RECONHECIMENTO MULTIMODAL DE GESTOS MANUAIS E MARCADORES FACIAIS GRAMATICAIS DA LÍNGUA BRASILEIRA DE SINAIS POR VISÃO COMPUTACIONAL EM APOIO À COMUNICAÇÃO DE PESSOAS COM DEFICIÊNCIA AUDITIVA

Débora dos Santos Figueiredo[1], Sergio Coral[2]

Resumo: A comunicação entre pessoas surdas e ouvintes que não dominam a Língua Brasileira de Sinais (Libras) ainda representa uma barreira de acessibilidade. Embora sistemas computacionais de reconhecimento de sinais tenham avançado, muitas abordagens priorizam os gestos manuais e deixam de considerar marcadores faciais gramaticais, importantes para a construção de sentido na língua. Diante dessa lacuna, este trabalho teve como objetivo desenvolver e avaliar um protótipo *web* para reconhecimento multimodal da Libras, integrando sinais manuais e marcadores faciais capturados por webcam convencional. A metodologia envolveu OpenCV e MediaPipe para processamento dos *frames* e extração de *landmarks*, scikit-learn para os classificadores supervisionados e fusão em nível de decisão entre os canais manual e facial. No canal manual, a Regressão Logística alcançou 81,84% de acurácia na validação por tomada e 56,32% na validação por sinalizador. No canal facial, o *ExtraTreesClassifier* obteve 94,31% de acurácia e 92,93% de *F1-score* na divisão controlada por *frames*. A integração multimodal permitiu combinar o reconhecimento lexical dos sinais com informações gramaticais faciais, produzindo uma interpretação mais contextualizada, embora limitada pela ausência de uma base multimodal pareada e pela variabilidade individual dos sinalizadores.

Palavras-chave: Libras; reconhecimento multimodal; gestos manuais; marcadores faciais gramaticais; visão computacional.

ABSTRACT: Communication between deaf people and hearing individuals who do not know Brazilian Sign Language (Libras) still represents an accessibility barrier. Although computational sign recognition systems have advanced, many approaches prioritize manual gestures and do not consider

[1]Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense (Unesc), Criciúma - Santa Catarina - Brasil. figsdebora26@unesc.net

[2]Orientador, Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense (Unesc), Criciúma - Santa Catarina - Brasil. sergiocoral@unesc.net

grammatical facial markers, which are important for meaning construction in the language. To address this gap, this study aimed to develop and evaluate a *web*-based prototype for multimodal Libras recognition, integrating manual signs and facial markers captured with a conventional webcam. The methodology used OpenCV and MediaPipe for frame processing and *landmark* extraction, scikit-learn for supervised classifiers, and decision-level fusion between the manual and facial channels. In the manual channel, Logistic Regression achieved 81.84% accuracy in take-based validation and 56.32% in signer-based validation. In the facial channel, *ExtraTreesClassifier* achieved 94.31% accuracy and a 92.93% *F1-score* in the controlled frame-level split. The multimodal integration made it possible to combine lexical sign recognition with facial grammatical information, producing a more contextualized interpretation, although limited by the absence of a paired multimodal dataset and by individual variability among signers.

Keywords: Brazilian Sign Language; multimodal recognition; manual gestures; grammatical facial markers; computer vision.

1 INTRODUÇÃO

A Língua Brasileira de Sinais (Libras) constitui um dos principais meios de comunicação de pessoas surdas no Brasil e representa um elemento fundamental para sua participação social, educacional e profissional. Seu reconhecimento legal pela Lei nº 10.436/2002 consolidou a Libras como meio de comunicação e expressão, caracterizado por natureza visual-motora e estrutura gramatical própria (Brasil, 2002). A Constituição Federal também assegura a igualdade de direitos e o acesso à educação como princípios fundamentais (Brasil, 1988). Entretanto, a garantia legal da língua não elimina as barreiras comunicacionais ainda presentes em situações cotidianas, especialmente quando a interação depende de pessoas ouvintes que não dominam Libras. Segundo o Censo Demográfico 2022, divulgado pelo IBGE, 2,6 milhões de pessoas no Brasil apresentavam dificuldade para ouvir, mesmo com o uso de aparelhos auditivos (IBGE, 2025). Esse indicador dimensiona a amplitude das demandas de acessibilidade auditiva no país e, articulado ao reconhecimento da Libras como direito linguístico, reforça a necessidade de soluções tecnológicas capazes de apoiar a comunicação em diferentes contextos sociais.

Embora frequentemente associada apenas aos movimentos das mãos, a Libras é uma língua de modalidade visuoespacial complexa, que

exige a articulação simultânea de múltiplos parâmetros fonológicos. Sua estrutura linguística não se restringe à configuração de mão, locação e movimento, mas depende também das expressões não manuais (ENM) (Quadros, 2019). Nesse contexto, as expressões faciais exercem um papel que ultrapassa a demonstração de emoções, atuando como marcadores gramaticais relevantes para a construção do enunciado. Elas podem contribuir para a definição de construções sintáticas interrogativas, negativas e condicionais, além de funcionarem como modificadores semânticos capazes de alterar a intensidade ou o aspecto verbal de um sinal (Paiva et al., 2020).

No campo da visão computacional, essa complexidade linguística evidencia uma limitação recorrente dos sistemas de reconhecimento automático de línguas de sinais: a predominância de abordagens centradas no rastreamento dos gestos manuais. Embora o avanço de modelos computacionais tenha ampliado a capacidade de identificar configurações, trajetórias e posições das mãos, a integração entre componentes manuais e não manuais permanece como um desafio relevante na literatura de SLR (*Sign Language Recognition*). Quando o canal facial é desconsiderado, o sistema tende a reconhecer apenas o item lexical, sem captar marcas gramaticais que podem alterar a função comunicativa do enunciado. Com isso, a interpretação automática pode tornar-se incompleta, ambígua ou distante da forma como a Libras é efetivamente produzida (Alaghband; Maghroor; Garibay, 2023).

