UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ – CAMPUS DE CASCAVEL
CENTRO DE EDUCAÇÃO, COMUNICAÇÃO E ARTES
CURSO DE PÓS-GRADUAÇÃO *STRICTO SENSU* EM LETRAS – NÍVEL DE MESTRADO E DOUTORADO
ÁREA DE CONCENTRAÇÃO EM LINGUAGEM E SOCIEDADE

KELI PEREIRA MALAQUIAS

# DIRETRIZES PARA A CONSTRUÇÃO DE *CORPORA* PARALELOS LIBRAS-PORTUGUÊS: UMA PROPOSTA

CASCAVEL – PR
2022

UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ – CAMPUS DE CASCAVEL
CENTRO DE EDUCAÇÃO, COMUNICAÇÃO E ARTES
CURSO DE PÓS-GRADUAÇÃO *STRICTO SENSU* EM LETRAS – NÍVEL DE MESTRADO E DOUTORADO
ÁREA DE CONCENTRAÇÃO EM LINGUAGEM E SOCIEDADE

KELI PEREIRA MALAQUIAS

# DIRETRIZES PARA A CONSTRUÇÃO DE *CORPORA* PARALELOS LIBRAS-PORTUGUÊS: UMA PROPOSTA

Dissertação apresentado à Universidade Estadual do Oeste do Paraná – UNIOESTE – para obtenção do título de Mestre em Letras, junto ao Programa de Pós-Graduação *Stricto Sensu* em Letras – Nível de Mestrado e Doutorado, área de concentração em Linguagem e Sociedade.

Linha de Pesquisa: Estudos da Linguagem: Descrição dos Fenômenos Linguísticos, Culturais, Discursivos e de Diversidade.

Orientador: Prof. Dr. Jorge Bidarra

CASCAVEL – PR

2022

Ficha de identificação da obra elaborada através do Formulário de Geração Automática do Sistema de Bibliotecas da Unioeste.

MALAQUIAS, Keli Pereira
Diretrizes para a construção de corpora paralelos Libras-Português: uma proposta / Keli Pereira MALAQUIAS; orientador Jorge Bidarra. -- Cascavel, 2022.
110 p.

Dissertação (Mestrado Acadêmico Campus de Cascavel) -- Universidade Estadual do Oeste do Paraná, Centro de Educação, Programa de Pós-Graduação em Letras, 2022.

1. Libras. 2. Linguística de Corpus. 3. Corpus paralelo Libras-Português. 4. Diretrizes para corpus representativo. I. Bidarra, Jorge, orient. II. Título.

**KELI PEREIRA MALAQUIAS**

Diretrizes para a construção de *Corpora* Paralelos Libras-Português: uma proposta

Dissertação apresentada ao Programa de Pós-Graduação em Letras em cumprimento parcial aos requisitos para obtenção do título de Mestra em Letras, área de concentração Linguagem e Sociedade, linha de pesquisa Estudos da Linguagem: Descrição dos Fenômenos Linguísticos, Culturais, Discursivos e de Diversidade, APROVADO(A) pela seguinte banca examinadora:

Jorge Bidarra
Universidade Estadual do Oeste do Paraná (UNIOESTE)

Patrícia Tuxi dos Santos
Universidade de Brasília (UnB)

Tânia Aparecida Martins
Universidade Estadual do Oeste do Paraná (UNIOESTE)

Alcione Tereza Corbari
Universidade Estadual do Oeste do Paraná (UNIOESTE)

Talita Serpa
Universidade Estadual Paulista (UNESP)

Cascavel, 02 de dezembro de 2021.

Aos meus pais, que são meu alicerce, Vanderlei e Mara.
À comunidade surda.

## AGRADECIMENTOS

Agradeço, sobretudo, a Deus pelo dom da vida, pela proteção, pela força para seguir em frente e, por tantas vezes, mesmo nas adversidades, fazer me sentir amada.

Aos meus pais, por todos os ensinamentos, por serem pessoas honestas e com o coração mais lindo que já vi neste mundo. Obrigada por sempre me apoiarem, me amarem como eu sou e estarem ao meu lado em todos os momentos. Vocês são meu porto seguro. Amo vocês!

Ao meu estimado orientador, professor Jorge Bidarra, por ter aceitado caminhar comigo esse tempo, por ser meu mentor para esse tema tão relevante para a área, pelas orientações, por tanta paciência em diversos momentos, pela sabedoria que nos conduz e por compartilhar as suas experiências que sempre nos impulsionam. Obrigada por tudo professor, principalmente por acreditar em mim, quando eu já não acreditava.

Ao Programa de Pós-Graduação em Letras (PPGL), da Universidade Estadual do Oeste do Paraná (UNIOESTE), em especial à professora e coordenadora, Dantielle Assumpção Garcia, e à assistente do programa, Magaly Lindbeck Guimarães, por me auxiliarem e atenderem as minhas necessidades quando mais precisei. Obrigada pelo apoio, pela paciência, pela compreensão e pelas novas oportunidades. Vocês foram fundamentais para a conclusão desta etapa.

Agradeço, também, a todos os professores que estiveram conosco nesses anos de Pós-Graduação, que nos auxiliaram em nossos estudos, ministraram disciplinas essenciais para a nossa formação acadêmica.

A minha Banca examinadora, que colaborou, com excelência, para o aperfeiçoamento deste trabalho e para a conclusão desta etapa.

Ao meu irmão Fernando (Nando), a minha cunhada Rutinéia (Ruth) e aos meus sobrinhos amados, Emanuel e Davi, por tanto amor, por deixarem meus dias mais leves, pelas orações para que eu continuasse e não desistisse de mim. Vocês são uma família muito amada e abençoada, são exemplos de perseverança, amor, fé e muita evolução. Sou grata a Deus por ter vocês.

Ao meu irmão Eduardo (nosso Dudu ou Duds), por trocar ideias comigo, mesmo não sendo sua área, por estar aqui do meu lado, por ser sempre tão prestativo e me ajudar em tantos momentos. Amo você incondicionalmente! A minha cunhada

Fernanda, por me ouvir tantas vezes, pelas palavras de apoio e por me ajudar quando precisei. Vocês são meus melhores vizinhos, amo tê-los por perto.

A minha tia Edna Zanatta (dinda do meu coração), por estar sempre ao nosso lado, por ser esse ser iluminado e nos dar tanto amor, minha família é sua família e vice-versa, desde sempre e para sempre. Te amo, minha segunda mãe.

A minha prima, irmã, amiga, minha metade, Pamela Zanatta, que mesmo distante geograficamente, está sempre do meu lado me apoiando e me incentivando a ser uma pessoa melhor, você é uma das minhas pessoas preferidas neste mundo.

Ao ex-aluno surdo da primeira turma que interpretei no Colégio Eleodoro, amigo e irmão do coração Harrison Gerotto Adams, por ser meu apoio e me deixar tão feliz com a sua trajetória, com as suas conquistas e por ser a prova de que o meu trabalho até aqui, na área da surdez, valeu a pena. Obrigada por ser esse amigo, confidente, meu Lorão querido, que sei posso contar e confiar sempre. Amo você!

A minha prima Letícia Pereira, que caminhou lado a lado comigo, em boa parte desse período, tão delicado, de pandemia. Saiba que suas palavras me incentivaram muito e seu abraço trouxe alento em muitos momentos. Não foram tempos fáceis, mas obrigada por, de alguma forma, mostrar a importância do trabalho que realizo, por trazer tanto esclarecimento e me mostrar tantos valores sobre essa caminhada. Te amo!

Aos meus amigos de vida e para toda a vida, Tânia Martins (minha mana) e Valdenir Pinheiro, meus sobrinhos do coração, Gustavo e Anthony, por serem minha segunda família e sempre me apoiarem, estarem sempre com a casa e o coração abertos para me receber, me ajudar e serem tantas vezes meu refúgio. Vocês são especiais para mim. Obrigada por me incentivar sempre!

As minhas amigas da primeira graduação, Ariel Tavares e Laura Wolfart, por me apoiarem tanto. Obrigada por esse longo tempo de amizade e cumplicidade, que nem o tempo e nem a distância foram capazes de consumir. Vocês são o que de mais preciso o curso de Letras me trouxe. Sou fã de vocês!

A minha amiga Maristela Godoy, que me acolheu num momento delicado e que até hoje me acolhe de alguma forma, me incentiva, me fortalece, me mostra outros caminhos, outras perspectivas. Maris da minha vida, você é grande demais! Obrigada por tanto. Te amo!

A minha amiga Larissa Freire, por ser minha confidente, por estar ao meu lado, por ser esse ser de luz, que traz sorriso e compreensão sem julgamento. Obrigada por ser esse ser humano incrível e com esse coração que não cabe no peito.

A minha amiga Leidiane Reis, por compartilhar comigo e confiar tantos momentos, e pelas meninas lindas Dani e Paula, que são reflexo da bondade do coração de vocês. Ao Daniel, que sempre nos recebe de braços abertos. Amo a sua família e essa amizade. Obrigada por sempre me receberem com tanto amor.

À Rosana de Fátima Janes Constâncio, por compartilhar tantos momentos de pós-graduação, que além de ser uma colega querida, que conheci nessa fase, tornou-se uma amiga que gostaria de levar para a vida. Obrigada por ser esse Ser de Luz tão especial, por ser tão sábia e me apoiar tanto nesta trajetória. Eu desejo a você muito sucesso!

Aos meus colegas e amigos de trabalho, Clescir Gonzatto e Flávio Kottwitz Júnior, pelo apoio nesses dias turbulentos. Obrigada pela paciência, por aguentar minhas distrações, minha ansiedade, meu estresse e até mesmo a minha ausência. Agradeço muito pelo incentivo e por tornarem, com boas risadas, os meus dias mais leves.

Ao professor Flávio Kottwitz Júnior, que foi o meu primeiro professor de Libras na Associação Cascavelense de Amigos de Surdos (ACAS), que não só me apoia, mas contribuí com a comunidade surda em geral. Sou sua fã! Obrigada pela amizade, por me ensinar tanto e fazer parte da minha profissionalização como tradutora e intérprete de Libras. Você foi e é imprescindível para meu aperfeiçoamento profissional e pessoal. Tenho muito a agradecer e aprender contigo!

À comunidade surda, todo meu respeito e gratidão, obrigada por desde o início serem tão receptivos, compartilhar comigo essa língua que amo e essa cultura linda. Sempre me receberam e me incentivaram a continuar a aprender e me profissionalizar, sem vocês esse sonho não se concretizaria.

Aos professores Celso Fernando Rocha e Talita Serpa, que ministraram uma disciplina referente a *corpus* na Universidade Estadual Pauulista (UNESP), *campus* São José do Rio Preto, muito esclarecedora para o andamento desse trabalho. Foi uma experiência maravilhosa aquela semana de estudo. Obrigada!

À professora Patrícia Tuxi, que desde o primeiro momento, sem nem mesmo me conhecer, me incentivou a dar continuidade neste projeto, obrigada por me apoiar e, após esse primeiro contato, aceitar fazer parte da minha banca, ler o meu texto,

destacar a importância deste trabalho à comunidade surda e realizar tantas contribuições.

Ao professor Nadir Antônio Prichler, que dispôs do seu tempo para compartilhar alguns conhecimentos metodológicos e me auxiliou muito para o andamento deste trabalho.

À Ionara Cristina Albani, que foi muito além de uma revisora deste trabalho, foi alguém que me deu suporte nesta fase final, sem você não conseguiria ter seguido em frente. Obrigada por todo o apoio, por acreditar e ser tão positiva sobre a conclusão desta dissertação.

Enfim, meus agradecimentos a todos aqueles que, de alguma forma, estão ou estiveram ao meu lado, com palavras de incentivo e conforto. Vocês fazem parte da minha vida, dos meus dias e me fazem acreditar que a vida é um grande presente de Deus e que tudo acontece por algum propósito maior. Deus está em tudo, do início ao fim! Obrigada Pai, por cada detalhe!

*“Antes de chorar sobre os limites que possui, antes de reclamar de suas inadequações, e fadar o seu destino ao fim, aceita o desafio de pousar os olhos sobre este aparente estado de fraqueza, e ouse acreditar, que mesmo em estradas de pavimentações precárias, há sempre um destino que poderá nos levar ao local onde o sol se põe tão cheio de beleza.”* - Pe. Fábio de Melo.

MALAQUIAS, Keli Pereira. **Diretrizes para a construção de corpora paralelos Libras-Português**: uma proposta. 2022. 110f. Dissertação (Mestrado em Letras) - Programa de Pós-Graduação em Letras, Universidade Estadual do Oeste do Paraná – UNIOESTE, Cascavel, 2022.

## RESUMO

Ao considerar a ascensão das pesquisas linguísticas na Libras e a necessidade de análises dos fenômenos linguísticos que ocorrem nas línguas de sinais, esta dissertação é fruto de estudos relacionados à Linguística de *Corpus* (LC), uma vez que, constata-se um aumento expressivo de *corpora* envolvendo essas línguas, tanto no Brasil como em outras partes do mundo. Assim, as principais questões que surgiram foram: (i) Quais seriam os critérios necessários para a constituição de um *corpus* paralelo entre línguas de sinais e línguas orais?; (ii) Que tipo de conteúdo deve prevalecer nesse *corpus*, a partir das análises que se pretende realizar?; e, mais especificamente, (iii) O que precisa ser considerado num *corpus* paralelo, em que, de um lado, o que tem são os dados linguísticos das línguas de sinais e, de outro, os dados equivalentes das línguas orais, sob a ótica das teorias que fundamentam a LC? Nesse contexto, o objetivo geral desta pesquisa é apresentar e/ou sugerir, de forma sistematizada, as diretrizes necessárias para a construção de um *corpus* paralelo Libras-Português. Para tanto, foram investigados, descritos e analisados os requisitos, bem como os tipos de informações necessárias, para a constituição de um *corpus* paralelo de língua de sinais em interface com as línguas faladas, de alguns *corpora* compilados ou em compilação em nível nacional e internacional. Para as análises, levou-se em consideração não só a origem, as definições e os conceitos que são caros à área, como também os contextos históricos envolvidos e as características necessárias para que um *corpus* linguístico seja, de fato, representativo. Com o intuito de contemplar o objetivo proposto e encontrar respostas aos problemas elencados, tomou-se como base de sustentação teórica, dentre outros, os seguintes autores: Aijmer e Altenberg (2013), Aluísio e Almeida (2006), Anderman e Rogers (2008), Baker (1995), Beber Sardinha (1999; 2000; 2004), Granger e Petch-Tyson (2003), Kennedy (1998), McEnery (1996), Olohan (2004), Quadros (2006; 2008; 2011; 2016; 2017; 2018; 2019), Sanchez (2005), Sinclair (2005), Shepherd (2012), Tagnin (2018). Assim, a proposta deste trabalho se constituiu por materiais de análises, com base em investigações teóricas e em *corpora* já constituídos ou em construção. É uma pesquisa básica, de cunho qualitativo, apoiada por um trabalho de revisão bibliográfica. Ao final, apresenta-se alguns programas que podem auxiliar na construção de *corpora* paralelos e propõe-se algumas diretrizes com o intuito de contribuir com estudos que vêm sendo desenvolvidos no âmbito da LC, notadamente, no que diz respeito às análises linguísticas da Libras em interface ao português.

**PALAVRAS-CHAVE:** Libras; Linguística de *Corpus*; *Corpus* paralelo Libras-Português; Diretrizes para *corpus* representativo.

MALAQUIAS, Keli Pereira. **Guidelines for the construction of parallel Libras-Portuguese corpora:** a proposal. 2022. 110f. Dissertation (Master's in Languages) - Graduate Program in Languages, State University of Western Paraná - UNIOESTE, Cascavel, 2022.

## ABSTRACT

When considering the rise of linguistic research in Libras and the need for analysis of linguistic phenomena that occur in sign languages, this dissertation is the result of studies related to Corpus Linguistics (LC), since there is a significant increase of corpora involving these languages, both in Brazil and in other parts of the world. Thus, the main questions that arose were: (i) What would be the necessary criteria for the constitution of a parallel corpus between sign languages and oral languages?; (ii) What type of content should prevail in this corpus, based on the analyses to be performed?; and, more specifically, (iii) What needs to be considered in a parallel corpus, where, on one side, there are linguistic data from sign languages and, on the other side, the equivalent data from oral languages, from the perspective of the theories that underlie LC? In this context, the general objective of this research is to present and/or suggest, in a systematized way, the necessary guidelines for the construction of a parallel Libras-Portuguese corpus. For this purpose, the requirements were investigated, described and analyzed, as well as the types of information required for the constitution of a parallel corpus of sign language in interface with the spoken languages, from some corpora compiled or being compiled at national and international scales. For the analyses, we took into consideration not only the origin, the definitions and the concepts that the area has been referring to, but also the historical contexts involved and the characteristics necessary for a linguistic corpus to be, in fact, representative. In an attempt to meet the proposed objective and find answers to the problems raised, the following authors, among others, were used as theoretical support: Aijmer and Altenberg (2013), Aluísio and Almeida (2006), Anderman and Rogers (2008), Baker (1995), Beber Sardinha (1999; 2000; 2004), Granger and Petch-Tyson (2003), Kennedy (1998), McEnery (1996), Olohan (2004), Quadros (2006; 2008; 2011; 2016; 2017; 2018; 2019), Sanchez (2005), Sinclair (2005), Shepherd (2012), Tagnin (2018). Thus, the proposal of this work consisted of analysis materials, based on theoretical investigations and on corpora already constituted or under construction. It is a basic research, of a qualitative nature, supported by a bibliographic review work. At the end, some programs that can help in the construction of parallel corpora are presented and propose some guidelines that can contribute to studies that have been developed within the LC, notably with regard to the linguistic analysis of Libras in an interface to Portuguese.

**Keywords:** Libras; Corpus Linguistics; Parallel corpus Libras-Portuguese; Guidelines for representative corpus.

## LISTA DE FIGURAS

## LISTA DE QUADROS

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| ACAS | Associação Cascavelense de Amigos de Surdos |
| AUSLAN | Língua de Sinais Australiana |
| BDTD | Biblioteca Digital Brasileira de Teses e Dissertações |
| BSL | Língua de Sinais Britânica |
| CAPES | Coordenação de Aperfeiçoamento de Pessoal de Nível Superior |
| CNPQ | Conselho Nacional de Desenvolvimento Científico e Tecnológico |
| DGS | Língua de Sinais Alemã |
| ELAR | Endangered Languages Archive |
| EPEEM | Grupo de Estudos de Pequenas Empresas e Empreendedorismo |
| EUA | Estados Unidos da América |
| INES | Instituto Nacional de Educação de Surdos |
| IPHAN | Instituto do Patrimônio Histórico e Artístico Nacional |
| IPOL | Instituto de Investigação e Desenvolvimento em Política Linguística |
| ISF | Idiomas Sem Fronteira |
| JSL | Língua de Sinais Japonesa |
| JSP | Língua de Sinais Japonesa |
| JSPS | Sociedade Japonesa para a Promoção da Ciência |
| LC | Linguística de *Corpus* |
| Libras | Língua Brasileira de Sinais |
| NGT | Língua de Sinais Holandesa |
| PJM | Língua de Sinais Polonesa |
| PLN | Processamento de Linguagem Natural |
| PPGL | Programa de Pós-Graduação em Letras |
| SEED | Secretaria de Educação |
| SEU | *Survey of English Usage* |
| UFRGS | Universidade Federal do Rio Grande do Sul |
| UFSC | Universidade Federal de Santa Catarina |
| UNESP | Universidade Estadual Paulista |
| UNIOESTE | Universidade Estadual do Oeste do Paraná |

**SUMÁRIO**

## 1 INTRODUÇÃO

O reconhecimento da Língua Brasileira de Sinais (Libras), por meio da Lei nº 10.436, em 2002[1], regulamentada pelo Decreto nº 5.626, em 2005[2], foi um importante avanço à comunidade surda no Brasil, uma vez que vem contribuindo para a inclusão das pessoas surdas nas mais diversas esferas sociais. De acordo com Quadros (2017, p. 34), para os surdos, a Libras é "um elemento constituidor de relações entre seus pares e na produção de significados a respeito de si, de seu grupo, dos outros e dos outros grupos". Do mesmo modo, essa conquista linguística também funcionou como uma mola propulsora para o surgimento de uma série de atividades que, dia após dia, vem se mostrando fundamental para o estabelecimento da Libras em território nacional, a começar pelo seu ensino nos diferentes níveis do sistema educacional brasileiro.

Com o advento e o fortalecimento dos cursos de Letras Libras, embora não sejam os únicos responsáveis, as contribuições que deles se originam têm sido notáveis, em especial no que se refere aos estudos linguísticos. Também, com o surgimento de muitos sinais, alguns deles realmente novos, outros oriundos dos regionalismos, os trabalhos que, reconhecidamente, já vinham sendo realizados pelos lexicógrafos e por especialistas de outras áreas do conhecimento, promoveram uma mudança significativa no *modus operandi*, que até bem pouco tempo atrás era colocado em prática.

Com efeito, não só a forma de os sinais serem coletados, como também a identificação de como e em que circunstâncias cada um deles tende se manifestar, fizeram com que os estudiosos lançassem mão de recursos mais eficientes e produtivos. É nesse contexto que os pressupostos teóricos defendidos pela Linguística de *Corpus* (LC) acabam encontrando um terreno fértil propício para os estudiosos da Libras e, também, das demais línguas de sinais espalhadas pelo mundo.

---

[1] Reconhece a Libras como meio de comunicação e expressão e outros recursos de expressão a ela associados. Disponível em: http://www.planalto.gov.br/ccivil_03/Leis/2002/L10436.htm. Acesso em: 4 out. 2017.

[2] Regulamenta a Lei nº 10.436/02, que discorre sobre as ações como meios a serem aplicados a políticas linguísticas e educacionais, com vistas à disseminação da Libras nas diferentes esferas. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2005/decreto/d5626.htm. Acesso em: 23 mar. 2019.

Motivados por esse novo olhar, o que vem se constatando é um aumento significativo na quantidade de *corpora* linguísticos envolvendo as línguas de sinais, não só no Brasil, como também em outros países. Todavia, o que a nossa pesquisa já nos vem apontando é que muitos dos *corpora* disponibilizados para acesso público ainda se mostram frágeis em relação a alguns aspectos: um deles relacionado à estruturação de dados; o outro concerne ao modo como as anotações básicas dos dados[3] foram realizadas, muitas vezes em desacordo com o preconizado pela Linguagem de *Corpus* (LC) e; por último, e talvez o mais relevante, refere-se à dificuldade de acesso às informações em Libras e à sistematização dos dados, pois, embora exista um amplo material disponível de forma eletrônica, essas informações estão em diferentes plataformas.

Portanto, as perguntas que regem o nosso trabalho são: (i) Quais seriam os critérios necessários para a constituição de um *corpus* paralelo entre línguas de sinais e línguas orais?; (ii) Que tipo de conteúdo deve prevalecer nesse *corpus*, a partir das análises que se pretende realizar?; e, mais especificamente, (iii) O que precisa ser considerado num *corpus* paralelo, em que, de um lado, o que tem são dados linguísticos das línguas de sinais e, de outro, os dados equivalentes das línguas orais, sob a ótica das teorias que fundamentam a LC?

Nesse contexto, o objetivo geral desta pesquisa é propor, de forma sistematizada, as diretrizes para a construção de um *corpus* paralelo Libras-Português, a partir dos fundamentos estabelecidos pela LC. No entanto, nesse momento, não foi criado um modelo de *corpus* para atender aos nossos anseios, mas, pretendeu-se contribuir com as pesquisas linguísticas em *corpus*, quanto a uma proposta para a sistematização dos critérios necessários para a constituição dele. Com isso em mente, buscamos atender a três objetivos específicos: (i) compreender a concepção de *corpus* linguístico da antiguidade até os tempos atuais, além dos *status* de *corpus* para a Linguística e a LC, bem como os pressupostos estabelecidos pela LC para a constituição de um *corpus* linguístico; (ii) investigar em que nível estão os *corpora* linguísticos já constituídos nas línguas de sinais, no Brasil e em alguns lugares do mundo, com o intuito de identificar e analisar os critérios elementares que eles apresentam na sua composição; (iii) a partir dos itens anteriores, definir quais elementos deveriam ser considerados, para que sejamos bem-sucedidos em relação

[3] Em linhas gerais, anotações básicas incluiriam glosas que são representações textuais das línguas de sinais e traduções de vídeos em Libras.

à proposição de um conjunto de diretrizes gerais, que permita-nos construir, de forma organizada e sistematizada, um *corpus* paralelo Libras-Português, que seja representativo aos estudos linguísticos.

Poderíamos elencar um conjunto de justificativas para investir nesta pesquisa, contudo, destacamos ao menos duas: (i) considerando os interessantes resultados que vêm sendo proporcionados pela área da LC para os estudos da linguagem e, ainda, (ii) a importância da Libras, não só para as pessoas surdas, mas para os estudiosos e aprendizes da língua, pois, investir nessa frente de trabalho é tão importante quanto necessário para o entendimento de como a língua funciona, seja do ponto de vista de seu uso pelos "falantes", seja em relação ao modo como a língua de sinais funciona e se estrutura gramaticalmente. Ao partir desses dois pontos, acreditamos que investir nessa área do conhecimento pode nos fornecer subsídios importantes para, num futuro próximo, repensarmos e formalizarmos uma proposta de modelo de *corpus* linguístico não apenas útil, mas sobretudo consistente e representativo às análises linguísticas dessa língua.

Essa pesquisa é de natureza básica e de cunho qualitativo. Adotou-se, em seu desenvolvimento, uma concepção metodológica baseada em revisão bibliográfica, com a análise de dados sustentada por pressupostos teóricos defendidos pela área da LC. Com base nisso, realizamos análises por meio da descrição de alguns *corpora* já constituídos no Brasil e em outros países, com o intuito de identificar e sistematizar os critérios necessários para a constituição de um *corpus* paralelo Libras-Português.

Organizada em quatro capítulos, no segundo capítulo, que segue adiante, apresentaremos a metodologia utilizada nesta pesquisa, bem como a sua perspectiva, como foi constituído o *corpus* de análise e o que delimitamos para que os objetivos supracitados fossem alcançados.

No terceiro capítulo, resgataremos um pouco da trajetória e da evolução dos estudos das línguas, tendo como suporte os chamados *corpora* linguísticos. Para tanto, tomaremos como ponto de partida as definições de *corpus* atribuídas desde os primórdios, até chegarmos aos dias atuais, momento em que os estudos linguísticos já consideram o que a área da LC preconiza a respeito. Na sequência, discorremos sobre a importância que *corpus* alcançou para a Ciência Linguística até, mais recentemente, a LC. Posteriormente, ressaltaremos sobre a ascensão e o fortalecimento da LC para os estudos e/ou análises linguísticas e, finalmente, apresentaremos os pressupostos estabelecidos pela LC.

No quarto capítulo, trataremos, mais especificamente, do desenvolvimento de *corpus* para as línguas de sinais, como e quando os estudos nessa área se iniciaram, quais os autores tidos como referência, o que pensam e o que propõem. Após, analisaremos os critérios que foram utilizados para a constituição de alguns bancos de dados existentes no Brasil. Por fim, apresentaremos um inventário descritivo de *corpora* de línguas de sinais, não só em âmbito nacional, mas também no que diz respeito ao que se tem produzido internacionalmente, para finalmente analisar sistematicamente os critérios para a constituição desses *corpora*.

Por conseguinte, no quinto e último capítulo, retomaremos os aspectos principais que foram abordados nos capítulos anteriores e passaremos a nossa proposta, qual seja, a apresentação de um conjunto de diretrizes voltadas para o mapeamento, a organização e a estruturação de dados linguísticos que, minimamente, deveriam constituir um *corpus* paralelo composto por dados linguísticos, de um lado, extraídos da Libras e, de outro, da Língua Portuguesa.

## 2 PERCURSOS METODOLÓGICOS

A proposta desta dissertação se desenvolveu com base numa perspectiva teórica de natureza básica, pois o nosso objetivo é propor os critérios necessários para a construção de um *corpus* linguístico paralelo Libras-Português, sem que haja a necessidade de construir um projeto piloto para a sua aplicação. Caracteriza-se como de cunho qualitativo, uma vez que buscamos investigar sobre como eram utilizados os *corpora* desde a antiguidade até os tempos atuais. Para tanto, foi realizada uma ampla revisão bibliográfica, orientada pela LC.

O *corpus* deste trabalho foi constituído por materiais descritivos para a realização das análises, com base em pesquisas que identificassem os critérios necessários para a constituição de um *corpus*. Com esta investigação, buscamos estabelecer, sistematicamente, quais são os critérios elementares para a construção de um *corpus* paralelo Libras-Português, útil e representativo aos pesquisadores da área.

Para o desenvolvimento da pesquisa, adotou-se como perspectiva teórica a LC, visto que o propósito deste trabalho foi descrever e analisar a evolução e os critérios utilizados na construção de *corpora* linguísticos em diversos países e no Brasil. Assim, fez-se necessário investigar e refletir sobre o que, basicamente, deve conter em um *corpus* de língua de sinais, para que, de fato, represente a língua e a comunidade que se investiga, além de serem úteis aos pesquisadores da área.

Do ponto de vista de sua natureza, recorreu-se à pesquisa básica, que orientou e deu suporte aos propósitos desse trabalho, analisando-se a evolução e os critérios utilizados na construção de um *corpus* linguístico de diversas línguas de sinais, com respaldo nos estudos da LC, a fim de estabelecer diretrizes para a construção de um *corpus* paralelo Libras-Português. Nesse contexto, Silveira e Córdova (2009, p. 34) afirmam que esse tipo de pesquisa “objetiva gerar conhecimentos novos, úteis para o avanço da Ciência, sem aplicação prática prevista. Envolve verdades e interesses universais”.

Portanto, para alcançar os propósitos desta pesquisa, não construímos um *corpus*, mas, ao propor os critérios, pretendemos contribuir com os estudos linguísticos dos pesquisadores da área, quanto ao processo necessário para constituí-lo. Caracteriza-se, assim, em uma pesquisa qualitativa, visto que ao definir o objeto de estudo, buscamos respaldo na literatura para desenvolver nossa proposta.

Quanto à perspectiva metodológica, foi pautada na pesquisa de revisão bibliográfica e, para isso, foi realizado um inventário bibliográfico, que consistiu na consulta de base de dados da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)[4], Biblioteca Digital Brasileira de Teses e Dissertações (BDTD)[5], Universidade Estadual Paulista (UNESP)[6], Universidade Federal do Rio Grande do Sul (UFRGS)[7], a partir das palavras-chave: "Linguística de *Corpus*", "O que significa *Corpus*", "Como se constrói um *Corpus*", "*Corpus* na perspectiva linguística", "*Corpus* aplicados à língua de Sinais", "*Corpus* em Libras", dentre outras. Assim, a pesquisa foi organizada com base em livros, dissertações, teses, artigos de periódicos, plataformas digitais e outros documentos, a fim de tomar conhecimento sobre o que a literatura especializada diz sobre o tema proposto. Nesse contexto, de acordo com Gil (2002, p. 38), a revisão bibliográfica é "desenvolvida com base em material já elaborado, que traz como principal vantagem o fato de permitir uma gama de fenômenos muito mais ampla do que aquela que se obteria ao se pesquisar diretamente".

## 2.1 *CORPUS* E CONTEXTO DA PESQUISA

Para as análises dos critérios que constitui um *corpus,* buscamos investigar diversos *corpora* linguísticos compilados ou em compilação nas línguas de sinais, no Brasil e internacionalmente. Dentre eles, mais especificamente: o *corpus* da Libras; o *corpus* da Língua de Sinais Australiana; o *corpus* de Língua de Sinais Britânica; o *corpus* da Língua de Sinais Alemã; o *corpus* da Língua de Sinais Holandesa; o *corpus* da Língua de Sinais Polonesa; e o *corpus* da Língua de Sinais Japonesa.

