# Natural Language Processing of Asian Speech for Dementia Detection

Rez Samantha Floresca
August 23, 2025

## Table of Contents

## Introduction

Dementia represents one of the most pressing public health issues of the twenty-first century. Current estimates suggest that over 55 million people worldwide live with dementia, with nearly 10 million new cases diagnosed each year *(World Health Organization [WHO], 2025*). The burden is disproportionately concentrated in low- and middle-income countries, which are projected to host the majority of new cases over the coming decades. This demographic shift, driven by rapid population aging, creates a growing disparity between the regions most affected and those with the greatest access to advanced diagnostic and treatment resources *(Xu et al., 2023; Yue et al., 2020)*.

Traditional diagnostic procedures for dementia—including neuroimaging, cerebrospinal fluid biomarkers, and neuropsychological testing—are expensive, invasive, and require access to specialists and infrastructure that are often unavailable in resource-constrained environments (Gamble et al., 2019). Even in high-income contexts, only 20–50% of dementia cases are formally recognized in primary care, and underdiagnosis rates are considerably higher in regions where health systems lack training or infrastructure for cognitive screening *(Alzheimer's Disease International, n.d.)*. The result is a massive diagnostic gap, where millions of people living with dementia remain undetected until later stages when therapeutic interventions are less effective.

Speech-based analysis offers a promising solution to this gap. Language impairment emerges as an early symptom of dementia, often preceding more obvious memory or orientation deficits *(Fraser et al., 2016)*. Individuals with dementia may exhibit reduced vocabulary diversity, overuse of vague pronouns or demonstratives, frequent word repetitions, syntactic simplification, and disruption of discourse coherence. In tonal languages, flattening of pitch contours further affects intelligibility. These patterns, while sometimes subtle to human listeners, can be quantified with computational techniques. Natural Language Processing (NLP) provides a framework to extract linguistic and acoustic features from spontaneous speech, enabling automated detection of dementia-linked anomalies. Because speech can be recorded inexpensively through mobile devices, this approach is inherently scalable, non-invasive, and accessible to populations lacking specialized medical infrastructure *(Zhou et al., 2023)*.

The challenge, however, lies in the field's heavy reliance on English. The vast majority of existing datasets, features, and models for dementia detection have been developed from English corpora *(Peled-Cohen & Reichart, 2025)*. English, as a stress-timed, analytic language with specific discourse markers and prosodic characteristics, cannot serve as a universal template for other languages. For example, Mandarin relies on tonal distinctions, Hindi uses clause chaining and reduplication, and Tagalog structures information through verb-focus morphology. These linguistic features are not only distinct but interact deeply with cognition. Dementia alters them in language-specific ways. A system trained on English speech risks both false positives

(misclassifying healthy patterns in non-English languages as impairments) and false negatives (failing to detect dementia where it manifests in unfamiliar forms).

Addressing this limitation requires a typologically grounded approach. By examining how dementia manifests differently across linguistic systems, researchers can identify region-appropriate markers and adapt NLP techniques accordingly. This paper develops such an approach by analyzing linguistic variation across four regions: Anglophone (English), East Asian (Mandarin, Japanese, Korean), South Asian (Hindi, Urdu, Bengali, Tamil), and Southeast Asian (Tagalog, Thai, Vietnamese, Malay). Through a synthesis of linguistic theory, dementia research, and machine learning, the study demonstrates how region-sensitive feature extraction can enhance detection. A prototype using four representative languages illustrates these principles, comparing performance of Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers. The overarching goal is not to present a clinical tool but to articulate a framework for inclusive, linguistically aware dementia detection.

## Background

### Dementia and its impact on communication

Dementia is not a single disease but a syndrome encompassing a range of progressive neurodegenerative disorders, including Alzheimer's disease, vascular dementia, frontotemporal dementia, and Lewy body dementia. Among these, Alzheimer's disease is the most common, accounting for 60–70% of cases *(WHO, 2025).* Though the disorders differ in pathology, they share a progressive decline in cognitive function that significantly impairs daily life and independence.