Abordagens recentes têm investigado arquiteturas profundas para reconhecimento de línguas de sinais, incluindo modelos baseados em *Transformers* e redes convolucionais em grafos, capazes de representar dependências espaciais e temporais em sequências de sinais (Cui et al., 2023; Naz et al., 2023). Embora apresentem resultados promissores, essas abordagens geralmente exigem maior volume de dados, maior custo computacional e protocolos mais amplos de validação, justificando, neste trabalho, a adoção de modelos clássicos como linha de base experimental para um vocabulário controlado.

Uma análise dos trabalhos correlatos evidencia os avanços e as limitações das abordagens unimodais centradas nas mãos. Fontana e Coral investigaram redes neurais convolucionais integradas à biblioteca MediaPipe para o reconhecimento do alfabeto manual estático da Libras. Embora o estudo tenha alcançado 98,67% de acurácia em ambiente controlado, os testes com webcam apresentaram variações associadas a ruídos no plano de fundo e a ângulos de captura desfavoráveis (Fontana; Coral, 2025).

Para abordar a natureza temporal da língua, Albino e Coral utilizaram o algoritmo *Dynamic Time Warping* (DTW) na classificação de sinais dinâmicos da Libras. A aplicação alcançou 97,04% de acurácia com a segmentação manual do gesto, mas esse resultado caiu para 69,35% na segmentação automática. Além da dificuldade de delimitar o início e o fim dos sinais, a abordagem não incorporou o canal facial ao processo interpretativo (Albino; Coral, 2023).

Reconhecendo a lacuna deixada pelas abordagens unimodais, outros pesquisadores direcionaram seus esforços para a análise dos componentes não manuais. Silva et al. propuseram a base Silfa (Sign Language Facial Action Database), voltada à anotação e detecção de unidades de ação facial em sinais da Libras. Em estudo posterior, Silva et al. reforçaram que a face constitui um componente linguístico relevante para tecnologias assistivas, especialmente por permitir a descrição sistemática de variações faciais associadas à produção dos sinais (Silva et al., 2022).

A análise facial isolada não resolve o reconhecimento do sinal como unidade comunicativa, pois a face não identifica o item lexical realizado pelas mãos. Em âmbito internacional, Kumar, Roy e Dogra combinaram gestos manuais e expressões faciais no reconhecimento da Indian Sign Language, utilizando os sensores Leap Motion e Microsoft Kinect (Kumar; Roy; Dogra, 2018). Long et al. também investigaram a combinação entre expressão facial e esqueleto das mãos (Long et al., 2023). Esses estudos reforçam o potencial da integração entre canais, mas não tratam da Libras e adotam condições experimentais distintas das empregadas nesta pesquisa.

Os trabalhos analisados indicam um hiato tecnológico e científico: a necessidade de uma solução nacional, voltada especificamente para a Libras, capaz de integrar as dimensões manual e facial de forma simultânea, mas operando com hardware acessível, isto é, utilizando apenas uma webcam convencional, sem sensores de profundidade ou dispositivos vestíveis.

Diante desse contexto, o trabalho parte da hipótese de que a integração entre gestos manuais e marcadores faciais gramaticais, capturados por webcam, pode produzir uma interpretação mais contextualizada do que abordagens unimodais, sem assumir a tradução irrestrita da Libras.

A contribuição da pesquisa está em avaliar a fusão entre reconhecimento manual e marcadores faciais gramaticais em um protótipo com webcam convencional, oferecendo evidências sobre as potencialidades e

limitações da multimodalidade em vocabulário controlado.

O objetivo geral foi desenvolver e avaliar um sistema *web* de reconhecimento multimodal da Libras, integrando gestos manuais e marcadores faciais gramaticais. Os objetivos específicos foram selecionar e processar bases públicas adequadas ao escopo experimental, extrair atributos visuais, treinar os classificadores, implementar a fusão contextual e avaliar o desempenho e a generalização da arquitetura.

2 MATERIAIS E MÉTODOS

Esta pesquisa caracterizou-se como aplicada, de base tecnológica e com delineamento experimental, pois teve como finalidade desenvolver e avaliar um protótipo computacional para reconhecimento multimodal da Libras. O estudo foi estruturado para analisar separadamente os canais manual e facial e, posteriormente, combiná-los em uma interpretação contextual.

O treinamento e a avaliação utilizaram duas bases públicas, uma para cada canal. O *pipeline* foi implementado com OpenCV, MediaPipe, NumPy e scikit-learn, conforme o Quadro 1, e o protótipo foi executado com webcam convencional, sem sensores corporais ou dispositivos vestíveis.

2.1 CARACTERIZAÇÃO DA PESQUISA

O desenvolvimento foi organizado em seis etapas: delimitação do escopo experimental, seleção das bases, extração dos atributos visuais, treinamento dos classificadores, integração multimodal e avaliação do desempenho. O processamento foi dividido em dois canais: o manual, responsável pelo reconhecimento lexical dos sinais, e o facial, destinado à identificação de marcadores gramaticais.

O canal facial foi delimitado para marcadores gramaticais, e não para emoções genéricas, em razão da função linguística das expressões não manuais na Libras (Paiva et al., 2020).

2.2 MATERIAIS E BASES

O Quadro 1 sintetiza os principais recursos empregados no desenvolvimento do protótipo.

Quadro 1 - Recursos utilizados no desenvolvimento da pesquisa

RECURSO	FUNÇÃO NA PESQUISA
Bases públicas	Treinamento e avaliação dos canais manual e facial.
Python, OpenCV e NumPy	Leitura dos vídeos e tratamento vetorial dos atributos.
MediaPipe	Extração dos *landmarks* das mãos e rastreamento facial no protótipo.
scikit-learn e joblib	Treinamento, avaliação e serialização dos classificadores.
HTML, CSS e JavaScript	Interface *web*, acesso à webcam e interação com o usuário.
Computador com webcam	Execução local do protótipo e testes em tempo real.

Fonte: Elaborado pela autora.

Foram utilizadas duas bases públicas, uma para cada canal de reconhecimento, conforme apresentado no Quadro 2.

