Integração Multimodal: Um Mapeamento Sistemático sobre Eye Tracking, TROG-2 e Machine Learning na Avaliação da Afasia

Jonas Cesconetto†
UNIVALI – Universidade do Vale do Itajaí, SC, Brasil
jonascesconetto@edu.univali.br

Alejandro Ramirez
UNIVALI – Universidade do Vale do Itajaí, SC, Brasil
ramirez@univali.br

ABSTRACT

Aphasia is an acquired language disorder that affects comprehension and/or production. This study expands previous systematic mappings by combining a theoretical synthesis with an experimental proof-of-concept, integrating eye tracking and TROG-2 for the empirical assessment of linguistic comprehension. The systematic mapping across eight databases (2020–2025), conducted under PRISMA guidelines, identified 28 studies distributed among five technological axes, highlighting the absence of full multimodal integration (Eye Tracking + TROG-2 + Machine Learning) in the literature. To address this gap, an exploratory analysis was conducted using a multimodal dataset collected from participants with and without aphasia performing TROG-2 tasks. Quantitative metrics derived from the recordings revealed consistent differences between groups, including longer response times (+81%) and a higher number of saccades (+67%) in aphasic participants ($p \approx 0.05$), indicating increased cognitive load and reduced visual efficiency during syntactic processing. The Mann-Whitney U test confirmed statistically significant intergroup differences for global measures of response time and visual search frequency. These findings empirically validate theoretical predictions of visual–linguistic interaction in aphasia and demonstrate the feasibility of combining Eye Tracking and TROG-2 in Portuguese-speaking populations. The results reinforce the potential of this multimodal approach as a foundation for future machine learning models capable of identifying oculomotor biomarkers of linguistic deficits, advancing precision medicine and the development of intelligent assistive systems for language rehabilitation.

KEYWORDS

Eye Tracking; TROG-2; Machine Learning; Multimodal Integration; Assistive Technology.

1 INTRODUÇÃO

A afasia é um distúrbio adquirido da linguagem que afeta compreensão e/ou produção, resultante de lesões cerebrais (pós-AVE) ou processos neurodegenerativos, como na Afasia Progressiva Primária (APP). Medidas comportamentais tradicionais (p.ex., acerto/erro) capturam apenas parte do fenômeno, motivando o uso de Eye Tracking para observar o processamento on-line em escala de milissegundos, revelando padrões de fixação e atenção visual durante tarefas linguísticas ([7]; [9]).

Como instrumentos de testes padronizados, o TROG-2 (Test for Reception of Grammar) – e sua versão brasileira TROG2-Br – mede compreensão morfossintática por blocos de itens com complexidade crescente ([2]).

Em paralelo, Inteligência Artificial (IA) e Machine Learning (ML) têm avançado na classificação de variantes de APP a partir de fala conectada, EEG e neuroimagem, alcançando concordância elevada com diagnósticos clínicos ([3]).

Entretanto, notou-se que não há estudos que integrem simultaneamente Afasia + Eye Tracking + TROG-2 + IA/ML na literatura. Este trabalho sistematiza o estado da arte apresentando alguns eixos e delineia fronteiras de pesquisa para integrar as três tecnologias. ([2]; [14]).

Este trabalho é caracterizado como um *mapeamento sistemático* da literatura, uma vez que, diante da escassez de estudos que integrem simultaneamente Eye Tracking, TROG-2 e Machine Learning, adotou-se uma estratégia de busca orientada por proximidade e organização dos achados em eixos temáticos. Essa abordagem é apropriada quando o objetivo principal é mapear o campo, identificar tendências, lacunas e oportunidades de pesquisa, em vez de sintetizar evidências comparáveis para responder a uma única questão com julgamento conclusivo.

2 METODOLOGIA

O presente estudo adota a abordagem de mapeamento sistemático, diferentemente de uma revisão sistemática tradicional. Enquanto revisões sistemáticas visam sintetizar evidências para responder questões clínicas específicas com protocolo rígido de avaliação de qualidade, o mapeamento sistemático tem como objetivo identificar e estruturar o escopo da literatura existente em determinado campo, mapeando lacunas e oportunidades de pesquisa. Essa escolha justifica-se pela ausência de estudos que integrem simultaneamente as três tecnologias investigadas, tornando o mapeamento a estratégia mais adequada para caracterizar o estado da arte e delimitar fronteiras de pesquisa.

2.1 Estratégia de Busca Sistemática

A busca sistemática foi conduzida em setembro de 2025 sob os princípios PRISMA, em oito bases de dados: Google Scholar, ScienceDirect, IEEE Xplore, MDPI Sensors, Wiley Online Library, SciELO, Springer Link e ACM Digital Library. A string principal combinava os termos: *((Afasia OR Aphasia) AND (rastreamento ocular OR "eye tracking") AND (TROG-2 OR TROG2 OR "Test for Reception of Grammar") AND (IA OR AI OR "Inteligência Artificial" OR "Artificial Intelligence") AND (Machine Learning OR "Aprendizado de Máquina") AND (2020-2025))*

2.2 Estratégia de Busca por Proximidade

Dado que a string completa não retornou estudos integrando todos os componentes simultaneamente, aplicou-se estratégia de busca por proximidade para mapear o estado da arte em cada eixo tecnológico isoladamente. As strings utilizadas foram derivadas da string principal por supressão progressiva dos componentes ausentes em cada eixo:

- Eixo 1: Eye tracking + Afasia (sem TROG-2 ou ML): *(Aphasia OR Afasia) AND ("eye tracking" OR "rastreamento ocular")*
- Eixo 2: TROG-2 + Afasia (sem eye tracking ou ML): *(Aphasia OR Afasia) AND (TROG-2 OR TROG2 OR "Test for Reception of Grammar")*
- Eixo 3: ML/IA + Afasia (sem eye tracking ou TROG-2): *(Aphasia OR Afasia) AND ("Machine Learning" OR "Aprendizado de Máquina" OR "Artificial Intelligence" OR "Inteligência Artificial")*
- Eixo 4: Eye tracking + TROG-2 + Afasia (sem ML): *(Aphasia OR Afasia) AND ("eye tracking" OR "rastreamento ocular") AND (TROG-2 OR TROG2 OR "Test for Reception of Grammar")*
- Eixo 5: ML + TROG-2 + Afasia (sem eye tracking): *(Aphasia OR Afasia) AND ("Machine Learning" OR "Aprendizado de Máquina") AND (TROG-2 OR TROG2 OR "Test for Reception of Grammar")*

2.3 Critérios de Seleção

Foram incluídos artigos publicados entre 1º de janeiro de 2020 e 31 de dezembro de 2025, redigidos em inglês ou português, que envolvessem pelo menos uma das três tecnologias-chave investigadas aplicada à afasia ou a distúrbios de linguagem relacionados. Quanto ao tipo de publicação, foram aceitos artigos revisados por pares, pré-prints disponibilizados em repositórios reconhecidos e datasets públicos devidamente documentados, desde que apresentassem metodologia claramente descrita.

Foram excluídos estudos anteriores a 2020, trabalhos sem acesso ao texto completo e pesquisas de natureza exclusivamente teórica, sem apresentação de dados empíricos. Também foram descartados resumos expandidos e anais de eventos que não contemplassem metodologia completa, bem como estudos voltados a distúrbios de linguagem sem relação direta com a afasia.

2.4 Processo de Triagem

A triagem seguiu quatro etapas sequenciais: execução das strings com remoção de duplicatas; avaliação independente de títulos e resumos por dois revisores (J.C. e A.R.G.R.), com discordâncias resolvidas por consenso; leitura completa dos pré-selecionados com aplicação rigorosa dos critérios; e busca manual por snowballing nas referências dos artigos incluídos.

2.6 Extração de Dados

Para cada estudo incluído, foram extraídos:

- Características metodológicas: Tamanho amostral, tipo(s) de afasia, instrumento de avaliação linguística, tecnologia de eye tracking (quando aplicável), algoritmos de ML (quando aplicável), métricas de desempenho.
- Resultados principais: Achados sobre diferenças entre grupos, acurácia de classificação/predição, validação clínica, limitações reportadas.