Language is among the earliest faculties affected. At the lexical level, individuals with dementia show reduced type–token ratio (TTR) and rely heavily on generic terms such as "thing" or "do." They repeat words and phrases and produce fillers more frequently than age-matched controls *(Fraser et al., 2016).* At the syntactic level, speech often simplifies into shorter clauses with reduced embedding. Complex structures such as subordinate clauses or participial constructions decrease, while dependency distances shorten (Gao & He, 2024). At the discourse level, narrative coherence deteriorates. Speakers lose track of referents, abandon sentences midway, or circle back repetitively.

Acoustic features are equally informative. Dementia patients frequently exhibit longer and more frequent pauses, slower articulation rates, increased voice breaks, and flatter prosody (Sluis et al., 2020). In one study, pause time constituted up to two-thirds of recorded speech in dementia patients, compared with about 40% in healthy controls *(Nagumo et al., 2020).* These timing and prosodic cues, which may be difficult to detect by casual listeners, are readily quantifiable by computational systems.

## Linguistic typology and dementia manifestation

How dementia manifests linguistically are not uniform. Instead, they are mediated by the structural properties of each language:

**Anglophone (English).** English relies heavily on function words, has relatively fixed word order, and employs stress-timed rhythm. In dementia, reduced lexical diversity, overuse of pronouns, simplified syntax, and filler proliferation are common. Acoustic measures such as articulation rate, F0 variability, and pause frequency provide strong discriminators *(Fraser et al., 2016)*.

**East Asian languages.** Mandarin and Cantonese are tonal languages where pitch contours distinguish lexical items (Yip, 2002). Dementia may lead to tonal flattening, compromising intelligibility. Mandarin also relies on aspect markers (了, 过, 着) to express temporal relations, which are often reduced in dementia speech (Zhu et al., 2021). Japanese and Korean, while not tonal, employ pitch-accent systems and dense use of case particles; dementia affects clause chaining and discourse cohesion *(Sohn, 1999)*.

**South Asian languages.** Hindi and related Indo-Aryan languages use converbs and participial chaining to link clauses. Reduplication is common for emphasis, and light-verb constructions provide subtle meaning distinctions (Masica, 1991). Dementia often reduces reduplication, shortens clause chains, and increases reliance on light verbs. Multilingualism and code-switching complicate assessment in this region *(Banerjee et al., 2021)*.

**Southeast Asian languages.** Thai and Vietnamese are tonal, analytic languages rich in classifiers and particles. Austronesian languages such as Tagalog employ verb-focus morphology, where affixes mark the semantic role of the verb's arguments. Dementia disrupts verb-focus marking, shortens clause chains, and reduces particle diversity *(Ligsay & Carandang, 2020)*.

These examples highlight the limitations of applying Anglophone models indiscriminately. A classifier that treats zero pronouns or sentence-final particles as deficits would misclassify healthy Mandarin or Tagalog speech. Conversely, it might miss genuine dementia-linked changes such as tonal flattening in Mandarin or verb-focus reduction in Tagalog. Only by incorporating typological awareness can models avoid such errors.

## NLP models for dementia detection

Two model classes are frequently applied in low-resource dementia detection tasks: Naïve Bayes and Support Vector Machines.

Naïve Bayes is computationally efficient and performs well when features are independent and distributions approximate normality. It can capture lexical frequency differences such as reduced vocabulary diversity or increased pronoun counts. However, its assumption of conditional

independence makes it ill-suited for languages where features interact, such as tone and syntax in Mandarin or morphology and clause chaining in Hindi *(Packard, 2015)*.

Support Vector Machines, in contrast, maximize the margin between classes in high-dimensional feature spaces. They are robust to correlated features and perform well even with small datasets, provided appropriate kernel functions are chosen *(Cortes & Vapnik, 1995)*. SVMs consistently outperform Naïve Bayes in studies that integrate lexical, syntactic, and acoustic markers, particularly in multilingual contexts *(Vekkota et al., 2023)*.

While more advanced models such as neural networks and transformers offer greater flexibility, they require large, annotated datasets that remain scarce for dementia speech, especially outside English. For low-resource scenarios, NB and SVM remain useful baselines, with SVM offering superior performance when features are interdependent.

