UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CLEISON CORREIA DE AMORIM

Uma Abordagem Linguística para o Reconhecimento de Línguas de Sinais

Recife

2022

CLEISON CORREIA DE AMORIM

Uma Abordagem Linguística para o Reconhecimento de Línguas de Sinais

Trabalho apresentado ao Programa de Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Orientador: Cleber Zanchettin

Recife

2022

Catalogação na fonte
Bibliotecária Nataly Soares Leite Moro, CRB4-1722

A524a Amorim, Cleison Correia de
Uma abordagem linguística para o reconhecimento de línguas de sinais / Cleison Correia de Amorim. – 2022.
86 f.: il., fig., tab.

Orientador: Cleber Zanchettin.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2022.
Inclui referências.

1. Inteligência computacional. 2. Língua de sinais. 3. Linguística. 4. Processamento de linguagem natural. I. Zanchettin, Cleber (orientador). II. Título

006.31 CDD (23. ed.) UFPE - CCEN 2022 – 199

Cleison Correia de Amorim

Uma Abordagem Linguística para o Reconhecimento de Línguas de Sinais

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional

Aprovado em: 05/07/2022.

BANCA EXAMINADORA

_____

Prof. Dr. Leandro Maciel Almeida
Centro de Informática/ UFPE

_____

Prof. Dr. Byron Leite Dantas Bezerra
Escola Politécnica de Pernambuco / UPE

_____

Prof. Dr. Cleber Zanchettin
Centro de Informática / UFPE
(**Orientador**)

Dedico este trabalho ao meu pai, que partiu ao longo da caminhada de meu mestrado, mas deixou um grande exemplo de integridade e simplicidade pelo qual me espelhar.

AGRADECIMENTOS

Primeiramente, agradeço a Deus por sua infinita bondade e por me conceber essa oportunidade de encerrar mais um importante ciclo em minha vida. É impossível não recordar como os últimos anos foram difíceis para todos nós, bem como nos trouxeram desafios e perdas que tivemos de superar. No entanto, hoje respirando mais aliviado e com esperanças renovadas, celebro esta conquista acadêmica.

Em segundo lugar, agradeço à minha família por todo apoio e compreensão nessa jornada. De um modo especial, sou muito grato aos meus pais por acreditarem, persistirem e ensinarem a cada um de seus filhos que a educação é o legado mais valioso que eles poderiam nos transmitir e que jamais alguém poderia nos tirá-lo. Seus ensinamentos são muito sábios e fontes de muitas realizações.

Por fim, agradeço ao meu orientador por acreditar em meu potencial desde o primeiro momento e por me apoiar compartilhando muito de seu tempo e conhecimento para que eu pudesse concretizar esta pesquisa.

"Aprenda suas teorias o melhor que puder, mas deixe-as de lado ao tocar o milagre de uma alma humana." (JUNG, 1928, p. 361, tradução nossa)

RESUMO

A língua de sinais é uma ferramenta essencial na vida do Surdo, capaz de assegurar seu acesso à comunicação, educação e desenvolvimento cognitivo e socio-emocional. Na verdade, ela é a principal força que une essa comunidade e o símbolo de identificação entre seus membros. Contudo, atualmente o número de indivíduos ouvintes que conseguem se comunicar por meio dessa língua é pequeno e, na prática, isso impõe obstáculos ao cotidiano do Surdo. Tarefas simples como utilizar o transporte público, comprar roupas, ir ao cinema ou obter assistência médica acabam tornando-se um desafio por conta dessa limitação. O Reconhecimento de Língua de Sinais é uma das áreas de pesquisa que objetiva desenvolver tecnologias capazes de reduzir essas barreiras linguísticas e facilitar a comunicação entre ambos indivíduos. Apesar disso, ao analisar sua evolução ao longo das últimas décadas, percebe-se que seu progresso ainda não é suficiente para disponibilizar soluções efetivamente aplicáveis ao mundo real. Isso ocorre principalmente porque várias pesquisas na área acabam não apropriando-se ou abordando adequadamente as particularidades linguísticas das línguas de sinais, decorrentes de sua natureza visual. Tendo isso em vista, este trabalho apresenta uma abordagem que aplica modelos sequenciais de aprendizagem de máquina para realizar o reconhecimento computacional dos sinais através de seus atributos linguísticos. Além disso, são introduzidos dois novos *datasets* para a língua de sinais, dentre os quais está um *dataset* de atributos linguísticos computados a partir do ASLLVD. Com isso, objetiva-se estabelecer uma direção capaz de conduzir a avanços mais efetivos para essa área e, consequentemente, contribuir com a superação dos obstáculos hoje enfrentados pelo Surdo.

Palavras-chaves: língua de sinais; linguística; processamento de linguagem natural.

ABSTRACT

Sign language is an essential resource to ensure the Deaf to have access to communication, education, as well as to cognitive and socio-emotional development. In fact, it is the main force that unites such community and the key trait that identifies its members. On the other hand, the number of hearing individuals who are able to communicate through this language is currently small and, in practice, this imposes obstacles to the daily life of the Deaf. Simple tasks like using public transportation, shopping for clothes, going to the movies, or getting medical assistance end up becoming challenges due to such limitation. The Sign Language Recognition, in turn, is one of the research areas dedicated to developing technologies that aim to reduce such language barriers and facilitate the communication between these individuals. However, when analyzing its evolution over the last decades, we realize that it has not progressed enough to provide solutions effectively applicable to the real world. This is mainly because several researches in this field do not appropriate or address the linguistic particularities presented by the sign languages, which stem from their visual nature. Considering this problem, the present work introduces an approach that adopts sequential machine learning models to recognize signs through their linguistic attributes. In addition, we introduce two new sign language datasets, among which is a novel dataset of linguistic attributes computed from the ASLLVD. Thus, we aim to establish a direction that can lead to more effective advances in this research area and, consequently, contribute to overcoming the obstacles faced by the Deaf today.

Keywords: sign language; linguistics; natural language processing.

LISTA DE FIGURAS

LISTA DE CÓDIGOS-FONTES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizagem de Máquina
AP	Aprendizagem Profunda
ArSL	*Arabic Sign Language* (Língua de Sinais Britânica)
ASL	*American Sign Language* (Língua de Sinais Americana)
ASLLRP	*American Sign Language Linguistic Research Project*
ASLLVD	*American Sign Language Lexicon Video Dataset*
AUSLAN	*Australian Sign Language* (Língua de Sinais Austrália)
BHOF	*Block-based Histogram of Optical Flow* (ou Histograma Baseado em Blocos de Fluxo Óptico)
BSL	*British Sign Language* (Língua de Sinais Britânica)
CEL	*Cross-Entropy Loss* (ou Perda de Entropia Cruzada)
CETENE	Centro de Tecnologias Estratégicas do Nordeste
CNN	*Convolutional Neural Network* (ou Rede Neural Convolucional)
CSL	*Chinese Sign Language* (Língua de Sinais Chinesa)
DGS	*Deutsche Gebärdensprache* (Língua de Sinais Alemã)
FLOPS	*Floating-Point Operations Per Second* (ou Operações de Ponto Flutuante por Segundo)
GB	Gigabyte
GRU	*Gated Recurrent Unit*
GSL	*Greek Sign Language* (Língua de Sinais Grega)
HCORF	*Hidden Conditional Ordinal Random Fields* (ou Campos Aleatórios Ordinais Condicionais Ocultos)
HEI	*Hand Energy Image* (ou Imagem de Energia da Mão)
HMM	*Hidden Markov Model* (ou Modelo Oculto de Markov)
HOF	*Histogram of Optical Flow* (ou Histograma de Fluxo Óptico)
IA	Inteligência Artificial

INES	Instituto Nacional de Educação de Surdos
IPA	*International Phonetic Alphabet* (ou Alfabeto Fonético Internacional)
ISL	*Indian Sign Language* (Língua de Sinais Indiana)
JSL	*Japanese Sign Language* (Língua de Sinais Japonesa)
Libras	Língua Brasileira de Sinais
LIS	*Lingua dei Segni Italiana* (Língua de Sinais Italiana)
LSA	*Lengua de Señas Argentina* (Língua de Sinais Argentina)
LSTM	*Long Short-Term Memory*
MB	Megabyte
MEI	*Motion Energy Image* (ou Imagem de Energia de Movimento)
MHI	*Motion History Image* (ou Imagem do Histórico de Movimento)
NGT	*Nederlandse Gebarentaal* (Língua de Sinais Holandesa)
PCA	*Principal Component Analysis* (ou Análise de Componente Principal)
PJM	*Polski Język Migowy* (Língua de Sinais Polonesa)
PLN	Processamento de Linguagem Natural
RLS	Reconhecimento de Língua de Sinais
RNA	Rede Neural Artificial
RNN	*Recurrent Neural Network* (ou Rede Neural Recorrente)
RNP	Rede Neural Profunda
SGD	*Stochastic Gradient Descent* (ou Gradiente Estocástico Descendente)
ST-GCN	*Spatial-Temporal Graph Convolutional Network* (ou Rede Convolucional de Grafos Espaço-Temporais)
TA	Tradução Automática
TID	*Türk İşaret Dili* (Língua de Sinais Turca)
TSL	*Taiwan Sign Language* (Língua de Sinais de Taiwan)
VC	Visão Computacional
VGT	*Vlaamse Gebarentaal* (Língua de Sinais Flamenga)

SUMÁRIO

1 INTRODUÇÃO

Segundo a World Health Organization (2021), o mundo possui hoje cerca de 1,5 bilhões de pessoas com algum grau de perda auditiva, o que corresponde a 19% da população mundial. Desse número, 450 milhões se referem a perda de grau moderado a total[1], as quais necessitarão de acesso a cuidados auditivos e outros serviços de reabilitação. No Brasil, o número dos que têm perda auditiva é de 10,7 milhões e o dos que apresentam perda moderada a total é de 2,3 milhões, de acordo com Agência Brasil (2019), IBGE (2021a), IBGE (2021b).

Ao analisar os dispositivos e intervenções que são capazes de auxiliar no diagnóstico e reabilitação desses indivíduos, a World Health Organization (2021) afirma que as últimas décadas testemunharam avanços revolucionários, como nos campos da tecnologia auditiva, diagnóstico e telemedicina, com inovações que possibilitam problemas relacionadas à audição serem identificadas em qualquer idade e ambiente. Contudo, uma vez que a grande maioria daqueles com perda auditiva vive em locais de baixa renda, onde profissionais especializados e serviços para cuidados auditivos não estão comumente disponíveis, existe uma disparidade no acesso a tais recursos:

> Cerca de 78% dos países de baixa renda têm menos de um otorrinolaringologista por milhão de habitantes; 93% têm menos de um audiologista por milhão; 17% têm um ou mais fonoaudiólogos por milhão; e 50% têm um ou mais professores para pessoas com perda auditiva por milhão.
>
> Mesmo em países com proporções relativamente altas desses especialistas, há uma distribuição desigual que, além de trazer desafios para essa população, impõe uma sobrecarga excessiva aos quadros que prestam esses serviços. (OPAS, 2021)

Essa carência de cuidados adequados gera deficiências estruturais que dificultam o acesso a oportunidades básicas como educação e emprego, e resultam numa pior qualidade de vida para eles. Isso estende-se a todas as gerações, afirmam Agência Brasil (2021), Agência Brasil (2019):

> Essas deficiências estruturais refletem-se na educação das crianças. Uma criança que ouve mal, aprende mal e torna-se um adulto menos capaz do que poderia ser, e assim por diante. (Agência Brasil, 2021)
>
> Uma vez que esses indivíduos têm menos oportunidades de estudar e acessar o mercado de trabalho do que a população ouvinte, o dinheiro para conseguir

[1] O grau de perda refere-se à intensidade mínima de som que um ouvido pode detectar. Na perda leve, essa intensidade está entre 20 e 34 dB; na moderada, entre 35 e 49 dB; na moderadamente severa, entre 50 e 64 dB; na severa, entre 65 e 79 dB; na profunda, entre 80 e 94 dB; e na total (ou surdez), 95 dB ou mais. (World Health Organization, 2021, p. 38)

> o aparelho auditivo é ainda mais difícil. Esse conjunto de preconceitos acaba criando um círculo vicioso que não possibilita que eles tenham as mesmas oportunidades de se dar bem na vida. (Agência Brasil, 2019)

Em meio a tantos desafios, a língua de sinais surge como uma ferramenta poderosa que é capaz de assegurar o desenvolvimento cognitivo, facilitar a comunicação, e possibilitar que esses indivíduos obtenham educação e desenvolvimento socio-emocional adequado (World Health Organization, 2021). Ela pode ser aprendida através de membros da própria família, da comunidade Surda, de conteúdos geralmente gratuitos na internet (como livros, cursos e vídeos) disponibilizados por instituições como o INES[2] ou em escolas públicas que ofertam seu ensino. Por conta disso, ela torna-se uma alternativa acessível para a inclusão desses indivíduos, uma vez que muitas das barreiras como a demanda por recursos financeiros ou profissionais especializados são removidas.

Stewart e Stewart (2021) afirmam que a língua de sinais é a chave para acessar a cultura Surda. O termo Surdo (escrito com "s" capitalizado), por sua vez, não refere-se apenas a uma condição clínica, mas a um grupo de indivíduos que, além de possuírem perda auditiva, utilizam a língua de sinais como principal meio de comunicação e compartilham experiências culturais associadas à surdez e ao uso dessa língua. Há um aspecto cultural fundamental, reiteram Pereira et al. (2011), acompanhado de um forte sentimento de identidade grupal que faz com que esses indivíduos compartilhem valores, crenças, comportamentos e uma língua própria.

Apesar disso, Bragg et al. (2019), Agência Senado (2019) observam que atualmente ainda são poucos os ouvintes que conseguem se comunicar por meio dessa língua. Isso traz obstáculos adicionais aos Surdos e transforma muitas de suas atividades corriqueiras num grande desafio. Por exemplo, no transporte público é difícil solicitar ajuda ou ter acesso às instruções divulgadas nos alto-falantes; em lojas, é raro encontrar vendedores preparados para interagir através dessa língua ou que não os trate com preconceito; no cinema, eles apenas podem consumir filmes estrangeiros, uma vez que os nacionais não dispõem de legenda; no serviço de saúde, não são raros os relatos de pacientes que saem de consultas com prescrições médicas erradas porque o médico não entendeu corretamente seus sintomas; entre outras situações.

Para contribuir com a superação desses desafios é importante, entre outros fatores, que a

[2] O Instituto Nacional de Educação de Surdos (INES), fundado em 1857, é o centro de referência nacional que subsidia a formulação de políticas públicas para o Surdo. Ele atende a estudantes da educação infantil até o ensino superior e também apoia a pesquisa de novas metodologias de ensino nesse contexto. (Ministério da Educação e Cultura, 2021)

comunidade acadêmica esteja mobilizada para impulsionar o desenvolvimento de alternativas e tecnologias. O Reconhecimento de Língua de Sinais (RLS) é um dos campos de pesquisa que se dedica a desenvolver algumas delas. Segundo Wadhawan e Kumar (2019), trata-se de uma área colaborativa e multidisciplinar que envolve Visão Computacional (SZELISKI, 2022), Processamento de Linguagem Natural (JURAFSKY; MARTIN, 2022), Reconhecimento de Padrões (BISHOP, 2006) e Linguística (QUADROS; KARNOPP, 2004) para construir métodos e algoritmos capazes de identificar sinais produzidos pelo articulador e compreender seu significado. Por meio deles, seria possível reduzir a barreira linguística entre Surdos e ouvintes permitindo que mensagens transmitidas utilizando-se a língua de sinais fossem transcritas automaticamente e compreendidas por aqueles que não a conhecem.

No entanto, apesar do potencial que o RLS possui, Selvaraj et al. (2022), Yin et al. (2021), Cooper, Holt e Bowden (2011) acreditam que o progresso apresentado por essa área ao longo das últimas décadas foi insuficiente para conduzir a avanços expressivos:

> Quando comparado com a pesquisa de Processamento de Linguagem Natural baseada em texto e fala, o progresso das pesquisas para línguas de sinais está significativamente atrasado. (SELVARAJ et al., 2022; YIN et al., 2021, tradução nossa)

> Enquanto sistemas de reconhecimento da fala avançaram ao ponto de estarem comercialmente disponíveis, o reconhecimento de sinais ainda está em sua infância. Atualmente, todos os serviços comerciais de tradução de sinais são baseados em humanos e requerem que pessoal especializado esteja disponível, o que os tornam caros e pouco acessíveis. (COOPER; HOLT; BOWDEN, 2011, tradução nossa)

Isso deve-se, de um modo geral, a um conjunto de particularidades que as línguas de sinais apresentam quando comparadas às línguas faladas, bem como à forma com que as pesquisas em RLS têm abordado elas, afirmam Bragg et al. (2019), Cooper, Holt e Bowden (2011). Diferentemente das faladas, as línguas sinalizadas possuem uma natureza visual e transmitem significado através de múltiplos canais ao mesmo tempo, como mãos, corpo, face, entre outros de granularidade ainda menor. Essa natureza faz com que sua linguística seja estruturada de uma forma muito específica, demandando que novas técnicas sejam desenvolvidas para abordar tais particularidades.

Contudo, segundo Cooper, Holt e Bowden (2011), Yin et al. (2021), um grande número de pesquisas nessa área trata o RLS como uma tarefa de reconhecimento de gestos não-estruturados ou poses de mãos estáticas, que são mapeados a partir de imagens RGB, dados de luvas eletrônicas ou coordenadas dos corpos dos indivíduos. Isso faz com que elas deixem de

abordar aspectos essenciais da língua de sinais – como sua linguística e suas particularidades – e desviem o foco para um conjunto de desafios pertinentes à área de Visão Computacional (VC), como a detecção, segmentação e rastreamento de partes do corpo; a interação entre mãos e oclusões decorrentes disso; variações de tom de pele; entre outros que comumente já são abordados ou solucionados por outras subáreas da VC. Como consequência, essas pesquisas acabam não produzindo avanços realmente efetivos para a RLS.

Tendo em vista isso, este trabalho busca contribuir com a área de Reconhecimento de Língua de Sinais (RLS) por meio de uma proposta que aborda a língua de sinais através de sua linguística, pela introdução de um novo *dataset* de atributos linguísticos e pela adoção de técnicas de Processamento de Linguagem Natural (PLN) nesse contexto afim de estabelecer uma direção que possa conduzir a avanços efetivos na área e, consequentemente, ajude a superar alguns dos desafios cotidianos atualmente encarados pelos Surdos.

1.1 OBJETIVOS

O objetivo geral deste trabalho consiste em propor uma abordagem de Reconhecimento de Língua de Sinais baseada na linguística, a qual considere as complexidades de sua natureza visual e preencha algumas das lacunas deixadas em aberto por pesquisas na área, afim de contribuir com avanços mais efetivos.

Como objetivos específicos, este trabalho busca alcançar:

- Propor uma estratégia para computar atributos linguísticos a partir de *datasets* existentes da língua de sinais;
- Disponibilizar um *dataset* de atributos linguísticos, o qual atualmente é inexistente, para suportar o desenvolvimento de novas técnicas para as línguas de sinais;
- Identificar, aplicar e avaliar algoritmos que possibilitem abordar o reconhecimento da língua de sinais através de sua linguística, acomodando as complexidades de sua natureza visual.

1.2 METODOLOGIA

A metodologia aplicada neste trabalho consistiu em primeiro compreender os desafios atuais do Surdo, da língua de sinais e da área de pesquisa para, em seguida, estabelecer uma

proposta capaz de abordar algumas das lacunas encontradas e avaliar seus resultados. As etapas percorridas incluem:

- Revisão do panorama do Surdo, das línguas de sinais e de sua linguística;
- Revisão da área de Reconhecimento de Língua de Sinais e das lacunas existentes;
- Elaboração de uma proposta que aborde as lacunas acima e produza artefatos que suportem novas pesquisas nessa direção;
- Realização de experimentos e análise dos resultados.

1.3 ORGANIZAÇÃO DO TRABALHO

Além do capítulo atual, esta dissertação está estruturada em mais quatro capítulos, que estão organizados da seguinte forma:

O Capítulo 2 apresenta o referencial teórico, que contém conceitos importantes que fundamentam a dissertação. A abordagem adotada nesta pesquisa é discutida no Capítulo 3, no qual também são apresentadas as hipóteses e técnicas utilizadas, bem como a preparação dos experimentos realizados. No Capítulo 4, por sua vez, são analisados os resultados desses experimentos. Por fim, no Capítulo 5 são discutidas as conclusões e um levantamento de propostas para trabalhos futuros, com base nas descobertas obtidas.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta conceitos fundamentais para compreensão do contexto das línguas de sinais e da abordagem proposta por esta pesquisa. Primeiro, será introduzido o contexto do Surdo e da língua de sinais na seção 2.1. Em seguida, a seção 2.2 se aprofundará na linguística e nas particularidades dessa língua. A seção 2.3, por sua vez, abordará a Aprendizagem de Máquina e alguns dos algoritmos aplicados ao Processamento de Linguagem Natural. Por fim, a seção 2.4 discutirá o panorama atual e os desafios existentes para a área de Reconhecimento de Língua de Sinais.

2.1 O SURDO E A LÍNGUA DE SINAIS

Segundo Pereira et al. (2011), a língua de sinais é a língua utilizada pela maioria dos Surdos em sua vida diária. Muito mais do que isso, ela é a principal força que une essa comunidade e o símbolo de identificação entre seus membros.

Quando utilizados os termos "Surdo" ou "comunidade Surda" (com "s" capitalizado) não nos referimos apenas a uma condição clínica, mas a um grupo de indivíduos que, além de possuírem perda auditiva, utilizam a língua de sinais como principal meio de comunicação e compartilham experiências culturais associadas à surdez e ao uso dessa língua. Esses elementos estão interconectados e não pode-se definir comunidade Surda sem considerá-los como um todo.

De fato, ter uma perda auditiva não significa que uma pessoa seja membro da comunidade Surda ou que automaticamente saiba sinalizar, embora certamente esses sejam requisitos importantes:

> Não há como apontar uma pessoa sentada num saguão lendo uma revista e dizer: "esta pessoa é Surda". Ainda que ela esteja utilizando aparelhos auditivos, não há como saber com qual comunidade ela se identifica.
>
> De maneira semelhante, não importa se ela é europeia, afro-americana, asiática ou de outra etnia; sua idade não é relevante, tampouco sua classe social ou gênero; a comunidade Surda não é moldada por nenhuma dessas características. (STEWART; STEWART, 2021, tradução nossa)

Há um aspecto cultural essencial acompanhado de um forte sentimento de identidade grupal, que fazem com que a surdez seja percebida como uma diferença e não como uma deficiência, reiteram Pereira et al. (2011).