Esses *corpora* estão organizados de forma descritiva, com intuito de realizar uma investigação minuciosa, sobre quais tipos de análises são realizadas e como eles estão organizados, para assim identificar o eles têm em comum e o que um *corpus*

[4] Disponível em: http://www.capes.gov.br/busca?searchword=linguistica%20de%20corpus&ordering=newest&searchphrase=all. Acessos entre: meses mar. e jul. 2018-2021.
[5] Disponível em: http://bdtd.ibict.br/vufind/Search/Results?lookfor=linguistica+de+corpus&type=AllFields. Acessos entre: meses mar. e jul. 2018-2021.
[6] Disponível em: https://repositorio.unesp.br/discover?rpp=10&etal=0&query=linguistica+de+corpus&group_by=none&page=2. Acessos entre: meses mar. e jul. 2018-2021.
[7] Disponível em: https://lume.ufrgs.br/discover. Acessos entre: meses mar. e jul. 2018-2021.

dever conter para ser considerado, de fato, representativo à língua que está sendo estudada. Com isso, buscamos refletir e analisar quais são as informações imprescindíveis num *corpus* paralelo para, finalmente, estabelecer uma proposta que contemple os critérios necessários para a criação de um *corpus* paralelo Libras-Português, que seja proficiente em relação ao trabalho do pesquisador e que beneficie as suas análises linguísticas.

## 2.2 TÉCNICAS OU PROCEDIMENTO DE COLETA DE DADOS

Para que os objetivos assinalados neste trabalho fossem satisfatórios, inicialmente, realizamos uma seleção de um amplo inventário bibliográfico relacionado à LC, sendo que esses dados foram extraídos de materiais referentes às línguas orais e às línguas de sinais. Esses aportes teóricos serviram de base para as análises sobre os critérios necessários para a construção de um *corpus.*

Buscamos, também, consultar vários projetos de *corpora online* compilados ou em compilação, no Brasil e internacionalmente, com o intuito de descrever e analisar a sua constituição. Assim, somente após todas as investigações e as reflexões acerca de *corpus* nas línguas de sinais, identificamos o que é importante considerar, quando se pretende construir um *corpus* paralelo de Libras em interface com o Português, que seja representativo aos estudos linguísticos.

## 3 *CORPUS* E LINGUÍSTICA DE *CORPUS*: DEFINIÇÕES E UM BREVE HISTÓRICO

Conforme a definição do dicionário eletrônico de Português[8], *corpus* é um substantivo masculino (*corpora*, no plural) descrito a partir de três sentidos: (i) como Coletânea – "reunião dos textos ou documentos sobre um assunto ou tema"; (ii) por Extensão – "repertório ou aquilo que registra toda a obra de um autor"; e (iii) na Linguística – "os registros orais que, colhidos no momento da fala, são utilizados para análise linguística".

O termo também é descrito em outras fontes lexicográficas. No dicionário eletrônico da Universidade de Cambridge[9], *corpus* é definido como um substantivo e/ou um banco de dados linguísticos, além de uma coleção de materiais escritos ou falados, armazenados em um computador, geralmente utilizados para identificar a língua em uso. Todos os dicionários, por exemplo, são retirados de um *corpus* de bilhões de palavras. O termo também é considerado como uma coleção de trabalhos de um único escritor ou sobre a escrita de um assunto particular.

No dicionário estadunidense Merriam-Webster[10], assim como nos demais, *corpus* é apresentado como um substantivo, cujo sinônimo pode ser "obra", e *corpora* como sua forma no plural. Em outras palavras, corresponde a todos os escritos ou a obras de um tipo particular, ou sobre um assunto em especial, por exemplo, as obras completas de um autor ou, ainda, a uma coletânea de conhecimento, de enunciados gravados, usados como base para a análise descritiva de um idioma. Para os aprendizes de língua Inglesa, *corpus* também pode ser definido como uma coleção de escritos, conversas, discursos, entre outros, que as pessoas usam para estudar e

[8] Disponível em: https://www.dicio.com.br/corpus/. Acesso em: 22 mar. 2019.

[9] "Corpus*,* noun [Language Database] A collection of written or spoken material stored on a computer and used to find out how language is used: All the dictionary examples are taken from a corpus of billions of words. *Corpus* noun [Collection of Writing] A collection of a single writer's work, or of writing about a particular subject. *Corpus* no Inglês Americano[9] - noun US /ˈkɔr•pəs/ plural corpora US/ˈkɔr•pər•ə/ - A collection of written and spoken language used in the study of language and in writing dictionaries". Definição de "*Corpus*" do Cambridge Advanced Learner's Dictionary & Thesaurus Cambridge University Press. Disponível em: https://dictionary.cambridge.org/pt/dicionario/ingles/corpus?q=corpus+. Acesso em: 23 mar. 2019.

[10] "Corpus noun. cor•pus \ˈkȯr-pəs\ plural corpora\ˈkȯr-p(ə-)rə\. [...] 3a: All the writings or works of a particular kind or on a particular subject especially: the complete works of an author. b: A collection or body of knowledge or evidence especially: A collection of recorded utterances used as a basis for the descriptive analysis of a language. Synonyms for corpus: Oeuvre. In English Language Learners Definition of corpus: a collection of writings, conversations, speeches, etc., that people use to study and describe a language: A collection of poems, paintings, songs, etc.". Disponível em: https://www.merriam-webster.com/dictionary/corpus#synonyms. Acesso em: 22 mar. 2019.

descrever uma linguagem, como uma coleção de poemas, de pinturas, de canções etc.

Para Sanchez, *corpus* é:

> Um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise (SANCHEZ, 1995, p. 8-9).

Com relação às definições de *corpus,* Kennedy (1998) faz duas observações: (i) a definição de um *corpus* como uma coleção de textos baseados em dados eletrônicos pode gerar muitos questionamentos, pois existem diferentes tipos de *corpora*; (ii) algumas definições, apresentadas em dicionários, sugerem que os *corpora* consistem necessariamente em coleções estruturadas de textos compilados para análise linguística, e que são extensos ou que buscam ser representativos como um todo. Para o autor, as questões relativas ao *corpus* vão muito além do que foi mencionado, uma vez que nem todos os *corpora* que podem ser utilizados para pesquisa linguística foram originalmente compilados para esse propósito.

Nesse contexto, ao apresentar algumas das muitas definições encontradas, observamos que *corpus* se refere a um banco de dados autênticos de uma língua ou uma variedade linguística, sejam eles frases, textos escritos ou, até mesmo, áudios selecionados com um propósito de análise ou de estudos linguísticos. Vale mencionar que essas definições contemplam as línguas orais, no entanto podemos adaptá-las a nossa proposição, que é investigar *corpus* nas línguas de sinais.

Nessa perspectiva, é importante ter em mente que trataremos de banco de dados em vídeos, considerando a tridimensionalidade das línguas de sinais. Portanto, a evolução tecnológica e a consequente construção de *corpora* eletrônicos são imprescindíveis para a construção de *corpora* em língua de sinais, pois, por meio de vídeos, sugere-se que possamos expressar, de uma maneira mais genuína, a língua de sinais, para que as análises linguísticas sejam representativas dessas línguas ou das variedades linguísticas.

Embora os *corpora*, atualmente, sejam necessariamente armazenados de forma eletrônica, historicamente nem sempre foi assim. Segundo Berber Sardinha

(2000b), já na Antiguidade e na Idade Média, os *corpora* eram utilizados tanto para registrar os aspectos linguísticos culturais de um povo, quanto informações de caráter religiosos, notadamente bíblicas. Na Grécia Antiga, por exemplo, um dos mais famosos é o *Corpus* Helenístico, criado e motivado por Alexandre, o Grande, ao longo de todo o seu reinado, no período de 336 a.C. a 323 a.C. Esses *corpora* eram coletados e analisados manualmente. Além disso, os trabalhos eram voltados mais ao ensino de línguas, diferentemente das funções estabelecidas atualmente, que são relativas à descrição da linguagem (BERBER SARDINHA, 2000b).

McCarthy e O'Keffee (2010) apontam que, até o século XIII, os trabalhos relacionados a *corpus* eram compostos, majoritariamente, por palavras e frases extraídas de diversos contextos e em grande quantidade de textos, realizados pelos estudiosos da Bíblia cristã. Eles, ao indexarem manualmente as suas palavras, tinham o intuito de especificar, para outros estudiosos bíblicos, as palavras organizadas de forma alfabética, junto com as citações e as passagens que nelas ocorreram, a fim de afirmar e repassar a ideologia divina que constava naquelas escrituras, portanto não eram apenas uma infinidade de textos com diversas fontes.

A partir do século XX, o uso de *corpora* abandona o modelo que até então vinha sendo adotado, voltando-se agora, mais especificamente, para os registros de informações linguísticas úteis para os estudos descritivos das línguas. Em 1921, Thorndike, por exemplo, registrou em seu *corpus* 4,5 milhões de palavras, no qual destacava as de maior frequência na língua inglesa – trabalho plausível, dada às condições da época, e que impulsionou mudanças no ensino da língua materna e estrangeira nos Estados Unidos e na Europa. Posteriormente, passados quase 25 anos, o mesmo autor tomou como base um *corpus* de 18 milhões de palavras e publicou uma obra com as 30 mil palavras mais frequente na língua inglesa. Apesar da relevância dessas publicações, muitas críticas foram tecidas, sobretudo, por se tratar de um trabalho feito manualmente, o que gerava dúvidas quanto à confiabilidade desses *corpora* (BERBER SARDINHA, 2000a).

O período crítico para os estudos baseados em *corpus* se sucedeu com a mudança de paradigma da Linguística, a partir das ideias de Chomsky, por volta de 1950. Pois, muitas críticas se basearam na preferência pelos estudos em teorias racionalistas da linguagem, especialmente as relacionadas à necessidade de se coletar dados empíricos e o meio pelo qual se realizavam as coletas e as análises de dados. Mais adiante, abordaremos sobre os estudos baseados em *corpora* que se

opõem à linguística Chomskyana, por priorizar uma análise empírica, ou seja, partiremos da abordagem Funcionalista baseada em Halliday e Sinclair.

Um dos argumentos, usados para embasar as críticas, dizia respeito à falta de confiabilidade em analisar manualmente grandes quantidades de dados linguísticos. Apesar desse panorama desfavorável, muitos pesquisadores continuaram seus estudos por meio de *corpora,* enquanto a grande maioria defendia a descrição da linguagem por meio de dados reais.

A despeito dos insistentes questionamentos que se faziam ao uso de *corpora* para os estudos linguísticos, em 1953, em Londres, Randolf Quirk e sua equipe compilaram o *corpus Survey of English Usage* (SEU), organizado em fichas de papel e planejado para ter o tamanho de 1 milhão de palavras. Cada papel continha uma palavra inserida em 17 linhas de texto, e cada ficha recebeu uma categoria gramatical. Esse conjunto de categorias serviu de base para os etiquetadores computadorizados contemporâneos.

A título de esclarecimento, as etiquetas utilizadas em *corpora* são códigos/abreviações criados de categorias gramaticais que, antes das ferramentas computacionais, eram realizados manualmente em papel *holerites*, e atualmente se utilizam de ferramentas computacionais.

Embora a construção e a manipulação dos trabalhos ainda fossem realizadas manualmente, foi o Survey, que serviu como base para o surgimento de muitos *corpora* computadorizados, dentre eles o corpus Brown (Brown University Standard Corpus of Present-Day American), ainda hoje, bastante explorado tanto por linguistas, quanto por estudiosos de línguas. Assim, em 1964, criou-se o primeiro *corpus* linguístico eletrônico com um milhão de palavras, o *corpus* Brown, que foi considerado como o fato propulsor para o desenvolvimento da LC*.*

Nesse momento, ainda existiam muitas dificuldades em informatizar esses dados. No caso do *corpus* supracitado, por exemplo, os textos foram transferidos por cartões perfurados um a um, trabalho admirável devido às condições da época. Já a transformação completa do Survey em *corpus* eletrônico só ocorreu em 1989, sendo que anteriormente apenas havia sido computadorizada uma parte de forma falada, que ficou conhecida como o *London-Lund Corpus* (BERBER SARDINHA, 2000a).

Ao mesmo tempo que os corpora eletrônicos tornavam-se uma importante fonte de dados para os estudos linguísticos, avanços tecnológicos na área da computação também aconteciam. Por volta dos anos 60, com o advento dos computadores de

grande porte (mainframes), a exploração e as vantagens proporcionadas pela nova tecnologia imprimiram avanços ainda mais notáveis para os estudos das línguas que já vinham sendo desenvolvidos. Tais avanços possibilitaram a realização de tarefas mais complexas e um aumento considerável na capacidade de armazenamento de dados, a começar pela substituição de cartões perfurados por fitas magnéticas.

Nos anos 1980, com a entrada de microcomputadores pessoais, a disseminação de *corpora* e o surgimento de ferramentas computacionais mais modernas, o fortalecimento das pesquisas linguísticas baseada em *corpus*" (BERBER SARDINHA, 2000a, p. 327) tornou-se uma realidade. (ALUÍSIO; ALMEIDA, 2006). Todos esses avanços deram mais visibilidade à área, atraindo estudiosos de diferentes campos do conhecimento, para além dos linguistas, que buscam subsídios em *corpus*, para suas mais diferentes análises.

Após a exposição sobre as definições e a evolução de *corpora* nas pesquisas relativas à linguagem, na subseção a seguir, discorremos sobre a relevância do(s) *corpus/corpora* para a Linguística, abordando, ainda, suas definições para essa ciência, em paralelo com o *status* que o *corpus* alcançou nas pesquisas linguísticas, com o surgimento da LC.

## 3.1 A RELEVÂNCIA DE CORPUS PARA A LINGUÍSTICA E PARA A LINGUÍSTICA DE CORPUS

Ao realizarmos investigações a respeito da evolução histórica da LC, constatamos que os *corpora* sempre foram recursos utilizados por diversos pesquisadores que, já nos séculos XV e XVII, com a publicação de diversas obras reconhecidas, beneficiavam-se de diversos exemplos de uso de práticas linguísticas, como para as palavras nomeadas no dicionário de Murakawa (2001, 2006). Esses recursos disponíveis deram o suporte necessário ao avanço de vários campos teóricos, especialmente no tocante aos estudos linguísticos.

Segundo Aluísio e Almeida (2006), sempre foram utilizados *corpora* como recursos para as pesquisas linguísticas. Ao abordarem as concepções de *corpus* para a Linguística e para LC, baseadas em definições retiradas de dicionários de Linguística e de alguns autores da área, observaram que uma das diferenças entre as concepções está basicamente no formato do *corpus*, pois, para a LC, os dados devem

estar em formato eletrônico, possibilitando serem compilados e processados por computador.

No caso da Linguística, de acordo com Galisson e Coste, o *corpus* é definido como:

> Um conjunto finito de enunciados tomados como objeto de análise. Mais precisamente, conjunto finito de enunciados considerados característicos do tipo de língua a estudar, reunidos para servirem de base à descrição e, eventualmente, à elaboração de um modelo explicativo dessa língua. Trata-se, pois, de uma coleção de documentos quer orais (gravados ou transcritos) quer escritos, quer orais e escritos, de acordo com o tipo de investigação pretendido. As dimensões do corpus variam segundo os objetivos do investigador e o volume dos enunciados considerados como característicos do fenômeno a estudar. Um corpus é chamado exaustivo quando compreende todos os enunciados característicos. E é chamado seletivo quando compreende apenas uma parte desses enunciados. (GALISSON; COSTE, 1983, p. 157).

Essa definição não menciona a exigência que as construções de *corpora* sejam eletrônicas para os estudos linguísticos, a questão do tamanho varia conforme as necessidades de cada análise e a questão da exaustividade pode ser flexibilizada. Um *corpus* exaustivo, por exemplo, compõe uma gama de enunciados com as mais diversas informações, enquanto um seletivo apresenta apenas uma parte desses enunciados, com objetivos previamente determinados pelo pesquisador.

Para a LC, conforme destaca Berber Sardinha (2004), a concepção de *corpus* está intimamente ligada a ferramentas computacionais. Para o autor, a LC é uma:

> Abordagem que se ocupa da coleta e da exploração de corpora, ou conjuntos de dados linguísticos textuais que foram coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por computador. (BERBER SARDINHA, 2004, p. 3).

Com base nos autores supracitados, verificamos que o trabalho com *corpora* apresenta duas características distintas entre a Linguística e a LC: (i) o armazenamento e a análise, para Linguística, era manual, enquanto para LC, computadorizado; (ii) a ênfase do trabalho era, em geral, voltada, no caso da Linguística, ao ensino de línguas e, para LC, à descrição de uma língua ou de uma variedade linguística.

No campo dos estudos linguísticos, a LC surge como uma abordagem mais sistematizada, pois é auxiliada pelo uso de ferramentas computacionais específicas. Esse fator foi indispensável, por exemplo, para o desenvolvimento dos estudos lexicográficos, mas também para outras múltiplas áreas da Linguística (SHEPHERD, 2012). Portanto, como afirma a autora, outras áreas da linguística também foram beneficiadas com as novas tecnologias, como os estudos relativos ao processo de ensino e aprendizagem de línguas ou variedades linguísticas, dos estudos de tradução e para o treinamento de tradução por máquina.

Na perspectiva da LC*,* Sinclair (2005), considerado o maior linguista de *corpus* da história, desenvolveu um trabalho de muita relevância na área do léxico, com a publicação do dicionário inglês COBUILD, criado a partir de uma parceria entre a Universidade de Birmingham (Grã-Bretanha) e a editora Collins.

No âmbito do COBUILD, foram produzidos vários dicionários, gramáticas e livros didáticos para o ensino do inglês. Atualmente, quase desativado, esse dicionário permanece como referência no desenvolvimento e na aplicação da pesquisa baseada em *corpus* com fins comerciais, pois foi o primeiro a ser compilado a partir de um *corpus* computadorizado.

Conforme destaca Sinclair (2005), um *corpus* consistiria em uma coleção de partes do texto de uma língua em formato eletrônico, cujos critérios de seleção seriam externos, a fim de representar, tanto quanto possível, uma língua ou uma variedade de idiomas como fonte de dados para pesquisa linguística.

Mediante a essas constatações, a principal diferença de *corpus* para LC e a Linguística se refere ao formato eletrônico, ou seja, para a LC são relevantes somente os dados que podem ser processados pelo computador. Dessa forma, somente se considera livros, revistas e textos impressos, para a LC, quando esses dados são readaptados para uma linguagem que seja reconhecida por computador.

De acordo com McEnery e Wilson (1996), o emprego do termo *corpus*, contemporaneamente, demanda quatro características específicas assim nomeadas: (i) amostragem e representatividade, ou seja, amostragem suficiente da língua e sua variedade para estudo e análise; (ii) tamanho finito, isto é, apresentação de um *corpus* com 1000 palavras, 200 palavras, dentre outros; (iii) formato eletrônico, diferente da ideia em referir-se apenas a textos impressos, pois há benefícios, como a pesquisa e a manipulação dos *corpora* mais rápida, e seu enriquecimento com acréscimo de informações, dentre outros; e (iv) constituição de um *corpus* como referência padrão

para representatividade de uma variedade de língua, disponibilizando-o para outros investigadores.

É importante salientar que para alguns autores, como citam Aluísio e Almeida (2006), diferentemente da Linguística, a LC não considera como *corpus* livros, revistas e outros textos na forma impressa e, também, desconsideram a *web*, neste caso, por ela ter dimensões desconhecidas, além de passar por mudanças constantes, o que torna questionável a questão da sua confiabilidade. Em outras palavras, para esses autores, para que seja considerado um *corpus*, além dos dados, necessariamente, estarem compilados num formato eletrônico e serem autênticos, devem ser coletados com um propósito de pesquisa e ter uma dimensão estabelecida.

Por outro lado, McEnery e Wilson (1996) mencionam que a *web*, ao avivar uma das características elencadas em seus estudos, tornou-se um instrumento necessário para a distribuição e o livre acesso de vários *corpora* apresentados em diversos projetos.

Kilgarriff e Grefenstette (2003) afirmam que a própria *web* pode ser considerada um *corpus*. Defendem que é imprescindível reconhecer que o surgimento do computador interferiu na concepção de *corpus* e na sua capacidade de armazenamento e sondagem, visto que uma série de conteúdos está disponibilizada em um curto espaço de tempo, além de possibilitar amostragem acelerada e eficiente de muitos fenômenos linguísticos. Em suma, ao contrário dos recursos manuais, com o computador há a possibilidade de armazenamento de inúmeros textos, sendo um fator importante para auxiliar na observação e na descrição de acontecimentos linguísticos que antes, sem o auxílio das ferramentas computacionais, poderiam ser imperceptíveis.

Desse modo, na década de 1990, as contribuições advindas da Computação e da Linguística Computacional passaram a exercer relevante papel para os estudos dos *corpora,* pois o aprimoramento e o desenvolvimento de ferramentas computacionais representam novos processamentos na língua, no caso do Brasil, por exemplo, na língua natural do português brasileiro e para o desenvolvimento dos estudos relacionados a *corpus.*

Em consonância com Trask (2004), a partir de *corpora*, pode-se fazer observações precisas sobre o comportamento linguístico de falantes reais, proporcionando informações altamente confiáveis e isentas de opiniões e de julgamentos prévios sobre os fatos de uma língua.

Nesse propósito, entende-se que é possível observar aspectos primordiais em uma investigação linguística, tais como: morfológicos, sintáticos, semânticos e pragmáticos. Além disso, é possível justificar o emprego das palavras, das expressões e das formas gramaticais, ou seja, por intermédio de um *corpus,* há a possibilidade e a perspectiva de se reconhecer a língua de uma forma objetiva, como afirma Beber Sardinha (2000).

Com propriedade, Fromm (2003) cita algumas análises linguísticas que podemos realizar por meio de um *corpus*, dentre elas: (i) a frequência das palavras mais comuns de uma língua ou variedade; (ii) a frequência das classes gramaticais; (iii) a comprovação de colocações; (iv) o reconhecimento e o detalhamento de lexias compostas e complexas de uma língua; (v) a regência dos verbos preposicionados; (vi) a composição mais provável das estruturas frasais cristalizadas, por exemplo, os provérbios e expressões idiomáticas; (vii) a seleção de nomenclatura para uma obra terminológica; (viii) a criação de dicionários gerais multilíngues; (ix) a verificação de modalidades de tradução em *corpus* bilíngue ou multilíngue; (x) a base de dados para tradutores; e (xi) o ensino de línguas estrangeiras.

Ao levar em consideração os aspectos observados, a respeito da constituição de um *corpus* como referência para futuras pesquisas, sabe-se que qualquer esforço proposto não será útil apenas para uma pesquisa presente, mas servirá de aporte para outros pesquisadores, haja vista que o formato computadorizado do *corpus* e a sua consequente disponibilização a outros observadores são os principais elementos que distinguem a Linguística e a LC.

## 3.2 ASCENSÃO E FORTALECIMENTO DA LINGUÍSTICA DE *CORPUS*

A partir dos anos 1960, a LC passou a ter grande influência na pesquisa científica em diversas áreas do conhecimento. Em estudos mais recentes, Tagnin (2018) destaca que desde a compilação do primeiro *corpus*, realizada por Henry Kučera e Winthrop Nelson Francis, *Corpus* Brown, em 1964, a LC passou a ter importância em áreas nem imaginadas na época. Além da construção desse *corpus* ser determinante aos estudos estatísticos da linguagem, o que veio a ser a LC, passou, também, a servir de modelo para a compilação de outros *corpora*, alguns desses já foram mencionados na primeira seção.

Atualmente, os *corpora* constituídos podem ultrapassar 1 bilhão de palavras, “como o News on the Web (NOW), com mais de 5 bilhões de palavras, ou o Global Web-Based English (GloWbE), o Wikipedia Corpus, o Hansard Corpus, e o Corpus del Español3” (TAGNIN, 2018, p. 11). A autora destaca que todos esses *corpora* são compostos por aproximadamente 2 bilhões de palavras. E, ainda, destaca que:

> O Corpus do Português faz parte do mesmo portal e é constituído de um corpus histórico com 45 milhões de palavras do século XIII ao século XX e um corpus extraído da Web com textos do Brasil, de Portugal, Moçambique e Angola. Não podemos deixar de mencionar o Corpus Brasileiro. Embora a língua inglesa ainda seja privilegiada em termos de variedade de corpora disponíveis, há um considerável número de corpora para outras línguas, tanto acessíveis on-line quanto off-line (VIANA, 2015), esses últimos em geral compilados por pesquisadores para um objetivo específico. (TAGNIN, 2018, p. 12).

Beber Sardinha (2000ª) já citava, nesse período, grandes centros de pesquisas, tais como os das Universidades na Grã-Bretanha, que se dedicavam à observação em *corpus* para a descrição das diversas questões relacionadas à língua, o que foi favorável para consolidar a criação de *corpora* e de materiais de apoio em diversas áreas do conhecimento.

Já nos Estados Unidos, o autor afirma que a LC ainda teria avanços modestos, mesmo contando com centros de pesquisa e com a facilidade em conquistar recursos para a área da informática. Isso se devia pelo fato de empresas de informática investirem altos recursos na pesquisa linguística com fins comerciais, tanto em âmbito acadêmico quanto industrial. Esse foi um fator que contribuiu mais para o desenvolvimento da pesquisa em Processamento da Linguagem Natural (PLN).

Vale ressaltar que a disciplina de PLN é fortemente atrelada à Ciência da Computação, no entanto, apesar dessa área compartilhar vários temas com a LC, tanto uma quanto a outra se mantêm independentes em termos de estudos e pesquisas.

No Brasil, conforme os estudos de Beber Sardinha (1999), a LC já estava sendo colocada em prática em centros focados no PLN, tais como a Lexicografia e a Linguística Computacional. No âmbito empresarial, o interesse crescente nos estudos baseados em *corpora* aumentou*,* pois os investimentos nas pesquisas são, geralmente, com finalidades comerciais, informatização de grandes bases de dados,

montagem de sistemas inteligentes de reconhecimento de voz e gerenciamento de informação.

Além disso, grandes empresas de telecomunicações e de produtos de informática, como a *Xerox*, a *Microsoft* e a *Canon*, têm associações de pesquisa de *corpus* e de PLN. Logo, a LC está ligada ao oferecimento de *corpora* eletrônicos e, consequentemente, condicionada à tecnologia com disponibilidade de ferramentas computacionais para estudo de *corpus*.

Portanto, evidencia-se que a LC se preocupa com diversos fenômenos comumente enfocados em outras áreas (léxico, sintaxe, texturas, dentre outras). Em suma, o crescimento da LC é notório, pois, na proporção em que aparecem mais pesquisadores, esses descobrem no *corpus* uma fonte riquíssima de informação.

Assim, em seu trabalho prático de exploração, sua ascendência tem ganhado reconhecimento em meio aos estudantes, linguistas e pesquisadores anônimos, os quais descobrem novas teorias, levantam mais questionamentos sobre a linguagem em práticas reais de uso efetivo, sedimentando um panorama de reflexão e novos relatos para a LC no Brasil.

Tagnin (2018) ressalta que a LC de *corpus* vem se destacando em diversas áreas. Menciona que, talvez, a primeira área a se beneficiar com LC foi a Lexicografia, com a publicação de dicionários baseados em *corpora* e, a partir daí, outras, como: (i) a Fraseologia, com ferramentas que permitem identificar as recorrências lexicais; (ii) a Terminologia, em seguida, a Terminologia Fraseológica, que somam incontáveis trabalhos; (iii) a Tradução, tanto para a técnica quanto para a literária, é fundamental o uso de "*corpora* comparáveis (textos originais em duas ou mais línguas) e *corpora* paralelos (originais e respectivas traduções em duas ou mais línguas)" (TAGNIN, 2018, p. 13); (iv) a Tradução Automática, principalmente a estatística, com o uso de grandes *corpora* multilíngue que, segundo a autora, contribuem até mesmo para um modelo de *corpus* baseado em regras de tradução para a Libras; (v) a Linguagem Oral, contando com falas espontâneas, que podem contribuir com o ensino de línguas estrangeiras; (vi) a Linguística Aplicada, que além de ter se desenvolvido no ensino, também foi importante na elaboração de gramáticas; e (vii) os Estudos de Gêneros textuais, para análises mais objetivas e detalhadas. A autora afirma que todas essas áreas se referem a linguistas, estudiosos da língua e das linguagens, no entanto a LC está conquistando uma abrangência muito maior.

Com relação aos estudos relativos aos *corpora* paralelos, se, em meados de 2003-2008, as abordagens relacionadas a eles (ou de tradução) eram menos documentadas, conforme afirmam Granger et al. (2003), Olohan (2004) e Anderman e Rogers (2008), atualmente, uma série de pesquisas tem indicado a sua utilidade em estudos de tradução e *corpora* eletrônicos paralelos.

Embora existam outros tipos de *corpora* eletrônicos, conforme citado por Baker (1995), como os comparáveis e os multilíngues, nesse momento, vamos nos ater apenas à definição, citada pela autora, referente a *corpus* paralelo, que são, basicamente, compostos de uma língua partida/origem, e suas respectivas traduções em uma outra língua. Esse tipo de *corpora* é fundamental para fornecer materiais para redação, treinamento para tradutores e aperfeiçoamento de sistemas de tradução automática (CAMARGO, 2012).

Ao considerar as especificidades inerentes a esse tipo *corpus*, essa foi a modalidade selecionada para a proposta principal deste trabalho, que se refere à proposição dos critérios necessários para a criação de um *corpus* paralelo entre a Libras e a Língua Portuguesa.

Nesse contexto, mediante ao fortalecimento e ao uso da LC na comunidade científica, há um amplo debate entre os seus praticantes, a respeito de seus *status*, principalmente, se ela pode ser considerada uma metodologia ou uma disciplina. O que podemos afirmar, de antemão, é que a LC não é uma disciplina como a Sociolinguística, a Psicolinguística ou a Semântica, pois o seu objeto de estudo não é delimitado como em outras áreas.

Portanto, podemos considerá-la, inicialmente, uma metodologia da qual outras áreas podem se utilizar. Todavia, a definição da LC como metodologia ou não dependerá da compreensão que temos do termo. Por exemplo, se entendermos metodologia como um modo típico de se aplicar um conjunto de pressupostos de caráter teórico, então a LC pode ser compreendida como uma metodologia, porque apresenta mais do que um simples instrumento computacional.

Em contrapartida, alguns autores afirmam que a LC não é uma metodologia, pois, por meio da sua consulta, seus participantes podem produzir um novo conhecimento, o que implicará aceitação ou não do que é proposto. Kennedy agrega à discussão ao afirmar que,

> Embora o escopo da Linguística de Corpus possa ser definido em termos do que as pessoas fazem com corpora, seria um engano assumir que Linguística de Corpus é somente um meio mais rápido de descrever como a linguagem funciona. A análise de um corpus pode revelar, e frequentemente revela, fatos a respeito de uma língua que nunca se pensou em procurar. (KENNEDY, 1998, p. 9).

Há autores, ainda, que trabalham com outra possibilidade, a de que a LC não é nem disciplina e nem metodologia. Hoey (1997) salienta que a LC não é um ramo da Linguística, mas a rota para a Linguística, ou seja, para se chegar à linguagem, portanto, não seria apenas instrumental, mas uma abordagem.

Isso significa que, como abordagem, a LC é vista de uma forma mais ampla, o que inclui um conjunto de posições, de crenças teóricas, de valores, um sentido mais filosófico a respeito da linguagem. Vale ressaltar que todas as abordagens têm suas críticas, assim como o ensino de línguas. Dessa forma, ao questionar se a LC pode ser considerada uma metodologia ou uma abordagem, constatamos que esse fator dependerá de como ela será utilizada na pesquisa.

Nesse contexto, ao refletir sobre a proposta que norteia esta pesquisa, podemos considerar a LC como abordagem, uma vez que ela serviu como base para nos orientar na reflexão de quais seriam os critérios necessários para a constituição de um *corpus* paralelo Libras-Português.

A partir dessas considerações, é possível afirmar que a LC é uma das áreas mais intensas voltadas ao estudo da linguagem. Nesse sentido, muitos trabalhos têm sido desenvolvidos, não só nas línguas orais, como também nas línguas de sinais, de modo a proporcionar grandes impactos na comunidade de linguistas, auxiliando para que a descrição fundamentada em *corpus* se torne uma norma e não exceção.