2.7 Análise e Síntese

Realizou-se síntese narrativa estruturada dos achados, organizando os resultados por eixo tecnológico, tipo de afasia e fase de validação. A análise identificou padrões, lacunas e oportunidades de integração entre os três componentes tecnológicos.

3 MAPEAMENTO DO ESTADO DA ARTE

3.1 Resultados da String de Busca

A busca sistemática inicial com a string completa não identificou estudos que integrassem simultaneamente eye tracking, TROG-2 e machine learning na avaliação da afasia. Este achado representa a lacuna crítica identificada por esta revisão.

A estratégia de busca por proximidade identificou 68 artigos potencialmente relevantes após remoção de duplicatas. Após triagem por título, resumo e texto completo, 28 estudos foram incluídos para análise, distribuídos conforme a Tabela 1:

Eixo	Descrição	N.	Principal Achado
1	Eye tracking + Afasia	16	Viabilidade de eye tracking web; tamanhos amostrais pequenos
2	TROG-2 + Afasia	8	Validação TROG2-Br; boa sensibilidade para APP
3	ML/IA + Afasia	12	Acurácias 73–100%; foco em APP
4	Eye tracking + TROG-2 + Afasia	3	Dataset brasileiro público com 21 participantes
5	ML + TROG-2 + Afasia	1	Prova de conceito; limitações metodológicas

Tabela 1 – Distribuição dos Estudos por Eixo Tecnológico

3.2 Síntese por Eixo Tecnológico

Eixo 1: Eye Tracking + Afasia (sem TROG-2 ou ML)

[9] conduziram revisão de escopo com 16 estudos sobre eye tracking e compreensão linguística na afasia, cobrindo paradigmas visual world e leitura de sentenças. O principal achado limitante foi o tamanho amostral reduzido nos grupos clínicos (média = 9 participantes afásicos, variação de 4 a 16), comprometendo a generalizabilidade dos resultados.

[14] compararam eye tracking laboratorial (EyeLink) com rastreamento via webcam (WebGazer.js) em 16 pacientes afásicos e 16 controles. O estudo demonstrou viabilidade do rastreamento online para detectar diferenças entre grupos, representando avanço significativo para democratização da tecnologia — embora os autores ressaltem que validações adicionais ainda sejam necessárias.

Eixo 2: TROG-2 + Afasia (sem eye tracking ou ML)

[2] publicaram estudo de validação do TROG2-Br (versão brasileira do Test for Reception of Grammar) para Afasia Progressiva Primária. Os objetivos foram testar a aplicabilidade da ferramenta para detectar déficits morfossintáticos em pacientes com APP, investigar associações entre desempenho no teste e variáveis sociodemográficas e clínicas, e caracterizar o desempenho das três variantes principais de APP e APP mista. A amostra compreendeu 74 participantes cognitivamente saudáveis (54 mulheres, idade 60+) e 34 indivíduos diagnosticados com APP, sendo 12 com variante semântica (APP-S), 6 com variante não-fluente/agramática (APP-NF/A), 5 com variante logopênica (APP-L) e 11 com APP mista (APP-Mx).

Os resultados em controles mostraram que os escores por bloco foram significativamente correlacionados com anos de educação formal (r de Spearman = 0,33, p = 0,004), mas não com idade. A mediana de blocos corretos no grupo controle foi 15, com escores para os percentis 10, 25, 75 e 90 de respectivamente 10, 13, 18 e 20 blocos. O padrão geral de resposta de adultos mais velhos no TROG2-Br foi de erros esporádicos, caracterizados por dar resposta incorreta para apenas uma sentença mas responder corretamente as três outras sentenças do bloco, sugestivo de dificuldades de processamento ao invés de déficits morfossintáticos genuínos.

Controles apresentaram desempenho significativamente superior no TROG2-Br comparados a indivíduos com APP (Mann-Whitney U = 402.0, p = 0.000). Trinta de 34 pacientes com APP apresentaram escores abaixo da mediana dos controles. O padrão de erros apontou para dificuldades leves de processamento cognitivo geral (atenção, memória de trabalho) nos controles, enquanto na APP os tipos de erros apontaram para déficits de processamento e morfossintáticos. Não foram encontradas diferenças significativas entre os quatro subgrupos de APP tanto considerando itens corretos (Kruskal-Wallis H=3.918, p=0.270) quanto número de blocos passados (Kruskal-Wallis H=2.724, p=0.436). Entretanto, análise qualitativa revelou que APP-S foi qualitativamente mais similar aos controles, com dificuldades de processamento e menor porcentagem de erros morfossintáticos. Apenas APP-S apresentou marginalmente menor porcentagem de erros consistentes e sistemáticos comparado a erros esporádicos e aleatórios. APP-NF/A apresentou erros apontando para déficits morfossintáticos evidentes, embora um paciente apresentasse compreensão de sentenças preservada. Todos os pacientes com APP-L apresentaram desempenho abaixo da mediana dos controles, com déficits relacionados à memória de trabalho fonológica. APP-Mx apresentou os escores mais baixos na amostra.

O TROG2-Br apresentou boa consistência interna (α de Cronbach = 0,87) e validade concorrente significativa com a versão brasileira validada do Token Test (r de Spearman = 0,765, p < 0,000). Os autores propuseram versão reduzida de 5 blocos (T-M-Q-K-L) com consistência interna de α = 0,82, útil para testar populações com atenção sustentada limitada ou em contextos de restrições de tempo. As estruturas sintáticas avaliadas por estes blocos incluem sentenças não-canônicas, orações relativas de objeto e sentenças com cláusulas encaixadas no centro.

Eixo 3: ML/IA + Afasia (sem eye tracking ou TROG-2)

[13] apresentaram modelo de machine learning baseado em redes neurais profundas (DNN) para subtipagem de pacientes com APP em três variantes principais (APP-NF, APP-S e APP-L), usando informações acústicas e linguísticas combinadas elicitadas automaticamente via análise acústica e linguística. O desempenho da DNN foi comparado à acurácia de classificação de Random Forests, Support Vector Machines e Decision Trees, bem como às classificações de clínicos especialistas. Os dados utilizados foram produções de fala conectada via tarefa simples e amplamente utilizada de descrição de figura: a descrição Cookie Theft do Boston Diagnostic Aphasia Examination. O modelo DNN superou os outros modelos de machine learning bem como as classificações de clínicos especialistas com 80% de acurácia de classificação. Importante destacar que 90% dos pacientes com APP-NF e 95% dos pacientes com APP-L foram identificados corretamente, fornecendo subtipagem confiável destes pacientes em suas variantes de APP correspondentes. Os autores concluíram que marcadores combinados de fala e linguagem de produções de fala conectada podem informar subtipagem de variantes em pacientes com APP, e que a abordagem automatizada end-to-end de machine learning pode permitir clínicos e pesquisadores fornecerem classificação fácil, rápida e de baixo custo de pacientes com APP.

[3] investigaram utilidade diagnóstica de biomarcador baseado em eletroencefalografia (EEG) em conjunto com machine learning para diagnóstico diferencial na APP. Indivíduos com APP semântica, logopênica ou não-fluente/agramática e controles saudáveis (n=10 por grupo) ouviram narrativa contínua enquanto respostas de EEG eram registradas. O envelope de fala e características linguísticas representando processos linguísticos centrais foram extraídos da fala narrativa e modelagem de função de resposta temporal (TRF) foi usada para estimar as respostas neurais a estas características. Os pesos beta de TRF resultantes

para o canal Cz foram usados como entrada para algoritmos de machine learning para classificação de APP vs. controles saudáveis, classificação três vias por subtipo de APP, classificação de um único subtipo de APP relativo aos outros dois, e classificação par a par por subtipo de APP. Os F1 scores foram mais altos para as tarefas de classificação par a par (F1's de 0,73 a 0,74), com classificação melhor que o acaso em todas as tarefas. Análises adicionais determinaram que os pesos beta de TRF melhoraram significativamente a classificação sobre formas de onda de EEG pré-processadas sozinhas para todas as tarefas exceto uma (APP vs. controles saudáveis). Os achados preliminares demonstram a utilidade potencial desta abordagem para diagnóstico diferencial de APP, garantindo investigação adicional.

