Figurative Language in Alzheimer's Discourse: Linguistic and Neural Alignment in Clinical Narratives

Diana Kylymnyk[1], **Vitória Hilgert**[2], **Helena Caseli**[2],
Ed Watkins[1], **Aline Villavicencio**[1,3,4], **Rodrigo Souza Wilkens**[1]
[1]University of Exeter (UK)
{d.kylymnyk, e.r.watkins, a.villavicencio, r.wilkens}@exeter.ac.uk

[2]Federal University of São Carlos (Brazil)
vitoriahilgert@estudante.ufscar.br, helenacaseli@ufscar.br
[3]University of Sheffield (UK)
[4]Federal University of Rio Grande do Norte (Brazil)

Abstract

Figurative language, including multiword expressions and metaphors, provides a sensitive lens on cognitive functioning but remains largely overlooked in computational studies of Alzheimer's Disease (AD). This work investigates figurative-language patterns in AD and whether they can help in distinguishing AD from non-clinical discourse and whether a neural model encodes comparable linguistic tendencies. We propose a two-step framework that combines relevant linguistic features with neural representations. Figurative expressions are automatically identified using Large Language Models focusing on idiomaticity and metaphor detection. These figurative language indicators are integrated with lexical, syntactic, and readability features and used to train classifiers on the ADReSS dataset. Correlation and proxy-model analyses reveal significant alignment between linguistic indicators and model predictions: participants with AD produce fewer figurative constructions, lower lexical diversity, and more concrete language. The results obtained demonstrate that contextual embeddings implicitly encode linguistic cues associated with cognitive decline and highlight the value of figurative-language metrics for transparent and linguistically grounded clinical NLP.

Keywords: Figurative language, Alzheimer's Disease, Multiword Expressions, Metaphor detection, Clinical NLP

1. Introduction

Natural Language Processing (NLP) has increasingly been applied to healthcare and clinical research, offering tools for automated diagnosis support, symptom monitoring, and patient communication analysis (Khurana et al., 2023; Hossain et al., 2023). In recent years, language-based approaches have shown interesting results in neurological and neurodegenerative disorders, where spontaneous speech provides insights into cognitive functioning (De la Fuente Garcia et al., 2020; Fraser et al., 2015). Among these conditions, Alzheimer's Disease (AD) has attracted growing attention, as linguistic alterations (e.g., lexical retrieval issues, syntactic simplification, and discourse incoherence) often emerge in early stages of the disease (Tóth et al., 2015; Luz et al., 2020). Detecting such patterns enables non-invasive and cost-effective screening (e.g., (Mirheidari et al., 2018; Paula et al., 2018)), complementing traditional clinical assessment methods and contributing to the early identification of cognitive decline.

In AD, executive control and the ability to integrate contextual cues are often impaired, affecting both literal and non-literal language processing (Amanzio et al., 2008; Papagno, 2001; Rassiga et al., 2009). Figurative language, therefore, offers an especially sensitive window into cognitive functioning. The comprehension and production of figurative language such as multiword expressions (MWEs), particularly idiomatic expressions, and metaphors require flexible reasoning, semantic control, and abstraction capabilities that gradually deteriorate in AD. Clinical studies show that individuals with AD tend to interpret idioms literally and struggle to understand novel metaphors, although familiar expressions may remain accessible through long-term memory (Papagno, 2001; Amanzio et al., 2008). From a cognitive-linguistic perspective, such figurative phenomena depend on conceptual mapping and domain transfer (Kövecses, 2010, 2005), mechanisms that weaken as executive and semantic control decline.

Despite this evidence, figurative language has been largely overlooked in computational studies of Alzheimer's detection. Existing NLP work has primarily focused on lexical, syntactic, and discourse-level features (Fraser et al., 2016; Luz et al., 2020), with few attempts to quantify how figurative language patterns change in clinical populations systematically. Moreover, little is known about how neural language models encode or respond to figurative expressions, or whether such representations contribute meaningfully to diagnostic prediction.

Building on this gap, this work examines figura-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11379–11389
11–16 May 2026.

tive language in Alzheimer's-related discourse and whether neural models reflect comparable linguistic tendencies. Using spontaneous speech transcripts, we combine idiomatic and metaphorical indicators with lexical, syntactic, and readability features to assess the ability of neural representations to encode cues of cognitive decline.

This work introduces a methodology for analysing figurative-language features in Alzheimer's speech, an evaluation framework quantifying their alignment with predictions from a classifier (in this case BERT (Devlin et al., 2019)), and a linguistic interpretation identifying which cues the model reflects.

This paper is organized as follows. Section 2 reviews related work on language-based Alzheimer's detection and figurative-language processing. Section 3 describes the dataset and methodological framework. Section 4 reports the main results and analyses, and Section 5 concludes with a summary and future directions.