## 3.3 OS PRESSUPOSTOS DETERMINADOS PELA LINGUÍSTICA DE *CORPUS*

Sobre a definição da LC, Berber Sardinha (2000b, p. 2) ressalta que a “Linguística de Corpus se ocupa da coleta e exploração de *corpora*, ou conjuntos de dados linguísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística”. Do mesmo modo, dedica-se à investigação da linguagem por meio de evidências empíricas, que são extraídas com auxílio de computador.

Para o autor, a história da LC está intimamente ligada à disponibilidade de *corpora* eletrônicos. Por outro lado, o pesquisador destaca que, nesse período, apesar da LC estar bastante avançada na Europa, fora dela o seu desenvolvimento ainda era lento, embora já existissem centros em que as pesquisas estariam consolidadas.

Mais recentemente, de acordo com Aijmer e Altenberg (2013), a LC é discutida com base nas vantagens que envolvem uma abordagem de *corpus*, cujos marcadores linguísticos e o uso de *corpora* permitem ao pesquisador ver o uso no contexto e descobrir regularidades e padrões de uso, com respeito a uma gama de informações, sejam elas, por exemplo, de classes gramaticais, tipos de texto e gêneros.

Sobre os *corpora* eletrônicos, Berber Sardinha (2000b) apresenta um levantamento sobre os que foram compilados ou os que estão em compilação e estreitamente ligados ao desenvolvimento da LC, dentre eles, destacam-se três como os principais de referência histórica: (i) o *corpus Brown*; (ii) o *Birmingham*; e o (iii) *corpus* BNC.

Conforme as pesquisas de Berber Sardinha (2000a), podemos observar que a LC já exercia grande influência nas pesquisas linguísticas há mais de 20 anos. A Grã-Bretanha, na década de 1990, destacava-se como um dos centros mais desenvolvidos, haja vista que várias universidades, tais como Birmingham, Brighton, Lancaster, Liverpool, Londres, dentre outras, dedicavam-se à pesquisa baseada em *corpus* para a descrição dos mais variados aspectos da linguagem. Do mesmo modo, os países escandinavos (Noruega, Suécia e Dinamarca) também foram, e ainda são, destaques quando se trata de centros estabelecidos dedicados à LC.

É de consenso, entre os pesquisadores supracitados, que a LC se concentra em analisar um conjunto de dados textuais com evidências empíricas, cuja coleta dos dados na base conta com o propósito de uma análise criteriosa de uma determinada língua ou de uma variedade linguística, por meio de computador.

Nessa perspectiva, Berber Sardinha (2000a) diz que a LC trabalha dentro de uma Abordagem Empirista e uma visão da linguagem como sistema probabilístico. O empirismo se baseia em dados obtidos por meio da observação da linguagem, opondo-se ao racionalismo, que verifica o funcionamento estrutural e o processamento cognitivo da linguagem.

De um lado, Halliday seguia a tradição empirista e via a linguagem como probabilidade, de outro, Chomsky, o maior expoente do racionalismo, enxergava a linguagem como possibilidade (KENNEDY, 1998). Assim, a LC, diferentemente da

Linguística Chomskyana, tem seu foco no desempenho linguístico e na descrição linguística a partir de uma visão mais empirista, ao invés de ter foco na competência linguística, em universais linguísticos e no racionalismo da pesquisa científica.

Diante disso, para a construção de um *corpus*, Berber Sardinha (2000) sugere que é necessária a produção de textos naturais "autênticos", ou seja, "não criados com o propósito de figurarem no *corpus"*. A produção deve-se ser "natural", o que significa dizer que deve ser produzido por humanos, jamais "provinda de programas de geração de textos" (BERBER SARDINHA, 2000, p. 336).

Nesse sentido, entende-se que um *corpus* é artificial quando se trata de "um objeto criado com fins específicos de pesquisa" (BERBER SARDINHA 2000a, p. 336). Assim, a melhor definição que incorpora a constituição de um *corpus* é feita por Berber Sardinha (2000a) com base em Sanchez (1995), que define *corpus* como um "conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade" (BERBER SARDINHA 2000a, p. 336).

Tais critérios devem estar organizados e levar em consideração o modo de representatividade, ou seja, "representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador" (BERBER SARDINHA 2000a, p. 336). Dessa maneira, a finalidade é propiciar resultados variados e úteis para a descrição e a análise de dados.

Shepherd (2012, p. 26), ao observar as mudanças tecnológicas que ocorreram ao longo do tempo, em relação aos estudos baseados em *corpora,* destaca que os linguistas de *corpus* estão "aprendendo a pensar como engenheiros de computação, usando e/ou desenvolvendo pequenas ferramentas para tarefas específicas em Linguística de Corpus". Sobre esse aspecto, o autor afirma que:

> As parcerias e interfaces da Linguística de Corpus com a área de Programação e Processamento de Linguagem Natural têm sido inúmeras e já produziram tecnologia para dicionários, analisadores sintáticos e morfológicos corretores ortográficos, interfaces em língua natural e ferramentas voltadas ao ensino. (SHEPHERD, 2012, p. 27).

Diante da importância que os *corpora* têm para as mais diversas áreas e do *status* que a LC vem alcançando nos últimos tempos, Berber Sardinha (2000a) sugere que o *corpus* seja uma amostra de uma população, a qual não se sabe a dimensão

de usuários da língua. Por isso, é importante tornar essa amostra a maior possível para que se aproxime da população e que seja representativa da língua em questão.

Fromm (2003) propõe que, antes mesmo de realizar a coleta do material, o pesquisador se questione sobre quais seriam os objetivos acerca do *corpus* que se pretende construir, a fim de economizar tempo, uma vez que trabalhará com uma gama de informações. Dessa forma, algumas reflexões como: (i) qual o tipo de pesquisa pretendemos aplicar no *corpus*?; (ii) a quem se destina o *corpus* a ser construído?; (iii) quais as fontes que serão trabalhadas?; (iv) qual o tamanho pretendido?; (v) Qual o meio que deverá ser publicado? são algumas questões prévias, que poderão colaborar com o andamento do trabalho.

Quanto ao tamanho do *corpus,* por vezes, também tem sido um problema, tendo em vista que ainda há divergências quanto à quantidade mínima de dados necessários para a formação de um *corpus*, uma vez que vem se discutindo que o tamanho é um critério subjetivo na sua definição.

Quanto à extensão, Berber Sardinha (2000a) destaca que ainda há poucos estudos sobre essa questão, mas define três abordagens sobre os critérios mínimos de extensão para a constituição de um *corpus* representativo de uma população: (i) a abordagem impressionística, que se baseia em constatações derivadas da prática da criação e da exploração de *corpora*, em geral feita por autoridades da área; (ii) a abordagem histórica, que se fundamenta na monitoração dos *corpora* efetivamente usados pela comunidade; e (iii) a abordagem estatística, baseada na aplicação de teorias estatísticas, que pode ser subdividida em três vertentes – interna, externa e relativa. Leva-se em consideração, ainda, a sua especificidade e a sua adequação, ou seja, se os dados são apropriados à investigação.

Quanto à composição interna de um corpus, Baker apresentou alguns critérios de seleção, dentre eles:

> (i) linguagem geral vs. domínio restrito
> (ii) linguagem escrita vs. falada
> (iii) sincrônico vs. diacrônico
> (iv) tipo em termos de variedade de fontes (escrita/falada) e gêneros (por exemplo, editoriais de jornais, entrevistas de rádio, ficção, artigos de jornal, audiências judiciais)
> (v) limites geográficos, por exemplo, o inglês britânico vs. o inglês americano

> (vi) monolíngue vs. bilíngue ou multilíngue. (BAKER, 1995, p. 229, tradução nossa).[11]

Posteriormente, de forma pontual, Berber Sardinha (2000a) define e apresenta, de maneira criteriosa, alguns pontos relevantes para a constituição de um *corpus*:

> A origem: os dados devem ser autênticos;
> O propósito: o *corpus* deve ter a finalidade de ser um objeto de estudo linguístico;
> A composição: o conteúdo do *corpus* deve ser criteriosamente escolhido;
> A formatação: os dados do *corpus* devem ser legíveis por computador;
> A representatividade**:** o *corpus* deve ser representativo de uma língua ou variedade;
> A extensão: o *corpus* deve ser vasto para ser representativo. (BERBER SARDINHA, 2000a, p. 338).

Ao levar em consideração todos esses aspectos, é possível observar a relação indissociável entre a LC e as ferramentas computacionais. Dessa forma, a partir da LC, torna-se possível analisar quais são os critérios necessários para construção de um *corpus,* de maneira que sejam realmente úteis aos pesquisares e representativos aos estudos linguísticos.

As discussões realizadas até o momento são fundamentais para nos direcionar à proposta principal desta pesquisa, referente aos critérios necessários para a construção de um *corpus* paralelo de Libras em interface com a Língua Portuguesa, de maneira que seja representativo aos pesquisadores e estudiosos da área.

---

[11] “(i) general language vs. restricted domain
(ii) written vs. spoken language
(iii) synchronic vs. diachronic
(iv) typocality in terms of range of sources (writes/speakers) and genres (e.g. newspaper editorials, radio interviews, fiction, journal articles, court hearings)
(v) geographical limits, e.g British vs. American English.
(vi) monolingual vs. bilingual or multilingual.” (BAKER, 1995, p. 229).

## 4 A UTILIZAÇÃO DE *CORPORA* NAS LÍNGUAS DE SINAIS

Ao percorrer sobre o desenvolvimento de *corpus* nas línguas de sinais, nota-se que, assim como nas línguas orais, a preocupação dos pesquisadores em estar buscando uma metodologia confiável e adequada às necessidades de análises linguísticas estão se intensificando, uma vez que cresceu o número de linguistas preocupados em descrever os fenômenos linguísticos que ocorrem na língua de sinais, bem como suas peculiaridades, no que se refere a uma modalidade gestual, visual e espacial[12].

Os avanços tecnológicos foram essenciais para a ascensão da LC, e trabalhar nessa perspectiva contribui com os estudos realizados nas línguas de sinais, pois possibilitam registros em vídeos, os quais são indispensáveis para essas línguas que se manifestam de maneira visual-espacial. Nesse contexto, vários países, inclusive o Brasil, vêm desenvolvendo pesquisas e ferramentas que viabilizam e tornam representativos os estudos executados nessa área (QUADROS, 2016).

Conforme Quadros (2019), há *corpora* construídos em línguas de sinais em vários países. Alguns destacados pela autora, incluindo o *corpus* criado/idealizado por ela e por toda uma equipe no Brasil, mais precisamente na Universidade Federal de Santa Catarina (UFSC), fomentado por pesquisas realizadas em nível de mestrado e doutorado. Seguem os apresentados pela autora:

- *Corpus* Libras: www.corpuslibras.ufsc.br;
- *Corpus* da Língua de Sinais Australiana: http://www.auslan.org.au/about/corpus/;
- *Corpus* de Língua de Sinais Britânica: http://www.bslcorpusproject.org/;
- *Corpus* da Língua de Sinais Alemã: https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html;
- *Corpus* da Língua de Sinais Holandesa: http://www.ru.nl/corpusngtuk/;

[12] Com relação às diferentes modalidades de língua, é de consenso entre os linguistas da área que as línguas orais se manifestam de forma oral e auditiva, e as línguas de sinais, de forma visual espacial/visuo-espacial (STOKOE, 1960; QUADROS, 2004; FERREIRA BRITO et. al., 1998), ou ainda, de forma visual, gestual e espacial (ZWITSERLOOD, 2014; RODRIGUES, 2018). Nesse sentido, as línguas de sinais são percebidas pelos olhos e expressadas pelas mãos e pelo corpo em um espaço de sinalização.

- *Corpus* da Língua de Sinais Polonesa: http://www.plm.uw.edu.pl/en/node/241;
- *Corpus* da Língua de Sinais Japonesa: http://reserch.nii.ac.jp/jsl-corpus/public/en/index.html.

Como se pode observar, apesar da linguística baseada em *corpus* ser predominantemente voltada às línguas orais escritas, alguns *corpora* eletrônicos já despontam também nas línguas de sinais. Se por um lado, nas línguas orais, observa-se o avanço e o domínio das pesquisas que florescem a todo o momento, aproveitando as capacidades das inúmeras ferramentas computacionais e *softwares* linguísticos, muitos deles voltados à aprendizagem de línguas, lexicografia, tradução automática, dentre outros; as línguas de sinais, por outro lado, são línguas visuais-gestuais e espaciais (ZWITSERLOOD, 2014; RODRIGUES, 2018). E além dessa diferença na modalidade, a escrita de sinais (*SignWriting*) ainda não é uma forma de escrita estabilizada até o momento, o que dificulta o seu processamento computacional. Como consequência, o registro e a organização dos *corpora* em línguas de sinais, necessariamente, são em vídeos que correspondem as mais diferentes línguas de sinais e, nesse aspecto, a sua descrição, que envolve desde a identificação de unidades básicas até a descrição da sintaxe ou semântica dessas línguas. Todos esses fatores tem sido um grande desafio para os linguistas da área e mais desafiador ainda para os linguistas de *corpus*. Nesse sentido, a LC, nas línguas de sinais, configura-se como um novo domínio linguístico e, portanto, bastante desafiador.

## 4.1 TRABALHOS DESENVOLVIDOS NO BRASIL REFERENTES A *CORPUS* DE LIBRAS EM FORMATO ELETRÔNICO

Nesta seção, apresentaremos trabalhos desenvolvidos no Brasil, referentes a alguns *corpora* ou, simplesmente, bancos de dados em formato eletrônico, constituídos em Libras, com o intuito de analisar quais são os critérios que já foram utilizados para a construção desses para, posteriormente, tecer comentários a respeito da possibilidade de utilizá-los para análises linguísticas, embasado nas concepções estabelecidas pela LC.

Quadro 1 - Registros em vídeos de dados em Libras, desenvolvidos no Brasil

| ***Corpus* – vídeos em libras** | **Conteúdo temático** | **Critérios identificados** |
|---|---|---|
| Manuário Acadêmico e Escolar – INES. Disponível em: http://www.manuario.com.br/indice-geral. | - Sinais relativos à área acadêmica. Estão organizados por área do conhecimento. | (i) Coleta de sinais junto a alunos surdos, professores e intérpretes do Instituto; |
| | | (ii) Sessões de validação desses sinais com professores surdos do INES e outros representantes da comunidade acadêmica; |
| | | (iii) Filmagem em estúdio dos sinais validados; |
| | | (iv) O repertório lexical pesquisado e registrado compreende conceitos e autores pertinentes ao universo escolar e acadêmico; |
| | | (v) Fundo azul; |
| | | (vi) *Links* por área, palavra por palavra de cada uma delas. |
| Sinalário Ilustrado de Química em Libras – DIDAPS[13]/INES. Disponível em: https://www.youtube.com/playlist?list=plbds0n-o6umocgyk3jdvixn2aqsmg6oct. | - Sinais relativos aos conteúdos da disciplina de Química.<br>- Está disponível na plaforma *youtube*, são relativos aos sinais do Manuário Acadêmico INES. | (i) Idem Manuário INES; |
| | | (ii) Também disponível na plataforma *youtube.* |
| Sinalário Disciplinar de Libras – CAS/SEED-PR App. Disponível para baixar em: https://play.google.com/store/apps/details?id=br.com.app.gpu1766632.gpu62fe9a3bd58b6fdb4b3dd202609a2594&hl=pt_BR<br><br>Disponível em: https://www.youtube.com/channel/UCoWGC5Tas9TTQhGRCHcpD2w | - Sinais correspondentes às disciplinas da Educação Básica das escolas públicas do estado do Paraná. | (i) Vídeo em Libras; |
| | | (ii) Soletração/datilologia da palavra em português; |
| | | (iii) Oralização da palavra; |
| | | (iv) Contextualização de sinais; |
| | | (v) Fundo azul; |
| | | (vi) Canal disponível na plataforma *youtube*; |

[13] O Desenvolvimento de Instrumentos Didáticos Acessíveis na Perspectiva Surda (DIDAPS/INES) é um Grupo de Pesquisa, liderado pelas professoras de Química Joana Saldanha e Jomara Fernandes, e pelas professoras de Libras Vanessa Lesser e Bárbara Carvalho, que conta com uma linha de pesquisa cujo grupo Sinalizando Química (SinQui) desenvolve trabalhos na criação de sinais de conceitos químicos em Libras. Este grupo tem um canal no *YouTube*, que disponibiliza material didático em Libras na área de Ciências, além de um rico sinalário ilustrado contendo o sinal, o conceito do elemento químico em Libras e uma imagem correspondente. Disponível em: http://www.manuario.com.br/dicionario-tematico/quimica. Acesso em: 20 out. 2019.

| | | |
|---|---|---|
| Informações disponíveis em: http://www.alunos.diaadia.pr.gov.br/modules/conteudo/conteudo.php?conteudo=1531. | | (vii) Aplicativo (ferramenta de apoio);<br>(viii) Ícone para contato – possibilidade de sugestões do usuário. |
| Sinalário de Libras I e II – UFPR. Disponível em: https://www.youtube.com/watch?v=mz0vJuG3WmQ. | - Um conjunto de sinais criados pelos acadêmicos do curso de Letras Libras da UFPR, para as disciplinas de Libras I e II. | (i) Vídeo em Libras;<br>(ii) Soletração/datilologia da palavra em português;<br>(iii) Sinal correspondente em Libras;<br>(iv) Fundo cinza neutro;<br>(v) Palavra escrita em português no canto superior direito do vídeo (na perspectiva do leitor);<br>(vi) Aspectos gerais do informante: quando mulher, apresenta-se com o cabelo preso, sem acessórios como brincos, correntes, etc., camiseta T-shirt preta com identificação (na cor branca) do curso Letras Libras UFPR. |
| Sinalário Instituto Phala. Disponível em: https://www.youtube.com/user/institutophala/videos. | - Contém um grande conjunto de sinais envolvendo mais de 40 categorias gramatical. Cada categoria compõe um vídeo com link específico.<br>- Também é composto de vários vídeos contendo Fábulas de Esopo, histórias e narrativas. | (i) Diversos vídeos em Libras;<br>(ii) Fundo cinza com ilustrações contextualizadas ao tema;<br>(iii) Título escrito relacionado ao conteúdo sinalizado;<br>(iv) Diferentes temas contextualizados;<br>(v) Vídeos apresentados em português, conta “janela” com intérprete de Libras. |
| Sinalário Técnico em Libras - Eletrônica com as mãos. Disponível em: https://vimeo.com/377126831. | - Conjunto de sinais voltados a vocabulários técnicos em Libras, relacionado ao Curso de Eletrônica do Instituto federal da Bahia (*campus* de Salvador). | (i) Vídeo referente a termos técnicos;<br>(ii) Fundo cinza;<br>(iii) Imagem referente ao sinal;<br>(iv) Palavra escrita em português no canto superior direito do vídeo (na perspectiva do leitor). |

| | | |
|---|---|---|
| Dicionário de Libras Biologia – Canal do *Youtube* – Disponível em: https://www.youtube.com/channel/UCP_FCqS6iCIFaHbGaSZ9cKQ/feed. | - Desenvolvido pelo EPEEM: Grupo de Estudos de Pequenas Empresas e Empreendedorismo, que atuam no setor metal-mecânico no estado do Paraná. | (i) Vídeo referente aos termos técnicos (Biologia); |
| | | (ii) Palavra escrita em português no canto superior esquerdo do vídeo (na perspectiva do leitor); |
| | | (iii) Sinalização do termo em Libras; |
| | | (iv) Fundo azul; |
| | | (v) Aspectos gerais do informante: quando mulher, apresenta-se com o cabelo preso, sem acessórios, camiseta T-shirt preta. Homens, com camiseta T-shirt preta. |
| O diário da Fiorella – canal do Youtube – Disponível em: https://www.youtube.com/channel/UC9g1xELVb53CLrS53UF4kuw/videos. | - Aquisição da linguagem por criança surda filha de pais surdos. | - Trata-se de vários vídeos (amadores), gravados no ambiente natural da criança, disponíveis na plataforma *youtube* desde meados de 2016. Retrata várias fases comunicativas da criança, até os dias atuais. |
| Vídeos em Libras – Disponível em: https://wp.ufpel.edu.br/materialibrasisf/videos-em-libras/. | - Conjunto de vídeos em Libras contendo algumas categorias temáticas e diálogos. Vídeos hospedados no site da UFPEL – Universidade Federal de Pelotas – RS. Direcionado ao curso Básico de Libras que compõe o Projeto Idiomas Sem Fronteira – ISF Libras. | (i) Soletração do tema seguido do sinal; |
| | | (ii) Exibição da palavra no português, seguido do sinal; |
| | | (iii) Fundo azul; |
| | | (iii) Homens e mulheres sem acessórios, camiseta T-shirt preta; |
| | | (iv) o repertório lexical compreende diversos temas e categorias. |

Fonte: Elaborado pela autora da pesquisa.

O Manuário Acadêmico e Escolar foi criado devido à necessidade de registrar e divulgar os sinais da Libras em dois contextos bem definidos: o Colégio de Aplicação e o Curso Bilíngue de Pedagogia do Instituto Nacional de Educação de Surdos (INES). Com a finalidade de contribuir com o fortalecimento da Libras, a coleta de sinais conta com uma equipe constituída por alunos, profissionais surdos e ouvintes do INES. Os sinais são divulgados no *site*, mencionado na tabela e estão organizados por área do conhecimento. Segundo os desenvolvedores, o objetivo é compartilhar com o público

uma rica produção lexical gerada em sala de aula. É válido dizer, também, que os organizadores apresentam como meta futura, a criação de um acervo sob forma de um dicionário bilíngue, acompanhado de verbetes, tanto na Libras quanto na Língua Portuguesa.

O Sinalário Disciplinar em Libras, criado pela Secretaria de Educação do Estado do Paraná (SEED-PR), é uma ferramenta desenvolvida para profissionais da educação, estudantes surdos, comunidade surda e demais interessados. É composto de sinais, datilologia das palavras, oralização e contextualização do termo apresentado no vídeo. Há no aplicativo, aproximadamente, 300 vídeos disponibilizados em Libras, com diversos termos referente as 13 disciplinas que compõem o currículo do Ensino Fundamental e do Ensino Médio, das quais: Filosofia, Sociologia, Ensino Religioso, Educação, Física, Ciências, Biologia, Artes, Química, Física, Matemática, Língua Portuguesa, Geografia e História.

Referente ao Sinalário de Libras I e II – Universidade Federal do Paraná (UFPR) – contém sinais criados pelos acadêmicos do curso de Letras Libras, para orientar as disciplinas de Libras I e II.

O Intitulo Phala – Centro de Desenvolvimento para Surdos se refere a uma Instituição sem fins lucrativos, fundada em 1999, por pais, familiares e profissionais na área da surdez. Seu objetivo é de oferecer atendimento à saúde, à educação, ao trabalho, à assistência social e à promoção de direitos e interesses, reivindicações e anseios das pessoas surdas de Itatiba e região. Lá são desenvolvidos vários projetos, dentre eles, um que vem ao encontro dos nossos interesses, referente a um *corpus* com aproximadamente 80 vídeos disponíveis na plataforma *YouTube*, com uma diversidade temática bastante interessante.

O Sinalário Técnico em Libras – Eletrônica com as mãos concerne a um conjunto de sinais voltados a vocabulários técnicos em Libras, relacionado ao Curso de Eletrônica do Instituto Federal da Bahia (*campus* de Salvador).

O Dicionário de Libras Biologia se refere a uma apresentação em vídeos, de termos relacionados à área da Biologia, desenvolvida pelo Grupo de Estudos de Pequenas Empresas e Empreendedorismo (EPEEM), que atua no setor metal-mecânico no estado do Paraná. O dicionário está disponível na plataforma *YouTube*, conforme consta na tabela.

O canal “O diário de Fiorella”, disponível no *YouTube*, trata de vídeos (amadores) referentes ao processo de aquisição da linguagem, de uma criança surda,

filha de pais surdos – um rico material que pode ser utilizado para análises sobre aquisição da linguagem.

No site da Universidade Federal de Pelotas – RS – (UFPEL), encontra-se o redirecionamento curso Básico de Libras, que compõe o Projeto Idiomas Sem Fronteira (ISF). Refere-se a um conjunto de vídeos em Libras, contendo diversas categorias temáticas e diálogos.

Como podemos notar, são vários os trabalhos realizados no Brasil, no entanto, o que está em pauta, nas próximas investigações, é se esses dados encontrados até o momento, são legíveis por máquina, e sendo legíveis, qual a possibilidade de estarem contribuindo com as análises linguísticas futuras?

De antemão, após percorrer a trajetória que compreende a LC, no capítulo anterior, esses bancos de dados, embora atendam aos objetivos aos quais se propõem e são úteis aos trabalhos desenvolvidos até o momento, não nos parece estarem de acordo com o que preconiza a LC. No entanto, a forma como serão manipulados futuramente, alguns deles, como por exemplo, “O diário de Fiorella”, disponível na plataforma *YouTube*, pode contribuir com trabalhos baseados em *corpora*, referentes à aquisição da linguagem, por se tratar de vídeos autênticos, onde a manifestação linguística acontece de forma natural. Para tanto, nas próximas seções, apresentaremos, também, algumas ferramentas computacionais que poderão auxiliar no desenvolvimento das pesquisas linguísticas baseadas em *corpus*.

Vale mencionar, que embora exista um amplo material em todo o território brasileiro, apresentados aqui apenas alguns deles, o acesso a essas informações nos parece bastante disperso, pois se encontram em diferentes plataformas, o que dificulta a sistematização dos dados.

Nesse contexto, na próxima seção, buscamos investigar como estão sendo desenvolvidos os grandes *corpora* em língua de sinais no Brasil e em algumas outras partes do mundo, com o intuito de sugerir o estabelecimento de um paralelo entre eles e considerar, minimamente, uma sistematização com base nos critérios estabelecidos pela LC, que contribua com os estudos linguísticos baseados em *corpora.*

## 4.2 INVENTÁRIO DESCRITIVO DE *CORPUS* EM LÍNGUA DE SINAIS

Após verificar as perspectivas dos trabalhos que já foram desenvolvidos no Brasil, nesta seção, buscaremos apresentar e descrever os *corpora* citados por

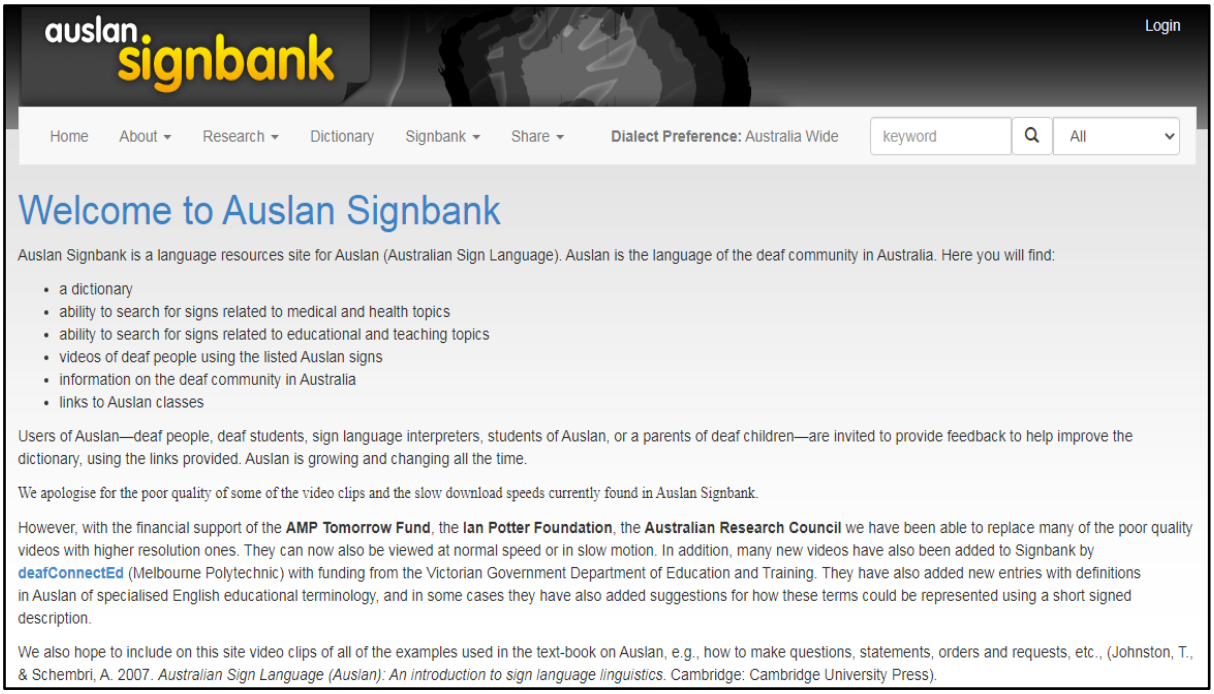
Quadros (2019) no início desse capítulo, pois aparentemente são *corpora* que estão sistematizados e condizem com os estudos estabelecidos pela LC. Portanto, nosso propósito é descrever e analisar as características que envolvem esses *corpora*, para na sequência organizar e propor as diretrizes para a constituição de um *corpus* que atenda ao objetivo principal dessa pesquisa. Nas próximas subseções, apresentamos um inventário descritivo mais detalhado de uma coleta realizada sobre os *corpora* nas línguas de sinais, em diversos países e no Brasil. São eles: o *corpus* da Língua de Sinais Australiana; o *corpus* de Língua de Sinais Britânica; o *corpus* da Língua de Sinais Alemã; o *corpus* da Língua de Sinais Holandesa; o *corpus* da Língua de Sinais Polonesa; e o *corpus* da Língua de Sinais Japonesa; o *corpus* da Libras.

### 4.2.1 *Corpus* da Língua de Sinais Australiana

O *Auslan Signbank*[14] é um site de recursos linguísticos para Auslan (Língua de Sinais Australiana), utilizado pela comunidade surda da Austrália. Contempla: (i) um dicionário; (ii) capacidade de pesquisar sinais relacionados a tópicos médicos e de saúde; (iii) capacidade de pesquisar sinais relacionados a tópicos educacionais e de ensino; (iv) vídeos de pessoas surdas usando os sinais de Auslan listados; (v) informações sobre a comunidade surda na Austrália; (vi) *links* para aulas de Auslan. Para ilustrar esses registros, na Figura 1, apresenta-se uma tela referente ao *Auslan Signbank*.

[14] Disponível em: http://www.auslan.org.au/about/corpus/. Acesso em: 22 set. 2020.

Figura 1 - Tela de representação do Auslan Signbank



Fonte: http://www.auslan.org.au/.

A Auslan é uma língua de sinais utilizada como meio de comunicação entre pessoas surdas, estudantes surdos, intérpretes de língua de sinais, alunos de Auslan ou pais de crianças surdas, que são convidados a fornecer um *feedback* para ajudar no desenvolvimento e na eficácia do dicionário. O *corpus* Auslan vem crescendo e mudando constantemente.

Para as questões referente à qualidade dos vídeos, conta-se com o apoio financeiro do *AMP Tomorrow Fund*, da *Lan Potter Foundation* e do *Australian Research Council*, também contam com a colaboração de adição de novos vídeos com a *DeafConnect Ed* (*Melbourne Polytechnic*), com financiamento do Departamento de Educação e Treinamento do Governo de Victoria. Além disso, acrescentaram novas entradas com definições em Auslan de terminologia educacional especializada em inglês e, em alguns casos, também há sugestões de como esses termos podem ser representados e, para isso, usa-se uma breve descrição.

O Auslan *Corpus* contém vídeos no Arquivo Auslan, vinculados com arquivos de anotação linguísticas. Em meados de 2008, esses arquivos foram depositados no *Endangered Languages Archive* (ELAR), e estão acessíveis desde 2012. Foi financiado por Trevor Johnston do *The Hans Rausing Endangered Languages Project*, com os propósitos de criar e garantir um arquivo de referência de Auslan, pois,

devido à redução de usuários surdos de língua de sinais, tornou-se uma língua ameaçada de extinção. Do mesmo modo, os pesquisadores objetivam criar um *corpus* linguístico, ou seja, uma coleção de textos e gravações em um idioma com as transcrições, que poderão colaborar com outros pesquisadores e alunos no processo de aprendizagem da língua.