[8] utilizaram técnicas de machine learning para prever cada variante de APP segundo resultados de avaliação linguística, utilizando algoritmo previamente desenvolvido baseado em análise de cluster hierárquico aglomerativo com método de ligação de Ward baseado em metabolismo regional de PET com FDG. Sessenta e oito pacientes com APP em estágios iniciais da doença e 20 controles saudáveis foram avaliados com protocolo abrangente de linguagem e cognição. Cinco variantes foram encontradas, com ambas as variantes não-fluente e logopênica sendo divididas em 2 subtipos. Os algoritmos de machine learning usando dados de testes de linguagem foram capazes de prever cada uma das 5 variantes de APP com grau relativamente alto de acurácia, permitindo possibilidade de diagnóstico automatizado auxiliado por máquina de variantes de APP. O estudo suporta a existência de 5 variantes de APP, mostrando algumas diferenças em características de linguagem e imagem PET com FDG, sugerindo que classificações futuras podem necessitar refinamento adicional para capturar a variabilidade clínica completa.

[12] utilizaram algoritmo de machine learning conhecido como SuStaIn (Subtype and Stage Inference) para descobrir perfis de progressão neuroanatômica baseados em dados de subtipos de APP e realizaram análise aprofundada subtipo-fenótipo para caracterizar a heterogeneidade da APP. A amostra incluiu 270 pacientes com diagnóstico de APP (APP-S, n=94; APP-NF, n=109; APP-L, n=51; APP-nos, n=16), com dados coletados de participantes inscritos em cinco estudos longitudinais no Dementia Research Centre, UCL, entre 1993 e 2020. O algoritmo SuStaIn combina modelagem de progressão de doença com machine learning não-supervisionado para identificar subgrupos de pacientes baseados em agrupamento em padrões similares de atrofia. Embora a variante semântica tenha perfil neuroanatômico claro, as variantes não-fluente/agramática e logopênica são difíceis de discriminar a partir de neuroimagem. O estudo identificou subtipos neuroanatômicos data-driven que não correspondem perfeitamente às classificações clínicas tradicionais, sugerindo que a heterogeneidade da APP pode ser melhor capturada através de abordagens baseadas em dados de neuroimagem.

Um estudo de 2024 explorou LLMs pré-treinados na classificação de afasias, usando valores de surpresa como variáveis preditoras em modelos de Decision Tree, Random Forest, Gradient Boosting e SVM. O SVM com embeddings do RoBERTa alcançou a maior acurácia, com os modelos baseados em transformers superando o acaso em acurácia balanceada, embora classificadores tradicionais com embeddings contextuais permanecessem competitivos..

Eixo 4: Eye Tracking + TROG-2 + Afasia (sem ML)

[1] desenvolveram dataset público consistindo de gravações de eye tracking obtidas de quinze pacientes com Afasia de Broca e seis indivíduos saudáveis. Os voluntários afásicos tinham boa função auditiva e ocular, permitindo uso de computador para comunicação. Os dados de participantes afásicos e saudáveis foram registrados uma vez em duas Clínicas de Fonoaudiologia parceiras localizadas na UNIVALI e UFSC. O grupo experimental (EG) teve idade média de 58 anos, enquanto o grupo controle (CG) teve idade média de 49 anos. Todos os participantes eram falantes nativos de português brasileiro com acuidade visual normal ou corrigida relatada.

O equipamento selecionado foi o PCEye Mini, com taxa de captura de 60 Hz, taxa de fluxo de dados de 30 frames por segundo, e valores de precisão e acurácia menores que 1,9° e 0,4°, respectivamente. O TROG-2, primeiro desenvolvido em inglês, foi culturalmente traduzido para o português brasileiro e utilizado neste trabalho porque permite avaliação abrangente da compreensão auditória e inclui variedade relativamente ampla de tipos de sentenças. Neste trabalho, oito blocos do TROG-2 foram considerados: três blocos de triagem variando em extensão (blocos A, D e F); dois com sentenças gramaticalmente simples (C e E); e três com sentenças gramaticalmente complexas (K, Q e S). Sentenças complexas desviam da ordem canônica e indivíduos afetados por afasia enfrentam mais desafios ao lidar com elas.

O desenho experimental envolveu trinta e duas sessões de gravação por participante, cada uma correspondendo à execução da segunda versão do TROG-2 clinicamente validado. Os estímulos visuais foram digitalizados e exibidos em computador em intervalos de tempo específicos, cuidadosamente sincronizados com pistas auditórias. Neste estudo, os participantes visualizaram imagens enquanto ouviam sentenças selecionadas. Uma cruz de fixação apareceu no centro da tela por 2 segundos antes de exibir o estímulo visual por 30 segundos. Subsequentemente, um prompt auditório foi apresentado por 4 segundos, começando 500 milissegundos após o estímulo visual terminar. Durante cada tentativa, os participantes foram requeridos a escolher a imagem que correspondesse ao prompt auditório.

Como resultado, este trabalho fornece arquivo em formato CSV consistindo de cinco colunas de dados separados por vírgulas: a tecla pressionada pelo participante, o timestamp do registro de eye-tracking, as coordenadas cartesianas do olhar no display, e finalmente o estímulo visual apresentado. As métricas foram extraídas via script MATLAB projetado para recuperar, processar e tabular os dados registrados no arquivo CSV. A saída do script consiste em dois arquivos de texto (.txt) para cada grupo, onde cada linha representa os dados de cada participante e métricas médias

específicas são reportadas após as oito colunas iniciais, apresentando acurácia média, tempo de resposta médio e número médio de fixações.

Estudos prévios sugerem que indivíduos com afasia têm menor acurácia que indivíduos não-afásicos. No estudo conduzido nas Clínicas de Fonoaudiologia, o Grupo Experimental mostra consistentemente menor acurácia média em todos os blocos do TROG-2. Os blocos contendo sentenças gramaticalmente mais simples (A, C, D, E) demonstram maior acurácia dentro do EG comparados às sentenças mais complexas (F, K, Q, S). As métricas observadas no CG seguem tendência similar, embora em maior extensão que no EG. A acurácia deteriora com sentenças gramaticalmente mais complexas.

A análise quantitativa das métricas de eye-tracking foi realizada por estatística descritiva. O teste de hipótese Kruskal-Wallis foi usado para acurácia, tempo de reação e número de fixações, aplicado às médias das métricas. O valor K-W para acurácia foi 7,61 ($p < 0,0058$), com valor de referência sendo 3,84 (para nível de significância alfa = 0,05). O tempo de resposta retornou valor de 9,28 ($p < 0,0023$), portanto ainda mais distante do valor crítico (3,84), enquanto o número de fixações foi a métrica com menor diferença estatística de 5,83 ($p < 0,0157$). Assim, aplicar o teste K-W refuta a hipótese nula, confirmando que os grupos têm desempenhos estatisticamente distintos. Outro fato importante evidenciado pelos gráficos é a maior dispersão dos dados do EG quando comparado ao CG.
Os autores concluem que testes de compreensão auditória são frequentemente pareados com eye tracking para medir como participantes com distúrbios neurológicos ouvem sentenças faladas e acompanham visualmente, fornecendo dados sobre tempo de reação, foco e compreensão. No entanto, não existe um único dataset público bem conhecido que combine explicitamente estes elementos: eye-tracking, afasia e compreensão auditória. Neste trabalho, o dataset bruto estará disponível para uso público ou propósitos de pesquisa mais amplos, constituindo recurso valioso para avançar terapia personalizada e melhor compreensão dos mecanismos subjacentes à afasia.