2. Related Work

Research on language and Alzheimer's disease has evolved from early feature-based analyses to neural models capable of capturing complex linguistic patterns. Non-literal language, particularly idiomatic expressions and metaphors may also reflect the cognitive flexibility and semantic control that deteriorate in Alzheimer's disease. However, figurative-language phenomena remain mostly absent from computational detection pipelines. In this section we review prior work from two complementary perspectives: Section 2.1 outlines language-based methods for Alzheimer's detection, while Section 2.2 examines figurative-language use in cognitive impairment

2.1. Language-based detection of Alzheimer's disease

Language provides one of the earliest and least intrusive indicators of cognitive decline (Forbes et al., 2024; Forbes-McKay and Venneri, 2005). Traditional approaches have relied on handcrafted linguistic features, capturing aspects such as lexical diversity, syntactic complexity, and discourse coherence, which tend to decrease with disease progression (Fraser et al., 2016; Tóth et al., 2015). Early work using feature-based machine learning demonstrated that measures like mean sentence length, type–token ratio, and pause frequency can differentiate between cognitively healthy and impaired speakers (Fraser et al., 2016; Luz et al., 2020), reflecting reductions in lexical diversity, syntactic complexity, and discourse coherence.

More recent studies have applied deep learning models, such as BERT (Devlin et al., 2019) and other transformer architectures, to the task of Alzheimer's detection (Luz et al., 2020; Balagopalan et al., 2020; Luz et al., 2021), capturing contextual dependencies from spontaneous-speech transcripts and reaching accuracies around 80-85% on benchmark datasets such as ADReSS. These models capture rich contextual information from spontaneous speech transcripts and have achieved competitive classification performance.[1]

Yet, most computational work still focuses on surface-level cues, overlooking figurative-language processing, even though psycholinguistic evidence shows its sensitivity to cognitive decline. In particular, idiomatic and figurative language, which depend on semantic flexibility and abstraction, are rarely examined in detection pipelines (Papagno, 2001; Amanzio et al., 2008; Rassiga et al., 2009).

2.2. Figurative Language in Cognitive Impairment

The processing of figurative language, including idioms, metaphors, and other multiword expressions, is a cognitively demanding activity that depends on executive control, semantic integration, and contextual flexibility (Glucksberg, 2003; Bambini et al., 2011; Cardillo et al., 2012). In Alzheimer's Disease (AD), these abilities deteriorate, leading to difficulties inhibiting literal interpretations and mapping conceptual meaning across domains (Papagno, 2001; Amanzio et al., 2008). As a result, AD patients often understand only the compositional, literal meaning of idiomatic phrases, even when they are familiar with the figurative sense stored in long-term memory. This tendency towards literalism has been linked to deficits in the central executive system, which mediates abstraction and reasoning (Roncero and de Almeida, 2014; Rassiga et al., 2009).

Psycholinguistic studies further show that novel metaphors pose a particular challenge, as their interpretation requires flexible conceptual mapping and the recruitment of inferential processes (Amanzio et al., 2008). In contrast, well-established idioms and conventional metaphors may remain relatively preserved, relying more on semantic memory than active executive processing. These findings suggest that figurative language can serve as a linguistic marker of cognitive decline, complementing traditional indicators such as lexical diversity or syntactic simplification (Kövecses, 2005; Rassiga et al., 2009).

[1] In this work, we consider only text-based models. For models that consider other models, see (Meghanani et al., 2021; Sarawgi et al., 2020; Farrús and Codina-Filbà, 2020).

From a computational perspective, MWEs and metaphors have long presented challenges to NLP systems due to their non-compositional semantics and context dependence (Shutova, 2010; Lai and Nissim, 2024). Traditional rule-based methods struggled to differentiate figurative from literal usage, while recent transformer-based and large language models (LLMs) offer improved contextual understanding (Lai and Nissim, 2024; Ge et al., 2023). Nevertheless, their internal representations remain opaque, and few studies have evaluated how well these models capture figurative phenomena in clinical contexts (Olivero, 2024; Tong et al., 2021). Bridging psycholinguistic insights with computational modelling, therefore, provides an opportunity to investigate whether changes in figurative language use can be automatically identified and interpreted as part of Alzheimer's detection pipelines.

The present study, therefore, examines whether these patterns, known to reflect cognitive flexibility, are also mirrored in the decision patterns of a neural language model trained for Alzheimer's detection.

3. Materials and Methods

This section presents the methodological framework used to investigate the relationship between figurative language and AD in spontaneous speech and to evaluate whether a neural model reflects comparable linguistic tendencies. As illustrated in Figure 1, the workflow consists of five main steps organized into three analytical phases. The first phase, figurative-language detection, involves data preprocessing (Section 3.1) and automatic identification of idiomatic and metaphorical expressions using an LLM (Section 3.2). The second phase, linguistic feature and model construction, comprises feature extraction integrating linguistic, readability, and figurative-language indicators (Section 3.3) and classification with a fine-tuned BERT model for Alzheimer's detection (Section 3.4). The final phase, linguistic alignment analysis, examines the correspondence between interpretable linguistic cues and model predictions (Section 3.5). Together, these stages address 2 research questions (RQs):

RQ1 To what extent can figurative-language indicators differ between AD and control participants?

RQ2 To what extent do neural model predictions align with these interpretable linguistic indicators?