Quadro 2 - Bases de dados utilizadas na pesquisa

CANAL	BASE	DADOS UTILIZADOS
Manual	MINDS-Libras Dataset (Almeida et al., 2019)	380 vídeos de 20 sinais da Libras, dos conjuntos `Sinalizador01` a `Sinalizador04`, com fundo *Chroma Key*, resolução 1920 $\times$ 1080 pixels e aproximadamente 30 quadros por segundo.
Facial	Grammatical Facial Expressions (Freitas; Barbosa; Peres, 2014)	27.936 *frames* processados localmente, organizados em 18 pares de arquivos, com coordenadas tridimensionais de 100 pontos faciais e rótulos binários por expressão gramatical.

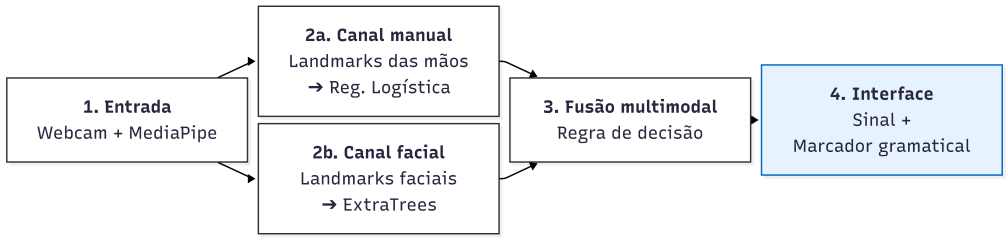
Fonte: Elaborado pela autora.

O canal manual reuniu 20 sinais da Libras. No facial, foram selecionados os marcadores “afirmação”, “negação”, “pergunta WH”, “pergunta sim/não” e “ênfase/foco”; a classe “sem marcador/neutro” representou a ausência predominante desses marcadores.

2.3 VISÃO GERAL DO *PIPELINE*

O *pipeline* transforma o fluxo de vídeo em representações numéricas. Os canais manual e facial são processados separadamente e suas saídas são combinadas por fusão em nível de decisão, conforme a Figura 1.

Figura 1 - *Pipeline* multimodal do protótipo.



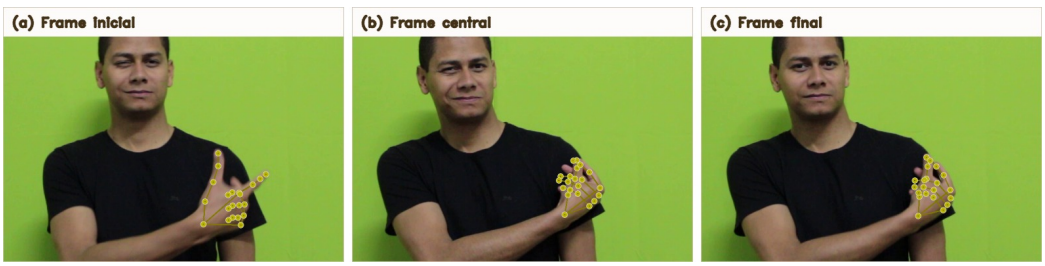
Fonte: Elaborado pela autora.

2.4 RECONHECIMENTO MANUAL

O reconhecimento manual foi baseado na extração de *landmarks* com MediaPipe Hands, solução de rastreamento de mãos em tempo real que estima o esqueleto da mão a partir de imagens RGB (Zhang et al., 2020). Cada vídeo do subconjunto da base MINDS-Libras Dataset foi lido com OpenCV e processado a cada três *frames*. Para cada mão detectada, os 21 pontos rastreados deram origem aos atributos utilizados pelo classificador: coordenadas normalizadas em relação ao punho, profundidade relativa, posição do punho, largura, altura e escala da região ocupada pela mão. Cada mão foi representada por 68 valores e, como o sistema considerou até duas mãos, cada *frame* foi descrito por 136 atributos.

Como os vídeos apresentavam durações distintas, cada amostra foi convertida em uma sequência de 24 *frames*, com interpolação linear quando necessária. Ao final do pré-processamento, cada vídeo foi representado por um vetor de 3.264 atributos, correspondente à combinação entre 24 *frames* e 136 atributos por *frame*. A Figura 2 exemplifica a extração dos pontos em três momentos do sinal “Vacina”, evidenciando a dimensão temporal considerada pelo modelo.

Figura 2 - Sequência de *frames* do sinal “Vacina” com *landmarks* das mãos extraídos pelo MediaPipe Hands.



Fonte: MINDS-Libras Dataset (Almeida et al., 2019).

O classificador principal foi implementado com Regressão Logística aplicada aos vetores temporais padronizados com *StandardScaler*. O

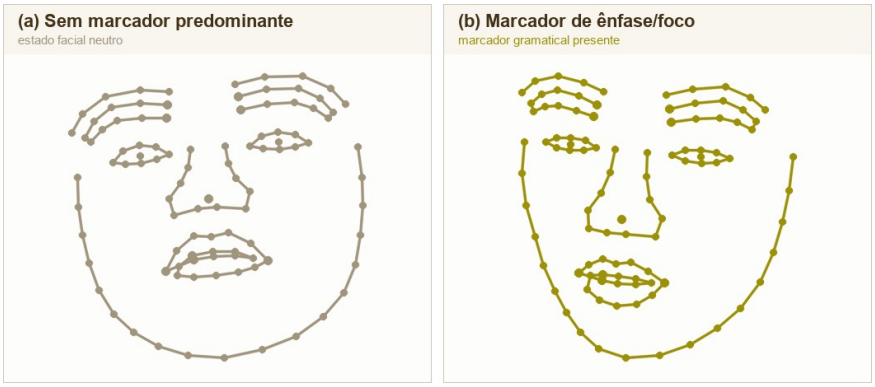
modelo realizou a classificação multiclasse, mas não aprendeu dependências sequenciais de forma explícita, pois os atributos foram concatenados em um vetor fixo.

Como referência por similaridade, também foi avaliado k-NN com $k = 1$. Para diferenciar os sinais “Medo” e “Ruim”, foi treinado um classificador especialista com características resumidas da sequência.

