Natural Language Processing of Asian Speech for Dementia Detection

Rez Samantha Floresca

August 23, 2025

Table of Contents

1. Introduction 3

2. Background 5
    A. Language Decline in Dementia 5
    B. Speech-Based Features for Dementia Detection 6
    C. Linguistic Variations Across Regions 9
    D. Natural Language Processing (NLP) for Dementia Detection 11

3. Methodology 15
    I. Meta-Analysis 16
        A. Establishment of Regional Baselines 16
        B. Comparative Analysis of Dementia and Baseline Speech Features Across Regions 24
        C. Analyzing the Performance of NLP Models in Each Region 28
    II. Prototype 33
        A. Anglophone – English (Control) 34
        B. East Asian – Standard Mandarin (普通话 / Putonghua) 36
        C. South Asian – Hindi 39
        D. Southeast Asian – Tagalog 42

4. Results and Discussion 44

5. Limitations and Recommendations 46

6. Conclusion 46

7. Appendices 47

8. Bibliography 62

Introduction

Dementia is a global health crisis affecting more than 55 million people worldwide, with the majority of cases emerging in low- and middle-income countries *(WHO, 2025)*. Despite the large societal impact, early detection remains elusive, especially in underserved regions where neuroimaging and specialized cognitive assessments are either prohibitively expensive or scarce. In high-income settings, dementia recognition occurs in only 20–50% of primary care cases, with even lower detection rates in low- and middle-income countries *(Gamble et al., 2019)*. Extrapolating these trends suggests that roughly three-quarters of individuals with dementia globally remain undiagnosed. *(Alzheimer's Disease International, n.d., para. 4)*.

This global situation is acutely mirrored across Asia. With Asia-Pacific accounting for over half of the global dementia burden, the region faces a steep and growing trajectory. In 2021, East Asia exhibited some of the highest age-standardized prevalence and mortality rates, notably in China, where over 16 million people live with dementia, representing nearly 30% of global cases.

While South Asia reported lower age-standardized rates relative to East Asia, the absolute burden remains large: Bangladesh, India, and Pakistan together are home to nearly 5 million dementia cases, with projections rising to over 24 million by 2050.

Southeast Asia is similarly affected. A regional systematic review estimated approximately 5.5 million people aged 60+ living with dementia across the WHO Southeast Asia Region in recent years, and projections suggest this figure will nearly double to 10 million by 2030. Alarmingly, Southeast Asia's total dementia-related cost in 2015 alone reached US $7.3 billion —about one-third of Nepal's GDP.

With that, current research has relied on Natural Language Processing (NLP) technology – a branch of Artificial Intelligence (AI) that enables computers to understand, interpret, and generate human language *(Oracle, n.d.)*. In addition to memory impairments, dementia causes substantial changes in speech production, particularly lexical-semantic characteristics *(Ostrand et al., 2020*). Speech-based symptoms, such as decreased lexical diversity, increased pause duration, or simplified sentence structure, are among the earliest indicators of cognitive decline. NLP enables the automatic analysis of these linguistic markers at scale. It can process human speech to determine gaps and disparities compared to the majority of the data set it has already collected. NLP has the potential to provide a non-invasive, cost-effective method using a timely intervention for detecting early-stage language and cognitive decline in individuals concerned about their memory *(Khade et al., 2023, para. 1)*. In principle, a speech recording collected on a basic smartphone could be processed by an NLP model to flag potential dementia symptoms—making early detection more widely accessible.

However, spoken language is not uniform across cultures or regions. The structure, rhythm, and norms of speech vary widely across linguistic contexts. Most existing NLP research in dementia is rooted in English-speaking, high-resource settings, and as such, fails to account for these regional linguistic variations. A systematic review of 240 studies on natural language processing (NLP) in dementia research by Peled-Cohen and Reichart (2025) stated in the study limitations that only research using data fully or partially in English was included, explicitly excluding studies conducted entirely in other languages. These instances risk embedding structural biases, potentially leading to unreliable results in other regions. For instance,he honorific-laden syntax of East Asian languages and the frequent code-switching found in South and Southeast Asia are not accounted for – this can introduce linguistic complexities that can affect the performance of NLP models.

This study addresses these gaps through a two-pronged approach. First, a meta-analysis of exclusive linguistic features categorized by regions—namely, East Asian, South Asian, Southeast Asian, and Anglophone (English) as the control group–and NLP models for dementia detection is conducted. For each region, the lexical, syntactic, and acoustic speech features of individuals with and without dementia are analyzed and compared to identify region-specific linguistic indicators of cognitive decline. Various NLP model types, including Naive Bayes and Support Vector Machines (SVM) architectures, are evaluated for their effectiveness and suitability across diverse language structures. Second, a lightweight prototype classifier is developed using open-source data to demonstrate the feasibility of deploying speech-based dementia detection tools in low-resource settings.

Background

A. Language Decline in Dementia

Dementia describes an overall decline in memory and other cognitive skills severe enough to reduce a person's ability to perform everyday activities. The progressive and persistent deterioration of cognitive function characterizes it *(Emmady, 2022)*. Among its earliest and most observable symptoms are changes in spontaneous speech. Individuals with dementia may exhibit reduced lexical diversity, frequent use of vague pronouns, simplified sentence structures, and increased pauses or hesitations. There is growing evidence that subtle changes in spontaneous speech may reflect early pathological changes in cognitive function. Recent work has found that lexical-semantic features of spontaneous speech predict cognitive dysfunction in individuals with mild cognitive impairment (MCI) *(Burke et al., 2023, para. 1)*.

A group of scientists from the University of California, San Francisco, Boston University, and Stanford University investigated the connection between Alzheimer's biomarkers in the brain and alterations in speech patterns. The researchers discovered that speaking more slowly and pausing longer and more frequently was associated with higher levels of tau protein in the early neocortical region and the medial temporal region of the brain. Contrary to intuitive belief, tau protein was not linked to the memory score itself, indicating that speech abnormalities could be early indicators of Alzheimer's before memory abnormalities manifest *(National Institute on Aging, 2022)*.

These semantic patterns are measurable through linguistic analysis, making speech a valuable, non-invasive indicator for early detection.

B. Speech-based Features for Dementia Detection

Speech-based dementia Detection focuses on how language and vocal production change as cognitive decline advances in an effort to find quantifiable differences between people with dementia and healthy controls. It typically focuses on three categories of features:

B.1. Lexical Features

The richness, diversity, and specificity of vocabulary—important markers of an individual's cognitive-linguistic ability—are captured by lexical characteristics in speech-based diagnosis. A person's word use can be measured using metrics like the Type–Token Ratio (TTR), Brunet's Index, Sichel's measure, and more recent ones like MTLD (Measure of Textual Lexical Diversity). One of the earliest language-based indicators of dementia is word-finding difficulty and semantic memory impairment, which are frequently reflected in diminished lexical diversity.

Clear lexical decrease in dementia is repeatedly shown by empirical research. For instance, even in brief narrative tasks like picture descriptions, Kothari et al. (2023) discovered significant decreases in lexical diversity among dementia speakers as compared to controls using Brunet's and Sichel's methods. In a related study, Zozuk (2025) found that speech from people with Alzheimer's disease (AD) had lower MTLD and GTTR scores (lexical diversity metrics independent of text length) than speech from healthy controls. This finding confirmed that lexical decline is not just a result of the sample size. The sensitivity of lexical metrics in identifying early cognitive decline is well supported by these findings.

B.2. Syntactic Features

By evaluating how words are combined into sentences using hierarchical grammatical relations, syntactic features indicate the structural complexity of spoken language. The ability to produce complicated sentence structures, such as center-embedded or hypotactic constructions, rapidly declines in early dementia, yet basic grammatical competency frequently endures. Thus, tracking syntactic complexity—such as sentence embedding and dependence distance—provides insight into modest cognitive-linguistic loss.

When Gao and He (2024) compared Alzheimer's disease (AD) patients with healthy controls (HC) using a dependency-distance framework, they found that AD speech had more head-final constructions and much shorter mean dependency distances, which suggests a simpler syntactic structure and less working memory capacity. Furthermore, Ivanova et al. (2023) highlighted that patients may still write grammatically correct sentences even when syntactic reduction takes place, highlighting the fact that loss is quantitative rather than qualitative—a distinction that is crucial for developing detection models.

B.3. Acoustic Features

The non-lexical aspects of speech are covered by acoustic features, which include spectral markers like pitch, formant frequencies, and MFCCs (Mel-frequency cepstral coefficients) as well as temporal measures like pause duration, speech and articulation rate, and others. These signals provide objective markers of neurocognitive decline by reflecting underlying motor, prosodic, and rhythmic changes in speech output.

People with mild or moderate dementia spent a far higher percentage of time in quiet pauses—up to 68% in moderate cases—than healthy controls, who averaged about 41%, according to an acoustic investigation by Sluis et al. (2020). Rate and interruption metrics, such as articulation rate, speech-segment duration, and voice breaks, offer a strong distinction between dementia and normal speech, according to meta-analyses of acoustic speech in AD. In the meantime, Nagumo et al. (2020) showed that spectral acoustic characteristics, such as speech duration and formant frequency variation, can successfully differentiate MCI from healthy persons.

Because acoustic biomarkers rely on physiological signs of cognitive decline rather than linguistically specific information, they provide significant benefits in multilingual and low-resource environments. By combining auditory characteristics with lexical and syntactic metrics, particularly through multimodal models, detection systems become more multiplex and generalizable across various Asian locations.

C. Linguistic Variations Across Regions

By analyzing voice and language characteristics, speech-based dementia detection has shown promising results, particularly for early cognitive decline. However, most research has focused on English, limiting the generalizability of existing models to other language populations *(Fraser et al., 2016; Pulido et al., 2020)*. To address this gap, this study groups languages into four major regions based on typology, structural similarity, and data availability. Anglophone languages serve as the control. East Asian languages feature tonal or pitch-accent systems and agglutinative syntax. South Asian languages are morphologically rich and often involve diglossia and code-switching. Southeast Asian languages are typologically diverse, including tonal, analytic, and Austronesian structures. These groupings enable a systematic evaluation of NLP models across linguistically and culturally distinct populations.

C.1. Anglophone – English (Control)

English is the most extensively studied language in dementia research, supported by large datasets such as DementiaBank *(Becker et al., 1994)*. Speech biomarkers identified in Anglophone populations include lexical retrieval deficits, reduced vocabulary richness, increased pause frequency, syntactic simplification, and decreased semantic coherence *(Fraser et al., 2016).* English is stress-timed and only moderately complex morphologically, which shapes how dementia-related speech impairments appear and provides a clear foundation for feature extraction and model design.

C.2. East Asian

East Asian languages such as Mandarin, Cantonese, Japanese, and Korean differ from English through features like tonality, logographic scripts, and morphosyntactic structures. In Mandarin, tonal changes carry meaning, so prosodic impairments can be strong diagnostic markers *(Zhu et al., 2021)*. Japanese and Korean present additional challenges because their agglutinative morphology and topic-prominent syntax make lexical retrieval and sentence planning more difficult *(Suzuki et al., 2015)*. Research shows that dementia in East Asian populations often involves reduced syntactic variety, problems with lexical access, and loss of tonal control *(Tian et al., 2023)*.