3.1. Dataset

Experiments use the ADReSS dataset (Luz et al., 2020), a publicly available benchmark dataset for Alzheimer's Disease (AD) detection, containing

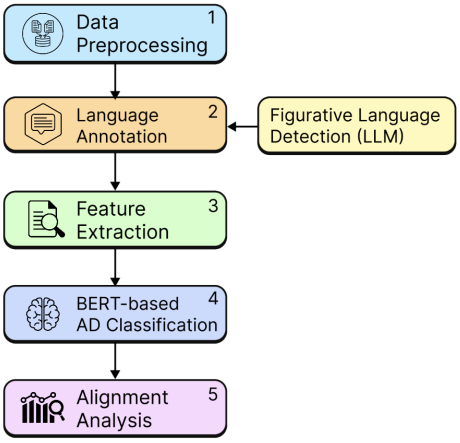

Figure 1: Overview of the experimental pipeline.

spontaneous speech transcripts of participants describing the *Cookie Theft* picture, including typical hesitations and disfluencies of cognitive impairment. This corpus is particularly suited for analyzing figurative-language patterns because it consists of spontaneous narrative speech where idiomatic and metaphorical usage naturally occurs.

The dataset contains a balanced sample of participants diagnosed with Alzheimer's Disease (AD) and cognitively healthy controls (CH), matched by age and gender to minimize demographic bias. Specifically, the dataset comprises 156 participants in total (78 AD and 78 CH). ADReSS provides predefined training and test splits (108 instances for training: 54 AD and 54 CH; 48 instances for testing: 24 AD and 24 CH) to ensure comparability across models and reproducibility of results. Ethical approval for data collection was handled by the ADReSS organizers, and all participant data have been anonymized to protect privacy. We used only the textual transcripts in this study, excluding acoustic features, and demographic metadata were included for matching purposes but not used as model input.

This corpus is particularly suitable for analysing figurative-language usage, since spontaneous narratives naturally elicit idiomatic and metaphorical expressions alongside literal ones. The dataset therefore supports both the linguistic analysis of figurative phenomena (addressing RQ1) and the evaluation of whether neural models reflect these linguistic distinctions (addressing RQ2).

3.2. Figurative Language Identification

To capture indicators of figurative language use, we implemented an automatic pipeline for identifying multiword expressions (MWEs) and metaphoric expressions in the ADReSS transcripts. Detection relied on a widely adopted open weight LLM, in

this case Llama 3 (Llama Team, AI @ Meta, 2024) prompted in a chain-of-thought style to analyse each sentence and output a structured JSON object containing counts and lists of idiomatic, literal, and metaphorical expressions.

Prompt engineering techniques were adapted from previous studies on idiomaticity detection (Phelps et al., 2024). Following this work, we tested several configurations, including zero-shot and few-shot prompting, as well as strategies such as expert impersonation and removing restrictive instructions. These adaptations led to improved accuracy and a more consistent reasoning pattern across sentences.

Two task-specific prompts were developed, one for MWEs and one for metaphors. Each prompt instructed the model to: (1) Identify candidate expressions that could be idiomatic or metaphorical; (2) Evaluate whether each expression is used figuratively or literally within its sentence context; and (3) Assign a global sentence label accordingly to the task (MWE or metaphor identification).

Few-shot examples were included in both prompts to encourage consistent reasoning, and outputs were formatted in JSON for transparent aggregation. Each example illustrated the reasoning process, from recognising an expression to classifying its usage, improving the model's reliability in ambiguous cases. An excerpt of the idiom-classification prompt is shown in Figure 2.

```
Your task is to decide whether a sentence contains
idiomatic usage ('i') or only literal usage ('l') of
multiword expressions (MWEs), also called potentially
idiomatic expressions (PIEs).

### Instructions:
1. Identify all PIEs (e.g., "kick the bucket",
    "break the ice").
2. For each PIE, compare how it's used in the sentence
    with its idiomatic and literal meanings.
3. If at least one PIE is used idiomatically, classify the
    whole sentence as 'i'.
4. If all PIEs are used literally, classify the sentence as
    'l'.
5. At the end, return a JSON with the following structure:
// JSON and examples
```

Figure 2: Idiom classification prompt used to label sentences.

These detection steps were first validated on two external benchmark datasets: MAGPIE (Haagsma et al., 2020) for the classification of idiomatic usage, and Metaphor-Paraphrase (Bizzoni and Lappin, 2018) for distinguishing metaphorical from literal sentences. Validation confirmed that the adapted prompting strategy yields consistent detection performance across MWEs and metaphors, providing a reliable basis for large-scale analysis of figurative language in spontaneous speech.

By combining prompt engineering, contextual reasoning, and structured outputs, this identification pipeline produces interpretable annotations of idiomatic and metaphorical language. These annotations form the basis for the quantitative linguistic features extracted in Section 3.3.

3.3. Linguistic Feature Extraction

We extracted linguistic, readability, and figurative-language features to characterize each transcript, enabling comparison between traditional and new indicators of cognitive decline. This stage aims to translate linguistic properties of each transcript into interpretable quantitative variables that can be used both for traditional classification and for the linguistic-alignment analyses described in Section 3.5.

Lexical and Structural measures quantify the overall complexity and richness of the lexical and syntactic content of the transcripts, which tend to decrease in Alzheimer's speech. These features comprise (L1) token count, (L2) sentence count, (L3) average sentence length, (L4) type-token ratio, (L5) moving-average TTR (Covington and McFall, 2006), (L6) part-of-speech counts, (L7) lexical diversity indicators (Lu, 2012).

Readability Indices estimate the cognitive load and syntactic predictability of a text, providing standardised metrics of linguistic accessibility that are often affected by cognitive impairment. We explore 5 indices: (R1) Flesch Reading Ease (Flesch, 1948), (R2) Flesch-Kincaid Grade (Kincaid et al., 1975), (R3) Gunning Fog Index (Gunning, 1952), (R4) SMOG (McLaughlin, 1969), (R5) Automated Readability Index (Kincaid et al., 1975).