2.5 RECONHECIMENTO FACIAL GRAMATICAL

A avaliação experimental utilizou a base pública Grammatical Facial Expressions. Cada *frame* contém coordenadas tridimensionais de 100 pontos faciais, totalizando 300 atributos. Antes do treinamento, os pontos foram centralizados em relação à ponta do nariz e normalizados conforme a amplitude facial observada em cada *frame*, reduzindo a influência de diferenças de posicionamento e escala. A Figura 3 apresenta uma representação bidimensional desses pontos em um *frame* sem marcador predominante e em outro com “ênfase/foco”, evidenciando as regiões faciais consideradas na extração dos atributos.

Figura 3 - Representação dos *landmarks* faciais em um *frame* sem marcador predominante e em outro com “ênfase/foco”.



Fonte: Grammatical Facial Expressions (Freitas; Barbosa; Peres, 2014).

Para cada marcador, foi treinado um classificador binário de Regressão Logística ou *ExtraTreesClassifier*. Este último utiliza árvores de decisão extremamente randomizadas e é adequado ao tratamento de relações não lineares em dados tabulares (Geurts; Ernst; Wehenkel, 2006; Grinsztajn; Oyallon; Varoquaux, 2022).

A execução facial no protótipo foi tratada separadamente da avaliação com a base UCI. Na aplicação *web*, o MediaPipe Face Mesh extrai métricas de sobrancelhas, olhos, boca e cabeça, estabilizadas em uma

janela temporal recente (Kartynnik et al., 2019). A classe “sem marcador/neutro” é atribuída quando nenhum marcador predomina.

2.6 INTEGRAÇÃO MULTIMODAL

A integração ocorreu por fusão em nível de decisão. O canal manual estimou o sinal mais provável dentro do vocabulário controlado, enquanto o canal facial acompanhou o marcador gramatical predominante durante sua execução. Ao final da captura, as saídas foram combinadas para compor a interpretação apresentada ao usuário.

Quando não houve marcador facial predominante, o sinal manual foi mantido em sua forma neutra. Nos demais casos, uma matriz semântica associou o item lexical ao tipo de marcador identificado, produzindo uma frase contextualizada em português. Essa etapa não pretende realizar tradução irrestrita da Libras, mas demonstrar como a análise conjunta de mãos e face pode enriquecer a interpretação em um vocabulário delimitado.

2.7 PROTOCOLO DE AVALIAÇÃO

A avaliação foi conduzida separadamente para os dois canais. No manual, foram utilizadas acurácia e matriz de confusão.

Para o modelo manual, foram aplicadas duas estratégias de validação cruzada por grupos, com novo treinamento a cada rodada. Na validação por tomada, cada uma das cinco execuções foi utilizada uma vez como teste, enquanto as demais compuseram o treinamento. Na validação por sinalizador, os dados de cada um dos quatro participantes foram isolados como teste em uma das rodadas, permitindo avaliar a generalização para indivíduos não observados no treinamento.

No canal facial, cada marcador foi tratado como um problema binário. Foram calculadas acurácia, acurácia balanceada e *F1-score*; esta última sintetiza precisão e revocação. A acurácia balanceada foi adotada devido à diferença entre as quantidades de *frames* positivos e negativos.

A avaliação facial empregou duas estratégias complementares. Na validação estrita entre sinalizadores, os dados de um participante da base UCI foram usados para treinamento, enquanto os dados do outro participante foram reservados para teste, com alternância dos papéis entre as rodadas. Na divisão controlada por *frames*, as amostras foram separadas em 75% para treinamento e 25% para teste, com estratificação e semente fixa. A primeira estratégia avaliou a generalização entre indivíduos; a segunda mediu o desempenho técnico em um cenário controlado dentro da

própria base.

O vídeo contendo a demonstração do protótipo em execução, incluindo o reconhecimento dos sinais manuais e dos marcadores faciais gramaticais, pode ser acessado por meio de vídeo publicado no YouTube[3].

3 RESULTADOS E DISCUSSÃO

Esta seção analisa o desempenho dos canais manual e facial gramatical, bem como a integração multimodal implementada no protótipo. A apresentação segue o fluxo do sistema: reconhecimento manual, reconhecimento facial e combinação das saídas, relacionando as métricas obtidas às decisões metodológicas e às limitações observadas.

3.1 RECONHECIMENTO MANUAL

O canal manual foi avaliado com 380 vídeos da base MINDS-Libras Dataset, distribuídos em 20 sinais da Libras. Cada amostra foi representada por sequências temporais de *landmarks* das mãos, padronizadas em 24 *frames* e convertidas em vetores de 3.264 atributos. A Tabela 1 apresenta os principais resultados obtidos. A acurácia corresponde ao valor agregado das predições; o desvio padrão (DP) foi calculado entre as rodadas; e o intervalo de confiança de 95% foi estimado pelo método de Wilson para a proporção de acertos. Os valores de DP são expressos em pontos percentuais (p.p.).

Tabela 1 - Resultados do reconhecimento manual

Estratégia / modelo	**Acurácia**	**DP entre rodadas (p.p.)**	**IC 95%**
Regressão Logística – validação por tomada	81,84%	4,20	77,65–85,39%
Regressão Logística – validação por sinalizador	56,32%	30,78	51,29–61,21%
k-NN – validação por tomada	71,32%	5,04	66,57–75,63%
k-NN – validação por sinalizador	37,63%	26,10	32,91–42,60%
Classificador especialista “Medo”/“Ruim”	87,50%	9,57	73,89–94,54%

Fonte: Elaborado pela autora.