A criação desse arquivo, conforme informações dos desenvolvedores, envolve gravações, comparações e descrições de um conjunto de amostras de sinais naturais, que são controlados e realizados por aproximadamente 100 falantes surdos, que usam essa língua desde a mais tenra idade, entre seus pares linguísticos em toda a Austrália. Um dos critérios para as gravações foi a organização da atividade em seções, cada uma com duração de três horas e dois participantes, totalizando 300 horas de vídeo. Outro critério utilizado foi a inclusão de informações etiquetadas, no arquivo, por um *software* de anotação multimídia, a fim de torná-lo um *corpus* legível por máquina. As gravações foram baseadas em entrevistas, produções de narrativas, pesquisas, conversações livres e outras respostas linguísticas motivadas por vários estímulos. Para tanto, as respectivas filmagens foram editadas com recortes específicos para o desenvolvimento de cada uma dessas atividades, assim, o arquivo consiste em vídeos com informações (metadados).

O *Corpus* Auslan contém dados de pesquisa de vídeos coletados como parte do projeto de variações sociolinguística em Auslan, financiado pelo Conselho de Pesquisa Australiano e pelo Instituto Real para Crianças Surdas e Cegas (# LP0346973), conduzido por Adam Schembri e Trevor Johnston (2003-2005). De 2008 a 2010, Adam Schembri liderou o projeto que criou o *Corpus British Sign Language* (BSL). Esse *corpus* conta com um arquivo de aproximadamente 357 vídeos em Auslan. Os arquivos são anotados e consideram os seguintes tipos de anotações: (i) identificador e IDglosa somente de substantivos e verbos; (ii) frequência de sinal e IDglosa para todos os sinais; (iii) marcação da classe gramatical que determinado item lexical pertence; identificação da direção do olhar no espaço enunciador; (iv) identificação da orientação da palma da mão na realização dos sinais; (v) identificação dos limites da sentença; (vi) identificação de argumentos verbais; (vii) marcação de argumentos verbais para funções macro e funções semânticas; (viii) marcação para a presença ou ausência de mudança no espaço; (ix) identificação de períodos de ação construída ('mudança de papel'); (x) tradução livre; e (xi) tradução literal.

Conforme afirmam os proponentes do *corpus*, incorre um longo tempo para a execução e a implementação do trabalho de anotações dos vídeos, além de que, os custos são onerosos, portanto, o *corpus* Auslan, assim como os próprios desenvolvedores afirmam, ainda levará um tempo para que seja suficientemente rico em anotações. Ressalta-se, ainda, que os arquivos existentes não estão disponíveis publicamente, e isso os inviabiliza em relação ao acesso e ao aprofundamento nas pesquisas acerca desse assunto.

Para entender melhor como funciona o acesso aos dados, o projeto fornece metadados relevantes em todas as mídias e arquivos de anotação. Esses metadados são armazenados em um arquivo de banco de dados, em campos que seguem as diretrizes de metadados IMDI[15], portanto incluem chaves para metadados de língua de sinais. Assim como o ELAN, ele foi desenvolvido no Instituto *Max Planck* de Psicolinguística em Nijmegen.

É relevante mencionar que esses metadados, primeiramente, são transferidos para o *software* IMDI, e somente após esse procedimento, o arquivo e o *corpus* ficam disponibilizados para pesquisas. Os metadados são compostos por informações sobre os vídeos/mídia, arquivos de anotação e consideram os seguintes aspectos: (i) ator (região, sexo, idade, educação, etc.); (ii) conteúdo (várias tarefas de linguagem, materiais usados como); (iii) mídia (formato e tipo); (iv) projeto (nome, idioma, metodologia); e (v) sessão (nome da tarefa, participantes, dentre outros); e (vi) recursos escritos relacionados (existência de um arquivo de anotação para um arquivo de mídia e o tipo de anotação concluída, se houver).

Essa base de dados apresenta uma proposta significativa no que diz respeito ao cruzamento de dados de duas ferramentas bastante eficazes em relação à construção de *corpus* em línguas de sinais. A combinação dos recursos de pesquisas contidos no ELAN com os critérios de metadados IMDI, de acordo com os organizadores/idealizadores, tem em vista fazer extensas investigações qualitativas e quantitativas de grandes amostras de mídia em Auslan. Desse modo, a convergência entre dados e metadados possibilita as evidências de construções gramaticais comuns ou regulares, variação sociolinguística e mudança lexical e gramatical. Tais possibilidades são características inovadoras para as análises linguísticas nas línguas de sinais.

[15] IMDI significa ISLE, ou seja, iniciativa de Dados Meta e é um padrão usado na descrição de *corpora* de linguagem.

### 4.2.2 *Corpus* de Língua de Sinais Britânica

O projeto de *corpus* de Língua de Sinais Britânica[16] (BSL) é produzido por sinalizantes surdos, cujas informações são apresentadas por um conjunto de vídeos que mostram pessoas surdas usando BSL, contém informações básicas sobre os sinalizates e descrições por escrito dos sinais (em trilhas ou etiquetas) no ELAN. Os vídeos foram coletados como parte do Projeto BSL *Corpus* original, financiado entre 2008 e 2011 pelo Conselho de Pesquisa Econômica e Social. O BSL *Corpus* é baseado no Centro de Pesquisa em Cognição e Linguagem em Surdez, *University College London*. Também, incluiu pesquisadores da *Bangor University* (País de Gales), *Heriot-Watt University* (Escócia), *Queens University Belfast* (Irlanda do Norte) e da *Universidade de Bristol* (Inglaterra).

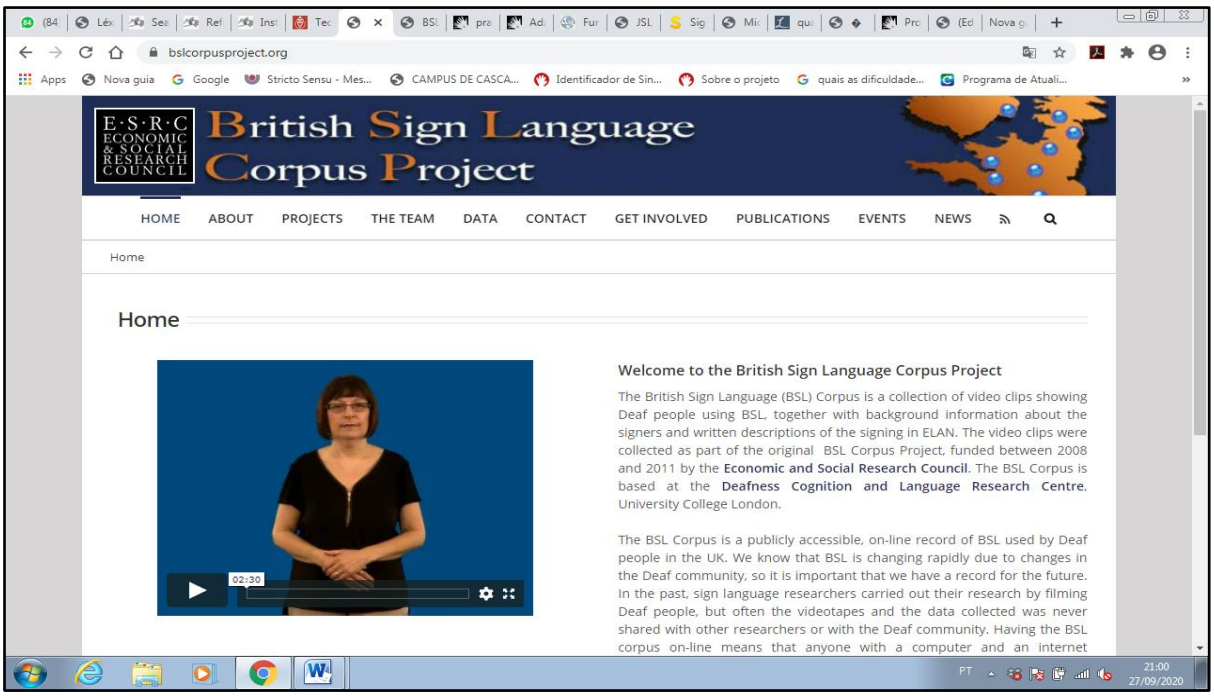
Por conseguinte, o BSL *Corpus* é um registro *online* de acesso público ao BSL, usado por pessoas surdas do Reino Unido. Menciona-se que, no passado, os vídeos e dados coletados de pessoas surdas em contexto comunicativo, raramente eram compartilhados, por isso o *corpus* vem colaborar com as trocas de informações entre os pesquisadores de língua de sinais em universidades e na comunidade surda. Além disso, o *corpus* é útil para a compreensão da estrutura e do uso da BSL, e importante para a educação de crianças surdas, para o treinamento de intérpretes e professores de BSL.

O projeto BSL *Corpus* original foi finalizado em 2011, sendo que os dados de vídeo e algumas anotações estão disponíveis, porém, não está concluído. Destaca-se que o projeto tinha dois objetivos principais: primeiro, criar um *corpus* BSL, ou seja, uma coleção de vídeos na internet, a fim de analisar pessoas surdas num contexto comunicativo da BSL, constando informações sobre os signatários (os sinalizadores) e as descrições anotadas (transcritas) em trilhas no ELAN; segundo, e não menos importante, utilizar esses dados para realizar pesquisas gramaticais, de vocabulários da BSL e das variações sociolinguísticas e mudanças de vocabulários que vem ocorrendo nessa língua. Verificou-se, ainda, que um segundo estudo foi desenvolvido com o intuito de descobrir quais são os sinais mais frequentes utilizados nas conversas em BSL. Para isso, realizou-se a análise de um conjunto de 25.000 sinais.

[16] Disponível em: https://bslcorpusproject.org/. Acesso em: 23 set. 2020.

É relevante mencionar, que a equipe de pesquisa do BSL *Corpus Project* filmou 249 surdos de oito cidades do Reino Unido (Londres, Bristol, Birmingham, Manchester, Newcastle, Glasgow, Cardiff e Belfast). Foram filmadas de 30 a mais pessoas, em cada uma das cidades mencionas. Esses vídeos incluíram homens e mulheres, adultos com pais surdos ou com pais ouvintes, sinalizantes considerados jovens e idosos, além de contemplar pessoas surdas com diferentes funções profissionais e de diversas etnias. A maior parte dos participantes atendeu a dois critérios do projeto, quais sejam: (i) ter aprendido a BSL antes dos sete anos de idade e (ii) ter residido na mesma cidade nos últimos 10 anos ou mais. Foram filmadas 249 pessoas em conversas com outra pessoa surda, por meio de entrevistas, contando histórias e mostrando os sinais que utilizavam para 102 conceitos-chave. Na Figura 2, pode-se observar como está disposto o *British Sign Language*.

Figura 2 - Tela de representação do British SignLanguage Corpus Project




Fonte: https://bslcorpusproject.org/.

Vale destacar, também, que ao explorar a plataforma desenvolvida, verificou-se que os criadores do projeto de *corpus* consideram que os trabalhos em BSL são necessários, pois quando se compara, por exemplo, ao o *British National Corpus of English*[17], que é uma amostra de um *corpus* grande, representativo, acessível e, o

[17] Disponível em: http://www.natcorp.ox.ac.uk/. Acesso em: 23 set. 2020.

mais importante, legível por máquina, na língua inglesa (oral), constata-se que o conjunto de dados referentes ao BSL *Corpus* ainda está em construção, sendo assim, é imprescindível que constem as anotações e as traduções para torná-lo legível por máquina e, consequentemente, um verdadeiro *corpus* da língua se sinais. Esse *corpus*, com as devidas anotações e traduções, além de colaborar significativamente com pesquisas sobre a estrutura e o uso da BSL, na formação de professores de BSL, intérpretes de língua de sinais e educadores de crianças surdas, também contribuirá com estudos comparativos entre a BSL e as línguas faladas, relacionadas ou não relacionadas em outros lugares do mundo.

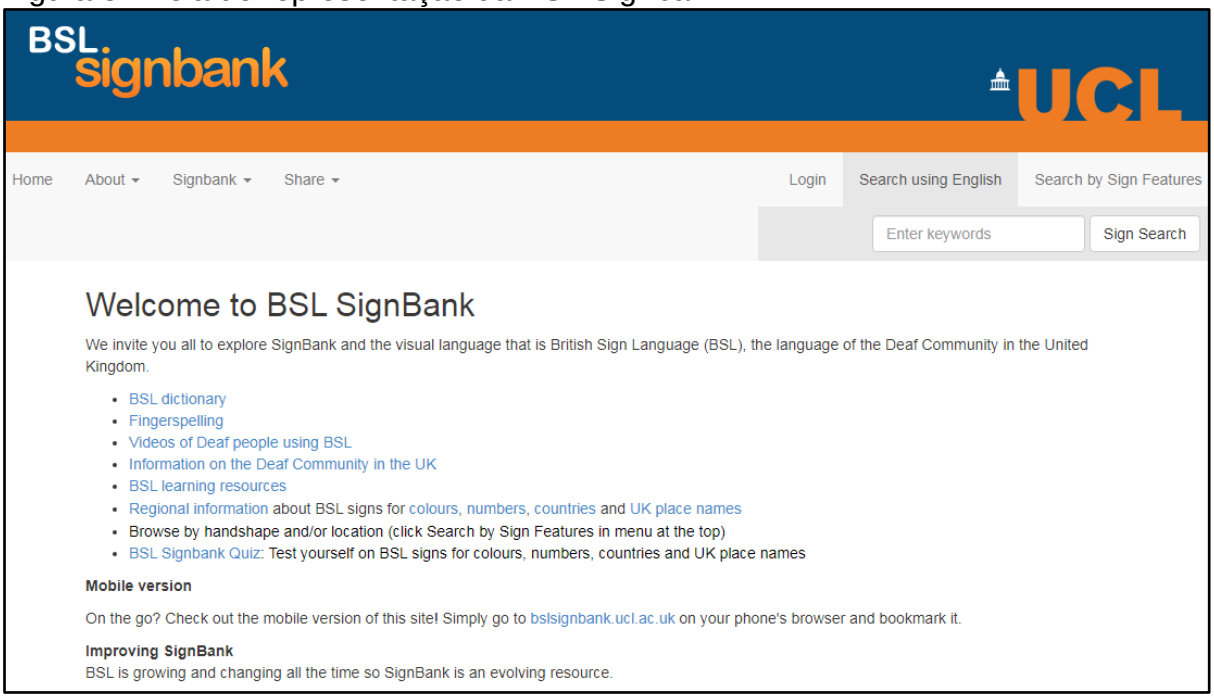
Por todos os aspectos mencionados, os desenvolvedores afirmam que a criação de *corpus* de línguas de sinais é o caminho para o futuro e que projetos semelhantes estão sendo desenvolvidos em países, como: Austrália, Brasil, Irlanda, Itália, França, Alemanha, Grécia, Holanda, Espanha e Suécia. Destaca-se que, alguns desses estudos são abordados nesta pesquisa.

A fim de descrever um pouco sobre os estudos realizados nesse âmbito, enfatiza-se que os estudos do BSL *Corpus* Project original (2008-2011) foram sobre variação fonológica, ou seja, estudos sobre as configurações de mãos; variação lexical, como ocorre nos sinais para cores, países, números e nomes de lugares no Reino Unido; além de pesquisas voltadas a frequência lexical (embasado em 25.000 etiquetas de sinais em conversas). Dentre os projetos que vem sendo desenvolvidos, destaca-se a criação de um banco de dados como parte de um estudo sobre frequência lexical, que documenta 50.000 sinais de quatro regiões diferentes (Bristol, Birmingham, Londres e Manchester), a partir de dados do BSL *Corpus* que foi transformado no BSL *SignBank*, entre os anos 2011 e 2015, disponível a partir de setembro de 2014, e que continua a evoluir até os dias atuais, conforme as pesquisas realizadas sobre a BSL.

O BSL *SignBank* contempla a língua de sinais da comunidade surda no Reino Unido. Na plataforma é possível consultar os seguintes conteúdos: (i) dicionário BSL; (ii) soletração (datilologia) das palavras; (iii) vídeos de pessoas surdas se comunicando em BSL; (iv) Informações sobre a comunidade surda do Reino Unido; (v) recursos de aprendizagem BSL; (vi) informações regionais sobre sinais BSL, além de cores, números, países e nomes de lugares no Reino Unido; (vii) possibilidade de navegação por configuração de mão e/ou localização (na parte superior do menu consta a opção de pesquisa por recurso de sinal); e há ainda (viii) ícone de jogo (Quiz),

que faz testes de sinais da BSL para as cores, números, países e nomes de lugares no Reino Unido. Na Figura 3, apresenta-se a plataforma BSL *Signbank*.

Figura 3 - Tela de representação da BSL Signbank



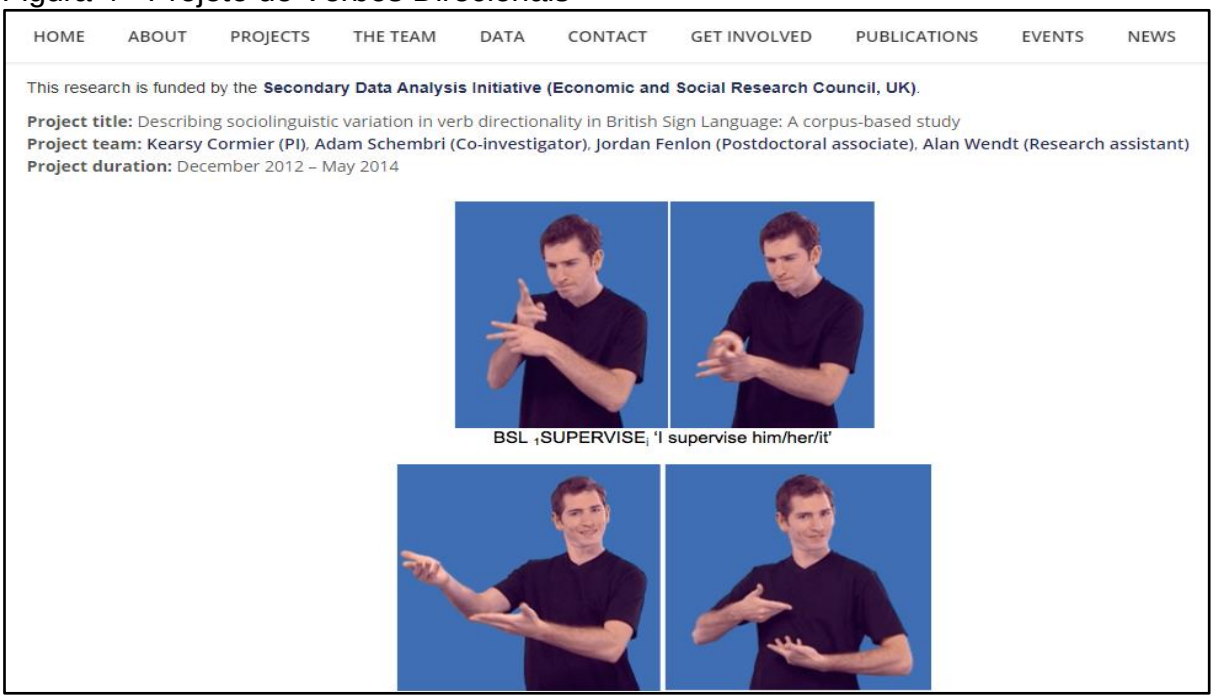

Fonte: https://bslsignbank.ucl.ac.uk/.

Apesar de conter vários recursos relevantes, na plataforma constam informações que o *SignBank* vem evoluindo constantemente, além disso, encontra-se um ícone, onde é possível que pessoas da comunidade surda contribuam com suas experiências e opiniões, configurando-o assim como um *corpus* colaborativo. Há grande incentivo para que esse *feedback* ocorra, pois, destaca-se que é para melhoria tanto do *Signbank* quando do BSL *Corpus*, e que essas ações são fundamentais para manter o acesso gratuito de ambos. É válido ressaltar, também, que outros projetos foram realizados no BSL *Corpus*, dentre eles, apresentaremos seis: (i) o Projeto de Verbos Direcionais; (ii) o Projeto *Digging into Signs*; (iii) o Projeto de Sintaxe BSL; (iv) o Projeto de Atitudes de Linguagem; (v) o Projeto ExTOL; e (vi) o Projeto de Promulgação.

O Projeto de Verbos Direcionais[18] (2012-2014) objetiva usar dados para investigar sobre a variação e a mudança no uso de verbos direcionais na BSL, conforme ilustrado na Figura 4.

[18] Disponível em: https://bslcorpusproject.org/projects/directional-verbs-project/. Acesso em: 23 set. 2020.

Figura 4 - Projeto de Verbos Direcionais

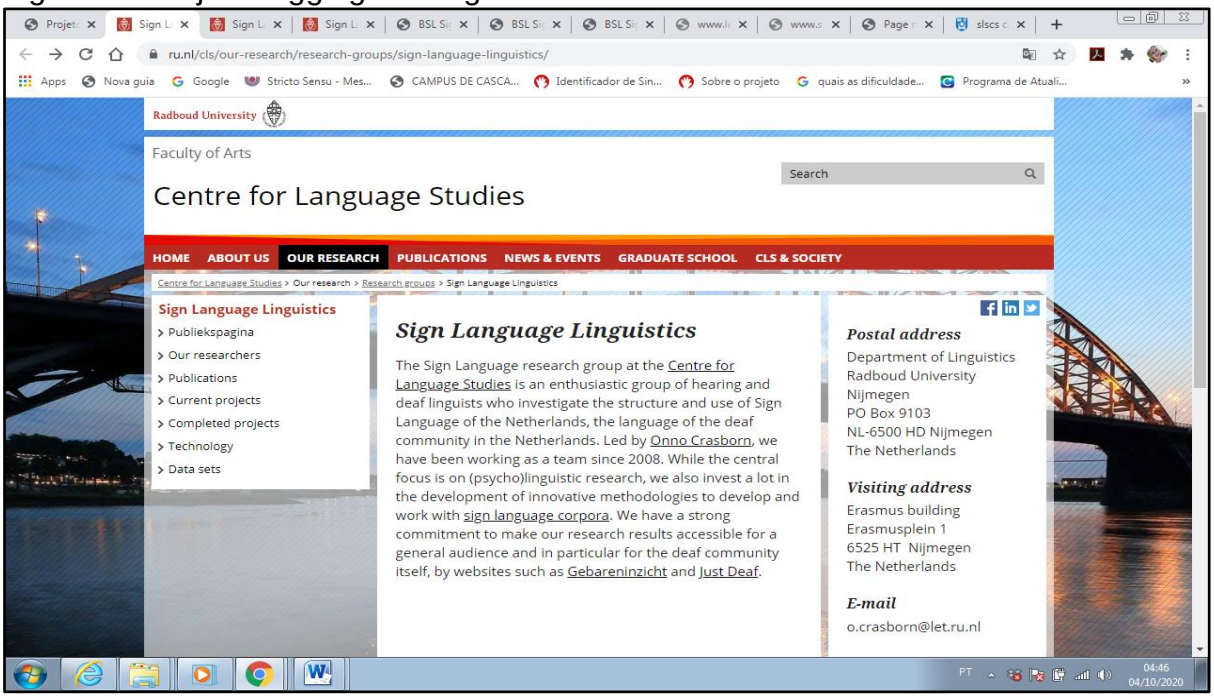



Fonte: https://bslcorpusproject.org/projects/directional-verbs-project/.

O Projeto *Digging into Signs*[19] (2014-2015) propõe, junto a uma equipe liderada por Onno Crasborn[20], desenvolver padrões de anotação entre *corpus* para dados das línguas de sinais, utilizando-se do BSL *Corpus* no Reino Unido e *Corpus* NGT na Holanda, além de aperfeiçoar as ferramentas de *softwares* para trabalhar com corpora de língua de sinais, conforme apresentado na Figura 5.

[19] Disponível em: https://www.ru.nl/cls/our-research/research-groups/sign-language-linguistics/. Acesso em: 23 set. 2020.
[20] Disponível em: https://www.ru.nl/personen/crasborn-o/. Acesso em: 23 set. 2020.

Figura 5 - Projeto Digging into Signs




*Fonte: https://www.ru.nl/cls/our-research/research-groups/sign-language-linguistics/.*

O Projeto *de* Sintaxe BSL[21] (2016-2020) tem a finalidade de documentar e descrever a ordem das palavras e os recursos não manuais em diversos tipos de frases na BSL. Em resumo, o *corpus* em questão tem em vista a condução de uma investigação detalhada do sistema gramatical BSL, com conversações espontâneas e abordagens fundamentadas em *corpus*, com base em teorias cognitivas/funcionais da gramática e das teorias sociolinguísticas. De forma genérica, esse projeto visa uma abordagem multidisciplinar ao que se refere à gramática da língua de sinais embasada em disciplinas relacionadas, novos métodos e tecnologias;

O Projeto de Atitudes de Linguagem[22] (2017-2019) tem a finalidade de estudar as atitudes linguísticas e a consciência da língua na comunidade surda britânica, para tanto, foram realizadas várias perguntas e analisadas as diferenças e as semelhanças existentes entre diferentes grupos sociais dentro da comunidade surda, por exemplo, em relação à idade ou ao gênero. Destaca-se, assim, que essas análises, de acordo com os pesquisadores, parecem ser indispensáveis não somente para estudos linguísticos futuros, mas também para um planejamento e uma política linguística,

[21] Disponível em: https://bslcorpusproject.org/projects/bsl-syntax-project/. Acesso em: 23 set. 2020.
[22] Disponível em: https://bslcorpusproject.org/projects/language-attitudes-project/. Acesso em: 23 set. 2020.

considerando o *status* legal da BSL no Reino Unido. Além de todos esses fatores, vale ressaltar que será um recurso valioso aos professores, uma vez que se propõe a explorar a atitude da comunidade surda em relação à BSL e ao inglês.

O Projeto ExTOL[23] (2018-2021) é financiado pelo EPSRC (EP/RO3298X/1), junto ao Centro de Visão, Fala e Processamento de Sinais da Universidade de Surrey, com o Grupo de Geometria Visual da Universidade de Oxford e o Centro de Pesquisa em Cognição e Linguagem de Surdez da UCL. ExTOL significa tradução de ponta a ponta da língua de sinais britânica, seu objetivo é obter dados da BSL *Corpus* e de outras fontes para construir um sistema capaz de observar sinais humanos e transformá-los em inglês escrito. Esse fator indica ser uma grande inovação mundial, pois os estudos linguísticos nas línguas de sinais começaram apenas na década de 1960, o que mostra ser bastante novo quando comparado às línguas orais-auditivas.

Assim, esse projeto visa construir ferramentas computacionais, que sejam capazes de identificar o movimento, a forma da mão, as expressões faciais e a postura corporal do sinalizante. Seus idealizadores também pretendem que essas ferramentas possam compreender como esses aspectos são agrupados em frases e de que forma serão traduzidos para a língua escrita ou oral. Conforme mencionado pelos desenvolvedores, a dificuldade em se estudar as línguas de sinais são justamente essas, no que se refere à necessidade de análises de imagens e vídeos, considerando, ainda, que essas línguas não possuem um sistema de escrita ou transcrição padrão, o que torna o trabalho mais complexo. Nessa perspectiva, mencionam que houve muitos avanços no reconhecimento de língua de sinais, citam as luvas dados e o sistema de captura de movimento como o *Kinect*[24], no entanto há pouco conhecimento dos cientistas da computação a respeito de como funcionam as línguas de sinais. Assim, especialistas da língua de sinais britânica, juntamente com engenheiros de *softwares* especializados em visão computacional e aprendizado de máquinas, buscam construir o primeiro sistema de tradução automática, que poderá ser funcional para qualquer outra língua de sinais, o que representa um grande marco ao que tange à comunicação entre surdos e ouvintes.

[23] Disponível em: https://cvssp.org/projects/extol/. Acesso em: 23 set. 2020.
[24] *Kinect* é um sensor de movimentos desenvolvido exclusivamente para os *consoles 360* **e** *XboxOne***,** *Xboxambos* da *Microsoft*. Disponível em: https://www.significados.com.br/kinect/#:~:text=O%20que%20%C3%A9%20Kinect%3A,controles%20ou%20joysticks%20para%20jogar. Acesso em: 24 set. 2020.

O Projeto de Promulgação[25] (2019-2020) visa investigar como as pessoas surdas reproduzem as ações, os enunciados, os pensamentos e os sentimentos sobre si mesmas, de outras pessoas, animais ou coisas, baseados em dados narrativos (encenações) pessoais, disponíveis no BSL *Corpus*. Para um melhor entendimento, cita-se como exemplo, a utilização de sinais equivalentes às palavras das línguas orais. Esses estudos irão colaborar com pesquisas sobre a estrutura linguística e o uso da BSL, além possibilitar uma referência em programas bilíngues de inglês/BSL para crianças surdas e adultos surdos ou ouvintes que estão em processo de aprendizagem da BSL.

### 4.2.3 *Corpus* da Língua de Sinais Alemã

O projeto *DGS-Korpus*[26] se refere à documentação e à pesquisa sobre a Língua de Sinais Alemã (DGS). É um projeto da Academia de Ciências de Hamburgo[27], desenvolvido no Instituto de Língua de Sinais Alemã e Comunicação de Surdos[28], da Universidade de Hamburgo[29], financiado pelo grupo de pesquisa do Governo Federal Alemão e dos Estados Federais no Programa de Academias e pelo Ministério Federal de Educação e Pesquisa. Na Figura 6, apresenta-se uma imagem do *DGS-Korpus*.

[25] Disponível em: https://bslcorpusproject.org/projects/enactment-project/. Acesso em: 24 set. 2020.
[26] Disponível em: https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html. Acesso em: 24 set. 2020.
[27] Disponível em: https://www.awhamburg.de/. Acesso em: 24 set. 2020.
[28] Disponível em: https://www.idgs.uni-hamburg.de/. Acesso em: 24 set. 2020.
[29] Disponível em: https://www.uni-hamburg.de/en/. Acesso em: 24 set. 2020.

Figura 6 - Tela de representação do DGS-Korpus


Fonte: https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html.

De acordo com os seus desenvolvedores, o principal objetivo desse projeto é a construção de um *corpus* representativo para a língua de sinais em uso no cotidiano, por pessoas surdas de todo o território alemão, parte da produção é apresentada como um *Corpus* Público da DGS. Afirma-se, ainda, que foram filmados 330 informantes aos pares para a coleta de dados, a pesquisa é realizada por uma equipe composta por membros surdos e ouvintes do Instituto de Língua de Sinais Alemã e Comunicação de Surdos da Universidade de Hamburgo. O grupo é responsável pela análise dos dados coletados com o propósito de torná-los acessíveis aos interessados em consultas aleatórias ou pesquisas da língua de sinais alemã.