[10] realizaram estudo preliminar investigando o uso de técnicas de eye tracking ao avaliar idosos cognitivamente saudáveis brasileiros usando o TROG-2. O estudo apresentou revisão de literatura para definir metodologia para condução do estudo e analisou o uso de eye tracking ao avaliar compreensão de sentenças em idosos. Os estudos reunidos na revisão de literatura forneceram subsídios para definir o processo de triagem, selecionar estímulos do TROG-2 e definir métricas de interesse usando eye tracking. Esta pesquisa foi fundamental para estabelecer parâmetros normativos para a população brasileira antes de aplicar a metodologia em populações clínicas.

[11] desenvolveram novo recurso para investigar processamento auditivo de sentenças por indivíduos afásicos no contexto brasileiro utilizando eye tracking. A solução foi baseada na aplicação da segunda versão do método clinicamente validado conhecido como Test for Reception of Grammar (TROG-2). Um piloto foi conduzido por equipe multidisciplinar, onde quatro indivíduos com afasia e quatro indivíduos sem afasia participaram. A análise de dados mostrou que eye tracking combinado com TROG-2 fornece informações complementares de insights individuais que levariam a programas de reabilitação mais personalizados e assertivos.

Eixo 5: ML + TROG-2 + Afasia (sem eye tracking)

[4] apresentaram trabalho com objetivo de classificar tipos de APP a partir dos resultados de indivíduos diagnosticados com APP no teste TROG-Br. A base de dados utilizada foi disponibilizada no trabalho de [2], compreendendo dados referentes a 23 pessoas diagnosticadas com variantes de APP: 5 com variante logopênica (L), 6 com variante agramática/não-fluente (NF/A) e 12 com variante semântica (S), além de 74 controles saudáveis. No conjunto de dados, para cada indivíduo são apresentados o diagnóstico da variante de APP, a avaliação do teste TROG2-Br em seus 20 blocos (de A a T) indicando o número de erros ou P para cada bloco correto (100% de acertos), o somatório do número de erros e o somatório dos blocos corretos.

Na fase de pré-processamento foram imputadas as linhas referentes a APP Mx (APP não identificada), por não ser alvo da classificação. Também foram substituídas as ocorrências de P por 0, indicando que não houve erros no respectivo bloco. Devido ao desbalanceamento do dataset resultante, contendo um total de 23 instâncias (5 da classe L, 6 da classe NF/A e 12 da classe S), foi aplicado o método de sobre amostragem SMOTE (Synthetic Minority Oversampling Technique). O SMOTE adiciona instâncias sintéticas às classes minoritárias, agrupando-as em torno das instâncias reais destas classes, reduzindo os problemas decorrentes do desequilíbrio de classes e melhorando o desempenho na classificação das classes minoritárias.

Os algoritmos de classificação utilizados foram Decision Tree (árvore de decisão), kNN (k vizinhos mais próximos), Naive Bayes e SVM (Support Vector Machine). As fases de pré-processamento, treinamento e teste dos modelos foram feitas utilizando a ferramenta Orange Data Mining. Os algoritmos Decision Tree e SVM apresentaram melhor performance comparados ao Naive Bayes e kNN, com o Decision Tree apresentando o melhor resultado com acurácia de 75% e F1-Score de 74,1%. A análise da matriz de confusão mostrou que ambos os modelos tiveram mais facilidade em classificar instâncias da classe L, classificando de forma similar instâncias das classes NF/A e S.

Uma das vantagens da Decision Tree é que elas são facilmente interpretáveis. A árvore de decisão gerada permitiu extrair regras para classificação correta de cada classe. Para a classe S, isto acontece quando o valor do atributo Q for ≤ 1 e o valor $M \leq 0$, classificando corretamente 100% das amostras. Para a classe NF/A, quando o valor de $Q \leq 1$ e o valor de $M > 0$, classific ando corretamente 80% das amostras. Para a classe L, quando Q for > 1, $K \leq 2$ e $H > 1$, classificando corretamente 85,7% das amostras. Os blocos Q (subject predicative adjective), M, K (reversible passive

voice) e H emergiram como mais discriminativos para diferenciação entre as variantes.

Os autores concluíram que os resultados são promissores e mostram que é possível classificar APP a partir da pontuação dos indivíduos no teste TROG-Br. Entretanto, reconhecem que se faz necessário um maior número de dados de treinamento para apresentar melhores resultados. Salientam que o TROG deve ser combinado com outros instrumentos para avaliar a afasia e assim obter resultados mais precisos. Como trabalhos futuros, consideram aplicar novamente os experimentos sobre uma base de dados com maior número de observações quando disponíveis, aplicar os experimentos sobre os resultados no teste TROG no idioma inglês onde o teste foi originalmente concebido, e utilizar outros algoritmos de classificação para análise de desempenho.

A Tabela 2 apresenta síntese comparativa dos principais estudos incluídos, organizados por eixo tecnológico, destacando características metodológicas centrais, principais resultados, limitações críticas e contribuições para o campo.

4 FRONTEIRAS A SEREM EXPLORADAS

A análise sistemática apresentada na Seção 3 revelou avanços significativos em cada eixo tecnológico individualmente, mas evidenciou lacuna crítica: a ausência de integração completa entre eye tracking, TROG-2 e machine learning na avaliação da afasia. Esta seção delineia as principais fronteiras científicas e tecnológicas que emergem desta revisão, organizadas em seis domínios estratégicos que podem transformar fundamentalmente a avaliação e tratamento da afasia nos próximos anos.

4.1 Integração Multimodal Completa: A Fronteira Central

A principal fronteira identificada por esta revisão é o desenvolvimento de sistema que integre simultaneamente as três tecnologias, criando ecossistema multimodal de avaliação. A arquitetura proposta compreenderia quatro camadas funcionais:

A camada de aquisição deveria priorizar acessibilidade e escalabilidade, utilizando eye tracking baseado em webcam (validado por [14]) combinado com apresentação digitalizada e temporalmente sincronizada do TROG-2. A interface deveria ser responsiva e acessível para diferentes perfis de usuários, incluindo pessoas com diversos graus de comprometimento motor e cognitivo. O protocolo temporal estabelecido por [1] - cruz de fixação (2s), estímulo visual (30s), intervalo (500ms), estímulo auditório (4s) - fornece ponto de partida validado, embora ajustes possam ser necessários para otimização baseada em dados empíricos subsequentes.

A extração de features deveria capturar múltiplas dimensões do processamento linguístico. Do eye tracking, métricas essenciais incluiriam: duração de fixações (primeiras fixações e fixações totais por área de interesse), número e sequência de fixações (scan path), latência até primeira fixação na imagem alvo, características de sacadas (número, amplitude, velocidade), proporção de tempo olhando para alvo versus distratores, padrões de regressões, e pupilometria quando tecnicamente viável. Do TROG-2, além da acurácia por bloco e tipo de estrutura sintática (já utilizadas por [4]), deveriam ser extraídos padrões de erro (esporádico, aleatório, consistente, sistemático conforme taxonomia de [2]), tempo de resposta por item, e efeitos de comprimento e complexidade sintática.

Criticamente, a integração multimodal permitiria extração de features derivadas da interação entre modalidades, representando dimensão inteiramente nova F não capturada por avaliações unimodais. Estas incluiriam: correlação entre padrões de fixação e acurácia por tipo de sentença, tempo até seleção de resposta após apresentação auditória completa, estratégias de busca visual para diferentes estruturas sintáticas, e mudanças na estratégia de olhar entre blocos simples e complexos. Por exemplo, indivíduos com APP-L, caracterizados por déficits de memória de trabalho fonológica [2], poderiam apresentar maior número de regressões visuais e maior latência em sentenças longas, padrão potencialmente distintivo desta variante.