Apesar disso, no decorrer da história os Surdos tiveram sua identidade estigmatizada e desvalorizada pela sociedade ouvinte, que não aceitava a língua de sinais. Segundo Hill, Lillo-Martin e Wood (2019), há até relativamente pouco tempo dizia-se a esses indivíduos que sua forma de comunicação era inferior, quebrada, sem importância ou insuficiente. Os sistemas educacionais e a comunidade majoritariamente ouvinte fizeram prevalecer a utilização da língua falada, mesmo às custas da língua de sinais.

De fato, atitudes como essas ainda persistem. Pereira et al. (2011) comentam que muitos ouvintes tentam diminuir os Surdos para que eles vivam isolados, tendo de assumir a cultura ouvinte como se ela fosse a única existente. Ser "normal" significaria, portanto, poder ouvir e falar oralmente:

> Os ouvintes não prestam atenção aos Surdos que se comunicam por meio da língua de sinais. Consequentemente, não acreditam que eles sejam capazes de estudar em faculdade ou realizar mestrado e doutorado, por exemplo. (PEREIRA et al., 2011)

> Os ouvintes veem os Surdos com curiosidade e, às vezes, zombam deles por serem diferentes. (STROBEL, 2016)

São muitas as lutas pelas quais os Surdos têm atravessado mas que, sobretudo nos últimos anos, têm conduzido a vitórias importantes como o reconhecimento das línguas de sinais oficialmente como línguas em diversos países, o direito a tradutores e intérpretes em eventos e canais públicos de comunicação e o acesso a uma educação bilíngue para as crianças Surdas, entre outras conquistas. No Brasil, por exemplo, a Língua Brasileira de Sinais (Libras) foi reconhecida como uma língua em 2002 e, nos Estados Unidos, a *American Sign Language* (Língua de Sinais Americana) (ASL) foi reconhecida ainda em 1989 (BRASIL, 2002; PEREIRA et al., 2011; JAY, 2011).

De acordo com Hill, Lillo-Martin e Wood (2019), Pereira et al. (2011), as línguas de sinais distinguem-se das línguas orais porque utilizam o canal visual-espacial em vez do oral-auditivo. Por esse motivo, são denominadas línguas de modalidade gestual-visual, onde a informação linguística é recebida pelos olhos e produzida no espaço pelas mãos, pelo movimento do corpo e pela expressão facial. Devido a disso, acrescentam Stewart e Stewart (2021), não existe uma forma escrita conveniente para essas línguas, mas apenas glosas que representam uma aproximação do significado de seus sinais.

Elas são consideradas como línguas naturais, ou seja, aquelas que emergem naturalmente quando indivíduos se reúnem formando uma comunidade. Assim como ocorre para as línguas

orais, não existe uma universalidade: cada país tem sua própria língua sinalizada e elas, por sua vez, refletem a cultura dos diferentes países em que são utilizadas.

Apesar das particularidades acima, línguas orais e línguas sinalizadas compartilham os mesmos princípios quanto ao fato de que possuem um léxico e uma gramática. Ou seja, ambas apresentam um conjunto de símbolos convencionais bem como um sistema de regras que rege a combinação desses símbolos em unidades maiores de significado. Dessa forma, a seção seguinte explorará em maior profundidade esses elementos linguísticos presentes nas línguas de sinais.

2.2 LINGUÍSTICA DA LÍNGUA DE SINAIS

Quadros e Karnopp (2004) afirmam que a linguística é o estudo científico das línguas naturais e humanas. Trata-se de uma ciência que procura desvendar os princípios independentes da lógica e da informação que determinam essa linguagem, bem como todas as formas criativas da comunicação. Ela busca respostas para problemas essenciais relacionados à linguagem como, por exemplo: "qual a natureza da linguagem humana?", "como a comunicação se constitui?", "quais os princípios que determinam a habilidade dos seres humanos de produzir e compreender a linguagem?".

Segundo Stewart e Stewart (2021), os primeiros estudos linguísticos sobre as línguas de sinais foram realizados pelo professor Dr. William C. Stokoe Jr. da Universidade de Gallaudet, em 1960. Seu primeiro artigo, intitulado "*Sign Language Structure*" (ou Estrutura da Língua de Sinais), foi seguido pela publicação do primeiro dicionário da *American Sign Language* (Língua de Sinais Americana) (ASL) em 1965 – o "*Dictionary of American Sign Language on Linguistic Principles*" (ou Dicionário da Língua de Sinais Americana em Princípios Linguísticos) – que foi compilado em parceria com dois colegas Surdos. Em 1971, por sua vez, ele estabeleceu o Laboratório de Pesquisa em Linguística da Universidade de Gallaudet.

Por conta disso, Stokoe ficou conhecido como o pai da linguística das línguas de sinais e seu trabalho teve um impacto profundo na conscientização acerca da ASL nos Estados Unidos e no restante do mundo.

Em seus estudos, ele comprovou que a ASL atendia a todos os critérios linguísticos de uma língua genuína – no léxico, na sintaxe e na capacidade de gerar uma quantidade infinita de sentenças. A análise de suas propriedades revelou que a língua de sinais apresenta organização formal nos mesmos níveis encontrados nas línguas faladas, incluindo um nível sublexical

de estruturação interna do sinal (análoga ao nível fonológico das línguas orais) e um nível gramatical (morfossintático), que especifica os modos como os sinais devem ser combinados para formarem frases e orações. Dessa forma, comprovou-se que os sinais não são meras imagens, mas símbolos abstratos com uma complexa estrutura interior. Aos estudos de Stokoe seguiram-se outros, que estenderam o escopo de análise às línguas de sinais utilizadas em diferentes países, como França, Itália, Uruguai, Argentina, Suécia e Brasil (STOKOE, 1960; QUADROS; KARNOPP, 2004; PEREIRA et al., 2011).

Serão abordados nas sessões seguintes aspectos importantes da organização gramatical da língua de sinais, segundo sua fonologia, morfologia e sintaxe.

2.2.1 Fonologia

Fonologia é o estudo das menores unidades constituintes de uma língua – denominados fonemas – e das regras que regem sua produção. Ela objetiva compreender essas unidades, bem como elas são articuladas, para compor unidades maiores com significado, como as palavras, de acordo com Quadros e Karnopp (2004), Hill, Lillo-Martin e Wood (2019).

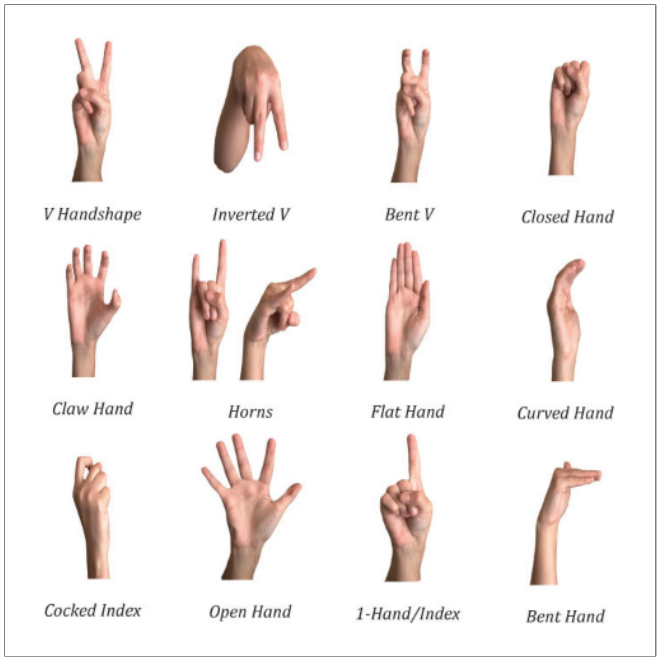
Stokoe (1960) definiu inicialmente três tipos de fonemas (ou parâmetros) para a língua de sinais, os quais são articulados simultaneamente para compor um sinal: configuração de mão, locação e movimento. Em 1974, BATTISON introduziu um quarto parâmetro, referente à orientação da palma da mão. Posteriormente, estudos como o de Baker e Padden (1978), adicionaram as expressões não-manuais, como expressões faciais, movimentos da boca e direção do olhar.

Dessa forma, atualmente a fonologia da língua de sinais compreende que os sinais são compostos pelos seguintes parâmetros (STEWART; STEWART, 2021; JAY, 2011; QUADROS; KARNOPP, 2004):

1. **Configuração de mão**: configuração assumida pelas mãos ao produzir o sinal, a qual pode permanecer estática ou variar durante a articulação do sinal. É possível que as mãos apresentem configurações distintas nesse processo. A Figura 1 ilustra algumas configurações utilizadas na ASL.

2. **Orientação**: direção apontada pelas palmas das mãos na articulação do sinal. Por exemplo, as palmas podem estar voltadas para o corpo, para fora, para o chão, para cima,

Figura 1 – Exemplos de configurações de mãos utilizadas na ASL



Fonte: Jay (2011, p. 72)

entre outras ilustradas na Figura 2. Cada uma das palmas pode também assumir uma orientação distinta.

Figura 2 – Exemplos de orientações que podem ser assumidas pelas palmas das mãos

Fonte: Quadros e Karnopp (2004, p. 59)

3. **Movimento**: corresponde à trajetória percorrida pelas mãos em relação ao corpo para

articular o sinal. É um parâmetro complexo que pode envolver uma ampla variedade de modos e direções, desde um sutil deslizar entre as mãos ou um movimento interno das mãos e punhos, até uma trajetória complexa desenhada no espaço, por exemplo. Além disso, os sinais podem envolver movimentos de uma ou de ambas as mãos.

4. **Locação**: é o local onde as mãos são posicionadas dentro do espaço de enunciação para articular o sinal. O espaço de enunciação, por sua vez, é uma área que contém todos os pontos possíveis dentro do raio de alcance das mãos, como ilustra a Figura 3. Nesse espaço, há um número limitado de locações, sendo que algumas são mais exatas – tais como a ponta do nariz –, e outras são mais abrangentes – como a frente do tórax. Por fim, as mãos podem permanecer fixas ou se deslocar de uma locação para outro durante a articulação de um sinal.

Figura 3 – Espaço de enunciação da língua de sinais

Fonte: Quadros e Karnopp (2004, p. 57)

5. **Expressões não-manuais**: consistem nas expressões faciais e movimentos corporais incorporados aos sinais para provê significado adicional. Elas desempenham duas funções essenciais: marcar construções sintáticas (como frases interrogativas, orações relativas, tópicos, concordância e foco) e diferenciar componentes lexicais (como referências específicas, referências pronominais, partículas negativas, advérbios, grau ou aspecto).

   De um modo geral, expressões faciais ajudam a prover mais clareza ou alterar o significado de um sinal. Movimentos corporais, por sua vez, são importantes para descrever pessoas em diferentes posições ou locais, ou narrar histórias envolvendo personagens com diferentes papéis, por exemplo.

2.2.2 Morfologia

Morfologia é o estudo da estrutura interna das palavras e das regras que determinam sua formação. Um morfema é a menor unidade indivisível de sintaxe que retém significado e, na língua de sinais, é tido como a combinação da configuração de mão, orientação, locação e movimento, afirmam Quadros e Karnopp (2004), Jay (2011), Hill, Lillo-Martin e Wood (2019).

Em línguas faladas, palavras complexas são muitas vezes formadas adicionando-se um prefixo ou sufixo a uma raiz. Por exemplo, o adjetivo "infeliz" é constituído de dois morfemas: o prefixo negativo *in-* e o adjetivo *feliz*; o substantivo "capacidade", por sua vez, é composto pelo adjetivo *capaz* acrescido do sufixo *-idade*; já o substantivo "guarda-chuva" é constituído pelos morfemas *guarda* e *chuva*.

De acordo com Klima e Bellugi (1975), Quadros e Karnopp (2004), nas línguas de sinais essas formações resultam frequentemente de processos em que uma raiz é enriquecida com movimentos e contornos no espaço de sinalização. Também são utilizadas expressões não-manuais, alterações nos parâmetros fonológicos ou sinais específicos para indicar tempo, grau, intensidade, pluralidade, aspecto, entre outros.

Serão discutidos a seguir os dois principais processos de formação de palavras apresentados pela morfologia tradicional, a derivação e a flexão, sob a perspectiva das línguas de sinais (QUADROS; KARNOPP, 2004; HILL; LILLO-MARTIN; WOOD, 2019; KLIMA; BELLUGI, 1979):

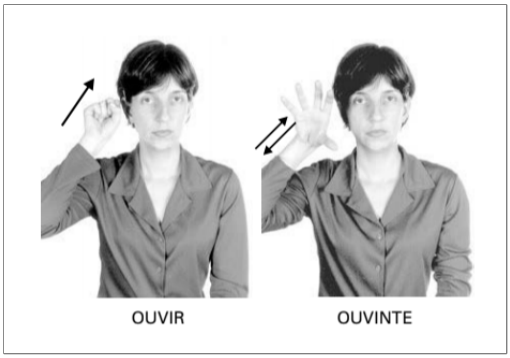
1. **Derivação**: consiste na formação de novas palavras a partir de uma mesma base lexical, como nos exemplos "infeliz" e "capacidade" introduzidos acima. No contexto das línguas de sinais, esses processos derivacionais podem incluir:

    a) Nominalização: consiste na derivação de substantivos a partir de verbos, e é um dos processos mais comuns para mudança de classe na morfologia.

    Na língua de sinais, os substantivos apresentam basicamente os mesmos parâmetros fonológicos que os verbos, mas diferenciam-se pela repetição (ou reduplicação) do seu movimento. A Figura 4 ilustra um exemplo de derivação do substantivo OUVINTE a partir do verbo OUVIR.

    b) Composição: consiste na criação de um novo sinal através da junção de duas bases preexistentes. Existem três regras para a composição de sinais: a regra do contato, na qual o contato existente no primeiro ou segundo sinal da composição

Figura 4 – O verbo OUVIR (à esquerda) é utilizado para derivar o substantivo OUVINTE (à direita)

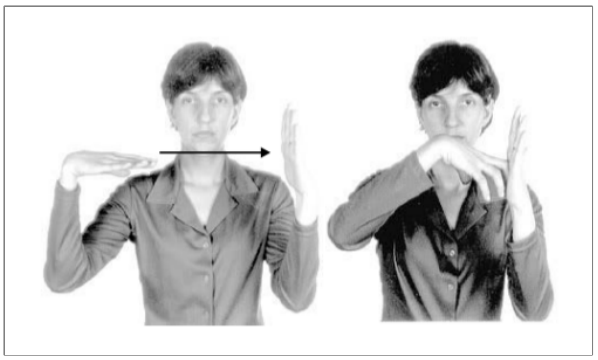


Fonte: Quadros e Karnopp (2004, p. 98)

é mantido; a regra da sequência única, na qual o movimento interno ou repetição dos sinais é eliminada para formar um composto; e a regra da antecipação da mão não-dominante, em que a mão passiva antecipa o segundo sinal no processo de composição.

Observe na Figura 5 o exemplo do sinal ACIDENTE, que é composto a partir dos sinais CARRO e BATER utilizando-se a regra da antecipação da mão não-dominante.

Figura 5 – Composição do sinal ACIDENTE a partir dos sinais CARRO e BATER



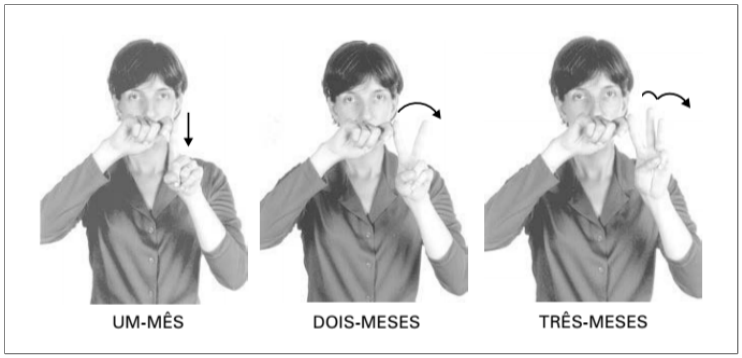
Fonte: Quadros e Karnopp (2004, p. 105)

c) Incorporação de numeral: combinação da configuração de mão de numeral a um sinal para especificar variação de quantidade em seu significado. Isso é útil, por exemplo, para representar número de anos, dias, horas, minutos, entre outros.

   A Figura 6 ilustra o uso desse mecanismo para especificar o número de meses no sinal.

d) Incorporação de negação: geração da contraparte negativa de um sinal através da

Figura 6 – Incorporação de numeral para especificar o número de meses no sinal

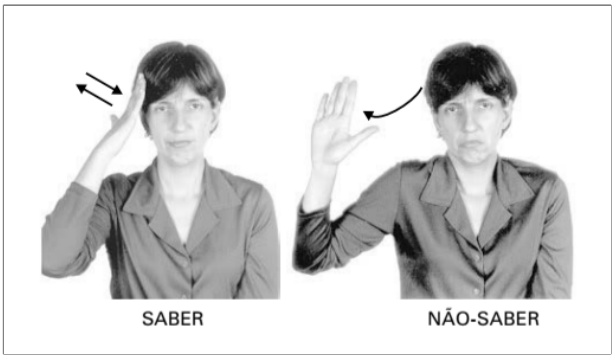


Fonte: Quadros e Karnopp (2004, p. 107)

alteração de um de seus parâmetros, que comumente é o seu movimento.

A Figura 7 ilustra a negação do sinal SABER adicionando-se um movimento e uma expressão não-manual específica de negação.

Figura 7 – Incorporação da negação ao sinal SABER



Fonte: Quadros e Karnopp (2004, p. 111)

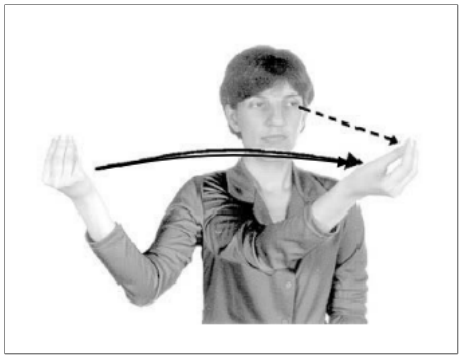
2. **Flexão**: consiste na adição de informação gramatical a palavras existentes para fazer com que elas se adequem melhor ao contexto em que são utilizadas. Nas línguas de sinais, alguns desses processos flexionais incluem:

    a) Pessoa: também conhecida como dêixis[1], consiste na modificação da referência de pessoa para os verbos. Na prática, isso é feito pelo interlocutor apontando-se para diferentes pontos no espaço à sua frente, os quais serão utilizados como referenciais que representam pessoas (ou objetos) envolvidas no discurso.

[1] Dêixis: palavra grega que significa apontar ou indicar, e representa uma forma de estabelecer referenciais no espaço que são utilizados para flexionar verbos com concordância. (QUADROS; KARNOPP, 2004)

A concordância do verbo se dará pela articulação do movimento partindo de um desses referenciais para o outro, conforme ilustrado na Figura 8.

Figura 8 – Flexão de pessoa para o verbo ENTREGAR, envolvendo dois referenciais



Fonte: Quadros e Karnopp (2004, p. 114)

b) Número: é utilizada para indicar: a forma plural do sinal, a qual é marcada pela sua repetição; ou a existência de múltiplos referentes no discurso, pela articulação da ação na direção dos respectivos referentes no espaço.

Veja na Figura 9 a flexão do verbo ENTREGAR para um, três e vários referentes.

Figura 9 – Flexão de número do verbo ENTREGAR para um (à esquerda), três (ao centro) e vários referentes (à direita)

Fonte: Quadros e Karnopp (2004, p. 120)

c) Grau: adiciona variação de grau ou intensidade ao sinal, a qual geralmente é transmitida utilizando-se expressões não-manuais. A Figura 10 ilustra a flexão do sinal LINDO para os graus de pouco (LINDINHO) e muito (LINDÍSSIMO).

d) Modo: especifica a maneira com que uma ação é realizada e também se utiliza de expressões não-manuais. Por exemplo, poderia-se detalhar que uma ação foi realizada "facilmente" ou "com dificuldade".

Figura 10 – Flexões de grau para o sinal LINDO

Fonte: Pereira et al. (2011, p. 65)

e) Aspecto temporal: determina a forma com que uma ação relaciona-se com o tempo, a qual pode ser uma das seguintes: incessante, ininterrupta, habitual (recorrente), contínua, ou duradoura (permanente).

   Observe na Figura 11 alguns exemplos de flexões do verbo CUIDAR para os aspectos temporais incessante, ininterrupto e habitual.

Figura 11 – Flexões temporal do verbo CUIDAR para os aspectos incessante (à esquerda), ininterrupto (ao centro) e habitual (à direita)

Fonte: Quadros e Karnopp (2004, p. 123)

f) Aspecto distributivo: determina a forma com que uma ação é distribuída entre diferentes pessoas ou objetos no discurso. Ela pode ser: exaustiva, quando a ação é repetida exaustivamente; específica, quando direciona-se a pessoas ou objetos específicos; não-específica, quando é generalizada ou indeterminada.

   Exemplos de flexão distributiva podem ser encontrados na Figura 9 para os aspectos específico (ao centro) e não-específico (à direita).

g) Reciprocidade: especifica que uma ação (ou relação) ocorre de forma mútua. Ela é representada pela duplicação do sinal, a qual é articulada simultaneamente. Observe

na Figura 12 um exemplo para o sinal OLHAR.

Figura 12 – Flexão de reciprocidade para o sinal OLHAR

Fonte: Quadros e Karnopp (2004, p. 122)

2.2.3 Sintaxe

Sintaxe é o estudo da construção de sentenças a partir das palavras, bem como dos princípios e regras envolvidos nesse processo, afirmam Hill, Lillo-Martin e Wood (2019), Jay (2011).

Nas línguas de sinais, a sintaxe é comunicada através da ordem das palavras e da utilização das expressões não-manuais. Observam-se a seguir alguns dos principais componentes presentes no processo de construção sintática dessas línguas (JAY, 2011; HILL; LILLO-MARTIN; WOOD, 2019; QUADROS; KARNOPP, 2004):

1. **Ordem das palavras**: existem diferentes possibilidades de ordenação das palavras nas sentenças da língua de sinais. Apesar disso, a ordem *Sujeito-Verbo-Objeto* (SVO) parece ser a mais básica dentre as demais. Ordenações como *Objeto-Sujeito-Verbo* (OSV), *Sujeito-Objeto-Verbo* (SOV) e *Verbo-Sujeito-Objeto* (VOS) são derivadas daquela primeira e resultam de operações sintáticas específicas associadas a alguma marcação como a concordância de verbos ou as expressões não-manuais.