A fim de atender ao objetivo proposto, o trabalho foi organizado em duas etapas, com um conjunto de aproximadamente 50 horas de dados publicados no *Public DGS Corpus*. (1) Coleta de dados da DGS por pessoas surdas, cuja finalidade seria torná-los acessíveis em um *corpus* anotado. Para tanto, essa etapa foi disponibilizada em dois portais distintos: (i) O *meine-dgs.de*[30]*:* destinado a membros da comunidade surda e demais interessados no DGS, com vistas a atender às

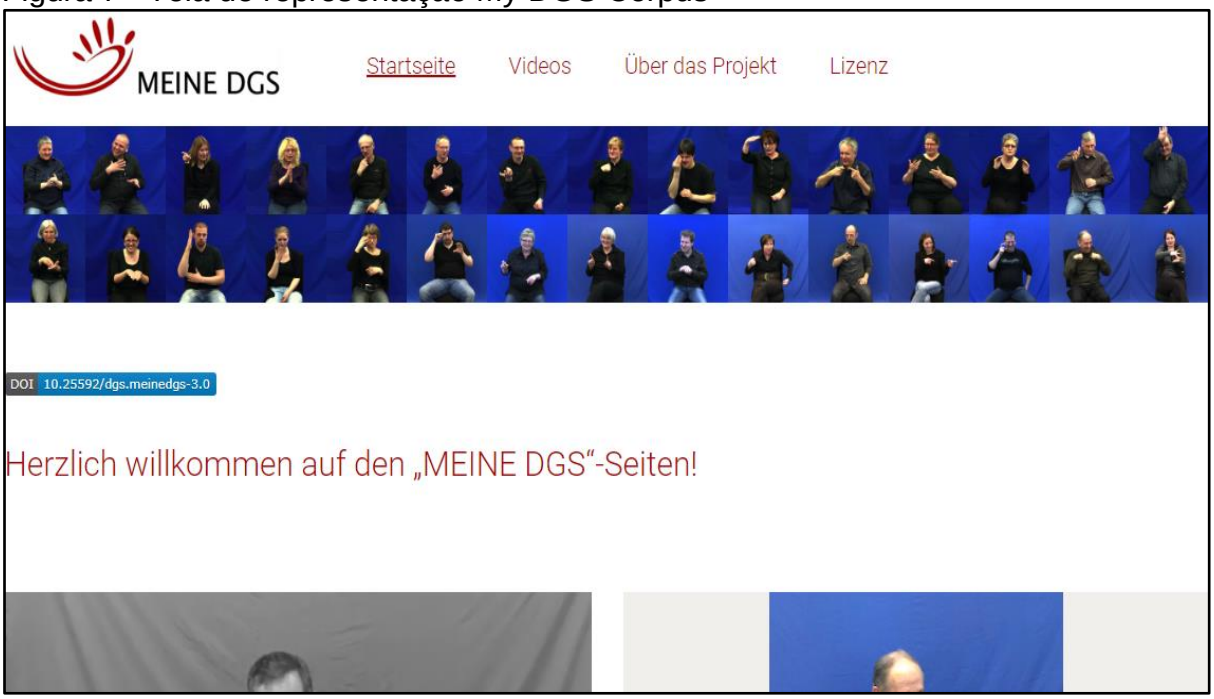
[30] Disponível em: https://www.sign-lang.uni-hamburg.de/meinedgs/overview/start.html. Acesso em: 24 set. 2020.

diversas necessidades de informação; (ii) O portal *ling.meine-dgs.de*[31]: contém dados de anotações detalhados e permite análises, por linguistas, dos diferentes aspectos do DGS em uma base empírica. (2) O desenvolvimento de um dicionário digital baseado em *corpus* (DW-DGS: "*Digitales Wörterbuch der Deutschen Gebärdensprache – Das korpusbasierte Wörterbuch DGS – Deutsch*")[32]. Esse dicionário contém descrições de sinais DGS, que são analisadas por meio dos itens lexicais em entradas que incluem informações linguísticas e extralinguísticas acerca de cada sinal. Constam informações que a versão final do dicionário estará disponível eletronicamente ao público em 2023 e será gratuita. Os verbetes são publicados periodicamente na plataforma.

Em linhas gerais, o *Corpus* DGS contém uma variedade de narrativas que, de um ponto de vista cultural, é de interesse para a comunidade surda (surdos, intérpretes, educadores e professores, pais e outras pessoas com interesse em Língua de Sinais Alemã). Assim, contém narrações e conversas em forma de vídeos DGS e suas respectivas traduções (legendas), que podem ser acionadas ou não pelos seus usuários, conforme a sua necessidade. Além disso, conta com uma coleção adicional de piadas. Na Figura 7, apresenta-se o portal *My DGS*.

[31] Disponível em: https://www.sign-lang.uni-hamburg.de/meinedgs/ling/start_en.html. Acesso em: 24 set. 2020.
[32] Disponível em: https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/dictionary.html. Acesso em: 25 set. 2020.

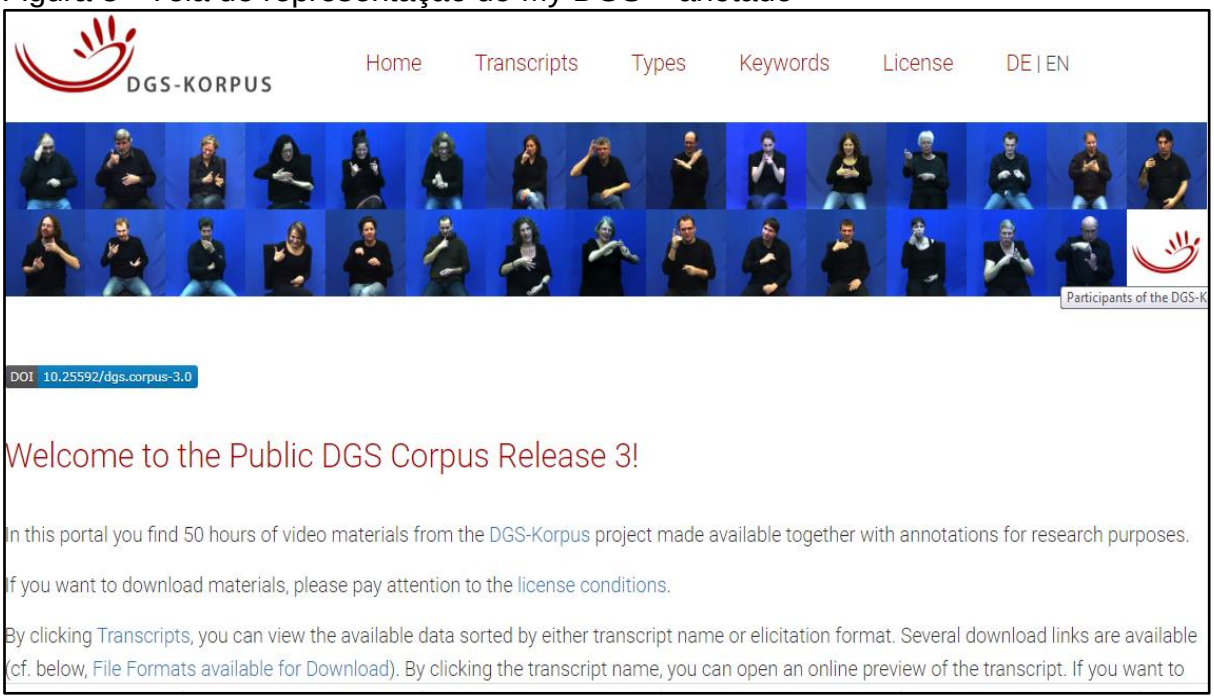
Figura 7 - Tela de representação My DGS Corpus




Fonte: https://www.sign-lang.uni-hamburg.de/meinedgs/overview/start.html.

O *MY DGS – anotado* é um portal de livre acesso e sem a necessidade de registro adicional por parte dos usuários. Oferece mais de 100 minutos de vídeos em DGS, com diferentes atividades na obtenção de dados (recontagem de histórias e narrações, conversas sobre o tema de sinais de trânsito, marcação de compromissos, descrições de sequências operacionais e curso de ação, com instruções). Todos esses vídeos, exceto os de piadas, que são apenas traduzidos, contêm anotações e podem ser consultados. As anotações incluem traduções, glosas e verbalizações. Esse *corpus* possibilita aos linguistas análises de diversos aspectos da DGS, numa base empírica. É relevante mencionar, ainda, que os vídeos e os arquivos de anotações, juntamente com as traduções, podem ser baixados pelos usuários. Os desenvolvedores esclarecem que apenas com a exceção dos verbetes anotados, disponíveis apenas em alemão, os *sites*, bem como todos os textos escritos estão na língua alemã e inglesa. Vale ressaltar, que em relação aos vídeos de piadas não constam informações sobre o motivo de não apresentarem anotações. Na Figura 8, ilustra-se o portal *My DGS – anotado*.

Figura 8 - Tela de representação do My DGS – anotado




Fonte: https://www.sign-lang.uni-hamburg.de/meinedgs/overview/start.html.

No que tange ao dicionário, constam informações na plataforma que se referem ao primeiro dicionário de DGS, baseado em *corpus* abrangente, com uma coletânea de dados de todo o território Alemão e com informações acessíveis, tanto em DGS quando na língua alemã. Nele contém registros de sinais em contexto comunicativo, desse modo, beneficiará vários grupos de usuários dessa língua, dentre eles: estudantes de DGS falantes de alemão como primeira língua; pais de crianças surdas; pessoas que ficaram surdas com idade pós-linguística; ouvintes que trabalham com surdos; intérpretes de língua de sinais; surdos nativos e estudiosos da teoria e estrutura da DGS (professores de língua de sinais ou linguistas).

Um dos diferenciais desse *corpus* é relativo à pesquisa corporativa de língua de sinais, nesse sentido possibilita a associação de outros projetos de *corpus* de língua de sinais em andamento e concluídos. Também conta com o apoio da comunidade linguística e incentiva a participação do maior número de pessoas surdas possíveis, para compartilharem seus conhecimentos, por meio do sistema *DGS-Feedback*.

### 4.2.4 *Corpus* da Língua de Sinais Holandesa

O *Corpus* da Língua de Sinais Holandesa[33] (NGT) é um conjunto de vídeos gravados de diversas histórias e conversas entre surdos, *online* e com acesso livre a filmes com anotações da Língua de Sinais da Holanda (abreviado como SLN ou NGT). Foi desenvolvido pelo grupo de língua de sinais, no departamento de Linguística, executado por Onno Crasborn, Inge Zwitserlood e Johan Ros da Radboud University[34], em Nijmegen. O projeto é financiado pela Organização Holandesa de Pesquisa Científica (NWO)[35]. O *corpus* foi gravado em maio de 2006 e concluído em 2008. Para a constituição desse *corpus*, pessoas surdas e ouvintes trabalharam juntas, com o propósito de garantir a sua construção com qualidade e o mais amplo possível.

Para realizar as gravações, várias pessoas surdas de todo o território holandês, que têm a NGT como primeira língua, foram convidadas a participar. Os pesquisadores destacam que objetivo era ter aproximadamente 100 participantes, dentre eles homens e mulheres, jovens e idosos. Foram realizadas várias atividades nas gravações, por exemplo, recontar histórias em quadrinhos e *videoclipes*; narrar um evento e discutir os temas específicos. Cada vídeo foi separado e a maioria está disponível na *internet*, exceto algumas gravações, em que os participantes não concederam permissão, e estão acessíveis apenas a pesquisadores. Na Figura 9, apresenta-se uma tela de representação do *corpus* NGT.

[33] Disponível em: http://www.ru.nl/corpusngtuk/. Acesso em: 25 set. de 2020.
[34] Disponível em: https://www.ru.nl/. Acesso em: 25 set. de 2020.
[35] Disponível em: https://www.nwo.nl/. Acesso em: 25 set. de 2020.

Figura 9 - Tela de representação Corpus NGT


Fonte: https://www.ru.nl/corpusngt/.

Nesse *site* é possível encontrar várias informações, dentre elas, instruções de como usar o *corpus,* com manuais para auxiliar na utilização das ferramentas, que inclui um guia para navegação no arquivo[36] e outra guia para as pesquisas de metadados IMDI[37]. Menciona-se, também, que é possível baixar o arquivo e visualizá-lo no navegador *web*, pelo recurso (arquivo de filme MPEG, arquivo de anotação EAF), ou também usar o *software* de metadados *Arbil*[38], para baixar dados maiores e importar para um *corpus* local do consulente.

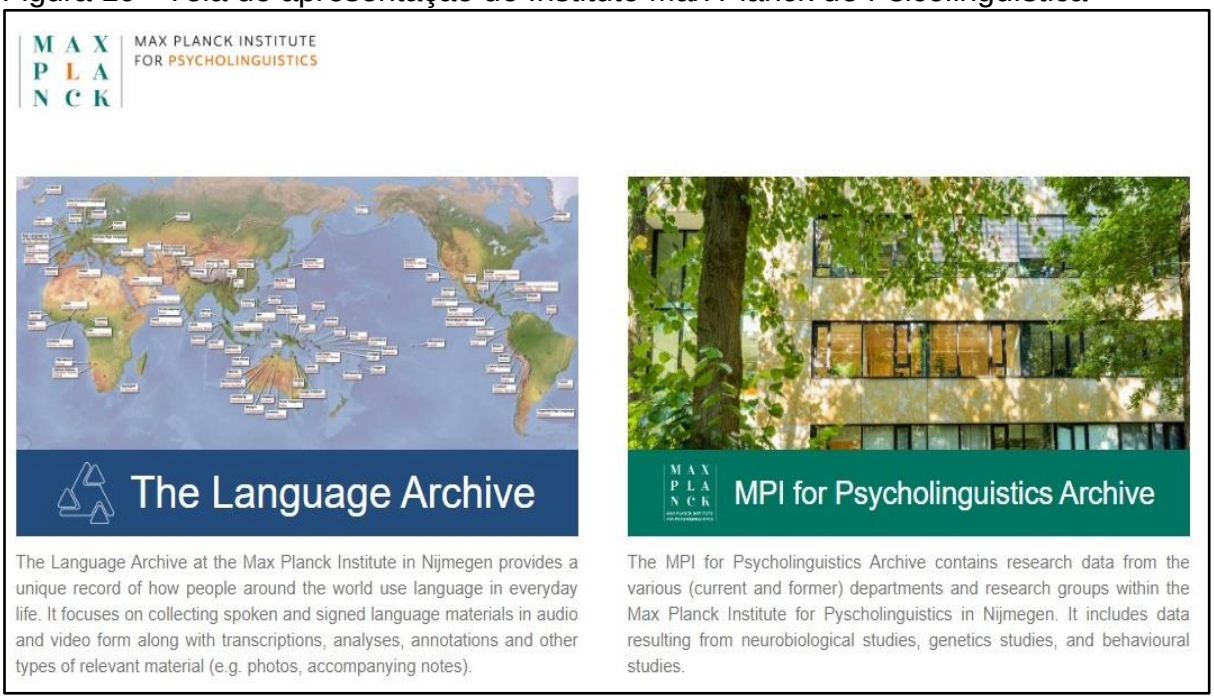
Destaca-se que é possível acessar todos os dados desse *corpus* no Instituto *Max Planck* de Psicolinguística[39], em Nijmegen. O acesso real ocorre por meio do navegador de *corpus online* IMDI[40]. No que se refere a esse instituto, constam informações que é o único dedicado inteiramente aos estudos psicolinguísticos, ou

[36] Disponível em: https://www.mpi.nl/corpus/a4guides/a4-guide-archive-browsing.pdf. Acesso em: 25 set. de 2020.
[37] Disponível em: https://www.mpi.nl/corpus/a4guides/a4-guide-metadata-search.pdf. Acesso em: 25 set. de 2020.
[38] Disponível em: https://archive.mpi.nl/forums/t/arbil-information-manuals-download/1045. Acesso em: 25 set. de 2020.
[39] Disponível em: https://www.mpi.nl/. Acesso em: 25 set. de 2020.
[40] Disponível em: https://archive.mpi.nl/?openpath=MPI319374%23. Acesso em: 25 set. de 2020.

seja, busca compreender de que forma produzimos e entendemos a língua e como ocorre a aquisição dela, seja com alunos de primeira ou de segunda língua.

Os desenvolvedores ressaltam, ainda, que esse instituto possui ampla experiência no que se refere a grandes bancos de dados linguísticos, nele há arquivos com registros de como as pessoas utilizam a língua no dia a dia. Concentram-se materiais, tanto das línguas orais quando das línguas sinais, em formato de áudio e vídeo, com transcrições, análises, anotações e outros elementos, como fotos, notas de acompanhamento, entre outras. A fim de ilustrar a plataforma que hospeda o *corpus* NGT, segue a representação na Figura 10.

Figura 10 - Tela de apresentação do Instituto Max Planck de Psicolinguística



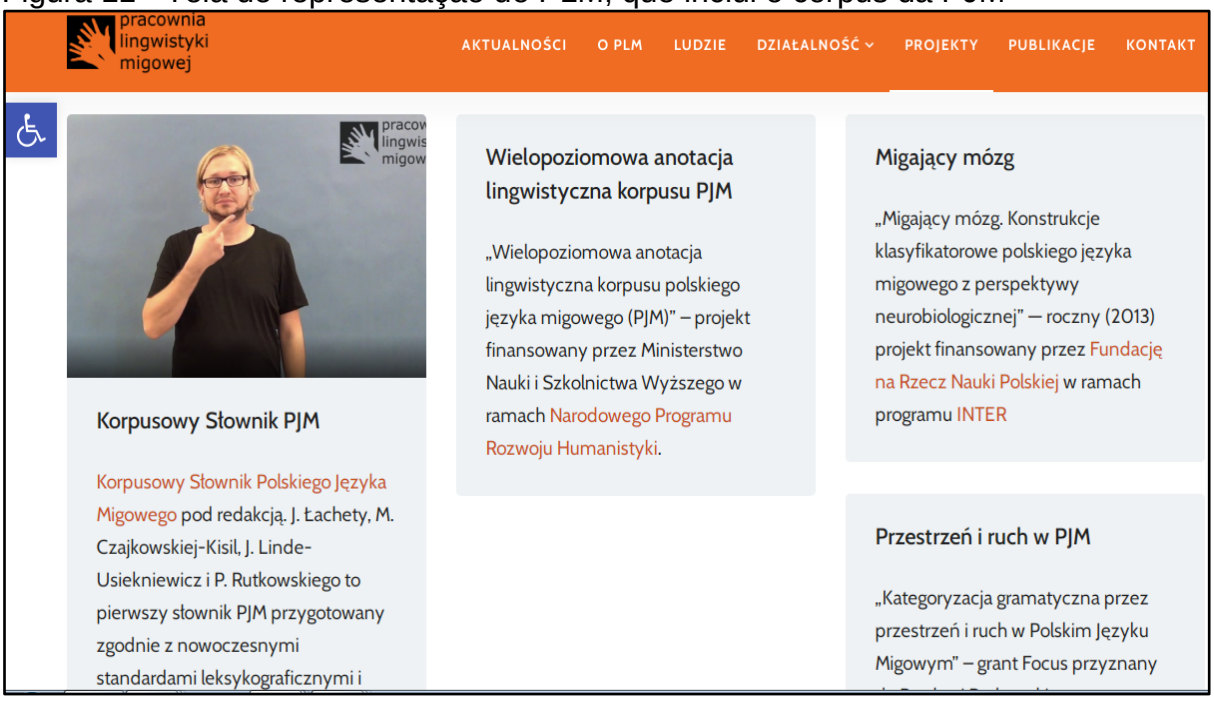

Fonte: https://archive.mpi.nl/.

Observa-se, nessa figura, que ao lado esquerdo está representado um mapa mundi com etiquetas, cuja função é indicar diversas línguas em diferentes regiões. Ao selecionar uma região, conforme mencionado, é possível encontrar materiais em línguas faladas e sinalizadas em formato de áudio e vídeo, associados a transcrições, além de outras possibilidades de pesquisas linguísticas relativas ao uso cotidiano.

### 4.2.5 *Corpus* da Língua de Sinais Polonesa

O *Corpus* da Língua de Sinais Polonesa[41] (PJM) é um projeto de coleta de dados linguísticos de usuários da PMJ, realizado pelo Laboratório de Linguística de Sinais[42] (PLM), que foi organizado nas estruturas da Faculdade de Estudos Poloneses da Universidade de Varsóvia, em 1 de junho de 2010. O projeto foi implementado, a partir de 2009, e possui um acervo de vídeos com aproximadamente 150 pessoas surdas de todo o território polonês. Com o intuito de que seja um *corpus* representativo, para a realização das gravações são convidadas pessoas surdas, maiores de 18 anos e usuárias da PJM, nascidas ou moradoras da Polônia, envolvendo um número equilibrado de homens e mulheres de faixas etárias diferentes e com diversos níveis de escolaridade. Na Figura 11, pode-se visualizar essa plataforma.

Figura 11 - Tela de representação do PLM, que inclui o corpus da PJM




Fonte: https://www.plm.uw.edu.pl/projekty/.

Para uma sessão de gravação, há sempre alguns informantes e um moderador surdo (se necessário realiza explicações dos conteúdos, porém não interfere na

[41] Disponível em: https://www.slownikpjm.uw.edu.pl/. Acesso em: 26 set. 2020.
[42] Disponível em: https://www.plm.uw.edu.pl/projekty/korpus-pjm/. Acesso em: 26 set. de 2020.

manifestação comunicativa dos participantes), o que garante a produção de diálogos naturais e de estruturas sintáticas para a análise. Os informantes não conhecem suas tarefas com antecedência e as desempenham em duplas, como, relatar matérias assistidas, discutir temas selecionados, fazer arranjos, entre outras atividades. Os materiais utilizados são desenhos, *clipes* de filmes, gráficos com fotos, mapas, dentre outros recursos. Cada filme é precedido por instruções de um moderador e iguais para todos os participantes, no momento da gravação.

Os dados são coletados e após são realizadas as anotações no programa iLex, que permite, entre outras funções, a pesquisa em *corpus*. Essas anotações são divididas em caracteres individuais, lematização, marcação, transcrição da articulação na notação HamNoSys[43] e tradução de sentenças individuais em PJM para o idioma polonês.

O Anotação Linguística Multinível do *Corpus*[44] de PJM é um projeto financiado pelo Ministério da Ciência e do Ensino Superior no âmbito do Programa Nacional para o Desenvolvimento Humano[45]. Os desenvolvedores enfatizam que esse *corpus* é extremamente valioso para os estudos gramaticais e estudos lexicais, sendo que o objetivo é detalhar os dados dos sinais coletados, além de retratar a cultura e a vida cotidiana das pessoas surdas polonesas, constituindo-se em um banco de dados – tesouro cultural. Afirmam, também, que a anotação detalhada é necessária, portanto, as seguintes etapas são planejadas: (1) dar continuidade nas gravações de vídeo para coleta de dados; (2) selecionar, segmentar e lematizar um novo material; e (3) realizar uma anotação multinível (incluindo transcrição, usando a notação *HamNoSys*, que auxilia na segmentação de frases, na tradução para o polonês e na interpretação gramatical básica).

Nesse contexto, é válido mencionar que os desenvolvedores dizem que esse *corpus* contém um dos maiores acervos desse tipo, já registrados no mundo, e que o resultado será um banco de dados com a possibilidade de busca e o processamento livre de todas as informações coletadas, por meio do programa iLex. O projeto PLM

[43] É um sistema de notação de Língua de Sinais de Hamburgo. *HamNoSys* é um sistema de transcrição "fonética", que tem sido utilizado desde que sua versão original foi publicada, baseada no sistema de *Stokoe*. Disponível em: https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/hamnosys-97.html. Acesso em: 26 set. 2020.
[44] Disponível em: https://www.plm.uw.edu.pl/projekty/wielopoziomowa-anotacja-lingwistyczna-korpusu-pjm/. Acesso em: 26 set. 2020.
[45] Disponível em: https://www.gov.pl/web/nauka/narodowy-program-rozwoju-humanistyki/. Acesso em: 26 set. 2020.

é assessorado pelo professor Trevor Johnston, da Austrália, um dos mais exímios especialistas nas pesquisas das línguas de sinais no mundo, professor do Departamento de Linguística da *Macquari University*[46], em Sydney.

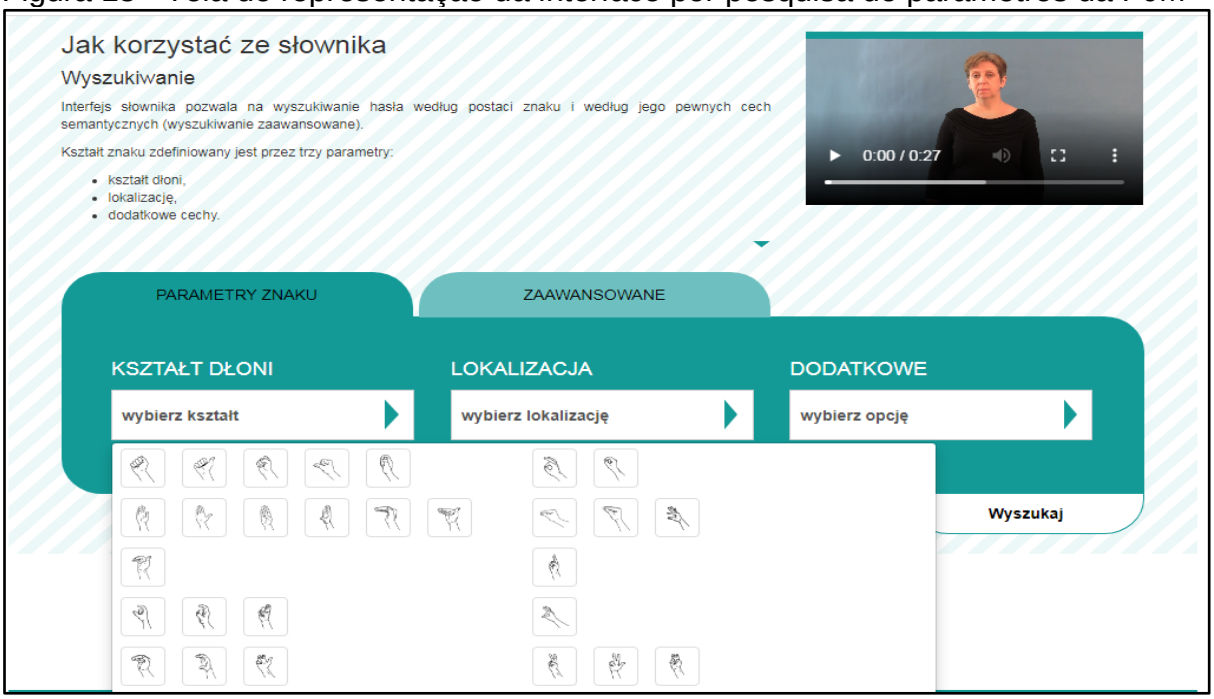
Ao acessar o Dicionário *Corpus* de Língua Gestual Polonesa[47], constam informações que ele é baseado no PJM *Corpus* e possui uma rica coleção de gravações de vídeos, sendo que esses registros são realizados por pessoas surdas usuárias da PJM, que falam sobre si mesmas e sobre tópicos de seus interesses, além de contar sobre histórias de imagens e vídeos. Há orientações, também, que para fazer uso do dicionário, a interface permite ao consulente a pesquisa de um termo por forma dos sinais e por alguns recursos semânticos, com a opção de pesquisa avançada. Para isso, são definidos três parâmetros: (i) configuração de mão; (ii) localização; e (iii) características adicionais. Na Figura 12, podemos visualizar essa plataforma, e a Figura 13 ilustra a interface de pesquisa com os parâmetros já assinalados.

Figura 12 - Tela de representação do Dicionário de PJM


Fonte: https://www.slownikpjm.uw.edu.pl/.

[46] Disponível em: https://www.mq.edu.au/. Acesso em: 26 set. 2020.
[47] Disponível em: https://www.slownikpjm.uw.edu.pl/. Acesso em: 26 set. 2020.

Figura 13 - Tela de representação da interface por pesquisa de parâmetros da PJM




Fonte: https://www.slownikpjm.uw.edu.pl/.

Nessa plataforma, os desenvolvedores enfatizam que o objetivo principal do projeto seria realizar um extenso *corpus* de pesquisa sobre gramática e vocabulário da PJM, pois, antes era uma língua pouco explorada, uma vez que não há a forma escrita. Assim, somente com os avanços tecnológicos foi possível registrar dados representativos da língua em uso. Esse projeto possibilitou análises gramaticais e lexicais da PJM com uma base em dados empíricos confiáveis, contando com aproximadamente 400 horas de filmagem e com a participação de quase 100 pessoas surdas em todo o território polonês.

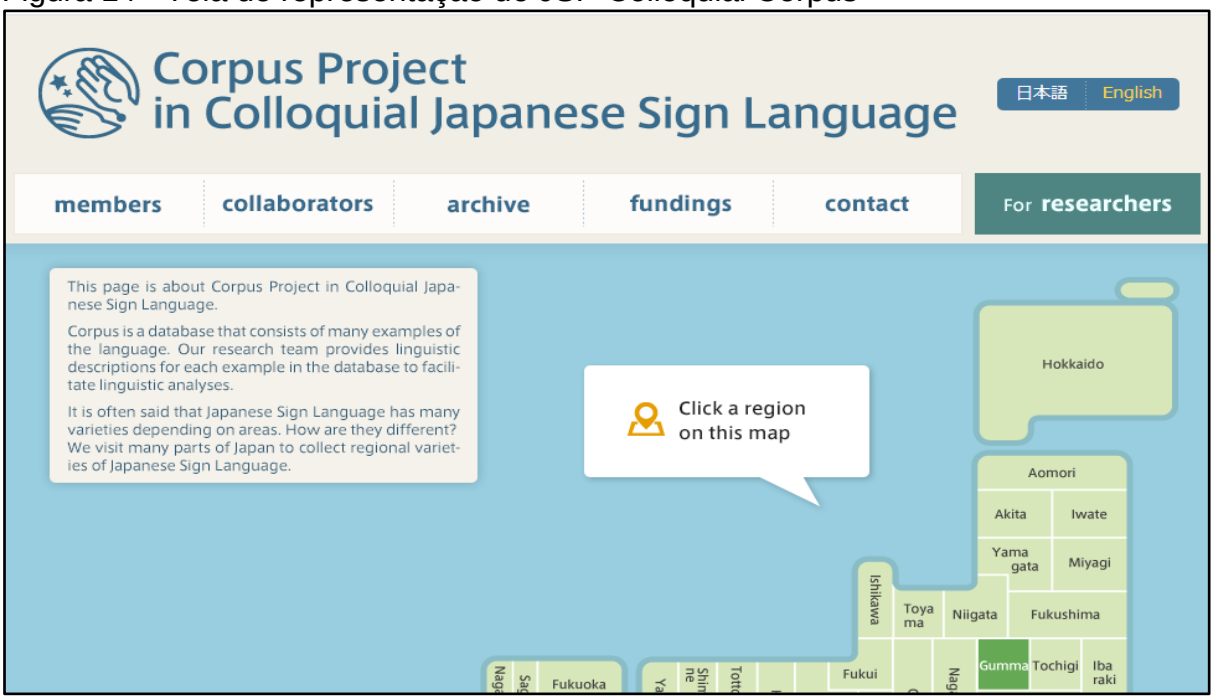
### 4.2.6 C*orpus* da Língua de Sinais Japonesa

A primeira fase do projeto de *corpus* (JSP *Colloquial Corpus*[48]), da Língua de Sinais Japonesa (JSL), foi financiada pela Sociedade Japonesa para a Promoção da Ciência (JSPS), entre os anos de 2011 e 2014. Foi o primeiro *corpus* criado com a finalidade acadêmica e acessível ao público. Para a criação do *corpus*, foi convidado o professor Adam Schembri, do projeto BSL que, com o apoio de seus colegas, colaborou para que esse projeto fosse efetivado. O *corpus* é colaborativo, assim

[48] Disponível em: http://research.nii.ac.jp/jsl-corpus/public/en/index.html. Acesso em: 26 set. 2020.

incentiva a colaboração de pessoas interessadas nesse projeto. Foram publicados na plataforma, vídeos segmentados e recortados (.mov), traduções (.pdf), manuais e dados anotados (.eaf). Na Figura 14, ilustramos o *corpus* em questão.