A validação por tomada alcançou 81,84% de acurácia (IC 95%: 77,65–85,39%), indicando que a representação temporal por *landmarks* foi capaz de reconhecer novas execuções de sinalizadores já presentes no

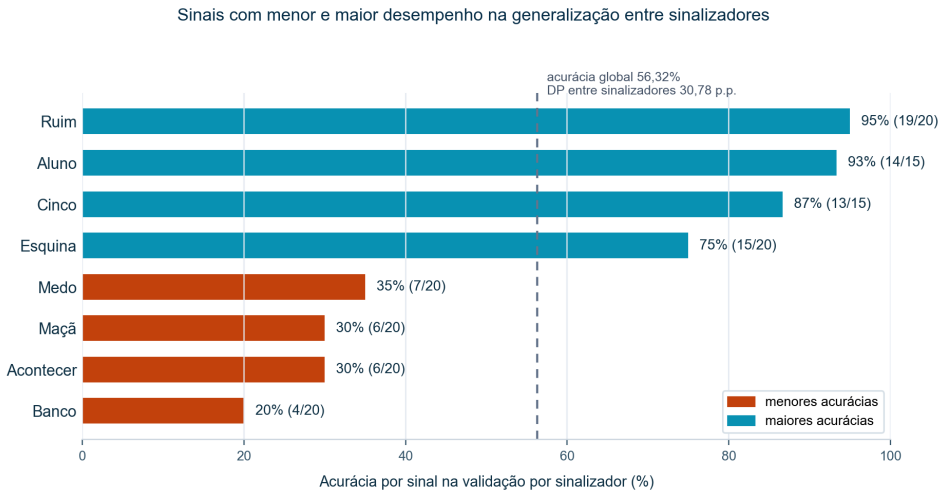
[3] <https://youtu.be/llYIiY7uLDY>

treinamento. Esse resultado sugere que a padronização das sequências em 24 *frames* preservou informações relevantes de configuração, posição e movimento das mãos.

Na validação por sinalizador, a acurácia foi reduzida para 56,32% (IC 95%: 51,29–61,21%), evidenciando maior dificuldade de generalização para indivíduos não observados. O desvio padrão de 30,78 pontos percentuais entre as quatro rodadas mostra variação expressiva de desempenho conforme o sinalizador mantido fora do treinamento. Essa queda é coerente com a variação natural na execução dos sinais, especialmente em aspectos como velocidade, amplitude, trajetória e posicionamento das mãos, fator também observado em abordagens de reconhecimento dinâmico da Libras (Albino; Coral, 2023).

A Figura 4 apresenta uma síntese visual dos sinais com menores e maiores desempenhos na validação por sinalizador. A escolha de destacar os extremos, em vez de representar todos os 20 sinais, favorece a leitura das classes mais estáveis e das mais críticas sem sobrecarregar a análise gráfica.

Figura 4 - Sinais com menor e maior acurácia na validação por sinalizador.



Fonte: Elaborado pela autora.

Conforme ilustrado na Figura 4, sinais como “Ruim”, “Aluno” e “Cinco” apresentaram maior estabilidade, enquanto “Banco”, “Acontecer”, “Maçã” e “Medo” concentraram os menores percentuais de acerto. Esses resultados indicam que parte dos erros está associada à proximidade visual entre trajetórias e regiões de execução, além da variação individual dos sinalizadores.

Os erros observados podem ser explicados por três fatores principais: semelhança geométrica entre sinais realizados em regiões próximas, perda parcial de dinâmica em movimentos curtos após a interpolação temporal e limitação de profundidade da captura por webcam monocular. Essas condições afetam especialmente sinais com trajetórias sutis ou visualmente próximas.

O k-NN apresentou desempenho inferior ao da Regressão Logística, sobretudo na validação por sinalizador, indicando menor robustez da comparação direta por similaridade diante da variação individual. Já o classificador especialista “Medo”/“Ruim”, com 87,50% de acurácia, mostrou que pares visualmente próximos podem se beneficiar de atributos resumidos da sequência, embora essa estratégia aumente a dependência de tratamentos específicos por classe. Na comparação exploratória entre Regressão Logística e k-NN, o teste pareado de Wilcoxon não indicou diferença estatisticamente conclusiva por tomada ($p = 0{,}0625$) nem por sinalizador ($p = 0{,}1250$), considerando o pequeno número de rodadas.

3.2 RECONHECIMENTO FACIAL GRAMATICAL

O canal facial foi avaliado com a base pública Grammatical Facial Expressions, da UCI, considerando os marcadores gramaticais selecionados para o estudo. A classe “sem marcador/neutro” representou a ausência predominante desses marcadores. A Tabela 2 apresenta os resultados médios obtidos nas estratégias de avaliação. As médias e os desvios padrão foram calculados entre os cinco marcadores, e os intervalos de confiança de 95% foram estimados pela distribuição t de Student.

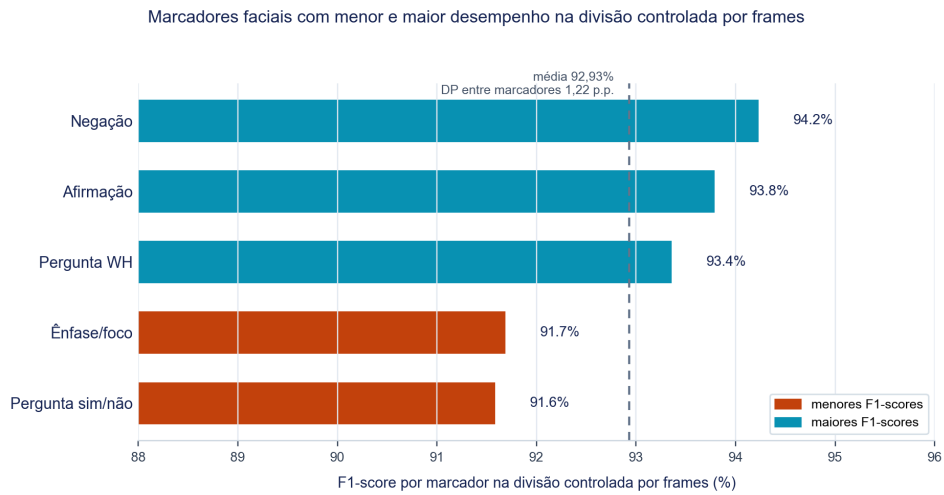
Tabela 2 - Resultados médios do reconhecimento facial gramatical

Métrica	**Média (%)**	**DP (p.p.)**	**IC 95%**
Validação estrita entre sinalizadores – Regressão Logística			
Acurácia	56,27	5,11	49,93–62,62%
Acurácia balanceada	57,22	7,09	48,42–66,02%
F1-score	50,72	12,22	35,54–65,89%
Divisão controlada por *frames* – Regressão Logística			
Acurácia	89,51	2,43	86,49–92,53%
Acurácia balanceada	89,39	2,19	86,67–92,10%
F1-score	87,57	1,57	85,62–89,52%
Divisão controlada por *frames* – *ExtraTreesClassifier*			
Acurácia	94,31	0,62	93,53–95,08%
Acurácia balanceada	93,92	0,72	93,03–94,82%
F1-score	92,93	1,22	91,41–94,45%

Fonte: Elaborado pela autora.