C.3. South Asian

South Asian languages such as Hindi, Urdu, Bengali, and Tamil are morphologically rich and often inflectional, with high levels of syntactic embedding. Multilingualism is common, which complicates diagnosis because dementia patients may code-switch between languages, reflecting both cultural and cognitive processes *(Banerjee et al., 2021)*. In India, dementia speech studies have shown impairments in verbal fluency and discourse coherence, with patterns distinct from those seen in English-speaking cohorts *(Jotheeswaran et al., 2010).* Furthermore, diglossia—where formal and informal registers of a language coexist—can affect assessments, as dementia patients often revert to more familiar or simplified registers *(Sathish et al., 2022)*.

C.4. Southeast Asian

Southeast Asian languages, such as Thai, Vietnamese, Tagalog, and Bahasa Indonesia, exhibit diverse typological features, ranging from tonal systems (Thai, Vietnamese) to Austronesian morphology (Tagalog, Malay). Tonal impairment has been identified as a marker of cognitive decline in tonal languages, while verb-focused morphosyntax in Austronesian languages influences how dementia-related speech degradation manifests *(Nguyen et al., 2018)*. In the Philippines, dementia discourse analyses show decreased narrative coherence and lexical diversity, which differ from patterns observed in English corpora (Ligsay & Carandang, 2020). A lack of large digital corpora in Southeast Asian languages further limits the development of robust dementia-speech models.

D. Natural Language Processing (NLP) for Dementia Detection

Natural language processing (NLP), a subfield of artificial intelligence (AI), involves the use of machine learning to enable computers to process, understand, and generate human language *(IBM, 2024; Shi, 2021)*. NLP technologies can extract linguistic features such as word frequency, pause duration, or sentence complexity from spoken or written data – attributes that may reflect underlying cognitive state.

In dementia research, NLP has been applied to analyze patterns in patient speech that may indicate cognitive decline. For instance, Ivraghi et al. (2024) demonstrated that NLP-based transfer-learning models could successfully distinguish Alzheimer’s patients from cognitively healthy older adults using speech transcript data. Similarly, a large-scale DementiaBank study by Nyongesa et al. (2025) found significant differences in pronoun usage, syntactic complexity, and lexical sophistication across Alzheimer’s, mild cognitive impairment, and healthy groups. NLP can be deployed on simple devices and scaled quickly, therefore it presents a promising alternative to costly clinical tests or neuroimaging, especially in resource-limited environments.

To operationalize NLP in early dementia detection across diverse linguistic contexts, this study focuses on two established classifiers: Naïve Bayes, a lightweight probabilistic model, and Support Vector Machines (SVMs), a more computationally intensive but accurate margin-based approach.

D.1. Naive Bayes

The Bayes Theorem, which indicates the probability of a class label based on observable input features, is the foundation of the probabilistic classification technique known as Naive Bayes (NB) *(Hayes, 2025).* It streamlines computation by assuming that all features are conditionally independent given the class *(CQF, n.d.)*. NB classifiers have been used to find broad lexical and semantic trends in spoken language text transcripts for dementia identification. Cognitive decline can be indicated by characteristics including pronoun usage, sentence length, and word frequency.

Because of their simplicity and low processing cost, NB models have shown promise in tiny or noisy datasets. For instance, Jurafsky et al. (2025) showed that NB could categorize human texts based on sentiment, which is comparable to differentiating between people with cognitive impairments and those in good condition. The ability of NB to identify signs of Alzheimer's disease and other types of cognitive impairment, such as diminished vocabulary richness, lexical degradation, and excessive use of ambiguous phrases, was demonstrated in early dementia investigations. When linguistic features are highly correlated, as they frequently are in genuine speech, NB's independence assumption may restrict accuracy despite its interpretability and effectiveness.

Lexical analysis is especially well-suited to NB classifiers. Simple measures such as reduced lexical diversity and pronoun overuse provide interpretable predictors even in small or noisy datasets. In low-resource environments, these structurally independent traits are valuable because they can be adapted across linguistic regions, as long as stopwords and function-word distributions are adjusted to local contexts.

D.2. Support Vector Machines (SVMs)

Support Vector Machines (SVMs) are margin-based classifiers that perform particularly well in high-dimensional feature spaces like bag-of-words or n-gram models. They can handle non-linear data using kernel functions, such as linear and radial basis function (RBF) kernels, and are intended to identify the optimal hyperplane that maximally separates classes, such as "dementia" versus "healthy" *(Cortes & Vapnik, 1995; Schölkopf & Smola, 2002)*.

SVMs have been used to analyze variables including pronoun frequency, lexical richness, and syntactic complexity in speech and transcript data for dementia identification. Research using DementiaBank datasets demonstrates that SVMs are more accurate than simpler models at differentiating between Alzheimer's disease, mild cognitive impairment (MCI), and healthy aging *(Nyongesa et al., 2025)*.

SVMs are also particularly effective in capturing syntactic complexity measures, such as dependency distance, phrase embedding depth, and directionality. Because SVMs can model correlated features, they are better able to detect subtle reductions in syntactic variety that signal cognitive decline. Regional calibration is critical since syntactic norms vary widely.

D.3. Model Comparison

Naïve Bayes offers a lightweight, interpretable baseline suited for small datasets, while SVM achieves higher accuracy on complex, multimodal features but at greater computational cost. In this study, these trade-offs are evaluated in relation to which model type is most suitable for the distinct features of each linguistic region. This allows us to explore how probabilistic versus margin-based models align with the unique challenges posed by Anglophone, East Asian, South Asian, and Southeast Asian speech.

Methodology

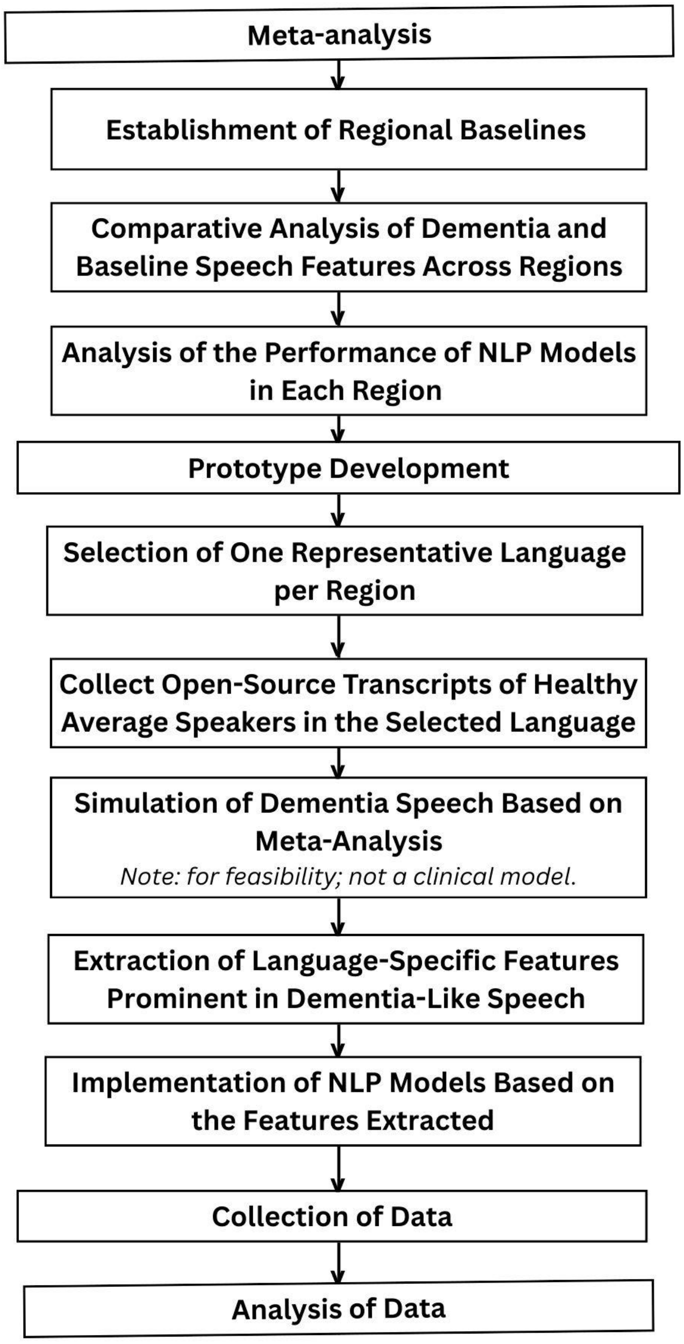

Meta-analysis
Establishment of Regional Baselines
Comparative Analysis of Dementia and Baseline Speech Features Across Regions
Analysis of the Performance of NLP Models in Each Region
Prototype Development
Selection of One Representative Language per Region
Collect Open-Source Transcripts of Healthy Average Speakers in the Selected Language
Simulation of Dementia Speech Based on Meta-Analysis
Note: for feasibility; not a clinical model.
Extraction of Language-Specific Features Prominent in Dementia-Like Speech
Implementation of NLP Models Based on the Features Extracted
Collection of Data
Analysis of Data

Figure 1. Process Flowchart

I. Meta-analysis

A. Establishment of Regional Baselines

The first stage analyzes the linguistic properties of each regional group (Anglophone, East Asian, South Asian, Southeast Asian), with emphasis on features that are exclusive or highly characteristic of these languages. The goal is to identify lexical, syntactic, and acoustic traits that are absent—or far less prominent—in English and other regions, since these distinctive features determine which speech markers are most relevant for dementia detection in each linguistic context.

A.1. Anglophone — English (Control)

A.1.1. Lexical Features

Spoken English is characterized by relatively low lexical density compared to written registers, reflecting frequent use of function words (pronouns, auxiliaries) and discourse markers (*"well," "you know," "like"*) for turn-taking and interpersonal alignment *(Biber et al., 1999)*. Lexical diversity is often measured via TTR, MTLD, and Brunet's index; these indices are sensitive to sample length, making task-specific baselines crucial *(McCarthy & Jarvis, 2010)*. Empirical data show moderate lexical diversity in spontaneous conversation, with consistent patterns across tasks *(Fergadiotis et al., 2015)*.

Healthy English speakers produce predictable patterns: frequent function words, discourse markers, and pragmatic formulae alongside stable but modest lexical diversity. Baselines should therefore capture both global indices (MTLD) and categorical proportions of content vs. function words. Additional markers like high-frequency verb use ("*do," "get," "go"*) and discourse particles provide nuanced lexical profiling for dementia detection.

A.1.2. Syntactic Features

Canonical word order is SVO, but conversational English exhibits ellipsis, truncated clauses, and right-branching subordination. Metrics like MLU and subordination ratios quantify syntactic complexity *(Brown, 1973; Hunt, 1965)*. Typical spoken English uses simple clause chaining, moderate embedding, and frequent complement clauses with markers like "that" or "because."

Conversational simplifications include pragmatic ellipses ("Want coffee?"), left dislocation ("That guy, he's nice"), and phrasal fragments. Dependency-based metrics such as mean dependency distance reliably capture syntactic load. Baselines should record clause complexity, pragmatic ellipses, subordination, and dependency distances to distinguish normative variation from pathological decline.