Figura 14 - Tela de representação do JSP Colloquial Corpus



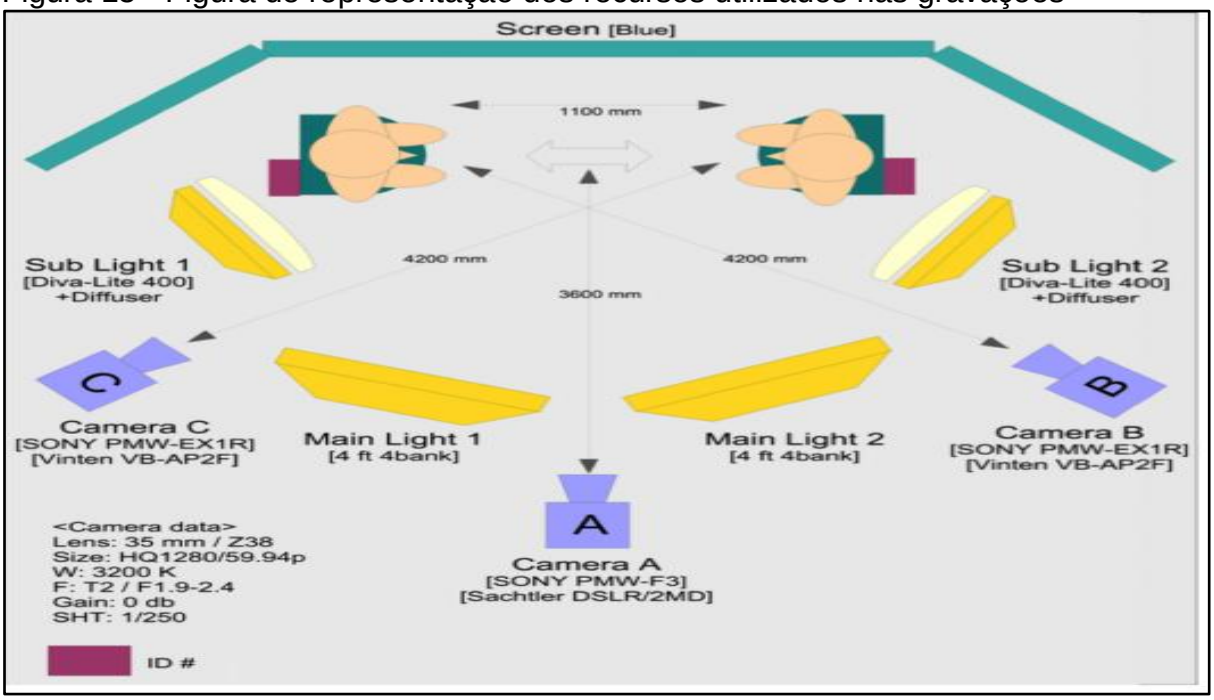

Fonte: http://research.nii.ac.jp/jsl-corpus/public/en/.

No *site*, há informações que foram filmados 40 sujeitos surdos em duas cidades, Gunma e Nara, cada uma com uma escola de pessoas surdas. Com isso, foi possível obter uma amostra balanceada, por idade, de indivíduos entre 30 e 70 anos, divididos em pares do mesmo sexo, em cada cidade. Para a coleta de dados foram usados três métodos: (i) entrevistas, com assistentes que viviam no mesmo território, conheciam os procedimentos com antecedência e realizavam perguntas aos participantes sobre suas vidas – essa etapa foi realizada apenas para fins introdutórios e não está disponível ao público; (ii) diálogos sobre animação, nesse momento, um participante memorizou uma história e, em seguida, explicou aos demais participantes; e (iii) elicitação lexical, ou seja, deixar que os participantes manifestassem, de forma natural, a língua de sinais, a fim de identificar as diferenças regionais e/ou geracionais de cada território.

O procedimento utilizado para as tarefas de diálogo foram: 1) dispositivos e configurações; 2) sincronização e corte de vários vídeos; 3) ID da região; 4) ID dos

participantes; 5) ID da sessão; e 6) nome dos arquivos. Em relação às filmagens, foram utilizadas três câmeras de alta definição, quatro dispositivos de iluminação, painéis azuis e cadeiras azuis, conforme ilustra a Figura 15.

Figura 15 - Figura de representação dos recursos utilizados nas gravações




Fonte: http://research.nii.ac.jp/jsl-corpus/research/data/manual/manual.html.

Nessa imagem, os desenvolvedores informam que: (i) a câmera A mostra os dois participantes dos joelhos para cima; (ii) a câmera B foca o participante à esquerda e as costas do outro participante; e (iii) a câmera C foca o participante à direita e as costas do outro participante. A disposição das câmeras foi projetada para melhor observação de todos os detalhes no diálogo, referentes aos sinais no espaço, à configuração de mãos, à direção do olhar, às expressões, enfim, às manifestações linguísticas para anotações e análises.

Para a construção desse *corpus* foram utilizados arquivos de vídeo, arquivos em *word* e arquivos do ELAN eaf. para as anotações. Para isso, foram realizadas duas etapas: (1) tradução em texto: compete aos intérpretes de língua de sinais a tradução da JSL para a língua japonesa escrita – foram criadas as glosas por ordem de palavras e traduções idiomáticas no *Microsoft Word*, em seguida houve a colaboração de pessoas surdas nativas de diferentes regiões do território japonês, a fim de verificar as traduções; (2) sobre esse texto criado em *word*, os sinalizantes nativos registram

no ELAN as anotações quanto às características dos movimentos das mãos para cada glosa e enunciado, a fim de observar as relações temporais entre ou dentro deles.

### 4.2.7 *Corpus* da Língua de Sinais Brasileira – Libras

O *corpus* da Língua de Sinais Brasileira[49] – Libras é um projeto desenvolvido na UFSC, com a finalidade de catalogar, de difundir e de proporcionar um ambiente de pesquisa da Libras, contribuindo de forma significativa para a valorização da cultura surda. A documentação vem sendo alimentada por meio do Inventário Nacional de Libras[50], financiada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ)[51], pelo Instituto do Patrimônio Histórico e Artístico Nacional[52] (IPHAN), em parceria com o Instituto de Investigação e Desenvolvimento em Política Linguística[53] (IPOL) e a UFSC[54]. Segundo dados disponíveis na plataforma, esse *corpus* se refere a um banco de produções de vídeos em Libras, que contempla diferentes gêneros textuais. O *corpus* inclui produções em Libras, de pessoas surdas de diversas idades, sendo homens e mulheres, de várias regiões do Brasil. O conteúdo desse *corpus* envolve conversas, narrativas, poemas, contos, entrevistas, listas de vocabulário, entre outros temas. Há possibilidade de alimentá-lo com mais e mais produções, indefinidamente.

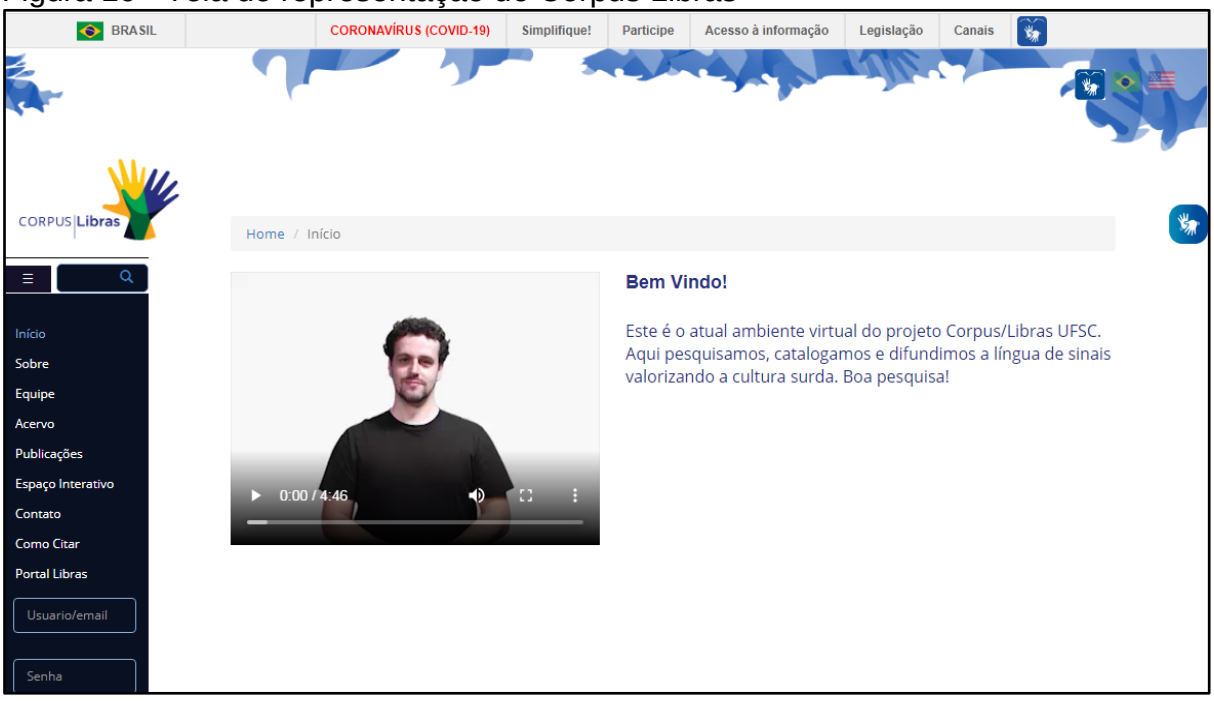
Os desenvolvedores desse projeto, ao considerar a relevância dos estudos linguísticos na Libras, mencionam que essa área carece de uma maior fundamentação empírica, em partes, devido aos grandes desafios que o registro e a manipulação de dados em uma língua de sinais demandam. Portanto, esse projeto tem a finalidade de contribuir com a reversão desse quadro, constituindo um *corpus* de Libras abrangente, representativo, consistente e vem sistematizando os procedimentos de registro, a documentação e a recuperação de dados e metadados relativos a Libras. A proposta

[49] Disponível em: http://www.corpuslibras.ufsc.br/. Acesso em: 27 set. 2020.
[50] O Inventário Nacional de Libras é um projeto ligado ao Instituto do Patrimônio Histórico e Artístico Nacional (IPHAN), do Ministério da Cultura. Faz parte do *corpus* e seu objetivo é coletar amostras de produções em Libras, identificar os usuários dessa língua no país, além de obter uma descrição básica dela, pois a considera um bem cultural brasileiro, que precisa ser preservado. Disponível em: http://ipol.org.br/inventario-nacional-da-lingua-brasileira-de-sinais-libras-levantamento-on-line/. Acesso em: 27 set. 2020.
[51] Disponível em: http://cnpq.br/. Acesso em: 27 set. 2020.
[52] Disponível em: http://portal.iphan.gov.br/indl. Acesso em: 27 set. 2020.
[53] Disponível em: http://ipol.org.br/. Acesso em: 27 set. 2020.
[54] Disponível em: http://ufsc.br/. Acesso em: 27 set. 2020.

dessa pesquisa, segundo os desenvolvedores, é oferecer resultados tanto no âmbito teórico quanto no âmbito aplicado, ou seja, primeiramente desenvolver diretrizes sobre a constituição de um *corpus* da Libras e, posteriormente, desenvolver uma proposta de estruturação de ensino da disciplina de Libras nas universidades brasileiras.

Nesse contexto, é relevante mencionar que a documentação da Libras teve início em 2013 e continua em andamento até os dias atuais. Os desenvolvedores pretendem que esse *corpus* seja cada vez mais alimentado com projetos de todo o território brasileiro, com a participação de pesquisadores surdos, pesquisadores bilíngues, instituições de fomento à pesquisa, órgãos governamentais e não governamentais. Acreditam, assim, que essas ações viabilizarão a documentação da Libras no país, que será amplamente socializada e concretizará uma política linguística, considerando as diferenças culturais e linguísticas de cada região. Na Figura 16, podemos visualizar a organização de tela inicial dessa plataforma.

Figura 16 - Tela de representação do Corpus Libras




Fonte: http://www.corpuslibras.ufsc.br/.

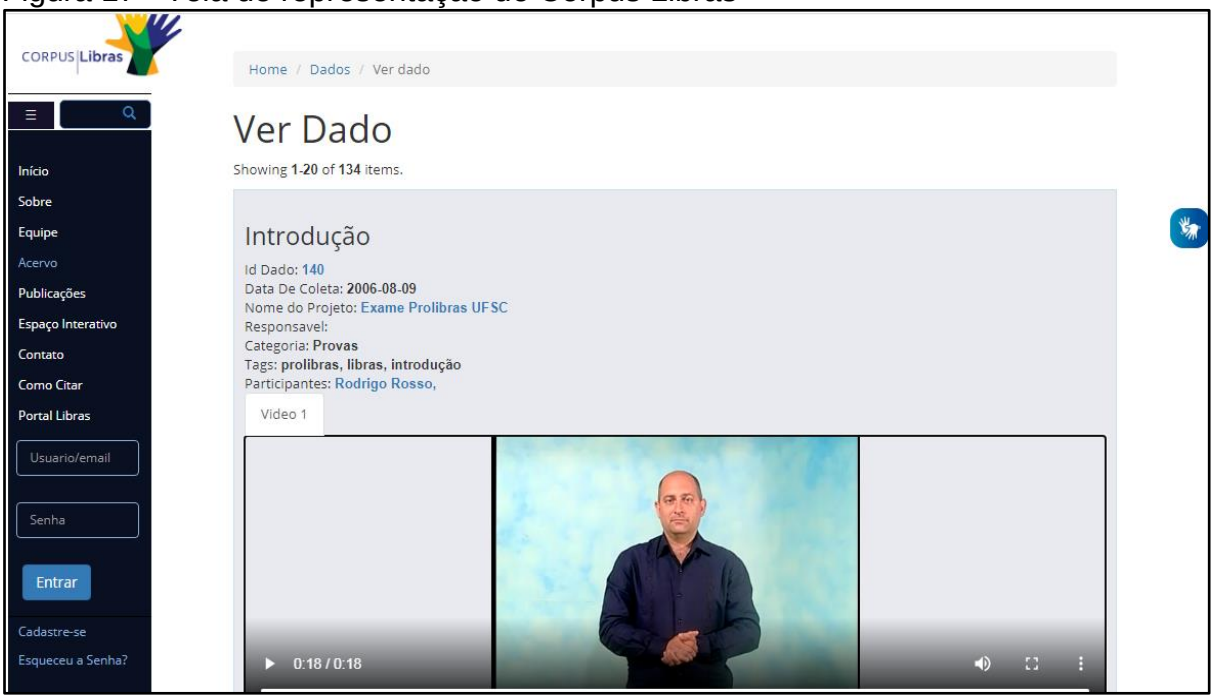
O *Corpus* de Libras envolve um conjunto de dados e metadados da Libras e está acessível a todas as pessoas interessadas. Descreveremos os objetivos elencados por seus desenvolvedores, a seguir:

> (i) disponibilizar um *corpus* gratuito e *online*, empiricamente abrangente e teórica/metodologicamente bem fundamentado a outros pesquisadores, profissionais que atuam com pessoas surdas e outras pessoas que desejam utilizá-lo para fins variados;
> (ii) oferecer diretrizes para a constituição de *Corpus* de Libras, no que tange ao registro, à documentação e à recuperação de dados para fins de análise linguística;
> (iii) disseminar as alternativas tecnológicas existentes, para fundamentar as pesquisas de forma empírica e consistente;
> (iv) realizar um importante registro linguístico, histórico e cultural, referente à vida cotidiana de pessoas surdas, visando contribuir com o processo de inclusão social no país.

Logo, nessa plataforma, segundo as informações, será possível encontrar: (i) um *Corpus* de Libras com registros em vídeos/fotos de situações eliciadas/motivadas e espontâneas de uso, para serem utilizados em pesquisas e/ou para outras finalidades aplicadas; (ii) um conjunto de diretrizes, para o registro e o arquivamento de dados e metadados, referentes ao uso da Libras; (iii) um formulário com campos e itens padronizados para sistematizar os resultados finais dos diferentes projetos associados ao *Corpus* de Libras.

A página do *Corpus* de Libras contém um material organizado por regiões, e ao clicar no ícone de cada estado é possível acessar os dados disponíveis de cada um, no entanto, não há informações em todas as regiões até o momento, uma vez que o *corpus* se encontra em construção. Os projetos disponíveis contemplam vários tipos de dados, como vídeos, imagens, textos e transcrições. A fim de ser ampliado, esse projeto oferece a possibilidade de estar se associando a outros em todo território brasileiro. Segundo os desenvolvedores, o objetivo é que as cinco regiões do país sejam representadas, posteriormente, todas as capitais dos estados. A região de Santa Catarina, onde se iniciou esse projeto, por exemplo, inclui o Inventário Nacional de Libras, poemas e contos produzidos por ex-alunos do curso Letras-Libras (turmas de 2006 e 2008) e exame Prolibras. Referente ao acervo disponível na plataforma, podemos visualizar na Figura 17.

Figura 17 - Tela de representação do Corpus Libras




Fonte: http://www.corpuslibras.ufsc.br/dados/dado/porprojeto/Exame%20Prolibras%20UFSC.

Para a utilização dos materiais disponíveis nesse *corpus*, segundo os desenvolvedores, não é necessário solicitar autorização, no entanto é preciso encaminhar o projeto de pesquisa ao comitê de ética para aprovação, informando que os materiais usados são de domínio público.

Para usar o C*orpus* de Libras, faz-se necessário utilizar o *software Handbrake,* para tanto, um *link*[55] de passo a passo é disponibilizado. Além disso, é possível ter acesso a vários tutoriais sobre a aplicação de materiais que podem ser empregados a vários projetos, a saber:

(i) Tutorial para transcrição[56] (Parte 1) – Explica-se como usar as funções iniciais do ELAN para transcrever sinais do *Corpus* de Libras;

(ii) Tutorial para transcrição[57] (Parte 2) – Demonstra-se como usar algumas funções avançadas do ELAN, a fim de transcrever dados para realizar análises das línguas de sinais;

[55] Disponível em: http://www.repositorio.ufsc.br/formatos-de-arquivos/conversao-de-videos-usando-handbrake/. Acesso em: 28 set. 2020.
[56] Disponível em: https://www.youtube.com/watch?v=VtmG8AQ1lD0. Acesso em: 28 set. 2020.
[57] Disponível em: https://www.youtube.com/watch?v=jsy6001wId4. Acesso em: 28 set. 2020.

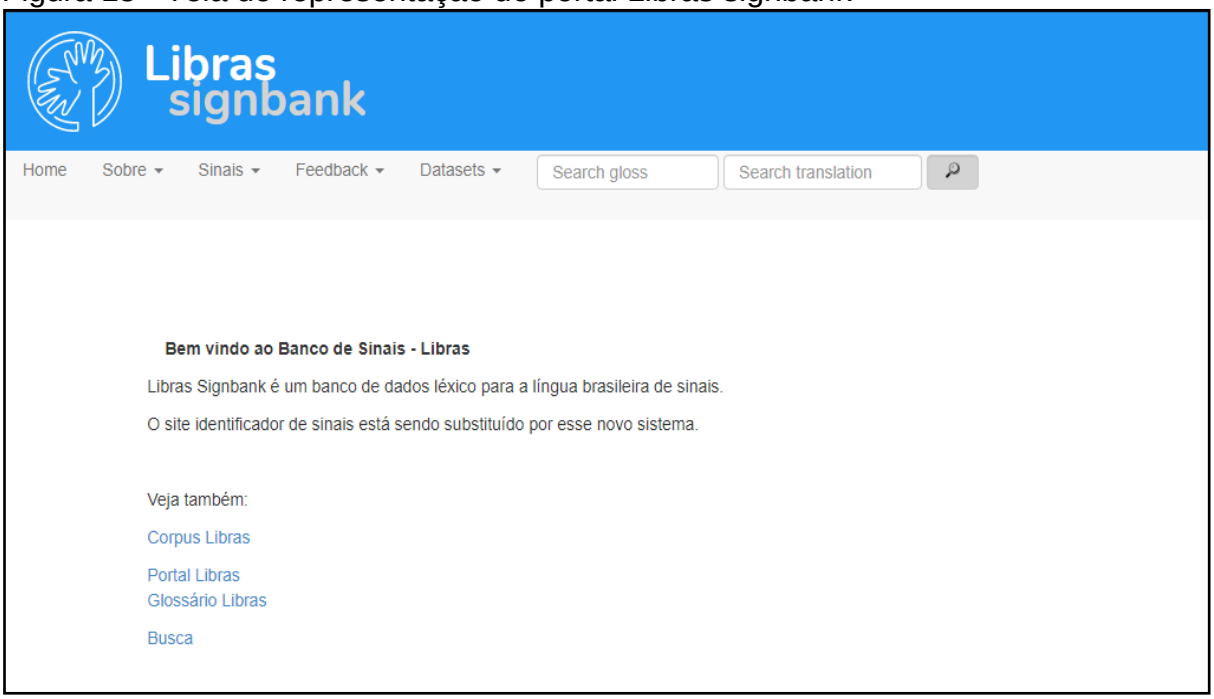
(iii) Tutorial de armazenamento e conversão de vídeos do *Corpus* de Libras[58] – constam instruções de como está organizado os dados em forma de vídeo, bem como o formato de conversão dos vídeos para realizar a transcrição no Sistema de Anotação do ELAN;

(iv) Manual de Transcrição do *Corpus* de Libras[59] – Apresenta-se as convenções definidas pelo Grupo de Pesquisa do *Corpus* de Libras a serem aplicadas na transcrição dos dados em língua de sinais.

É relevante mencionar, também, que além desses manuais, em especial o "Manual do Transcritor", os transcritores fazem uso do ID que, mais recentemente, vem sendo substituído pelo *Libras signbank*[60]*.* Essa ação foi estabelecida pela equipe, pois esse banco de sinais pode ser associado ao ELAN para busca de termos, tanto para transcrição dos sinais, como para tradução para o português e para o inglês. Além disso, o intuito é integrar os estudos referentes à língua de sinais internacionalmente (QUADROS, 2019). Na Figura 18, ilustramos a representação dessa plataforma.

[58] Disponível em: https://repositorio.ufsc.br/bitstream/handle/123456789/169879/TUTORIAL%20VIDEOS%20NALS%20%281%29.pdf?sequence=1&isAllowed=y. Acesso em: 28 set. 2020.
[59] Disponível em: https://repositorio.ufsc.br/bitstream/handle/123456789/169881/2015%202905%20MANUAL_CORPUS%20transcric%cc%a7a%cc%83o.pdf?sequence=1&isAllowed=y. Acesso em: 28 set. 2020.
[60] Disponível em: http://signbank.libras.ufsc.br/. Acesso em: 29 set. 2020.

Figura 18 - Tela de representação do portal Libras signbank




Fonte: http://signbank.libras.ufsc.br/.

Por fim, na subseção a seguir, após a investigação e a descrição dos *corpora* acima, apresentaremos as análises dos critérios identificados até o momento, a fim de contemplar a proposta desta pesquisa.

## 4.3 ANÁLISE DOS CRITÉRIOS UTILIZADOS NA ORGANIZAÇÃO E NA ESTRUTURAÇÃO DE *CORPORA* DE LÍNGUAS DE SINAIS

Nesta seção, apresentamos alguns aspectos-chave referentes aos critérios que observamos no processo de documentação e de registros baseados em *corpus* para pesquisas em língua de sinais. Dentre outras línguas de sinais pesquisadas, destacaremos aqui os critérios mais frequentes utilizados nas línguas de sinais: Australiana – Auslan; Britânica – BSL; Alemã – DGS; Norueguesa – NGT; Polonesa – PJM; Japonesa – JSL; e Brasileira – Libras.

Em primeiro lugar, enfatizamos que o campo da LC nas línguas de sinais é um termo, ainda, definido cuidadosamente como *corpus*, uma vez que, o surgimento da tecnologia que tornou possível essa nova abordagem para a pesquisa em língua de sinais se encontra em vias de buscar meios para a construção de *corpora* paralelos,

ou seja, que envolvam uma ou mais línguas orais com uma ou mais línguas de sinais, com metadados associados.

Em segundo lugar, é notável, em cada um desses *corpora* investigados, que alguns critérios definidos, principalmente em relação à seleção dos informantes e dos participantes, condizem com os pressupostos teóricos da LC.

Conforme observado, na seção anterior, o primeiro projeto de *corpus* de língua de sinais começou na Austrália, por volta de 2004, com um arquivo organizado e estrurado, a partir de vídeos que foram capturados por meio de gravações de um grupo de 100 surdos nativos ou aprendizes precoces e/ou sinalizantes, quase nativos de Auslan. Assim como os demais, a expectativa sempre foi tornar os *corpora* legíveis por máquina.

Nesse sentido, é possível verificar que, ao longo dos tempos, os desenvolvedores de projetos voltados aos *corpora* de línguas de sinais lançaram mão de metodologias e critérios na construção do *corpus*.

Desse modo, percebemos, no decorrer da nossa pesquisa, que é comum encontramos pelo menos três etapas que contribuem significativamente para a seleção de critérios que estruturam os *corpora*, quais sejam: (1) vídeos – recrutamento de participantes, seleção das atividades linguísticas para a produção do *corpus* e configuração para as filmagens; e (2) metadados (dados sobre dados) – anotação para os *corpora*, geralmente com foco no uso de glosas e etiquetas, que possibilitam informações sobre os dados. Essa etapa é considerada uma das mais demoradas e trabalhosas na criação de *corpora* de língua de sinais, uma vez que apresenta anotações de informações variadas, por exemplo, sobre o participante (região, sexo, idade), entre outros. Já em relação à terceira etapa, percebemos que está mais voltada (3) à acessibilidade e ao uso dos dados disponíveis nos *corpora*. No entanto, tais dados parecem não atender às necessidades relacionadas às pesquisas voltadas às análises linguísticas envolvendo *corpora* paralelos de línguas de sinais x línguas de sinais e línguas de sinais x línguas orais.

Sobre esse terceiro momento, retomaremos no próximo capítulo, onde iremos refletir e propor os critérios necessários para a construção de um *corpus* paralelo Libras-Português, para fins de análises linguísticas.

Até o momento, na pesquisa realizada nos *corpora* das línguas de sinais: Auslan, BSL, DGS, NGT, PJM, JSL e Libras, identificamos pelo menos sete critérios que são comuns entre os desenvolvedores para a organização de vídeos. A fim de

verificarmos as semelhanças e as diferenças na primeira etapa de seleção dos critérios, organizamos a Tabela 1, apresentada a seguir.

Tabela 1 - Critérios identificados em corpora de língua de sinais: organização dos vídeos

| **Critérios para a organização dos vídeos** | ***Corpus* AUSLAN** | ***Corpus* BSL** | ***Corpus* DGS** | ***Corpus* NGT** | ***Corpus* PJM** | ***Corpus* JSL** | ***Corpus* LIBRAS** |
|---|---|---|---|---|---|---|---|
| 1. Recrutamento de informantes surdos sinalizantes. | X | X | X | X | X | X | X |
| 2. Distribuição regional dos registros filmados. | X | X | X | X | X | X | X |
| 3. Gravações baseadas em entrevistas. | X | X | X | X | X | X | X |
| 4. Produções de narrativas (espontâneas ou motivadas). | X | X | X Piadas sem anotações. | X | X | X | X |
| 5. Pesquisas variadas (conversações livres e outras respostas linguísticas motivadas por estímulos diversos). | X | X | X | X | X | X | X |
| 6. Possibilitar pesquisas de sinais relacionados a diferentes temas (saúde, educação etc.). | X | X | X Materiais escolares. | X | X | X | X |
| 7. Distribuição de câmeras em diferentes posições. | Informações não disponíveis sobre o nº de câmeras. | Informações não disponíveis sobre o nº de câmeras. | Gravação em estúdio móvel por toda a Alemanha. Uso simultâneo de **três** câmeras. | Uso simultâneo de **quatro** câmeras. | Uso simultâneo de **cinco** câmeras. | Uso simultâneo de **três** câmeras. | Uso simultâneo de **quatro** câmeras. |

Fonte: Elaborada pela autora da pesquisa.

Ao observarmos os critérios identificados na Tabela 1a, de modo geral, notamos que há dados em comum em relação à seleção de informantes, aos conteúdos utilizados nas filmagens e à disponibilização de dados por regiões, a fim de contemplar as variações linguísticas ou a língua local. Quanto aos números de câmeras utilizadas nas filmagens, observamos, nos *corpora* que encontramos essa informação, que há, pelo menos, três ou mais câmeras distribuídas em diferentes posições para as gravações.

Nessa etapa, ainda, justificamos que não informamos/identificamos o tamanho do *corpus* como um critério a ser analisado, uma vez que concordamos com a questão da flexibilidade, já discutida sobre o tamanho que deveria ter um *corpus*, ou seja, quanto à quantidade de enunciados/dados, à quantidade de vídeos, no que concerne ao tempo de cada vídeo ou, ainda, quanto ao tempo total que cada *corpus* já constituído possui. Embora seja uma questão pertinente, notamos que quanto a esse quesito não há uma padronização, pois o/as tamanho/amostras, assim como nos estudos da LC em relação às línguas orais, deve ser representativa a língua que está sendo investigada pelo pesquisador. Logo a questão de tamanho/amostras nos parece uma questão bem subjetiva.

Em relação à segunda etapa, identificamos ao menos seis critérios em comum nos *corpora* pesquisados. Para tanto, organizamos na Tabela 2 os dados encontrados, relacionados aos metadados para a estruturação de um *corpus* linguístico, conforme segue.

Tabela 2 - Critérios identificados em corpora de língua de sinais: estruturação dos metadados para um corpus linguístico

| **Critérios Para estruturação dos metadados para um *corpus* linguístico** | ***Corpus* AUSLAN** | ***Corpus* BSL** | ***Corpus* DGS** | ***Corpus* NGT** | ***Corpus* PJM** | ***Corpus* JSL** | ***Corpus* Libras** |
|---|---|---|---|---|---|---|---|
| 1. Uso do *ELAN* como ferramenta computacional de transcrição e de anotação. | X | X | X | X | ----- | X | X |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2. Uso de outras ferramentas computacionais de transcrição e anotação. | IMDI (ISLE) – Para descrição. | EXTOL – Desenvolvimento SW. Kinect. | ILEX. MaxQDA. OpenPose. | IMDI (Tolk) recursos para voz (não eletrônica) SW Arbil. | ILEX e HamNoSys. | Handbrake WORD. | Handbrake. |
| 3. Criação de trilhas com anotações diversas com informações acerca dos informantes (comunidade surda, nível de escolaridade, dentre outras). | X | X | X | X | X | X | X |
| 4. Criação de trilhas com anotações diversas, contendo informações linguísticas e extralinguísticas. | X | X | X | X | X | X | X |
| 5. Arquivos anotados com glosas. | X | X | X | X | X | X | X |
| 6. Informações etiquetadas /anotações. | X | X | X | X | X | X | X |

Fonte: Elaborada pela autora da pesquisa.

Ao considerar a heterogeneidade que encontramos nos tipos de *corpus* em línguas de sinais, bem como as comunidades de pesquisa envolvidas, a partir desse estudo e dos elementos que elencamos na tabela 1b, observamos que, até o momento, parece não haver algum método ou critérios que sejam amplamente dominantes ou sirvam como um padrão oficial, como por exemplo, fazer anotações nos vídeos com registros em línguas de sinais. Por outro lado, o número de ferramentas e os formatos, utilizados na construção dos *corpora*, são bastante variados, e apesar dessas variações é possível notar que existem algumas semelhanças conceituais entre eles, por exemplo, as informações contidas nas anotações.

No Brasil, a UFSC é considerada uma das referências em relação aos estudos linguísticos da Libras. Ronice Muller de Quadros é uma das idealizadoras do projeto de *corpus*, que envolve diferentes projetos para registros de metadados em Libras.

Esse *corpus* está sendo constantemente alimentado e se encontra disponível no portal de Libras da UFSC (QUADROS, 2016).

Ao acessar esse *corpus*, a autora apresenta as decisões tomadas para se estabelecer os procedimentos necessários para a realização das transcrições básicas do *Corpus* de Libras e das anotações sugeridas para o desenvolvimento de análises mais específicas dos dados em Libras (QUADROS, 2016).

Conforme Quadros (2016), vários documentos foram e ainda podem ser registrados nesse *corpus* de Libras, com diferentes finalidades. Segundo a autora:

> Essa coleta de dados objetiva ser replicada em todo o Brasil para o estabelecimento de um Corpus da Libras com dados que permita análises comparáveis da Libras de diferentes regiões do país. A metodologia usada para o Inventário Nacional de Libras compreende interações de surdos em pares divididos em três grupos, por idade e por gênero. Todos os procedimentos para a coleta dos dados, organização dos dados e metadados e transcrição dos dados foram aplicados e ajustados para serem usados em todo o país e permitirem essa coleta de dados objetiva ser replicada em todo o Brasil para o estabelecimento de um Corpus da Libras com dados que permita análises comparáveis da Libras de diferentes regiões do país. (QUADROS, 2016, p. 12).

