

MultiConAD: A Unified Multilingual Conversational Dataset for Early Alzheimer’s Detection

Arezo Shakeri
University of Stavanger
Stavanger, Norway
arezo.shakeri@uis.no

Mina Farmanbar
University of Stavanger
Stavanger, Norway
mina.farmanbar@uis.no

Krisztian Balog
University of Stavanger
Stavanger, Norway
krisztian.balog@uis.no

Abstract

Dementia is a progressive cognitive syndrome with Alzheimer’s disease (AD) as the leading cause. Conversation-based AD detection offers a cost-effective alternative to clinical methods, as language dysfunction is an early biomarker of AD. However, most prior research has framed AD detection as a binary classification problem, limiting the ability to identify Mild Cognitive Impairment (MCI)—a crucial stage for early intervention. Also, studies primarily rely on single-language datasets, mainly in English, restricting cross-language generalizability. To address this gap, we make three key contributions. First, we introduce a novel, multilingual dataset for AD detection by unifying 16 publicly available dementia-related conversational datasets. This corpus spans English, Spanish, Chinese, and Greek, and incorporates both audio and text data derived from a variety of cognitive assessment tasks. Second, we perform finer-grained classification, including MCI, and evaluate various classifiers using sparse and dense text representations. Third, we conduct experiments in monolingual and multilingual settings, finding that some languages benefit from multilingual training while others perform better independently. This study highlights the challenges in multilingual AD detection and enables future research on both language-specific approaches and techniques aimed at improving model generalization and robustness.

CCS Concepts

• **Information systems** → *Multilingual and cross-lingual retrieval; Environment-specific retrieval*; **Test collections; Clustering and classification**; • **Applied computing** → *Health informatics*.

Keywords

Alzheimer’s, Early Alzheimer’s detection, Dementia detection, Cognitive impairment, Conversational dataset, Multilingual dataset

ACM Reference Format:

Arezo Shakeri, Mina Farmanbar, and Krisztian Balog. 2025. MultiConAD: A Unified Multilingual Conversational Dataset for Early Alzheimer’s Detection. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’25)*, July 13–18,

2025, Padua, Italy. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3726302.3730313>

1 Introduction

Dementia is a progressive clinical cognitive decline that impairs daily activities and independence [14]. Cases are expected to rise from 57 million in 2019 to 152 million by 2050 [16]. According to the World Health Organization, in 2019, it was the seventh leading cause of death globally and the second in high-income countries [22]. Alzheimer’s disease (AD) accounts for 50%–75% of cases with prevalence roughly doubling every five years after the age 65 [28]. The lack of effective medical treatments underscores the need for early dementia and prevention [16].

Traditional clinical methods for AD detection, including MRI, PET imaging, and cerebrospinal fluid analysis, are costly, time-consuming, and impractical for large-scale early screening. This has led to efforts to find more accessible and cost-effective alternatives [56]. Recent studies have increasingly identified language dysfunction as an early indicator of cognitive decline, making speech and language features as valuable biomarkers for the early detection of dementia [24, 43]. This growing body of research suggests that analyzing conversational patterns can provide critical insights into the onset of AD and other forms of dementia. However, a major limitation of existing studies is their reliance on single-language datasets, predominantly in English [1, 3, 42], with frequently utilized datasets including DementiaBank’s Pitt corpus, ADReSS, and ADReSSo. This focus on a single language restricts the cross-linguistic generalizability of findings, making it difficult to apply these methods in multilingual or culturally diverse settings. Furthermore, the lack of multilingual datasets poses challenges for developing diagnostic tools that can be used globally [11]. Expanding this scope, the recent Interspeech TAUADIAL Challenge aims to investigate speech as a marker of cognitive function in a global health context [5]. This initiative provides data in two major languages, Chinese and English, to enhance understanding of the relationship between speech and cognition across diverse linguistic settings. The Signal Processing Grand Challenge (SPGC) is another effort aimed at exploring the transferability of speech features across languages (English and Greek) for AD prediction [35]. Despite these efforts, cross-linguistic speech-based detection of AD remains an open gap in the field. Additionally, most prior research has framed AD detection as a binary classification problem (AD vs. Healthy Controls (HC)), limiting the ability to identify Mild Cognitive Impairment (MCI)—a crucial intermediate stage for early intervention [49]. Detecting MCI is particularly important because individuals at this stage have a higher likelihood of progressing to AD, yet timely interventions can potentially slow or prevent further decline [46]. However, the lack of

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR ’25, Padua, Italy

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1592-1/2025/07
<https://doi.org/10.1145/3726302.3730313>

datasets with labeled MCI cases has restricted the development of models that can distinguish between these three cognitive states.

To address this significant gap, we present MultiConAD, a unified, multilingual conversational dataset for Alzheimer’s detection. MultiConAD consolidates 16 publicly available dementia-related conversational datasets across four languages: English, Spanish, Chinese, and Greek. These languages represent some of the most widely spoken languages in the world, with Mandarin Chinese and Spanish being the most spoken, followed by English [49]. The corpus incorporates both audio and text modalities; most datasets include both, while a few contain only audio or only text. Crucially, these datasets feature a diverse range of cognitive assessment tasks, such as picture descriptions, story recall tasks, and verbal and semantic fluency tests. As our second contribution, we perform a more nuanced classification approach at a finer level of granularity by considering MCI as an intermediate category between AD and HC. Third, we conduct extensive experiments on multiple language settings using different text representations and classification algorithms. The overarching objective of this research is to investigate the impact of classification approach (binary vs. multiclass), language, and dataset composition (monolingual, multilingual, and translated) on the performance of Alzheimer’s detection models using conversational data.

Our results reveal interesting language-specific trends: while some languages benefit from multilingual training, suggesting shared markers of cognitive decline, other languages perform better when trained separately, indicating unique language-dependent patterns. These findings highlight the importance of tailoring AD detection models to specific language while also demonstrating the potential advantages of leveraging multilingual training for improved generalization. Beyond these immediate findings, MultiConAD facilitates future research in several key areas, including optimizing language-specific models through deeper linguistic analysis and leveraging cross-lingual transfer learning.

In summary, this paper presents the following contributions: (1) we introduce and release a unified multilingual conversational dataset for early Alzheimer’s detection; (2) we experimentally evaluate and compare various approaches for leveraging monolingual, multilingual, and translated datasets for both binary and multiclass classification tasks; and (3) we offer a thorough analysis of the results, emphasizing the key challenges associated with this problem. All resources developed in this study, including instructions on how to obtain the dataset and model implementations, are publicly accessible at <https://github.com/ArezoShakeri/MultiConAD>.

2 Related Work

To provide the necessary context for the unified Alzheimer’s detection dataset introduced in this paper, we discuss cognitive tasks used for Alzheimer’s detection, datasets, and computational approaches.

2.1 Cognitive Task Analysis in Alzheimer’s Detection

Cognitive Task Analysis is an essential method used in AD studies to evaluate various aspects of communication, memory, executive function, and overall cognitive ability as the disease progresses [15]. Boston Diagnostic Aphasia Examination (BDAE) is a

well-known method for assessing linguistic abilities and language impairments through tasks such as word fluency tests, sentence construction, and narrative storytelling, helping clinicians and researchers understand how language deficits manifest in different stages of Alzheimer’s [12]. One widely used component of the BDAE is the **Cookie Theft Picture (CTP)** [7], an image depicting a busy kitchen scene where participants are asked to describe the situation in detail. This task assesses a person’s ability to produce coherent speech and reveals cognitive and language dysfunctions, such as difficulties in organizing thoughts or recalling key details from memory. Another prominent task is **Story Recall**, where participants are instructed to memorize the short story to be told, and are asked to recall the story when it ends [8]. Additionally, **Verbal Fluency Tests** are commonly used to evaluate an individual’s lexical retrieval and executive function [30]. These tasks require participants to generate as many words as possible within a time limit, either from a specific letter (e.g., F, A, S) or a semantic category (e.g., animals, fruits). The **Semantic Fluency Task** is also widely employed in clinical settings to identify challenges in speech production, executive functioning, and semantic memory performance [30]. In this task, participants are asked to produce as many words as possible in a given semantic category (e.g., animals) and time frame.