Psycholinguistic Features capture conceptual specificity and abstraction levels in language use, which are known to differ between cognitively healthy and impaired speakers. These features were derived from concreteness and abstractness norms for English words (Brysbaert et al., 2014). We computed three indicators: (A1) average concreteness score, (A2) average abstractness score, and (A3) concreteness coverage, representing the proportion of words in a transcript that could be assigned a concreteness or abstractness value.

Discourse Markers measures provide insight into each narrative's cohesion and logical structure, reflecting how speakers organize and relate ideas across utterances. Discourse features capture the use of connectives (e.g., *however*, *but*, *and*), following (Fraser, 1999). Each of the nine connectives was represented as a distinct feature (D1-D9).

Figurative-language identified by the LLM-based identification step (Section 3.2) is post-processed to generate quantitative indicators of figurative usage. From the model's structured JSON

output, we extracted the following counts and proportions for each transcript:

- **PIE-related features:** total number of identified potential idiomatic expressions (P1), number of expressions used idiomatically (P2), number used literally, and sentence-level classifications (i.e., *Idiomatic* or *Literal*) (P3) according to the usage of its PIEs.
- **Metaphor-related feature:**total number of metaphorical expressions (M1), number of literal interpretations (M2), and sentence-level classifications (M3); i.e., *Metaphoric* or *Literal*.

From these, we aggregated document-level measures: the proportion of sentences containing MWEs (P4), the proportion of sentences containing metaphors (M4), and the ratios of idiomatic/metaphoric to literal usages (P5 and M5).

These four feature groups (i.e., lexical, readability, discourse, and figurative) form a multidimensional representation of each transcript that combines traditional linguistic analysis with cognitive-linguistic insights. This feature set is used in the modelling experiments and the linguistic alignment analyses.

3.4. Models

This stage evaluates how linguistic and contextual representations distinguish between Alzheimer's and control narratives, providing the modelling foundation for the linguistic alignment analyses described in Section 3.5.

We compare both traditional feature-based classifiers and transformer-based neural models, enabling a direct contrast between interpretable linguistic cues and deep contextual embeddings. These models are evaluated using accuracy and weighted F1-score. Given the fixed separation between the training and test sets, the bootstrap resampling (1000 iterations) is used to estimate confidence intervals (95% confidence).

Feature-based Models: We trained a range of machine learning classifiers using the handcrafted linguistic features derived in Section 3.3, including Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machines (SVM). Model performance was evaluated using accuracy, F1-score, and bootstrap resampling to compute 95% confidence intervals, ensuring statistical robustness. These models rely exclusively on interpretable linguistic features and serve as a transparent baseline for assessing how well human-readable linguistic indicators predict AD.

Transformer-based Models[2]: We fine-tuned the BERT base uncased model for the neural approach, which captures rich contextual information from text. Three variations of the BERT model are explored. (1) *Frozen BERT:* we extracted the *[CLS]* token embedding as a semantic vector and averaged these representations across sentences to obtain document-level embeddings. This representation feeds a Logistic Regression. (2) *Fine-tuned BERT:* a classic fine-tuning approach for an end-to-end binary classification (AD or control). (3) *Custom BERT:* we further explore BERT's ability to capture Alzheimer-related patterns, by using a custom classification head composed of a two-layer multilayer perceptron instead of the standard one-layer. This deeper classification head increases representational capacity, but may require more data than the standard BERT.

In summary, this modelling phase compares a highly interpretable feature-based baseline, which encodes linguistic features related to symptoms of AD, with highly performing transformer-based models. The resulting predictions feed the linguistic alignment analyses presented in Section 3.5, addressing RQ2 on whether model decision patterns reflect interpretable linguistic indicators.

3.5. Linguistic Consistency Assessment

This stage investigates whether the neural model's decision patterns correspond to interpretable linguistic indicators, thereby addressing RQ2 and testing the second hypothesis of linguistic alignment between model predictions and human-observed features. To this end, we conducted a two-stage analysis to quantify the correspondence between linguistic variables and model behaviour.

Correlation analysis: Spearman correlation is computed between each linguistic feature and both (a) the gold labels and (b) the predicted labels by the best-performing model. This comparison allows us to determine whether linguistic dimensions, such as figurative-expression density, lexical diversity, concreteness, and readability, exhibit similar tendencies when correlated with human annotations and model outputs. In other words, the goal is to assess how closely the model's predictions reflect the linguistic distinctions that also separate AD from control speech.

All correlations are calculated using Spearman's ρ (rank-based) coefficients to capture relationships between linguistic features and model outputs. Significance levels were set at $p - value < 0.05$. Weak associations ($|\rho| < 0.2$) are excluded to focus on moderate to strong relationships, ensuring that only linguistically meaningful correlations are retained for analysis.

Feature Alignment via Proxy Modelling: To

[2]Optimal hyperparameters identified through a grid search conducted across combinations of learning rate (2e-5, 3e-5), batch size (8, 16), number of epochs (2, 3), and dropout rates (0.1, 0.3).

estimate how much of the neural model's predictive behaviour could be explained by explicit linguistic cues, we trained a Random Forest proxy model to approximate the BERT classifier's predicted probabilities using the handcrafted linguistic features as input. This procedure provides a quantifiable measure of linguistic alignment, indicating the degree to which the model's decisions can be inferred from interpretable variables. Feature importance from the proxy model is then analysed to identify which linguistic properties, such as figurative-language density, lexical diversity, or readability indices, most strongly aligned with the model's outputs.