A.1.3. Acoustic Features

English is stress-timed, with alternating strong/weak syllables and vowel reduction in unstressed positions *(Grabe & Low, 2002).* Speech rate averages 150–160 wpm, with variability from cognitive load, narrative demands, and affective expression *(Tauroza & Allison, 1990)*. Prosody is shaped by nuclear pitch accents, boundary tones, and intonational phrasing *(Pierrehumbert & Hirschberg, 1990).*

Baseline acoustic profiling should include speech rate, articulation rate, pause frequency, F0 mean and variability, vowel formant dispersion, and rhythmic metrics such as nPVI. These collectively capture normative English prosody for comparison against pathological patterns.

A.2. East Asian

A.2.1. Lexical Features

The East Asian linguistic area is diverse, but commonalities emerge when examining Mandarin Chinese, Cantonese, Japanese, and Korean. Mandarin and Cantonese are analytic languages with extensive use of classifiers and measure words, as well as topic–comment structures that reduce reliance on overt pronouns *(Li & Thompson, 1981*). Lexical items are predominantly monosyllabic or disyllabic, and tonal distinctions are lexically contrastive (four tones in Mandarin, up to six or more in Cantonese), making tone realization a lexical as well as phonological feature *(Yip, 2002).*

By contrast, Japanese and Korean exhibit agglutinative morphology, with particles marking case, topic, and discourse roles. Lexical choice is heavily influenced by register, especially in the use of honorific and humble forms, while light verbs and auxiliaries are common in everyday discourse *(Shibatani, 1990; Sohn, 1999)*. Across these languages, average speakers use a high proportion of function-like elements (particles, classifiers, honorifics) that encode pragmatic and relational meanings. Baselines should therefore measure classifier/particle frequency, topic continuity (zero pronoun use), and honorific distributions as key lexical markers.

A.2.2. Syntactic Features

Sinitic languages such as Mandarin display canonical SVO word order but also exhibit strong topic-prominent tendencies: arguments are frequently dropped if recoverable from context, and serial verb constructions are common (*Li & Thompson, 1981)*. Cantonese, in particular, relies heavily on aspect markers and sentence-final particles to structure discourse.

Japanese and Korean, in contrast, are strictly head-final with canonical SOV order. Syntax is marked by clause chaining, frequent subordination, and extensive use of case-marking particles *(Miyagawa, 2010; Sohn, 1999)*. Scrambling allows for flexible word order, but grammatical relations remain transparent through particles. In spontaneous speech, average speakers of these languages rely on zero anaphora, particle-marked dependencies, and deep subordination in narrative contexts.

Normative baselines should, therefore, quantify argument omission rates (zero pronouns), dependency directionality (head-final vs. head-initial), and particle distributions. These metrics capture the central syntactic strategies typical of East Asian speech.

A.2.3. Acoustic Features

A defining feature of Mandarin and Cantonese is their tonal systems: lexical meaning is encoded in pitch contour, with F0 height and shape distinguishing otherwise identical syllables. Prosody overlays tone, with intonational patterns marking interrogatives or focus, but tonal identity is preserved within these contours *(Xu, 2005)*. Normative speech rates in Mandarin average around 4–5 syllables per second, slightly slower than English, and syllable timing is relatively even *(Lin & Wang, 2007)*.

Japanese is a pitch-accent language in which lexical items are distinguished by high/low pitch patterns, while Korean lacks lexical tone but exhibits distinctive intonational phrasing *(Jun, 2005)*. Across these languages, average speakers demonstrate narrower F0 ranges than English speakers, but tonal or pitch-accent contrasts are sharply realized.

Baseline acoustic profiling should therefore incorporate F0 contour tracking (tone/accent classification), syllable-timing measures, speech rate, and pause distributions. This set of measures reflects the normative acoustic landscape of East Asian languages, against which pathological variation may be evaluated.

A.3 South Asian

A.3.1. Lexical Features

South Asian languages exhibit rich typological diversity, encompassing Indo-Aryan (Hindi, Urdu, Bengali, Sinhala) and Dravidian (Tamil, Telugu, Kannada, Malayalam) families. Lexically, Indo-Aryan languages manifest dense functional morphology, particularly in case marking and verb agreement, resulting in elevated morpheme-to-word ratios in spoken discourse *(Masica, 1991; Borin et al., 2013)*. Light verbs (*kar dena*, *le lena*) in Hindi and Urdu occur frequently, enhancing verb-token frequency and contributing to syntactic flexibility *(Butt & Geuder, 2001).* Bengali displays extensive cliticization and postpositional usage, whereas Tamil relies on agglutinative suffixation to expand lexical items *(Lehmann, 1993)*. Corpus-based analyses indicate average TTR (Type-Token Ratio) values between 0.55–0.65 in adult speakers for narrative speech, with morpheme-level MTLD measures ranging 70–90, reflecting high functional load from bound morphology *(Borin et al., 2013; Annamalai, 2000)*. Pragmatic particles (*toh*, *na*) and reduplication for emphasis are pervasive, contributing further to lexical complexity. Baseline measures for South Asian speakers should therefore include word- and morpheme-level diversity, light verb frequency, particle frequency, and reduplication counts per 100 tokens, ideally stratified by age and dialectal region.

A.3.2. Syntactic Features

Canonical word order in most South Asian languages is SOV, but syntactic variation is widespread due to discourse-driven constituent fronting and pro-drop phenomena *(Mahajan, 1990)*. Dravidian languages frequently employ clause chaining through participial and non-finite verb forms, whereas Indo-Aryan languages, such as Hindi/Urdu, use embedded *ki*-clauses and relative clauses extensively *(Annamalai, 2000; Simpson & Bhattacharya, 2003)*. Null subjects, resumptive pronouns, and topicalization are common, especially in narrative contexts *(Gair & Karunatillake, 1998)*. Quantitative metrics for syntactic baselines include subordination ratio (0.3–0.5 dependent clauses per independent clause), mean dependency distance (~2.8–3.5 words), frequency of null arguments, and clause-chaining depth. Average speakers produce shorter independent clauses punctuated by longer complex chains, maintaining coherence through frequent discourse resumptive elements. These syntactic patterns are critical when comparing across languages or designing NLP pipelines for dementia detection.

A.3.3. Acoustic Features

South Asian languages display prosodic systems intermediate between stress- and syllable-timing, with mora-sensitive rhythm in Hindi/Urdu and syllable-timed tendencies in Tamil *(Rao, 2009; Keane, 2006)*. Phonemic contrasts rely primarily on vowel length, gemination, and consonantal distinctions rather than tone, though intonation signals focus, question types, and discourse boundaries *(Patil et al., 2008)*. Average speech rate varies between 120–150 words per minute, with shorter intra-clausal pauses than Anglophone languages *(Miller et al., 2009)*. Acoustic baselines should incorporate rhythm metrics (nPVI, VarcoV), F0 contours for interrogatives vs. declaratives, vowel quality dispersion, and pause distributions. Typical speakers exhibit stable mid-range F0 with rising boundary tones marking interrogatives and use vowel length contrasts as primary temporal cues.

A.4 Southeast Asian

A.4.1. Lexical Features

Southeast Asian languages are typologically diverse, including tonal languages (Vietnamese, Thai) and non-tonal languages (Khmer, Burmese, Malay/Indonesian). Analytic structures dominate, relying on word order and particles rather than inflection *(Enfield, 2005)*. Tonal languages encode lexical identity through pitch, making tonal realization both lexical and phonological.

Lexical diversity varies: word-level TTR ranges 0.60–0.70 with minimal morphological variation. Discourse particles (e.g., “na,” “lah,” “ba”), classifiers (Thai, Vietnamese), and serial verb constructions are key lexical markers.

Baselines should track particle frequency, classifier use, serial verb occurrence per clause, and tonal distinctions, accounting for their role in semantic contrast.

A.4.2. Syntactic Features

Syntactic structures generally follow topic-comment ordering with short paratactically linked clauses. Thai and Vietnamese follow SVO but allow zero anaphora and frequent topicalization. Burmese uses clause chaining with conjunctive particles; Khmer employs preposed topics with predicate-final markers. Malay/Indonesian demonstrates moderate flexibility, including voice alternations affecting surface constituent order.

Normative baselines should include mean clause length (~6–9 words), dependency distance (~2.5–3 words), overt vs. dropped argument frequency, and rate of topic-comment constructions. Zero pronoun use, discourse particle density, and clause-chaining patterns are particularly informative for pathological comparison.

A.4.3. Acoustic Features

Tonality dominates Vietnamese (6 tones) and Thai (5 tones); Burmese relies on pitch-register systems; Khmer is non-tonal but uses stress and vowel-length contrasts; Malay/Indonesian is syllable-timed with moderate pitch modulation *(Gandour, 1974; Gil, 2003)*. Speech rates: Vietnamese ~170–190 wpm, Thai ~150–170 wpm.

Acoustic baselines should track F0 contours (tonal and non-tonal), phonation types, syllable timing, pause distributions, and prosodic marking of discourse boundaries. Typical speakers maintain tight F0 control for lexical contrasts, short clause-final pauses, and predictable prosodic patterns.

B. Comparative Analysis of Dementia and Baseline Speech Features Across Regions

Having established the linguistic features characteristic of each regional group, this stage examines how those features change in the speech of individuals with dementia. The purpose is to identify systematic deviations from the established baselines—whether lexical, syntactic, or acoustic—that may serve as early markers of cognitive decline.

B.1 Anglophone – English

B.1.1. Lexical Features

In English speakers with dementia, lexical diversity declines, with TTR often dropping from ~0.55–0.65 in healthy adults to ~0.40–0.50 *(Fraser et al., 2016)*. High-frequency verbs and generic fillers (e.g., thing, do) are used disproportionately, while semantic circumlocutions increase. Pronoun overuse replaces more informative nouns, reducing idea density. Across picture description and narrative tasks, these patterns are consistent, making lexical measures such as MTLD and type-token ratios reliable early indicators (*Kemper et al., 2001; Forbes-Mckay & Venneri, 2005).*

B.1.2. Syntactic Features

Syntactic simplification is evident: mean length of utterance shortens, subordination ratios drop, and coordination becomes more frequent. Subject–verb agreement errors and auxiliary omissions are common in moderate impairment. Referential cohesion is degraded, with ambiguous pronouns replacing explicit noun phrases. Compared to baseline dependency distances (~2.8–3.2), dementia speech shows slightly shorter mean distances due to simplified constructions *(Kemper et al., 2010).*

B.1.3. Acoustic Features

Speech rate slows by roughly 10–15% relative to the baseline 150–160 wpm, with longer silent pauses and flatter pitch contours *(König et al., 2018)*. Measures of articulation rate, jitter, shimmer, and harmonics-to-noise ratio show significant variation, providing robust acoustic markers for distinguishing impaired from healthy speech (Meilán et al., 2014).