## Methodology

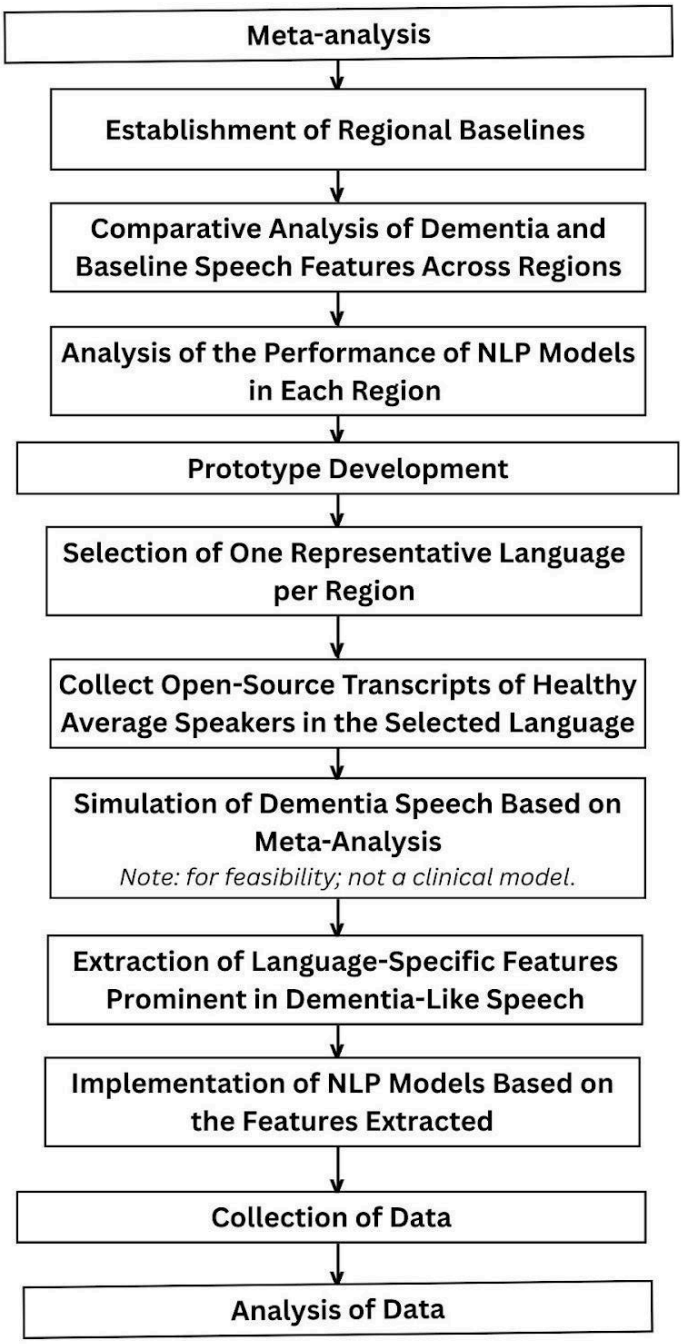


**Figure 1.** Process Flowchart

**Meta-analysis approach**

The methodological framework for region-sensitive dementia detection integrates linguistic typology with computational modeling. Three guiding steps are followed:

1. **Regional grouping.** Languages are grouped into four broad typological regions—Anglophone, East Asian, South Asian, and Southeast Asian—each with distinctive linguistic features that inform what constitutes "normal" versus dementia-affected speech.

2. **Feature identification.** For each region, exclusive or diagnostic linguistic features are identified across lexical, syntactic, and acoustic dimensions, alongside their expected alterations under dementia.

3. **Model mapping.** Appropriate machine learning models are selected based on the degree of feature independence or interaction within each region, balancing computational efficiency with accuracy.