4. Findings and Analysis

This section presents the main empirical results and analyses examining how linguistic and figurative-language indicators contribute to the detection of AD from spontaneous speech. The goal is to evaluate both the reliability of the figurative-language features and their relation to model performance and linguistic interpretation. We first evaluate the automatic detection of MWEs and metaphors that underpin the figurative-language features used in later analyses (Section 4.1). We then report and compare results of Alzheimer's detection models, including both feature-based and transformer-based approaches (Section 4.2). Finally, we analyse how these features, particularly figurative-language indicators, relate to model predictions and reflect patterns of language impairment (Section 4.3).

4.1. Figurative-Language Detection

To assess the reliability of the figurative-language (i.e., MWE and metaphors) indicators introduced in Section 3.2, the LLM-based detection pipeline was evaluated on two benchmark datasets. On the MAGPIE idiom corpus, the model achieved an F1-score of 0.69, showing a slight tendency to under-predict idiomatic usage (see Table 1). For the Metaphor corpus (Bizzoni and Lappin, 2018), performance was comparable (F1 = 0.70) as can be seen in Table 1. These results indicate stable and consistent detection across both tasks, suggesting that the pipeline provides sufficiently accurate outputs for large-scale analysis of figurative language in clinical speech.

	MWE	**Metaphor**
Accuracy	0.70	0.70
Precision	0.70	0.75
Recall	0.68	0.72
F1	0.69	0.69

Table 1: Results of Idiomatic classification and Metaphor detection using Llama 3.

Detection accuracy was slightly higher for MWEs than for metaphors, which is consistent with the broader linguistic complexity of metaphor recognition. Because metaphors are more context-dependent and often novel or creative, they require deeper semantic abstraction and are thus more challenging for an LLM to classify consistently. In contrast, idioms exhibit clearer lexical patterns and more fixed forms, likely facilitating recognition.

A closer inspection of the classification outcomes revealed a conservative bias toward literal interpretation. For example, expressions such as *break the ice* and *lose track*) were sometimes labeled as literal even when a figurative reading was possible.

These results can be compared with the performances reported in Phelps et al. (2024), in which the best score for the MAGPIE dataset was an F1 of 0.90 with GPT-4, while other models could not achieve similar performance: Llama2 (13B) achieved 0.43 while Gemini 1.0 Pro had a 0.72 F1. The evaluation therefore confirms that the LLM-based pipeline produces sufficiently reliable figurative-language features for corpus-level analyses. While individual classifications may be conservative, the aggregated measures capture systematic differences in figurative expression frequency that are relevant for modelling Alzheimer's speech.

Manual inspection revealed two main sources of error. (1) The model often produced false positives, tagging single words as multiword expressions (171 cases involving 75 distinct items). This suggests that lexical boundaries are not properly encoded, and the model may rely excessively on contextual or semantic cues rather than structural ones. (2) Inconsistent identification across similar contexts, observed in 83 cases. For instance, *cookie jar* was detected in "*...and the cookie jar lid is off...*" but missed in "*...the boy is taking cookies out of the cookie jar...*", while *running out of* was recognized in "*...water's running out of the sink...*" but not in "*...water is running out of the faucet...*". These inconsistencies might indicate that the model is sensitive to minor surface variations and may over-rely on contextual cues. The model also identified sequences without idiomatic meaning (e.g., "*I guess*", possibly revealing a confusion between lexical frequency, formulaicity, and non-compositional meaning.

Despite these limitations, the pipeline achieves reliability comparable to previous LLM-based idiomaticity studies and provides a robust foundation for subsequent analyses. The resulting figurative-language features (i.e., P4, M4, P5 and M5) form the basis for the linguistic and modelling analyses presented in Sections 4.2 and 4.3.

4.2. AD Detection

To evaluate how linguistic and contextual representations differentiate AD from control narratives, two

complementary families of models were compared: feature-based classifiers trained on interpretable linguistic variables and transformer-based models with contextual embeddings from BERT. The goal is to determine to what extent traditional linguistic indicators account for diagnostic distinctions and whether contextual representations capture additional, cognitively meaningful information.

Feature-based Models: Table 2 summarizes the results for models trained exclusively on linguistic features. Among these approaches, the Random Forest classifier achieved the highest performance (F1 = 0.716), confirming the robustness of ensemble methods for handling heterogeneous linguistic dimensions. Logistic Regression and Support Vector Machines obtained moderate results (F1 = 0.63), while Gradient Boosting and XGBoost performed slightly lower (F1 = 0.59 – 0.68). These outcomes suggest that surface-level linguistic indicators already provide informative signals of cognitive impairment, though they capture only part of the variation present in spontaneous speech.

Transformer-based Models: Transformer-based classifiers achieved stronger and more consistent results across evaluations. The best configuration, frozen BERT, reached an accuracy of 0.816 and F1 of 0.815. This setting outperformed the feature-based models and the fine-tuned and custom BERT variants. These findings indicate that contextual semantics encoded in BERT's representations provide discriminative information about language impairment even without end-to-end optimization, whereas full fine-tuning offers limited gains under data-constrained clinical conditions. The improvement over feature-based models highlights the capacity of contextual embeddings to integrate lexical, syntactic, and pragmatic cues relevant to Alzheimer-related discourse. Furthermore, we attribute the inferior performance of the fine-tuned BERT models (both standard architecture and modified classification head architecture) to the hyperparameters used, especially since Balagopalan et al. (2020) obtained 0.83 of F1 using fine-tuned BERT.