A camada de modelagem deveria empregar abordagem hierárquica começando com algoritmos tradicionais para estabelecer baseline (SVM, Random Forest, XGBoost), progredindo para deep learning quando justificado pela complexidade dos dados. Estudos como [13] demonstraram superioridade de redes neurais profundas (80% acurácia) sobre métodos tradicionais em dados de fala, sugerindo potencial similar para dados multimodais. Modelos ensemble combinando múltiplas arquiteturas poderiam capturar diferentes aspectos dos dados. Transfer learning de modelos pré-treinados em tarefas relacionadas (como os LLMs aplicados à afasia em 2024) poderia acelerar desenvolvimento quando dados específicos forem limitados.

As tarefas de predição deveriam abranger espectro de aplicações clínicas: classificação binária (presença vs. ausência de afasia), classificação multiclasse (variantes de APP: S, NF/A, L), classificação de severidade, predição de estruturas sintáticas problemáticas específicas para orientar intervenção, e predição de resposta a intervenções específicas. Esta última representa aplicação particularmente promissora ainda não explorada na literatura revisada.

Diferentemente de muitos estudos revisados que tratam ML como "caixa preta", sistema clínico robusto requer interpretabilidade transparente. Técnicas como SHAP (SHapley Additive exPlanations) values, amplamente utilizadas em aplicações de ML em saúde, permitiriam explicar contribuição de cada feature para predições individuais. Visualizações intuitivas de features mais importantes, relatórios automatizados formatados para clínicos, recomendações personalizadas para intervenção baseadas em perfil do paciente, e dashboards de monitoramento com visualização de progresso seriam componentes essenciais da interface.

Estudo	Eixo	Tecnologias/Métodos	Limitações Críticas	Contribuição Principal
[9]	1	Eye tracking (visual world, leitura)	Amostras pequenas; generalização limitada	Mapeamento sistemático do campo
[14]	1	EyeLink vs. WebGazer.js	Validação inicial; necessita mais testes	Democratização da tecnologia
[2]	2	TROG2-Br	Amostras pequenas por variante; normas por escolaridade necessárias	Validação TROG2-Br; protocolo padronizado
[13]	3	DNN + análise acústica/linguística	N não reportado; sem validação externa	Classificação automatizada via fala
[3]	3	ML + EEG/TRF	N pequeno (10/grupo); preliminar	Biomarcadores neurofisiológicos
[8]	3	ML + PET-FDG	Requer neuroimagem; custo elevado; acesso limitado	Refinamento de taxonomia
[12]	3	SuStaIn + MRI longitudinal	Complexidade de interpretação clínica	Maior dataset APP; modelagem de progressão
[1]	4	Eye tracking (PCEye) + TROG2-Br	N pequeno; apenas Broca; dataset não disponível ainda	Dataset multimodal brasileiro público
[10]	4	Eye tracking + TROG-2	Sem população clínica	Parâmetros normativos brasileiros
[11]	4	Eye tracking + TROG-2	N muito pequeno; prova de conceito	Viabilidade da abordagem integrada
[4]	5	ML (DT, kNN, NB, SVM) + TROG2-Br	N=23 pré-SMOTE; sem validação externa; dados sintéticos	Primeira aplicação ML ao TROG2-Br

Tabela 2 – Síntese Comparativa dos Principais Estudos por Eixo Tecnológico

A integração multimodal apresenta desafios técnicos não triviais. A sincronização temporal precisa entre modalidades é crítica - misalinhamentos de centenas de milissegundos podem comprometer análises de processamento em tempo real. O gerenciamento de dados de alta dimensionalidade (centenas ou milhares de features de eye tracking + TROG-2) requer técnicas apropriadas de redução de dimensionalidade e seleção de features para evitar overfitting, especialmente dado tamanhos amostrais tipicamente pequenos em populações clínicas. A fusão de features em múltiplos níveis (early fusion vs. late fusion) requer experimentação sistemática para determinar abordagem ótima. Finalmente, a validação cruzada apropriada em contexto multimodal, garantindo que dados de teste sejam verdadeiramente independentes em todas as modalidades, apresenta complexidade adicional.

4.2 Datasets Públicos e Infraestrutura de Open Science

A revisão revelou escassez crítica de dados, com a maioria dos estudos de ML operando com amostras $n < 50$. O dataset de [1], embora representando avanço importante, inclui apenas 15 indivíduos com afasia de Broca. Para que abordagens de ML multimodal atinjam seu potencial, é necessária mudança de paradigma em direção a datasets substancialmente maiores, compartilhados publicamente e enriquecidos com anotações clínicas detalhadas.

Datasets de próxima geração deveriam incluir mínimo de 200+ participantes (100+ afásicos, 100+ controles) para permitir particionamento adequado em conjuntos de treino, validação e teste com poder estatístico suficiente. Coleta multicêntrica em 5-10 instituições garantiria diversidade geográfica, socioeconômica e linguística, mitigando vieses de site único. Dados longitudinais com avaliações repetidas (trimestrais ou semestrais) ao longo de pelo menos 18-24 meses capturariam trajetórias de progressão (em APP) ou recuperação (em afasia pós-AVC), dimensão crítica ausente na maioria dos estudos revisados.

Anotações ricas seriam essenciais, incluindo metadados demográficos completos (idade, sexo, educação, profissão, lateralidade), informações clínicas detalhadas (tipo de afasia segundo critérios padronizados como Gorno-Tempini et al., 2011 para APP; severidade via Western Aphasia Battery ou similar; tempo desde início; etiologia; localização de lesão quando disponível), resultados de avaliações complementares (neuropsicológica, fonoaudiológica, funcional), dados de neuroimagem estrutural e funcional quando disponíveis, e histórico de tratamentos e resposta a intervenções.

O compartilhamento de dados clínicos sensíveis enfrenta desafios éticos e legais significativos. Anonimização robusta de dados identificáveis é tecnicamente complexa, especialmente para dados de vídeo (potencialmente incluídos em gravações de eye tracking) e neuroimagem (onde reconstrução facial é possível). O consentimento informado deve explicitamente incluir permissão para compartilhamento público, o que pode reduzir taxa de participação. Conformidade com Lei Geral de Proteção de Dados (LGPD) no Brasil e regulações internacionais equivalentes (GDPR na Europa, HIPAA nos EUA) requer expertise jurídica

especializada. Protocolos de acesso controlado (tiered access) podem ser necessários para dados particularmente sensíveis, equilibrando abertura com proteção de privacidade.

4.3 Medicina de Precisão e Personalização de Intervenções

A medicina tradicional opera predominantemente com protocolos padronizados aplicados uniformemente a grupos diagnósticos. A medicina de precisão, habilitada por ML em dados multimodais, promete mudar paradigma para tratamentos personalizados baseados em perfis individuais detalhados.

Sistema integrado geraria perfil abrangente caracterizando: (1) Perfil de compreensão sintática - estruturas preservadas vs. comprometidas com gradiente de dificuldade individual, recursos cognitivos disponíveis inferidos de padrões de eye tracking (memória de trabalho refletida em regressões visuais, atenção sustentada em consistência de fixações); (2) Perfil de estratégias de processamento - estratégias visuais empregadas (scan patterns que revelam se busca é sistemática ou aleatória), tempo de processamento por tipo de estrutura identificando gargalos específicos, adaptabilidade a diferentes níveis de complexidade; (3) Perfil preditivo - risco de progressão baseado em padrões baseline, resposta esperada a diferentes modalidades de intervenção predita por modelos treinados em dados históricos, áreas prioritárias para intervenção identificadas por análise de features mais deficitárias.
Baseado em perfis, sistema recomendaria: alvos terapêuticos específicos (estruturas sintáticas prioritárias identificadas por maior déficit relativo e maior potencial de melhora), nível de complexidade apropriado para início (determinado por perfil de desempenho, evitando frustração de tarefas muito difíceis ou tédio de tarefas muito fáceis), progressão individualizada de dificuldade (adaptativa baseada em desempenho contínuo), estratégias de intervenção mais adequadas ao perfil (abordagens top-down vs. bottom-up, quantidade de suporte visual necessário, tipo de scaffolding apropriado), e parâmetros de dosagem (intensidade e duração de tratamento personalizadas).