2.2 Datasets

The dataset developed in this study is multilingual, comprising four languages: Chinese, Spanish, English, and Greek, sourced from multiple countries and regions. The combined multilingual dataset includes participants from North America (United States), Europe (Spain, Greece), and Asia (China). This selection of languages ensures that the dataset captures a broad spectrum of speech patterns and cognitive variations seen in different linguistic groups.

To build this diverse dataset, data were sourced from 16 publicly available resources that provide conversational speech related to AD. This corpus incorporates both audio and text formats. All datasets utilized in this study, with the exception of two Chinese datasets, are publicly accessible through DementiaBank, representing the most prominent publicly available datasets for Alzheimer’s detection. Access to DementiaBank is restricted to members of the DementiaBank consortium and is password-protected. Established researchers and clinicians specializing in dementia may apply for membership to gain access. The two Chinese datasets are available in the following GitHub repositories: <https://github.com/lzy1012/Alzheimer-s-disease-datasets> and https://github.com/lzl32947/NCMMSC2021_AD_Compensation.

The participant groups included in these datasets consist of individuals with dementia (DM), AD, MCI, and HC. Among these, only the two English datasets, Lu and VAS, contain data for dementia. It is important to note that dementia encompasses various types of diseases, including AD, Parkinson’s disease, vascular dementia, and others. Table 2 presents small parts from picture description conversations in the Pitt and Taukdial datasets. The Taukdial transcripts were originally in audio format and were converted to text using automatic transcription.

Table 1 provides an overview of all datasets, followed by a detailed description of each dataset below.

Table 1: Overview of datasets used in this study. Tasks: (PD) Picture Description, (FT) Fluency Task, (SR) Story Retelling, (FC) Free Conversation, (NA) Narrative. Labels: (DM) Dementia, (AD) Alzheimer’s Disease, (MCI) Mild Cognitive Impairment, (HC) Healthy Control.

Language	Source	Modality		Task					Labels				
		Text	Audio	PD	FT	SR	FC	NA	DM	AD	MCI	HC	
English	Pitt	✓	✓	✓	✗	✗	✗	✗	✗	✗	255	42	243
	Lu	✓	✓	✓	✗	✗	✗	✗	✗	6	16	2	27
	VAS	✓	✓	✗	✗	✗	✓	✗	✗	30	✗	35	36
	Baycrest	✓	✓	✗	✗	✓	✗	✗	✗	✗	3	7	✗
	Kempler	✓	✓	✓	✗	✗	✗	✗	✗	✗	7	✗	✗
	WLS	✓	✓	✓	✓	✗	✗	✗	✗	✗	263	✗	1106
	Delware	✓	✓	✓	✗	✗	✗	✓	✗	✗	✗	61	34
	Taukadiadial	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	95	74
Spanish	Ivanova	✓	✓	✗	✗	✗	✗	✓	✗	74	90	197	
	PerLA	✓	✗	✓	✓	✗	✓	✗	✗	21	✗	✗	
Chinese	NCMMSE	✗	✓	✓	✓	✗	✓	✗	✗	79	93	108	
	iFkyTek	✓	✗	✓	✗	✗	✗	✗	✗	68	144	111	
Greek	DS3	✗	✓	✓	✓	✗	✗	✗	✗	76	✗	19	
	DS5	✗	✓	✓	✓	✗	✗	✗	✗	26	35	31	
	DS7	✗	✓	✓	✓	✗	✗	✗	✗	27	35	2	
	ADReSS-M	✗	✓	✓	✗	✗	✗	✗	✗	22	✗	24	

Table 2: Excerpts from transcripts of picture descriptions from the Pitt and Taukadiadial datasets. “INT” and “PAR” denote interviewer and participant utterances in the Pitt corpus; Taukadiadial has only participant-side utterances.

Dataset	Transcript
Pitt	INT: just tell me everything that you see happening in that picture. PAR: alright. there’s &-um a young boy that’s getting a cookie jar. PAR: and it he’s &-uh in bad shape because &-uh the thing is PAR: and in the picture the mother is washin(g) dishes and doesn’t see [...]
Taukadiadial	What I see is a young lady was riding her tricycle with her cat, and somehow the cat got up in the tree. Her dog is trying to chase the cat, can’t get up the tree. Her father climbed the tree and is sitting on the limb [...]

- **Pitt corpus** [4] is part of DementiaBank, a multimedia database designed to facilitate the study of individuals with dementia, supported by grants NIA AG03705 and AG05133. This dataset encompasses various types of conversational data, including tasks such as the Cookie Theft picture (CTP) description [13], fluency tasks, story recall, and sentence construction. For this study, data were exclusively drawn from Probable AD, PossibleAD, MCI, and HC participants for the CTP task.
- **Lu** [29] is an English-language dataset from DementiaBank, comprising conversations from 26 HC participants and 28 individuals with dementia, specifically focusing on CTP conversations from the United States.
- **VAS** [31] consists of voice commands collected from 40 older adults, aged 65 or older, who were either HC participants or MCI participants. These participants were grouped based on their Montreal Cognitive Assessment (MoCA) [19] scores. The data

was collected through Amazon Alexa, a Voice-Assistant System, and includes daily spontaneous voice commands issued by users to seek assistance with everyday tasks.

- **Baycrest** [23] includes retellings of the Cinderella story and other discourse tasks. For the Cinderella task, participants were provided with a storybook featuring Disney illustrations of the tale. The text was obscured, and pages deemed less relevant were taped together to prevent visibility.
- **Kempler** [26] originally was collected from individuals diagnosed with probable AD at the UCLA Geriatric Outpatient Clinic and the West Los Angeles Veterans Administration Hospital. This dataset contains conversations from individuals with MCI, focusing on topics such as family, profession, personal history, and questions from a neuropsychological interview [26].
- **WLS** (The Wisconsin Longitudinal Study) [18] follows a random sample of Wisconsin high school graduates from 1957 (N = 10,317), born between 1938 and 1940. Initially focused on educational and occupational aspirations, later surveys in 1964, 1975, 1992, 2004, and 2011 increasingly focused on health and life course experiences as participants aged [18].
- **Delaware** [29], part of DementiaBank, investigates cognitive-linguistic features in neurotypical adults and individuals with MCI due to possible AD. It employs the DementiaBank discourse protocol, which includes picture description, story narrative, procedural discourse, and personal narrative tasks.
- **Taukadiadial** [34] consists of spontaneous speech samples from cognitively normal subjects and individuals with MCI, recorded while describing pictures. The dataset includes English and Chinese audio recordings.
- **Ivanova** [20] comprises recordings from participants over 60 years old, all native European Spanish speakers with at least six

Table 3: Summary of related work in Alzheimer’s detection.