Model	Accuracy	F1
Random Forest	**0.72**	**0.72**
XGBoost	0.68	0.68
SVM (linear)	0.63	0.63
Gradient Boosting	0.59	0.59
Logistic Regression	0.58	0.58
Frozen BERT	**0.81**	**0.81**
BERT (fine-tuned)	0.68	0.68
Modified BERT	0.59	0.54
(Balagopalan et al., 2020)	**0.83**	**0.82**
(Luz et al., 2020)	0.56	0.58

Table 2: Comparison of model performance on classification task.

Overall, the frozen BERT demonstrated the best balance between generalization and interpretability, surpassing the feature-based Random Forest by approximately 10 percentage points in F1. Compared with state-of-the-art results on the same corpus (F1 = 0.82), the proposed model achieves competitive performance while remaining computationally efficient, as seen in Table 2. The moderate gap between transformer- and feature-based approaches suggests that linguistic features still encode a substantial portion of the information exploited by deep contextual models.

4.3. Linguistic Interpretation of Model Behaviour

We analyse how linguistic and figurative-language patterns differentiate between AD and control narratives, and whether comparable tendencies appear in the model's predictions. The aim is not to explain the internal mechanics of the classifier, but to examine whether the linguistic regularities proposed, particularly those concerning figurative-language use, are also reflected in the model's behaviour. Two complementary analyses were conducted: (1) correlations between linguistic features and diagnostic labels (Section 4.3.1) and (2) correlations between linguistic features and model predictions (Section 4.3.2). The final subsection (4.3.3) compares these structures to assess convergence between human-observed and model-reflected tendencies.

4.3.1. Language Features and Diagnostic Labels

Addressing RQ1, this analysis quantifies how interpretable linguistic indicators vary between AD and control participants. Table 3 presents Spearman correlation for all features. Lexical-diversity measures such as the moving-average TTR (L5) exhibit moderate negative correlations($\rho \approx -0.34$ with the AD label, confirming the well-known reduction of vocabulary variability in AD speech. Readability indices, including the Flesch Reading Ease (R1) and Flesch-Kincaid Grade (R2), correlate positively ($\rho \approx 0.33$), indicating simpler and more predictable syntax in impaired speakers.

Most notably, figurative-language features display clear patterns. The proportion of sentences containing MWE (P4) and metaphors (M4) is consistently lower in the AD group ($\rho \approx -0.29$ and $\rho \approx -0.16$, respectively). These tendencies are illustrated in Figures 3 and 4, where idiomatic and metaphorical densities show lower distributions for AD transcripts. The pattern suggests a preference for literal formulations and a reduced use of figurative or abstract expression, aligning with cognitive-linguistic theories that associate figurative processing with executive and semantic control.

Feature ID	Labels	Predictions
Reading Ease (R1)	0.32	0.31
MWEs proportion (P4)	0.29	0.32
Idioms (P2)	0.21	0.24
Abstractness (A2)	0.20	0.25
Metaphors (M4)	0.19	0.23
Avg. Concreteness (A1)	-0.20	-0.25
Flesch-Kincaid (R2)	-0.27	-0.23
Readability (R5)	-0.27	-0.21
Concreteness Cov. (A3)	-0.28	-0.29

Table 3: Correlation values of features with gold labels and predicted probabilities ($p<0.05$).

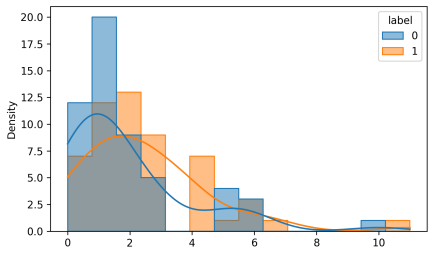

Figure 3: Distribution of idiomatic expressions.

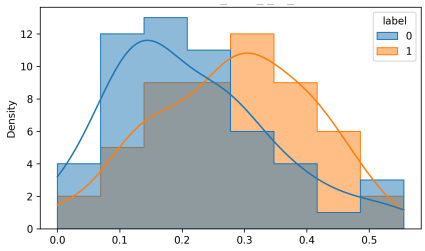

Figure 4: Distribution of metaphorical expressions.

4.3.2. Language Features and Model Predictions

To address RQ2, we examine whether similar linguistic tendencies are reflected in the model's output probabilities. Correlations between BERT-based predictions and the same set of features (Table 3, right column) reveal a broadly similar structure to that observed for the gold labels. These results suggest the classifier are sensitive to linguistic dimensions that also characterize human distinctions between healthy and impaired language.

To estimate the extent to which the model's behavior can be inferred from explicit cues, a Random Forest proxy model was trained to approximate BERT's predicted probabilities using the handcrafted linguistic features as input. The proxy achieved 0.73 accuracy, confirming that a substantial share of the neural model's decisions can be reconstructed from interpretable variables. Feature-importance analysis (Figure 5) highlights lexical diversity, concreteness, and figurative-language density as the strongest predictors, followed by readability indices.