O *ExtraTreesClassifier* apresentou o melhor desempenho na divisão controlada por *frames*, com *F1-score* médio de 92,93%, desvio padrão de 1,22 ponto percentual e IC 95% de 91,41–94,45%. Esse resultado indica que os atributos faciais extraídos da base foram capazes de representar padrões relevantes dos marcadores gramaticais em ambiente controlado. A Figura 5 detalha esse desempenho por marcador, permitindo observar quais classes ficaram acima ou abaixo da média geral.

Figura 5 - Marcadores faciais com menor e maior *F1-score* na divisão controlada por *frames*.



Fonte: Elaborado pela autora.

Conforme apresentado na Figura 5, os melhores resultados ocorreram nos marcadores “negação”, “afirmação” e “pergunta WH”. Já “pergunta sim/não” e “ênfase/foco” ficaram abaixo da média geral, embora ainda tenham apresentado desempenho elevado na divisão controlada. Essa diferença pode estar associada à sutileza de alguns marcadores faciais e à dependência de variações individuais na movimentação de sobrancelhas, olhos, boca e cabeça.

Apesar do bom desempenho no cenário controlado, a validação estrita entre sinalizadores apresentou *F1-score* médio de 50,72%, evidenciando menor capacidade de generalização entre indivíduos. Esse resultado é relevante para a análise multimodal, pois mostra que o canal facial contribui para contextualizar a saída do sistema, mas sua robustez depende da diversidade dos dados utilizados no treinamento e da variabilidade das expressões produzidas pelos sinalizadores. Como a divisão controlada separa *frames* da própria base, seus resultados não devem ser interpretados como evidência de generalização para novos sinalizadores.

3.3 INTEGRAÇÃO MULTIMODAL

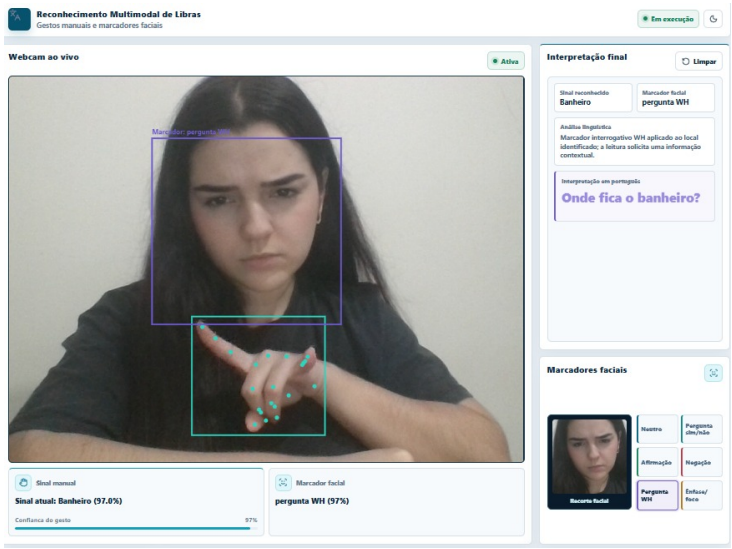
A integração multimodal reuniu as saídas dos canais manual e facial em uma interpretação contextual. O canal manual reconheceu o sinal dentro do vocabulário controlado, enquanto o canal facial indicou o marcador gramatical predominante durante sua execução. A saída final foi definida pela combinação entre essas informações, permitindo representar o sinal em forma neutra, afirmativa, negativa, interrogativa ou enfática.

Quando não houve marcador facial predominante, o sinal manual foi mantido em sua forma neutra. Nos demais casos, uma matriz semântica associou o item lexical ao tipo de marcador identificado, produzindo uma frase contextualizada em português. Essa etapa não pretende realizar tradução irrestrita da Libras, mas demonstrar como a análise conjunta de mãos e face pode enriquecer a interpretação em um vocabulário delimitado.

Não foi calculada uma acurácia multimodal fim a fim, pois os canais foram treinados com bases independentes. A base manual não possui anotações faciais gramaticais associadas aos vídeos, e a base facial não contém sinais manuais correspondentes. Portanto, uma métrica única de desempenho multimodal seria metodologicamente inadequada. Ainda assim, a execução funcional do protótipo em ambiente *web* permitiu verificar a viabilidade da composição entre os canais, preservando a avaliação separada dos modelos e aproximando a saída do sistema da natureza visuoespacial da Libras (Rastgoo; Kiani; Escalera, 2021).

A aplicação funcionou localmente com webcam e servidor Python. Como o tempo de execução não foi instrumentado, os resultados não permitem uma conclusão sobre a latência. A Figura 6 apresenta o sinal manual “Banheiro” combinado ao marcador facial “pergunta WH”, produzindo uma interpretação interrogativa em português.

Figura 6 - Exemplo de integração multimodal no protótipo *web*.



Fonte: Elaborado pela autora.

3.4 DISCUSSÃO E DESAFIOS

Os resultados mostram que a principal limitação do protótipo está na generalização para sinalizadores não observados. No canal manual, a diferença entre a validação por tomada (81,84%) e por sinalizador (56,32%) indica sensibilidade a variações individuais de velocidade, amplitude, trajetória e posicionamento das mãos. Essa queda de desempenho já era esperada ao se adotar modelos de classificação clássicos em vez de arquiteturas de modelagem temporal profunda (como *Transformers*). A escolha, contudo, foi consciente, visto que o objetivo principal desta pesquisa não foi superar o estado da arte em acurácia, mas demonstrar a viabilidade técnica da integração multimodal em um protótipo de baixo custo computacional.

Esse comportamento dialoga com os trabalhos de Fontana e Coral, que obtiveram alta acurácia em ambiente controlado, mas relataram instabilidades em testes com webcam (Fontana; Coral, 2025), e de Albino e Coral, cujo sistema também apresentou queda de desempenho diante de desafios de segmentação temporal (Albino; Coral, 2023). Em relação a esses estudos, o presente trabalho mantém o vocabulário controlado, mas avança ao integrar marcadores faciais gramaticais ao reconhecimento manual.