Reavaliações periódicas automatizadas (mensais ou trimestrais) permitiriam detecção precoce de plateaus ou declínio, ajuste dinâmico de metas e estratégias, e validação ou refinamento de predições iniciais. Visualizações de trajetória ao longo do tempo facilitariam comunicação de progresso a pacientes, familiares e equipe.

Personalização efetiva requer modelos treinados em amostras suficientemente grandes e diversas para capturar heterogeneidade de respostas, dados longitudinais vinculando características baseline a desfechos de tratamento, e validação de que recomendações personalizadas efetivamente melhoram desfechos comparadas a abordagens padronizadas (requerendo ensaios clínicos randomizados com braço personalizado vs. padrão). Adicionalmente, há risco de overfit - modelos muito personalizados podem capturar ruído idiossincr ático ao invés de padrões genuinamente preditivos.

4.4 Expansão Tecnológica e Metodológica

Os estudos de [3] com EEG e [8] com PET demonstraram valor de biomarcadores cerebrais. Próxima geração de sistemas multimodais poderia integrar: Eletroencefalografia (EEG): Componentes de potenciais relacionados a eventos (ERPs) como N400 (sensível a violações semânticas) e P600 (sensível a violações sintáticas) durante apresentação do TROG-2 forneceriam marcadores temporais de alta resolução do processamento linguístico, complementando resolução temporal limitada do eye tracking (tipicamente 30-60 Hz). Correlação entre features de eye tracking e atividade neural elucidaria mecanismos subjacentes a estratégias visuais observadas.

Neuroimagem Funcional por Espectroscopia de Infravermelho Próximo (fNIRS): Alternativa portátil e relativamente acessível ao fMRI, fNIRS mede oxigenação cerebral durante tarefas cognitivas, fornecendo resolução espacial moderada (centímetros) de ativação cortical em regiões linguísticas frontais e temporais.

Análise Acústica de Fala: Integração com análise computacional de produções de fala (utilizando técnicas demonstradas eficazes por [13]), capturando dimensão expressiva da linguagem complementar à compreensão avaliada por TROG-2.
Biomarcadores Multimodais Integrados: Modelos de ML multimodais integrando dados comportamentais (eye tracking + TROG-2), neurofisiológicos (EEG/fNIRS) e de fala poderiam alcançar acurácia diagnóstica superior a qualquer modalidade isolada, fornecer insights sobre mecanismos neurais subjacentes a déficits comportamentais, e identificar biomarcadores precoces de progressão (particularmente relevante em APP).

4.5 Direções Inovadoras em Machine Learning

Dados clínicos sensíveis frequentemente não podem ser centralizados devido a restrições regulatórias, institucionais ou éticas. Aprendizado federado oferece solução elegante: modelos são treinados localmente em cada instituição com seus próprios dados, e apenas parâmetros do modelo (não dados brutos) são compartilhados e agregados centralmente. Isso viabiliza colaboração multi-institucional em larga escala preservando privacidade. Implementação requer infraestrutura técnica para coordenação (servidor central de agregação), protocolos de segurança (criptografia de parâmetros transmitidos), e governança de dados (acordos claros sobre propriedade intelectual, autoria, uso de modelos resultantes). Desafios incluem heterogeneidade de dados entre sites (diferentes protocolos de coleta, equipamentos, populações) e necessidade de suficientes dados locais para treinamento inicial efetivo.

Sistemas de ML em saúde enfrentam exigência crescente de explicabilidade. Técnicas como SHAP values, LIME (Local Interpretable Model-agnostic Explanations), e Attention mechanisms em redes neurais visualizam quais features influenciam predições. Modelos intrinsecamente interpretáveis (Decision Trees com profundidade limitada, Generalized Additive Models, rule-based systems) podem ser preferíveis a "caixas

pretas" complexas quando interpretabilidade é prioritária, mesmo com sacrifício modesto de acurácia. Validação clínica da interpretabilidade - apresentando explicações a especialistas e avaliando concordância com intuições clínicas - é componente crítico frequentemente negligenciado. Explicações podem gerar novos insights clínicos, identificando padrões ou combinações de features não previamente reconhecidas como clinicamente significativas.

Dado tamanhos amostrais pequenos típicos em populações clínicas especializadas, técnicas que aproveitam dados não-anotados ou transferem conhecimento de domínios relacionados são promissoras. Aprendizado auto-supervisionado em grandes datasets de eye tracking de indivíduos saudáveis realizando tarefas linguísticas poderia pré-treinar modelos para capturar padrões gerais de processamento visual de linguagem, que são então refinados (fine-tuned) com dados menores mas anotados de populações clínicas. Transfer learning de modelos treinados em línguas com mais recursos (inglês) para português, ou de APP para afasia pós-AVC, poderia acelerar desenvolvimento quando dados diretos são escassos.

5 ANÁLISE EXPLORATÓRIA DE DADOS EXPERIMENTAIS (PROVA DE CONCEITO CONSIDERAÇÕES

Com o objetivo de aproximar o levantamento teórico da aplicação prática, realizou-se uma análise exploratória inédita sobre o dataset público disponibilizado por [1]. Importa destacar que os dados brutos utilizados são de autoria de [1], enquanto o pipeline de processamento, as métricas extraídas, a análise estatística comparativa e as visualizações apresentadas nesta seção constituem contribuição original dos autores do presente trabalho, não havendo sobreposição com os resultados publicados na obra original. E integra o mesmo eixo de pesquisa desta revisão (Eye-Tracking + TROG-2 + Afasia). Essa iniciativa teve como propósito validar empiricamente os padrões de comportamento visual descritos na literatura, por meio de registros reais de participantes com e sem afasia.

Foram analisadas amostras provenientes do banco de dados experimental: participantes do grupo de controle (sem afasia) e do grupo afásico, ambos realizando tarefas do teste TROG-2. Os movimentos oculares foram capturados por um Tobii PC Eye Mini (60 Hz), com coordenadas normalizadas (0–1), e processados por um script em *Python* padronizado que inclui: (i) limpeza e remoção de outliers (IQR), (ii) segmentação por estímulo (até 32 itens), (iii) detecção de sacadas e fixações pelo método I-VT (limiar adaptativo no percentil 85 da velocidade, fixações ≥100 ms), e (iv) extração de métricas temporais e espaciais por paciente e por estímulo.

5.1 Padrão de Fixações Oculares

A Figura 1 apresenta o padrão de fixações do participante do grupo de controle, caracterizado por maior concentração espacial e trajetória econômica do olhar, indicando foco visual eficiente e estabilidade atencional. Em contraste, a Figura 2 ilustra o padrão do grupo afásico, evidenciando dispersão mais ampla e maior varredura do campo visual, sugerindo esforço compensatório para a compreensão gramatical.