2. **Tipos de sentenças**: diferentes tipos de sentenças são produzidos pela utilização de marcadores nas línguas de sinais, como as expressões não-manuais, e podem influenciar a forma com que as palavras são ordenadas nessas sentenças. Entre os principais tipos, podem-se enumerar:

a) Interrogativa: é geralmente marcada por expressões que combinam movimentos de levantar ou abaixar as sobrancelhas, a inclinação da cabeça, ou a sustentação do último sinal articulado. Observe na Figura 13 (à direita) a ênfase fornecida pela expressão facial ao sinal interrogativo "QUEM".

Figura 13 – Sentença interrogativa "JOÃO GOSTAR QUEM?"

Fonte: Quadros e Karnopp (2004, p. 187)

b) Declarativa: denota uma declaração afirmativa, negativa, ou neutra, a qual também podem ser marcadas através de expressões faciais. A Figura 15 (à direita) ilustra uma declaração afirmativa e a Figura 16 (à direita) uma negativa.

c) Condicional: construção condicional do tipo "se ... então", que é marcada pela elevação das sobrancelhas seguida de uma expressão afirmativa ou interrogativa
A Figura 14 ilustra um exemplo equivalente à sentença "se chover hoje, então o jogo será cancelado".

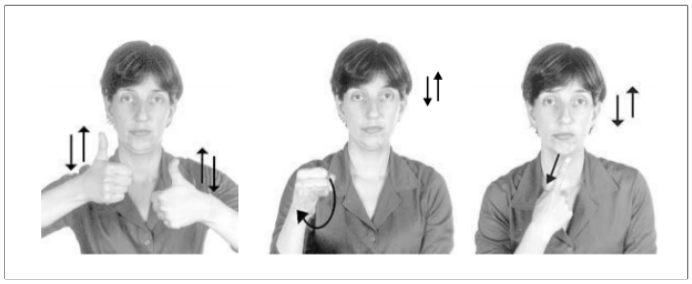
Figura 14 – Sentença condicional "CHOVER HOJE, JOGO CANCELAR"

Fonte: Jay (2011, p. 121)

d) Topicalizada: estabelece uma voz passiva, movimentando o objeto (ou tópico em questão) para o início da sentença e transformando sua ordem para *Objeto-Sujeito-Verbo* (OSV). Com isso, o objeto é geralmente demarcado por uma expressão não-manual diferente do restante da sentença.

Observe na Figura 15 o exemplo da sentença "FUTEBOL JOÃO GOSTAR", onde o termo futebol é o tópico central.

Figura 15 – Frase topicalizada "FUTEBOL JOÃO GOSTAR"



Fonte: Quadros e Karnopp (2004, p. 147)

3. **Negação**: existem diversas formas de construir negações de enunciados ou proposições nas línguas de sinais. Podem-se enumerar entre as principais: sinalizar NÃO antes ou após outro sinal; balançar a cabeça em negação enquanto sinaliza; utilizar uma orientação oposta para alguns sinais; ou franzir a testa enquanto sinaliza.

   Observe na Figura 16 o uso da expressão de balançar a cabeça em negação e a adição do sinal NÃO à parte final da sentença "FUTEBOL JOÃO GOSTAR".

Figura 16 – Negação da sentença "FUTEBOL JOÃO GOSTAR"

Fonte: Quadros e Karnopp (2004, p. 147)

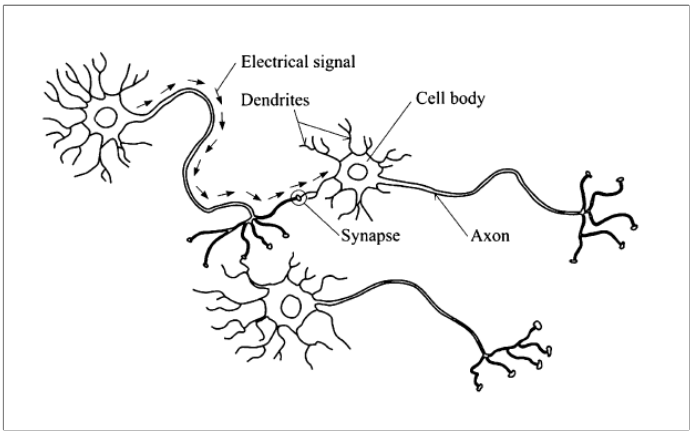
2.3 APRENDIZAGEM DE MÁQUINA

Quiza, López-Armas e Davim (2012), Russell e Norvig (2010) definem formalmente a Inteligência Artificial (IA) como um ramo da ciência da computação que visa estudar e projetar agentes inteligentes, ou seja, sistemas capazes de perceber seu ambiente e realizar ações que maximizam suas chances de sucesso. A Aprendizagem de Máquina (AM), por sua vez, é

definida por Murphy (2012), Goodfellow, Bengio e Courville (2016) como uma área que estuda um conjunto de algoritmos de IA capazes de detectar automaticamente padrões a partir de dados e, em seguida, utilizá-los para prever dados futuros ou realizar tomadas de decisões. Esses algoritmos comumente adotam modelos estatísticos para realizar análises e inferências sobre esses dados e não necessitam que instruções explícitas sejam programadas para que consigam detectar esses padrões.

Segundo Mitchell (1997), Bishop (2006), os algoritmos de AM constituem-se de um tipo de estrutura denominado Rede Neural Artificial (RNA). Essa estrutura é inspirada pela observação dos sistemas de aprendizado biológico, os quais são compostos de redes muito complexas de neurônios interconectados entre si, conforme ilustra a Figura 17.

Figura 17 – Exemplo de neurônios interconectados, que compõem os sistemas biológicos de aprendizado

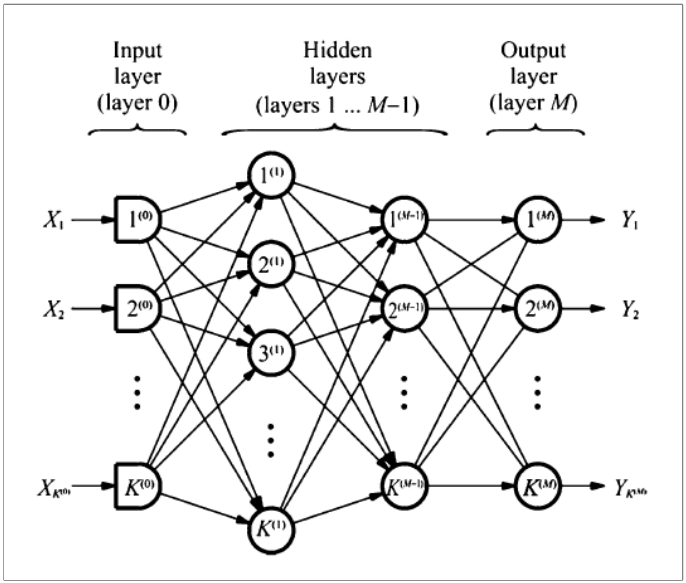


Fonte: Quiza, López-Armas e Davim (2012, p. 40)

De uma maneira análoga, a RNA é composta por um conjunto de unidades simples densamente interconectadas e organizadas em camadas, que exercem o papel dos neurônios, conforme ilustra a Figura 18. Cada unidade recebe como entrada um conjunto de números reais – geralmente produzidos por unidades da camada anterior – e produz como saída um outro conjunto de números reais – que poderão ser utilizados pelas unidades da camada seguinte. Além disso, cada conexão entre essas unidades recebe um peso, que é um número real que representa a importância daquela conexão. Esses pesos são ajustados à medida em que a RNA adapta-se aos padrões existentes nos dados, ou seja, conforme ela aprende acerca do domínio do problema em questão.

LeCun, Bengio e Hinton (2015), Russell e Norvig (2010) comentam que a forma mais utilizada de aprendizado pelas RNAs é o aprendizado supervisionado, no qual a partir da observação de um conjunto de amostras de pares de entrada-saída, a rede aprende uma função

Figura 18 – Exemplo de Rede Neural Artificial (RNA) com 4 camadas interconectadas: os dados $\{X_1, X_2, \ldots, X_k\}$ são recebidos pelas unidades da camada de entrada (ou *input layer*), processados pelas camadas ocultas (ou *hidden layers*) e pela camada de saída (ou *output layer*) que, por sua vez, produz as saídas $\{Y_1, Y_2, \ldots, Y_k\}$



Fonte: Quiza, López-Armas e Davim (2012, p. 43)

que mapeia daquele tipo de entrada para aquela saída respectiva, ajustando seus pesos de acordo.

Por exemplo, imagine a construção de uma RNA para identificar em imagens a existência de casas, carros, pessoas ou animais de estimação. O primeiro passo consistiria em coletar um grande conjunto de imagens de casas, carros, pessoas e animais de estimação, cada uma identificada quanto à existência de um desses elementos – formando assim os pares de entrada-saída mencionados acima que, nesse contexto, seriam pares de "imagem-elemento". Esse conjunto de imagens (ou amostras) é dividido em dois subconjuntos: um de treinamento – utilizado para que a RNA aprenda a identificar esses elementos –, e outro de testes – utilizado para avaliar o sucesso da rede nesse processo de aprendizado.

O próximo passo consistiria em realizar o treinamento da RNA, para o qual será aplicada a técnica de aprendizado supervisionado e utilizado o subconjunto de imagens de treinamento estabelecido acima. Nesse processo, a rede visitará cada uma dessas imagens e produzirá uma pontuação com a probabilidade de existência um dos elementos possíveis naquela imagem – casa, carro, pessoa ou animal de estimação. Objetiva-se que aquele elemento que está presente na imagem receba a maior pontuação e os demais elementos recebam pontuações menores, mas isso apenas ocorrerá à medida que os pesos da rede forem sendo ajustados para tal. Esse processo de ajuste de pesos é denominado otimização da RNA e, para que ele ocorra,

é necessário que seja introduzida uma função capaz de computar o erro ou a distância entre as pontuações informadas pela rede e aquelas que refletem o padrão correto desejado. A essa função atribui-se o nome de função objetivo – a qual também é conhecida como função de perda ou função de custo.

LeCun, Bengio e Hinton (2015), Goodfellow, Bengio e Courville (2016) afirmam que um dos algoritmos de otimização mais utilizados é o *Stochastic Gradient Descent* (ou Gradiente Estocástico Descendente) (SGD). De uma forma simplificada, ele consiste num processo de apresentar pequenos *batches* (ou lotes) de amostras para a RNA, computar as respectivas pontuações e erros (pela função objetivo), computar os gradientes para cada um dos pesos da rede (os quais indicam o quanto o erro aumentaria ou diminuiria se os pesos fossem ajustados em uma pequena quantidade) e ajustar os pesos na direção oposta à dos gradientes. Esse processo é repetido para vários *batches* de amostras do subconjunto de treinamento até que o erro médio pare de diminuir. Uma vez que o menor erro médio é encontrado, assume-se que o treinamento da RNA está finalizado.

Por fim, o desempenho da RNA será avaliado utilizando-se a mesma função objetivo para calcular o erro médio porém para um subconjunto diferente de amostras – o subconjunto de testes, estabelecido acima. Isso permitirá conhecer a capacidade de generalização da rede, ou seja, o quanto ela é bem-sucedida ao tentar produzir respostas sensatas para amostras nunca antes observadas.

Apesar disso, Goodfellow, Bengio e Courville (2016), Bishop (2006) argumentam que essa divisão das amostras em apenas dois subconjuntos fixos, de treinamento e de testes, pode ser problemático principalmente se isso resultar em um subconjunto de testes pequeno. Isso implicaria incerteza estatística em torno do erro médio de testes estimado, tornando difícil afirmar que um algoritmo funciona melhor que outro na tarefa em questão. Para lidar com esse problema, é possível que seja aplicado um procedimento conhecido como validação cruzada, o qual permite com que todas as amostras sejam utilizadas nesse processo de testes ou estimativa de erro médio.

Os algoritmos de validação cruzada baseiam-se na ideia de repetir as etapas de treinamento e testes utilizando subconjuntos diferentes de amostras, escolhidos aleatoriamente a partir do conjunto original, e computando o erro a partir da média dos erros obtidos em cada repetição. O mais comum desses algoritmos é o k-*fold*, que funciona dividindo o conjunto de dados original em k subconjuntos não sobrepostos e realizando k repetições de treinamento e testes. A cada repetição i, o i-ésimo subconjunto é utilizado como subconjunto de testes e o restante

dos dados é utilizado como o de treinamento, e assim por diante. O erro médio é então estimado tomando-se a média dos erros obtidos nessas k repetições.

Segundo Goodfellow, Bengio e Courville (2016), a maioria das RNAs também apresenta parâmetros que controlam diferentes aspectos do seu comportamento, os quais são denominados hiperparâmetros. Exemplos deles incluem o número de camadas da rede, as dimensões dessas camadas, a taxa de aprendizagem e o tamanho dos *batches* adotados no treinamento, entre outros que variam de acordo com o tipo de RNA. Contudo, diferentemente do que ocorre para os pesos da rede, esses hiperparâmetros não são aprendidos automaticamente durante o processo de treinamento. Ao contrário, a escolha dos valores de hiperparâmetros que melhor otimizam a capacidade da rede é geralmente uma tarefa complexa porque demanda uma compreensão mais profunda acerca do papel que eles exercem sobre o desempenho da rede.

Por conta disso, é comum que sejam adotadas abordagens automáticas para essa finalidade, a exemplo do algoritmo *grid search* (ou busca de grade) – que será utilizado mais adiante nos experimentos deste trabalho. O *grid search* funciona com base num conjunto pequeno e finito de valores a serem explorados para cada um dos hiperparâmetros da rede, os quais são informados pelo usuário ao início da busca. Com base nisso, o algoritmo gera o produto cartesiano de todas as combinações possíveis desses valores de hiperparâmetros e, em seguida, treina uma rede neural para cada uma dessas combinações. Ao final desse processo, a rede neural que apresentar o menor erro médio é então considerada aquela que encontrou a melhor combinação de valores de hiperparâmetros. Obviamente, um problema relevante desse algoritmo é o seu custo computacional, que cresce exponencialmente em função do número de hiperparâmetros e de valores a serem explorados.

Para essa seleção automática de hiperparâmetros é introduzido um novo subconjunto de amostras denominado subconjunto de validação. Ele é importante para evitar que nesse processo sejam utilizadas as mesmas amostras para treinar e validar o desempenho das redes para cada combinação de hiperparâmetros. Sendo assim, esse subconjunto é criado a partir de uma divisão do subconjunto de treinamento na qual, tipicamente, 20% das amostras são direcionadas para validação de hiperparâmetros e 80% permanecem sendo utilizadas para treinamento. O subconjunto de testes, por sua vez, permanece inalterado e não é envolvido na seleção de hiperparâmetros.

2.3.1 Aprendizagem profunda e modelos sequenciais

Segundo Goodfellow, Bengio e Courville (2016), LeCun, Bengio e Hinton (2015), a Aprendizagem Profunda (AP) consiste num tipo de Aprendizagem de Máquina baseado em RNAs que adotam um número muito grande de camadas de processamento para extrair progressivamente representações de níveis mais elevados a partir dos dados. A esse tipo particular de RNA atribui-se o nome de Rede Neural Profunda (RNP).

Por conta de sua estrutura robusta, essas redes têm sido capazes de lidar com problemas muito complexos ao longo das últimas décadas e produzir progressos extremamente promissores para problemas que resistiram por muito tempo às melhores tentativas de avanço pela comunidade de IA, principalmente aqueles envolvendo linguagem – e que são abordados pela área de Processamento de Linguagem Natural (PLN).

Jurafsky e Martin (2022), por sua vez, definem a linguagem como sendo um fenômeno inerentemente temporal. Os autores afirmam que ela pode ser compreendida como uma sequência de eventos que desdobram-se ao longo do tempo como um fluxo contínuo de dados. Dessa forma, para que fosse possível abordar esse aspecto temporal e lidar com dados organizados de maneira sequencial, foram estabelecidas arquiteturas específicas de RNPs que atualmente são conhecidas como redes neurais sequenciais ou modelos sequenciais. Dentre as mais populares dessas arquiteturas estão a *Recurrent Neural Network* (ou Rede Neural Recorrente) (RNN) (e suas extensões, como o *Long Short-Term Memory* (LSTM) e a *Gated Recurrent Unit* (GRU)) e o *Transformer*.

As RNNs baseiam-se no trabalho de Rumelhart, Hinton e Williams (1986) e consistem de redes neurais que contêm ciclos (ou recorrências) em suas conexões, os quais fazem com que o valor de suas unidades sejam direta ou indiretamente dependentes de suas próprias saídas anteriores. De um modo geral, elas funcionam processando cada palavra da sequência e combinando ela com o contexto ou estado oculto anterior para tentar prever a próxima palavra da sequência. Esse contexto, por sua vez, é capaz de representar as informações de todas as palavras anteriores daquela sequência.

Contudo, LeCun, Bengio e Hinton (2015), Goodfellow, Bengio e Courville (2016), Graves (2012) ressaltam que essas redes apresentaram limitações em armazenar informações por um período muito longo de tempo, dentre as quais estão os problemas conhecidos de *gradient vanishing* (ou desaparecimento do gradiente) e *gradient exploding* (ou explosão do gradiente). Isso demandou com que extensões dessa arquitetura fossem desenvolvidas no decorrer dos anos

com o intuito abordar melhor essas questões.

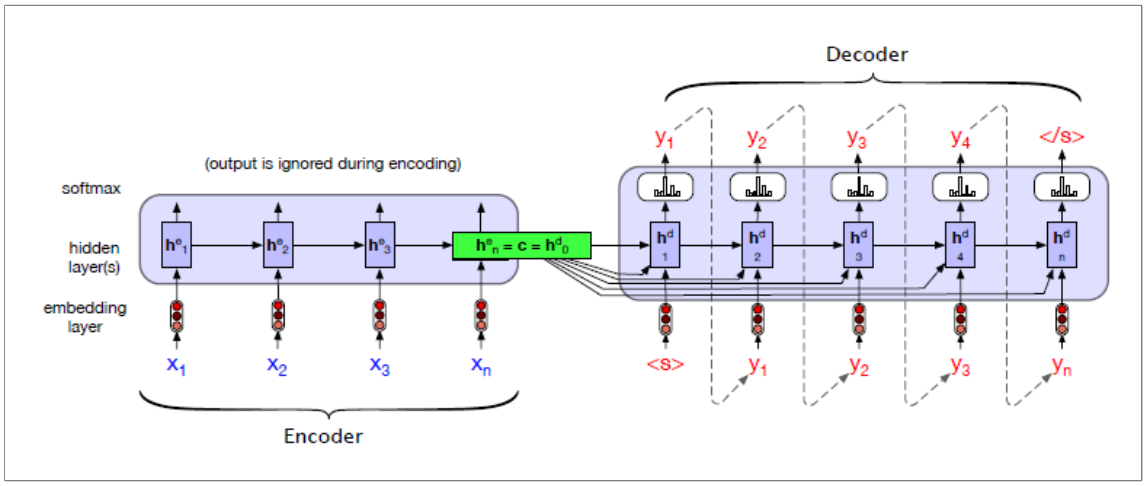
Uma das principais extensões é a LSTM, que foi introduzida por Hochreiter e Schmidhuber (1997). De acordo com Graves (2012), Jurafsky e Martin (2022), a principal inovação dessa rede é a capacidade de aprender a gerenciar o contexto de forma automática, decidindo quando informações são necessárias e quando podem ser removidas, sem necessitar que uma estratégia explícita seja codificada para isso. Ele utiliza uma camada específica para representar esse contexto e um conjunto de portas, as quais controlam o fluxo de informações para dentro e para fora de suas células. Segundo Goodfellow, Bengio e Courville (2016), LeCun, Bengio e Hinton (2015), essas redes são extremamente bem-sucedidas em diferentes tipos de aplicações, como reconhecimento e geração de caligrafia, reconhecimento de fala, Tradução Automática (TA), legendagem de imagens e análise sintática.

A GRU é também uma extensão muito popular das RNNs e foi criada por Cho et al. (2014a) com o intuito de simplificar o desenho das unidades internas do LSTM. De acordo com Goodfellow, Bengio e Courville (2016), Ravanelli et al. (2018), elas diferenciam-se apenas pela forma como controlam o fluxo de informações entre suas camadas: enquanto o LSTM utiliza três portas em suas células internas (*update gate*, *forget gate* e *output gate*), o GRU propõe a adoção de apenas duas portas para isso (*update gate* e *reset gate*).

Um outro tipo de arquitetura bastante utilizada no PLN é a *Sequence-to-Sequence* (ou Sequência para Sequência), também conhecida como *Encoder-Decoder* (ou Codificador Decodificador). Ela foi apresentada por Cho et al. (2014b), Sutskever, Vinyals e Le (2014) e sua estrutura é composta por duas redes neurais: uma codificadora, que recebe uma sequência de entrada e gera uma representação contextualizada dela – que seria o contexto; e uma decodificadora, que produz uma sequência de saída específica para a tarefa em questão, conforme ilustrado na Figura 19. Esses redes codificadoras e decodificadoras são geralmente implementados utilizando-se RNNs, como o LSTM e o GRU. Além disso, algumas otimizações dessa arquitetura consideram a adição de uma camada de *attention* (ou atenção) antes do decodificador com o objetivo de eliminar um gargalo observado ali por Bahdanau, Cho e Bengio (2015).