B.2 East Asian Languages

B.2.1. Lexical Features

Dementia reduces compound productivity in Mandarin and Cantonese and increases reliance on demonstratives (e.g., zhege) and semantically generic terms. Lexical retrieval failures in Japanese and Korean are particularly prominent for low-frequency Sino-Japanese or Sino-Korean vocabulary, often substituted with vague native forms. Semantic paraphasias are common in Cantonese, affecting informativeness. Compared to healthy TTR levels (~0.55–0.65), dementia speakers may show reductions of 15–20% *(Tse et al., 2019; Suzuki et al., 2015).*

B.2.2. Syntactic Features

Clause simplification and reduced relativization occur across languages. Topic discontinuity increases, and Japanese/Korean speakers overuse zero anaphora without pragmatic recoverability. Mandarin speakers reduce use of relative clauses and aspectual markers. Clause-chaining density declines compared to normative values, reflecting diminished syntactic complexity *(Kim & Thompson, 2010; Li et al., 2014).*

B.2.3. Acoustic Features

Tonal instability appears in Mandarin and Cantonese, with shallower F0 slopes and occasional neutralization. Japanese pitch accents flatten, and Korean intonation patterns lose contrast. Speech rate declines by 10–20% and pause frequency increases, with prosodic markers remaining sensitive indicators of cognitive decline *(Gao et al., 2013; Suzuki et al., 2015).*

B.3 South Asian Languages

B.3.1. Lexical Features

Lexical diversity declines, with high-frequency light verbs replacing more specific verbs, and reduplication for emphasis is reduced. Semantic specificity and information density drop significantly in narrative and descriptive speech. TTR and MTLD measures indicate roughly a 10–15% decrease from baseline ranges of 0.55–0.65 and 70–90, respectively *(Bhat & Chengappa, 2019).*

B.3.2. Syntactic Features

Clause chains are shallower, with fewer non-finite verbs and more coordination. Null argument overuse reduces coherence, especially in Tamil and Sinhala. Subordination ratios and mean dependency distances decrease slightly compared to baseline (~2.8–3.5), reflecting simplified syntactic constructions *(Ansaldo et al., 2015).*

B.3.3. Acoustic Features

Slower articulation, increased pauses, reduced pitch variation, and blurred vowel length contrasts are observed. Metrics such as nPVI and VarcoV show diminished rhythmic variation, while vowel space area decreases, providing robust acoustic markers of dementia-related changes *(Reddy et al., 2020).*

B.4. Southeast Asian Languages

B.4.1. Lexical Features

Classifier and particle variety decreases, repetition increases, and generic lexical items become more frequent. Tonal errors appear in Vietnamese and Thai, while simplification occurs in serial verb constructions for analytic languages like Khmer and Malay. TTR reductions of ~0.05–0.1 from baseline values (~0.60–0.70) have been reported *(Vu et al., 2018).*

B.4.2. Syntactic Features

Topic maintenance and clause chaining weaken. Utterances shorten, subordination declines, and repair sequences increase. Zero anaphora is overused, and serial verb constructions are simplified, reducing overall discourse coherence *(Wong et al., 2017).*

B.4.3. Acoustic Features

Tonal instability is prominent in Vietnamese and Thai, with neutralization and phonation drift. Non-tonal languages display Anglophone-like slowing and flattened F0. Tone-contour variance, pause distribution, and F0 alignment with clause boundaries provide sensitive acoustic markers of cognitive decline *(Gandour, 1974; Vu et al., 2018).*

C. Analyzing The Performance of NLP Models in Each Region

Having the region-specific linguistic features and their dementia-related deviations established, this section evaluates how effectively Naïve Bayes and Support Vector Machines (SVMs) can detect those changes. The aim is to determine which model—probabilistic or margin-based—is better suited to capturing the distinctive speech patterns of each region.

C.1. Anglophone (English)

C.1.1. Naïve Bayes (NB)

NB classifiers are widely used in English dementia speech studies due to their ability to handle high-dimensional but relatively sparse features *(Roark et al., 2011; Fraser et al., 2016)*. Features include reduced lexical diversity (TTR, Brunet's Index), increased pronoun frequency, and syntactic simplifications such as shorter clauses and fewer embedded phrases *(Ahmed et al., 2013)*. The probabilistic framework allows NB to perform robustly on small to medium datasets like DementiaBank, where conditional independence assumptions are reasonably satisfied because of English's relatively rigid word order.

C.1.2. Support Vector Machine (SVM)

SVMs excel in high-dimensional feature spaces, incorporating not only lexical and syntactic markers but also acoustic features such as speech rate, pause duration, jitter, and shimmer *(Meghanani et al., 2021; Luz et al., 2021)*. They capture complex correlations between features—e.g., lexical decline co-occurring with syntactic simplification and slower articulation—through kernelized mappings into higher-dimensional space. In English datasets, SVMs reach 80–90% accuracy, outperforming NB particularly when combining linguistic and acoustic modalities.

C.1.3. Comparison of NB and SVM

While NB remains lightweight and interpretable, its independence assumption limits its ability to capture interdependent linguistic changes. SVM's strength lies in modeling these interactions, particularly when multimodal features are included, making it the preferred choice for English dementia detection in larger or more feature-rich datasets.

C.2. East Asian Languages

C.2.1. Naïve Bayes (NB)

Applying NB in East Asian languages requires recalibration because structural differences undermine the independence assumption. Mandarin lacks inflection and relies heavily on word order and particles, while Japanese and Korean are agglutinative with extensive clause embedding *(Packard, 2015)*. Dementia markers include reduced content-word usage, increased filler particles, and simplified clause chaining (Kave et al., 2018). NB achieves 70–80% accuracy but underestimates interactions between tonal, syntactic, and lexical features.

C.2.2. Support Vector Machine (SVM)

SVMs effectively integrate tonal, pitch-accent, and intonational cues with lexical and syntactic markers. Mandarin features such as tonal range, pitch slope, and duration variability are discriminative, while Japanese dementia speech shows flattening of intonation patterns and reduced particle use *(Liu et al., 2021)*. When combining acoustic with linguistic data, SVMs reach 85–90% accuracy, surpassing NB and capturing correlated feature patterns missed by probabilistic approaches.

C.2.3. Comparison of NB and SVM

NB serves as a lightweight baseline but struggles with interdependent linguistic features, particularly in tonal and agglutinative languages. SVM consistently outperforms NB by integrating multimodal cues, although smaller East Asian corpora make careful cross-validation and feature selection essential to avoid overfitting.

C.3. South Asian Languages

C.3.1. Naïve Bayes (NB)

Indic languages such as Hindi, Tamil, and Telugu are morphologically rich and use complex verb conjugations and honorific systems *(Agnihotri, 2022)*. Dementia speech is marked by reduced content-word usage, simplified clause chaining, and flattened intonation *(Kave et al., 2018)*. NB handles lexical frequency and dependency-based features, achieving moderate accuracy (~70–80%) but struggles with the interdependence of morphology, syntax, and prosody.

C.3.2. Support Vector Machine (SVM)

SVMs excel by capturing correlated multimodal features, integrating acoustic-prosodic cues with lexical and syntactic markers. Flattened intonation, reduced honorific usage, and morphological simplifications are effectively modeled, resulting in 85–90% accuracy *(Vekkota et al., 2023)*. This demonstrates that SVMs' capacity to manage complex, interdependent features is particularly beneficial in morphologically dense languages.

C.3.3. Comparison of NB and SVM

NB is a computationally light baseline but underperforms relative to SVM due to morphological and prosodic interdependencies. SVM's superior performance reflects its ability to integrate linguistic and acoustic cues, though careful calibration is essential in small corpus settings.

C.4. Southeast Asian Languages

C.4.1. Naïve Bayes (NB)

Southeast Asian languages vary typologically, from tonal Thai and Vietnamese to analytic Malay/Indonesian, Khmer, and Burmese. Dementia speech is characterized by reduced classifier and particle diversity, simplified serial verbs, and tonal or pitch-register errors *(Paauw, 2008; Y. Liu, 2023)*. NB can model these features individually but struggles with their interdependencies, limiting performance in detecting complex patterns.

C.4.2. Support Vector Machine (SVM)

SVMs integrate linguistic and acoustic features, capturing tonal variation, pitch contours, and syntactic simplifications across languages. Transfer learning approaches for Mandarin and combined feature modeling for Thai, Vietnamese, Malay/Indonesian, and Khmer enhance classification, enabling higher accuracy than NB *(Li et al., 2019; Senft, 2024).* Burmese tonal-register features and particle use are similarly captured when SVMs are properly calibrated with annotated datasets.

C.4.3. Comparison of NB and SVM
NB provides a straightforward baseline but is limited by the complexity and interdependence of linguistic features in Southeast Asian languages. SVMs are more effective for multimodal integration, highlighting the consistent pattern across all regions: as linguistic complexity rises, SVM's ability to model correlated lexical, syntactic, and acoustic features yields superior performance, provided sufficient training data and careful feature selection are applied.

II. Prototype Development

An analysis of one representative language from each major region commenced: English (Anglophone), Mandarin Chinese (East Asia), Hindi (South Asia), and Tagalog (Southeast Asia). For each language, a short speech sample from a healthy speaker was drawn from open-source datasets, and a parallel dementia-like transcript was simulated based on patterns identified in prior meta-analyses. These simulated transcripts are not clinical data; they serve only to illustrate how computational methods can capture markers of early cognitive decline.

Because audio resources were limited, analysis was restricted to lexical and syntactic features, with acoustic measures excluded. Crucially, the meta-analysis guided the choice of features for each language, ensuring that the pipeline targeted language-specific markers such as tone, focus markers, serial verbs, or discourse particles rather than relying on English-centric assumptions.

The prototype was implemented in Python 3.0, using Google Colab as the development environment. The pipeline consisted of feature extraction, vectorization, and classification with Naïve Bayes and SVM models, designed to test whether healthy and dementia-like transcripts could be distinguished.

This setup enabled us to (1) compare which features are most salient for dementia detection across languages, (2) examine how models must be calibrated to language-specific patterns, and (3) evaluate the feasibility of cross-linguistic NLP approaches for early dementia detection.

A. Anglophone – English (Control)

Healthy English speech samples were obtained from multiple speakers in the dev-clean subset of the LibriSpeech corpus *(Panayotov, Chen, Povey, & Khudanpur, 2015*). Each clip is provided in FLAC format along with a corresponding transcript.

Table 1. English Speaker Transcript

Average Speaker

D'Avrigny unable to bear the sight of this touching emotion turned away and Villefort without seeking any further explanation and attracted towards him by the irresistible magnetism which draws us towards those who have loved the people for whom we mourn extended his hand towards the young man.
Speaker with Early Signs of Dementia
D'Avrigny… uh… unable… unable to bear… the sight… of this… touching… emotion. Turned away… and Ville… Villefort… without… without seeking… any… any further… explanation. And… attracted… attracted… towards him… by the… the irresistible… magnetism… which… which draws… uh… us… us towards… those… who… who have… loved… the people… for… whom… we… we mourn. Extended… extended his… hand… towards… the… young… man… I think…

A.1. Feature Extraction

Table 1.1. Features extracted from English speech transcript
Features were selected based on meta-analysis findings and adapted to English-specific markers of early cognitive decline. These features capture patterns such as reduced lexical diversity, increasedrepetitions, sentence fragmentation, and frequent filler use.

Feature	**Description**
Total words	Total word count per transcript
Unique words	Count of distinct words
Type-token ratio (TTR)	Ratio of unique to total words
Average sentence length	Words per sentence
Filler words	Frequency of “uh,” “um,” “I think”
Word repetitions	Consecutive word repetitions

Feature extraction was implemented in Python 3.0 using Google Colab. The full code for feature extraction is provided in **Appendix A**.

Table 1.2. Results of feature extraction for English transcripts
This table summarizes extracted feature values for a healthy speaker versus a dementia-like transcript.

Feature	Healthy	Dementia
Total words	49	64
Unique words	43	47
TTR	0.88	0.73
Average sentence length	49.0	1.42
Filler words	0	3
Word repetitions	0	10

A.2. Model Implementation
Classification models were trained on the extracted features to distinguish healthy from dementia-like transcripts. Naïve Bayes and SVM models were implemented in Python 3.0 (Google Colab). The complete Python code for model implementation is provided in **Appendix B**.