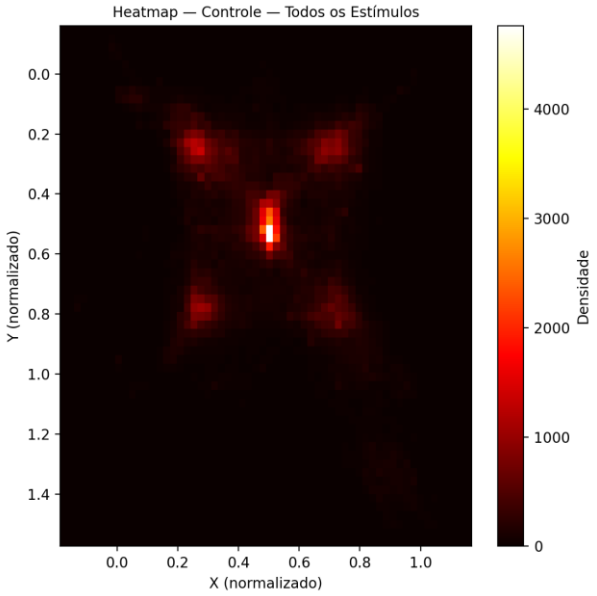

Figura 1: Padrão de Fixações Oculares – Participante Controle (TROG-2)

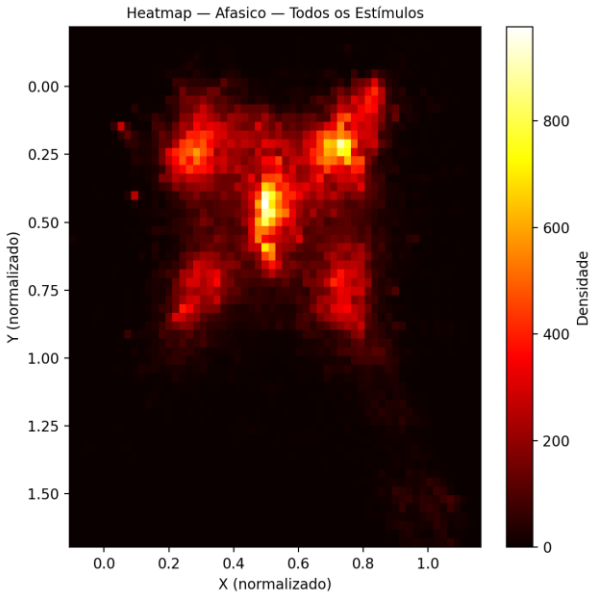

Figura 2: Padrão de Fixações Oculares – Participantes Afásicos (TROG-2)

5.2 Indicadores Quantitativos

Na Tabela 3, os valores de U e p-valor correspondem aos resultados do teste estatístico de Mann-Whitney U, utilizado para comparar as distribuições entre os dois grupos. O U representa o valor da estatística do teste (indicando a posição relativa dos escores dos

grupos), enquanto o p-valor expressa a probabilidade de que a diferença observada tenha ocorrido ao acaso. Assim, valores de $p < 0,05$ indicam diferenças estatisticamente significativas entre os grupos para aquela métrica.

Observa-se um aumento consistente no tempo de resposta e no total de sacadas no grupo afásico, com valores próximos ao limiar de significância estatística ($p \approx 0,05$). As variáveis de microdinâmica — tempo médio de sacada e duração média de fixação — não apresentaram diferenças globais significativas, embora mostrem variações pontuais conforme o tipo de estímulo.

Na análise segmentada por estímulo (n = 32), identificaram-se diferenças mais pronunciadas nos itens de maior complexidade sintática. Os indicadores mais sensíveis ao efeito da afasia foram o número de sacadas e o tempo de resposta, ambos com diferenças estatisticamente significativas em quase metade dos estímulos avaliados. Métricas associadas à microdinâmica do olhar, como a duração média das fixações, mostraram menor sensibilidade isoladamente, mas contribuíram para a caracterização global do esforço cognitivo.

Esses achados sugerem que o efeito da afasia se manifesta com maior força em métricas de "custo global" (tempo total e frequência de sacadas) e varia conforme a complexidade sintática do estímulo (ver subitens 11–25 para maiores contrastes). Já os índices de microdinâmica do movimento ocular, quando agregados, tendem a ser menos sensíveis como marcadores isolados.

Para referência cruzada com a redação anterior: os valores aproximados reportados na literatura (ex.: +15–20% em sacadas/fixações e tempo de resposta) são compatíveis com os efeitos observados aqui quando se olha por estímulo; no agregado, a significância aparece sobretudo para tempo de resposta e nº de sacadas ($p \approx 0,05$).

5.3 Considerações

A análise experimental reforça a viabilidade técnica e clínica da integração entre Eye-Tracking e TROG-2 para o estudo de déficits mais sacadas para processar a cena, especialmente em itens com exploração mais custoso: demoram mais para responder e realizam linguísticos em falantes do português brasileiro. No plano comportamental-oculomotor, os afásicos exibem padrão de maior demanda sintática. Esses resultados são convergentes com relatos prévios (por exemplo, [2]; [14]), que descrevem maior variabilidade espacial e incremento na alternância do foco em populações com afasia.

Do ponto de vista metodológico, a análise exploratória (I-VT com limiar adaptativo, fix≥100 ms) mostrou-se robusta para: (i) segmentar a exploração antes da resposta (Key≠0), (ii) comparar grupos por estímulo e (iii) gerar produtos reprodutíveis (CSVs, heatmaps e gráficos por item). Os resultados fornecem evidências empíricas que sustentam o Eixo 4 deste artigo (Eye-Tracking + TROG-2 + Afasia) e criam condições para a continuidade no Eixo 5, com integração a modelos de Machine Learning. Em particular, o tempo de resposta e o número de sacadas emergem como características candidatas a *features* discriminativas em modelos preditivos de desempenho linguístico, com potencial aplicação em sistemas assistivos inteligentes para quantificação de compreensão e orientação de intervenções clínicas personalizadas.

A inclusão de um modelo de classificação com validação cruzada formal excede o escopo desta prova de conceito, dadas as restrições de tamanho amostral que comprometeriam a validade de qualquer estimativa de generalização. Esta limitação é reconhecida e aponta diretamente para a necessidade de datasets maiores como condição para avançar nessa direção, conforme discutido na seção 4.2

6 CONCLUSÃO

Esta revisão sistemática investigou o estado da arte da integração entre eye tracking, TROG-2 e machine learning na avaliação da afasia, mapeando 28 estudos publicados entre 2020 e 2025 através de estratégia de busca sistemática em oito bases de dados acadêmicas. A análise revelou avanços significativos em cada eixo tecnológico individualmente, mas evidenciou lacuna crítica: nenhum estudo integra simultaneamente as três tecnologias, representando a principal oportunidade científica identificada por esta revisão.

O Eixo 1 (Eye tracking + Afasia, n=16 estudos) estabeleceu viabilidade técnica do rastreamento ocular para investigar processamento linguístico em afasia, com avanço crítico na democratização via sistemas baseados em webcam [14]. Entretanto, tamanhos amostrais persistentemente pequenos (média < 10 participantes clínicos) limitam generalizabilidade e desenvolvimento de modelos preditivos robustos.

O Eixo 2 (TROG-2 + Afasia, n=8 estudos) forneceu fundamento psicométrico sólido através da validação do TROG2-Br [14], demonstrando sensibilidade para detectar déficits morfossintáticos em APP com propriedades adequadas (α=0,87; r=0,765 com Token Test). Este eixo estabelece protocolo padronizado essencial, embora revele necessidade de normas ajustadas por escolaridade.

O Eixo 3 (ML/IA + Afasia, n=12 estudos) apresentou maior diversidade metodológica, variando de deep learning ([13]: 80% acurácia) a biomarcadores neurofisiológicos ([3]: F1 0,73-0,74) e neuroimagem ([12]: n=270). Acurácias reportadas entre 73-95% demonstram potencial, mas ausência sistemática de validação externa limita aplicabilidade clínica imediata.

Indicador	Controle (média ± DP)	Afásico (média ± DP)	U	p-valor	Observações
Tempo médio de resposta (s)	10,23 ± 1,57	12,39 ± 2,30	33	0,0508	Afásicos mais lentos
Total de sacadas	745,31 ± 205,91	1.099,00 ± 447,39	33	0,0508	Mais movimentos oculares
Tempo médio de sacada (s)	0,0490 ± 0,0178	0,0374 ± 0,0175	84	0,2513	Diferença não significativa
Total de fixações (≥100 ms)	494,54 ± 112,62	583,10 ± 102,56	38	0,094	Tendência a mais fixações
Duração média de fixação (s)	0,5265 ± 0,0747	0,5258 ± 0,1499	66	0,9753	Sem diferença
Dispersão média (área)	0,8419 ± 0,2536	0,7862 ± 0,2976	77	0,4757	Sem diferença global

Tabela 3 – Métricas oculomotoras agregadas por grupo (todos os estímulos)

O Eixo 4 (Eye tracking + TROG-2 + Afasia, n=3 estudos) representa avanço qualitativo importante. O dataset brasileiro de [1] (2024, n=21) estabelece precedente ao combinar duas modalidades, com confirmação estatística de diferenças significativas entre grupos (Kruskal-Wallis $p<0,01$), validando sensibilidade da abordagem combinada.