O *Transformer*, por sua vez, consiste num tipo de arquitetura que não é recorrente e, ao invés disso, baseia-se num mecanismo de *attention* para estabelecer dependências globais entre os dados de entrada e saída. Ele foi introduzido por Vaswani et al. (2017) e baseia-se na estrutura do *Encoder-Decoder*, porém seus codificadores e decodificadores são compostos por blocos empilhados de redes multicamadas que combinam camadas lineares simples, redes

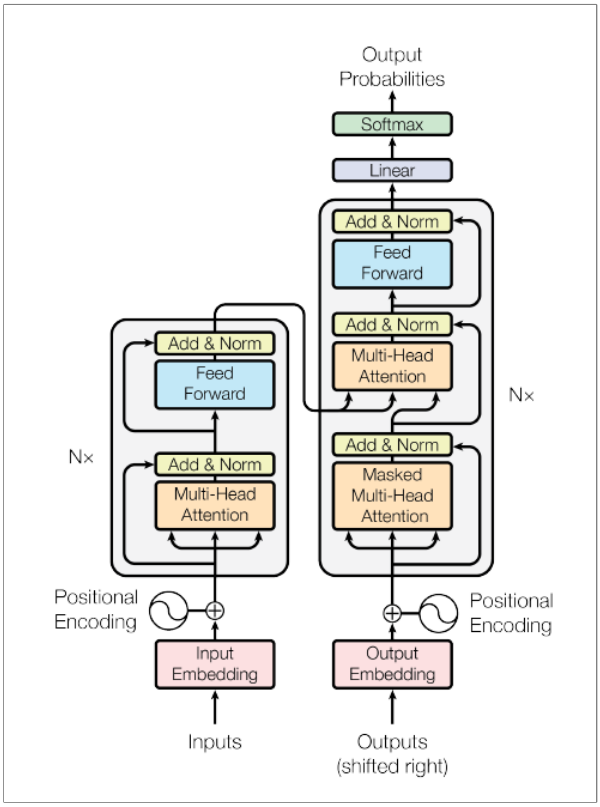
Figura 19 – Arquitetura do *Encoder-Decoder*: a sequência de entrada $\{x_1, x_2, x_3, x_n\}$ é recebida pelo *encoder* (à esquerda) e utilizada para gerar o contexto c (em verde), o qual é utilizado pelo *decoder* (à direita) para produzir a sequência $\{y_1, y_2, y_3, y_n\}$



Fonte: Jurafsky e Martin (2022, p. 220)

feed-forward e camadas de *self-attention* (ou auto-atenção) – as quais são a principal inovação aqui –, conforme ilustra a Figura 20.

Figura 20 – Arquitetura do *Transformer*: são utilizados blocos empilhados que combinam redes *feed-forward* e camadas de *self-attention* para o *encoder* (à esquerda) e o *decoder* (à direita); os *embeddings* (abaixo) recebem uma codificação posicional para que seja considerada a ordem de suas sequências



Fonte: Vaswani et al. (2017, p. 3)

Segundo Wolf et al. (2020), o *Transformer* é escalável e capaz de capturar o contexto de sequências muito longas, e isso possibilitou a construção e aplicação de modelos de maior capacidade para uma grande variedade de tarefas. Devido a disso, ele tornou-se rapidamente a arquitetura dominante no PLN, superando o desempenho de redes alternativas como as RNNs e a *Convolutional Neural Network* (ou Rede Neural Convolucional) (CNN) para tarefas de compreensão e geração de linguagem natural.

Apesar disso, mesmo com as inovações introduzidas pelo *Transformer*, é possível observar pela Figura 20 e Figura 19 que sua estrutura ainda preserva muitas semelhanças com o *Encoder-Decoder*, da qual origina-se. Em ambas estruturas, percebem-se dois grandes blocos interconectados, que interagem para codificar uma representação de contexto a partir de uma sequência de dados de entrada, e para decodificá-lo gerando uma nova sequência de saída. Essas sequências, por sua vez, podem possuir tamanhos distintos e permitem com que sejam modelados diferentes tipos de representações de linguagem como, por exemplo, uma sequência de texto que é traduzida de um idioma para outro, uma representação de fala que é transformada em texto (ou vice-versa), ou ainda uma resposta que é gerada para uma pergunta fornecida como entrada.

Devido a essa afinidade para lidar com tarefas de linguagem, ambas arquiteturas *Transformer* e *Encoder-Decoder* serão consideradas para avaliação da abordagem introduzida mais adiante neste trabalho.

2.4 RECONHECIMENTO DE LÍNGUAS DE SINAIS

De acordo com Wadhawan e Kumar (2019), Cooper, Holt e Bowden (2011), o Reconhecimento de Língua de Sinais (RLS) é uma área de pesquisa colaborativa e multidisciplinar que tem por objetivo elaborar métodos e algoritmos para identificar os sinais articulados pelos usuários dessa língua e compreender seu significado.

É uma área capaz de contribuir com a quebra de barreiras existentes para os usuários dessa língua e facilitar a comunicação cotidiana entre Surdos e ouvintes, segundo Rastgoo, Kiani e Escalera (2021), Papastratis et al. (2021). Isso é importante porque, além de promover a inclusão dos Surdos em sociedade, aborda o problema atual de que as tecnologias de comunicação são, em sua maioria, desenvolvidas para suportar línguas faladas ou escritas, mas excluem as línguas sinalizadas. Por exemplo, o WhatsApp, Telegram e iMessage tornaram-se ferramentas imprescindíveis em nossas vidas, porém, a comunidade Surda enfrenta diversos desafios para

utilizá-las.

Ainda segundo os autores, apesar dessas necessidades terem sido identificadas há muito tempo pela comunidade acadêmica, apenas recentemente a área de RLS passou a receber mais atenção. Isso deve-se principalmente aos avanços ocorridos nas tecnologias de sensoriamento e dos algoritmos de IA, que abriram caminho para o desenvolvimento de aplicações capazes de abordar tais demanda de maneira mais efetiva. Além disso, o advento das arquiteturas de AP proporcionou uma melhora significativa no desempenho dos algoritmos utilizados nesta área.

Koller (2020) realizou uma análise baseada nos estudos mais relevantes em RLS publicados desde 1983, a qual possibilita delinear melhor essa evolução recente e o estado da arte atual. Além disso, as revisões apresentadas por Rastgoo, Kiani e Escalera (2021), Papastratis et al. (2021), Wadhawan e Kumar (2019) também contribuem para estender essa análise. Esse panorama será discutido na seção a seguir.

2.4.1 Breve panorama

Nesta seção será discutido brevemente o panorama observado para a RLS ao longo das últimas décadas, tomando como base as revisões literárias apresentadas por Koller (2020), Rastgoo, Kiani e Escalera (2021), Papastratis et al. (2021), Wadhawan e Kumar (2019).

Primeiramente, observa-se na Tabela 1 uma perspectiva geral do número de estudos publicados nessa área até 2020, conforme identificado por Koller (2020). Percebe-se que esse número praticamente dobrou a cada intervalo de 5 anos e, com o advento de novos dispositivos e algoritmos de IA por volta dos anos 2010, cresceu de forma ainda mais expressiva. Isso evidencia a ênfase maior recebida pela área recentemente.

Apesar disso, a tabela também nos revela que a maioria desses estudos aborda o RLS utilizando vocabulários muito pequenos, com menos de 200 sinais. Isso corresponde a conjuntos limitados que são geralmente selecionados para simplificar as pesquisas mas que, na prática, acabam limitando também a representatividade da língua de sinais perante seu contexto real de aplicação. Apenas a partir de 2015 percebe-se um crescimento mais significativo na adoção de vocabulários com mais de 1.000 sinais.

A Tabela 2, por sua vez, divide os estudos acima entre aqueles que abordam a língua utilizando sinais isolados – os quais são reconhecidos separadamente do contexto do discurso –, e os que utilizam sinais contínuos – onde sequências completas são utilizadas nesse processo. Percebe-se que a maior parte desses estudos seguem pela linha dos sinais isolados, uma vez

Tabela 1 – Estudos em RLS publicados até 2020, agrupados por intervalos de cinco anos (na horizontal) e tamanho de vocabulário modelado (na vertical)

Vocabulário	Ano							Total
	<1990	1990 - 1995	1995 - 2000	2000 - 2005	2005 - 2010	2010 - 2015	2015 - 2020	
0 - 50	2	5	12	12	40	35	50	156
50 - 200	0	1	6	10	27	22	51	117
200 - 500	0	1	3	2	7	15	25	53
500 - 1000	0	0	0	0	1	13	13	27
>1000	0	0	1	6	3	4	40	54
Total	2	7	22	30	78	89	179	

Fonte: Koller (2020, p. 3)

que a abordagem de sinais contínuos é certamente mais complexa e apenas teve seus primeiros *datasets* disponibilizados por volta dos anos 2015. Os dados também nos mostram que, ao passo em que estudos com sinais isolados comumente modelam vocabulários pequenos, aqueles que optam por sinais contínuos passaram a preferir vocabulários acima de 1.000 sinais.

Tabela 2 – Estudos em RLS publicados até 2020, segmentados entre aqueles que abordaram sinais isolados (na primeira tabela) e sinais contínuos (na segunda tabela); os números são agrupados por intervalos de cinco anos (na horizontal) e tamanho de vocabulário (na vertical)

Sinais isolados

Vocabulário	Ano							Total
	<1990	1990 - 1995	1995 - 2000	2000 - 2005	2005 - 2010	2010 - 2015	2015 - 2020	
0 - 50	2	4	7	7	27	29	43	119
50 - 200	0	1	2	8	12	17	34	74
200 - 500	0	1	2	1	3	6	19	32
500 - 1000	0	0	0	0	1	11	12	24
>1000	0	0	1	3	2	2	6	14
Total	2	6	12	19	45	65	114	

Sinais contínuos

Vocabulário	<1990	1990 - 1995	1995 - 2000	2000 - 2005	2005 - 2010	2010 - 2015	2015 - 2020	Total
0 - 50	0	1	5	5	13	6	7	37
50 - 200	0	0	4	2	15	5	17	43
200 - 500	0	0	1	1	4	9	6	21
500 - 1000	0	0	0	0	0	2	1	3
>1000	0	0	0	3	1	2	34	40
Total	0	1	10	11	33	24	65	

Fonte: Koller (2020, p. 3)

Ao analisar as línguas de sinais que foram abordadas pelas pesquisas em RLS, conforme Tabela 3, percebe-se que a ASL foi predominante dentre as demais. Isso provavelmente deve-se ao pioneirismo recebido por ela nos estudos de Stokoe (1960) que, além de produzir maior clareza acerca de sua estrutura, viabilizou o desenvolvimento de recursos importantes como *datasets*, que suportaram tais pesquisas. As línguas *Chinese Sign Language* (Língua de Sinais Chinesa) (CSL) e *Deutsche Gebärdensprache* (Língua de Sinais Alemã) (DGS) aparecem em

seguida na tabela com uma participação também relevante, a qual desde 2010 apresenta números muito próximos àqueles da ASL. A Língua Brasileira de Sinais (Libras), por sua vez, aparece de forma mais modesta nesse levantamento realizado por Koller (2020).

Tabela 3 – Línguas de sinais abordadas pelos estudos em RLS

Língua de sinais	Ano							Total
	<1990	1990 - 1995	1995 - 2000	2000 - 2005	2005 - 2010	2010 - 2015	2015 - 2020	
ASL	1	5	5	11	30	21	46	119
CSL	0	0	2	10	4	21	35	72
DGS	0	0	1	2	4	17	40	64
BSL	2	2	13	3	0	0	0	20
ArSL	0	0	0	0	6	7	4	17
JSL	1	1	4	1	2	1	1	11
ISL	0	0	0	0	0	2	9	11
GSL	0	0	0	0	1	10	2	13
TID	0	0	0	0	5	0	4	9
NGT	0	0	2	0	2	0	5	9
VGT	0	0	0	0	0	0	7	7
LIS	0	0	0	0	2	2	1	5
AUSLAN	0	1	3	0	0	0	1	5
LSA	0	0	0	0	0	0	5	5
TSL	0	0	3	1	0	0	0	4
PJM	0	0	0	1	1	2	0	4
Libras	0	0	0	0	1	2	1	4
...				...				...
Outras	0	0	2	0	7	1	15	15
Total	4	9	35	29	65	86	176	

Fonte: Koller (2020, p. 9)

Na Tabela 4 observam-se os tipos de dados de entrada utilizados por tais estudos. Imagens ou *frames* de vídeos em RGB aparecem em destaque e vêm sendo adotados desde os anos iniciais até a atualidade. Ao longo da história da RLS eles dividiram espaço com outros tipos de dados mas, a partir de 2005, tornaram-se predominantes provavelmente pela maior adoção de técnicas baseadas em Visão Computacional desde então. Luvas eletrônicas estiveram presentes principalmente nos anos iniciais da RLS, quando técnicas envolvendo dispositivos conectados ao corpo dos indivíduos foram muito utilizadas. Apesar disso, elas foram gradativamente cedendo espaço a outros tipos como as luvas coloridas e ao *mocap*[2] até por volta de 2005.

O surgimento do Kinect em 2010 representou uma revolução para a área de RLS, devido à capacidade que ele introduziu de rastrear os corpos dos indivíduos e de fornecer dados como

[2] *Mocap* (*motion capture* ou captura de movimento): é uma técnica de captura de movimentos que utiliza equipamentos específicos como marcadores ou trajes especiais afixados ao corpo dos indivíduos ou objetos (KITAGAWA; WINDSOR, 2017).

coordenadas e imagens profundidade além daquelas RGB, o qual não existia na época. Nos anos posteriores ao seu lançamento, vários outros dispositivos também foram introduzidos com o mesmo propósito. Na Tabela 4, é possível perceber claramente uma mudança para a adoção de dados RGB combinados aos dados de profundidade, decorrente disso.

Tabela 4 – Proporção de tipos de dados de entrada utilizados pelos estudos em RLS

Dados de entrada	Ano						
	<1990	1990 - 1995	1995 - 2000	2000 - 2005	2005 - 2010	2010 - 2015	2015 - 2020
RGB	50%	29%	36%	33%	72%	85%	87%
Profundidade	0%	0%	0%	0%	1%	38%	22%
Luva colorida	0%	0%	18%	10%	18%	4%	3%
Luva eletrônica	50%	71%	41%	50%	10%	7%	4%
Mocap	0%	29%	23%	57%	6%	7%	6%

Fonte: Koller (2020, p. 4)

Por fim, a Tabela 5 apresenta os tipos de *features* utilizadas pelos estudos acima. Elas estão categorizados entre *features* manuais, que correspondem aos parâmetros de configuração de mão, orientação, movimento e locação introduzidos na subseção 2.2.1; não-manuais, que referem-se às expressões não-manuais introduzidas na mesma seção; e globais, que são aquelas que capturam uma visão completa do corpo dos indivíduos, como coordenadas, imagens RGB, de profundidade, ou de fluxo óptico.

Percebe-se uma predominância da adoção de *features* manuais desde os anos iniciais até por volta de 2015. Isso explica-se principalmente pelo fato de que um grande número dessas pesquisas aborda os sinais através de recortes das mãos dos indivíduos, muitas vezes estáticas e fora do seu contexto original. A esse tipo de abordagem Koller (2020) enquadrou em seu levantamento como sendo refente à configuração de mão e, consequentemente, uma *feature* manual. *Features* relacionadas à locação e ao movimento das mãos também são encontradas modeladas de diferentes maneiras e contribuem para essa predominância.

Com o uso do Kinect e das técnicas de Aprendizagem Profunda a partir de 2010, as *features* globais passaram a assumir uma posição bastante expressiva nesses estudos. Isso porque, ao mesmo tempo em que este dispositivo passou a fornecer novos tipos de informações, como coordenadas corporais e dados de profundidade combinados com RGB, essas técnicas introduziram um poder de processamento que possibilitou com que essas informações fossem utilizadas diretamente como *features* de entrada para eles. Por outro lado, as *features* não-manuais não chegaram a apresentar uma adoção significativa perante os demais tipos, ao analisá-los de um modo geral.

Tabela 5 – Tipos de *features* utilizadas pelos estudos em RLS

Tipos de features	Ano						
	<1990	1990 - 1995	1995 - 2000	2000 - 2005	2005 - 2010	2010 - 2015	2015 - 2020
Manuais	100%	100%	100%	100%	99%	100%	47%
Não-manuais	0%	0%	5%	0%	13%	17%	8%
Globais	0%	0%	0%	0%	1%	37%	66%

Fonte: Koller (2020, p. 7)

2.4.2 Desafios da área

Nesta seção, serão discutidos os principais desafios identificados para o Reconhecimento de Língua de Sinais (RLS) segundo perspectivas apresentadas por Bragg et al. (2019), Cooper, Holt e Bowden (2011). Apesar de serem constatados vários avanços e um aumento no interesse por essa área ao longo dos últimos anos, muitos desses desafios ainda representam barreiras importantes que limitam a produção de progressos mais expressivos.

Dessa forma, essa discussão ajudará a conscientizar acerca deles e compreender a jornada ainda a ser percorrida pela área, bem como a direção que precisa ser adotar para isso. Sendo assim, esses desafios podem ser enumerados como:

1. **Abordagem inadequada**: de acordo com Bragg et al. (2019), muitas pesquisas em RLS ocorrem em silos disciplinares e, por conta disso, acabam por não abordar esse problema de forma abrangente. Por exemplo, publicações que limitam-se a rotular imagens de mãos estáticas em determinadas configurações deixam de considerar o contexto em volta delas ou a forma com que aquilo se conecta à linguística. Em outros casos, Cooper, Holt e Bowden (2011) afirmam que o RLS é tratado como uma tarefa simples de reconhecimento de gestos, na qual os sinais são considerados apenas um conjunto de possibilidades bem definidas a serem rotuladas.

   Um outro fator apontado por Bragg et al. (2019) é de que equipes desenvolvendo algoritmos para as línguas sinais comumente carecem de maior propriedade acerca delas, seja pelo envolvimento de membros Surdos, pela exposição a essa cultura, pelo aprofundamento das complexidades linguísticas que esses algoritmos deveriam considerar, ou pela empatia com os problemas reais que eles deveriam resolver.

   Muito frequentemente também, utilizam-se *datasets* que não refletem contextos reais de uso para treinar tais algoritmos, como no caso daqueles que envolvem indivíduos

não-nativos na língua ou que são coletados da internet e cuja procedência não pode ser confirmada.

2. **Natureza visual**: as línguas de sinais apresentam o desafio de serem multicanais, ou sejam, elas transmitem significado através de vários canais visuais simultâneos que precisam ser considerados – como os movimentos das mãos, do corpo, expressões da face, entre outros.

   Dessa forma, Bragg et al. (2019) comentam que não é suficiente simplesmente replicar para o RLS as técnicas utilizadas com sucesso por outras áreas, como o reconhecimento de línguas faladas ou escritas, uma vez que elas assumem a existência de um canal único. Além disso, técnicas assim comumente adotam algum tipo de anotação no formato de texto como entrada, no entanto, as línguas de sinais não possuem uma forma escrita ou uma anotação padrão.

   Sendo assim, existe uma necessidade de que técnicas específicas sejam estabelecidas com o intuito de acomodar adequadamente as particularidades dos sinais.

3. **Linguística particular**: conforme introduzimos na seção 2.2, as línguas de sinais possuem um sistema linguístico próprio e complexo desenvolvido em torno do espaço visual, o qual as diferencia das línguas faladas. Abordar essa linguística é um fator fundamental para tratá-las efetivamente como línguas e não apenas como sistemas de gestos, o que certamente conduzirá a avanços mais concretos no RLS. Contudo, quando são analisados os estudos nessa área percebe-se que são poucos aqueles que utilizam esse tipo de abordagem.

   Cooper, Holt e Bowden (2011) também ressaltam a pouca atenção que é recebida pelos parâmetros não-manuais nessas pesquisas, que comumente se concentram apenas em recortes das mãos dos indivíduos – conforme discutimos na subseção 2.4.1. Apesar disso, a seção 2.2 nos mostra que muito da carga linguística dos sinais é transmitida através desses parâmetros e isso os torna elementos essenciais da língua que precisam ser endereçados.

4. ***Datasets* limitados**: segundo Cooper, Holt e Bowden (2011), Bragg et al. (2019), os *datasets* disponíveis publicamente para as línguas de sinais costumam ser limitados em quantidade e qualidade. Como consequência, isso acaba também limitando a variedade e o desempenho dos algoritmos que poderiam ser aplicados ao RLS, uma vez que técnicas

recentes como a Aprendizagem Profunda demandam uma grande quantidade de dados para funcionar adequadamente. Em áreas como o reconhecimento da fala, por exemplo, o estado da arte foi alcançado graças à utilização de corpus com milhões de amostras de palavras; no caso das línguas de sinais, no entanto, os *datasets* não costumam ultrapassar 10.000 amostras.

Bragg et al. (2019) também destacam que há uma representatividade limitada de indivíduos e situações nesses *datasets*, que ocorre porque eles geralmente são criados em ambientes controlados envolvendo consultores ou intérpretes da língua. Representatividade nesse contexto refere-se à variedade de características como idade, geografia, tom de pele, deficiência, proficiência, condições do ambiente, qualidade da câmera, ângulos de captura, entre outras.

Além disso, os autores ressaltam a importância de se desenvolverem mais *datasets* com sinais contínuos, envolvendo sequências completas de diálogos e enunciados mais longos, uma vez que isso reflete melhor contextos reais de utilização da língua.

3 MATERIAIS E MÉTODOS

Neste capítulo será discutida em mais detalhes a abordagem proposta, as justificativas para sua adoção, bem como as técnicas que foram aplicadas e a preparação dos experimentos realizados.

Esta pesquisa propõe-se a realizar o reconhecimento da linguagem de sinais a partir de uma perspectiva estritamente linguística, baseada nos constituintes fonológicos mínimos que descrevem os sinais, e centrada no aprendizado das complexidades e regras que convém contexto e dá significado a eles.

Isso assemelha-se à forma como hoje outras línguas são abordadas com sucesso pelo Processamento de Linguagem Natural (PLN) e diferencia-se daquela predominante no Reconhecimento de Língua de Sinais (RLS) que, por sua vez, trata os sinais como gestos não-estruturados, mapeados a partir de dados brutos capturados dos indivíduos – como pixels de imagens ou *frames* de vídeos, pontos lidos de luvas eletrônicas, coordenadas 2D ou 3D, entre outros – e colocam em segundo plano a importância linguística do sinal.

A hipótese deste trabalho assume que, além de deixar de abordar uma parte muito importante dessa língua, lidar com esse tipo de dados brutos traz complexidades adicionais que extrapolam o escopo que deveria ser efetivamente abordado pelo RLS. Em outras palavras, esse foco inadequado faz com que pesquisas em RLS deixem de solucionar um problema intrinsecamente de PLN e passem a investir esforços consideráveis tentando lidar com um conjunto de desafios pertinentes à área de Visão Computacional (VC) – os quais comumente já estão abordados ou solucionados por uma de suas subáreas, como a detecção, segmentação e rastreamento de partes do corpo em imagens e vídeos; a interação entre mãos e oclusões decorrentes disso; as variações de tom de pele, cores de roupa e luminosidade do ambiente; entre outros listados nas revisões literárias elaboradas no decorrer da última década para a RLS (PAPASTRATIS et al., 2021; RASTGOO; KIANI; ESCALERA, 2021; KOLLER, 2020; BRAGG et al., 2019; WADHAWAN; KUMAR, 2019; SUHARJITO et al., 2018; JOKSIMOSKI et al., 2022; COOPER; HOLT; BOWDEN, 2011).