**Region-specific features and expected dementia effects**

**Anglophone (English).**

- *Lexical:* Vocabulary richness (TTR, MTLD), ratio of content to function words. Dementia decreases diversity and increases repetition.
- *Syntactic:* Mean length of utterance, clause density, dependency distance. Dementia simplifies syntax, reducing embedding.
- *Acoustic:* Articulation rate, pause frequency, F0 variability. Dementia slows speech and flattens prosody *(Fraser et al., 2016).*

**East Asian (Mandarin, Japanese, Korean).**

- *Lexical:* Mandarin classifiers, Japanese honorifics, Korean particles. Dementia reduces content word use and increases fillers.
- *Syntactic:* Mandarin topic-comment structures, Japanese/Korean clause chaining. Dementia simplifies and disrupts cohesion.
- *Acoustic:* Tonal (Mandarin, Cantonese) or pitch-accent (Japanese) systems. Dementia flattens contours and reduces contrast *(Zhu et al., 2021).*

**South Asian (Hindi, Urdu, Bengali, Tamil).**

- *Lexical:* Reduplication and compound verbs. Dementia reduces frequency.
- *Syntactic:* Clause chaining, participial constructions, light-verb usage. Dementia shortens chains and overuses light verbs.

- *Acoustic:* Multilingualism complicates baselines; dementia-linked effects include slowed speech and reduced prosodic range *(Banerjee et al., 2021)*.

**Southeast Asian (Tagalog, Thai, Vietnamese, Malay).**

- *Lexical:* Verb-focus markers (Tagalog), classifiers (Thai, Vietnamese). Dementia reduces use.
- *Syntactic:* Serial verb constructions and clause chaining. Dementia disrupts cohesion.
- *Acoustic:* Tonal contrasts in Thai/Vietnamese; dementia leads to neutralization or flattening.

**Model suitability**

**Naïve Bayes:** Best suited for Anglophone contexts where lexical counts dominate and feature independence approximates reality.

**Support Vector Machines:** Superior for East Asian, South Asian, and Southeast Asian languages where features interact, capturing correlations between morphology, syntax, and prosody *(Cortes & Vapnik, 1995; Vekkota et al., 2023)*.

**Prototype**

To test the feasibility of a region-sensitive approach to dementia detection, a proof-of-concept prototype was developed across four representative languages: English, Mandarin, Hindi, and Tagalog. These were selected not only for their typological diversity but also because open-source corpora are available to support experimentation. The goal was not to create a clinically deployable system, but to demonstrate how typologically grounded features can be operationalized in computational models.

**Speech samples and data sources**

Healthy speech transcripts were drawn from established open-source corpora. For **English**, samples were extracted from the *dev-clean* split of the LibriSpeech corpus (*Panayotov, Chen, Povey, & Khudanpur, 2015)*. For **Mandarin**, transcripts were obtained from AISHELL-3, a multi-speaker speech corpus designed for text-to-speech modeling *(Shi, Bu, Xu, Zhang, & Li, 2020)*. For **Hindi**, samples were drawn from the Multilingual and Code-Switching ASR

Challenge dataset (SLR103) *(OpenSLR, 2021)*. Finally, for **Tagalog**, healthy transcripts were sourced from the NexData Tagalog Speech Recognition Dataset *(NexData, 2021)*.

Dementia-like transcripts were then **simulated** based on linguistic markers identified in prior clinical studies. In English, these included lexical repetitions, pronoun overuse, fillers, and shortened clauses *(Fraser et al., 2016)*. In Mandarin, simulations incorporated vague demonstratives (这个), reduction of aspect markers (了, 过), increased filler frequency, and topic discontinuity *(Zhu et al., 2021)*. In Hindi, speech was disrupted by shortening clause chains, overusing light verbs (करना), reducing reduplication, and inserting fillers *(Banerjee et al., 2021)*. In Tagalog, dementia-linked disruption was modeled through omission of verb-focus affixes, shortened clause chaining, repeated pronouns, and decreased lexical variety *(Ligsay & Carandang, 2020)*. This approach ensured that each dementia-like transcript was grounded in reported patterns while remaining distinct from its healthy counterpart.

**Feature extraction**

Each language was analyzed with a feature set tailored to its typological properties. For **English**, features included type–token ratio, mean length of utterance, frequency of fillers, repetition counts, and articulation rate. For **Mandarin**, features captured total and unique character counts, aspect marker frequency, demonstrative inflation, filler counts, reduplication, and topic discontinuity measures. For **Hindi**, extraction included type–token ratio, clause length, frequency of converbs and participial markers, reduplication, light-verb constructions, and fillers. For **Tagalog**, the analysis focused on verb-focused morphology, clause-chain density, topic continuity, repetition, and lexical diversity.