Reference	Dataset	Categories	Accuracy	Method
[6]	ADReSS-M	AD, HC	69.6%	Acoustic, linguistic, and para-linguistic features and SVM
[9]	Taukadal	MCI, HC	77.5%	Linguistic, acoustic, and data augmentation
[10]	Pitt + Swedish dataset	MCI, HC	63% (Pitt) 72% (Swedish)	Multilingual word embeddings
[21]	Pitt	AD, HC	90.6%	Domain-specific FastText word embeddings and 1D-CNN + BLSTM
[27]	VAS	DM, MCI, HC	74.7%	Linguistic and acoustic features and ML classifiers
[33]	iFlyTek	AD, HC	83.7%	Transformer
[40]	Ivanova	AD, MCI, HC	73.03%	Data augmentation on audio files and CNN
[47]	A custom conversational AD dataset	AD, HC	95.8%	Linguistic, acoustic and ANN
[52]	ADReSS-M	AD, HC	82.6%	Acoustic features and ANN
[54]	A corpus from DementiaBank	AD, HC	92.2%	PoS features and a transformer model
[57]	NCMMSE	AD, MCI, HC	89.1%	Linguistic and acoustic features and SVM

years of primary education, ensuring literacy and minimizing cognitive load during a reading task.

- **PerLA** [51] is part of the Clinical Linguistics PERLACH Corpus. Collected between 2012 and 2014, it includes 27 transcriptions from conversations with 21 AD patients, some interviewed twice.
- **Lu** [37], available through DementiaBank, consists of interview recordings from 52 Mandarin-speaking AD patients. These recordings include tasks such as the CTP description, category fluency, and picture naming exercises [44].
- **NCMMSE** [41], from the National Conference on Man-Machine Speech Communication (NCMMSC), consists of audio recordings where participants engage in various tasks, such as picture descriptions and fluency exercises. It features 79 subjects with AD, 93 with MCI, and 108 HC.
- **IFlytek** [32], derived from the Predictive Challenge of AD in 2019, comprises transcripts of spontaneous speech collected during CTP description tasks, a common diagnostic tool for neurological disorders [32]. Participants, ranging in age from 41 to 98, included 68 individuals with AD, 144 with MCI, and 111 healthy controls.
- **Dem@Care** [25] datasets consist of data collected through lab and home-based experiments conducted at the Greek Alzheimer’s Association for Dementia and Related Disorders in Thessaloniki, Greece, and participants’ homes. These datasets include video and audio recordings alongside data from physiological sensors. Additionally, it incorporates data from sleep, motion, and plug sensors, offering a comprehensive view of participants’ behavioral and physiological patterns. We used the Ds3, Ds5, and Ds7 datasets from Dem@Care project.
- **ADReSS-M** [35] comprises spontaneous speech samples from cognitively normal individuals and AD patients. The training set includes English audio recordings of participants describing the CTP from the Boston Diagnostic Aphasia Examination. The test set features speech samples in Greek, where participants describe a different picture. For this study, we utilized only the test set including 24 HC, 22 AD, as the dataset creator advises against using the training set alongside other English datasets from DementiaBank due to potential overlap.

2.3 Computational Approaches to Alzheimer’s Detection

Numerous studies have explored cognitive task-based approaches for Alzheimer’s detection, using speech and language analysis as key indicators of cognitive decline. These methods consider linguistic and paralinguistic features, such as lexical richness, syntactic complexity, fluency, and acoustic properties, to differentiate between AD, MCI, and HC across various languages and datasets [49]. We review related work focusing on single-language and multilingual conversation-based Alzheimer’s detection for both binary and multiclass classification, which are summarized in Table 3.

For binary AD detection, Duan et al. [9] analyze the multilingual Taukadal dataset, extracting language-agnostic speech features to classify MCI and HC. Fraser et al. [10] show that multilingual approaches outperform monolingual ones, achieving up to 72% accuracy in MCI detection using narrative speech in English and Swedish. Jain et al. [21] demonstrate that domain-specific word embeddings improve AD vs. HC classification on the Pitt corpus.

Regarding multiclass AD detection, Orozco-Chavez et al. [40] address the underrepresentation of Spanish speakers, using a CNN model on the Ivanova dataset from DementiaBank and achieving 73.03% accuracy. Due to data limitations, they augmented the training set with synthetic samples. Kurtz et al. [27] utilize the VAS dataset, achieving 74.7% accuracy in differentiating DM, MCI, and HC, highlighting the potential of voice assistant systems for passive monitoring. Ying et al. [57] integrate acoustic and linguistic features from Wav2Vec2.0 and BERT for early AD detection in Chinese (NCMMSC dataset), achieving 89.1% accuracy.

Liu et al. [33] propose a transformer-based model with a feature purification network, finding that semantic features generalize well across English and French. Chen et al. [6] show that paralinguistic features outperform pre-trained ones in cross-lingual AD detection, achieving 69.6% accuracy when training on English and testing on Greek. Tamm et al. [52] integrate acoustic features with demographic covariates, achieving 82.6% accuracy for AD classification. Wen et al. [54] identify 12 PoS features distinguishing AD from HC, reaching 92.2% accuracy. Finally, Sadeghian et al. [47] propose a non-invasive, speech-based diagnostic tool for AD detection in clinical and home settings.

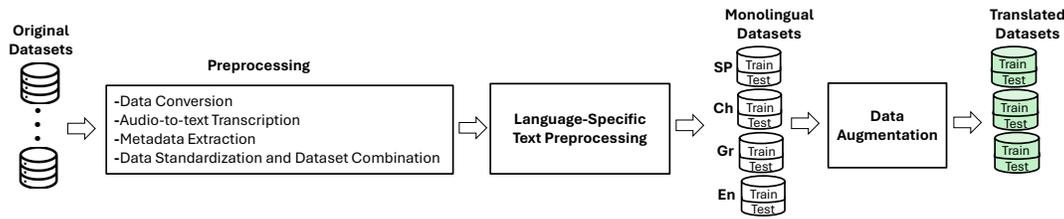


Figure 1: Flowchart illustrating the process of dataset generation.

Table 4: Distribution of diagnostic categories (AD, MCI, HC) across languages with corresponding train and test splits.

Language	Train Split			Test Split		
	AD	MCI	HC	AD	MCI	HC
Spanish	78	70	156	20	18	39
Chinese	93	200	187	23	51	47
Greek	266	51	87	66	13	22
English	512	239	1450	63	60	87
Total	949	560	1880	172	142	195

3 Dataset Creation

MultiConAD is constructed from 16 individual datasets presented in Table 1, which vary significantly in format, modality, and associated metadata. To create a unified multilingual dataset, we performed a series of preprocessing and normalization steps. This involved converting data to a standard format, transcribing audio data, extracting relevant metadata, standardizing data structures, and performing language-specific text cleaning. Figure 1 illustrates the process. The overall goal was to create a consistent and well-structured dataset suitable for training and evaluating multilingual AD detection models. A detailed breakdown of the number of conversation transcripts in each dataset, along with the train-test split statistics, is provided in Table 4.

3.1 Preprocessing

3.1.1 Data Conversion. The datasets exist in various formats, reflecting differences in modalities and data sources. All text-based datasets from DementiaBank¹ are provided in the CHAT format [36]. The Chinese iFlyTek text-based dataset is available in TSV format. All audio files are provided in WAV format.

3.1.2 Audio-to-text Transcription. Several datasets consisted solely of audio recordings. We transcribed them using a state-of-the-art Transformer-based multilingual speech recognition system. Specifically, we employed the Whisper-Large multilingual model (Large-V3) [45] to transcribe recordings in Chinese, Greek, and Spanish. For speaker diarization, we implemented whisper-large-v3 along with NVIDIA’s NeMo, as NeMo offers a robust framework for distinguishing speakers within a conversation. However, upon manual inspection, we found that the diarization results were often inaccurate, likely due to background noise or suboptimal audio

quality. Consequently, we omitted the speaker diarization step and focused solely on transcribing the audio files.