Sobre os critérios indicados por Quadros (2006) para a realização das transcrições em Libras, de acordo com as análises estabelecidas junto ao grupo de pesquisa do *Corpus* de Libras, quatro decisões importantes foram tomadas:

1) O uso do ELAN para a transcrição de dados do *Corpus* de Libras;

2) A anotação apenas por meio de glosas de sinais produzidos, exclusão de informações morfológicas com a utilização do Identificador de Sinais (ID) de cada sinal (evita-se o problema em definir o que constituiria a sentença na língua de sinais);

3) A anotação de sinal por sinal de ambas as mãos: mão direita e mão esquerda;

4) A tradução livre do texto em Libras para a Língua Portuguesa, no formato de texto, com segmentação por meio de sentenças enquanto unidades de sentido (aqui a questão da sentença é determinada pelo sentido, e não por razões sintáticas).

O ELAN é um *software* gratuito disponível para *download*[61]*.* Essa ferramenta oferece diferentes recursos para transcrever as anotações de fala e/ou sinais associadas às gravações em vídeo. É muito utilizada por outros autores, como Nonhebel, Crassborn e Van Der Kooij (2004), nas discussões sobre as convenções para a transcrição (QUADROS, 2004). Além do uso do ELAN, seguindo a linha de outros pesquisadores (JOHNSTON, 1991; PIZZUTO; PIENTRANDREA, 2001; NONHEBEL; CRASBORN; VAN DER KOOJI, 2004), essa plataforma consiste em busca de sinais ou glosas, pois, considera-se esse um meio mais simples para a transcrição (QUADROS, 2004). Assim, foi criado o ID, o qual usa como parâmetro a Configuração de Mão e a Localização do Sinal. Além disso, é possível encontrar o sinal ou glosa pelo nome em português, ou parte dele.

Na próxima seção, retomaremos com mais detalhes sobre essa ferramenta, pois verificamos ser a mais utilizada como recurso nos *corpora* pesquisados.

Vale mencionar, também, a dissertação de Veras (2014) em relação ao C*orpus* de Libras, que trata sobre os procedimentos metodológicos na formação de *corpus* de línguas de sinais, a partir da plataforma *youtube.com*, tendo a formação de um *corpus* de gêneros da Libras como projeto piloto para a elaboração e a identificação das principais questões que cercam a constituição daquele, a partir de vídeos da rede. Nesse prisma, Veras (2014) aborda sobre questões referentes à ética na formação de *corpus*, além de questões metodológicas como: o procedimento para a extração de vídeos; a organização do *corpus* em vídeos; a limpeza do *corpus* em vídeo; a seleção de vídeos e os gêneros em *corpora*.

Conforme Veras (2014), embora seja necessário um tratamento mais criterioso, o material coletado contém aproximadamente 1.100 vídeos e foi classificado e autorizado para o uso em pesquisas. O autor reconhece que é uma plataforma limitada em relação “a gêneros específicos, categorias, e informantes de grupos específicos” (VERAS, 2014, p. 130).

Assim, notamos que ainda há muitos desafios atribuídos aos estudiosos dessa área, pois mesmo com os avanços tecnológicos, existe um grande caminho a ser percorrido para obter resultados consideráveis ao que tange à constituição de um *corpus* paralelo, em língua de sinais em interface com as línguas orais. Além disso, a disposição desses *corpora* exige um grande trabalho em equipe, que envolve,

[61] Disponível em: em http://www.latmpi.eu/tools/elan/. Acesso em: 20 out. 2019.

inevitavelmente, linguistas e profissionais da tecnologia de informação. É válido mencionar, também, que se faz necessária uma tecnologia de ponta, além de altos investimentos financeiros.

Na sequência, abrimos uma nova seção referente aos/às programas/ferramentas computacionais que vêm sendo utilizados pelos pesquisadores, para fins de construção e manipulação de *corpora* aplicados às línguas de sinais, de um modo geral, e à Libras, de modo particular.

## 4.4 PROGRAMAS UTILIZADOS PARA A CONSTRUÇÃO DE UM *CORPUS* PARALELO LIBRAS-PORTUGUÊS

Apresentaremos, nesta seção, de forma sucinta, alguns *softwares* utilizados nas línguas de sinais, identificados na descrição dos *corpora* abordados no decorrer deste trabalho. Em nossas análises, constatamos que praticamente todos os *corpora* apresentados utilizam o ELAN, como ferramenta, para o desenvolvimento das transcrições. Destacamos que não encontramos a informação do uso do ELAN somente no *corpus* da PJM.

É válido mencionar que os *softwares* que serão apresentados e descritos, quanto as suas funcionalidades, não são obrigatórios para o desenvolvimento de novos *corpora*, ou seja, há possibilidades de serem criados outros *softwares* com esse propósito. No entanto, o intuito dessa proposição é colaborar com os pesquisadores, que tenham interesse em realizar análises linguísticas em *corpora* e buscam uma ferramenta que possa auxiliá-los em suas análises. Notamos, portanto, que o trabalho dos linguistas e pesquisadores da língua está, cada dia mais, interligado ao trabalho dos cientistas da computação.

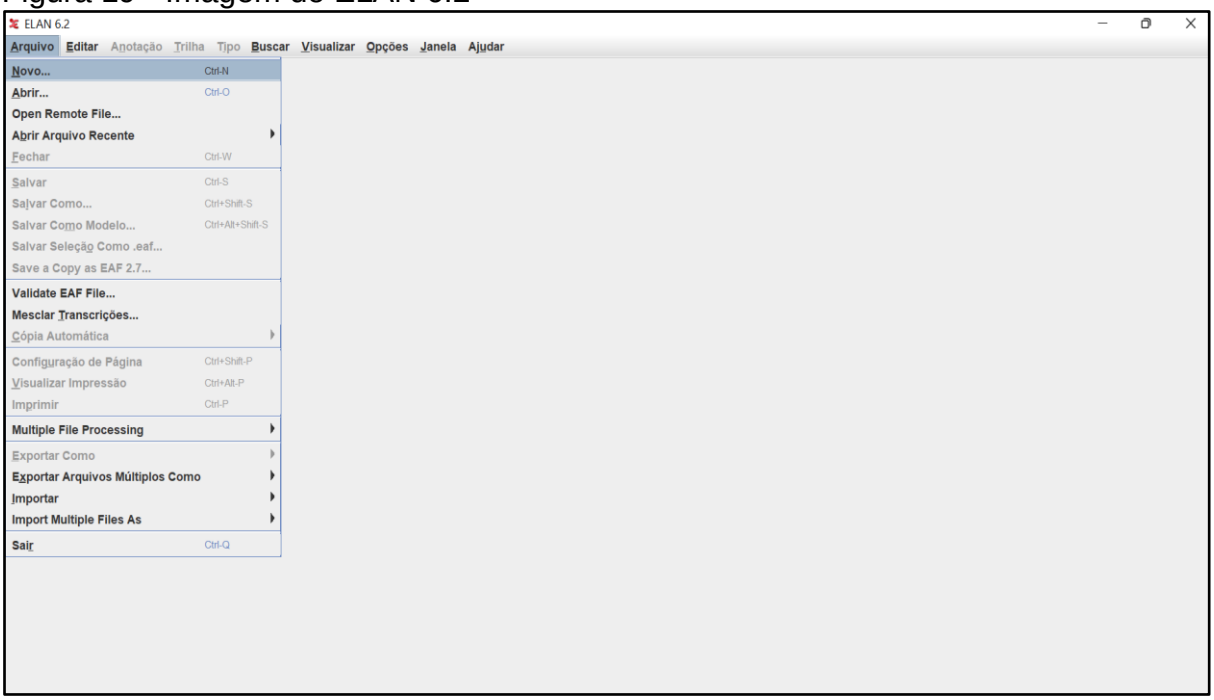
Nas próximas subseções, serão descritos os programas identificados na Tabela b, a fim de mostrar alguns recursos encontrados e direcionar os pesquisadores a se aprofundarem nos recursos que lhes interessarem, para futuros trabalhos. Mencionamos, ainda, que o ELAN, por ser o mais utilizado, despertou-nos interesse e, por isso, foi baixado com o intuito de verificarmos as suas funcionalidades. Dessa forma, iremos apresentá-lo de forma mais detalhada, neste trabalho, para esclarecimentos e como uma possível sugestão de recurso aos pesquisadores, uma vez que, frequentemente, ele vem sendo atualizado e aperfeiçoado com novas versões que acompanham as necessidades de análises linguísticas dos

pesquisadores. No entanto, destacamos, também, que há vários tutoriais que poderão servir como base para aqueles que se interessarem em realizar suas análises com esse programa.

### 4.4.1 ELAN (Versão 6.2) – EUDICO *Language Annotator*[62]: ferramenta para transcrição/anotação

Quanto ao ELAN, verificamos que o seu uso é recorrente para transcrição das línguas de sinais, na maior parte dos *corpora* descritos neste trabalho, podendo, consequentemente, ser uma importante ferramenta para a constituição de *corpora* em línguas de sinais. Conforme McCleary, Viotti e Leite (2010), o ELAN é uma ferramenta para descrições e análises linguísticas multimodais. Dentre outras vantagens, esse *software* está sendo constantemente atualizado, portanto, com as novas versões é possível a correção de problemas e a disponibilização de novos recursos. Para a visualização da interface desse programa, podemos conferir a Figura 19.

Figura 19 - Imagem do ELAN 6.2

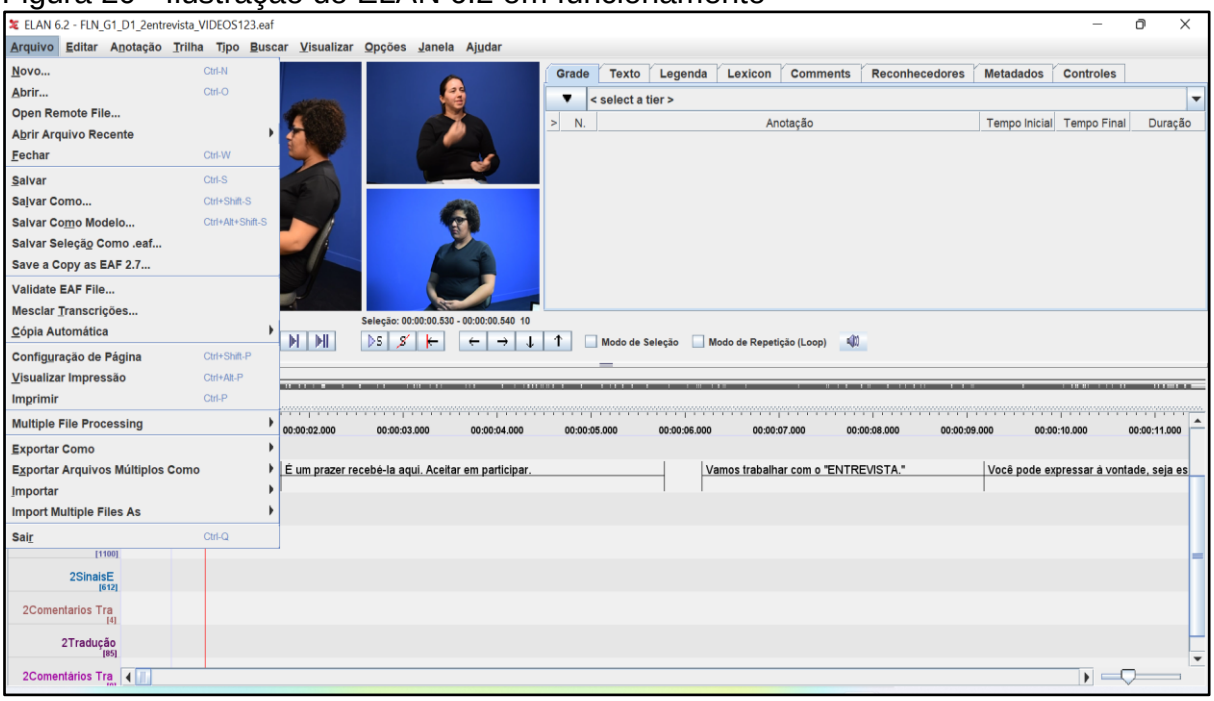



Fonte: Imagem referente ao Programa ELAN 6.2.

[62] Versão mais recente disponível em: https://archive.mpi.nl/tla/elan. Acesso em: 30 out. 2021.

O programa ELAN foi desenvolvido pelo *Max Planck Institute of Psycholinguistics*, da Holanda. A versão ilustrada nessa etapa é a 6.2 e foi lançada em 6 de julho de 2021. Esse programa é, constantemente, utilizado como uma ferramenta para realizar a importação de vídeo para transcrição, uma vez que apresenta um sistema complexo de buscas e a capacidade de operar com até quatro câmeras. Com esse programa é possível criar trilhas (glosa, gestos, marcação não manual e marcação cultural), levantamento bibliográfico para embasamento teórico e análise descritiva do texto em Libras e Português. Além disso, está disponível gratuitamente. O conteúdo das anotações consiste em texto Unicode, e os documentos de anotação são armazenados em um formato XML (EAF). Baixamos um vídeo disponível para consulta no *signbank*, com o intuito de ilustrar um pouco dos mecanismos disponíveis após um vídeo ser baixado para as análises (FIGURA 20).

Figura 20 - Ilustração do ELAN 6.2 em funcionamento

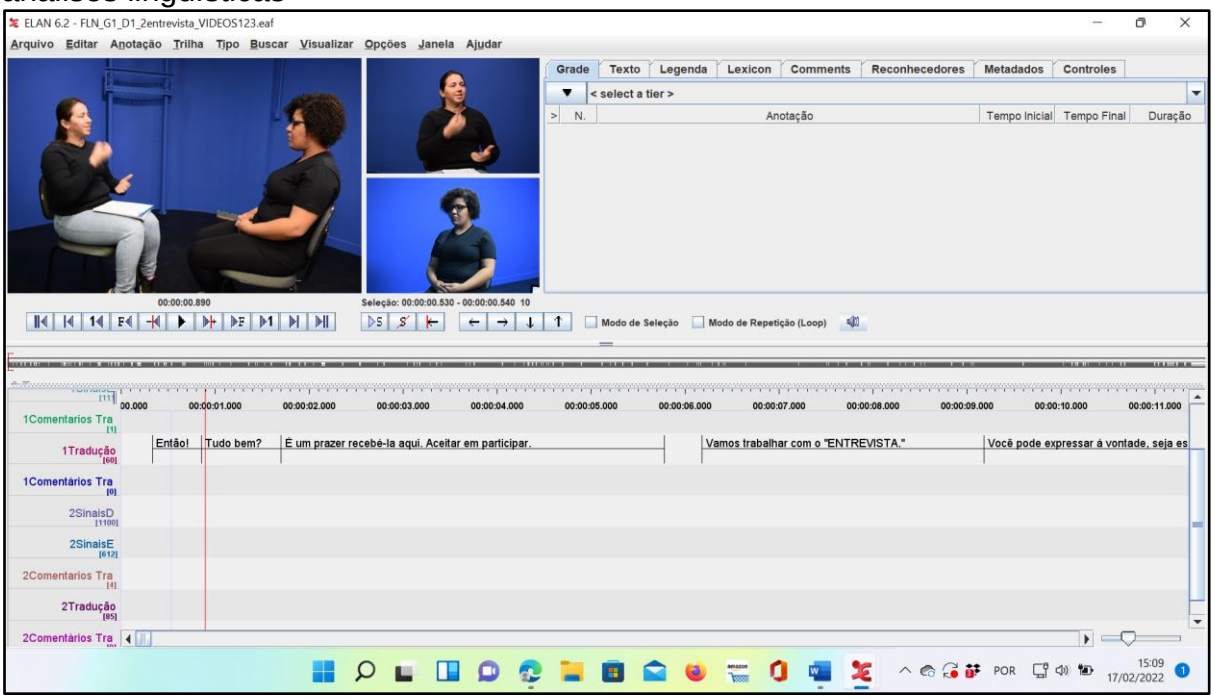



Fonte: Imagem referente ao Programa ELAN 6.2.

Essa ferramenta, ainda, possibilita o uso de arquivos de vídeos e de áudio, o que é importante para uma transcrição e uma análise bimodal (surdos e ouvintes/Libras-Português). Também, oferece a possibilidade de visualizar dois ou mais vídeos simultaneamente e a criação de várias trilhas, ou seja, com diversas informações e análises sobre o/a vídeo/manifestação linguística, sem acarretar

problemas no processo operacional do programa. Outro fator existente, e não menos importante, é que o ELAN apresenta flexibilidade, ao pesquisador, quanto à escolha das trilhas para a transcrição e/ou as análises que são do seu interesse, ocultando aquelas que não correspondem a sua pesquisa. Para ilustrar as trilhas criadas após um vídeo ser baixado, criamos um exemplo, a fim de exemplificar a disposição do vídeo e das trilhas no programa. Na Figura 21, podemos verificar como se estabelece essa relação.

Figura 21 - Ilustração dos vídeos e trilhas criadas, por nós, como exemplo para as análises linguísticas




Fonte: Imagem referente ao Programa ELAN 6.2

A versão 6.3[63] é a mais recente, foi atualizada no dia 03 de fevereiro de 2022 e está disponível para *download*. É uma versão que contém alguns novos recursos e várias correções de *bugs*.

[63] Disponível para Download em: https://archive.mpi.nl/tla/elan/download. Acesso em: 20 fev. 2022.

### 4.4.2 EXTOL[64] – Integrador para IBM i (EEI)

O EXTOL EDI Integrator for i (EEI) é um aplicativo de *software* do Sistema I, usado por empresas para o processamento de dados críticos de transação B2B, com base em qualquer combinação de cargas de arquivos planas X12, EDIFACT ou proprietárias. O aplicativo facilita o acesso, a pesquisa e a visualização de dados B2B pelos usuários. Isso significa dizer que é gasto menos tempo para as solicitações e mais tempo para projetos estratégicos. Além disso, é fácil de instalar, configurar e usar, no entanto é um *software* que exige um orçamento para adquirir.

### 4.4.3 Kinect[65] – Sensor de movimentos

É um sensor de movimentos desenvolvido para o Xbox 360 e Xbox One, junto com a empresa *Prime Sense*. Essa tecnologia incorpora câmeras RGB, projetores infravermelhos e detectores que mapeiam a profundidade por meio de cálculos estruturados de luz, como por exemplo, os usados em scanners 3D. Além disso, o sensor também possui microfones juntamente com *software* e a inteligência artificial da *Microsoft*. Isso possibilita que o dispositivo realize reconhecimento de gestos em tempo real, reconhecimento de voz e detecção esquelética corporal de até quatro pessoas. Isso viabiliza que o sensor seja utilizado como um dispositivo de interface natural de usuário sem mãos para interagir com um sistema de computador. Além disso, o aplicativo é gratuito para projetos sem fins lucrativos, a fim de garantir que os interessados desenvolvam conteúdos úteis para a humanidade, como por exemplo, o caso de universitários que estudam o uso do *Kinect* em cirurgias.

[64] Disponível em: https://softwareconnect.com/edi/extol-edi-integrator-for-ibm-i-eei/. Acesso em: 1 nov. 2021.
[65] Disponível em: https://docs.microsoft.com/en-us/previous-versions/windows/kinect/dn782041(v=ieb.10)?redirectedfrom=MSDN. Acesso em: 1 nov. 2021.

### 4.4.4 IMDI (ISLE)[66] – ISLE Metadata Initiative[67]

A ISLE Meta Data Initiative (IMDI) é um padrão de metadados que descreve recursos de linguagem multimídia e multimodal. O objetivo principal desse projeto é promover a acessibilidade e a disponibilidade das pesquisas linguísticas. O padrão fornece uma capacidade de interação entre sistemas, referente à estrutura de *corpus* pesquisáveis e descrições de recursos. Os metadados podem ser criados a partir de formulários na *Web* na interface de depósito ou criados com um editor de metadados CMDI[68], como Arbil, que posteriormente serão carregados na interface de depósito.

### 4.4.5 SW Arbil[69] – Editor geral de metadados

O Arbil é um editor geral de metadados, ferramenta *browser & organizer* para IMDI, CMDI e formatos semelhantes de metadados. Foi projetado para que possa ser usado *off-line* em locais remotos. Quanto aos dados, esses podem ser inseridos em qualquer estágio, em parte ou como um todo. Arbil é um aplicativo para organizar materiais de pesquisas e metadados relacionados a um formato apropriado para o arquivamento. No entanto, destaca-se que não há desenvolvimento ativo ou apoio para esse aplicativo.

### 4.4.6 ILEX[70] – Ferramenta para lexicografia da língua de sinais e análise de *corpus*

É uma ferramenta que estabelece as características encontradas na lexicografia empírica e na transcrição do discurso da língua de sinais. Ele suporta que o usuário realize a construção de léxico enquanto trabalha na transcrição de um *corpus*. Embora essa ferramenta tente atingir um nível de compatibilidade com outras ferramentas de anotação multimídia, ela oferece diversas características únicas[71], consideradas essenciais devido à natureza específica das línguas de sinais.

---

[66] Disponível em: https://archive.mpi.nl/forums/t/imdi-metadata-information/2933. Acesso em: 2 nov. 2021.
[67] Disponível em: https://www.mpi.nl/ISLE/. Acesso em: 2 nov. 2021.
[68] Disponível em: https://www.clarin.eu/content/component-metadata. Acesso em: 2 nov. 2021.
[69] Disponível em: https://archive.mpi.nl/forums/t/arbil-information-manuals-download/1045. Acesso em: 2 nov. 2021.
[70] Disponível em: https://www.researchgate.net/figure/The-iLex-Annotation-Software_fig2_312420117/download. Acesso em: 3 nov. 2021.
[71] Disponível em: http://ilex.sourceforge.net/. Acesso em: 19 fev. 2022.

Como, por exemplo, dentre outras características, ser: (i) multilíngue: a maior base de palavras que contém é o inglês, mas há uma quantidade de, pelo menos, uns 12 idiomas, e se pretende expandi-lo para outros idiomas, com buscas de dicionários *onlines;* (ii) internacionalizado: está em UTF-8[72] e busca lidar com o maior número de línguas possíveis, além de atender dialetos, como por exemplo, inglês britânico e inglês americano; (iii) *interlinked:* as palavras são interligadas dentro de cada idioma e, também, entre outros idiomas, por exemplo, a palavra estará com as suas variações dentro do próprio idioma e com as respectivas traduções para os outros idiomas; (iv) ser uma fonte aberta: lançada para todos usarem livremente.

### 4.4.7 MaxQDA[73] – *Software* para análise de dados qualitativos

MAXQDA é um *software* desenvolvido para análise de dados qualitativos, quantitativos e métodos mistos em pesquisas acadêmicas, científicas e comerciais. O *software* está disponível como uma aplicação universal para sistemas operacionais *Windows* e *macOS* e é desenvolvido pela empresa *VERBI Software*, em Berlim, na Alemanha. A ênfase em ir além da pesquisa qualitativa pode ser observada na presença de ferramentas estatísticas e na habilidade do *software* em lidar de forma relativamente rápida com uma grande quantidade de entrevistas. Está disponível para *download* e disponibiliza 30 dias para teste. Após esse período, caso o pesquisador decida dar continuidade ao uso da ferramenta, serão atribuídos valores, que podem variar conforme o revendedor.

### 4.4.8 OpenPose[74] – O primeiro sistema multitarefa em tempo real a detectar conjuntamente os pontos-chave do corpo humano (mão, face e pé)

Esse sistema detecta, em tempo real, pontos-chave do corpo humano, mão, face e pé, totalizando 135 pontos-chave em imagens únicas. Os autores desse sistema inovador são: Ginés Hidalgo, Zhe Cao, Tomas Simon, Shih-En Wei, Yaadhav

---

[72] É um tipo de codificação binária, que define um mapeamento entra *bytes* e texto. Disponível em: https://developer.mozilla.org/pt-BR/docs/Glossary/UTF-8. Acesso em: 19 fev. 2022.
[73] Disponível em: https://www.maxqda.com/brasil/software-analise-qualitativa. Acesso em: 4 nov. 2021.
[74] Disponível em: https://cmu-perceptual-computing-lab.github.io/openpose/web/html/doc/. Acesso em: 4 nov. 2021.

Raaj, Hanbyul Jooe Yaser Sheikh. É mantido por Ginés Hidalgo e Yaadhav Raaj. Segundo informações, na página, é possível instalar o programa a partir da fonte. Caso o usuário opte por não instalar ou escrever qualquer código, basta baixar para usar a versão mais recente do *OpenPose* no *Windows*. Esse sistema está disponível gratuitamente e pode ser redistribuído, desde que não seja para uso comercial.

### 4.4.9 Tolk – Recurso para voz não eletrônica

É um recurso para voz não eletrônica, destinado para intérpretes de língua de sinais, utilizado no *corpus* da NGT. Esse recurso é baseado no padrão de *Design Spring* para filtrar e traduzir dados de fontes de dados, como por exemplo, conexões TCP/IP, arquivos de texto, bancos de dados e serial (rs232, usb, BT). As aplicações típicas são manipular dados e gerenciar dispositivos como leitores RFID IPICO. Ele pode ser executado *online*, no provedor de hospedagem gratuita OnWorks, para estações de trabalho.

### 4.4.10 HamNoSys[75] – Sistema de Notação da Língua de Sinais de Hamburgo

É um sistema de transcrição fonético, que está em uso generalizado desde que sua versão original, na tradição dos sistemas baseados em *Stokoe*, foi publicada. A versão atual 4 leva em conta experiências práticas no uso do sistema, não só referentes a ASL, mas também em outras línguas de sinais. Vale mencionar, que esse sistema não se baseia na ortografia nacional diversificada dos dedos e, portanto, pode ser aplicada internacionalmente. Esse sistema também é uma base para uma série de controle de avatar e está disponível para *download* de forma gratuita

### 4.4.11 Handbrake[76] – Conversor de vídeos para MPEG

Segundo informações, na plataforma, o HandBrake é uma das melhores ferramentas em relação à modificação de vídeos. Há uma matriz quase infinita de ajustes possíveis, que vai do formato à codificação ou faixas de áudio. É um programa

[75] Disponível em: https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/hamnosys-97.html Acesso em: 4 nov. 2021.
[76] Disponível em: https://handbrake.fr/downloads.php. Acesso em: 4 nov. 2021.

multiplataforma e multitarefa em código aberto, licenciado em GPL, de conversão de arquivos de vídeo e DVD e Bluray para MPEG. O programa conta com uma farta biblioteca livre e está disponível aos interessados que desejam alterá-lo ou estudá-lo.

### 4.4.12 WORD[77] – Processamento de texto

É um programa de processamento de texto, projetado para a criação de documentos de qualidade profissional. Com as melhores ferramentas de formatação de documento, o *Word* ajuda a organizar e escrever os documentos com mais eficiência.

Após as descrições e as análises realizadas, referentes aos possíveis critérios utilizados na construção de *corpora* em línguas de sinais, bem como os possíveis recursos que podem colaborar com o desenvolvimento nas análises linguísticas dos pesquisadores, seguimos, agora, para o último capítulo da nossa proposição, que é, por fim, sugerir e apresentar uma proposta relativa aos critérios necessários para a construção de um *corpus* paralelo Libras-Português.

[77] Disponível em: https://support.microsoft.com/pt-br/office/tarefas-b%C3%A1sicas-no-word-87b3243c-b0bf-4a29-82aa-09a681999fdc#:~:text=O%20Microsoft%20Word%202010%20%C3%A9%20um%20programa%20de,organizar%20e%20escrever%20os%20documentos%20com%20mais%20efici%C3%AAncia. Acesso em: 5 nov. 2021.

# 5 A PROPOSTA: DIRETRIZES PARA A CONSTITUIÇÃO DE *CORPORA* PARALELOS LIBRAS-PORTUGUÊS

Neste capítulo, retomamos os aspectos principais que foram abordados nos capítulos anteriores, cujas informações foram tomadas como base para a proposta de um conjunto de diretrizes voltadas para a elaboração de um *corpus* paralelo, do qual tomem parte, de um lado, línguas de modalidade visual-gestual e, de outro, línguas orais. Para fins ilustrativos, nesta nossa apresentação, vamos considerar duas línguas, a Libras e o Português. O modelo, que ora apresentamos, vale-se, não somente do que já fora mencionado a respeito de um conjunto de diretrizes voltadas para o seu mapeamento, mas, também, de critérios de organização e estruturação de dados linguísticos que consideramos importantes para fins de análises linguísticas.

Por meio de elementos e critérios que identificamos e, nesse momento, registramos a respeito de *corpora* de línguas de sinais, buscamos responder, pelo menos, as três primeiras indagações que nos conduziram no decorrer desta pesquisa, quais sejam: (i) Quais seriam os critérios necessários para a constituição de um *corpus* paralelo entre línguas de sinais e línguas orais?; (ii) Que tipo de conteúdo deve prevalecer nesse *corpus*, a partir das análises que se pretende realizar?; e, mais especificamente, (iii) O que precisa ser considerado num *corpus* paralelo, em que, de um lado, o que tem são dados linguísticos das línguas de sinais e, de outro, os dados equivalentes das línguas orais, sob a ótica das teorias que fundamentam a LC?

Para a constituição de um *corpus* paralelo Libras-Português, que seja embasado na LC, alguns critérios importantes precisam ser levados em consideração, de modo a garantir que os dados registrados sejam, não apenas representativos, mas, sobretudo, afinados com os requisitos estabelecidos pelas análises linguísticas que o pesquisador pretenda realizar. Para que isso aconteça, faz-se necessário, primeiro, que os dados coletados sejam extraídos de fontes não controladas, ou seja, obtidos de produções espontâneas, tais como, jornais, revistas, artigos, entrevistas, dentre outras modalidades, pois, somente assim é possível garantir a isenção e a não manipulação dos dados utilizados pelo pesquisador para as suas análises. Além disso, muitos outros cuidados devem ser observados: um deles é de que as evidências empíricas que subjazem à formação desses dados sejam legíveis por máquina.

O que será apresentado nas próximas seções são apenas alguns critérios, que foram pensados a partir da aplicação de pressupostos teóricos preconizados pela LC.

Nesse sentido, importa lembrar que, como tal, não se trata de apresentarmos aqui uma relação exaustiva, mas, antes, tão somente um guia inicial de orientação, à qual, sem dúvida, poderão/deverão ser acrescidos muitos outros[78].

Ao buscar imprimir uma maior clareza para o que apresentaremos, optamos por separar esse guia por tópicos, começando pelos cuidados que precisam ser tomados durante a coleta de dados até o momento da disponibilização deles. A seguir, seguem as diretrizes.

## 5.1 COLETA DE DADOS

Consideramos que a etapa inicial para a construção de um *corpus* é a coleta de dados, uma vez que as informações devem ser coletadas criteriosamente com o propósito de pesquisas linguísticas, por meio de evidências empíricas, com o auxílio de ferramentas computacionais – algumas descritas no capítulo anterior. Nesse momento, vale lembrar que estamos sugerindo um *corpus* paralelo, que parte, primeiramente, da língua sinais para uma língua oral, nesse caso a Libras para o Português.

Colocamos em evidência a questão da direcionalidade do *corpus* (Libras-Português), uma vez que essa coleta partirá primeiramente da Libras, para depois serem transcritas as glosas e realizadas as respectivas traduções para o Português. Portanto, faz-se necessário um cuidadoso planejamento. Apresentaremos, a seguir, algumas diretrizes que consideramos importantes para que que a coleta de dados alcance êxito nas etapas que a sucedem:

(i) Refletir e estabelecer, primeiramente, o que se pretende pesquisar e analisar;
(ii) Estabelecer o público-alvo que será beneficiado com esse *corpus*;
(iii) Definir quais serão os materiais/conteúdos que serão utilizados no momento da gravação;

[78] N.A.: Vale mencionar que esse trabalho não se encerra aqui. Com efeito, tomando-o como ponto de partida, a nossa equipe de pesquisadores vem aprofundando o assunto, não só para atender as nossas necessidades, mas também com o objetivo de auxiliar outros pesquisadores/estudiosos que pretendam desenvolver seus próprios *corpora* linguísticos, notadamente quando se pretende estabelecer uma interface entre línguas orais e de sinais.