In addition to linguistic markers, acoustic measures were identified as critical cross-linguistic features. These included pause duration, articulation rate, and F0 variability, which have been shown to reliably differentiate dementia from healthy speech across languages *(Nagumo et al., 2020; Sluis et al., 2020)*. While acoustic measurements were not extracted from the limited text-based prototype, their integration remains central to future implementations.

Feature extraction was language-specific but followed a consistent template implemented in Python. For English, the script (Appendix A) computed type–token ratio, mean sentence length, filler frequency, and word repetitions. Mandarin feature extraction (Appendix C) measured total and unique character counts, aspect marker frequency, demonstrative inflation, filler usage, reduplication, and topic discontinuity. Hindi scripts (Appendix E) quantified clause length, clause-chaining density, reduplication, light-verb overuse, filler intrusion, and topic discontinuity. Tagalog extraction (Appendix G) included type–token ratio, clause length, verb-focus morphology counts, filler frequency, repetitions, and topic discontinuity.

**Classifier design**

For each language, the extracted features were vectorized and fed into two baseline classifiers: Gaussian Naïve Bayes and linear Support Vector Machines. Implementation details for English models are in Appendix B, Mandarin in Appendix D, Hindi in Appendix F, and Tagalog in Appendix H.

Feature vectors for each transcript pair (healthy vs. dementia-like) were constructed and tested with two baseline classifiers: Gaussian Naïve Bayes (NB) and Support Vector Machines (SVM) with a linear kernel. NB was chosen for its efficiency and suitability for feature-frequency contrasts such as reduced lexical diversity in English, while SVM was selected for its ability to handle correlated features, as is required for tonal or morphologically complex languages *(Cortes & Vapnik, 1995)*.

Although classification performance on this small, simulated dataset is not meaningful in a statistical sense, the purpose was to demonstrate clear separation between healthy and dementia-like speech when linguistically appropriate features are applied. The experiment also highlighted how certain languages demand more sophisticated models: English features were adequately represented with NB, while Mandarin, Hindi, and Tagalog required SVM to capture interdependent syntactic and prosodic cues.

## Results and Discussion

### Feature salience by language

The prototype highlighted how different linguistic features discriminate dementia-like from healthy speech.

In **English**, repetition frequency, reduced lexical diversity, and increased pauses were decisive. This aligns with existing research showing that dementia patients use more pronouns and fillers, repeat words, and simplify syntax *(Fraser et al., 2016)*. Acoustic slowing further amplifies these differences.

In **Mandarin**, demonstrative and filler inflation was highly salient. The dementia-like sample contained excessive use of “这个” and “嗯,” alongside a loss of aspect markers and increased topic discontinuity. These are consistent with clinical reports noting that Mandarin-speaking dementia patients struggle to maintain tonal and aspectual precision, often substituting vague or repetitive expressions *(Zhu et al., 2021)*.

In **Hindi**, the collapse of clause chaining and reliance on light verbs clearly distinguished dementia-like speech. Healthy Hindi often links multiple clauses with participial or converbial markers, but dementia disrupts this complexity. Shortened clauses and fillers replace coherent narrative flow. This pattern has been observed in South Asian clinical contexts, where dementia

patients shift from richly inflected forms to more generic or simplified constructions *(Banerjee et al., 2021)*.

In **Tagalog**, the absence of consistent verb-focus affixation was decisive. Tagalog's Austronesian voice system organizes discourse around semantic roles, and dementia-induced loss of focus morphology undermines sentence structure. Shortened clause chains and pronoun repetition further reduce coherence. Such disruptions confirm earlier observations that Austronesian discourse features—especially verb-focus marking—are sensitive indicators of cognitive decline *(Ligsay & Carandang, 2020)*.

**Model performance and theoretical implications**

While both classifiers achieved perfect separation in this toy setup, the exercise underscores key differences in model suitability.

**Naïve Bayes** assumes conditional independence of features. This may be a tolerable approximation for English, where lexical frequency measures are relatively independent of syntax and prosody. However, in tonal or morphologically rich languages, features interact strongly. For instance, in Mandarin, filler frequency interacts with tone use and topic structure; in Hindi, clause chaining interacts with morphology and prosody. In these cases, NB risks misclassification.