3.1.3 Metadata Extraction. We extracted all relevant information, including patient IDs, cognitive scores, demographic details, and conversational data—and standardized them into a unified format. Mini-Mental State Examination (MMSE) [55] and MoCA [39] are two commonly reported cognitive scores in the datasets used for this study. MMSE is a widely used cognitive screening tool designed to assess cognitive impairment, particularly in dementia. A score below 24 typically indicates the need for further evaluation, though factors such as age, education, and cultural background can influence interpretation [55]. MoCA is a cognitive screening tool designed to detect MCI by evaluating eight cognitive domains through rapid and sensitive tasks. A score below 26 indicates cognitive impairment, with the MoCA demonstrating higher sensitivity than the MMSE in detecting early cognitive decline [39]. This information was typically provided in supplementary files, most commonly in Excel format. The dataset metadata, including details on the original data modality, underlying task, dataset name, language, test date, test duration, and testing environment, is also incorporated into the final normalized dataset. Comprehensive documentation of this process can be found in the GitHub repository accompanying this work.

3.1.4 Data Standardization and Dataset Combination. To facilitate integration into a single multilingual dataset, all extracted data and transcriptions were stored as JSONL files. Certain dataset-specific considerations were made during standardization: for the Pitt Corpus, only transcripts from the CTP task were included, and for the Chinese NCMMSE dataset, only the training set was used due to test set label unavailability. Datasets within the same language were combined, resulting in four standardized datasets in English, Spanish, Chinese, and Greek. To ensure fair comparisons, an 80%-20% train-test split was applied across all datasets, with the exception of the WLS dataset, which was used exclusively for training.

WLS dataset. Given its extensive number of transcripts and its demonstrated utility in AD detection, as highlighted by Guo et al. [17], we incorporated the WLS dataset to enhance model performance even if it does not include explicit dementia diagnoses but provides cognitive test scores and health-related responses.

Participants in the WLS dataset completed two verbal fluency tasks, naming words in the categories of animals and food within one minute. Verbal fluency tests are highly sensitive to AD detection, with a suggested screening cutoff of 15 words in the animal

¹<https://dementia.talkbank.org/>

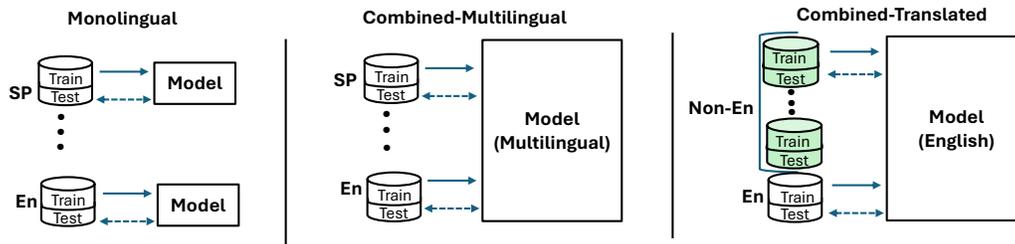


Figure 2: Flowchart depicting the experimental scenarios. Monolingual; Combined-Multilingual; Combined-Translated

category [17]. To establish a comparable control subgroup within WLS, we applied age- and education-adjusted verbal fluency thresholds: 16 words for participants under 60, 14 words for those aged 60–79, and 12 words for individuals over 79. Prior studies indicate that animal fluency scores below 15 are strongly indicative of AD, with a sensitivity of 0.88 and specificity of 0.96 [17]. Since no normative data exist for the food category, we adopted the same thresholds as the animal category due to their similar score distributions, as suggested by Guo et al. [17].

3.2 Language-Specific Text Preprocessing

We developed four distinct text preprocessing pipelines, one for each language, to process the unified datasets generated during data collection. A major part of the preprocessing effort involved removing unnecessary symbols and marks in the transcripts, such as *, &=laugh, and &=nodes. Additionally, we created visualizations of transcript lengths and their frequency distributions to identify and remove potential outliers. The unified dataset comprises both the raw and preprocessed transcripts.

3.3 Data Augmentation

To broaden the usability of the dataset and facilitate addressing the research objectives outlined in the introduction, we generate English translations of all Spanish, Greek, and Chinese texts. These translations are then leveraged in a combined English-translated version of the dataset (see Section 3.4).

We perform the translation with help of GPT-4 model accessed via the OpenAI API. Specifically, we used the prompt:

Translate the following {source_lang} text to {target_lang}:
{text}

Translation:

To ensure accuracy and consistency, we set the temperature parameter to zero, minimizing variability in responses.

3.4 Dataset Variants

We define three dataset variants to enable a range of evaluation scenarios, investigating the impact of multilingual data and translation on model performance. Figure 2 illustrates the corresponding experimental scenarios using these dataset variants.

- **Monolingual:** Models are trained and evaluated solely on the train and test sets of the *same* language. This represents a baseline scenario where only data from a single language is used.
- **Combined-Multilingual:** Models are trained on a combination of the training sets from *all* languages. This allows the model to

potentially learn cross-lingual patterns. Evaluation is performed on a single language’s test set at each iteration.

- **Combined-Translated:** We utilize the machine-translated versions of all non-English datasets. The model is then trained on a combined dataset that includes both the original English and translated-to-English training data, while evaluation is performed on the translated test sets. This setting tests the effectiveness of using translation to create a unified English-only dataset.

4 Experiments

This section details the experiments conducted on the unified MultiConAD dataset. We explored two classification tasks: binary and multiclass classification. In the binary classification setting, the MCI group was excluded, concentrating on the distinction between HC and AD. The multiclass classification, which addresses a research gap identified in our recent literature review [49], involved classifying instances into HC, MCI, and AD categories. To investigate multilingual patterns related to AD pathology, all classification models were trained on three different variants of the dataset—monolingual, combined-multilingual, and combined-translated—as described in Section 3.4.

4.1 Research Questions

We address the following research questions.

- **RQ1:** How does the performance of Alzheimer’s detection models differ when framed as a binary classification problem (AD vs. HC) compared to a multiclass problem (AD vs. MCI vs. HC)?
- **RQ2:** For multiclass classification, do the confusion matrices of the models reveal specific challenges in distinguishing between different stages of cognitive decline (AD vs. MCI)?
- **RQ3:** Does training language-specific models on a combined dataset in multiple languages (combined-multilingual) improve overall performance compared to training only on a single language (monolingual)?
- **RQ4:** How does the performance of models trained on translated and combined data (combined-translated) compare to models trained on the original language (monolingual) and on the combined dataset without translation (combined-multilingual)?

4.2 Baseline methods

We consider two text representations, sparse and dense, and a diverse set of established classification algorithms. We chose a TF-IDF

Table 5: Binary classification (AD vs. HC) results comparing Sparse and Dense text representations and classifiers (DT, RF, SVM, LR) across dataset variations. Values represent classification accuracy. Best results are in boldface. The arrows denote a performance improvement \uparrow or degradation \downarrow relative to the Monolingual setting.