Alguns exemplos populares desses problemas sendo consistentemente endereçados dentro da VC incluem as ferramentas OpenPose (WEI et al., 2016; CAO et al., 2017; SIMON et al., 2017) e MediaPipe (LUGARESI et al., 2019; BAZAREVSKY et al., 2019; VAKUNOV et al., 2020; BAZAREVSKY et al., 2020), desenvolvidas pela Carnegie Mellon University e Google Research,

respectivamente. Ambas são o resultado de anos de pesquisa em torno de tais questões, as quais alcançaram um nível de maturidade elevado capaz de abordar em tempo real tarefas de estimativa de pose e rastreamento do corpo, mãos e face (inclusive envolvendo múltiplos indivíduos) de forma robusta a variações corporais, de luminosidade e de ambientes, utilizando apenas uma câmera comum RGB. Elas estão disponíveis abertamente[1] e a reutilização desse conhecimento nas etapas para capturar e gerar *features* de níveis mais elevados para o RLS certamente contribuirá para progressos mais efetivos.

Fazendo uma analogia com outras tarefas de PLN, abordar a língua de sinais por meio dos dados brutos como discutido acima e lidar com os desafios apresentados, por exemplo, possui uma complexidade equivalente a tentar interpretar textos manuscritos apenas rastreando-se o movimento da mão do autor enquanto ele desenha as letras no papel – ao invés de simplesmente escanear o texto final escrito como entrada para isso; ou ainda, tentar reconhecer a fala de um indivíduo apenas realizando a detecção e o rastreamento dos movimentos de seus lábios – ao invés de considerar os sinais de áudio capturados para tal.

Como resultado desse enquadramento inadequado por parte das pesquisas em RLS, no decorrer das últimas décadas constata-se um progresso pouco expressivo dessa área sobretudo nos aspectos da linguagem e aplicabilidade no mundo real, acerca do qual Selvaraj et al. (2022), Yin et al. (2021), Cooper, Holt e Bowden (2011) reiteram:

> Quando comparado com a pesquisa de Processamento de Linguagem Natural baseada em texto e fala, o progresso das pesquisas para línguas de sinais está significativamente atrasado. (SELVARAJ et al., 2022; YIN et al., 2021, tradução nossa)

> Enquanto sistemas de reconhecimento da fala avançaram ao ponto de estarem comercialmente disponíveis, o reconhecimento de sinais ainda está em sua infância. Atualmente, todos os serviços comerciais de tradução de sinais são baseados em humanos e requerem que pessoal especializado esteja disponível, o que os tornam caros e pouco acessíveis. (COOPER; HOLT; BOWDEN, 2011, tradução nossa)

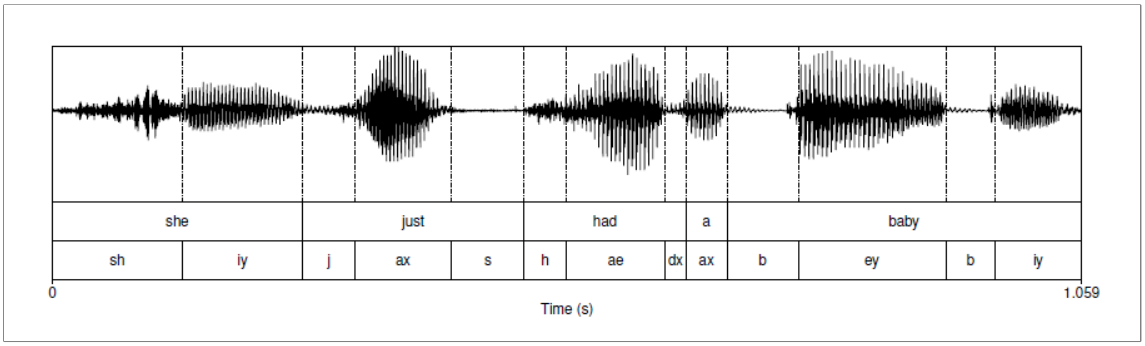
Dessa forma, considerando a discussão desenvolvida até aqui, nesta pesquisa a RLS será posicionada como uma tarefa de PLN, delimitando-se seu escopo ao âmbito da linguística e representando-se os sinais através de seus fonemas. Além disso, será aplicado o conhecimento disponibilizado pelas ferramentas acima para criar *features* que representem estes fonemas, viabilizando este processo. Com isso, objetiva-se eliminar o escopo pertinente a outras áreas

[1] O OpenPose está disponível em <https://github.com/CMU-Perceptual-Computing-Lab/openpose> e o MediaPipe em <https://mediapipe.dev/>.

de pesquisa e concentrar a capacidade dos modelos aplicados ao aprendizado das regras e restrições linguísticas da língua de sinais.

Tal estratégia de abordar a linguagem por meio de suas unidades constituintes mínimas é também observada em outras tarefas de PLN. Jurafsky e Martin (2022) afirmam que a ideia da palavra falada ser composta por unidades menores da fala é adotada, por exemplo, por algoritmos utilizados em tarefas de reconhecimento de fala e de conversão de texto em voz. Observe na Figura 21 o exemplo ilustrado pelos autores da forma de onda da fala para a sentença em inglês "*she just had a baby*" (ou "ela acabou de ter um bebê"). Cada trecho é rotulado na linha inferior com suas respectivas partículas mínimas de som (ou "fones"), as quais são transcritas utilizando-se o ARPAbet[2]. Esse tipo de partícula é comumente utilizado como *feature* de entrada para tarefas envolvendo o processamento da fala.

Figura 21 – Formas de onda da fala para a sentença "*she just had a baby*" (primeira linha) rotuladas com suas respectivas partículas de som transcritas em ARPAbet (linha inferior)



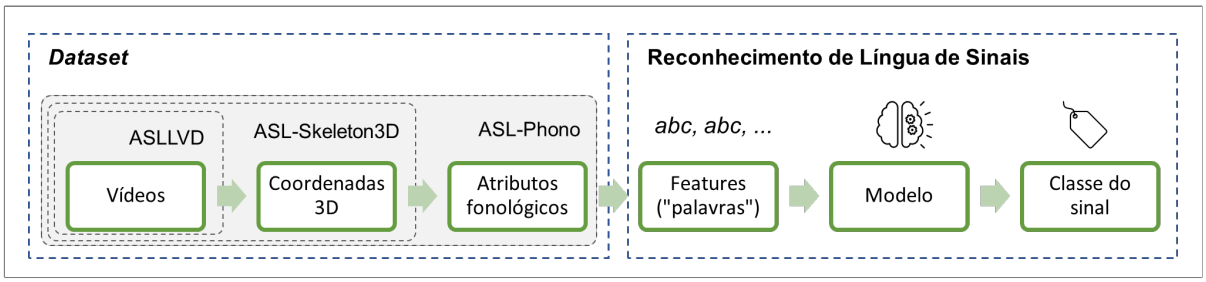
Fonte: Jurafsky e Martin (2022, p. 586)

No caso da abordagem proposta aqui, essas partículas serão substituídas por alguns dos parâmetros fonológicos introduzidos na seção 2.2. Uma vez que não há *datasets* disponíveis com esse tipo de representação para as línguas de sinais, o primeiro passo consistirá em gerar esse *dataset*. Para isso, o ASLLVD será adotado como *dataset* base e suas amostras serão processadas utilizando-se o OpenPose, o qual fornece coordenadas que possibilitam a extração de *features* fonológicas de nível semântico mais elevado. Em seguida, serão aplicados modelos de Aprendizagem de Máquina (AM) comumente utilizados em tarefas de PLN, com o intuito de avaliar seu desempenho neste contexto e a eficácia da abordagem proposta.

[2] ARPAbet é um alfabeto fonético simples introduzido por Shoup (1980) que utiliza símbolos ASCII para representar um subconjunto do IPA que se refere ao idioma inglês-americano. O *International Phonetic Alphabet* (ou Alfabeto Fonético Internacional) (IPA), por sua vez, é a representação fonética padrão para a transcrição das línguas ao redor do mundo (JURAFSKY; MARTIN, 2022).

A Figura 22 ilustra essa abordagem, a qual é divida em duas etapas. No bloco à esquerda, observa-se o processo envolvido na geração do *dataset*, que inicia-se pelas amostras do ASLLVD, contempla a obtenção de coordenadas 3D por meio de ferramentas de VC e finaliza com a geração do ASL-Phono, que contém os respectivos atributos fonológicos. No bloco à direita, está a etapa de Reconhecimento de Língua de Sinais, que engloba a preparação das *features*, o processamento dessas *features* pelos modelos de AM e a classificação dos sinais. Todas essas etapas serão discutidas em detalhes nas seções a seguir.

Figura 22 – Etapas envolvidas na abordagem proposta



Fonte: Elaborada pelo autor.

3.1 CRIAÇÃO DOS *DATASETS*

O primeiro passo para desenvolver e avaliar a abordagem proposta consiste em estabelecer um *dataset* que viabilize essa forma de modelar o problema. Como a proposta apresentada aqui é nova, não existem *datasets* diretamente compatíveis com ela e, por este motivo, será derivado um novo a partir de outro já existente – o *American Sign Language Lexicon Video Dataset* (ASLLVD).

O ASLLVD (ATHITSOS et al., 2008; NEIDLE; THANGALI; SCLAROFF, 2012) é um *dataset* público[3] amplo da ASL que contém aproximadamente 2.745 sinais representados em cerca de 9.763 sequências de vídeo. Esses sinais são articulados por indivíduos Surdos nativos na língua e foram capturados por meio de quatro câmeras distintas sincronizadas: uma visão frontal em alta resolução a meia velocidade, outra visão frontal, uma visão lateral e uma visão da face, conforme ilustrado na Figura 23.

Para que fosse possível extrair *features* no formato de parâmetros fonológicos a partir das amostras do ASLLVD, que são compostas essencialmente de sequências de vídeos com *frames* em RGB, foi necessário realizar um processo em duas etapas.

[3] Disponível em <http://www.bu.edu/asllrp/av/dai-asllvd.html>

Figura 23 – Exemplo de três perspectivas capturadas pelo ASLLVD para o sinal MERRY-GO-ROUND

Fonte: Athitsos et al. (2008, p. 2)

Na primeira, foi realizada a estimativa das coordenadas dos esqueletos dos indivíduos de cada *frame* para duas das perspectivas fornecidas para as amostras, os quais foram combinados em seguida para compor um esqueleto tridimensional final. A saída dessa etapa deu origem ao *dataset* intermediário denominado ASL-Skeleton3D. Na segunda etapa, foi aplicado um conjunto de operações algébricas sob esse esqueleto tridimensional para finalmente computar os parâmetros fonológicos, processo esse que originou o *dataset* ASL-Phono.

Esse processo de extração de *features* envolveu alguns desafios relevantes, dentre os quais podem-se enumerar:

1. Definição de uma estratégia para representar indivíduos no espaço tridimensional utilizando apenas *frames* de vídeo bidimensionais simples, bem como para contornar a ausência de perspectivas ou a baixa qualidade de algumas dessas amostras;

2. Estabelecimento de um subconjunto inicial de atributos fonológicos capaz de capturar e representar variações significativas na articulação dos sinais, e que ao mesmo tempo pudessem ser modelados computacionalmente nessa primeira iteração da abordagem proposta.

3. Identificação de técnicas matemáticas e medidas antropométricas que pudessem fundamentar a modelagem e o cálculo desses atributos.

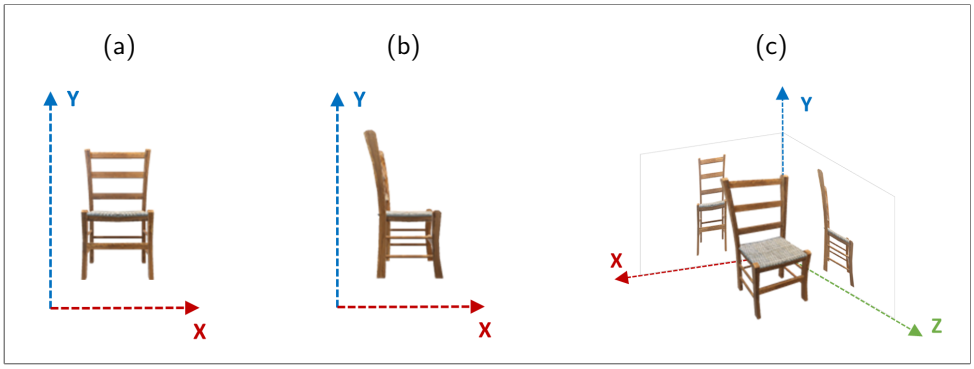
4. Demanda por recursos computacionais significativos para processar duas perspectivas distintas para cada uma das quase 10.000 amostras contidas em cada *dataset*. Em média, isso consumiu cerca de 120 horas contínuas de processamento distribuído com GPUs e gerou mais de 1 TB de dados cada vez que os *datasets* precisaram ser gerados novamente.

3.1.1 ASL-Skeleton3D

O ASL-Skeleton3D é um *dataset* intermediário que introduz a representação em coordenadas 3D das amostras do ASLLVD. Esse tipo de informação fornece detalhes mais precisos acerca do corpo dos indivíduos enquanto eles articulam os sinais, o que possibilita a pesquisadores em RLS extrair diferentes tipos de *features*, explorar novas técnicas, ou ainda derivar outros novos *datasets*. Além disso, essa representação é genérica e pode ser replicada para outras línguas de sinais.

Para que fosse possível projetar as amostras do ASLLVD dentro do espaço tridimensional, foi adotada uma estratégia que consiste em combinar duas perspectivas 2D perpendiculares entre si – a vista frontal e a vista lateral – para reconstruir uma perspectiva 3D, conforme ilustra a Figura 24. Com isso, assume-se que enquanto a vista frontal fornece os eixos x e y, a vista lateral fornecerá a dimensão de profundidade, correspondente ao eixo z.

Figura 24 – Estratégia adotada para representar as amostras no espaço 3D: as perspectivas frontal (a) e lateral (b) são posicionadas perpendicularmente para reconstruir uma perspectiva 3D (c)



Fonte: Elaborada pelo autor.

Uma vez definida essa estratégia, foi considerado o processo descrito por Amorim, Macêdo e Zanchettin (2019) para realizar a estimativa dos esqueletos para os indivíduos nas amostras. Esse processo é composto pelas etapas de obtenção de amostras, segmentação dos sinais, estimativa e normalização dos esqueletos, as quais sofreram adaptações no contexto do presente trabalho para acomodar a composição de esqueletos 3D e os desafios encontrados para isso. Serão discutidas a seguir as adaptações realizadas para superar as limitações dos *datasets* existentes:

1. **Obtenção das amostras**: nessa etapa as amostras de vídeos são recuperadas a partir dos servidores do ASLLVD, mas no contexto atual, isso é realizado para ambas as

câmeras frontal e lateral.

A saber, existem dois formatos nos quais os vídeos dessas câmeras podem ser disponibilizados: o *mov*, que é compacto e mais fácil de processar; e o *.vid*, que consiste no vídeo bruto, que é maior e mais pesado para baixar e processar. Ao analisar as amostras, identificou-se que parte delas possuía ambas câmeras disponíveis nos dois formatos; para outras, cada câmera estava em um formato distinto; contudo, nos piores casos uma das câmeras estava ausente ou corrompida, fazendo com que essas amostras fossem perdidas. Lidar com essa falta de homogeneidade acrescentou uma complexidade inesperada, mas que foi contornada resultando numa perda de apenas 0,16% das amostras originais – ou seja, 16 amostras de um total inicial de 9.763.

2. **Segmentação dos sinais**: compreende a segmentação das sequências de vídeos do ASLLVD em pedaços menores, contendo um único sinal. Nessa etapa também reduziu-se a taxa de quadros de 60 para 3 FPS, uma vez que considera-se ser pouco provável a articulação de mais de três movimentos relevantes em um único segundo. Isso contribuiu para reduzir em cerca de 20 vezes o número de *frames* a serem processados nas etapas seguintes.

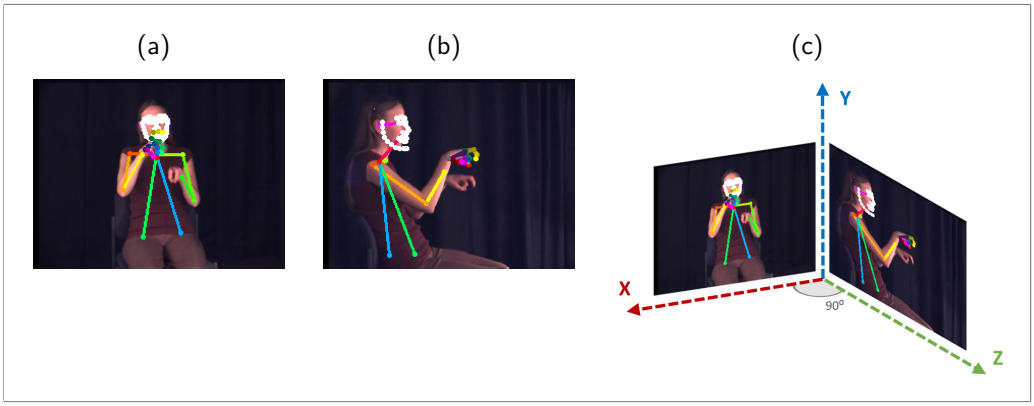
3. **Estimativa dos esqueletos 3D**: nesse momento os esqueletos dos sinalizadores são estimados utilizando-se o OpenPose para ambas as câmeras frontal e lateral. Dois esqueletos 2D são obtidos a partir disso:

    a) *Esqueleto frontal* (Figura 25a), contendo as coordenadas plotadas sobre o eixos x e y que correspondem às mesmas coordenadas x e y de quando imagina-se o indivíduo representado no espaço tridimensional (vide Figura 25c).

    b) *Esqueleto lateral* (Figura 25b), que também é estimado com um par de coordenadas x e y mas que, quando posicionados perpendicularmente ao esqueleto frontal (vide Figura 25c), denotam a dimensão de profundidade e correspondem aos eixos z e y do indivíduo no espaço tridimensional, respectivamente. Uma vez que y já foi obtido por meio do esqueleto frontal, ele pode ser então descartado.

Dessa forma, são combinadas as coordenadas x, y e z obtidas para dar origem ao esqueleto 3D.

Figura 25 – Os esqueletos 2D frontal (a) e lateral (b) são posicionados perpendicularmente (c) e combinados para compor o esqueleto 3D final utilizado aqui

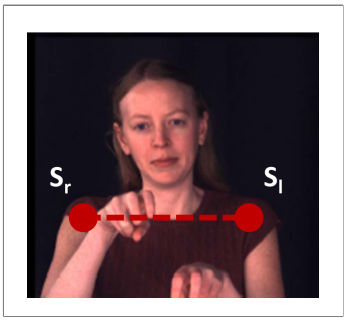


Fonte: Elaborada pelo autor.

4. **Normalização dos esqueletos 3D**: por fim, os esqueletos 3D são normalizados para remover variações decorrentes do posicionamento das câmeras e dos corpos dos indivíduos. Isso é importante porque o ASLLVD foi capturado através de diferentes seções e envolveu diferentes sinalizadores.

   Adotou-se como referência para essa normalização a largura entre os ombros dos sinalizadores (vide Figura 26), a qual foi inspirada pela medida antropométrica de *diâmetro biacromial* apresentada por Stoudt, Damon e McFarland (1970). Dessa forma, a largura entre os ombros $W_{shoulders}$ é definida aqui como a distância euclidiana d entre as coordenadas do ombro esquerdo S_l e do ombro direito S_r, conforme Equação 3.1:

Figura 26 – A largura entre ombros foi utilizada para normalizar as coordenadas nos esqueletos 3D



Fonte: Elaborada pelo autor.

$$W_{shoulders} = d\left(S_l, S_r\right) \tag{3.1}$$

Utilizando-se $W_{shoulders}$ é possível transformar as coordenadas K do esqueleto 3D em coordenadas normalizadas K_{norm}, conforme Equação 3.2:

$$K_{norm} = \frac{K}{W_{shoulders}} \tag{3.2}$$

O Código-fonte 1 exemplifica uma amostra do ASL-Skeleton3D resultante do processamento acima, bem como suas propriedades. Nele, atributos com sufixo "dh" referem-se à *dominant hand* (mão dominante) e aqueles com "ndh" referem-se à *non-dominant hand* (mão não-dominante). Observa-se no início do arquivo informações básicas extraídas do ASLLVD, como rótulo, nome do indivíduo, sessão, cena, *frames* de início e fim, entre outras. Na propriedade "*frames*" estão listados os *frames* para aquela sequência e o esqueleto 3D estimado para cada um deles. Cada esqueleto contém grupos referentes ao "*body*" (corpo), "*face*" (face), "*hand left*" (mão esquerda) e "*hand right*" (mão direita) que, por sua vez, contêm propriedades que listam o nome da coordenada, o "*score*" (ou acurácia) de sua estimativa e os respectivos eixos x, y e z. Por exemplo, ao observar o primeiro índice das propriedades do grupo "*body*" percebe-se que ele corresponde à coordenada "*nose*" (nariz), apresenta um *score* de aproximadamente 90% e coordenadas x, y e z localizadas em 4,488, 1,696 e 2,872.

Código-fonte 1 – Exemplo de amostra do *dataset* ASL-Skeleton3D

```
// Exemplo de amostra do ASL-Skeleton3D
{
    label: "merry-go-round",
    gloss: "MERRY-GO-ROUND",
    consultant: "Liz",
    session: "ASL_2008_01_25",
    scene: 5,
    frame_start: 100,
    frame_end: 129,
    handshape_dh_start: "crvd-U",
    handshape_dh_end: "crvd-U",
    handshape_ndh_start: "crvd-U",
    handshape_ndh_end: "crvd-U",
    passive_arm: "N",
    fps: 3,
    mode: "3d",
    frames: [
        {
            frame_index: 100,
            skeleton: {
                body: {
                    name: [
                        "nose", "neck", "shoulder_right", ...
                    ],
                    score: [
                        0.8980225, 0.62372, 0.5912965, ...
```

```
                    ],
                    x: [
                        4.488181140105, 4.468695935888, 3.477285931555, ...
                    ],
                    y: [
                        1.696915084399, 2.871106152409, 2.871303114762, ...
                    ],
                    z: [
                        2.872175376611, 1.898548259016, 1.622871308373, ...
                    ]
                },
                face:       { ... },
                hand_left:  { ... },
                hand_right: { ... }
            }
        },
        // ...
    ]
}
```

Fonte: Elaborada pelo autor.