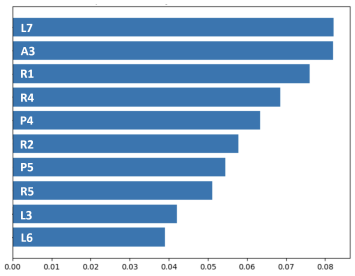

Figure 5: Feature-importance analysis.

4.3.3. Alignment of Human and Model-Reflected Patterns

We compared the two correlation structures (features × labels vs. features × predictions) to assess the convergence between human-observed and model-reflected tendencies. The comparison reveals a moderate but systematic overlap. Lexical diversity (L5), concreteness (A1), and readability (R1-R3) maintain consistent correlation directions in both perspectives, indicating that these dimensions are robust linguistic correlates of cognitive decline captured by both humans and the model.

Figurative-language indicators behave in a particularly informative way. While idiomatic and metaphorical proportions correlate negatively with the diagnostic labels, their associations with the model's predictions are slightly stronger in magnitude. This suggests that the model is even more responsive to the absence of figurative constructions than the human annotation captures, possibly because figurative-language reduction co-occurs with other semantic and syntactic simplifications that the model encodes contextually. In this sense, figurative-language metrics act as a bridge between interpretable linguistic observations and the latent semantic representations learned by the model.

Overall, these results demonstrate linguistic (not algorithmic) alignment. The model's predictive patterns mirror known linguistic characteristics of Alzheimer's speech, particularly reductions in figurative density and lexical variation. Rather than serving as an explainability technique, this analysis offers a linguistic interpretation of model behaviour, showing that the proposed figurative-language features capture aspects of the communicative simplification associated with cognitive decline and are partly echoed in the model's decision tendencies.

5. Conclusion

This work investigated how figurative-language indicators contribute to Alzheimer's Disease detec-

tion in spontaneous speech and whether a neural model reflects comparable linguistic tendencies. We provided a linguistically grounded view of language impairment by integrating LLM-derived measures of idiomatic and metaphorical usage with lexical, syntactic, and readability features. Results showed that AD participants used fewer figurative expressions, displayed lower lexical diversity, and produced more concrete language, which may be partly linked to deficits in abstraction and semantic control. However, more investigation is needed. To provide transparency and illustrate the behaviour of our extraction method, some examples of correctly and incorrectly identified figurative expressions are provided in Appendix A. Moreover, BERT-based predictions correlated with these linguistic dimensions, suggesting that contextual embeddings implicitly capture cues linked to cognitive decline.

The alignment between interpretable linguistic indicators and model behaviour indicates that neural representations encode meaningful patterns of figurative and literal expression without explicit modelling. This highlights the potential of figurative-language metrics as bridges between human-interpretable and neural features. While our findings currently focus on English-speaking participants, cross-linguistic generalization remains future work. Future work will expand this to larger and multilingual datasets, examining cross-linguistic variability and multimodal cues in cognitive impairment.

6. Acknowledgements

This study was partly funded by AIM-Health project, UKRI (grant MR/U506734/1), and by the São Paulo Research Foundation (FAPESP, Brazil, grants 2024/10233-7 and 2025/05422-8). This research was also funded by the National Institute for Health and Care Research (NIHR) Exeter Biomedical Research Centre (BRC) (NIHR203320).[3]

7. Bibliographical References

Martina Amanzio, Giuliano Geminiani, Daniela Leotta, and Stefano Cappa. 2008. Metaphor comprehension in alzheimer's disease: Novelty matters. *Brain and language*, 107(1):1–10.

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

Valentina Bambini, Claudio Gentili, Emiliano Ricciardi, Pier Marco Bertinetto, and Pietro Pietrini. 2011. Decomposing metaphor processing at the cognitive and neural level through functional magnetic resonance imaging. *Brain research bulletin*, 86(3-4):203–216.

Yuri Bizzoni and Shalom Lappin. 2018. Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the workshop on figurative language processing*, pages 45–55.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Eileen R Cardillo, Christine E Watson, Gwenda L Schmidt, Alexander Kranjec, and Anjan Chatterjee. 2012. From novel to familiar: tuning the brain for metaphors. *Neuroimage*, 59(4):3212–3221.

Michael A. Covington and John D. McFall. 2006. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of Quantitative Linguistics*, 13(2):93–109.

Sofia De la Fuente Garcia, Craig W Ritchie, and Saturnino Luz. 2020. Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: a systematic review. *Journal of Alzheimer's Disease*, 78(4):1547–1574.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Mireia Farrús and Joan Codina-Filbà. 2020. Combining prosodic, voice quality and lexical features to automatically detect alzheimer's disease. *arXiv preprint arXiv:2011.09272*.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Malcolm Forbes, Mojtaba Lotfaliany, Mohammadreza Mohebbi, Charles F Reynolds, Robyn L Woods, Suzanne Orchard, Trevor Chong, Bruno Agustini, Adrienne O'Neil, Joanne Ryan, et al. 2024. Depressive symptoms and cognitive decline in older adults. *International psychogeriatrics*, 36(11):1039–1050.

[3]The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Katrina E Forbes-McKay and Annalena Venneri. 2005. Detecting subtle spontaneous language decline in early alzheimer's disease with a picture description task. *Neurological sciences*, 26(4):243–254.