Language	Text Repr.	Monolingual				Combined-Multilingual				Combined-Translated			
		DT	RF	SVM	LR	DT	RF	SVM	LR	DT	RF	SVM	LR
Spanish	Sparse	0.73	0.73	0.78	0.78	0.80 \uparrow	0.76 \uparrow	0.66 \downarrow	0.80 \uparrow	0.75 \uparrow	0.80 \uparrow	0.75 \downarrow	0.73 \downarrow
	Dense	0.71	0.71	0.80	0.80	0.66 \downarrow	0.78 \uparrow	0.66 \downarrow	0.80	0.59 \downarrow	0.76 \uparrow	0.78 \downarrow	0.76 \downarrow
Chinese	Sparse	0.67	0.69	0.70	0.70	0.67	0.69	0.67 \downarrow	0.69 \downarrow	0.70 \uparrow	0.90 \uparrow	0.89 \uparrow	0.84 \uparrow
	Dense	0.69	0.76	0.83	0.80	0.76 \uparrow	0.86 \uparrow	0.67 \downarrow	0.81 \uparrow	0.66 \downarrow	0.84 \uparrow	0.80 \downarrow	0.84 \uparrow
Greek	Sparse	0.68	0.78	0.77	0.78	0.68	0.76 \downarrow	0.60 \downarrow	0.73 \downarrow	0.58 \downarrow	0.67 \downarrow	0.69 \downarrow	0.51 \downarrow
	Dense	0.65	0.75	0.78	0.77	0.64 \downarrow	0.75	0.75 \downarrow	0.73 \downarrow	0.62 \downarrow	0.66 \downarrow	0.70 \downarrow	0.64 \downarrow
English	Sparse	0.73	0.78	0.77	0.75	0.67 \downarrow	0.74 \downarrow	0.58	0.75	0.73	0.74 \downarrow	0.76	0.61 \downarrow
	Dense	0.65	0.75	0.81	0.79	0.67 \uparrow	0.77 \uparrow	0.58 \downarrow	0.67 \downarrow	0.71 \uparrow	0.73 \downarrow	0.83 \uparrow	0.70 \downarrow

Table 6: Multiclass classification (AD vs. MCI vs. HC) results comparing Sparse and Dense text representations and classifiers (DT, RF, SVM, LR) across dataset variations. Values represent classification accuracy. Best results are in boldface. The arrows denote a performance improvement \uparrow or degradation \downarrow relative to the Monolingual setting.

Language	Text Repr.	Monolingual				Combined-Multilingual				Combined-Translated			
		DT	RF	SVM	LR	DT	RF	SVM	LR	DT	RF	SVM	LR
Spanish	Sparse	0.61	0.60	0.61	0.61	0.51 \downarrow	0.62 \uparrow	0.51 \downarrow	0.58 \downarrow	0.47 \downarrow	0.58 \downarrow	0.56 \downarrow	0.56 \downarrow
	Dense	0.52	0.61	0.61	0.61	0.47 \downarrow	0.61	0.61	0.57 \downarrow	0.60 \uparrow	0.58 \downarrow	0.56 \downarrow	0.51 \downarrow
Chinese	Sparse	0.36	0.35	0.40	0.39	0.42 \uparrow	0.39 \uparrow	0.39 \downarrow	0.40 \uparrow	0.45 \uparrow	0.59 \uparrow	0.68 \uparrow	0.62 \uparrow
	Dense	0.51	0.58	0.59	0.56	0.43 \downarrow	0.62 \uparrow	0.60 \uparrow	0.60 \uparrow	0.43 \downarrow	0.64 \uparrow	0.60 \uparrow	0.45 \downarrow
Greek	Sparse	0.59	0.74	0.67	0.71	0.57 \downarrow	0.71 \downarrow	0.53 \downarrow	0.66 \downarrow	0.64 \uparrow	0.65 \downarrow	0.69 \uparrow	0.60 \downarrow
	Dense	0.54	0.66	0.73	0.73	0.54	0.66	0.65 \downarrow	0.67 \downarrow	0.62 \uparrow	0.61 \downarrow	0.60 \downarrow	0.42 \downarrow
English	Sparse	0.59	0.62	0.65	0.65	0.59	0.58 \downarrow	0.41 \downarrow	0.66 \uparrow	0.50 \downarrow	0.61 \downarrow	0.66 \uparrow	0.64 \downarrow
	Dense	0.51	0.62	0.65	0.63	0.50 \downarrow	0.62	0.65	0.63	0.50 \downarrow	0.57 \downarrow	0.66 \uparrow	0.41 \downarrow

weighting for sparse representation due to its effectiveness in highlighting less common, more informative words. This approach has been utilized in similar studies, such as [2, 38, 48, 50]. We conducted experiments on all datasets both with and without stopword removal. However, we observed that eliminating stopwords led to a decline in model performance. Based on this finding, we decided to forgo stopword removal in our final approach. For dense representation, we employ a multilingual embedding model `multilingual-intfloat/multilingual-e5-large` [53], which is initialized from `xlm-roberta-large` and trained on a mixture of multilingual datasets. It supports 100 languages from `xlm-roberta`. The `e5-large` model, based on XLM-RoBERTa, features 24 layers, 16 attention heads, and a 1024 hidden size. It uses GELU activation, 0.1 dropout, and a 250,002-token vocabulary, with a 4096 intermediate size for capturing complex linguistic structures.

After obtaining feature representations through the aforementioned sparse and dense approaches, the resulting features were fed into four machine learning classifiers: Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR). To optimize the performance of classifiers, we employed grid search in combination with 5-fold cross-validation to evaluate various hyperparameter configurations. DT was tuned for different depths (`max_depth`: 10, 20, 30), while the RF was evaluated with

varying numbers of estimators (`n_estimators`: 50, 100, 200). For SVM, we experimented with different regularization strengths (`C`: 0.1, 1, 10) and kernel types (linear, rbf). Similarly, LR was tested with different `C` values (0.1, 1, 10) to control regularization. The best hyperparameters were chosen based on accuracy, and the final model was trained on the full training set.

4.3 Results

4.3.1 Binary vs. Multiclass Classification. We first ask how the performance of the models compare when Alzheimer’s detection is framed as a binary vs. multiclass classification problem (RQ1). Tables 5 and 6 present the results of binary and multiclass classification based on both dense and sparse representations, evaluated across the monolingual, multilingual-combined, and combined-translated settings. The experimental results indicate that AD detection is a significantly easier problem when formulated as a binary classification task (AD vs. HC) compared to a multiclass classification (AD vs. MCI vs. HC). Across all models and datasets, binary classification consistently yields higher accuracy, with peak performance reaching 0.90 on the combined-translated dataset (Chinese, sparse, RF). In contrast, multiclass classification demonstrates a decline in performance, with the highest observed accuracy not exceeding

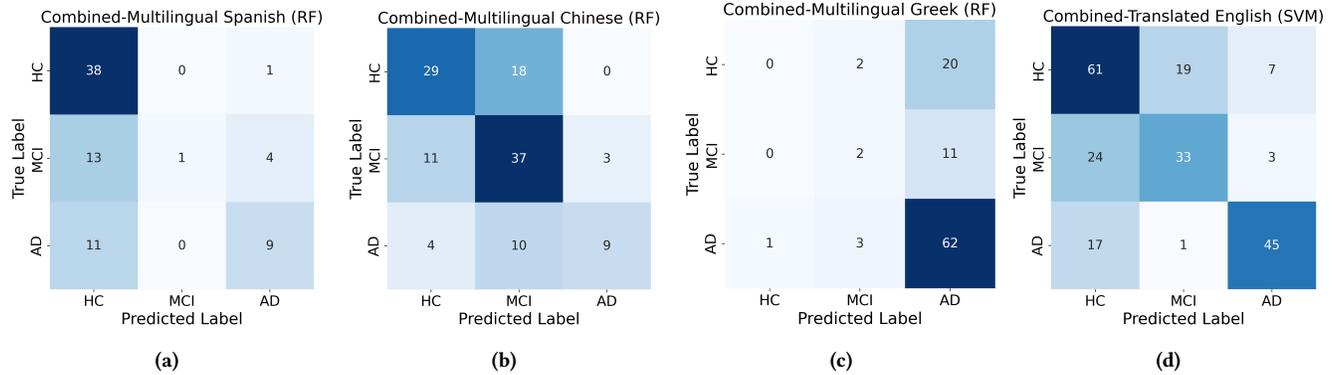


Figure 3: Confusion matrices for multiclass classification using the best-performing dataset variant and classifier for each language.