(iv) Organizar uma documentação que autorize o uso de imagem dos participantes, bem como esclarecer sobre a forma que esses dados serão aplicados na pesquisa e quais serão os meios de divulgação, respeitando questões éticas que um trabalho como esse impõe. Essa etapa nem sempre é simples, uma vez que é necessária a autorização dos participantes para tornar o *corpus* disponível;

(v) Produzir vídeos autênticos, de forma que a manifestação da língua ocorra de naturalmente no momento da gravação, com o intuito de que o *corpus* seja representativo aos estudos língua ou da variedade linguística, caso contrário esse *corpus* pode ser considerado artificial, portanto, não estaria condizente com os pressupostos da LC;

(vi) Selecionar informantes nativos da língua ou, dependendo do que se pretende pesquisar, poderão ser selecionadas pessoas fluentes na Libras e que pertencem à comunidade surda, por exemplo, um trabalho com profissionais tradutores e intérpretes de Libras/Português para análise de dados que competem aos estudos da tradução;

(vii) Ou ainda, podem-se criar sub*corpus*[79] destinados aos estudos das Libras, para aprendizes da língua;

(viii) Estabelecer quais ferramentas computacionais serão utilizadas, como as que apresentamos na seção anterior, ou ainda, desenvolver junto aos cientistas da computação uma ferramenta que atenda aos interesses das análises que serão realizadas;

(ix) Utilizar o ELAN como ferramenta, pois conforme mencionamos no decorrer desse trabalho, parece-nos bem interessante e é uma ferramenta computacional muito utilizada para a construção de *corpora* em línguas de sinais, além disso, conforme descrito na seção anterior, está disponível gratuitamente para *download*, em constantes atualizações para correções de *bugs*.

Quanto a esse momento, é extremamente importante que se estabeleça, antecipadamente, um plano de trabalho para que se chegue ao propósito das análises, determinada pelos pesquisadores, com mais agilidade. Após as sugestões

[79] Em um grande corpus há a possiblidade de serem criados vários subcorpora com diferentes temáticas/abordagem.

estabelecidas para o planejamento da coleta de dados, na próxima seção, apresentaremos orientações sobre a organização dos vídeos para a constituição do *corpus* em Libras, que poderão servir como base para que o pesquisador prossiga com os demais mecanismos para as análises que se pretende.

## 5.2 ORGANIZAÇÃO DE VÍDEOS

A etapa de organização de vídeos é bastante trabalhosa, pois requer uma grande quantidade de profissionais envolvidos, tecnologia avançada, além de uma quantidade considerável de informantes nativos/fluentes na língua de sinais, para que as amostras na Libras sejam representativas ao que se propõe. No decorrer da nossa pesquisa, observamos, também, que os *corpora* constituídos, que foram apresentados e descritos, recebem apoio financeiro, pois, deve-se considerar que para realizar um trabalho dessa dimensão, os custos são dispendiosos. Para essa fase do projeto, acreditamos ser importante:

> (i) Recrutar informantes surdos ou pessoas sinalizantes, pertencentes à comunidade surda, ou manifestações linguísticas envolvendo aprendizes da Libras. As produções deverão ser, impreterivelmente, criadas por humanos para que a manifestação da língua ou variedade linguística aconteça de maneira natural, ou seja, jamais utilizar-se de ferramentas computacionais, como por exemplo, avatar em 3D, para produções em Libras;
> (ii) Distribuir regionalmente os registros filmados, observando que assim como nas línguas orais, as línguas de sinais também têm suas variações linguísticas;
> (iii) Definir o conteúdo das gravações. Nessa etapa o mais importante é estabelecer a quantidade de vídeos e o tempo aproximado de cada um deles, a fim de que amostras sejam satisfatórias ao que se pretende analisar, considerando que não há necessidade de uma infinidade de vídeos que possam vir a atrapalhar as análises, ao invés de contribuir com o pesquisador. Assim, o arquivo deverá ter informações (metadados) para essa finalidade;
> (iv) Estabelecer se as gravações serão com base em entrevistas ou em produções de narrativas (espontâneas ou motivadas). Em relação a esse tipo de conteúdo, é necessária uma discussão sobre a autenticidade desses dados;

(v) Utilizar outros conteúdos, como pesquisas variadas (conversações livres, piadas, contação de histórias ou outros gêneros textuais);

(vi) Possibilitar pesquisas de sinais relacionados a diferentes temas/área (saúde, educação, cultura, política etc.);

(vii) Preparar um estúdio, não amador, com os recursos necessários para manter a qualidade nas gravações, com cores neutras de fundo. A esse respeito notamos que, predominantemente, a cor azul é bastante utilizada nos estúdios de gravações;

(viii) Distribuir as câmeras em diferentes posições, com o intuito de capturar o movimento das mãos em diferentes ângulos, dependendo do que se pretende analisar e com dispositivos de iluminação necessária para manter a qualidade dos vídeos;

(ix) Dispor de *drives* com uma grande capacidade de memória e que comportem vídeos. Para isso, faz-se necessário tecnologias de ponta;

(x) Organizar uma equipe de técnicos e/ou profissionais para produção e edição de vídeos.

Após propormos algumas diretrizes para a organização, os registros e o armazenamento de vídeos em Libras, o que ainda está em pauta e que pode vir a ser discutido em trabalhos futuros é a questão da autenticidade dos conteúdos no momento das gravações, pois, ao analisarmos os *corpora*, no capítulo anterior, observamos que eles têm em comum as produções de entrevistas espontâneas ou motivadas. Esse pode ser um fator para as próximas discussões, pois, ao se constituírem em produções motivadas e organizadas com o propósito de se criar um *corpus*, poderiam estar descaracterizando uma produção natural da língua. Por isso, acreditamos que essa ainda será uma questão levantada pelos pesquisadores da LC.

Nas próximas seções, buscaremos apresentar algumas sugestões sobre a disposição, além de informações referente aos dados que podem facilitar e permitir análises linguísticas de um *corpus* paralelo Libras-Português.

## 5.3 TRANSCRIÇÃO DE LÍNGUA DE SINAIS[80]

Após a edição dos vídeos em Libras, sugerimos que os pesquisadores iniciem o processo de transcrição. O processo de transcrição se refere ao momento que uma equipe, geralmente surdos nativos e estudiosos da Libras, inicia as análises dos vídeos, realizando as respectivas transcrições em glosas[81.]. No programa ELAN, por exemplo, verificamos uma ferramenta que possibilita acelerar ou diminuir o ritmo do vídeo, para então ir transcrevendo, em trilhas, as glosas que se referem ao vídeo em Libras, como podemos visualizar na seção anterior, onde descrevemos o ELAN e o ilustramos. Também serão coladas em trilhas paralelas ao vídeo as respectivas traduções na língua portuguesa. Para esse momento, consideramos:

> (i) Usar glosas para a transcrição dos sinais manuais da Libras, encontrados nos vídeos, por meio de trilhas, com o intuito de facilitar o acesso aos sinais para as análises;
> (ii) Optar pelo uso do Identificador de Sinais (ID)[82], que "é uma ferramenta que disponibiliza os nomes dados aos sinais para as glosas utilizadas nos sistemas de transcrição" (QUADROS, 2016, p. 24). No entanto, mais recentemente esse material foi transferido e está disponível para consulta no *Libras signbank*[83] (QUADROS, 2019), sendo que nessa plataforma é possível ter acesso a dicas de transcrições, por meio de um tutorial[84].
> (iii) Criar trilhas para identificar cada um dos enunciadores, a fim de analisar as manifestações linguísticas de cada um deles;

80 Considerando que as línguas de sinais não apresentam, até o momento, um sistema de escrita consolidado entre os seus usuários, o que dificulta o processo de transcrição das línguas de sinais, os pesquisados buscam recorrer ao uso de glosas.
81 Quadros (2016) apresenta algumas convenções propostas pelos autores Nonhebel, Crasborn e Van Der Kooij (2004), esses autores propuseram a criação de trilhas independentes para a mão direita e mão esquerda, já utilizando-se do ELAN, com "um sistema de anotação que permite a inclusão de trilhas para cada aspecto transcrito associado diretamente ao vídeo" (QUADROS, 2016, p. 17).
82 Conforme Quadros (2016), o Identificador de Sinais está disponível de forma aberta e gratuita na página: http://www.idsinais.libras.ufsc.br.
83 Disponível em: http://signbank.libras.ufsc.br/. Acesso em: 02 nov. 2021.
84 Disponível em: file:///C:/Users/kelin/AppData/Local/Temp/2019%20TUTORIAL_CORPUS%20transcric%CC%A7a%CC%83o-15835183035e62925f439b03.92987599.pdf. Acesso em: 5 nov. 2021.

(iv) Organizar trilhas, uma para transcrever os sinais realizados com a mão direita e outra para identificar os sinais realizados com a mão esquerda, utilizando-se das glosas;

(v) Produzir trilhas para as traduções realizadas na língua portuguesa, considerando que buscamos critérios para um *corpus* paralelo Libras-Português;

(vi) Elaborar trilhas para estabelecer as análises ou os comentários, uma realizada pelos transcritores de Libras e/ou Português e outra realizada pelos tradutores de Libras e/ou Português.

Identificamos, nesta seção, alguns mecanismos que poderão auxiliar nas análises linguísticas da Libras em interface ao Português, como a produção de glosas, com base nos vídeos produzidos, por meio de trilhas para transcrição da Libras, dispondo, também, de trilhas referentes à tradução para o Português. Assim, após definir as trilhas necessárias para as análises, podemos realizar as anotações/etiquetas com informações sobre cada sinal, sejam elas linguísticas ou extralinguísticas. Na próxima seção, buscaremos sugerir critérios/orientações referentes às anotações, sejam estruturais ou linguísticas.

## 5.4 ANOTAÇÕES ESTRUTURAIS E ANOTAÇÕES LINGUÍSTICAS

Após a definição das trilhas, podemos seguir para o processo de anotações, estruturais e/ou linguísticas, que compõe um conjunto de informações, uma etapa que também exige trabalho em equipe, devido a sua complexidade. Segundo Aluísio e Almeida (2006), resumidamente, a anotação estrutural se refere à “marcação de dados externos e internos dos textos”. Quanto aos dados externos, entendem que são as documentações do *corpus*, que incluem metadados, como: tamanho do arquivo, autoria, tipo de arquivo, informação sobre a distribuição do *corpus*. Quanto aos dados internos, consideram: marcação da estrutura geral (informações sobre o *corpus*, títulos etc.) e marcação da estrutura de subparágrafos, ou seja, elementos que são de interesse linguístico, como sentenças, citações, palavras, no nosso caso, por exemplo, referimo-nos aos sinais. Vale mencionar, que essas informações facilitam a geração de um *subcorpus*, como selecionar um determinado autor, a época, o gênero, dentre outras. Em relação às anotações linguísticas, conforme Aluísio e Almeida

(2006), refere-se aos níveis linguísticos, tal como: “níveis morfossintático, sintático, semântico, discursivo etc.”. Em relação às anotações, é possível:

> (i) Estabelecer identificação para os arquivos de anotações referentes à Libras, disponíveis nos vídeos e, também, realizar as anotações referente às trilhas de glosas e tradução para o Português. Essas anotações/etiquetas são informações adicionadas às trilhas, com o intuito de facilitar a busca baseada nos interesses dos pesquisadores, ao definir suas análises linguísticas;
> (ii) Incluir arquivos específicos para análises de substantivos e verbos;
> (iii) Administrar a frequência dos sinais, após as informações estarem etiquetadas;
> (iv) Anotar as marcações da classe gramatical, para qual um item lexical pertence;
> (v) Criar uma trilha específica com informações/anotações sobre a direção do olhar no momento da sinalização;
> (vi) Etiquetar com informações sobre os parâmetros[85] linguísticos na Libras, quanto: à orientação/direcionalidade da palma da mão; à configuração de mão; à localização; ao movimento; à expressão facial e/ou corporal, no momento que cada sinal é realizado;
> (vii) Produzir anotações a respeito da marcação para mudança de movimento/posição no espaço, conhecido como “*Role Shift*”[86];
> (viii) Estabelecer anotações para a mudança de informante no momento da comunicação;
> (ix) Gerar trilha com anotações a respeito das informações extralinguísticas;
> (x) Estabelecer arquivos anotados com glosas, ou seja, com as transcrições dos vídeos em Libras;
> (xi) Criar arquivos anotados com a tradução para o Português.

85 Quanto aos parâmetros da Libras, Ferreira Brito (1995) identificou os seguintes: Configuração das mãos (CM) que são as diversas formas que a mão assume para realização de um sinal; Movimento (M) que é um parâmetro complexo, envolvendo uma grande quantidade de formas e direções; Ponto de Articulação (PA) que refere-se ao espaço em frente ao corpo ou uma região do próprio corpo, onde os sinais são articulado; Orientação da Mão (OR) que refere-se a direção para a qual a palma da mão aponta na produção do sinal e; segundo Quadros e Karnopp (2004), temos ainda as Expressões não-manuais (ENM), que referem-se ao movimentos da face, dos olhos, da cabeça ou do tronco; contudo, a combinação desses parâmetros constituem um sinal e podem ser comparadas com fonemas ou morfemas nas línguas de sinais.

86 *Role Shift* é um mecanismo nas línguas de sinais, marcado por uma mudança na posição do corpo ou direção do olhar, para estabelecer uma mudança pronominal pelo enunciador.

Para além dessas características de anotações/etiquetas, pode-se acrescentar mais informações, conforme as necessidades de análises realizadas pelo pesquisador. Essa etapa, conforme as pesquisas que observamos, é bastante trabalhosa, demanda tempo, e os custos também são onerosos. Além disso, é difícil estabelecer critérios, pois demanda flexibilidade, uma vez que os interesses dos pesquisadores são diversos. Na próxima seção, apresentaremos informações pertinentes aos metadados produzidos pelos pesquisadores em *corpus* linguístico.

## 5.5 ESTRUTURAÇÃO DOS METADADOS[87] PARA UM *CORPUS* LINGUÍSTICO

Para a produção dos metadados, que em outras palavras significa dados sobre outros dados, ou seja, informações sobre os dados já estabelecidos, usam-se programas para a descrição das informações, alguns deles identificados e descritos nos *corpora* apresentados na seção anterior. Esses, quando combinados a outros programas, por exemplo, com o ELAN, possibilitam a criação das trilhas de transcrições e informações linguísticas e extralinguísticas, sendo que os dados e metadados oportunizam evidências ousadas e inovadoras para análises linguísticas nas línguas de sinais. Em relação aos metadados, sugerimos:

> (i) Criar etiquetas com anotações diversas acerca dos informantes, como: comunidade surda, região, escolaridade, idade, gênero, dentre outros;
> (ii) Informar, por meio das etiquetas, sobre os conteúdos, ou seja, em relação ao tipo de material utilizado para as gravações;
> (iii) Produzir etiquetas com informações referentes ao projeto: o nome, o idioma ou a variedade linguística a ser trabalhada, bem como a metodologia e as análises que serão aplicadas;
> (iv) Elaborar etiquetas sobre a mídia, informando sobre o formato e o tipo de arquivo;
> (v) Gerar etiquetas com informações referentes ao arquivo de anotação, ao arquivo de mídia e para o tipo de anotação concluída.

87 Os metadados referem-se às informações sobre os vídeos/mídia e arquivos de anotação.

Por fim, neste capítulo, buscamos apresentar e sugerir alguns mecanismos e algumas diretrizes que podem auxiliar estudiosos e pesquisadores que tenham interesse em construir um *corpus* paralelo Libras-Português, com base nas concepções da LC. Para tanto, levamos em consideração os argumentos e os critérios indicados pela LC e, também, o que identificamos em relação às descrições e às análises dos *corpora* investigados no capítulo 4. Nesse contexto, nosso propósito foi expressar e reforçar a importância de se construir *corpora* com dados autênticos e naturais da língua e/ou variedade linguística, a fim de torná-los confiáveis e representativos às análises linguísticas na Libras em interface com a Língua Portuguesa.

## 6 CONSIDERAÇÕES FINAIS

Ao considerarmos os estudos realizados, relativos às premissas estabelecidas pela LC e, também, as descrições dos *corpora* do Brasil e de outras partes do mundo, aqui realizadas, foi possível notar que todos esses *corpora* apresentados em língua de sinais são compostos, basicamente, por vídeos com produções realizadas por pessoas surdas, usuárias das línguas de sinais, e todos partem de projetos e financiamentos para a constituição da documentação. Conforme Quadros (2019), a metodologia utilizada nesses *corpora,* em geral, é motivada pela documentação desenvolvida, primeiramente, na Língua de Sinais Australiana e, posteriormente, aperfeiçoada pelas Línguas de Sinais Britânica e Alemã, com o intuito de disponibilizar, publicamente, esses dados para análises linguísticas nas línguas de sinais.

Em linhas gerais, em relação ao tipo de conteúdo que deve prevalecer nesses *corpora,* os dados nos mostraram que os conteúdos podem variar significativamente, ou seja, alguns são direcionados para temáticas das esferas acadêmica, saúde, jurídica, dentre outras; e outros são mistos, isto é, compõem uma gama variada de vídeos em línguas de sinais, envolvendo as mais diversas temáticas, como piada, literatura, diálogos, entrevistas etc. Vale dizer que um *corpus* cuja temática é direcionada pode auxiliar mais diretamente o pesquisador, conforme a necessidade do que se pretende encontrar.

Ainda, ao que concerne ao conteúdo, é válido mencionar que os *corpora* constituídos em Libras ou nas demais línguas de sinais, quando possuem entrevistas/conteúdos motivados ou, até mesmo, quando as gravações são direcionadas com o intuito de constituir um *corpus* para fins de análises linguísticas, podem estar colocando em jogo a questão da autenticidade dos dados para a LC. Pois, um dos pressupostos determinados por ela, visto no decorrer desta pesquisa, é que as análises devem partir da manifestação natural da língua, em contextos espontâneos de comunicação entre os falantes de uma comunidade linguística e/ou variedade linguística.

Em contrapartida, consideramos importante refletir sobre a validação desses *corpora*, ou seja, mesmo aqueles que vêm sendo criados de forma motivada ou espontânea, em estúdios, com câmeras posicionadas estrategicamente e com falantes das línguas de sinais, entendemos serem úteis e representativos aos

pesquisadores, conforme os seus objetos de investigação. Por exemplo, acreditamos que se o objetivo da investigação contemplar aspectos fonológicos ou morfológico de um sinal, esses *corpora* poderão servir como base para esse tipo de análise linguística, uma vez que os sinais poderão ser analisados com mais detalhes em relação aos seus parâmetros, ou seja, em relação à formação de um sinal (configuração de mão, movimento, ponto de articulação etc.), mesmo que a questão da autenticidade desses *corpora* seja um ponto a ser discutido.

Em relação à dimensão de um *corpus* é de comum acordo, entre os autores que discutem sobre a LC, que quando se fala em tamanho*,* não existe uma regra determinada. Assim, compactuamos que é necessário que as amostras coletadas sejam representativas à língua estudada e estejam de acordo com as propostas de análise conduzidas pelo pesquisador.

Dessa forma, acreditamos que os *corpora* em línguas de sinais, que têm sido criados no Brasil e em outras partes do mundo, podem contribuir, positivamente, com alguns pesquisadores, de forma que nos parece ser importante ou até mesmo necessário um novo olhar dos pesquisadores da LC sobre a constituição e a representatividade de *corpora* em línguas de sinais.

Mesmo diante de grandes desafios, consideramos que em breve teremos bons resultados e dados cada vez mais autênticos para a constituição de *corpora* em Libras, uma vez que ela vem sendo utilizada com mais frequência nos diversos meios de comunicação, como por exemplo, jornais, programas políticos, provas oficiais (ENEM, Prova Brasil, concursos e outros). Com isso, ponderamos que há possibilidades reais de ampliar a criação de bancos de dados que sejam autênticos e representativos para as análises linguísticas da Libras, de acordo com os pressupostos estabelecidos pela LC.

Enfim, levando-se em consideração os aspectos descritos, observados e analisados no decorrer deste trabalho, na última seção, buscamos identificar e sugerir, didaticamente, alguns critérios para auxiliar e contribuir com os pesquisadores interessados em construir um *corpus* paralelo Libras-Português, conforme a proposta principal desta pesquisa. Todavia, notadamente, não se trata de uma tarefa fácil, mas que demanda uma gama de profissionais, dentre eles, estudiosos surdos, tradutores e intérpretes de Libras-Português, linguistas, equipe técnica para a produção e edições dos vídeos, informantes dispostos a colaborar com as gravações para

análises e que, sobretudo, concordem em tornar os dados coletados disponíveis para consultas.

Em virtude dos aspectos analisados, esses fatores, além de demandar um exaustivo trabalho e um tempo considerável dedicado para esse propósito, também exigem altos investimentos, como foi possível observar nos *corpora* das línguas de sinais descritos e disponíveis no capítulo 4, pois, em geral, recebem apoio financeiro e requerem tecnologia de última geração. Por isso, para muito além de se ter uma equipe disposta a trabalhar com a construção de um *corpus* paralelo Libras-Português, é incontestável a necessidade de parcerias que estejam receptivas a altos investimentos, visto que é necessária uma estrutura adequada para a realização de um trabalho como esse, que tem se mostrado, cada dia mais, imprescindível para auxiliar nas análises linguísticas, seja nas línguas orais, seja nas línguas de sinais.

## REFERÊNCIAS

AIJMER, K.; ALTENBERG, B. Advances in Corpus-based Contrastive Linguistics: Studies in Honour of Stig Johansson. **John Benjamins Publishing CompanyAmsterdam**, Philadelphia, 2013.

ALUÍSIO, S. M.; ALMEIDA, G. M. B. O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa lingüística. **Calidoscópio**, São Leopoldo, RS, v. 4, p. 156–178, 2006.

ANDERMAN, G.; ROGERS, M. (Eds). **Incorporating Corpora.** The Linguist and Translator. Clevedon: MultilingualMatters, 2008.

BAKER, Mona. Corpus in Translation Studies: an overview and some suggestions for future research. **Target**, Amsterdam, v. 7, n. 2, 1995.

BERBER SARDINHA, A. P. Processamento Computacional do Português. In: SIMPÓSIO, 9. São Paulo, 1999. **Anais**... São Paulo: InPLA – PUCSP, 1999.

BERBER SARDINHA, T. Linguística de Corpus: histórico e problemática. **D.E.L.T.A.,** São Paulo, v. 16, n. 2, p. 323-367, 2000a.

BERBER SARDINHA, T. **Linguística de corpus**. São Paulo: Manole, 2004.

BERBER SARDINHA, T. **O que é um corpus representativo?** São Paulo: DIRECT papers 44, 2000b. Disponível em: https://doczz.com.br/doc/385196/o-que-%C3%A9-um-corpus-representativo.---lingu%C3%ADstica-aplicada-.... Acesso em: 16 out. 2021.

BRASIL. **Lei 10.436/02.** Reconhece a LIBRAS como Língua Oficial no Brasil. Brasília, DF. Disponível em: http://www.planalto.gov.br/ccivil_03/Leis/2002/L10436.htm. Acesso em: 4 out. 2017.

BRASIL. **Decreto nº 5.626 de 22 de dezembro de 2005**. Regulamenta a Lei nº 10.436 de 24 de abril, que dispõe sobre a Língua Brasileira de Sinais. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2005/decreto/d5626.htm. Acesso em: 23 mar. 2019.

CAMARGO, Diva Cardoso de; ROCHA, Celso Fernando; PAIVA, Paula Tavares Pinto (Org.). **Pesquisas em estudos da tradução e corpora eletrônicos no Brasil.** São Paulo: Editora Unesp, 2012. Disponível em: http://hdl.handle.net/11449/113720. Acesso em: 18 out. 2021.

CORPUS. **Dicionário Online de Português.** Porto: 7Graus, 2018. Disponível em: https://www.dicio.com.br/corpus/. Acesso em: 22 mar. de 2019.

CORPUS. **Dicionário Online de Inglês.** Cambridge Academic Content Dictionary. Cambridge University Pre, United Kingdom, 2019. Disponível em:

https://dictionary.cambridge.org/pt/dicionario/ingles/corpus?q=corpus+. Acesso em: 23 mar. 2019.

CORPUS. **Merriam-Webster Online Dictionary.** Incorporated, 2019. Disponível em: [https://www.merriam-webster.com/dictionary/corpus#synonyms]. Acesso em: 22 mar. 2019.

ELAN EUDICO. **LinguisticAnnotator.** Palhoça, SC: SEPEI, 2014. Disponível emhttps://eventoscientificos.ifsc.edu.br/index.php/sepei/sepei2014/paper/viewFile/406/526. Acesso em: 16 out. 2021.

**ELAN (VERSÃO 6.3) [SOFTWARE DE COMPUTADOR].** Nijmegen: Instituto Max Planck de Psicolinguística, 2002. Disponível em: https://archive.mpi.nl/tla/elan. Acesso em: 20 fev. 2022.

FERREIRA-BRITO, L. **Por uma gramática de Língua de Sinais.** Rio de Janeiro: Tempo Brasileiro, 1995.

FONSECA, J. J. S. **Metodologia da pesquisa científica**. Fortaleza: UEC, 2002.

FROMM, Guilherme. O uso de corpora na análise linguística. **Revista Factus**, São Paulo, v. 1, n. 1, p. 69-76, 2003.

GALISSON, R.; COSTE, D. **Dicionário de didática das línguas.** Coimbra: Livraria Almedina, 1983.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

GRANGER, S. L. J.; PETCH-TYSON, S. (Eds.). **Corpus-based Approaches to Contrastive Linguistics and Translation Studies**. Amsterdam: Rodopi. 2003.

HOEY, M. From concordance to text structure: new uses for computer corpora. In: LEWANDOSWKA-TOMASZCZYK, B.; MELIA, P. J. (Orgs.). **PALC'97 – Practical Applications in Language Corpora.** Lodz: Lodz University Press, 1997.

KENNEDY, G. D. **An introduction to corpus linguistics.** London, England: Studies in language and linguistics, 1998.

KILGARRIFF, A.; GREFENSTETTE, G. Introduction to the Special Issue on Web as Corpus. **Computational Linguistics**, Brighton, v. 29. n.3, 2003.

**LETRAS LIBRAS.** Florianópolis, [s.d]. Disponível em: https://libras.ufsc.br/libras-distancia/. Acesso em: 10 jul. 2018.

MCCARTHY, M.; O'KEEFFE, A. Historical perspective: what are corpora and how have they evolved? In: MCCARTHY, M.; O'KEEFFE, A. The Routledge handbook of corpus linguistics. By Routledge: 2 Park Square, Milton Park, Abingdon, OX14 4RN, 2010.

McCLEARY, L.; VIOTTI, E.; LEITE, T. A. Descrição das línguas sinalizadas: a questão da transcrição dos dados. **Alfa**, São Paulo, v. 54, n.1, p. 265-289, 2010.

McENERY, T.; WILSON, A. **Corpus linguistics**. Edinburgh, Edinburgh University Press, 1996.

MURAKAWA, C. A. A. Tradição lexicográfica em língua portuguesa. In: OLIVEIRA, A. M. P. P.; ISQUERDO, A. N. (Orgs.). **As ciências do léxico:** lexicologia, lexicografia e terminologia. 2. ed. Campo Grande: Ed. UFMS, 2001. p. 153-159.

MURAKAWA, C. A. A. **Lexicógrafo da língua portuguesa.** Araraquara: Laboratório Editorial FCL/UNESP; São Paulo: Cultura Acadêmica Editora, 2006.

OLOHAN, M. **Introducing Corpora in Translation Studies.** London: Routledge, 2004.

QUADROS. R. M. (Org.). **Estudos Surdos I.** Petrópolis, RJ: Editora Arara Azul, 2006. Disponível em: http://www.editora-araraazul.com.br/EstudosSurdos.php. Acesso em: 21 jun. 2019.

QUADROS, R. M.; KARNOPP, L. B. **Língua de Sinais Brasileira**: estudos linguísticos. Porto Alegre: ARTMED, 2004.

QUADROS, R. M.; SOUZA, S. X. Aspectos da tradução/encenação na língua de sinais brasileira para um ambiente virtual de ensino: práticas tradutórias do curso de Letras Libras. In: QUADROS, R. M. (Org.). **Estudos Surdos III.** Petrópolis: Editora Arara Azul, 2008.

QUADROS, R. M.; STUMPF, M.; OLIVEIRA, J. Avaliação de Surdos na Universidade. In: HEINING, O.; FRONZA, C. (Orgs.). **Diálogos entre linguística e educação.** Blumenau: Edifurb, 2011.

QUADROS, R. M. A transcrição de textos do Corpus de Libras. **Revista Leitura**, Alagoas, v. 1, n. 57, p. 8–34, 2016.

QUADROS, R. M. Línguas de Sinais: abordagens teóricas e aplicadas A transcrição de textos do Corpus de Libras. **Revista Leitura**, Alagoas, v. 1, n. 57, p. 8-34, jan./jun. 2016.

QUADROS, R. M. **Língua de Herança:** Língua Brasileira de Sinais. Porto Alegre: Penso, 2017.

QUADROS, R. M.; SCHMITT, D.; LOHN, J. T.; LEITE, T. A. **Corpus de Libras.** Florianópolis, [s.d]. Disponível em: http://corpuslibras.ufsc.br/. Acesso em: 30 jun. 2018.

QUADROS, R. M. Tecnologia para o estabelecimento de documentação de língua de sinais. In: CORRÊA, Y.; CRUZ, C. R. (Orgs.). **Língua Brasileira de Sinais e Tecnologias Digitais.** Porto Alegre: Penso, 2019.

RODRIGUES, C. H. Translation and signed language: highlighting the visual-gestural modality. **Cad. Trad.,** Florianópolis, v. 38, n. 2, p. 294-319, maio/ago. 2018.

SANCHEZ, A. et al. (Orgs.) **CUMBRE – Corpus Linguistico del Espanol Contemporaneo –** Fundamentos, Metodologia, y Aplicaciones. Madrid: SGEL, 1995.

SINCLAIR, J. Corpus and Text - Basic Principles. In: M. WYNNE (Ed.). **Developing Linguistic Corpora:** a Guide to Good Practice. Oxford: Oxbow Books, 2005. p. 1-16. Disponível em: http://ahds.ac.uk/linguistic-corpora/. Acesso em: 10 jul. 2020.

SHEPHERD, T. M. G.; BERBER SARDINHA, T.; PINTO, M. V. (Orgs.). **Caminhos da Linguística de Corpus.** Campinas, SP: Mercado de Letras, 2012.

SILVEIRA, D. T.; CÓRDOVA, F. P. A pesquisa científica. In: GERHARDT, T. E.; SIVEIRA, D. T. **Métodos de pesquisa.** Porto Alegre: Editora da UFRGS, 2009

STOKOE, W. **Sign and Culture:** a Reader for Students of American Sign Language. Maryland: Linstok Press, 1960.

TAGNIN, S. E a Linguística de Corpus vai desbravando novos horizontes. In: FINATTO, M. J. B. et al. (Orgs.) **Linguística de Corpus:** perspectivas. Porto Alegre: Instituto de Letras - UFRGS, 2018.

TRASK, R. L. **Dicionário de Linguagem e Lingüística**. São Paulo: Contexto, 2004.

VERAS, E. C. **Procedimentos metodológicos para a compilação de um corpus de língua de sinais a partir da rede:** reflexões com base em um corpus piloto de gêneros na plataforma youtube. 2014. 183 f. Dissertação (Mestrado em Linguística) – Instituição de Ensino, Universidade Federal de Santa Catarina (UFSC), Florianópolis, 2014.

ZWITSERLOOD, I. Meaning at the feature level in sign languages. The case of name signs in Sign Language of the Netherlands. In: KAGER, R.; GRIJZENHOUT, J.; SEBREGTS, K. (Ed.). **Where the Principles Fail.** A Festschrift for WimZonneveld on the occasion of his 64th birthday.Utrecht: Utrecht Instituteof Linguistics, 2014. p. 241-251.