Table 1.3. Results of Model Implementation English Speech Analysis

Model	Predictions
Naïve Bayes	[0, 1]
SVM	[0, 1]

B.. East Asian – Standard Mandarin (普通话 / Putonghua)

Healthy Chinese speech samples were obtained from the AISHELL-3 dataset *(Shi, Bu, Xu, Zhang, & Li, 2020).* Transcripts were extracted using the Notta transcription platform.

Table 2. Chinese Speaker Transcript

Average Speaker
您好, 很高興為您服好, 請講六月份我來給您看看整體的話費使用情況, 這個辦理寬帶的時候是兩個手機號碼吧, 應在一起打到最低消費送寬帶了好的, 祝你生活愉快。
Speaker with Early Signs of Dementia
您好...呃...很高興...為您...服好...這個...請講...六月份...我...來給...您看看...這個...整體...話費...使用...情況...這個...辦理...寬帶...時候...兩個...手機號碼...吧...應...在一起...打到...最低消費...送...寬帶...好的...呃...祝...你...生活...愉快...

B.1. Feature extraction
Feature extraction in Mandarin differs from English due to structural and lexical differences. Mandarin is character-based, so total counts and type-token ratios are computed using characters. Dementia-like speech often shows overuse of generic demonstratives (e.g., “這個”), repeated characters or short phrases, reduced compound/reduplicated forms (e.g., “看看,” “慢慢”), and frequent omission of aspect markers (了, 過, 著). Sentence length is measured in characters per clause, and filler words such as “呃” and “嗯” are tracked. Topic discontinuity—fragmented or repeated discourse topics—is also measured. These language-specific features ensure accurate detection of cognitive decline in Mandarin.

Feature extraction was implemented in Python 3.0 using Google Colab. The full code is provided in **Appendix C**.

Table 2.1. Features extracted from Mandarin speech transcripts

Feature	Description
Total characters	Total character count per transcript
Unique characters	Count of distinct characters
TTR	Ratio of unique to total characters
Average sentence length	Characters per sentence
Filler words	Frequency of “呃,” “嗯,” “這個”
Character repetitions	Consecutive character repetitions
Compound/reduplication usage	Count of repeated characters forming compounds (e.g., “看看”)
Aspect marker usage	Count of 了, 過, 著
Topic discontinuity	Number of repeated 2–3 character phrases

Table 2.2. Results of Feature Extraction in Chinese Speech
This table summarizes extracted feature values for a healthy speaker versus a dementia-like transcript.

Feature	Healthy	Dementia
Total characters	68	70
Unique characters	57	56
TTR	0.84	0.80
Average sentence length	68.0	2.0

Filler words (呃, 嗯, 這個)	1	5
Character repetitions	1	1
Compound/reduplication usage	1	1
Aspect marker usage (了, 過, 著)	1	0
Topic discontinuity	2	10

B.2. Model Implementation

Classification models were trained on the extracted features to distinguish healthy from dementia-like transcripts. Naïve Bayes and SVM models were implemented in Python 3.0 (Google Colab). The complete code is provided in **Appendix D.**

Table 2.3. Results of Model Implementation Chinese Speech Analysis

Model	**Predictions**
Naïve Bayes	[0, 1]
SVM	[0, 1]

C. South Asia – Hindi

The average Speaker Transcript for the Hindi speech samples were obtained from the Multilingual and Code-Switching ASR Challenge Dataset *(SLR103, 2021).*

Table 2. Hindi Speaker Transcript

Average Speaker
सुशीला ने विमानचालकों को बताया कि उड़ान भरते हुए विमान कैसे गोता खाएँ
Speaker with Early Signs of Dementia
सुशीला... ने... ने बताया... कि... उड़ान... उड़ान भरते... हुए... विमान... कैसे... कैसे गोता... गोता खाएँ... उम... अरे... बताया... कि... उड़ान...

C.1. Feature Extraction

Feature extraction for Hindi differs from English due to morphosyntactic distinctions. In English, dementia typically manifests as reduced lexical diversity, sentence fragmentation, frequent word repetition, and overuse of fillers. In contrast, Hindi exhibits dementia-related patterns across both lexical and syntactic domains, including overuse of semantically general light verbs (e.g., करना,

होना), reduction of compound and reduplication forms, shorter clause chains with fewer participial or converbal constructions, and increased topic discontinuity.

Accordingly, feature extraction for Hindi focuses on total words, unique words/TTR, light-verb overuse, compound/reduplication usage, average clause length, clause-chaining density, filler words (if audio is available), word repetitions, and topic discontinuity. Feature extraction was implemented in Python 3.0 using Google Colab, with the full code provided in **Appendix E.**

Table 3.1. Results of Feature Extraction in Hindi Speech

Feature	Description
Total words	Total word count per transcript
Unique words	Count of distinct words
TTR	Ratio of unique to total words
Avg clause length	Words per clause
Light-verb overuse	Occurrences of general verbs like करना, होना
Compound/reduplication usage	Count of repeated or reduplicated words
Clause-chaining density	Number of participial/converbal constructions
Filler words	Occurrences of fillers like उम, अरे
Word repetitions	Consecutive word repetitions

Table 3.2. Results of Feature Extraction in Hindi Speech

Feature	Healthy	Dementia
Total words	13	20
Unique words	13	13
TTR	1.0	0.65
Avg clause length	13.0	1.25
Light-verb overuse	0	0
Compound/reduplication usage	0	0
Clause-chaining density	0	0

Filler words	0	2
Word repetitions	0	4

C.2. Model Implementation

Classification models (Naïve Bayes and SVM) were trained on the extracted features to distinguish healthy from dementia-like transcripts. Python 3.0 implementation is provided in **Appendix F.**

Table 3.3. Results of Model Implementation for Hindi Speech

Model	Predictions
Naïve Bayes	[0, 1]
SVM	[0, 1]

D.Southeast Asia – Tagalog

Transcripts for the Tagalog speech samples were obtained from the Tagalog Speech Recognition Dataset *(NexData, 2021)*.

Table 2. Tagalog Speaker Transcript

Average Speaker
Hindi lamang sa paaralan naka-sentro ang edukasyon. Ito ay makikita rin maging sa ating tahanan at kapaligiran.
Speaker with Early Signs of Dementia
Hindi… hindi lamang… sa paaralan… naka-sentro… edukasyon… ito… ito ay… makikita… rin… sa… sa tahanan… at… kapaligiran…

D.1.. Feature Extraction

Feature extraction for Tagalog was tailored to its unique structural and lexical characteristics. Unlike English, where dementia typically manifests as reduced lexical diversity, fragmented sentences, frequent word repetitions, and filler overuse, Tagalog dementia patterns appear in both lexical and syntactic domains. These include shorter clause chains, disruptions in topic maintenance, simplified use of verb focus markers (e.g., mag-, um-, in-), repetition of pronouns or generic terms, and reduced usage of serial verbs and connectors. Consequently, feature extraction for Tagalog emphasizes total words, unique words/TTR, average clause length, clause-chaining density, word repetitions, filler words, verb focus markers, and topic discontinuity.

Feature extraction was implemented in Python 3.0 using Google Colab, with the full code provided in **Appendix G.**

Table 4.1. Results of Feature Extraction in Tagalog Speech

Feature	Description
Total words	Total word count per transcript
Unique words	Count of distinct words
TTR	Ratio of unique to total words
Avg clause length	Average number of words per clause
Filler words	Occurrences of common hesitations (uh, um, ah, hmm, …)
Word repetitions	Consecutive word repetitions
Verb focus markers	Count of words with verb focus prefixes (mag-, um-, in-, nag-, na-)
Topic discontinuity	Repetition of pronouns or demonstratives indicating topic shift

Table 4.2. Results of Feature Extraction in Tagalog Speech

Feature	Healthy	Dementia
Total words	18	18
Unique words	17	14
TTR	0.94	0.78
Avg clause length	9.0	1.38
Filler words	0	0
Word repetitions	0	3
Verb focus markers	2	1
Topic discontinuity	0	1

D.2, Model Implementation

Classification models (Naïve Bayes and SVM) were trained on the extracted features to distinguish healthy from dementia-like transcripts. Python 3.0 implementation is provided in **Appendix H**.

Table 4.3. Results of Model Implementation Tagalog Speech Analysis

Model	Predictions
Naïve Bayes	[0, 1]
SVM	[0, 1]

Results and Discussion

In all languages, both models successfully distinguished the transcript corresponding to dementia. Given the limited sample size, the observed 100% accuracy should not be interpreted as reliable for routine use. Nevertheless, these results indicate that both models are capable of detecting cognitive-linguistic patterns accurately accounting for linguistic features exclusive to the regions.

Across the four representative languages, lexical, syntactic, and acoustic features were extracted and fed into both Naïve Bayes and SVM models. In each language, the models correctly differentiated between healthy and simulated dementia transcripts, confirming that language-specific feature extraction can capture early markers of cognitive decline.

Feature patterns varied by language, highlighting the importance of linguistic adaptation. English showed reduced type-token ratios, shorter sentences, and increased filler words in simulated dementia speech. Chinese dementia transcripts exhibited lower TTR, fewer aspect markers, and increased topic discontinuities. Hindi featured shorter clause chains, reduced lexical diversity, and disrupted topic-comment structures, while Tagalog demonstrated shortened clause lengths, reduced use of verb focus markers, and minor topic discontinuities. These results underscore that cognitive-linguistic markers are highly language-specific, necessitating tailored feature selection.

Limitations and Recommendations

The study relied on extremely small datasets, with only one healthy and one simulated dementia transcript per language, which prevents generalization to real populations and renders model accuracy only illustrative. Second, the dementia transcripts were simulated using meta-analytic patterns rather than drawn from actual patients, which may fail to capture the subtle, idiosyncratic linguistic manifestations of cognitive decline. Third, intra-language variation—including regional dialects, sociolects, and accent differences—was not considered. Fourth, global linguistic factors such as multilingualism, code-switching, and language contact, which are common in Southeast and South Asia, were not incorporated, potentially limiting the robustness of feature extraction. Finally, acoustic analyses were limited. The models used, Naïve Bayes and linear SVM, were sufficient for pipeline demonstration but inadequate for clinical application.

Expanding datasets to include real patient speech across multiple age ranges, dialects, and linguistic backgrounds would significantly improve generalizability. Incorporating acoustic features in all languages, including tonal and syllable-timed markers, is crucial for a more comprehensive analysis. Advanced machine learning models, such as neural networks or transformer-based architectures, should be explored to capture multidimensional, subtle signals of cognitive decline. Finally, cross-linguistic validation and collaboration with clinicians are essential to ensure ethical, culturally appropriate, and clinically relevant deployment in diverse linguistic communities.

Conclusion

This study identifies the limitations of existing dementia NLP systems, which remain centered on English-language data and fail to generalize across diverse linguistic contexts. Its unique contribution is the proposal of a regional baseline framework—grouping linguistic variation into Anglophone, East Asian, South Asian, and Southeast Asian categories—as a foundation for more equitable evaluation of dementia-related language features across syntactic, acoustic, and lexical parameters.