O *dataset* final e o código-fonte utilizado para o processamento apresentado nesta seção estão disponíveis publicamente na URL listada abaixo[4].

3.1.2 ASL-Phono

O ASL-Phono é um *dataset* que introduz uma representação baseada na linguística da língua de sinais e a descreve em termos de seus atributos fonológicos. Ele é produzido a partir dos esqueletos fornecidos pelo *dataset* ASL-Skeleton3D e, assim como este, também apresenta 9.747 amostras correspondentes a 2.650 sinais distintos. Além disso, a abordagem utilizada para computar esses atributos pode ser replicada para outras línguas de sinais.

Por se tratar de uma versão inicial da representação proposta, neste trabalho será selecionado apenas um subconjunto dos parâmetros fonológicos introduzidos na seção 2.2 para que, dessa forma, seja possível validar sua efetividade e traçar uma direção para iterações futuras. Sendo assim, descrevem-se a seguir esses parâmetros, bem como a estratégia utilizada para calculá-los e representá-los computacionalmente:

1. **Configuração de mão**: é a configuração de mão utilizada pelo sinalizador na articulação do sinal.

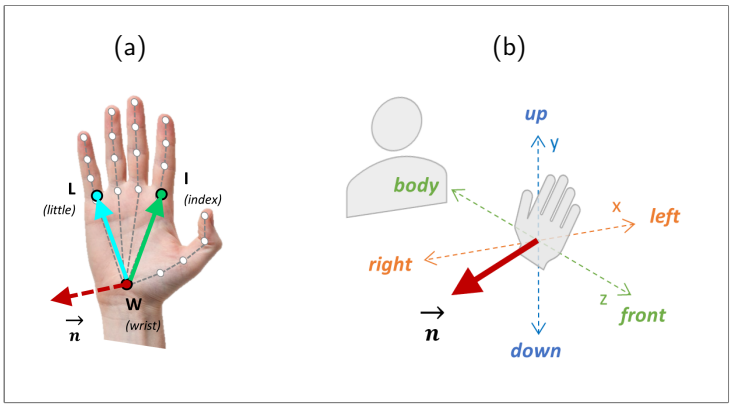
[4] Disponível em <http://www.cin.ufpe.br/~cca5/asl-skeleton3d>

O ASLLVD fornece originalmente as configurações de mão inicial e final para cada sinal, descrita de acordo com as 88 opções apresentadas por Neidle e Opoku (2020) no *American Sign Language Linguistic Research Project* (ASLLRP)[5]. Foi utilizada essa mesma informação para o ASL-Phono, porém adicionou-se um passo extra para distribuir essas configurações entre todos os *frames*, e não apenas o inicial e final. Dessa forma, dividiu-se a sequência de *frames* em duas metades e associou-se à primeira delas a configuração inicial provida pelo ASLLVD e, à segunda, a configuração final.

2. **Orientação**: é a direção apontada pelas palmas das mãos na articulação dos sinais.

   Para calculá-la, foi utilizada álgebra linear para explorar a relação das mãos com o espaço tridimensional em que suas coordenadas estão inseridas. Primeiramente, assumiu-se que cada palma é um plano cartesiano que atravessa as coordenadas estimadas para as mãos (vide Figura 27a). Selecionaram-se então três dessas coordenadas para descrever o plano: W, que corresponde à coordenada do pulso; L, localizada na base do dedo mínimo; e I, localizada na base do dedo indicador.

Figura 27 – As coordenadas W, L e I são utilizados para obter a normal $\overrightarrow{n}$ da palma da mão (a), a qual é utilizada para calcular a orientação da palma O_{palm} (b)



Fonte: Elaborada pelo autor.

A partir dessas coordenadas, Anton e Rorres (2005) afirmam que pode-se estabelecer dois vetores auxiliares em termos dos quais esse mesmo plano cartesiano também é descrito (vide Figura 27a): $\overrightarrow{WL}$, indicado pela seta azul e $\overrightarrow{WI}$, indicado pela seta verde. Por meio deles, calculou-se o vetor normal $\overrightarrow{n}$ utilizando a Equação 3.3 (para a mão esquerda) e Equação 3.4 (para a mão direita). $\overrightarrow{n}$, que é perpendicular à palma da mão, é ilustrado na Figura 27a como uma seta tracejada vermelha.

[5] Disponível em <http://www.bu.edu/asllrp>

$$\overrightarrow{n}_{left} = \overrightarrow{WI} \times \overrightarrow{WL} \tag{3.3}$$

$$\overrightarrow{n}_{right} = \overrightarrow{WL} \times \overrightarrow{WI} \tag{3.4}$$

Por fim, utilizou-se os valores das coordenadas de $\overrightarrow{n}$ para definir a orientação da palma O_{palm}, conforme Equação 3.5. Essa orientação consiste na combinação de até três das seguintes direções: *right* (direita), *left* (esquerda), *up* (para cima), *down* (para baixo), *body* (voltada para o corpo) ou *front* (para frente). Por exemplo, "*right_front_down*" seria uma orientação válida indicando que a palma da mão está inclinada, conforme ilustrado na Figura 27b.

$$O_{palm} = \begin{cases} right & \text{if } \overrightarrow{n}_x < -k \\ left & \text{if } \overrightarrow{n}_x > k \\ up & \text{if } \overrightarrow{n}_y < -k \\ down & \text{if } \overrightarrow{n}_y > k \\ body & \text{if } \overrightarrow{n}_z < -k \\ front & \text{if } \overrightarrow{n}_z > k \end{cases} \tag{3.5}$$

Na Equação 3.5, k é definido empiricamente como 0,30 para filtrar variações pouco significativas em $\overrightarrow{n}$.

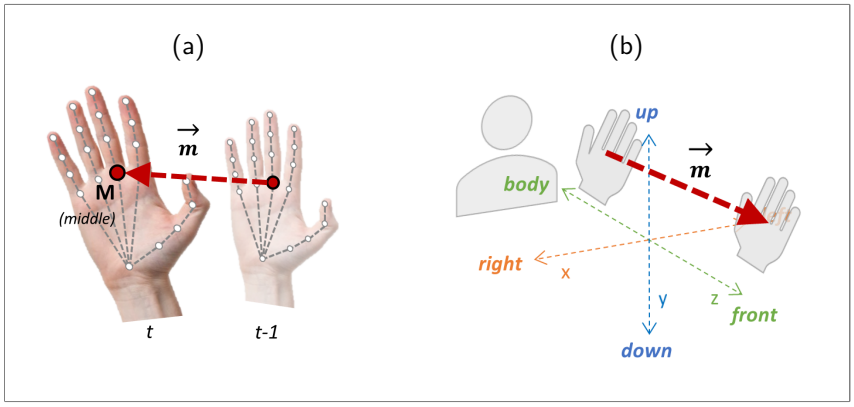
3. **Movimento**: é o deslocamento realizado pelas mãos na articulação do sinal.

   Para calculá-lo, primeiro será selecionada como referência a coordenada M localizada na base do dedo médio (vide Figura 28a). Em seguida, será obtido seu deslocamento entre os *frames* anterior (tempo $t-1$) e atual (tempo t) utilizando a Equação 3.6 que, por sua vez, fornecerá o vetor de movimento $\overrightarrow{m}$ (indicado pela seta tracejada vermelha na Figura 28a).

$$\overrightarrow{m} = M_t - M_{t-1} \tag{3.6}$$

A partir de $\overrightarrow{m}$, pode-se então calcular o movimento da mão V_{hand} através da Equação 3.7, que consiste numa operação semelhante àquela utilizada para a orientação da mão. Com

Figura 28 – O vetor de movimento $\overrightarrow{m}$ é obtido através da trajetória da coordenada M entre os frames anterior $(t-1)$ e atual (t) (a); $\overrightarrow{m}$ é então utilizado para calcular o movimento da mão V_{hand} (b)



Fonte: Elaborada pelo autor.

isso, V_{hand} também consistirá na combinação de até três direções: *right* (direita), *left* (esquerda), *up* (para cima), *down* (para baixo), *body* (para o corpo) ou *front* (para frente). A Figura 28b ilustra um movimento categorizado com a direção "*front*".

$$V_{hand} = \begin{cases} right & \text{if } \overrightarrow{m}_x < -k \\ left & \text{if } \overrightarrow{m}_x > k \\ up & \text{if } \overrightarrow{m}_y < -k \\ down & \text{if } \overrightarrow{m}_y > k \\ body & \text{if } \overrightarrow{m}_z < -k \\ front & \text{if } \overrightarrow{m}_z > k \end{cases} \tag{3.7}$$

Na Equação 3.7 o limiar k foi também estabelecido empiricamente como 0,30, para filtrar movimentos com baixa relevância.

4. **Expressão não-manual (abertura da boca)**: captura o grau de abertura da boca para cada *frame* na articulação do sinal que, por sua vez, denota a existência de expressão não-manual envolvendo ela.

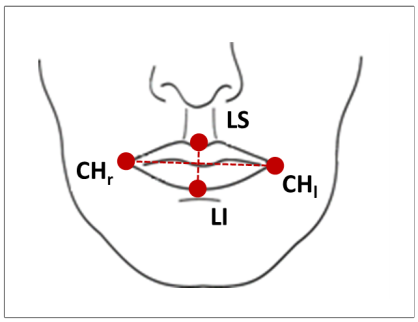
   Para computar essa *feature*, utilizou-se como referência o trabalho de Ferrario et al. (2000), que analisa e propõe medidas antropométricas para os lábios. Dentre elas, foi selecionada a *vermilion height to mouth width* (ou altura dos lábios com relação à largura da boca) para estabelecer a abertura da boca P_{mouth}, uma vez que essa medida é capaz de capturar a proporção de abertura dos lábios em termos de um único número.

A Equação 3.8 define formalmente o cálculo de P_{mouth}, que consiste na proporção entre a altura e a largura dos lábios:

$$P_{mouth} = \frac{d(LS, LI)}{d(CH_r, CH_l)} \tag{3.8}$$

A altura dos lábios é dada pela distância d entre as coordenadas do *labiale superius* LS e do *labiale inferius* LI, que são os pontos mais externos aos lábios superior e inferior, respectivamente. A largura, por sua vez, consiste na distância entre as coordenadas do *cheilion* direito CH_r e do *cheilion* esquerdo CH_l, que são os pontos situados nos cantos direito e esquerdo dos lábios, conforme ilustra a Figura 29.

Figura 29 – A abertura da boca P_{mouth} é obtida a partir da medida antropométrica *vermilion height to mouth width* que utiliza quatro coordenadas dos lábios para calcular uma única proporção



Fonte: Ferrario et al. (2000)

Essas são as *features* extraídas para o ASL-Phono. O Código-fonte 2 exemplifica uma amostra resultante desse processo, bem como a disposição dos atributos fonológicos em suas propriedades. Atributos com sufixo "dh" referem-se à *dominant hand* (mão dominante) e aqueles com "ndh" referem-se à *non-dominant hand* (mão não-dominante). Sua estrutura é muito semelhante àquela apresentada para o ASL-Skeleton3D, exceto pela propriedade "*frames*" que, ao invés de coordenadas 3D, aqui contém os atributos computados acima. Além do seu respectivo valor computado, cada atributo pode apresentar também um *score*, que é obtido a partir da precisão estimada para as coordenadas envolvidas no seu cálculo.

Código-fonte 2 – Exemplo de amostra do *dataset* ASL-Phono

```
// Exemplo de amostra do ASL-Phono
{
    label: "merry-go-round",
    gloss: "MERRY-GO-ROUND",
    consultant: "Liz",
```

```
    session: "ASL_2008_01_25",
    scene: 5,
    frame_start: 100,
    frame_end: 129,
    handshape_dh_start: "crvd-U",
    handshape_dh_end: "crvd-U",
    handshape_ndh_start: "crvd-U",
    handshape_ndh_end: "crvd-U",
    passive_arm: "N",
    fps: 3,
    mode: "3d",
    phono_attributes: [
        {
            frame_index: 100,
            movement_dh:        { value: "right_up",
                                  score: 0.837191500001 },
            movement_ndh:       { value: "left_front_down",
                                  score: 0.36356175     },
            orientation_dh:     { value: "left",
                                  score: 0.72990875     },
            orientation_ndh:    { value: "back_right",
                                  score: 0.317478625    },
            handshape_dh:       { value: "bent-B-L"     },
            handshape_ndh:      { value: "bent-B-L"     },
            non_manual: {
                mouth_opening:  { value: 0.028637266132,
                                  score: 0.876049       }
            }
        },
        // ...
    ]
}
```

Fonte: Elaborada pelo autor.

A Tabela 6, por sua vez, apresenta estatísticas calculadas para o *dataset* resultante que nos fornecem um panorama da distribuição final de suas amostras, bem como do número de movimentos, orientações, configurações de mão e variação na abertura de boca computados acima.

Existem três agrupamentos contidos nessa tabela:

- *Dataset*: sumariza o total de amostras, de sinais e de cada um dos atributos computados para o *dataset*. Por exemplo, há um total de 9.747 amostras e 26 movimentos possíveis para a mão dominante.
- Por amostra: fornece estatísticas de mínimo, máximo, média e desvio padrão calculadas agrupando-se o *dataset* por amostra. Por exemplo, há em média 3,02 *frames* por amostra

Tabela 6 – Estatísticas calculadas a partir do ASL-Phono, as quais são visualizadas para todo o *dataset* e segundo agrupamentos por amostra e por sinal. (D) refere-se à mão dominante e (ND) refere-se à mão não-dominante

		Nº amostras	Nº sinais	Nº frames	Movimento		Orientação		Config. mão		Abertura
					(D)	(ND)	(D)	(ND)	(D)	(ND)	da boca
Dataset	Total	9.747	2.650	-	26	26	26	26	85	78	-
Por amostra	Mín	-	-	1	0	0	0	0	0	0	0,01
	Máx	-	-	12	10	8	6	5	2	2	2,19
	Média	-	-	3,02	1,94	1,26	2,20	1,18	1,17	0,72	0,13
	Desvio	-	-	0,87	0,83	1,15	0,79	1,07	0,38	0,58	0,12
Por sinal	Mín	1	-	-	1	0	1	0	1	0	0,02
	Máx	59	-	-	24	16	22	14	8	8	0,99
	Média	3,68	-	-	6,04	3,66	5,37	2,63	1,81	1,17	0,13
	Desvio	2,67	-	-	2,93	3,18	2,29	2,35	0,99	1,14	0,07

Fonte: Elaborada pelo autor.

e esse número varia de 1 a 12 *frames*; de maneira semelhante, as amostras têm em média 1,94 movimentos distintos para a mão dominante, o que varia entre 0 e 10 movimentos.

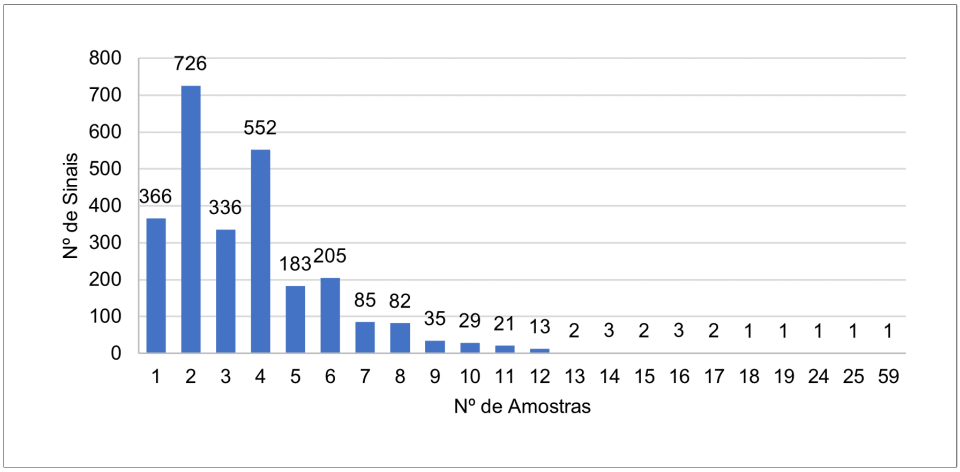
- Por sinal: fornece estatísticas de mínimo, máximo, média e desvio padrão calculadas agrupado-se o *dataset* por sinal. Por exemplo, cada sinal tem em média 3,68 amostras e isso pode variar de 1 até 59 – no entanto, o desvio padrão nos revela que tal variação concentra-se muito mais em torno da média; nota-se também que a mão dominante realiza em média 6,04 movimentos distintos em um único sinal, o que pode variar de 1 a 24 para outros.

Por fim, a Figura 30 detalha a relação entre o número de sinais e o número de amostras existentes no ASL-Phono. Por meio desta análise, percebe-se mais claramente que a maior parte dos sinais concentra-se na faixa de 1 e 6 amostras, que é exatamente o intervalo descrito pela média e desvio padrão da tabela acima. No entanto, há sinais que apresentam 7 ou mais amostras e ainda outros casos atípicos em que 1, 2 ou 3 sinais que concentram sozinhos um número muito grande de amostras – de 13 até 59. Será discutida na seção seguinte a abordagem utilizada neste trabalho para lidar com esse desbalanceamento e homogeneizar as amostras antes de prosseguir com os experimentos.

O *dataset* e o código-fonte resultantes do processamento apresentado nesta seção estão disponíveis publicamente na URL listada abaixo[6].

[6] Disponível em <http://www.cin.ufpe.br/~cca5/asl-phono>

Figura 30 – Relação entre número de sinais e número de amostras disponíveis no ASL-Phono



Fonte: Elaborada pelo autor.

3.2 PREPARAÇÃO DOS DADOS

Conforme introduzido na seção anterior, o número de amostras disponíveis por sinal não está balanceado homogeneamente no ASL-Phono e, como consequência, isso poderia influenciar de maneira indesejada o desempenho dos modelos utilizados nos experimentos adiante, fazendo com que algumas classes fossem extremamente favorecidas e, outras, severamente penalizadas.

Devido a isso, serão aplicados dois procedimentos para tentar equalizar essa proporção. Primeiro, serão descartados aqueles sinais que apresentam apenas 1 amostra disponível, uma vez que esse número é insuficiente para permitir o modelo aprender e generalizar tais sinais, sobretudo porque o *dataset* será particionado em mais de um subconjunto durante seu treinamento e todas as classes precisam estar igualmente representadas neles.

Em seguida, será realizada uma reamostragem do *dataset* com o intuito de balancear melhor a proporção de amostras. Será utilizado para isso uma reamostragem *Naive Random Under-Sampling* (ou sub-amostragem aleatória ingênua), que reduz o número de amostras super-representadas selecionando aleatoriamente algumas delas e, em seguida, uma *Naive Random Over-Sampling* (ou sobre-amostragem aleatória ingênua) que, por sua vez, aumenta o número de amostras sub-representadas replicando aleatoriamente algumas existentes (HE; MA, 2013).

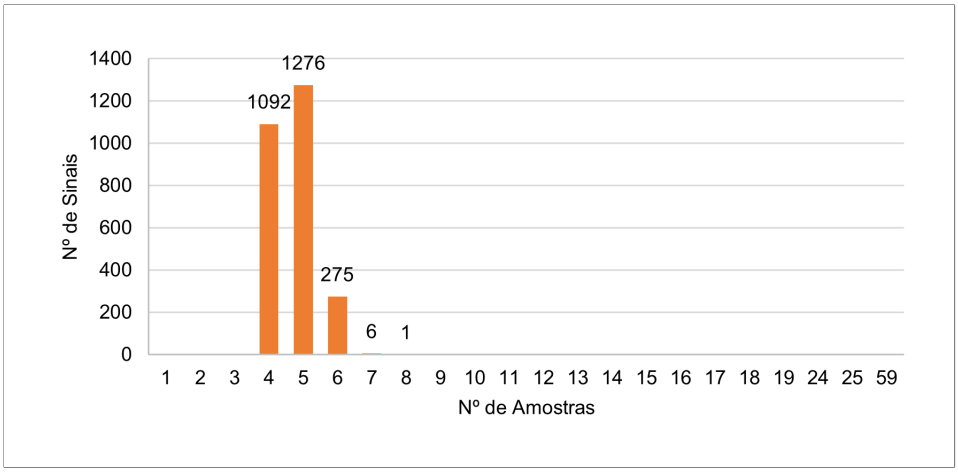
A Equação 3.9 define a operação aplicada para definir o número de amostras n' a ser obtido para cada sinal por processo de reamostragem. Nela, $\overline{m}$ refere-se à média de amostras

por sinal e n é o número atual de amostras para aquele sinal:

$$n' = round(\overline{m} + \ln(n)) \tag{3.9}$$

Observe na Figura 31 a nova distribuição das amostras no *dataset*. De uma forma resumida, ao comparar com a Figura 30, percebe-se que a reamostragem homogeneizou a relação de amostras por sinal em torno da nova média do conjunto, trazendo para essa região também os antigos *outliers* (valores atípicos ou pontos fora da curva).

Figura 31 – Distribuição do número de amostras por sinal após a reamostragem



Fonte: Elaborada pelo autor.

Por fim, será aplicada uma transformação às amostras do ASL-Phono para tornar a estrutura apresentada no Código-fonte 2 compatível com a entrada dos modelos que serão aplicados mais adiante. Para isso, serão compactados os valores informados para os atributos fonológicos de cada *frame* em uma "palavra" única, fazendo com que a sequência de *frames* seja então representada como uma sequência dessas palavras.

Por exemplo, considerando-se um *frame* contendo dois atributos com os valores "*valor_atributo_1*" e "*valor_atributo_2*", ao compactá-los, eles primeiro seriam abreviados para "*va1*" e "*va2*" e, em seguida, concatenados para formar uma palavra "*va1-va2*". A sequência de *frames* da amostra tornaria-se, portanto, algo semelhante a uma sequência de palavras *{"va1-va2", "va3-va4", . . . , "vaN-vaN"}*.

Observe na Tabela 7 um exemplo mais próximo do contexto real para esse processo. Na primeira linha estão os valores originais dos atributos do *frame*; na segunda, os valores abreviados; e, na terceira, a palavra formada através da concatenação deles.