Bruce Fraser. 1999. Lexical cohesion and discourse. *Topics in Cognitive Science*, 1(1):101–120.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2015. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's disease*, 49(2):407–422.

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

Mengshi Ge, Rui Mao, and Erik Cambria. 2023. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artificial Intelligence Review*, 56(Suppl 2):1829–1895.

Sam Glucksberg. 2003. The psycholinguistics of metaphor. *Trends in cognitive sciences*, 7(2):92–96.

R. Gunning. 1952. The technique of clear writing. *McGraw-Hill*.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. Magpie: A large corpus of potentially idiomatic expressions. In *12th Language Resources and Evaluation Conference: LREC 2020*, pages 279–287. European Language Resources Association (ELRA).

Elias Hossain, Rajib Rana, Niall Higgins, Jeffrey Soar, Prabal Datta Barua, Anthony R Pisani, and Kathryn Turner. 2023. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. *Computers in biology and medicine*, 155:106649.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2023. Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744.

J.P. Kincaid, R.P. Fishburne, R.L. Rogers, and B.S. Chissom. 1975. Derivation of readability formulas for navy enlisted personnel. Technical report, Naval Technical Training Command.

Zoltán Kövecses. 2005. *Metaphor in culture: Universality and variation*. Cambridge university press.

Zoltán Kövecses. 2010. Metaphor and culture. *Acta Universitatis Sapientiae, Philologica*, 2(2):197–220.

Huiyuan Lai and Malvina Nissim. 2024. A survey on automatic generation of figurative language: From rule-based systems to large language models. *ACM Computing Surveys*, 56(10):1–34.

Llama Team, AI @ Meta. 2024. The llama 3 herd of models.

Xiaofei Lu. 2012. The automatic measurement of syntactic complexity in child language. *International Journal of Corpus Linguistics*, 17(3):357–374.

Saturnino Luz, Faheem Haider, Davida Fromm, and Brian MacWhinney. 2021. Troubadour at the adress challenge 2021: Acoustic and linguistic features fusion for alzheimer's dementia recognition through spontaneous speech. In *Proceedings of Interspeech 2021*, pages 3820–3824, Brno, Czech Republic. ISCA.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. In *Proceedings of Interspeech 2020*, pages 2172–2176. ISCA.

G.H. McLaughlin. 1969. Smog grading: A new readability formula. *Journal of Reading*, 12(8):639–646.

Amit Meghanani, C S Anoop, and A G Ramakrishnan. 2021. Recognition of alzheimer's dementia from the transcriptions of spontaneous speech using fasttext and cnn models. *Frontiers in Computer Science*, 3:624558.

Bahman Mirheidari, Daniel Blackburn, Traci Walker, Annalena Venneri, Markus Reuber, and Heidi Christensen. 2018. Detecting signs of dementia using word vector representations. In *Interspeech*, pages 1893–1897.

Susanna Olivero. 2024. *Figurative language understanding based on large language models*. Ph.D. thesis, Politecnico di Torino.

Costanza Papagno. 2001. Comprehension of metaphors and idioms in patients with alzheimer's disease: A longitudinal study. *Brain*, 124(7):1450–1460.

Felipe Paula, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. Similarity measures for the detection of clinical conditions with verbal fluency tasks. In *Proceedings of the 2018 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 231–235.

Dylan Phelps, Thomas Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the times: Evaluating the use of large language models for idiomaticity detection. *arXiv preprint arXiv:2405.09279*.

Cecilia Rassiga, Federica Lucchelli, Franca Crippa, and Costanza Papagno. 2009. Ambiguous idiom comprehension in alzheimer's disease. *Journal of clinical and experimental neuropsychology*, 31(4):402–411.

Carlos Roncero and Roberto G de Almeida. 2014. The importance of being apt: metaphor comprehension in alzheimer's disease. *Frontiers in Human Neuroscience*, 8:973.

Utkarsh Sarawgi, Wazeer Zulfikar, Nouran Soliman, and Pattie Maes. 2020. Multimodal inductive transfer learning for detection of alzheimer's dementia and its severity. *arXiv preprint arXiv:2009.00700*.

Ekaterina Shutova. 2010. Models of metaphor in nlp. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 688–697.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4673–4686.

László Tóth, Gábor Gosztolya, Veronika Vincze, Ildikó Hoffmann, Gréta Szatlóczki, Edit Biró, Fruzsina Zsura, Magdolna Pákáski, and János Kálmán. 2015. Automatic detection of mild cognitive impairment from spontaneous speech using asr.

A. Examples of Automatically Detected Metaphorical Expressions

The metaphor detection procedure occasionally produced literal descriptions or transcription artefacts. Tables 4 and 5 illustrate examples of correctly identified figurative expressions and common false positives extracted from the dataset.

Expression	Comment
"it's going to be like total catastrophe"	metaphorical exaggeration describing the situation
"that's a mess"	evaluative metaphor describing disorder in the scene
"the stool's going to collapse"	metaphorical framing of instability
"the water's overrunning"	figurative intensification of the overflow

Table 4: Examples of correctly detected metaphorical expressions.

Expression	Issue
"the boy is reaching for a cookie"	literal description of an action
"standing on a stool"	literal physical activity
"the wind is blowing the curtains"	literal event in the image
"Invalid"	transcription or extraction artefact

Table 5: Examples of incorrect metaphor detections (false positives).