problem, judgment and problem solving ability of the subject, community affairs of the subject, home life and hobbies of the subject, and personal questions related to the subject. The set for subject includes memory-related questions, orientation-related questions, and questions about judgment and problem-solving ability.