0.74 on the monolingual dataset (Greek, sparse, RF). These findings suggest that distinguishing between AD and HC is relatively straightforward, whereas differentiating between AD, MCI, and HC presents greater challenges. The reduced performance in the multiclass setting is likely attributable to the clinical and cognitive overlap between MCI and both AD and HC, leading to increased misclassification rates. This is reflected in the consistently lower accuracy observed in multiclass classification compared to the binary classification scenario.

In classification tasks, predictions made by a machine-learned classifier are subject to error, which manifests as two primary types: Type I error (false positives) and Type II error (false negatives). In AD detection, a Type I error means incorrectly classifying a HC individual as having AD, while a Type II error means incorrectly classifying an individual with AD as healthy. While both types of error are undesirable, the consequences of a Type II error in a clinical setting are far more severe. Missing a diagnosis of AD (a Type II error) delays potential treatment and intervention, potentially leading to worse patient outcomes. Conversely, a Type I error, while still causing concern and requiring further testing, allows for proactive monitoring and reduces the risk of undetected disease progression. Therefore, while we acknowledge both error types, this study prioritizes Type II error, which we measure in terms of *false negative rate*: the proportion of non-HC participants incorrectly classified as HC, relative to the total number of HC participants (i.e., $FN/(FN+TP)$). Table 7 presents the false negative rates for the best-performing classifier across all languages using the dense representation model for multiclass classification. The results for monolingual and combined-translated models are based on SVM, while the Multilingual-Combined model results are based on the RF classifier. In multiclass classification, Spanish has the highest error (0.78), followed by English (0.37), while Greek (0.06) and Chinese (0.24) perform better. The combined-multilingual dataset reduces errors, especially in Spanish (0.63) and Greek (0.01), but translation increases errors in some cases. Overall, false negative rate is high in the multiclass setting, particularly due to the difficulty of distinguishing MCI from HC, while multilingual training provides inconsistent improvements across languages.

4.3.2 Different Stages of Cognitive Decline. Next, we take a closer look at the difficulty of in differentiating between various stages

Table 7: False negative rate in multiclass classification using SVM for Monolingual and Combined-Translated, and RF for Combined-Multilingual. The arrows denote a performance improvement \uparrow or degradation \downarrow relative to the Monolingual setting.

Language	Monolingual	Combined-Multilingual	Combined-Translated
Spanish	0.78	0.63 \uparrow	0.71 \uparrow
Chinese	0.24	0.20 \uparrow	0.31 \downarrow
Greek	0.06	0.01 \uparrow	0.13 \downarrow
English	0.37	0.48 \downarrow	0.33 \uparrow

of cognitive decline in the multiclass setting. Figure 3 presents the confusion matrices for the four languages, using the models with the lowest false negative rate for each language, as shown in Table 7. Specifically, Spanish, Chinese, and Greek are based on combined-multilingual with RF, while English is based on Translated-Combined with SVM. A common issue across all models is the misclassification of MCI, which is frequently confused with either HC or AD, highlighting the progressive and overlapping nature of cognitive decline. In the Spanish model, HC cases are classified with high accuracy, but distinguishing MCI and AD remains problematic. Similarly, in the Chinese model, a substantial number of HC cases are misclassified as MCI, and while MCI is relatively well classified, some cases are incorrectly labeled as AD. The Greek model exhibits the most significant challenge in HC classification, with HC cases frequently misclassified as AD, suggesting potential dataset imbalances or language-specific difficulties in speech-based diagnosis. In contrast, AD cases in Greek are classified with high accuracy, indicating clearer distinguishing features at later disease stages. The English model, trained on the combined-translated dataset, demonstrates strong performance in identifying HC cases but struggles to distinguish MCI, with a considerable number of MCI cases being misclassified as HC or AD. Overall, the results underscore the inherent difficulty in differentiating MCI from both HC and AD, with varying degrees of classification accuracy across languages.

Another key observation is that the Spanish model predominantly predicts HC across all true labels, indicating a strong bias

that limits its ability to distinguish MCI and AD cases. In contrast, the Greek model exhibits severe misclassification toward AD, suggesting overconfidence in AD predictions while misclassifying HC and MCI cases. This pattern may result from class imbalance or insufficient training data, which could be influencing the model’s decision-making process.

4.3.3 Monolingual vs. Multilingual Dataset. Tables 5 and 6 highlight the results from the best-performing classifier in bold for each language and dataset variant. In binary classification, the impact of multilingual training varies across languages, with some benefiting while others experience a decline. The best-performing model for Chinese shows a notable improvement when trained on the combined-multilingual dataset, increasing accuracy from 0.83 to 0.86. However, Spanish sees no change, maintaining a peak accuracy of 0.80 in both monolingual and multilingual settings. In contrast, Greek and English experience declines, with Greek dropping from 0.78 to 0.76 and English decreasing from 0.81 to 0.77. These results indicate that multilingual training does not consistently enhance binary classification performance and that its effectiveness is highly language dependent.

For multiclass classification, training language-specific models on the combined-multilingual dataset generally leads to performance improvements, except for Greek. Spanish achieves a slight increase in accuracy, improving from 0.61 in the monolingual setting to 0.62 in the multilingual setting. Similarly, Chinese benefits from a boost from 0.59 to 0.62, while English sees a marginal gain from 0.65 to 0.66. However, Greek, which performs best in a monolingual setting with an accuracy of 0.74, drops to 0.71 when trained in a multilingual setting. This suggests that while multilingual training is beneficial for certain languages in multiclass classification, it does not universally enhance performance across all cases.

4.3.4 Translated Dataset. In binary classification, training on the combined-translated dataset has varying effects across languages, occasionally boosting performance but often leading to declines. The best-performing model for English benefits from translation-based augmentation, increasing accuracy to 0.83 compared to 0.81 in the monolingual setting and 0.77 in the combined-multilingual setting. Chinese experiences a raise, increasing from 0.86 with combined-multilingual training and 0.83 in monolingual to 0.90 with combined-translation. Greek sees the most significant decline, falling from 0.78 in the monolingual setting and 0.76 in the combined-multilingual setting to just 0.70 with translation-combined. Spanish remains stable at 0.80 across all datasets, indicating that translation does not provide additional value for this language.

In multiclass classification, training on the combined-translated dataset enhances accuracy for Chinese, increasing to 0.68 compared to 0.59 in the monolingual setting and 0.62 in the combined-multilingual setting. English maintains its performance at 0.66, matching the combined-multilingual setting and slightly improving over the 0.65 accuracy in the monolingual setting. However, Spanish sees a slight decline (0.60 vs. 0.61 monolingual, 0.62 combined-multilingual), while Greek experiences a more noticeable drop (0.69 vs. 0.74 monolingual, 0.71 combined-multilingual), suggesting that translation does not provide uniform benefits across languages.

5 Conclusion and Future Work

This paper has introduced a geographically diverse, multilingual conversational dataset for Alzheimer’s detection, comprising 16 publicly available datasets across English, Spanish, Chinese, and Greek, and featuring a range of cognitive assessment tasks. An English-translated version of the multilingual data is also provided to assess the impact of translation on detection performance. Key contributions include unifying heterogeneous datasets into a consistent text format and examining how task framing (binary vs. multiclass) and dataset composition (monolingual, combined-multilingual, and combined-translated) influence performance and the ability to distinguish between stages of cognitive decline.