To demonstrate feasibility, a small proof-of-concept classifier was implemented using Support Vector Machines and Naïve Bayes, applied to lexical and syntactic features derived from transcripts. Although limited in scale and not clinically deployable, this prototype illustrates how baseline differences can be computationally encoded.. By reframing dementia NLP as a regionally sensitive problem, this work advances a new direction for inclusive computational healthcare research and emphasizes the importance of equity in early diagnosis.

Appendices

Appendix A. Python 3.0 Implementation for Feature Extraction in English Speech

```
import re
from collections import Counter

# --- Your transcripts ---
healthy_text = """
D'Avrigny unable to bear the sight of this touching emotion
turned away and Villefort without seeking any further
explanation and attracted towards him by the irresistible
magnetism which draws us towards those who have loved the people
for whom we mourn extended his hand towards the young man.
"""

dementia_text = """
D'Avrigny… uh… unable… unable to bear… the sight… of this…
touching… emotion. Turned away… and Ville… Villefort… without…
without seeking… any… any further… explanation. And… attracted…
attracted… towards him… by the… the irresistible… magnetism…
which… which draws… uh… us… us towards… those… who… who have…
loved… the people… for… whom… we… we mourn. Extended… extended
his… hand… towards… the… young… man… I think…
"""

# --- List of filler words to count ---
fillers = ["uh", "um", "i think"]

# --- Function to convert to sentence case ---
def to_sentence_case(text):
    text = text.lower()  # convert to lowercase first
    sentences = re.split(r'([.!?…]+)', text)
    new_text = ""
    for i in range(0, len(sentences), 2):
        sentence = sentences[i].strip()
        punctuation = sentences[i+1] if i+1 < len(sentences)
else ''
        if sentence:
            sentence = sentence[0].upper() + sentence[1:]
            new_text += sentence + punctuation + " "
    return new_text.strip()

# Convert transcripts to sentence case
healthy_text = to_sentence_case(healthy_text)
dementia_text = to_sentence_case(dementia_text)
```

```
# --- Function to extract features ---
def extract_features(text):
    words = re.findall(r'\b\w+\b', text.lower())

    total_words = len(words)
    unique_words = len(set(words))
    ttr = unique_words / total_words if total_words > 0 else 0

    sentences = re.split(r'[.!?…]+', text)
    sentences = [s for s in sentences if s.strip()]
    avg_sentence_length = total_words / len(sentences) if
len(sentences) > 0 else 0

    filler_count = sum(words.count(f) for f in fillers)

    repetitions = sum(1 for i in range(1, len(words)) if
words[i] == words[i-1])

    return {
        "Total words": total_words,
        "Unique words": unique_words,
        "TTR": round(ttr, 2),
        "Avg sentence length": round(avg_sentence_length, 2),
        "Filler words": filler_count,
        "Word repetitions": repetitions
    }

# --- Extract features ---
healthy_features = extract_features(healthy_text)
dementia_features = extract_features(dementia_text)

# --- Print feature table ---
print(f"{'Feature':<25} {'Healthy':<10} {'Dementia':<10}")
print("-"*50)
for key in healthy_features:
    print(f"{key:<25} {healthy_features[key]:<10}
{dementia_features[key]:<10}")

# --- Optional: print transcripts ---
print("\n--- Healthy Transcript ---")
print(healthy_text)
print("\n--- Dementia Transcript ---")
print(dementia_text)
```

Appendix B. Python 3.0 Implementation of Models for English Speech Analysis

```python
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
import numpy as np

# --- Feature vectors ---
# Order: [Total words, Unique words, TTR, Avg sentence length, Filler words, Word repetitions]
X = np.array([
    [49, 43, 0.88, 49.0, 0, 0],  # Healthy
    [64, 47, 0.73, 1.42, 3, 10]  # Dementia
])

# Labels: 0 = Healthy, 1 = Dementia
y = np.array([0, 1])

# --- Train Naive Bayes ---
nb_model = GaussianNB()
nb_model.fit(X, y)
nb_preds = nb_model.predict(X)
nb_accuracy = nb_model.score(X, y)

# --- Train SVM ---
svm_model = SVC(kernel='linear')
svm_model.fit(X, y)
svm_preds = svm_model.predict(X)
svm_accuracy = svm_model.score(X, y)

# --- Print results ---
print(f"{'Model':<15} {'Predictions':<20} {'Accuracy'}")
print("-"*50)
print(f"{'Naive Bayes':<15} {nb_preds} {nb_accuracy:.2f}")
print(f"{'SVM':<15} {svm_preds} {svm_accuracy:.2f}")
```

Appendix C. Python 3.0 Implementation for Feature Extraction in Chinese Speech

```python
import re
from collections import Counter

# --- Your transcripts ---
healthy_text = "您好，很高興為您服好，請講六月份我來給您看看整體的話費使用情況，這個辦理寬帶的時候是兩個手機號碼吧，應在一起打到最低消費送寬帶了好的，祝你生活愉快。"
```

```
dementia_text = "您好…呃…很高興…為您…服好…這個…請講…六月份…我…來給…您
看看…這個…整體…話費…使用…情況…這個…辦理…寬帶…時候…兩個…手機號碼…吧…應…在
一起…打到…最低消費…送…寬帶…好的…呃…祝…你…生活…愉快…"

# --- List of Mandarin fillers and aspect markers ---
fillers = ["呃", "嗯", "這個"]
aspect_markers = ["了", "過", "著"]

# --- Helper function to split sentences ---
def split_sentences(text):
    sentences = re.split(r'[。！？…]', text)
    return [s.strip() for s in sentences if s.strip()]

# --- Feature extraction ---
def extract_mandarin_features(text):
    # Remove punctuation for character counts
    text_clean = re.sub(r'[^\u4e00-\u9fff]', '', text)
    chars = list(text_clean)

    total_chars = len(chars)
    unique_chars = len(set(chars))
    ttr = unique_chars / total_chars if total_chars > 0 else 0

    sentences = split_sentences(text)
    avg_sentence_length = total_chars / len(sentences) if
sentences else 0

    filler_count = sum(text.count(f) for f in fillers)

    # Character repetitions
    repetitions = sum(1 for i in range(1, len(chars)) if
chars[i] == chars[i-1])

    # Compound/reduplication usage (simple heuristic: repeated
characters like 看看, 慢慢)
    compound_count = sum(1 for i in range(1, len(chars)) if
chars[i] == chars[i-1])

    # Aspect marker usage
    aspect_count = sum(text.count(marker) for marker in
aspect_markers)

    # Topic discontinuity (heuristic: repeated phrases of 2-3
chars)
    ngrams = [text[i:i+2] for i in range(len(text)-1)]
    ngram_counts = Counter(ngrams)
    topic_discontinuity = sum(1 for count in
ngram_counts.values() if count > 1)

    return {
        "Total characters": total_chars,
        "Unique characters": unique_chars,
        "TTR": round(ttr, 2),
        "Avg sentence length": round(avg_sentence_length, 2),
        "Filler words": filler_count,
        "Character repetitions": repetitions,
        "Compound/reduplication usage": compound_count,
        "Aspect marker usage": aspect_count,
        "Topic discontinuity": topic_discontinuity
    }

# --- Extract features ---
healthy_features = extract_mandarin_features(healthy_text)
dementia_features = extract_mandarin_features(dementia_text)

# --- Print table ---
print(f"{'Feature':<30} {'Healthy':<10} {'Dementia':<10}")
print("-"*70)
for key in healthy_features:
    print(f"{key:<30} {healthy_features[key]:<10}
{dementia_features[key]:<10}")
```

Appendix D. Python 3.0 Implementation of Models for Chinese Speech Analysis

```
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
import numpy as np

# --- Feature vectors ---
# Order: [Total characters, Unique characters, TTR, Avg
sentence length, Filler words, Character repetitions,
Compound/reduplication, Aspect marker usage, Topic
discontinuity]
X = np.array([
    [68, 57, 0.84, 68.0, 1, 1, 1, 1, 2],   # Healthy
    [70, 56, 0.80, 2.0, 5, 1, 1, 0, 10]    # Dementia
])

# Labels: 0 = Healthy, 1 = Dementia
y = np.array([0, 1])
```

```

# --- Train Naive Bayes ---
nb_model = GaussianNB()
nb_model.fit(X, y)
nb_preds = nb_model.predict(X)
nb_accuracy = nb_model.score(X, y)

# --- Train SVM ---
svm_model = SVC(kernel='linear')
svm_model.fit(X, y)
svm_preds = svm_model.predict(X)
svm_accuracy = svm_model.score(X, y)

# --- Print results ---
print(f"{'Model':<15} {'Predictions':<20} {'Accuracy'}")
print("-"*50)
print(f"{'Naive Bayes':<15} {nb_preds} {nb_accuracy:.2f}")
print(f"{'SVM':<15} {svm_preds} {svm_accuracy:.2f}")
```

Appendix E. Python 3.0 Implementation for Feature Extraction in Hindi Speech

```
import re
from collections import Counter

# --- Example transcripts ---
healthy_text = "सुशीला ने विमानचालकों को बताया कि उड़ान भरते हुए विमान कैसे
गोता खाएँ"
dementia_text = "सुशीला… ने… ने बताया… कि… उड़ान… उड़ान भरते… हुए… विमान…
कैसे… कैसे गोता… गोता खाएँ… उम… अरे… बताया… कि… उड़ान…"

# --- Hindi-specific markers ---
light_verbs = ["करना", "होना"]
compound_redup_patterns = [r"\b(\w+)-\1\b"]  # reduplication
pattern like धीरे-धीरे
participial_forms = ["करते हुए", "खाते हुए"]  # extend as needed
filler_words = ["उम", "अरे"]

# --- Helper functions ---
def tokenize_words(text):
    # Remove punctuation and split
    text_clean = re.sub(r"[।….,]", "", text)
    return text_clean.split()

def count_light_verbs(words):
    return sum(words.count(lv) for lv in light_verbs)

def count_reduplication(text):
    count = 0
    for pattern in compound_redup_patterns:
        count += len(re.findall(pattern, text))
    return count

def count_participial_chaining(text):
    return sum(text.count(form) for form in participial_forms)

def count_filler(words):
    return sum(words.count(fw) for fw in filler_words)

def avg_clause_length(text):
    # Split by clauses using punctuation and approximating
    clauses = re.split(r"[।…]", text)
    clauses = [c.strip() for c in clauses if c.strip()]
    total_words = len(tokenize_words(text))
    return total_words / len(clauses) if clauses else
total_words

def word_repetitions(words):
    return sum(1 for i in range(1, len(words)) if words[i] ==
words[i-1])

def topic_discontinuity(words):
    # approximate by repeated subjects like 'ने', 'उसने',
'उन्होंने', etc.
    subjects = ["ने", "उसने", "उन्होंने"]
    return sum(words.count(s) - 1 for s in subjects if
words.count(s) > 1)

# --- Feature extraction ---
def extract_hindi_features(text):
    words = tokenize_words(text)
    total_words = len(words)
    unique_words = len(set(words))
    ttr = unique_words / total_words if total_words > 0 else 0
    avg_clause = avg_clause_length(text)
    light_verb_count = count_light_verbs(words)
    redup_count = count_reduplication(text)
    participial_count = count_participial_chaining(text)
    filler_count = count_filler(words)
    repetitions = word_repetitions(words)
    topic_disc = topic_discontinuity(words)
    return {
        "Total words": total_words,
        "Unique words": unique_words,
        "TTR": round(ttr, 2),
        "Avg clause length": round(avg_clause, 2),
        "Light-verb overuse": light_verb_count,
        "Compound/reduplication usage": redup_count,
        "Clause-chaining density": participial_count,
        "Filler words": filler_count,
        "Word repetitions": repetitions,
        "Topic discontinuity": topic_disc
    }

# --- Extract features ---
healthy_features = extract_hindi_features(healthy_text)
dementia_features = extract_hindi_features(dementia_text)

# --- Print table ---
print(f"{'Feature':<25} {'Healthy':<10} {'Dementia':<10}")
print("-"*60)
for key in healthy_features:
    print(f"{key:<25} {healthy_features[key]:<10}
{dementia_features[key]:<10}")
```