Tabela 7 – Exemplo de compactação dos atributos fonológicos do *frame* de uma amostra do ASL-Phono em uma "palavra"

	Atributos					
	Mão dominante			Mão não-dominante		
	Movimento	Orientação	Config. mão	Movimento	Orientação	Config. mão
Valores originais	*right_up*	*left*	*bentBL*	*left_front_down*	*back_right*	*bentBL*
Valores abreviados	*ru*	*l*	*bentBL*	*lfd*	*br*	*bentBL*
Palavra	*ru-l-bentBL-lfd-br-bentBL*					

Fonte: Elaborada pelo autor.

3.3 PREPARAÇÃO DOS MODELOS

Tomando como referência a discussão introduzida na subseção 2.3.1, serão adotados nos experimentos deste trabalho três das principais arquiteturas utilizadas em tarefas de Processamento de Linguagem Natural (PLN): o *Encoder-Decoder* em uma versão com *Long Short-Term Memory* (LSTM) e outra com *Gated Recurrent Unit* (GRU), além do *Transformer*.

Para estabelecer os parâmetros dessas arquiteturas, as estratégias de otimização e de treinamento, bem como as métricas utilizadas nos experimentos, foram consideradas as discussões apresentadas por Goodfellow, Bengio e Courville (2016) e pela seção 2.3.

Dessa forma, o algoritmo de otimização dos modelos será definido como o *Stochastic Gradient Descent* (ou Gradiente Estocástico Descendente) (SGD) com *momentum* de 0,9 (ROBBINS, 2007). Ele será combinado a uma estratégia de redução da taxa de aprendizagem por um fator de 0,2 sempre que o valor do erro médio calculado atingir um platô por 5 épocas seguidas.

A função objetivo (ou função de perda), por sua vez, será a *Cross-Entropy Loss* (ou Perda de Entropia Cruzada) (CEL) (MITCHELL, 1997), que é apresentada na Equação 3.10. Nela, p representa as probabilidades ou pontuações estimadas pelo modelo para as amostras e y corresponde ao valor correto esperado para essas estimativas:

$$L_{\log}(y, p) = -(y \log(p) + (1 - y) \log(1 - p)) \tag{3.10}$$

Os dados serão particionados numa proporção de 15% para o subconjunto de validação, 15% para o de testes e os 70% restantes para o subconjunto de treinamento. Os *batches* (ou lotes), por sua vez, possuirão tamanho de 50 amostras.

A seleção dos hiperparâmetros dos modelos foi realizada utilizando-se o algoritmo *Grid Search* (ou busca em grade) com validação cruzada de 5 *folds*. O conjunto de valores de

Tabela 8 – Hiperparâmetros otimizados para o *Encoder-Decoder* (LSTM), *Encoder-Decoder* (GRU) e o *Transformer* com os respectivos erros médios calculados (*Cross-Entropy Loss* (ou Perda de Entropia Cruzada) (CEL))

Parâmetro		Erro médio (CEL)		
		Encoder-Decoder (LSTM)	**Encoder-Decoder (GRU)**	**Transformer**
Taxa de aprendizagem	0,001	7,734328	7,278577	2,846646
	0,01	7,732102	4,797364	0,020484
	0,1	5,390974	4,903971	0,019960
Dropout	0,1	5,390974	4,797364	0,019960
	0,5	5,570855	4,904458	0,024303
Tamanho embeddings	128	7,733950	5,722228	0,027742
	512	5,794343	5,064633	0,021442
	1024	5,390974	4,797364	0,019960
Tamanho camadas ocultas	128	5,565219	5,722228	0,020796
	256	5,390974	5,064633	0,019960
	512	5,571651	4,797364	0,020773
Nº de camadas	2	5,390974	4,797364	0,019960
	4	7,734451	6,197899	0,021727
	6	7,734467	7,152559	0,068949
Nº de cabeças	4	N/A	N/A	0,020086
	8			0,019960

Fonte: Elaborada pelo autor.

hiperparâmetros utilizados na busca estão apresentados na Tabela 8 e as combinações que melhor reduziram o erro médio para cada modelo foram as seguintes:

- *Encoder-Decoder* com LSTM: taxa de aprendizagem de 0,1; *dropout* de 0,1; *embeddings* com dimensão de 1024; camadas ocultas com dimensão de 256; e utilização de 2 camadas de LSTM no codificador e no decodificador.
- *Encoder-Decoder* com GRU: taxa de aprendizagem de 0,01; *dropout* de 0,1; *embeddings* com dimensão de 1024; camadas ocultas com dimensão de 512; e utilização de 2 camadas de GRU no codificador e no decodificador.
- *Transformer*: taxa de aprendizagem de 0,1; *dropout* de 0,1; *embeddings* com dimensão de 1024; camadas ocultas com dimensão de 256; utilização de 2 camadas de codificadores e decodificadores e de 8 cabeças de *attention*.

Para a execução dos experimentos – que contemplaram ciclos completos de busca de parâmetros, treinamento, validação e testes de cada um dos modelos, bem como a análise de custo computacional a ser discutida adiante – foram adotadas as seguintes configurações:

- Para o *Encoder-Decoder* (LSTM) e o *Encoder-Decoder* (GRU) foi utilizado o *cluster* Neumann II, que é provido pelo Centro de Tecnologias Estratégicas do Nordeste (CETENE)[7]. Por meio dele, cada modelo teve à disposição 4 vCPUs, 32 GB de memória e 2 GPUs do tipo GeForce GTX 980 Ti (com 6 GB de memória cada).

- Para o *Transformer*, no entanto, não foi possível utilizar o Neumann II uma vez que esse modelo necessitou de recursos de GPU e de tempo total de processamento que excediam as restrições estabelecidas pelo *cluster*. Por conta disso, foram adotadas 3 máquinas virtuais do tipo *Standard* NV6 providas pela Azure[8]. Cada uma dessas máquinas disponibilizou 6 vCPUs, 56 GB de memória e 3 GPUs do tipo Tesla M60 (com 8 GB de memória cada).

O código-fonte utilizado nos experimentos deste trabalho foi disponibilizado através do endereço indicado abaixo[9].

[7] O CETENE é uma unidade de pesquisa do Ministério da Ciência, Tecnologia e Inovações do Brasil que objetiva desenvolver, introduzir e aperfeiçoar inovações tecnológicas estratégicas para o desenvolvimento do Nordeste. Mais detalhes em <http://antigo.cetene.gov.br/cluster>.

[8] A Azure é uma plataforma de computação em nuvem composta por mais de 200 produtos e serviços providos pela Microsoft. Mais detalhes em <https://azure.microsoft.com/>

[9] Disponível em <https://www.cin.ufpe.br/~cca5/sl-nlp>.

4 AVALIAÇÃO EXPERIMENTAL

Nos capítulos anteriores, foi introduzida a abordagem proposta e as técnicas utilizadas na preparação e execução dos experimentos deste trabalho. Neste capítulo, portanto, serão analisados os resultados apresentados por esses experimentos. Isso será realizado em duas partes: primeiro, serão discutidos os resultados coletadas para cada um dos modelos utilizados; em seguida, serão comparados esses resultados com aqueles obtidos por outras pesquisas em Reconhecimento de Língua de Sinais (RLS) que também adotaram o *American Sign Language Lexicon Video Dataset* (ASLLVD) como base de análise, mas que seguiram por abordagens distintas.

4.1 ANÁLISE DOS RESULTADOS

A Tabela 9 apresenta o desempenho dos modelos utilizados nos experimentos deste trabalho, respectivamente o *Encoder-Decoder* implementado com LSTM, o *Encoder-Decoder* implementado com GRU e o *Transformer*. Para cada modelo, são listadas as métricas[1] de acurácia, precisão, *recall* e *F1-score*, bem como o erro médio calculado.

Primeiramente, observa-se que os modelos baseados na arquitetura *Encoder-Decoder* apresentaram resultados muito semelhantes entre si, apesar de ainda não muito expressivos de um modo geral.

Dentre eles, percebe-se também que a versão implementada utilizando codificador e decodificador baseados em redes GRU obteve uma pequena vantagem em comparação àquela que utilizou redes LSTM – ao passo que a primeira alcançou uma acurácia de 42,78%, a segunda obteve 42,56%. O valor das demais métricas e do erro médio computado para ambos os casos replicaram esse mesmo comportamento e reforçam tal análise.

Por outro lado, quando analisados os resultados do *Transformer* observa-se um desempenho bastante expressivo com uma acurácia de 99,56%, a qual excede aquelas apresentadas pelos *Encoder-Decoders* e é reiterada consistentemente pelo valor das demais métricas e também do erro médio computado.

Isso nos remete à argumentação de Wolf et al. (2020), Jurafsky e Martin (2022) citada na

[1] Uma vez que o reconhecimento de sinais consiste numa classificação multi-classes, os valores das métricas binárias de precisão, *recall* e *F1-score* foram consolidados utilizando-se a média ponderada pelo número de amostras em cada classe.

Tabela 9 – Resultados dos modelos utilizados neste trabalho

Etapa	Métrica	**Encoder-Decoder (LSTM)**	**Encoder-Decoder (GRU)**	**Transformer**
Treinamento	Acurácia	97,49%	99,65%	100,00%
	Precisão	97,37%	99,67%	100,00%
	Recall	97,49%	99,65%	100,00%
	F1-score	97,12%	99,63%	100,00%
	Erro médio (CEL)	0,871327	0,206063	0,000405
Validação	Acurácia	46,30%	48,57%	99,90%
	Precisão	39,83%	41,92%	99,85%
	Recall	46,30%	48,57%	99,90%
	F1-score	41,51%	43,74%	99,87%
	Erro médio (CEL)	4,673247	4,229719	0,008175
Testes	Acurácia	42,56%	42,78%	99,56%
	Precisão	45,26%	46,95%	99,53%
	Recall	42,56%	42,78%	99,56%
	F1-score	41,47%	42,48%	99,54%
	Erro médio (CEL)	5,353844	4,783055	0,036416

Fonte: Elaborada pelo autor.

subseção 2.3.1, que afirma que essa arquitetura tem se mostrado extremamente bem-sucedida entre tarefas de Processamento de Linguagem Natural (PLN), superando arquiteturas como as RNNs. Exatamente por isso, elas são atualmente dominantes nessa área.

No contexto deste trabalho, acredita-se que alguns motivos particulares contribuíram para os resultados acima. Em primeiro lugar, entende-se que a escolha pelo uso da abordagem linguística dos sinais nos permitiu trabalhar num nível semântico muito mais elevado do que seria possível fazer com dados brutos como os pixels de imagens RGB ou coordenadas aleatórias no espaço. Conforme discutido no Capítulo 3, essa é uma abordagem comum em tarefas envolvendo linguagem no PLN e que permite produzir *features* de melhor qualidade para ensinar os modelos acerca da estrutura dessas línguas.

Em segundo lugar, a representação introduzida aqui foi capaz de transformar um conjunto de canais linguísticos complexos dos sinais do ASLLVD, em *features* discretas (ou "palavras", como aqui as denominamos) mais simples e bem definidas a serem consumidas pelos modelos. Isso deu origem a um vocabulário que, por sua vez, é muito menos complexo do que aqueles com os quais arquiteturas como o *Transformer* foram originalmente projetadas para lidar. Além disso, nesse processo foi selecionado um número menor de atributos fonológicos da língua de sinais e isso também contribuiu para tornar o vocabulário utilizado aqui ainda mais compacto.

Dessa forma, entende-se que a combinação desse conjunto de *features* semanticamente

mais coerentes e simplificadas, com a utilização de modelos robustos no processamento de linguagens como o *Transformer* foram fatores que conduziram ao desempenho favorável acima. Contudo, para os *Encoder-Decoders* esse desempenho foi mais modesto.

A Tabela 10 apresenta o custo computacional aferido para os modelos durante a etapa de testes que, por sua vez, contempla a inferência dos sinais para as 1.821 amostras que compõem o respectivo subconjunto de testes. Na tabela, são listadas as métricas de tempo de processamento (em segundos) e memória (em Megabytes (MBs)) consumidas em ambas CPU e GPU, além do número de *Floating-Point Operations Per Second* (ou Operações de Ponto Flutuante por Segundo) (FLOPS) (em MFLOPS, que corresponde a 1×10^6 FLOPS).

Tabela 10 – Custo computacional calculado para os modelos utilizados neste trabalho durante a etapa de testes

Etapa	Métrica		**Encoder-Decoder (LSTM)**	**Encoder-Decoder (GRU)**	**Transformer**
Testes	Tempo	CPU	1,609s	2,032s	1,883s
		GPU	0,249s	0,478s	0,687s
	Memória	CPU	50,326 MB	50,327 MB	50,615 MB
		GPU	2.163,952 MB	3.605,349 MB	7.088,787 MB
	Operações		10.758 MFLOPS	36.791 MFLOPS	656.413 MFLOPS

Fonte: Elaborada pelo autor.

Ao analisar a métrica de tempo de processamento em CPU, percebe-se que o *Encoder-Decoder* (GRU) foi o modelo com o maior consumo (2,032 segundos), seguido pelo *Transformer* (1,883 segundos) e pelo *Encoder-Decoder* (LSTM) (1,609 segundos). No entanto, quando considerado o processamento em GPU, o *Transformer* passa a figurar como aquele com o maior tempo (0,687 segundos), seguido pelo *Encoder-Decoder* (GRU) (0,478 segundos) e pelo *Encoder-Decoder* (LSTM) (0,249 segundos).

Com relação ao consumo de memória de CPU, observa-se que os três modelos apresentaram números muito semelhantes entre si (em torno de 50 MB) – sendo o consumo do *Transformer* de 50,615 MB, o do *Encoder-Decoder* (GRU) de 50,327 MB e o do *Encoder-Decoder* (LSTM) de 50,326 MB. Em contrapartida, ao observar a memória de GPU constata-se uma diferença mais significativa, onde o *Transformer* apresenta-se com o consumo mais elevado (de aproximadamente 7,089 GB) e é seguido pelo *Encoder-Decoder* (GRU) (com aproximadamente 3,605 GB) e pelo *Encoder-Decoder* (LSTM) (com cerca de 2,164 GB).

Finalmente, ao analisar o número de operações executadas, observa-se que o *Transformer* realizou uma quantidade significativamente maior que os demais modelos (aproximadamente

656.413 MFLOPS). Na sequência, aparecem o *Encoder-Decoder* (GRU) com 36.791 MFLOPS e o *Encoder-Decoder* (LSTM) (com 10.758 MFLOPS). Essa diferença ilustra de maneira mais concreta a discussão acima que ressalta que a arquitetura *Transformer* apresenta um grau elevado de complexidade e robustez, sobretudo quando comparado às demais arquiteturas *Encoder-Decoders* utilizadas neste trabalho.

4.2 COMPARAÇÃO DOS RESULTADOS

Para estabelecer um comparativo entre os resultados deste trabalho e outras pesquisas dentro da área de Reconhecimento de Língua de Sinais (RLS), foram selecionados estudos que também utilizaram o ASLLVD como referência em seus experimentos. Esses estudos são enumerados a seguir:

- **Theodorakis, Pitsikalis e Maragos (2014)**: utiliza técnicas não-supervisionadas e também *Hidden Markov Model* (ou Modelo Oculto de Markov) (HMM) para gerar subunidades de movimento e pausa (chamadas 2-S-U) que são utilizadas para reconhecer os sinais. Para isso, os autores processam os *frames* das amostras concentrando-se apenas nas mãos dos indivíduos e selecionam um subconjunto de 97 sinais do ASLLVD.

- **Lim, Tan e Tan (2016)**: introduz a técnica de *Block-based Histogram of Optical Flow* (ou Histograma Baseado em Blocos de Fluxo Óptico) (BHOF), que gera histogramas do fluxo óptico das mãos dos indivíduos a partir dos *frames* das amostras. Em seus experimentos, os autores selecionam um subconjunto de apenas 20 sinais do ASLLVD.

  Além disso, eles também comparam os resultados do BHOF com aqueles obtidos pelas técnicas *Motion Energy Image* (ou Imagem de Energia de Movimento) (MEI), *Motion History Image* (ou Imagem do Histórico de Movimento) (MHI), *Principal Component Analysis* (ou Análise de Componente Principal) (PCA) e *Histogram of Optical Flow* (ou Histograma de Fluxo Óptico) (HOF) para o mesmo subconjunto de sinais.

- **Metaxas, Dilsizian e Neidle (2018)**: combina uma variedade de técnicas com o intuito de produzir diferentes *features* referentes a configuração de mão inicial e final; número de mãos envolvidas; distância entre mãos; coordenadas 3D do corpo, face, mãos e braços; e contato das mãos com o corpo. Essas *features* são utilizadas como entrada para um modelo baseado em *Hidden Conditional Ordinal Random Fields* (ou Campos

Aleatórios Ordinais Condicionais Ocultos) (HCORF) para reconhecer um subconjunto selecionado de 350 sinais do ASLLVD.

- **Lim et al. (2019)**: introduz uma representação chamada *Hand Energy Image* (ou Imagem de Energia da Mão) (HEI) que também concentra-se nas mãos dos indivíduos e é utilizada como entrada para uma rede *Convolutional Neural Network* (ou Rede Neural Convolucional) (CNN). Os autores adotam o mesmo subconjunto de 20 sinais utilizados por Lim, Tan e Tan (2016) acima.

- **Amorim, Macêdo e Zanchettin (2019)**: utiliza grafos para modelar a dimensão espacial das coordenadas 2D do corpo dos indivíduos bem como a relação temporal dos seus movimentos, os quais são fornecidos como entrada para uma rede *Spatial-Temporal Graph Convolutional Network* (ou Rede Convolucional de Grafos Espaço-Temporais) (ST-GCN). Os autores avaliam os resultados para o mesmo subconjunto de 20 sinais definidos por Lim, Tan e Tan (2016), mas também para o ASLLVD inteiro.

Como pode-se perceber pela introdução acima, a maioria desses estudos aborda o RLS através do processamento de dados brutos – como *frames* RGB ou coordenadas 2D ou 3D –, conforme discutido ao longo da seção 2.4. Apesar de em alguns casos eles serem utilizados para gerar *features* intermediárias – como subunidades de movimento e pausa, imagens de fluxo óptico ou de energia de movimento, histogramas, grafos, entre outras – estas, por sua vez, apresentam ainda um nível semântico muito menos informativo quanto à língua de sinais e à sua linguística do que aquelas que introduzimos neste trabalho.

Além disso, parte desses trabalhos concentra-se apenas nas mãos dos indivíduos e isso nos remete a um dos problemas discutidos na subseção 2.4.2, que se refere à abordagem inadequada no reconhecimento da língua de sinais. Como resultado, traços linguísticos importantes – como expressões não-manuais, locação de mãos, movimentos do corpo e interações entre mãos e corpo – acabam sendo desconsiderados.

Observa-se também que todos esses trabalhos modelam vocabulários que correspondem a subconjuntos muito pequenos do ASLLVD, conforme discutido na subseção 2.4.1. Esses subconjuntos, por sua vez, representam menos de 13% do vocabulário total de 2.745 sinais disponibilizado por esse *dataset* e isso faz com que eles acabem não abrangendo uma amplitude significativa da língua de sinais.

Compreende-se que todos esses fatores têm por objetivo simplificar o tamanho e a complexidade do escopo abordado pelas pesquisas acima. No entanto, é inevitável ressaltar após a discussão realizada ao longo deste trabalho que recortes assim distanciam tais pesquisas do contexto real de uso da língua de sinais e, consequentemente, limitam o nível das contribuições que efetivamente agregam avanços à área de RLS.

A Tabela 11 relaciona os resultados apresentados pelos estudos introduzidos acima e os obtidos pelos modelos utilizados nos experimentos deste trabalho. De uma maneira geral, observa-se que os modelos *Encoder-Decoders* utilizados aqui posicionaram-se com um desempenho mediano em relação às demais pesquisas: ao mesmo tempo em que sua acurácia de aproximadamente 43% ultrapassou técnicas como HEI, MEI e MHI, outras como HCORF, BHOF, HOF e ST-GCN mostraram-se superiores alcançando valores de até 93,30%. Contudo, a acurácia de 99,56% registrada pelo *Transformer* posicionou-se consistentemente superior aos demais resultados da tabela.

Tabela 11 – Comparação dos resultados de nossos experimentos com outras pesquisas em RLS que também basearam-se no ASLLVD

Autor	Técnica	Nº de Sinais	Acurácia
Theodorakis, Pitsikalis e Maragos (2014)	2-S-U + HMM	97	63,15%
Lim, Tan e Tan (2016)	BHOF	20	85,00%
	HOF	20	70,00%
	PCA	20	45,00%
	MEI	20	25,00%
	MHI	20	10,00%
Metaxas, Dilsizian e Neidle (2018)	HCORF	350	93,30%
Lim et al. (2019)	HEI	20	31,50%
Amorim, Macêdo e Zanchettin (2019)	ST-GCN	20	61,04%
	ST-GCN	2.745	16,48%
Proposta atual	**Encoder-Decoder (LSTM)**	**2.650**	**42,56%**
	Encoder-Decoder (GRU)	**2.650**	**42,78%**
	Transformer	**2.650**	**99,56%**

Fonte: Elaborada pelo autor.

Quando considerados apenas os estudos que modelaram *features* referentes ao corpo inteiro do indivíduo (mesmo que indiretamente através de dados brutos como coordenadas) em vez de apenas suas mãos, tem-se uma perspectiva diferente. Metaxas, Dilsizian e Neidle (2018) e Amorim, Macêdo e Zanchettin (2019) se enquadram nesses critérios e, pela tabela, observa-se que ambas as técnicas de HCORF e ST-GCN superaram os resultados das implementações de *Encoder-Decoders* utilizadas em nossos experimentos. Enquanto aquelas primeiras registraram acurácia de 93,30% e 61,04%, respectivamente, os *Encoder-Decoders* alcançaram valores em

torno de 43%.

É importante salientar que a maioria dos estudos na tabela modelaram vocabulários muito pequenos do ASLLVD, o que difere-se do que foi realizado neste trabalho. Conforme discutido acima, esse tipo de abordagem fornece um recorte simplificado do problema e favorece para que tais estudos alcancem acurácias mais elevadas. Os números apresentados por Amorim, Macêdo e Zanchettin (2019) ajudam a ilustrar a diferença decorrente disso: ao modelar os 20 sinais utilizados pela maioria dos estudos na tabela, o ST-GCN obteve uma acurácia de 61,04%; no entanto, quando utilizados os 2.745 sinais, esse número caiu para 16,48%. Dessa forma, entende-se que mesmo as acurácias de 42,56% e 42,78% obtidas pelos *Encoder-Decoders* nos experimentos deste trabalho representam resultados bastante significativos em comparação com os demais estudos na tabela, uma vez que aqui abordou-se um vocabulário complexo de 2.650 sinais e os demais consideraram apenas recortes de 20, 97 ou 350 sinais.