No canal facial, a queda entre a divisão controlada por *frames* e a validação entre participantes reforça que os marcadores gramaticais variam entre indivíduos. Esse resultado é coerente com estudos sobre ex-

pressões faciais em Libras, que destacam a relevância linguística da face e a complexidade de sua detecção automática (Silva et al., 2020). Assim, o canal facial contribui para contextualizar a saída do sistema, mas ainda depende de bases mais diversas para maior robustez.

Comparado a abordagens multimodais internacionais, que muitas vezes utilizam sensores específicos ou tratam outras línguas de sinais (Kumar; Roy; Dogra, 2018), este protótipo diferencia-se por focar na Libras e operar com webcam convencional. Essa escolha favorece a acessibilidade, mas também impõe limitações de iluminação, enquadramento, oclusão das mãos e ausência de profundidade real. Arquiteturas baseadas em *Transformers* e redes em grafos podem modelar dependências espaciais e temporais de forma mais explícita (Cui et al., 2023; Naz et al., 2023), mas sua comparação com os resultados deste trabalho exige bases, vocabulários e protocolos equivalentes.

Os intervalos de confiança e desvios padrão apresentados quantificam a variabilidade disponível nos registros do experimento. Entretanto, a ausência de predições pareadas amostra a amostra impediu a aplicação do teste de McNemar entre todos os classificadores, permanecendo como limitação da análise estatística.

4 CONCLUSÃO

Este trabalho teve como objetivo desenvolver e avaliar um protótipo *web* para reconhecimento multimodal da Libras por visão computacional. A separação entre reconhecimento lexical e informação gramatical facial permitiu combinar as saídas dos canais em uma interpretação contextual, sem assumir a tradução irrestrita da Libras.

No canal manual, a validação por tomada alcançou 81,84% de acurácia e a validação por sinalizador, 56,32%. No canal facial, o *ExtraTreesClassifier* obteve 94,31% de acurácia e *F1-score* médio de 92,93% na divisão controlada; a validação entre sinalizadores, contudo, confirmou limitações de generalização.

As limitações do estudo envolvem o vocabulário restrito, a quantidade limitada de sinalizadores nas bases, a ausência de uma base multimodal pareada e a dependência de condições adequadas de iluminação, enquadramento e visibilidade das mãos e da face. Diferenças individuais de velocidade, amplitude, trajetória e posicionamento das mãos também dificultaram a generalização entre sinalizadores. Como continuidade, recomenda-se ampliar as bases e o vocabulário, incluir dados multimodais

sincronizados, testar o sistema com mais sinalizadores e em cenários reais, investigar modelos temporais mais robustos, incluindo *Transformers*, redes em grafos e modelos híbridos, e preservar predições por rodada para análises estatísticas mais completas.

DECLARAÇÃO DE USO DE INTELIGÊNCIA ARTIFICIAL

Os autores declaram que ferramentas de inteligência artificial generativa foram utilizadas exclusivamente como apoio auxiliar à elaboração deste trabalho, incluindo assistência na revisão linguística, estruturação textual e organização de ideias. Nenhuma ferramenta de IA foi utilizada para substituição da autoria intelectual, análise científica, interpretação dos resultados ou tomada de decisões acadêmicas. A responsabilidade integral pelo conteúdo, originalidade, veracidade das informações e conformidade com os princípios de integridade científica permanece integralmente atribuída aos autores.

REFERÊNCIAS

ALAGHBAND, M.; MAGHROOR, H. R.; GARIBAY, I. **A survey on sign language literature**. Machine Learning with Applications, Elsevier, v. 14, p. 100504. 2023, Disponível em: <https://doi.org/10.1016/j.mlwa.2023.100504>. Acesso em: 03 jun. 2026.

ALBINO, M. d. S.; CORAL, S. Trabalho de Conclusão de Curso (Ciência da Computação), **Reconhecimento de Linguagem de Sinais Utilizando o Algoritmo Dynamic Time Warping para Auxiliar Surdos**. Criciúma: [s.n.], 2023.

ALMEIDA, S. G. M. et al. **MINDS-Libras Dataset**. 2019. Zenodo. Dataset. Disponível em: <https://zenodo.org/records/2667329>. Acesso em: 17 maio 2026.

Brasil. **Constituição da República Federativa do Brasil de 1988**. 1988. Diário Oficial da União, Brasília, DF. Disponível em: <https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>. Acesso em: 05 maio 2026.

Brasil. **Lei n. 10.436, de 24 de abril de 2002. Dispõe sobre a Língua Brasileira de Sinais – Libras e dá outras providências**. 2002. Diário Oficial da União, Brasília, DF. Disponível em: <https://www.planalto.gov.br/ccivil_03/leis/2002/l10436.htm>. Acesso em: 03 jun. 2026.

CUI, Z. et al. **Spatial-temporal transformer for end-to-end sign language recognition**. Complex & Intelligent Systems, Springer, v. 9, p. 4645–4656.

2023, Disponível em: <https://doi.org/10.1007/s40747-023-00977-w>. Acesso em: 03 jun. 2026.

FONTANA, G. A. d. S.; CORAL, S. Trabalho de Conclusão de Curso (Ciência da Computação), **Redes Neurais Convolucionais na Detecção e Interpretação da Linguagem de Sinais para a Assistência a Indivíduos com Deficiência Auditiva**. Criciúma: [s.n.], 2025.

FREITAS, F.; BARBOSA, F.; PERES, S. **Grammatical Facial Expressions**. 2014. UCI Machine Learning Repository. Dataset. Disponível em: <https://archive.ics.uci.edu/dataset/317/grammatical+facial+expressions>. Acesso em: 05 maio 2026.

GEURTS, P.; ERNST, D.; WEHENKEL, L. **Extremely randomized trees**. Machine Learning, Springer, v. 63, n. 1, p. 3–42. 2006, Disponível em: <https://doi.org/10.1007/s10994-006-6226-1>. Acesso em: 15 maio 2026.