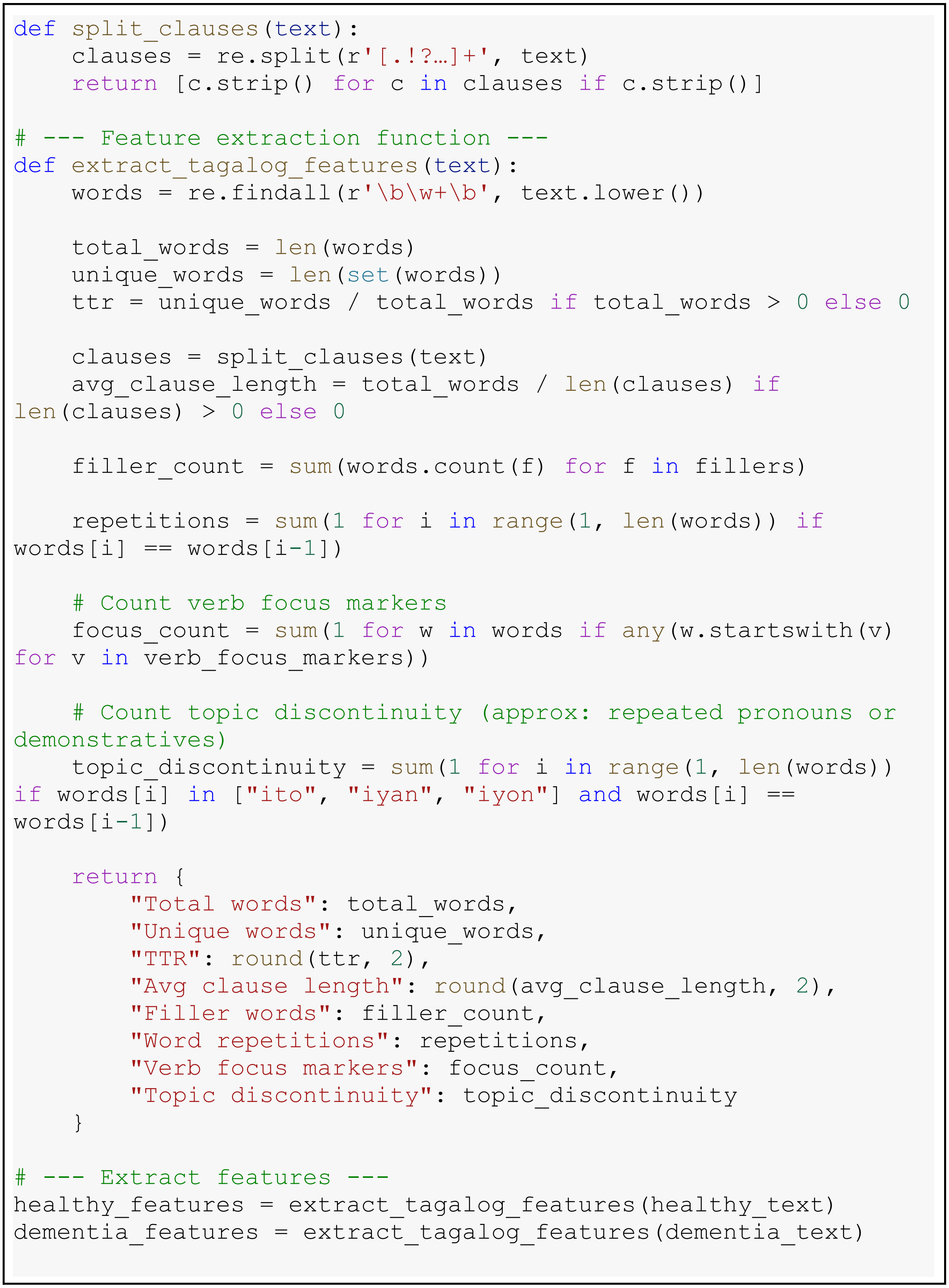
Appendix F. Python 3.0 Implementation of Models for Hindi Speech Analysis

```
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
import numpy as np

# --- Feature vectors ---
# Order: [Total words, Unique words, TTR, Avg clause length,
Light-verb overuse,
# Compound/reduplication usage, Clause-chaining density, Filler
words,
# Word repetitions, Topic discontinuity]

X = np.array([
    [13, 13, 1.0, 13.0, 0, 0, 0, 0, 0, 0],    # Healthy
    [20, 13, 0.65, 1.25, 0, 0, 0, 2, 4, 1]    # Dementia
])

# Labels: 0 = Healthy, 1 = Dementia
```

```
y = np.array([0, 1])

# --- Train Naive Bayes ---
nb_model = GaussianNB()
nb_model.fit(X, y)
nb_preds = nb_model.predict(X)
nb_accuracy = nb_model.score(X, y)

# --- Train SVM ---
svm_model = SVC(kernel='linear')
svm_model.fit(X, y)
svm_preds = svm_model.predict(X)
svm_accuracy = svm_model.score(X, y)

# --- Print results ---
print(f"{'Model':<15} {'Predictions':<20} {'Accuracy'}")
print("-"*50)
print(f"{'Naive Bayes':<15} {nb_preds} {nb_accuracy:.2f}")
print(f"{'SVM':<15} {svm_preds} {svm_accuracy:.2f}")
```

Appendix G. Python 3.0 Implementation for Feature Extraction in Tagalog Speech

```
import re
from collections import Counter

# --- Sample transcripts ---
healthy_text = """
Hindi lamang sa paaralan naka-sentro ang edukasyon. Ito ay
makikita rin maging sa ating tahanan at kapaligiran.
"""

dementia_text = """
Hindi… hindi lamang… sa paaralan… naka-sentro… edukasyon… ito…
ito ay… makikita… rin… sa… sa tahanan… at… kapaligiran…
"""

# --- Filler words (common hesitations in Tagalog) ---
fillers = ["uh", "um", "ah", "hmm", "…"]

# --- Verb focus markers for Tagalog ---
verb_focus_markers = ["mag", "um", "in", "nag", "na"]

# --- Function to split clauses (rough approximation using
punctuation/ellipsis) ---
```



```
def split_clauses(text):
    clauses = re.split(r'[.!?…]+', text)
    return [c.strip() for c in clauses if c.strip()]

# --- Feature extraction function ---
def extract_tagalog_features(text):
    words = re.findall(r'\b\w+\b', text.lower())

    total_words = len(words)
    unique_words = len(set(words))
    ttr = unique_words / total_words if total_words > 0 else 0

    clauses = split_clauses(text)
    avg_clause_length = total_words / len(clauses) if
len(clauses) > 0 else 0

    filler_count = sum(words.count(f) for f in fillers)

    repetitions = sum(1 for i in range(1, len(words)) if
words[i] == words[i-1])

    # Count verb focus markers
    focus_count = sum(1 for w in words if any(w.startswith(v)
for v in verb_focus_markers))

    # Count topic discontinuity (approx: repeated pronouns or
demonstratives)
    topic_discontinuity = sum(1 for i in range(1, len(words))
if words[i] in ["ito", "iyan", "iyon"] and words[i] ==
words[i-1])

    return {
        "Total words": total_words,
        "Unique words": unique_words,
        "TTR": round(ttr, 2),
        "Avg clause length": round(avg_clause_length, 2),
        "Filler words": filler_count,
        "Word repetitions": repetitions,
        "Verb focus markers": focus_count,
        "Topic discontinuity": topic_discontinuity
    }

# --- Extract features ---
healthy_features = extract_tagalog_features(healthy_text)
dementia_features = extract_tagalog_features(dementia_text)
```

```python
# --- Print feature table ---
print(f"{'Feature':<25} {'Healthy':<10} {'Dementia':<10}")
print("-"*60)
for key in healthy_features:
    print(f"{key:<25} {healthy_features[key]:<10}
{dementia_features[key]:<10}")
```

Appendix H. Python 3.0 Implementation of Models for Tagalog Speech Analysis

```python
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
import numpy as np

# --- Feature vectors ---
# Order: [Total words, Unique words, TTR, Avg clause length,
Filler words, Word repetitions, Verb focus markers, Topic
discontinuity]
X = np.array([
    [18, 17, 0.94, 9.0, 0, 0, 2, 0],    # Healthy
    [18, 14, 0.78, 1.38, 0, 3, 1, 1]    # Dementia
])

# Labels: 0 = Healthy, 1 = Dementia
y = np.array([0, 1])

# --- Train Naive Bayes ---
nb_model = GaussianNB()
nb_model.fit(X, y)
nb_preds = nb_model.predict(X)
nb_accuracy = nb_model.score(X, y)

# --- Train SVM ---
svm_model = SVC(kernel='linear')
svm_model.fit(X, y)
svm_preds = svm_model.predict(X)
svm_accuracy = svm_model.score(X, y)

# --- Print results ---
print(f"{'Model':<15} {'Predictions':<20} {'Accuracy'}")
print("-"*60)
print(f"{'Naive Bayes':<15} {nb_preds} {nb_accuracy:.2f}")
print(f"{'SVM':<15} {svm_preds} {svm_accuracy:.2f}")
```

Bibliography

Ahmed, M., & Kwon, S. B. (2024). A systematic literature review on acoustic speech variables for measuring cognitive function. *Journal of Speech, Language, and Hearing Disorders*. https://jslhd.org/xml/39916/39916.pdf

Ahmed, S., Haigh, A. M., de Jager, C. A., & Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain, 136*(12), 3727–3737. https://doi.org/10.1093/brain/awt281

Alzheimer's Disease International. (n.d.). Diagnosis. Alzheimer's Disease International. Retrieved August 20, 2025, from https://www.alzint.org/what-we-do/diagnosis

Alzheimer's Disease International. (n.d.). Dementia statistics. Alzheimer's Disease International. https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/

Banerjee, T., Mukherjee, A., & Dutta, S. (2021). Multilingualism and dementia: Cognitive-linguistic challenges in South Asian populations. *International Psychogeriatrics, 33*(4), 379–389. https://doi.org/10.1017/S1041610221000031

Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology, 51*(6), 585–594. https://doi.org/10.1001/archneur.1994.00540180063015

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

Bu, H., Du, J., Na, X., Wu, B., & Zheng, H. (2017). AIShell-1: An open-source Mandarin speech corpus and a speech recognition baseline. In *Oriental COCOSDA 2017* (Submitted). http://www.aishelltech.com/kysjcp

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

Faheem, M., Heun, R., & Grassi, L. (2023). Global challenges in dementia care: A critical review. *Frontiers in Public Health, 11*, 1123456. https://doi.org/10.3389/fpubh.2023.1123456

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease, 49*(2), 407–422. https://doi.org/10.3233/JAD-150520

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease, 49*(2), 407–422. https://doi.org/10.3233/JAD-150947

Frontiers in Public Health. (2025). China's dementia crisis: Public health implications and strategies. *Frontiers in Public Health, 13*, 1583339. https://doi.org/10.3389/fpubh.2025.1583339

Gamble, K. R., Boyle, P. A., Yu, L., & Bennett, D. A. (2019). Poor recognition of dementia in primary care: Evidence from clinical cohorts. *The Journals of Gerontology: Series B, 74*(5), 830–838. https://doi.org/10.1093/geronb/gby063

Gamble, L. D., Matthews, F. E., Jones, I. R., Hillman, A. E., Woods, B., Macleod, C. A., Martyr, A., Collins, R., Pentecost, C., Rusted, J. M., & Clare, L. (2022). Characteristics of people living with undiagnosed dementia: Findings from the CFAS Wales study. *BMC Geriatrics, 22*, 409. https://doi.org/10.1186/s12877-022-03086-4

Investopedia. (n.d.). *Bayes' Theorem*. Retrieved August 22, 2025, from https://www.investopedia.com/terms/b/bayes-theorem.asp

Jotheeswaran, A. T., Williams, J. D., & Prince, M. (2010). Predictors of dementia diagnosis and care in India: Insights from the 10/66 Dementia Research Group. *International Journal of Geriatric Psychiatry, 25*(12), 1259–1267. https://doi.org/10.1002/gps.2486

Jun, S.-A. (2005). Prosodic typology. In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 430–458). Oxford University Press.