O Eixo 5 (ML + TROG-2 + Afasia, n=1 estudo) constitui prova de conceito pioneira (Freitas et al., 2023: 75% acurácia com Decision Tree) mas metodologicamente limitada (n=23 pré-SMOTE, sem validação externa). A identificação dos blocos Q, M, K e H como discriminativos fornece insight clínico valioso sobre estruturas sintáticas informativas.

Múltiplos estudos ([8]; [12]) revelam que subtipos data-driven baseados em biomarcadores não correspondem perfeitamente a classificações clínicas tradicionais, sugerindo necessidade de refinamento taxonômico futuro. Processamento Multimodal: A integração de eye tracking com TROG-2 oferece janela única para processos cognitivos subjacentes à compreensão sintática, permitindo dissociar déficits morfossintáticos genuínos de limitações em processamento geral (atenção, memória de trabalho).

Machine Learning como Ferramenta de Descoberta: Além de aplicações diagnósticas, ML pode gerar hipóteses científicas identificando padrões não aparentes (como blocos TROG-2 mais discriminativos), potencialmente revelando mecanismos subjacentes aos déficits.

Potencial Transformador: Um sistema integrado validado poderia revolucionar avaliação clínica oferecendo: objetividade (métricas quantitativas automáticas), eficiência (avaliação potencialmente mais rápida), acessibilidade (eye tracking via webcam permite avaliação remota), sensibilidade (detecção de déficits sutis), e personalização (perfis individualizados orientando intervenções).
Barreiras Atuais: Implementação enfrenta desafios significativos incluindo tamanhos amostrais inadequados, ausência de validação externa sistemática, falta de padronização de protocolos, questões éticas de compartilhamento de dados, e necessidade de aprovações regulatórias.

A afasia afeta milhões globalmente, com impacto devastador na qualidade de vida. Esta revisão demonstrou que, embora progressos substanciais tenham ocorrido em eye tracking, TROG-2 e machine learning individualmente, a integração completa permanece inexplorada. Esta lacuna representa simultaneamente desafio e oportunidade transformadora.
O contexto brasileiro apresenta posição favorável para liderança, com validação do TROG2-Br, desenvolvimento pioneiro de dataset multimodal, e primeira aplicação de ML a dados do TROG2-Br. Investimento continuado e colaboração interdisciplinar podem posicionar o Brasil na vanguarda desta fronteira tecnológica emergente.

O presente mapeamento apresenta limitações que devem ser consideradas na interpretação dos resultados. A delimitação temporal de 2020 a 2025, embora justificada pela necessidade de capturar avanços tecnológicos recentes em eye tracking e machine learning, pode ter excluído estudos fundacionais relevantes, especialmente para instrumentos consolidados como o TROG-2, cuja literatura de validação é anterior a esse período. A ausência de instrumento formal de avaliação de risco de viés, substituída por avaliação qualitativa estruturada, representa outra limitação inerente ao desenho de mapeamento adotado. Por fim, a busca por proximidade, embora necessária diante da ausência de estudos que integrem simultaneamente as três tecnologias, introduz heterogeneidade metodológica entre os eixos, o que limita comparações diretas entre estudos.

A jornada da pesquisa básica à implementação clínica é longa, mas os benefícios potenciais - avaliação mais precisa, prognóstico mais confiável, intervenções personalizadas - justificam plenamente o esforço. Este mapeamento sistemático estruturou o campo, identificou o caminho à frente, e espera inspirar a próxima geração de pesquisas que transformarão aspiração em realidade para pessoas vivendo com afasia.

AGRADECIMENTOS

Agradecemos à Universidade do Vale do Itajaí e a Fundação de Amparo à Pesquisa e Inovação no Estado de Santa Catarina (FAPESC), TO 2024TR2195.

REFERÊNCIAS

[1] ALVES, Michael Douglas Cabral et al. Exploring sentence processing in people with aphasia in the Brazilian context using eye-tracking. Authorea Preprints, 2024. DOI: 10.22541/au.171204756.68362058/v1.

[2] CARTHERY-GOULART, Maria Teresa et al. Sentence comprehension in primary progressive aphasia: a study of the application of the Brazilian version of the Test for the Reception of Grammar (TROG2-Br). Frontiers in Neurology, v. 13, p. 815227, maio 2022. DOI: 10.3389/fneur.2022.815227.

[3] DIAL, Heather et al. Application of machine learning and temporal response function modeling of EEG data for differential diagnosis in primary progressive aphasia. Scientific Reports, v. 15, n. 1, p. 29539, 2025. DOI: 10.1038/s41598-025-13000-8.

[4] FREITAS, Maurício de et al. Uso de aprendizado de máquina para identificar o tipo de afasia progressiva primária a partir do desempenho no TROG-2Br. In: COMPUTER ON THE BEACH, 14., 2023, Florianópolis. Anais... Florianópolis: Univali, 2023. p. 512-514. DOI: 10.14210/cotb.v14.p512-514.

[5] GORNO-TEMPINI, M. L. et al. Classification of primary progressive aphasia and its variants. Neurology, v. 76, n. 11, p. 1006-1014, mar. 2011. DOI: 10.1212/WNL.0b013e31821103e6.

[6] KNILANS, Jessica; DEDE, Gayle. Online sentence reading in people with aphasia: evidence from eye tracking. American Journal of Speech-Language Pathology, v. 24, n. 4, p. S961-S973, 2015. DOI: 10.1044/2015_AJSLP-14-0140.

[7] MATIAS-GUIU, Jordi A. et al. Advances in primary progressive aphasia. Brain Sciences, v. 12, n. 5, p. 636, 2022. DOI: 10.3390/brainsci12050636.

[8] MATIAS-GUIU, Jordi A. et al. Machine learning in the clinical and language characterisation of primary progressive aphasia variants. Cortex, v. 119, p. 312-323, 2019. DOI: 10.1016/j.cortex.2019.05.007.

[9] SHARMA, Saryu et al. Eye tracking measures for studying language comprehension deficits in aphasia: a systematic search and scoping review. Journal of Speech, Language, and Hearing Research, v. 64, n. 3, p. 1008-1022, mar. 2021. DOI: 10.1044/2020_JSLHR-20-00287.

[10] SOUZA, Paulo Henrique de et al. Eye tracking application in the evaluation of aphasic Portuguese speakers. In: IBERIAN CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGIES (CISTI), 16., 2021, Chaves, Portugal. Proceedings... [S.l.]: IEEE, 2021. ISBN 978-989-54659-1-0. DOI: 10.23919/CISTI52073.2021.9476298.

[11] SOUZA, Paulo Henrique de; FREITAS, Maria Isabel d'Ávila; RAMIREZ, Alejandro Rafael Garcia. Assessment of aphasics fluent in Portuguese language using eye-tracking. In: IBERIAN CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGIES (CISTI), 18., 2023, Aveiro, Portugal. Proceedings... [S.l.]: IEEE, 2023. p. 1-4. DOI: 10.23919/CISTI58278.2023.10211448.

[12] TAYLOR, Beatrice et al. Data-driven neuroanatomical subtypes of primary progressive aphasia. Brain, v. 148, n. 3, p. 955-968, 2025. DOI: 10.1093/brain/awae314.

[13] THEMISTOCLEOUS, Charalambos et al. Automatic subtyping of individuals with primary progressive aphasia. Journal of Alzheimer's Disease, v. 79, n. 3, p. 1185-1194, 2021. DOI: 10.3233/JAD-201101.

[14] VAN BOXTEL, Willem S. et al. Online eye tracking for aphasia: a feasibility study comparing web and lab tracking and implications for clinical use. Brain and Behavior, v. 14, n. 11, e70112, 2024. DOI: 10.1002/brb3.70112.