GRINSZTAJN, L.; OYALLON, E.; VAROQUAUX, G. **Why do tree-based models still outperform deep learning on typical tabular data?** In: **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2022. 2022. v. 35, p. 507–520. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2022/hash/0378c7692da36807bdec87ab043cdadc-Abstract-Datasets_and_Benchmarks.html>. Acesso em: 03 jun. 2026.

IBGE. **Censo Demográfico 2022: Pessoas com deficiência e pessoas diagnosticadas com transtorno do espectro autista: resultados preliminares da amostra**. Rio de Janeiro: [s.n.], 2025. Disponível em: <https://www.ibge.gov.br/estatisticas/todos-os-produtos-estatisticas/22827-censo-demografico-2022.html?edicao=43453>. Acesso em: 03 jun. 2026.

KARTYNNIK, Y. et al. **Real-time facial surface geometry from monocular video on mobile gpus**. arXiv preprint arXiv:1907.06724. 2019, Disponível em: <https://arxiv.org/abs/1907.06724>. Acesso em: 02 jun. 2026.

KUMAR, P.; ROY, P. P.; DOGRA, D. P. **Independent bayesian classifier combination based sign language recognition using facial expression**. Information Sciences, Elsevier, v. 428, p. 30–48. 2018, Disponível em: <https://doi.org/10.1016/j.ins.2017.10.046>. Acesso em: 05 maio 2026.

LONG, Z. et al. **Sign language recognition based on facial expression and hand skeleton**. In: **2023 38th Youth Academic Annual Conference of Chinese Association of Automation**. [s.n.], 2023. 2023. Disponível em: <https://arxiv.org/abs/2407.02241>. Acesso em: 03 jun. 2026.

NAZ, N. et al. **Mipa-resgcn: a multi-input part attention enhanced residual graph convolutional framework for sign language recognition**. Computers and Electrical Engineering, Elsevier, v. 112, p. 109009. 2023, Disponível em: <https://doi.org/10.1016/j.compeleceng.2023.109009>. Acesso em: 03 jun. 2026.

PAIVA, F. A. d. S. et al. **Análise do papel das expressões não manuais na intensificação em libras**. DELTA: Documentação de Estudos em Linguística Teórica e Aplicada, v. 36, n. 4. 2020, Disponível em: <https://www.scielo.br/j/delta/a/ZkwFT3Nh3ncD4z8TpzjDK4d/>. Acesso em: 28 out. 2025.

QUADROS, R. M. d. **Libras: Conhecimentos básicos para a educação**. São Paulo: Parábola Editorial. 2019.

RASTGOO, R.; KIANI, K.; ESCALERA, S. **Sign language recognition: A deep survey**. Expert Systems with Applications, Elsevier, v. 164, p. 113794. 2021, Disponível em: <https://doi.org/10.1016/j.eswa.2020.113794>. Acesso em: 05 maio 2026.

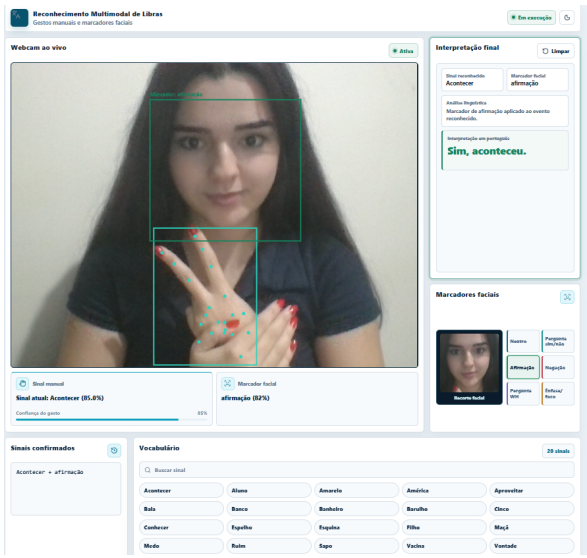
SILVA, E. P. et al. **Silfa: Sign language facial action database for the development of assistive technologies for the deaf**. In: **2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)**. [s.n.], 2020. 2020. p. 688–692. Disponível em: <https://ieeexplore.ieee.org/document/9320274>. Acesso em: 04 abril. 2026.

SILVA, E. P. d. et al. **Facial action unit detection methodology with application in brazilian sign language recognition**. Pattern Analysis and Applications, Springer, v. 25, p. 549–565. 2022, Disponível em: <https://doi.org/10.1007/s10044-021-01024-5>. Acesso em: 05 maio 2026.

ZHANG, F. et al. **Mediapipe hands: On-device real-time hand tracking**. arXiv preprint arXiv:2006.10214. 2020, Disponível em: <https://arxiv.org/abs/2006.10214>. Acesso em: 03 jun. 2026.

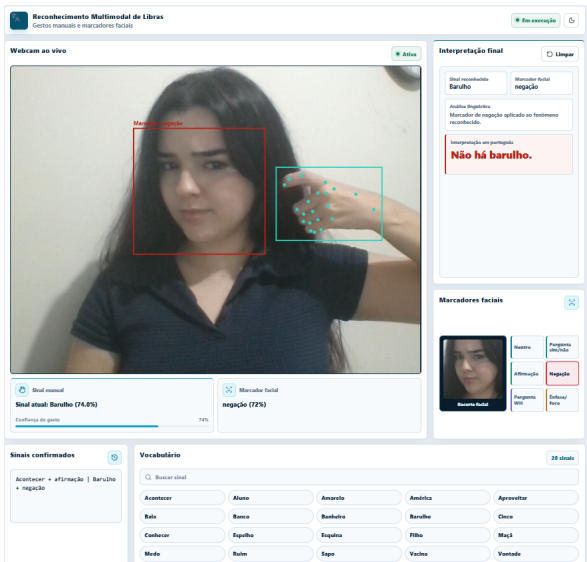
APÊNDICE A - IMAGENS DO PROTÓTIPO EM FUNCIONAMENTO

Figura A.1 - Reconhecimento do sinal “Acontecer” associado ao marcador facial de “afirmação”.



Fonte: Elaborado pela autora.

Figura A.2 - Reconhecimento do sinal “Barulho” associado ao marcador facial de “negação”.



Fonte: Elaborado pela autora.