5 CONSIDERAÇÕES FINAIS

Foi introduzida ao longo deste trabalho uma abordagem que concentra-se nos atributos linguísticos da língua de sinais para realizar o seu reconhecimento de forma automática. Apesar dessa ser uma estratégia comum entre tarefas de Processamento de Linguagem Natural (PLN), ela difere-se da abordagem que tem sido utilizado pela maioria das pesquisas em Reconhecimento de Língua de Sinais (RLS) que, por sua vez, concentra-se principalmente no processamento de dados brutos como pixels RGB ou coordenadas.

Dessa maneira, o foco foi trocado do domínio da Visão Computacional (VC) para o domínio do PLN, permitindo que os modelos aplicados aprendessem acerca da estrutura linguística em vez das relações entre pixels e coordenadas, por exemplo. Pelos resultados apresentados no Capítulo 4, percebe-se que essa abordagem apresentou uma eficácia bastante satisfatória sobretudo quando combinada com o modelo *Transformer* que, por sua vez, é uma arquitetura amplamente utilizada e bem-sucedida ao lidar com tarefas envolvendo linguagens.

Entre as principais contribuições do trabalho, pode-se enumerar que ele:

- Aborda a língua de sinais efetivamente como uma língua, ao invés de considerá-la como um conjunto de gestos ou posições estáticas de mãos, conforme também observa-se em muitas pesquisas em RLS. Isso posiciona a linguística como o pilar fundamental para essa área e abrange suas complexidades.

- Conscientiza o leitor acerca da linguística e das particularidades da natureza visual das línguas de sinais. Isso permite que pesquisadores apropriem-se desses temas e motivem-se a desenvolver novas pesquisas que contribuam com avanços seguindo essa mesma direção.

- Traz clareza acerca das principais lacunas que têm limitado progressos expressivos na área de RLS ao longo das últimas décadas, os quais poderiam viabilizar soluções efetivamente aplicáveis ao contexto real da língua de sinais e do Surdo.

- Contribui com um novo *dataset* de atributos linguísticos para a língua de sinais, que viabiliza com que novas pesquisas em RLS sejam desenvolvidas seguindo a mesma abordagem que introduzimos aqui. Além disso, este trabalho também contribui com um *dataset* de coordenadas 3D que permite pesquisadores derivar outros tipos de represen-

tações linguísticas. Ambos baseiam-se no ASLLVD, que é um importante *dataset* da ASL.

- Introduz uma estratégia para computar atributos linguísticos a partir de dados brutos, como *frames* de vídeos ou coordenadas, a qual foi aplicada na geração do *dataset* acima. Nela, as complexidades relacionadas à Visão Computacional são delegadas para ferramentas que implementam seus respectivos estados da arte. Como consequência, essa estratégia é compatível com diferentes *datasets* que, por sua vez, poderiam ser processados e combinados para produzir volumes ainda maiores de dados, que são atualmente escassos para as línguas de sinais.

- Contribui com dois artigos para a área de RLS: em Amorim, Macêdo e Zanchettin (2019), além do ST-GCN, é introduzido o processamento do ASLLVD utilizado como base para a obtenção dos *datasets* apresentados neste trabalho; em Amorim e Zanchettin (2022) (que foi submetido e encontra-se em revisão), são discutidos os detalhes da criação desses *datasets*.

Por outro lado, enxergam-se como passos futuros para a abordagem deste trabalho os seguintes itens:

- Avaliar a eficácia dessa abordagem para o cenário de sinais contínuos que, por sua vez, é mais complexo por não haver uma segmentação clara entre os sinais – como no nosso contexto. Apesar disso, sinais contínuos são muito mais próximos do contexto real de uso da língua de sinais e, por conta disso, possuem uma grande sinergia com a linha de pesquisa iniciada aqui.

- Explorar outros atributos linguísticos que atualmente não considerados no subconjunto incipiente adotado, para ampliar as *features* do *dataset* introduzido aqui. Isso contribuirá para prover mais contexto sobre a estrutura linguística dos sinais aos modelos, os quais precisarão tornar-se gradualmente mais robustos à medida em que são aprofundados cenários mais reais de uso da língua – como no exemplo dos sinais contínuos.

- Produzir um corpus de sinais mais amplo e diversificado para a língua de sinais, utilizando a estratégia de anotação de atributos linguísticos introduzida aqui. Conforme discutido anteriormente, os *datasets* atualmente disponíveis para as línguas de sinais

são pequenos e pouco diversificados em comparação àqueles utilizados pelo Processamento de Linguagem Natural (PLN) para alcançar o estado da arte em outros tipos de linguagens. Além disso, técnicas modernas e mais robustas de Aprendizagem Profunda são orientadas a dados e demandam vocabulários cada vez maiores.

- Por fim, avaliar o desempenho da abordagem deste trabalho no cenário online (ou em tempo real) de reconhecimento de sinais. Uma vez que é um objetivo de longo prazo contribuir para que a área de RLS consiga prover ferramentas para contextos reais da língua, é importante conhecer sua eficácia também para cenários online afim de assegurar que a direção percorrida pela linha de pesquisa permanece alinhada com esse propósito conforme evolui.

REFERÊNCIAS

Agência Brasil. *País tem 10,7 milhões de pessoas com deficiência auditiva, diz estudo*. 2019. <https://agenciabrasil.ebc.com.br/geral/noticia/2019-10/brasil-tem-107-milhoes-de-deficientes-auditivos-diz-estudo>. Acessado: 2021-12-22.

Agência Brasil. *OMS estima 2,5 bilhões de pessoas com problemas auditivos em 2050*. 2021. <https://agenciabrasil.ebc.com.br/saude/noticia/2021-03/oms-estima-25-bilhoes-de-pessoas-com-problemas-auditivos-em-2050>. Acessado: 2021-12-30.

Agência Senado. *Baixo alcance da língua de sinais leva surdos ao isolamento*. 2019. <https://www12.senado.leg.br/noticias/especiais/especial-cidadania/baixo-alcance-da-lingua-de-sinais-leva-surdos-ao-isolamento>. Acessado: 2022-01-07.

AMORIM, C. C. de; MACÊDO, D.; ZANCHETTIN, C. Spatial-temporal graph convolutional networks for sign language recognition. In: TETKO, I. V.; KŮRKOVÁ, V.; KARPOV, P.; THEIS, F. (Ed.). *ICANN: International Conference on Artificial Neural Networks*: Artificial neural networks and machine learning - ICANN 2019: Workshop and special sessions. [S.l.]: Springer, Cham, 2019. v. 11731, p. 646–657. ISBN 978-3-030-30493-5.

AMORIM, C. C. de; ZANCHETTIN, C. ASL-Skeleton3D and ASL-Phono: Two novel datasets for the american sign language. arXiv, 2022. Disponível em: <https://arxiv.org/abs/2201.02065>.

ANTON, H.; RORRES, C. *Elementary linear algebra: applications version*. 9. ed. [S.l.]: John Wiley & Sons, 2005. ISBN 0471669598.

ATHITSOS, V.; NEIDLE, C.; SCLAROFF, S.; NASH, J.; STEFAN, A.; YUAN, Q.; THANGALI, A. The american sign language lexicon video dataset. In: *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. [S.l.: s.n.], 2008. p. 1–8. ISSN 2160-7508.

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural machine translation by jointly learning to align and translate. In: BENGIO, Y.; LECUN, Y. (Ed.). *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. [s.n.], 2015. Disponível em: <http://arxiv.org/abs/1409.0473>.

BAKER, C.; PADDEN, C. Focusing on the nonmanual components of american sign language. *Understanding Language through Sign Language Research*, Academic Press, Nova York, NY, p. 27–57, 1978.

BATTISON, R. Phonological deletion in american sign language. *Sign Language Studies*, Gallaudet University Press, v. 5, p. 1–19, out. 1974.

BAZAREVSKY, V.; GRISHCHENKO, I.; RAVEENDRAN, K.; ZHU, T.; ZHANG, F.; GRUNDMANN, M. Blazepose: On-device real-time body pose tracking. In: . [S.l.: s.n.], 2020.

BAZAREVSKY, V.; KARTYNNIK, Y.; VAKUNOV, A.; RAVEENDRAN, K.; GRUNDMANN, M. Blazeface: Sub-millisecond neural face detection on mobile gpus. In: *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*. [S.l.: s.n.], 2019.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2006. ISBN 978-0387-31073-2.

BRAGG, D.; KOLLER, O.; BELLARD, M.; BERKE, L.; BOUDREAULT, P.; BRAFFORT, A.; CASELLI, N.; HUENERFAUTH, M.; KACORRI, H.; VERHOEF, T.; VOGLER, C.; MORRIS, M. R. Sign language recognition, generation, and translation: An interdisciplinary perspective. In: *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. New York, NY, USA: Association for Computing Machinery, 2019. (ASSETS '19), p. 16–31. ISBN 978-1450366762. Disponível em: <https://doi.org/10.1145/3308561.3353774>.

BRASIL. Lei nº 10.436, de 24 de abril de 2002. *Diário Oficial da República Federativa do Brasil*, Brasília, DF, 2002. Acessado: 2022-01-07. Disponível em: <https://www.planalto.gov.br/ccivil_03/leis/2002/l10436.htm>.

CAO, Z.; SIMON, T.; WEI, S.-E.; SHEIKH, Y. Realtime multi-person 2d pose estimation using part affinity fields. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2017. p. 1302–1310.

CHO, K.; MERRIËNBOER, B. van; BAHDANAU, D.; BENGIO, Y. On the properties of neural machine translation: Encoder-decoder approaches. In: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Doha, Qatar: Association for Computational Linguistics, 2014. p. 103–111. Disponível em: <https://aclanthology.org/W14-4012>.

CHO, K.; MERRIËNBOER, B. van; GULCEHRE, C.; BAHDANAU, D.; BOUGARES, F.; SCHWENK, H.; BENGIO, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1724–1734. Disponível em: <https://aclanthology.org/D14-1179>.

COOPER, H.; HOLT, B.; BOWDEN, R. Sign language recognition. In: MOESLUND, T. B.; HILTON, A.; KRÜGER, V.; SIGAL, L. (Ed.). *Visual Analysis of Humans*. [S.l.]: Springer, London, 2011. p. 539–562.

FERRARIO, V. F.; SFORZA, C.; SCHMITZ, J. H.; CIUSA, V.; COLOMBO, A. Normal growth and development of the lips: a 3-dimensional study from 6 years to adulthood using a geometric model. *Journal of Anatomy*, Wiley Online Library, v. 196, n. 3, p. 415–423, 2000.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>. ISBN 978-0262035613.

GRAVES, A. *Supervised Sequence Labelling with Recurrent Neural Networks*. 1. ed. [S.l.]: Springer Berlin, Heidelberg, 2012. 146 p. (Studies in Computational Intelligence). ISSN 1860-949X. ISBN 978-3-642-24797-2.

HE, H.; MA, Y. (Ed.). *Imbalanced Learning: Foundations, Algorithms, and Applications*. 1. ed. [S.l.]: Wiley - IEEE Press, 2013. ISBN 978-1118074626.

HILL, J. C.; LILLO-MARTIN, D. C.; WOOD, S. K. *Sign Languages: Structures and Contexts*. New York, NY: Routledge, 2019. ISBN 978-0-429-02087-2.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural Computation*, v. 9, n. 8, p. 1735–1780, 1997.

IBGE Instituto Brasileiro de Geografia e Estatística. *Pesquisa Nacional de Saúde 2019: Ciclos de vida*. Rio de Janeiro, RJ: Instituto Brasileiro de Geografia e Estatística - IBGE, 2021. ISBN 978-65-87201-76-4.

IBGE Instituto Brasileiro de Geografia e Estatística. *Projeção da população do Brasil e das Unidades da Federação*. 2021. <https://www.ibge.gov.br/apps/populacao/projecao/index.html>. Acessado: 2021-12-27.

JAY, M. *Don't Just Sign... Communicate!: A Student's Guide to Mastering American Sign Language Grammar*. Los Angeles, CA: Routledge, 2011. ISBN 978-0-9845294-4-5.

JOKSIMOSKI, B.; ZDRAVEVSKI, E.; LAMESKI, P.; PIRES, I. M.; MELERO, F. J.; MARTINEZ, T. P.; GARCIA, N. M.; MIHAJLOV, M.; CHORBEV, I.; TRAJKOVIK, V. Technological solutions for sign language recognition: A scoping review of research trends, challenges, and opportunities. *IEEE Access*, v. 10, p. 40979–40998, 2022. ISSN 2169-3536.

JUNG, C. G. *Contributions to Analytical Psychology*. [S.l.]: K. Paul, Trench, Trubner & Co. Ltd. (New York) and Harcourt, Brace and Company (London), 1928.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3. ed. Stanford University, 2022. Draft. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/>.

KITAGAWA, M.; WINDSOR, B. *MoCap for Artists: Workflow and Techniques for Motion Capture*. New York, NY: Routledge, 2017. ISBN 978-0-240-81000-3.

KLIMA, E. S.; BELLUGI, U. Wit and poetry in american sign language. *Sign Language Studies*, Gallaudet University Press, v. 8, p. 203–223, sep 1975.

KLIMA, E. S.; BELLUGI, U. *The Signs of Language*. [S.l.]: Harvard University Press, 1979. 417 p. ISBN 978-06-748-0796-9.

KOLLER, O. *Quantitative Survey of the State of the Art in Sign Language Recognition*. arXiv, 2020. Disponível em: <https://arxiv.org/abs/2008.09918>.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, v. 521, n. 7553, p. 436–444, maio 2015.

LIM, K. M.; TAN, A. W.; TAN, S. C. Block-based histogram of optical flow for isolated sign language recognition. *J. Vis. Comun. Image Represent.*, Academic Press, Inc., USA, v. 40, n. PB, p. 538–545, oct 2016. ISSN 1047-3203.

LIM, K. M.; TAN, A. W. C.; LEE, C. P.; TAN, S. C. Isolated sign language recognition using convolutional neural network hand modelling and hand energy image. *Multimedia Tools and Applications*, v. 78, n. 14, p. 19917–19944, jul 2019.

LUGARESI, C.; TANG, J.; NASH, H.; MCCLANAHAN, C.; UBOWEJA, E.; HAYS, M.; ZHANG, F.; CHANG, C.-L.; YONG, M.; LEE, J.; CHANG, W.-T.; HUA, W.; GEORG, M.; GRUNDMANN, M. Mediapipe: A framework for perceiving and processing reality. In: *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*. [s.n.], 2019. Disponível em: <https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf>.

METAXAS, D.; DILSIZIAN, M.; NEIDLE, C. Linguistically-driven framework for computationally efficient and scalable sign recognition. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018.

Ministério da Educação e Cultura. *Conheça o INES*. 2021. <https://www.gov.br/ines/pt-br/acesso-a-informacao-1/institucional/conheca-o-ines>. Acessado: 2022-05-30.

MITCHELL, T. M. *Machine Learning*. [S.l.]: McGraw-Hill, 1997. ISBN 0070428077.

MURPHY, K. P. *Machine Learning: A Probabilistic Perspective*. London, UK: The MIT Press, 2012. ISBN 978-0-262-01802-9.

NEIDLE, C.; OPOKU, A. *A User's Guide to the American Sign Language Linguistic Research Project (ASLLRP) Data Access Interface (DAI) 2 — Version 2*. 2020. Disponível em: <http://www.bu.edu/asllrp>.

NEIDLE, C.; THANGALI, A.; SCLAROFF, S. Challenges in development of the american sign language lexicon video dataset (ASLLVD) corpus. In: *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC 2012, Istanbul, Turkey*. [s.n.], 2012. Disponível em: <https://open.bu.edu/handle/2144/31899>.

OPAS Organização Pan-Americana da Saúde. *OMS estima que 1 em cada 4 pessoas terão problemas auditivos até 2050*. 2021. <https://www.paho.org/pt/noticias/2-3-2021-oms-estima-que-1-em-cada-4-pessoas-terao-problemas-auditivos-ate-2050>. Acessado: 2021-12-22.

PAPASTRATIS, I.; CHATZIKONSTANTINOU, C.; KONSTANTINIDIS, D.; DIMITROPOULOS, K.; DARAS, P. Artificial intelligence technologies for sign language. *Sensors*, v. 21, n. 17, 2021. ISSN 1424-8220. Disponível em: <https://www.mdpi.com/1424-8220/21/17/5843>.

PEREIRA, M. C. da C.; CHOI, D.; VIEIRA, M. I.; GASPAR, P.; NAKASATO, R. *Libras: Conhecimento Além dos Sinais*. 1. ed. São Paulo, SP: Pearson Universidades, 2011. ISBN 978-85-7605-878-6.

QUADROS, R. M. de; KARNOPP, L. B. *Língua de Sinais Brasileira: Estudos Linguísticos*. [S.l.]: Artmed Editora, 2004. ISBN 978-85-363-1174-6.

QUIZA, R.; LÓPEZ-ARMAS, O.; DAVIM, J. P. Finite element in manufacturing processes. In: _____. *Hybrid Modeling and Optimization of Manufacturing: Combining Artificial Intelligence and Finite Element Method*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 13–37. ISBN 978-3-642-28085-6.

RASTGOO, R.; KIANI, K.; ESCALERA, S. Sign language recognition: A deep survey. *Expert Systems with Applications*, v. 164, p. 113794, 2021. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S095741742030614X>.

RAVANELLI, M.; BRAKEL, P.; OMOLOGO, M.; BENGIO, Y. Light gated recurrent units for speech recognition. *IEEE Transactions on Emerging Topics in Computational Intelligence*, v. 2, n. 2, p. 92–102, 2018.

ROBBINS, H. E. A stochastic approximation method. *Annals of Mathematical Statistics*, v. 22, p. 400–407, 2007.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. *Nature*, v. 323, n. 6088, p. 533–536, oct 1986.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 3. ed. [S.l.]: Prentice Hall, 2010. ISBN 978-0-13-604259-4.

SELVARAJ, P.; NC, G.; KUMAR, P.; KHAPRA, M. OpenHands: Making sign language recognition accessible with pose-based pretrained models across languages. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022. p. 2114–2133. Disponível em: <https://aclanthology.org/2022.acl-long.150>.

SHOUP, J. E. Phonological aspects of speech recognition. In: LEA, W. A. (Ed.). *Trends in Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1980. p. 125–138.

SIMON, T.; JOO, H.; MATTHEWS, I.; SHEIKH, Y. Hand keypoint detection in single images using multiview bootstrapping. In: *CVPR*. [S.l.: s.n.], 2017.

STEWART, D. A.; STEWART, J. *Barron's American Sign Language: A Comprehensive Guide to ASL 1 and 2*. New York, NY: Barrons Educational Services, 2021. ISBN 978-1-5062-6827-9.

STOKOE, W. C. *Sign Language Structure*. Silver Springs, MD: Linstok Press, 1960.

STOKOE, W. C.; CASTERLINE, D. C.; CRONEBERG, C. G. *Dictionary of American Sign Language on Linguistic Principles*. Washington, DC: Gallaudet College Press, 1965. ISBN 978-09-321-3000-6.

STOUDT, H. W.; DAMON, A.; MCFARLAND, R. A. Skinfolds, body girths, biacromial diameter, and selected anthropometric indices of adults, united states, 1960-1962. *Vital and Health Statistics*, National Center for Health Statistics, v. 35, n. 1000, fev. 1970.

STROBEL, K. *As imagens do outro sobre a cultura surda*. 4. ed. [S.l.: s.n.], 2016. 146 p. ISBN 978-8532807786.

SUHARJITO; WIRYANA, F.; KUSUMA, G. P.; ZAHRA, A. Feature extraction methods in sign language recognition system: A literature review. In: *2018 Indonesian Association for Pattern Recognition International Conference (INAPR)*. [S.l.: s.n.], 2018. p. 11–15.

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. Sequence to sequence learning with neural networks. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. Cambridge, MA: MIT Press, 2014. (NIPS'14), p. 3104–3112.

SZELISKI, R. *Computer Vision: Algorithms and Applications*. 2. ed. [S.l.]: Springer, 2022. ISBN 978-3-030-34371-2.

THEODORAKIS, S.; PITSIKALIS, V.; MARAGOS, P. Dynamic–static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition. *Image and Vision Computing*, v. 32, n. 8, p. 533–549, 2014.

VAKUNOV, A.; CHANG, C.-L.; ZHANG, F.; SUNG, G.; GRUNDMANN, M.; BAZAREVSKY, V. Mediapipe hands: On-device real-time hand tracking. In: . [S.l.: s.n.], 2020. Https://mixedreality.cs.cornell.edu/workshop.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention is all you need. In: . Red Hook, NY: Curran Associates Inc., 2017. (NIPS'17), p. 6000–6010. ISBN 9781510860964.

WADHAWAN, A.; KUMAR, P. Sign language recognition systems: A decade systematic literature review. *Archives of Computational Methods in Engineering*, Springer, v. 28, p. 785–813, 2019.

WEI, S.-E.; RAMAKRISHNA, V.; KANADE, T.; SHEIKH, Y. Convolutional pose machines. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. p. 4724–4732.

WOLF, T.; DEBUT, L.; SANH, V.; CHAUMOND, J.; DELANGUE, C.; MOI, A.; CISTAC, P.; RAULT, T.; LOUF, R.; FUNTOWICZ, M.; DAVISON, J.; SHLEIFER, S.; PLATEN, P. von; MA, C.; JERNITE, Y.; PLU, J.; XU, C.; SCAO, T. L.; GUGGER, S.; DRAME, M.; LHOEST, Q.; RUSH, A. Transformers: State-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, 2020. p. 38–45. Disponível em: <https://aclanthology.org/2020.emnlp-demos.6>.

World Health Organization. *World Report on Hearing*. Geneva, CH, 2021. Disponível em: <https://www.who.int/publications/i/item/world-report-on-hearing>.

YIN, K.; MORYOSSEF, A.; HOCHGESANG, J.; GOLDBERG, Y.; ALIKHANI, M. Including signed languages in natural language processing. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, 2021. p. 7347–7360. Disponível em: <https://aclanthology.org/2021.acl-long.570>.