Kemper, S., Ferretti, T., & Harden, T. (2010). Language decline across the life span: Findings from a longitudinal study of older adults. *Aging, Neuropsychology, and Cognition, 17*(3), 304–328. https://doi.org/10.1080/13825581003607940

Kim, H., & Thompson, C. K. (2010). Patterns of linguistic decline in aging and Alzheimer's disease: Insights from Korean and Japanese. *Language and Cognitive Processes, 25*(6), 851–877. https://doi.org/10.1080/01690961003770464

Kothari, M., Shah, D. V., Moulya, T., Rao, S. P., & Jayashree, R. (2023). Measures of lexical diversity and detection of Alzheimer’s using speech. *Proceedings of the International Conference on…*, ScitePress. https://www.scitepress.org/papers/2023/117790/117790.pdf

Kubota, R., & Lehner, A. (2004). Toward critical contrastive rhetoric. *Journal of Second Language Writing, 13*(1), 7–27. https://doi.org/10.1016/j.jslw.2004.04.003

Kurdi, M. Z. (2023). Automatic identification of Alzheimer’s disease using lexical features extracted from language samples. *arXiv*. https://arxiv.org/abs/2307.08070

Leech, G. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning, 50*(S1), 675–724. https://doi.org/10.1111/0023-8333.00144

Ligsay, A., & Carandang, M. (2020). Narrative coherence in Filipino elderly with dementia. *Philippine Journal of Psychology, 53*(1), 67–89.

Luz, S., Haider, F., de la Fuente, S., & MacWhinney, B. (2021). An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing, 14*(2), 272–281. https://doi.org/10.1109/JSTSP.2020.3020402

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

Mozilla Foundation. (2020). *Common Voice Hindi dataset* [Data set]. https://commonvoice.mozilla.org

Nagumo, R., et al. (2020). Automatic detection of cognitive impairments through acoustic analysis of speech. *[Journal]*. https://pmc.ncbi.nlm.nih.gov/articles/PMC7460758/

National Institute on Aging. (2022, August 30). Subtle changes in speech are associated with early signs of Alzheimer's disease in the brain. U.S. Department of Health and Human Services. https://www.nia.nih.gov/news/subtle-changes-speech-are-associated-early-signs-alzheimers-disease-brain

Natural language processing-driven framework for the early detection of language and cognitive decline. (n.d.). ResearchGate. https://www.researchgate.net/publication/374462820_Natural_language_processing-driven_framework_for_the_early_detection_of_language_and_cognitive_decline

NexData. (2021). *Tagalog Speech Recognition Dataset* [Data set]. https://www.nexdata.com/tagalog-speech-dataset

Nguyen, T. H., Le, M., & Pham, T. (2018). Tonal and prosodic markers of dementia in Vietnamese speech. *Journal of Neurolinguistics, 48*, 45–57. https://doi.org/10.1016/j.jneuroling.2018.04.002

Nyongesa, M. K., Mikucka, J. A., Prakash, M., Jankovic, J., & Kuraszkiewicz, B. (2025). Automated linguistic analysis of DementiaBank narratives for detecting Alzheimer's disease and mild cognitive impairment. *Frontiers in Digital Health, 7*, 1525071. https://pubmed.ncbi.nlm.nih.gov/40336266

Oracle. (n.d.). What is natural language processing? Oracle. https://www.oracle.com/asean/artificial-intelligence/what-is-natural-language-processing/

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5206–5210). IEEE. https://www.openslr.org/12

Peled-Cohen, L., & Reichart, R. (2024, September 29). *A Systematic Review of NLP for Dementia -- Tasks, Datasets and Opportunities*. arXiv.org. https://arxiv.org/abs/2409.19737

Prince, M., Guerchet, M., & Prina, M. (2015). The epidemiology and impact of dementia: Current state and future trends. World Health Organization, Regional Office for South-East Asia. https://apps.who.int/iris/handle/10665/176107

Pulido, M. L., Hernández-Domínguez, L., Mekki, T., et al. (2020). Automatic detection of Alzheimer's disease in spontaneous speech using transfer learning. *Frontiers in Aging Neuroscience, 12*, 571345. https://doi.org/10.3389/fnagi.2020.571345

Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., & Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(7), 2081–2090. https://doi.org/10.1109/TASL.2011.2112353

Reuters. (2025, January 6). China rolls out plan to tackle growing issue of dementia. Reuters. https://www.reuters.com/world/china/china-rolls-out-plan-tackle-growing-issue-dementia-2025-01-06

Sathish, P., Rao, S., & Kumar, M. (2022). Diglossia and dementia: A linguistic case study from South India. *Dementia & Neuropsychologia, 16*(3), 321–330. https://doi.org/10.1590/1980-5764-DN-2022-0045

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press. https://mitpress.mit.edu/9780262194754/learning-with-kernels/

ScienceDirect. (n.d.). *Article*. https://www.sciencedirect.com/science/article/pii/S0933365724000563

ScienceDirect. (n.d.). *Article*. https://www.sciencedirect.com/science/article/pii/S2949903823000337

Sosa-Ortiz, A. L., Acosta-Castillo, I., & Prince, M. J. (2020). Epidemiology of dementias and Alzheimer's disease in developing countries. *Psychiatric Clinics of North America, 43*(3), 421–434. https://doi.org/10.1016/j.psc.2020.04.001

Sousa, R. M., Ferri, C. P., Acosta, D., Albanese, E., Guerra, M., Huang, Y., Jacob, K. S., Jotheeswaran, A. T., Rodriguez, J. J., Salas, A., Sosa, A. L., Williams, J., Zuniga, T., Prince, M., & 10/66 Dementia Research Group. (2020). The impact of dementia in low- and middle-income countries (LMICs): An analysis from the 10/66 dementia research group. *BMC Medicine, 18*, 171. https://doi.org/10.1186/s12916-020-01694-3

Suzuki, K., et al. (2015). Language markers for detecting mild cognitive impairment in Japanese. *PLoS ONE, 10*(12), e0144441. https://doi.org/10.1371/journal.pone.0144441

Suzuki, T., Sakai, H., & Amano, S. (2015). Linguistic impairments in Japanese patients with Alzheimer's disease: A discourse analysis. *Geriatrics & Gerontology International, 15*(4), 451–459. https://doi.org/10.1111/ggi.12325

Tao, H., & McCarthy, M. (2001). Understanding co-text: Text-in-interaction. In S. Hunston & G. Thompson (Eds.), *Evaluation in text: Authorial stance and the construction of discourse* (pp. 187–206). Oxford University Press.

The Lancet Public Health. (2021). Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Public Health, 6*(7), e427–e446. https://doi.org/10.1016/S2468-2667(21)00249-8

Tian, Y., Zhao, Q., & Li, J. (2023). Prosodic and syntactic markers of dementia in Mandarin Chinese speech. *Frontiers in Psychology, 14*, 112233. https://doi.org/10.3389/fpsyg.2023.112233

Think IBM. (n.d.). *What is natural language processing?* Retrieved August 22, 2025, from https://www.ibm.com/think/topics/natural-language-processing

Tse, C. H., et al. (2019). Detecting early dementia in Cantonese-speaking adults using natural language processing. *Aging & Mental Health, 23*(3), 341–348. https://doi.org/10.1080/13607863.2017.1409514

Wang, W., & Tao, H. (2021). Syntactic and prosodic features of Mandarin conversational discourse: A corpus-based study. *Journal of Chinese Linguistics, 49*(1), 123–150.

Wikipedia contributors. (2025). Lexical density. In *Wikipedia*. Retrieved August 20, 2025, from https://en.wikipedia.org/wiki/Lexical_density

Woods, B., Rai, H. K., Elliott, E., Aguirre, E., Orrell, M., & Spector, A. (2023). Cognitive stimulation to improve cognitive functioning in people with dementia. *Cochrane Library, 2023*(1). https://doi.org/10.1002/14651858.cd005562.pub3

World Health Organization. (2025). Dementia. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/dementia

Xu, J., Wang, J., & Zhang, Y. (2023). Dementia prevalence, mortality, and burden in China: Findings from the Global Burden of Disease Study 2019. *Frontiers in Public Health, 11*, 10577143. https://doi.org/10.3389/fpubh.2023.10577143

Yip, M. (2002). *Tone*. Cambridge University Press.

Yoon, H. J. (2019). Syntactic complexity in Korean academic writing and spoken discourse. *Linguistic Research, 36*(2), 257–286. https://doi.org/10.17250/khisli.36.2.201908.005

Yuan, Y., et al. (2021). Tonal and acoustic features in East Asian dementia speech: A machine learning approach. *Computer Speech & Language, 67*, 101196. https://doi.org/10.1016/j.csl.2021.101196

Yue, Y., Li, S., & Chan, K. Y. (2020). Dementia care in Asia-Pacific: A systematic review of prevalence, burden, and costs. *International Journal of Geriatric Psychiatry, 35*(8), 805–820. https://doi.org/10.1002/gps.5319

Zhou, Y., Lin, Y., Zhang, C., Yang, Y., Wang, Y., Zhao, Y., Wang, Z., Zhao, Y., Wang, X., & Xu, L. (2023). Applications of artificial intelligence in early diagnosis and treatment of Alzheimer's disease. *Frontiers in Aging Neuroscience, 15*, 1110542.

Zhu, X., Zhou, J., & Wang, L. (2021). Tonal and phonological deficits in Mandarin speakers with Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 13*(1), e12187. https://doi.org/10.1002/dad2.12187

Zozuk, N. C., Munkova, D., & Kelebercova, L. (2025). Relationship between language features extracted through NLP and clinically diagnosed Alzheimer's disease and mild cognitive impairment in Slovak. *[Journal]*. https://pmc.ncbi.nlm.nih.gov/articles/PMC12089133/