The results of this study highlight the fundamental challenges in Alzheimer’s detection when transitioning from a binary to a multiclass classification approach. While distinguishing between cases of Alzheimer’s Disease and Healthy Controls is often achievable with reasonable accuracy, achieving the same level of accuracy in a multiclass setting that also includes patients with Mild Cognitive Impairment (MCI) proves significantly more difficult. This is largely due to the clinical and cognitive overlap between MCI and the other two categories, leading to increased misclassification rates. Our findings also shed light on the potential benefits and limitations of multilingual and translated datasets. While some languages, such as Chinese, show notable improvements with combined-multilingual or translated training, others, such as Greek, experience a decline in performance. This suggests that the effectiveness of multilingual learning is highly language-dependent, likely influenced by underlying cognitive tasks conducted in the dataset collection step, dataset size, linguistic features, and model adaptation. These findings emphasize the need for tailored approaches in multilingual Alzheimer’s detection, considering language-specific challenges and dataset characteristics to optimize diagnostic accuracy.

This multilingual conversational dataset opens up significant avenues for future research in Alzheimer’s detection. While our results demonstrate the feasibility of automatic detection, the substantial headroom for improvement suggests several promising directions. First, future research may focus on optimizing language-specific approaches, tailoring methods to the unique characteristics of each language. A deeper investigation into syntactic, semantic, and lexical features could reveal subtle, language-specific indicators of cognitive decline. Second, advanced preprocessing techniques to explicitly identify participant roles (i.e., patient vs. interviewer) within the conversations could help create more nuanced models that specifically target the patient’s language use. Third, investigating cross-lingual patterns and leveraging transfer learning techniques could enhance model generalization and robustness, particularly for languages with limited data. Fourth, incorporating more advanced deep learning methods—such as transformer-based architectures or hierarchical models—along with ensemble techniques that combine diverse models or language variants, could improve robustness and capture richer linguistic patterns. These approaches may also enhance generalization across languages and diagnostic groups. Finally, exploring advanced data augmentation methods to address the class imbalance across diagnostic groups (HC, MCI, and AD) could help improve classification performance.

References

- [1] 2021. NCMMS AD2021: Alzheimer's disease recognition evaluation 2021. <https://github.com/THUstlab/AD2021>
- [2] Surabhi Adhikari, Surendrabikram Thapa, Usman Naseem, Priyanka Singh, Huan Huo, Gnana Bharathy, and Mukesh Prasad. 2022. Exploiting linguistic information from Nepali transcripts for early detection of Alzheimer's disease using natural language processing and machine learning techniques. *International Journal of Human-Computer Studies* 160 (2022), 102761.
- [3] Felix Agbavor and Hualou Liang. 2022. Predicting dementia from spontaneous speech using large language models. *PLOS Digit Health* 1 (2022).
- [4] JT Becker, F Boller, OL Lopez, J Saxton, and KL McGonigle. 1994. Dementia bank dataset. English Pitt Corpus. *The natural history of Alzheimer's disease. Description of study cohort and accuracy of diagnosis. Arch Neurol* 51 (1994), 585–94.
- [5] Barrera-Altuna Benjamin, Lee Daeun, Zarnaz Zaima, Han Jinyoung, and Kim Seungbae. 2024. The Interspeech 2024 TAUADIAL Challenge: Multilingual Mild Cognitive Impairment Detection with Multimodal Approach. *Proc. Interspeech 2024* (2024), 967–971.
- [6] Xuchu Chen, Yu Pu, Jinpeng Li, and Wei-Qiang Zhang. 2023. Cross-lingual Alzheimer's disease detection based on paralinguistic and pre-trained features. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–2.
- [7] Louise Cummings. 2019. Describing the Cookie Theft picture: Sources of breakdown in Alzheimer's dementia. *Pragmatics and Society* 10 (2019), 151–174.
- [8] Wechsler David. 2009. *WMS-IV: Wechsler memory scale*. Pearson/PsychCorp, Pearson, San Antonio, Tex.
- [9] Junwen Duan, Fangyuan Wei, Hong-Dong Li, and Jin Liu. 2024. Pre-trained Feature Fusion and Matching for Mild Cognitive Impairment Detection. In *Proc. Interspeech 2024*. 962–966.
- [10] Kathleen C Fraser, Kristina Lundholm Fors, and Dimitrios Kokkinakis. 2019. Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Computer Speech & Language* 53 (2019), 121–139.
- [11] 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)*. Association for Computational Linguistics, 3659–3670.
- [12] Harold Goodglass and Edith Kaplan. 1972. *The assessment of aphasia and related disorders*. Philadelphia, Boston: Lea & Febiger.
- [13] Harold Goodglass, Edith Kaplan, and Sandra Weintraub. 2001. *BDAE: The Boston diagnostic aphasia examination*. Lippincott Williams & Wilkins Philadelphia, PA.
- [14] Leslie Grasset, Fiona E Matthews, Karine Péres, Alexandra Foubert-Samier, Catherine Helmer, Jean-François Dartigues, and Carol Brayne. 2018. Evolution of dementia diagnosis over time (1988–2013): Evidence from French and English cohorts. Implication for secular trends analyses. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 10 (2018), 490–497.
- [15] Angela Guarino, Francesca Favieri, Ilaria Boncompagni, Francesca Agostini, Micaela Cantone, and Maria Casagrande. 2019. Executive Functions in Alzheimer Disease: A Systematic Review. *Front Aging Neurosci* (2019).
- [16] Jing Guo, Bin Gao, Yun Huang, and Suhang Song. 2024. Trajectory of multimorbidity before dementia: A 24-year follow-up study. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 16, 1 (2024), e12523.
- [17] Yue Guo, Changye Li, Carol Roan, Serguei Pakhomov, and Trevor Cohen. 2021. Crossing the "Cookie Theft" corpus chasm: applying what BERT learns from outside data to the ADReSS challenge dementia detection task. *Frontiers in Computer Science* 3 (2021), 642517.
- [18] Pamela Herd, Deborah Carr, and Carol Roan. 2014. Cohort profile: Wisconsin longitudinal study (WLS). *International journal of epidemiology* 43, 1 (2014), 34–41.
- [19] John Hobson. 2015. The montreal cognitive assessment (MoCA). *Occupational Medicine* 65, 9 (2015), 764–765.
- [20] Olga Ivanova, Juan José G Meilán, Francisco Martínez-Sánchez, Israel Martínez-Nicolás, Thide E Llorente, and Nuria Carcavilla González. 2022. Discriminating speech traits of Alzheimer's disease assessed through a corpus of reading task for Spanish language. *Computer Speech & Language* 73 (2022), 101341.
- [21] Minni Jain, Rishabh Doshi, Vibhu Sehra, and Divyashikha Sethia. 2021. Exploring the Effects of Different Embedding Algorithms and Neural Architectures on Early Detection of Alzheimer's Disease.. In *ISIC*. 376–383.
- [22] Bente Johnsen, Ieva Martinaityte, Tom Wilsgaard, and Henrik Schirmer. 2023. Incidence of dementia over a period of 20 years in a Norwegian population. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 15, 4 (2023), e12479.
- [23] Phillip R Johnston, Anthony R McIntosh, and Jed A Meltzer. 2023. Spectral slowing in chronic stroke reflects abnormalities in both periodic and aperiodic neural dynamics. *NeuroImage: Clinical* 37 (2023), 103277.
- [24] Maria Kaltsa, Tsolaki Anthoula, Ioulietta Lazarou, Ilias Mittas, Mairi Papageorgiou, Despina Papadopoulou, Ianthi Maria Tsimpli, and Magda Tsolaki. 2024. Language Markers of Dementia and Their Role in Early Diagnosis of Alzheimer's Disease: Exploring Grammatical and Syntactic Competence via Sentence Repetition. *Journal of Alzheimer's disease reports* 8 (2024).
- [25] Anastasios Karakostas, Alexia Briassouli, Konstantinos Avgerinakis, Ioannis Kompatsiaris, and Magda Tsolaki. 2016. The dem@ care experiments and datasets: a technical report. *arXiv preprint arXiv:1701.01142* (2016).
- [26] Daniel Kempler, Susan Curtiss, and Catherine Jackson. 1987. Syntactic preservation in Alzheimer's disease. *Journal of Speech, Language, and Hearing Research* 30, 3 (1987), 343–350.
- [27] Eli Kurtz, Youxiang Zhu, Tiffany Driesse, Bang Tran, John A Batsis, Robert M Roth, and Xiaohui Liang. 2023. Early detection of cognitive decline using voice assistant commands. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
- [28] Christopher A Lane, John Hardy, and Jonathan M Schott. 2018. Alzheimer's disease. *European journal of neurology* 25, 1 (2018), 59–70.
- [29] Alyssa M Lanzi, Anna K Saylor, Davida Fromm, Houjun Liu, Brian MacWhinney, and Matthew L Cohen. 2023. DementiaBank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology* 32, 2 (2023), 426–438.
- [30] Muriel Deutsch Lezak. 2004. *Neuropsychological assessment*. Oxford University Press, USA.
- [31] Xiaohui Liang, John A Batsis, Youxiang Zhu, Tiffany M Driesse, Robert M Roth, David Kotz, and Brian MacWhinney. 2022. Evaluating voice-assistant commands for dementia detection. *Computer Speech & Language* 72 (2022), 101297.
- [32] Ning Liu and Zhenming Yuan. 2021. Spontaneous language analysis in Alzheimer's disease: evaluation of natural language processing technique for analyzing lexical performance. *Journal of Shanghai Jiaotong University (Science)* (2021), 1–8.
- [33] Ning Liu, Zhenming Yuan, and Qingfeng Tang. 2022. Improving Alzheimer's disease detection for speech based on feature purification network. *Frontiers in Public Health* 9 (2022), 835960.
- [34] Saturnino Luz, Sofia De La Fuente Garcia, Fasih Haider, Davida Fromm, Brian MacWhinney, Alyssa Lanzi, Ya-Ning Chang, Chia-Ju Chou, and Yi-Chien Liu. 2024. Connected Speech-Based Cognitive Assessment in Chinese and English. (2024).
- [35] Saturnino Luz, Fasih Haider, Davida Fromm, Ioulietta Lazarou, Ioannis Kompatsiaris, and Brian MacWhinney. 2023. Multilingual alzheimer's dementia recognition through spontaneous speech: a signal processing grand challenge. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–2.
- [36] Brian MacWhinney. 2017. Tools for analyzing talk part 1: The chat transcription format. *Carnegie.[Google Scholar]* 16 (2017).
- [37] Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. 2011. AphasiaBank: Methods for studying discourse. *Aphasiology* 25, 11 (2011), 1286–1307.
- [38] Matej Martinc and Senja Pollak. 2020. Tackling the ADReSS Challenge: A Multimodal Approach to the Automated Recognition of Alzheimer's Dementia.. In *Interspeech*. 2157–2161.
- [39] Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. 2005. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society* 53, 4 (2005), 695–699.
- [40] Isabel Orozco-Chavez, Moisés Martínez-Estrada, and Benjamin A Itzá-Ortiz. 2024. An automatic Alzheimer's disease classifier based on reading task for Spanish language. *The European Physical Journal Special Topics* (2024), 1–15.
- [41] David Ortiz-Perez, Manuel Benavent-Lledo, Jose Garcia-Rodriguez, David Tomás, and M Flores Vizcaya-Moreno. 2024. Deep Insights into Cognitive Decline: A Survey of Leveraging Non-Intrusive Modalities with Deep Learning Techniques. *arXiv preprint arXiv:2410.18972* (2024).
- [42] Paula Andrea Pérez-Toro, Philipp Klumpp, Abner Hernandez, Tomas Arias, Patricia Lillo, Andrea Slachevsky, Adolfo Martin Garcia, Maria Schuster, Andreas K Maier, Elmar Noeth, et al. 2022. Alzheimer's Detection from English to Spanish Using Acoustic and Linguistic Embeddings.. In *Interspeech*. 2483–2487.
- [43] Ulla Petti, Simon Baker, and Anna Korhonen. 2020. A systematic literature review of automatic Alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association* 27, 11 (2020), 1784–1797.
- [44] Xiaoke Qi, Qing Zhou, Jian Dong, and Wei Bao. 2023. Noninvasive automatic detection of Alzheimer's disease from spontaneous speech: a review. *Frontiers in Aging Neuroscience* 15 (2023), 1224723.
- [45] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.
- [46] Knopman David S. and Petersen Ronald C. 2014. Mild cognitive impairment and mild dementia: a clinical perspective. *Mayo Clinic proceedings* (2014), 1452–1459.
- [47] Roozbeh Sadeghian, J David Schaffer, and Stephen A Zahorian. 2021. Towards an automatic speech-based diagnostic test for Alzheimer's disease. *Frontiers in Computer Science* 3 (2021), 624594.