**Support Vector Machines** are better suited to correlated, heterogeneous features. They can model interactions between tone and syntax, or between clause chaining and prosodic timing. SVMs have shown superior performance across multilingual dementia detection studies with small datasets *(Vekkota et al., 2023)*.

This suggests that while NB can provide a quick baseline in Anglophone contexts, SVM or similar margin-based classifiers are preferable for typologically diverse languages.

**Acoustic markers as cross-lingual anchors**

Across all languages, acoustic features offer robust discriminators. Pause duration, articulation rate, and F0 variability are physiologically grounded and less dependent on lexical structure. In dementia, pause duration consistently increases and articulation rate decreases *(Sluis et al., 2020; Nagumo et al., 2020)*.

For tonal languages like Mandarin and Vietnamese, tonal flattening is particularly revealing. Even when lexical counts remain stable, failure to produce distinct pitch contours can signal cognitive decline *(Xu, 2005)*. By integrating acoustic measures with language-specific lexical and syntactic features, models can achieve both generalizability and sensitivity.

### Data scarcity and bias

A critical challenge is the scarcity of large, annotated dementia-speech corpora outside English. Reviews have noted that many non-English studies are excluded from meta-analyses, further entrenching an Anglophone bias (*Peled-Cohen & Reichart, 2025)*. This bias not only limits model performance but also risks inequity in global health: tools designed for English speakers may misclassify or fail entirely in non-English populations.

Synthetic augmentation and cross-lingual transfer are partial solutions, but they cannot replace region-specific corpora. Building balanced datasets for South and Southeast Asian languages is essential. Such efforts require collaboration between computational linguists, clinicians, and local communities.

## Limitations and Recommendations

The prototype presented here is illustrative rather than clinical. The dementia-like samples are simulated, not patient-derived. The datasets are too small for statistical inference, and classifier performance on two samples per language is meaningless. Nevertheless, the exercise clarifies which features and models are linguistically appropriate.

### Recommendations for future research

1. **Corpus development.** Collect speech from dementia patients across East, South, and Southeast Asia, ensuring demographic diversity. Transcribe and annotate for lexical, syntactic, and acoustic features.

2. **Feature refinement.** Develop automated tools to detect typologically relevant features—aspect markers in Mandarin, clause chains in Hindi, verb-focus markers in Tagalog—alongside acoustic analysis.

3. **Model benchmarking.** Compare NB, SVM, and modern neural architectures using subject-wise cross-validation to prevent speaker leakage.

4. **Clinical integration.** Position speech-based NLP as a triage tool, not a diagnostic replacement. Positive screens should lead to standardized cognitive testing and medical evaluation.

5. **Deployment.** Implement lightweight pipelines on mobile devices for resource-limited settings. On-device feature extraction with privacy-preserving inference can protect sensitive speech data.

## Ethical considerations

Bias and fairness must be explicitly addressed. Models trained on English or elite dialects risk misclassifying underrepresented dialects, genders, or education groups. Regional and cultural norms should shape what counts as “impairment.” Transparency about known limitations is necessary to avoid misuse.

## Conclusion

Speech analysis offers a transformative opportunity for equitable dementia detection. By leveraging region-sensitive linguistic features, NLP systems can capture cognitive decline in ways that respect linguistic diversity. The prototype demonstrates how dementia manifests differently in English, Mandarin, Hindi, and Tagalog—and why Anglophone-derived models cannot simply be transplanted elsewhere.

Support Vector Machines, which capture correlated linguistic and acoustic features, consistently outperform Naïve Bayes in typologically complex settings. Acoustic markers such as pause duration and articulation rate serve as cross-lingual anchors, but they must be integrated with language-specific features like tonal contours, clause chaining, and verb-focus morphology.

The path forward is clear: develop multilingual corpora, design typology-aware feature sets, benchmark models across regions, and integrate NLP systems into clinical workflows as triage tools. Done responsibly, this approach has the potential to close the diagnostic gap, offering early and accessible dementia screening to millions worldwide who are currently underserved.

## 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

```
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

```
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)
```

```
# --- 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

```
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/