- [48] Thomas Searle, Zina Ibrahim, and Richard Dobson. 2020. Comparing natural language processing techniques for Alzheimer's dementia prediction in spontaneous speech. *arXiv preprint arXiv:2006.07358* (2020).
- [49] Arezo Shakeri and Mina Farmanbar. 2025. Natural language processing in Alzheimer's disease research: Systematic review of methods, data, and efficacy. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 17, 1 (2025), e70082.
- [50] Arezo Shakeri, Shaima Ahmad Freja, Yeganeh Hallaj, and Mina Farmanbar. 2024. Uncovering Linguistic Patterns: A Machine Learning Exploration for Early Dementia Detection in Speech Transcripts. In *2024 4th International Conference on Applied Artificial Intelligence (ICAPAI)*. 1–8.
- [51] Alejandro Suárez Rodríguez. 2024. Is Alzheimer's Disease a Cross-Linguistic Issue? Comparing Corpora from a Role and Reference Grammar Perspective. *Círculo de lingüística aplicada a la comunicación* (2024).
- [52] Bastiaan Tamm, Rik Vandenberghe, and Hugo Van Hamme. 2023. Cross-Lingual Transfer Learning for Alzheimer's Detection from Spontaneous Speech. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–2. doi:10.1109/ICASSP49357.2023.10096770
- [53] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint arXiv:2402.05672* (2024).
- [54] Bingyang Wen, Ning Wang, Koduvayur Subbalakshmi, Rajarathnam Chandramouli, et al. 2023. Revealing the Roles of Part-of-Speech Taggers in Alzheimer Disease Detection: Scientific Discovery Using One-Intervention Causal Explanation. *JMIR Formative Research* 7, 1 (2023), e36590.
- [55] Michael Woodward and M Galea. 2005. Mini-mental state examination (MMSE). *Aust J Physiother* 51, 3 (2005), 198.
- [56] Qin Yang, Xin Li, Xinyun Ding, Feiyang Xu, and Zhenhua Ling. 2022. Deep learning-based speech analysis for Alzheimer's disease detection: a literature review. *Alzheimer's Research & Therapy* 14, 1 (2022), 186.
- [57] Yangwei Ying, Tao Yang, and Hong Zhou. 2023. Multimodal fusion for alzheimer's disease recognition. *Applied Intelligence* 53, 12 (2023), 16029–16040.